

Gesture Recognition Using Image Comparison Methods

Philippe Dreuw, Daniel Keysers, Thomas Deselaers, and Hermann Ney

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

Abstract. We introduce the use of appearance-based features, and tangent distance or the image distortion model to account for image variability within the hidden Markov model emission probabilities to recognize gestures. No tracking, segmentation of the hand or shape models have to be defined. The distance measures also perform well for template matching classifiers. We obtain promising first results on a new database with the German finger-spelling alphabet. This newly recorded database is freely available for further research.

1 Introduction

Work in the field of vision-based gesture recognition usually first segments parts of the input images, for example the hand, and then uses features calculated from this segmented input like shape or motion. Problems with this approach are tracking, occlusion, lighting or clothing constraints. Results in the field of object recognition in images suggest that this intermediate segmentation step is not necessary and even hindering, as e.g. segmentation or tracking is never perfect. The question addressed in our research is if appearance based features are competitive for gesture recognition and if we can use similar models of image variability as in object recognition. We have integrated distance measures known from image and optical character recognition (e.g. being invariant against affine transformations) into the hidden Markov model classifiers.

Most of the common systems [2,8,9,10] assume a constant environment, e.g. persons wearing non-skin-colored clothes with long sleeves and a fixed camera position under constant lighting conditions. The presented systems are often highly person-dependent and the gestures used exhibit great differences to be easily recognizable. We aim at overcoming these shortcomings with this work.

2 Appearance-Based Features for Gesture Recognition

In appearance-based approaches the image itself and simple transformations (filtering, scaling, etc.) of the image are usually used as features. In this paper, we denote an original image X in a sequence at time $t = 1, \dots, T$ by X_t , and the pixel value at the position (x, y) by $X_t(x, y)$.

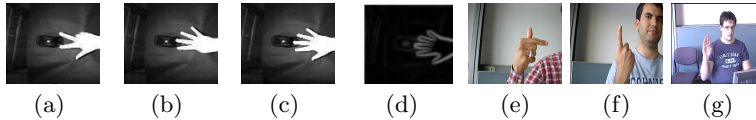


Fig. 1. Infrared images of the gesture “Five”. (a)-(d): original and spatial derivatives image features. (e)-(g) are examples of the i6-Gesture database.

When working, for example, with gray valued images (e.g. infrared images like in Fig. 1(c)), *original images* or their *spatial derivatives* can be used as features. *Skin probability images* have been created according to their skin probability maps [5]. Other features have been analyzed in [3].

3 Hidden Markov Models

The ability of hidden Markov models (HMM) to compensate time and amplitude variations has been proven for speech recognition, gesture recognition, sign language recognition and human actions [4,8,9,10]. In particular we focus on distance measures being invariant against slight affine transformations or distortions. The idea of a HMM is to represent a signal by a state of a stochastic finite state machine. A more detailed description can be found in [4].

In each state s of an HMM, a distance is calculated. We assume pooled variances over all classes and states, i.e. we use $\sigma_{sdk} = \sigma_d$. The negative logarithm of $p(X|s)$ can be interpreted as a distance $d(p(X|s))$ and is used as emission score:

$$-\log(p(X|s)) = \frac{1}{2} \left(\sum_{d=1}^D \left(\underbrace{\left(\frac{X_d - \mu_{sd}}{\sigma_d} \right)^2}_{\text{distance}} + \underbrace{\log(2\pi\sigma_d^2)}_{\text{normalization factor}} \right) \right)$$

When working with image sequences, we calculate a distance between two images, e.g. we compare the current observation image X_t (or any transformed image \tilde{X}_t) with the mean image μ_s at this state. Simply comparing the pixel values is quite often used in object recognition but different methods have been proposed to do this.

Tangent Distance. Because the Euclidian distance does not account for affine transformations such as scaling, translation and rotation, the tangent distance (TD), as described in [7], is one approach to incorporate invariance with respect to certain transformations into a classification system. Here, invariant means that image transformations that do not change the class of the image should not have a large impact on the distance between the images. Patterns that all lie in the same subspace can therefore be represented by one prototype and the corresponding tangent vectors. Thus, the TD between the original image and any of the transformations is zero, while the Euclidean distance is significantly greater than zero.

Image Distortion Model. The image distortion model [6] is a method which allows for small local deformations of an image. Each pixel is aligned to the

pixel with the smallest squared distance from its neighborhood. These squared distances are summed up for the complete image to get the global distance. This method can be improved by enhancing the pixel distance to compare sub images instead of single pixels only. Further improvement is achieved by using spatial derivatives instead of the pixel values directly.

4 Databases

LTI-Gesture Database. The LTI-Gesture database was created at the Chair of Technical Computer Science at the RWTH Aachen [1]. It contains 14 dynamic gestures, 140 training and 140 testing sequences. An error rate of 4.3% was achieved on this database. Fig. 1(c) shows an example of a gesture.

i6-Gesture Database. We recorded a new database of fingerspelling letters of German Sign Language. Our database is freely available on our website¹. The database contains 35 gestures and consists of 700 training and 700 test sequences. 20 different persons were recorded under non-uniform daylight lighting conditions, without any restrictions on the clothing while gesturing. The gestures were recorded by one webcam (320x240 at 25 fps) and one camcorder (352x288 at 25 fps), from two different points of view. Fig. 1(e)-Fig. 1(g) show some examples of different gestures. More information is available on our website.

5 Results

In [1], an error rate of 4.3% has been achieved using shape and motion features in combination with forearm segmentation. Using the centroid features as presented in [8], we have achieved an error rate of 14.2%, and we can conclude that these features should only be used to describe motion patterns instead of more complex hand shapes. Using original image features on the LTI-Gesture database, we have achieved an error rate of 5.7% which has been improved to 1.4% in combination with the tangent distance [3] or the image distortion model (see Tab. 1).

On the i6-Gesture database, we have used only the webcam images to test our system. It is obvious that this database contains gestures of very high complexity, and that additional methods are needed for feature extraction or other distance measures. Using a camshift tracker to extract position independent features (note that we do *not* try to segment the hand), we could improve the error rate from 87.1% to 44.0%.

Using a two-sided tangent distance we have improved the error rate to the currently best result of 35.7%, which shows the advantage of using distance mea-

¹ <http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html>

Table 1. Error rates [%] on the LTI-Gesture database.

Features	Euclidian	Tangent	IDM
COG [8]	14.2	-	-
original	5.7	1.4	1.4
magnitude Sobel	7.1	1.4	1.4

Table 2. Error Rates [%] on the i6-Gesture database.

Feature	Euclidian	Tangent
original thresholded by skin color prob.	87.1	-
+ camshift tracking (no segmentation)	44.0	35.7

tures that are invariant against small affine transformations and the possibility of recognizing gestures by appearance-based features (see Tab. 2).

6 Conclusion

At this point, some questions still remain unanswered, e.g. not all distance measures and camera streams were completely analyzed on the i6-Gesture database which are expected to improve the error rate. The best achieved error rate on the i6-Gesture database is 35.7% and shows the high complexity of this database. Nevertheless, this result is promising because only a simple webcam without any restriction for the signer has been used and some signs are visually very similar, as for example the signs for “M”, “N”, “A”, and “S”.

The use of tangent distance and image distortion models as appropriate models of image variability in combination with appearance-based features has been investigated and compared to the Euclidian distance on other databases. Using these distance measures, the error rate has been reduced on all regarded databases, especially on the LTI-Gesture database. This shows the power of integrating these distance measures into the HMM emission probabilities for recognizing gestures.

References

1. S. Akyol, U. Canzler, K. Bengler, and W. Hahn. Gesture Control for Use in Automobiles. In *IAPR MVA Workshop*, Tokyo, Japan, pages 349–352, Nov. 2000. 3
2. R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In J. M. Tomas Pajdla, editor, *ECCV*, volume 1, Prague, Czech Republic, pages 391–401, May 2004. 1
3. P. Dreuw. Appearance-Based Gesture Recognition. Diploma thesis, RWTH Aachen University, Aachen, Germany, Jan. 2005. 2, 3
4. F. Jelinek. *Statistical Methods for Speech Recognition*. Cambridge, MA, Jan. 1998. 2
5. M. Jones and J. Rehg. Statistical Color Models with Application to Skin Color Detection. Technical Report CRL 98/11, Compaq Cambridge Research Lab, 1998. 2
6. D. Keysers, J. Dahmen, H. Ney, B. Wein, and T. Lehmann. Statistical Framework for Model-based Image Retrieval in Medical Applications. *Journal of Electronic Imaging*, 12(1):59–68, Jan. 2003. 2
7. D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in Statistical Pattern Recognition using Tangent Vectors. *PAMI*, 26(2):269–274, Feb. 2004. 2
8. G. Rigoll, A. Kosmala, and S. Eickeler. High Performance Real-Time Gesture Recognition using Hidden Markov Models. In *Int. Gesture Workshop*, volume 1371, Bielefeld, Germany, pages 69–80, Sep. 1998. 1, 2, 3

9. T. Starner, J. Weaver, and A. Pentland. Real-time ASL recognition using desk and wearable computer based video. *PAMI*, 20(12):1371–1375, Dec. 1998. 1, 2
10. M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll. Gesture Components for Natural Interaction with In-Car Devices. In *Int. Gesture Workshop*, volume 2915 of *LNAI*, Gif-sur-Yvette, France, pages 448–459, Mar. 2004. 1, 2