A Hybrid Approach to Sign Language Recognition

Jeroen Lichtenauer [a] Emile Hendriks [b] Marcel Reinders [b]

[a] *Imperial College London, 180 Queen's Gate, London SW7 2AZ*
[b] *Delft University of Technology, P.O.Box 5031, 2600 GA Delft*

Abstract

Methods commonly used for speech and sign language recognition often rely on outputs of Hidden Markov Models (HMM) or Dynamic Time Warping (DTW) for classification, which are merely factorized observation likelihoods. Instead, we propose to use Statistical DTW (SDTW) only for warping, while classifying the synchronized features with either of two proposed discriminants. This hybrid approach is shown to outperform HMM and SDTW. However, we have found that combining likelihoods of multiple models in a second classification stage degrades performance of the proposed classifiers, while improving performance with HMM and SDTW. A proof-of-concept experiment, combining DFFM mappings of multiple SDTW models with SDTW likelihoods, shows that also for model-combining, hybrid classification can provide significant improvement over SDTW.

1 Introduction

A recent modification of HMM, called Statistical DTW (SDTW) [1], incorporates the warping flexibility of the exemplar-based DTW in the statistical framework of HMM. Since in [1] SDTW was shown to outperform HMM in on-line handwriting recognition, an improvement over HMM can also be expected when SDTW is applied to sign language recognition. Our results show that this is indeed the case. However, we further improve upon SDTW, based on our main hypothesis:

Proposition *The maximized likelihood that results in the optimal signal warping is not the optimal conditional likelihood estimation of the signal class*.

Instead of relying directly on the likelihoods obtained from (S)DTW or HMM, we consider SDTW primarily as a registration method. The synchronized feature sets are classified by Combined Discriminative Feature Detectors (CDFD) or Quadratic classification of Discriminative Features transformed by a Fisher Mapping (Q-DFFM). Details about these classification methods and the stereo hand tracking can be found in [2].

Our experiments are limited to hand motion trajectories and apparent hand-size change in isolated signs. These are the few components that the current state of the art in human motion analysis allows to track in reasonably soft-constrained situations. We assume that if sign language recognition by motion and hand size change benefits from a hybrid approach, this will certainly be the case if even more parallel aspects (e.g. detailed hand/body pose and facial expression) are considered.

2 Results

Sign classification is evaluated on a set of 120 different signs of the Dutch Sign Language (DSL), each performed by 75 different persons. The images are captured at 640x480 pixels and 25 frames per second. Evaluation is done by 5-fold cross-validation, with a separation of persons between the train and test sets (person-independent), and a separation of non-target classes between train and test sets, in the case of target-class classification (rejection of unseen classes). Our hybrid methods (SDTW + CDFD or Q-DFFM) are compared to SDTW and a 40 state HMM with Bakis topology. Three types of classifications are considered: 1) Target-class classification with a model of the target class, 2) Target-class classification by combining

Table 1: Average classification results for 120 signs and 5 cross-validations. The results for target-class classification are measured in $\mathrm{pAUC}_{0.1}$. For multi-class classification, the rate of correct classification is given.

		target-cl. single model	target-cl. multi-m comb.	multi-cl. multi-m comb.
a	HMM	84.61%	96.97%	90.8%
b	SDTW	90.54%	97.22%	90.8%
c	SDTW+CDFD	95.46%	90.86%	76.0%
d	SDTW+Q-DFFM	**96.62%**	94.84%	83.7%
e	SDTW&DFFM5		**97.50%**	**92.3%**

models of multiple classes, 3) Multi-class classification by combining the SDTW models of all classes. In target-class classification, a sign is either detected as the target class, or rejected (binary decision). The rejection threshold determines the trade-off between true positives and false positives (operating point). All possible operating points for the trained classifier of a target class are represented by the Receiver Operating Characteristic (ROC) curve. For evaluation of the trained classifiers, we computed the partial Area Under the Curve (AUC) of the ROC curves, between false positive rates of 0 and 0.1 ($\mathrm{pAUC}_{0.1}$).

The first two result columns in table 1 show the average $\mathrm{pAUC}_{0.1}$ over the 120 target classes in the 5 cross-validations. Both hybrid methods clearly outperform HMM and SDTW when a single model is used. In the second experiment, the likelihood outputs of multiple (96) models are combined by training a 2nd stage classifier (Fisher) in the output space of all classifiers trained in the previous experiment. Hence, the previous target-class classifier for the real target class, but also target-class classifiers for other classes, are combined together, to classify the real target class. Results are shown in the second results column of table 1. While HMM and SDTW benefit from the model combining, the hybrid methods do not. This is because the target-class classifiers trained with our approach are too specific to be used as inputs for 2nd stage detection of other classes. Because the Fisher Mapping, that is part of Q-DFFM, does contain information about other classes, we combined the SDTW outputs with the first 5 Fisher dimensions of each of the single-model SDTW+Q-DFFM target-class classifiers, increasing the dimensionality for the 2nd stage Fisher classifier by 500% over SDTW alone. This method is indicated by SDTW&DFFM5 in table 1(e). Despite the increase of dimensionality, the richer description increases the partial ROC surface from 97.22% to 97.50%. The significance of this improvement is indicated by a p-value of 0.009 in a paired t-test of $\mathrm{pAUC}_{0.1}$ over all individual classifiers. In the multi-class classification experiment, we have combined the single-model target-class classifiers for all 120 sign classes in a single feature space, using Fisher to discriminate between the 120 classes in the 2nd stage. Also here, the SDTW&DFFM5 method outperforms HMM and SDTW significantly (3rd column in table 1). The improvement over HMM has a p-value of 0.019 in a paired t-test of the classification rates over the 5 cross-validation folds. Again, this confirms the benefit of a hybrid approach to sign language recognition.

3 Acknowledgements

This work is a part of the PhD work of Jeroen Lichtenauer, conducted at TU Delft, the Netherlands. Current work of Jeroen Lichtenauer is supported by MAHNOB project of Maja Pantic, funded by the European Research Council under the ERC Starting Grant agreement No. ERC-2007-StG-203143.

References

[1] C. Bahlmann and H. Burkhardt. The writer independent online handwriting recognition system *frog on hand* and cluster generative statistical dynamic time warping. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 26(3):299–310, March 2004.

[2] J.F. Lichtenauer, E.A. Hendriks, and M.J.T. Reinders. Sign language recognition by combining statistical dtw and independent classification. *Submitted to: Transactions on Pattern Analysis and Machine Intelligence*, 2008.