Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney

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

Abstract. In this paper, we employ a zero-order local deformation model to model the visual variability of video streams of American sign language (ASL) words. We discuss two possible ways of combining the model with the tangent distance used to compensate for affine global transformations. The integration of the deformation model into our recognition system improves the error rate on a database of ASL words from 22.2% to 17.2%.

1 Introduction

In sign language recognition, as in other disciplines of pattern recognition, a considerable number of errors are due to the variability of the input signal. Each signer may utter a word differently, depending on his individual signing style or the predecessor and successor of the uttered word. Therefore, a large visual variability of utterances for each word exists. To model the variability of utterances, the tangent distance (TD) [1, 2] and the image distortion model (IDM) [3, 4] can be used to account for global and local variations, respectively.

The BOSTON50 database is a publicly available database in which the words are signed with high visual variability. In this paper we present experiments on this database using Hidden Markov Models (HMM) in a nearest neighbor manner. In [5] we have shown that the error rate using TD instead of the Euclidean distance in this setup is 22.2%.

Although tangent distance compensates for global affine transformations, it is sensitive to local image deformations. Therefore, we apply two different ways to combine the image distortion model with the tangent distance to compensate for local deformations. This combination enables the classifier to compensate for both local and global transformations. Using this combination, the error rate is considerably reduced from 22.2% to 17.2%, which is an improvement of 23 percent relative.

Our system is designed to recognize sign language words using appearance-based features extracted directly from standard cameras. This means that the system works without any explicit segmentation or tracking of the hands. Thus, the recognition can be expected to be more robust in cases where tracking and segmentation are difficult.

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

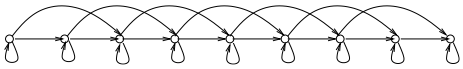

Fig. 2. The topology of the employed HMM.

2 Database and Extracted Features

The BOSTON50 database was created from the database of ASL sentences published by the National Center for Sign Language and Gesture Resources at Boston University[1]. It consists of 483 utterances of 50 American sign language words. The movies were recorded at 30 frames per second and the size of the frames is 195×165 pixels. The sentences were signed by three different signers, one man and two women. The signers were dressed differently. The frames of two cameras were used. One of the cameras shows a front view of the signer, the other one shows a side view. Sample images of the different views and the signers are shown in Figure 1. These two views are merged into one feature vector. According to the experiments reported in [6], the features of the front and side cameras are weighted with 0.38 and 0.62, respectively.

The feature vectors we have used consist of combinations of down-sampled original images which are multiplied by binary skin-intensity images and vertical and horizontal derivatives of these images. The images are down-scaled to 13×11 pixels and multiplied by a binary skin-intensity image. Afterwards, the derivatives are computed by applying Sobel filters. Thus, the feature vector x_t is between 143 and 3×143=429 dimensional. Note that the multiplication with the binary skin-intensity image is not an explicit segmentation step because it is a simple pixel-based transformation similar to e.g. Sobel filters.

3 Decision Process

The decision making of our system employs HMMs to recognize the sign language words. This approach is inspired by the success of the application of HMMs in speech recognition [7] and also most sign language recognition systems [8–11]. The recognition of sign language words is similar to spoken word recognition in the modelling of sequential samples. The topology of the HMM used is shown in Figure 2. There is a transition loop at each state and the maximum allowed transition distance is two, which means that one state, at most, can be skipped.

In [6] we presented a nearest neighbor approach using HMMs for recognition of sign language. The decision rule used to classify an observation sequence

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

$x_1^T = x_1,\ldots, x_t,\ldots x_T$ in this approach is

$$r(x_1^T) = \arg\max_w \left\{ \max_{n=1,\ldots,N_w} \left\{ Pr\left(u_{wn}|x_1^T\right)\right\}\right\}$$
$$= \arg\max_w \left\{ \max_{n=1,\ldots,N_w} \left\{ Pr\left(u_{wn}\right) \cdot Pr\left(x_1^T|u_{wn}\right)\right\}\right\},$$

where u_{wn} is the n-th utterance of word w and $Pr(u_{wn})$ is the a-priori probability of an utterance, uniformly distributed, and can thus be neglected. $Pr\left(x_1^T|u_{wn}\right)$ is approximated using maximum approximation to be

$$Pr\left(x_1^T|u_{wn}\right) = \sum_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1},u_{wn}) \cdot Pr(x_t|s_t,u_{wn}) \right\}$$
$$\cong \max_{s_1^T} \left\{ \prod_{t=1}^{T} Pr(s_t|s_{t-1},u_{wn}) \cdot Pr(x_t|s_t,u_{wn}) \right\}.$$

Here, the transition probability $Pr(s_t|s_{t-1},u_{wn})$ is uniformly distributed, and the emission probability $Pr(x_t|s_t,u_{wn})$ is a Gaussian with $\mu_{s_t} = u_{wnt}$, and a diagonal covariance matrix with $\sigma_d^2 = \sigma^2 = 0.1$:

$$Pr(x_t|s_t,u_{wn}) = \frac{1}{\sqrt{2\pi\sigma^2}^D} \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{td}-\mu_{s_td})^2}{\sigma^2}\right).$$

The feature vector x_t is a down-sampled image at time t. Therefore, the sum $\sum_{d=1}^{D}(x_{td}-\mu_{s_td})^2/\sigma^2$ is the distance between the observation image at time t and the prototype image μ_{s_t} of the state s_t that is scaled by the variances σ^2. This scaled distance can be replaced by other distance functions like the TD or IDM distance, which we introduce in the following section.

As the number of utterances in the database for each word is not large enough to have a separate training and test set, the experiments have been performed using leaving-one-out. That is, one observation is classified using all the other ones as training observations and this is repeated for all utterances from the database.

4 Invariant Distances

Because of the visual variability of utterances of each word, invariance of distance functions is an important aspect in sign language recognition. An invariant distance measure ideally takes into account transformations of the patterns, yielding small distances for patterns which differ by a transformation that does not change the class-membership. We briefly describe the *tangent distance* and the *image distortion model* which compensate for global and local transformations respectively.

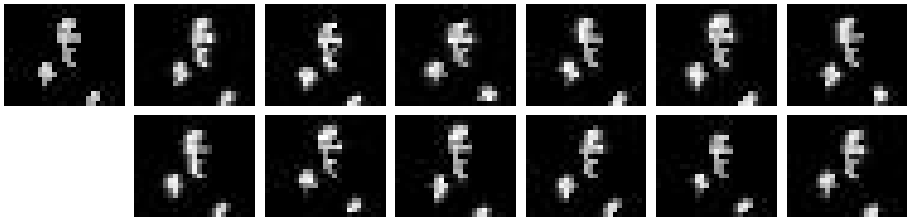

Fig. 3. Example of first-order approximations of affine transformations. (Left to right: original image, ± horizontal translation, ± vertical translation, ±axis deformation, ± diagonal deformation, ± scale, ± rotation)

4.1 Overview of Tangent Distance

Let $\mu \in \mathbb{R}^D$ be a class specific prototype pattern, and $\mu(\alpha)$ denote a transformation of μ that depends on a parameter L-tuple $\alpha \in \mathbb{R}^L$. Here we assume that the transformation does not change class membership for small α. Now, the set of all transformed patterns is a manifold $\mathcal{M}_\mu = \{\mu(\alpha) \mid \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D$ in pattern space. The distance between two patterns can then be defined as the minimum distance between the pattern x_t and the manifold $\mathcal{M}_\mu$ of a class specific prototype pattern μ. However, the calculation of this distance is a hard non-linear optimization problem in general. The manifold can be approximated by a tangent subspace $\widehat{\mathcal{M}_\mu}$. The tangent vectors μ_l that span the subspace are the partial derivatives of $\mu(\alpha)$ with respect to α_l. Thus, a first-order approximation of $\mathcal{M}_\mu$ is obtained as

$$\widehat{\mathcal{M}_\mu} = \{\mu + \sum_{l=1}^{L} \alpha_l \mu_l \ : \ \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D.$$

The approximation is valid for small values of α, which is sufficient in many applications. All patterns depicted in Figure 3 lie in the same subspace and can therefore be represented by one prototype and the corresponding tangent vectors. Therefore, the tangent distance between the original image and any of these transformed images is zero, while the Euclidean distance is significantly greater than zero. Using the squared Euclidean norm, the TD is defined as

$$d_{TD}(x_t, \mu) = \min_{\alpha \in \mathbb{R}^L} \left\{ ||x_t - (\mu + \sum_{l=1}^{L} \alpha_l \mu_l)||^2 \right\} .$$

A double-sided TD can also be defined, where both of the manifolds of the reference and observation are approximated.

4.2 Image Distortion Model

We briefly review an image distortion model that is able to compensate for local displacements. The efficiency of the model in handwritten character recognition is shown in [3]. In this model, to calculate the distance between the image frame

x_t and the class-specific prototype image μ, instead of computing the squared error between the pixels x_{ij} and μ_{ij}, we compute the minimum distance between x_{ij} and $\mu_{i'j'}$, where $(i', j') \in R_{ij}$, and R_{ij} is a certain neighborhood of (i, j). According to this definition, the invariant distance can be calculated by

$$d_{IDM}(x_t, \mu) = \sum_{ij} \min_{(i',j') \in R_{ij}} \{d(x_{ij}, \mu_{i'j'})\}.$$

The accuracy of the IDM depends on choosing a suitable R_{ij} which leaves the class-membership unchanged. If R_{ij} is too small, too few deformations can be compensated. If it is too large, deformations that change class membership are tolerated. In both cases, the error rate of the classifier will increase. To increase accuracy in finding the optimal displacement, local image context is used [3]. This leads to a mapping of edges to edges in the alignment. In informal experiments we found 7×7 sub images to be optimal. Therefore, the local distance $d(x_{ij}, \mu_{i'j'})$ between pixel x_{ij} and $\mu_{i'j'}$ is chosen to be

$$d(x_{ij}, \mu_{i'j'}) := \sum_{m=-3}^{3} \sum_{n=-3}^{3} ||x_{i+m,j+n} - \mu_{i'+m,j'+n}||^2.$$

Figure 4 shows the effect of the IDM. The figure consists of an image frame and three image pairs. Each image pair includes the transformed image that results from the IDM distance calculation (left) and an artificially distorted image (right). The difference from the original image is shown in the second row. The distorted image frames are created artificially by one pixel displacements of the left hand of the signer. The near-zero difference images for the left images of the pairs show that the IDM effectively compensates for the artificial distortion.

5 Combination of TD and IDM

To compensate for global and local variations simultaneously, we propose to combine TD and IDM. Here we apply two different methods for combination.

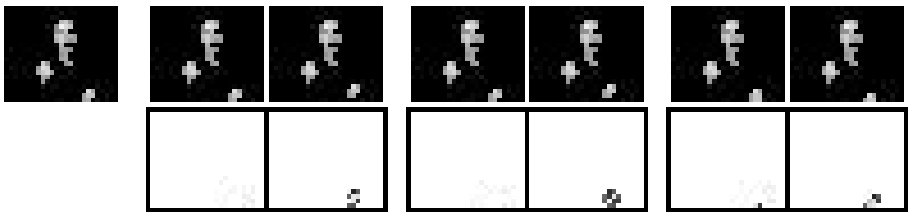

Fig. 4. Example of the image distortion model. (first row: original image and image pairs including the transformed image that results from the IDM distance (left) and artificially distorted image by displacement of the left hand of the signer (right); second row: difference images to the original image).

Method A: TD to compare sub images in IDM. In the IDM, only the displacement of the sub images is allowed. If we calculate the TD instead of the Euclidean distance $d(x_{ij}, \mu_{i',j'})$ between sub images, small affine transformations in the sub-images are considered. The image frames where only the left hand of the signer is distorted by axis deformation and where there is diagonal deformation and scaling of the left hand of the signer are shown in Figure 5. These distortions are tolerated by the proposed combination method, and the transformed images that result from the combination method are also shown. From the difference images, it can be seen that the method accounts for these transformations. Note that the distorted images are different from Figure 4.

Method B: IDM to compare TD-transformed images. Another possible way to combine the two invariant distances is the use of the TD before employing the image distortion model to find the closest image frames in the linear subspaces as proposed in [12]. We employ the one-sided tangent distance using the tangent vectors of the prototype image μ. The closest image frame in the subspace $\widehat{\mathcal{M}}_\mu$ to the observation image frame x_t is calculated by $\hat{\mu} = \mu + \sum_{l=1}^{L} \hat{\beta}_l \mu_l$ where

$$\hat{\beta} = \arg\min_{\beta} \left\{ ||x_t - (\mu + \sum_{l=1}^{L} \beta_l \mu_l)||^2 \right\}, \qquad \beta \in \mathbb{R}^L.$$

Thus, $\hat{\mu}$ compensates for small global affine transformations. This $\hat{\mu}$ is then used in the IDM instead of μ. In this combination, we first account for global transformations and then for local deformations, yielding a distance function that is invariant with respect to global transformations and local deformations.

6 Experimental Results

In the experiments, we have used HMMs in a nearest neighbor manner, i.e. each training observation forms its own HMM, where variance is fixed globally. The achieved results for using the different distances are given in Table 1. It can be seen that both combination methods of TD and IDM improve the accuracy of the

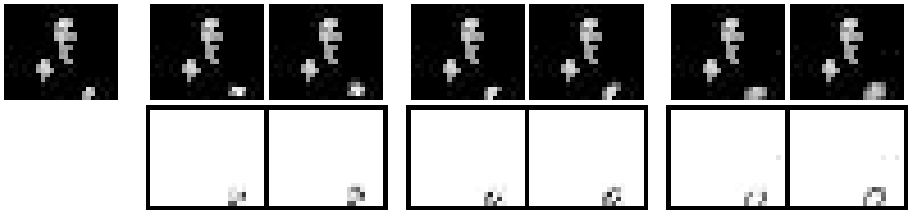

Fig. 5. Example of the first combination method. (first row: original image and image pairs including the transformed image that results from the combination method and distorted image by axis deformation, diagonal deformation and scaling of the left hand of the signer, second row: the difference image between the transformed or distorted image and the original image)

Table 1. Error rates [%] of the classifier with different distances.

Distance	Original image	Horizontal gradient	Vertical gradient	Horizontal & vertical gradients
TD [5]	22.2	22.8	23.4	21.3
IDM	21.9	21.5	24.6	23.4
IDM+TD (Method A)	**17.2**	18.8	18.2	18.4
IDM+TD (Method B)	20.3	21.1	21.5	20.9

classifier compared to using one of the invariant distances alone. Consistently, the best error rate is obtained with method A, which enables the classifier to model global transformations in the sub images of the image frames.

More experiments have been performed to investigate how to weigh the importance of the local sub images of the original image with respect to the gradient images. Figure 6 shows the error rate of the classifier using IDM and the two combination methods of TD and IDM depending on the relative weight of original images and derivatives. A relative weight of zero means that only gradient images are used. The graphs show that the best results are achieved when only the local sub images of the original images are used. The best error rate of 17.2% is obtained using combination method A, which is an improvement of 23% relative. About 65% of the remaining misclassified utterances of the data are due to very strong visual differences from the other utterances in the database, which means they are always different from all training utterances when classified.

Unfortunately, a direct comparison with results of other research groups is 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.

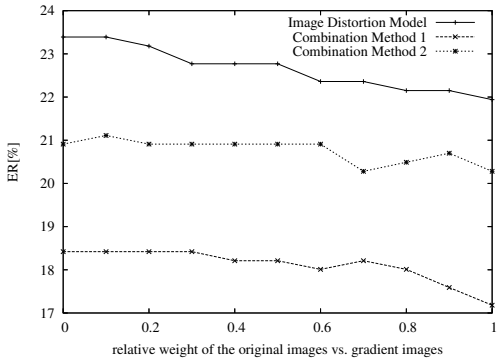

Fig. 6. Error rate of the system with respect to the gradient weight.

7 Conclusion

In this paper we have presented two different ways of combining IDM and TD in sign language recognition to account for visual variabilities. These methods allow for recognizing sequences with variations in position, orientation, or size of the hands or the head of the signer. The TD, accounting for global affine transformations, and the IDM, accounting for local deformation, complement each other and allow for compensating a combination of global and local transformations. The recognition results are consistently improved using the combined method.

In the future, we plan to use these methods on larger databases and in the recognition of continuous sign language.

References

1. P. Simard, Y. Le Cun, J. Denker. Efficient Pattern Recognition Using a New Transformation Distance. In *Advances in NIPS*, pp. 50–58, Morgan Kaufmann, 1993.
2. D. Keysers, W. Macherey, H. Ney, J. Dahmen. Adaptation in Statistical Pattern Recognition Using Tangent Vectors. In *TPAMI*, 26(2):269–274, Feb. 2004.
3. D. Keysers, C. Gollan, H. Ney. Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *ICPR*, vol. 4, pp. 511–514, Cambridge, UK, Aug. 2004.
4. P. Dreuw, D. Keysers, T. Deselaers, H. Ney. Gesture Recognition Using Image Comparison Methods. In *Gesture in Human-Computer Interaction and Simulation*, Vannes, France, May 2005. In press.
5. M. Zahedi, D. Keysers, H. Ney. Pronunciation Clustering and Modeling of Variability for Appearance-based Sign Language Recognition. In *Gesture in Human-Computer Interaction and Simulation*, Vannes, France, May 2005. In press.
6. M. Zahedi, D. Keysers, H. Ney. Appearance-based Recognition of Words in American Sign Language. In *Iberian Conf. on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005. In press.
7. L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proc. of the IEEE*, 77(2):267–296, Feb. 1989.
8. R. Bowden, D. Windridge, T. Kadir, A. Zisserman, M. Brady. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In *ECCV*, pp. 390–401, Prague, Czech Republic, May 2004.
9. B. Bauer, H. Hienz, K.F. Kraiss. Video-Based Continuous Sign Language Recognition Using Statistical Methods. In *ICPR*, pp. 463–466, Barcelona, Spain, Sep. 2000.
10. T. Starner, J. Weaver, A. Pentland. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. In *TPAMI*, 20(12):1371–1375, Dec. 1998.
11. C. Vogler and D. Metaxas. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *Proc. Int. Conf. on Systems, Man and Cybernetics*, pp. 156–161, Orlando, FL, Oct. 1997.
12. D. Keysers, J. Dahmen, H. Ney, B. Wein, and T.M. Lehmann Statistical Framework for Model-based Image Retrieval in Medical Applications In *Journal of Electronic Imaging* 12(1):59–68, Jan. 2003.