Rapid Signer Adaptation for Continuous Sign Language Recognition Using a Combined Approach of Eigenvoices, MLLR, and MAP

Ulrich von Agris, Christoph Blömer, and Karl-Friedrich Kraiss
Institute of Man-Machine Interaction, RWTH Aachen University, Germany
{vonagris,bloemer,kraiss}@mmi.rwth-aachen.de

Abstract

Current sign language recognition systems are still designed for signer-dependent operation only and thus suffer from the problem of interpersonal variability in production. Applied to signer-independent tasks, they show poor performance even when increasing the number of training signers. Better results can be achieved with dedicated adaptation methods. In this paper, we describe a vision-based recognition system that quickly adapts to new signers. For rapid signer adaptation it employs a combined approach of eigenvoices, maximum likelihood linear regression, and maximum a posteriori estimation. An extensive evaluation was performed on a large sign language corpus, that contains continuous articulations of 25 native signers. The proposed adaptation approach significantly increases accuracy even with a small amount of adaptation data. Supervised adaptation with only 10 adaptation utterances yields a recognition accuracy of 75.8%, which is a relative error rate reduction of 30.2% compared to the signer-independent baseline.

1 Introduction

Deaf and hearing impaired people use sign language for everyday communication among themselves. Since sign languages are non-verbal languages, information is conveyed visually, using a combination of manual and non-manual means such as the signer's hands and facial expressions. Although different in form, they serve the same functions as a spoken language.

Research in the field of sign language recognition has made remarkable advances in recent years. Present achievements provide the basis for future applications with the objective of supporting the integration of deaf people into the hearing society. Translation systems, for example, could facilitate communication between deaf and hearing people in public situations. Further applications such as user interfaces and automatic indexing of signed videos become feasible.

All applications mentioned before have in common that they must operate in a user-independent scenario. Current systems for sign language recognition achieve excellent performance for signer-dependent operation, but their recognition rates decrease significantly if the signer's articulation deviates from the training data. This performance drop results from the strong interpersonal variability in production of sign languages.

Better results can be obtained with dedicated adaptation methods successfully applied in automatic speech recognition. Speaker adaptation aims to improve the general performance level for a new speaker approaching that of an signer-dependent system for that speaker, but avoiding the need for large amounts of training data.

In this paper, we propose a new combined approach for rapid signer adaptation, which is an advancement of our former approach introduced in [8].

2 Related Work

The current state in automatic sign language recognition is roughly 30 years behind speech recognition, which corresponds to a gradual transition from isolated to continuous recognition for small vocabulary tasks. Research efforts were mainly focused on robust feature extraction and statistical modeling of signs. However, current recognition systems are still designed for signer-dependend operation under laboratory conditions. The reader interested in a thorough survey on sign language recognition is directed to [6].

Only a few recognition systems were reported which support signer-independent operation. Generally, they show poor performance since the reference models are built on a small training population. Although signer adaptation can significantly improve the performance for an untrained signer, solely two publications describe the application of adaptation methods so far.

The system, found in [5], applies Bayesian networks and hidden Markov models to recognize a vocabulary of 20 distinct gestures on the basis of visual features that reflect aspects of sign language grammar. Supervised maximum a posteriori adaptation to one new signer with a set of all 20 gestures yields 88.5% accuracy.

In our previous work [8], we employed a combined approach of maximum likelihood linear regression and maximum a posteriori estimation for signer adaptation. Algorithms were modified to consider the specifics of sign languages. On a vocabulary of 153 isolated signs, supervised adaptation to four new signers with a set of 80 / 160 signs averages 78.6% / 94.6% accuracy.

In summary, it can be stated that only a few publications addresses the problem of interpersonal variance in signing on the classification level so far.

3 System Design

Figure 1 shows a schematic of the vision-based sign language recognition system that constitutes the basis for our ongoing research work. A detailed description of the system is provided in [2]. Since sign languages make use of manual and facial means of expression both channels are employed for recognition.

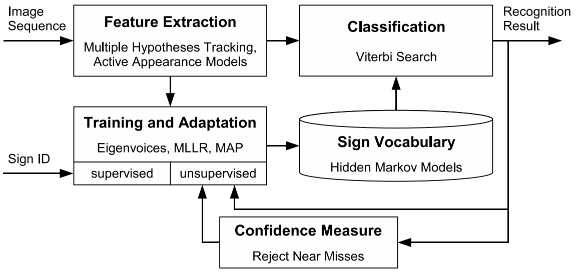

Figure 1. Schematic of the adaptive sign language recognition system.

For mobile operation in uncontrolled environments sophisticated algorithms were developed that robustly extract manual and facial features. The extraction of manual features relies on a multiple hypotheses tracking approach to resolve ambiguities of hand positions. For facial feature extraction an active appearance model is applied to identify areas of interest such as the eyes and mouth region. Finally, a numerical description of the signer's manual configuration, facial expression, and lip pattern is computed to compose the feature vector.

The classification stage uses hidden Markov models and is designed for recognition of isolated signs as well as of continuous sign language. Emission probabilities are represented by Gaussian mixture models. Training and classification apply the Viterbi algorithm.

4 Signer Adaptation

Selected adaptation methods from automatic speech recognition are combined for the use in sign language recognition tasks in order to improve the performance of the signer-independent recognizer. Based on some adaptation data, the adaptation process reduces the mismatch between the signer-independent model and the observations recorded from a new signer.

Various adaptation methods have already been investigated in the context of speech recognition. Due to the obvious similarities between speech and sign language recognition, some are applicable for signer adaptation. Following three conventional model-based methods are combined: maximum a posteriori (MAP) estimation, maximum likelihood linear regression (MLLR), and the eigenvoice (EV) approach. All methods are employed in current speech recognition systems and have proven to perform excellent in the speech domain.

4.1 Maximum A Posteriori Estimation

The maximum a posteriori estimate $\tilde{\mu}_{\text{MAP}}$ for the Gaussian mean μ_m of a mixture component m is a linear interpolation between a-priori knowledge derived from a signer-independent model and the observations from the adaptation sequences. During Viterbi alignment of an adaptation sequence with its corresponding model, the feature vectors mapped to a certain component can be recorded, yielding the empirical mean $\bar{x}_m$ of the mapped vectors. According to [4], the MAP estimate is

$$\tilde{\mu}_{\text{MAP}} = \frac{\tau}{\tau + N} \cdot \mu_m + \left(1 - \frac{\tau}{\tau + N}\right) \cdot \bar{x}_m \qquad (1)$$

where N is the number of feature vectors aligned to component m and τ is a weight for the influence of the a-priori knowledge. If N approaches infinity, the influence of the signer-independent model approaches zero and the adapted parameter equals the empirical mean.

Thus MAP performs well on large sets of adaptation data, but its pure form can only be used to update seen components. This can be solved by using the MLLR-adapted model as prior knowledge, replacing the signer-independent mean by the already transformed mean.

4.2 Maximum Likelihood Linear Regression

The mixture components of the signer-independent HMMs are clustered into a set of regression classes $C = 1, \ldots, R$ such that each Gaussian component m belongs to one class $c \in C$. A linear transformation

W_c for each class c is then estimated from the adaptation data. Estimation of the transformation matrices follows the maximum likelihood paradigm, so the transformed models best explain the adaptation sequences. Reestimation formulae for W_c based on the iterative expectation-maximization algorithm are given in [1].

The Gaussian mean μ_m of each component m from class c is then transformed with the corresponding matrix W_c, yielding the adapted parameter

$$\tilde{\mu}_m = W_c \cdot \bar{\mu}_m \tag{2}$$

where $\bar{\mu}_m$ is the extended mean vector

$$\bar{\mu}_m^T = \begin{bmatrix} 1 & \mu_m^T \end{bmatrix} \tag{3}$$

A component from a model which has not been observed in adaptation data can thus be transformed based on the observed components from the same class.

As proposed in [1], a regression class tree is used to improve the clustering of the mixture components, where the number of regression classes depends on the available amount of adaptation data. Each node c of the tree corresponds to a regression class and a transformation W_c is associated with the node.

However, whereas MLLR becomes efficient only after a certain number of adaptation utterances have been articulated, the eigenvoice approach can improve the performance of a sign language recognition system even after a few adaptation utterances.

4.3 Eigenvoices

The eigenvoice approach [3] constrains the adapted model to be a linear combination of a small number of basis vectors obtained offline from a set of R reference speakers, and thus greatly reduces the number of free parameters to be estimated from adaptation data. These basis vectors, called *eigenvoices*, are orthogonal to each other and represent the most important components of variation between the reference speakers.

The adapted model is located in the *speaker space*, that is obtained by applying a dimensionality reduction technique, such as principal component analysis, to a set of R supervectors of dimension D extracted from R well-trained speaker-dependent models. In order to get the reduced speaker space, only the K first eigenvectors $e_1, e_2, \cdots, e_K$ with $K < R << D$ are kept. Related to the mean supervector e_0, these K eigenvoices, which capture most of the variation of the training data, span the reduced speaker space of dimension K.

A supervector is composed of the model parameters that have to be adapted to the new speaker. Typically, it consists in the concatenation of all the Gaussian mean vectors of a speaker-dependent model. Those parameters that are not represented in the supervector, e.g. variances and transition probabilities, must be obtained from an speaker-independent model.

Finally, a new speaker can be located in the speaker space by a vector of $K + 1$ weights $w_0, w_1, \cdots, w_K$. All Gaussian mean vectors $\hat{\mu}_i$ of the adapted models are then updated using the equation

$$\hat{\mu}_i = \sum_{k=0}^{K} w_k e_k \tag{4}$$

with $i = 1, 2, \cdots, N$, where N is the total number of Gaussians of the speaker-adapted system. The weights w_k are generally estimated using *maximum likelihood eigen-decomposition* [3] to maximize the likelihood of the adaptation data.

4.4 Combined Adaptation Approach

Each of the adaptation method described before has specific benefits and drawbacks. MAP estimation can approach speaker-dependent performance but only updates the parameters of models that are observed in the adaptation data, thus large amounts of adaptation data are generally required. MLLR allows much faster adaptation by even updating unseen models, but suffers from the problem of just as fast saturation. Finally, the eigenvoice approach provides by far the fastest adaptation but faces an even more serious saturation problem.

Consequently, a combined approach concatenating these three adaptation methods is supposed to achieve better results. The proposed approach uses the models obtained by eigenvoice estimation as prior for MLLR, and the models adapted by MLLR again as prior for final MAP estimation. This approach will be referred to as EV+MLLR+MAP in the following and is expected to combine the benefits of each method, hence yielding rapid adaptation without fast performance saturation.

5 Adaptation Experiments

Database In contrast to speech recognition, there is no standardized benchmark that meets the requirements for signer-independent continuous recognition. For this reason we recorded a new database, as described in [7]. The corpus is based on a vocabulary of 450 basic signs in German Sign Language and comprises 780 sentences (603 for training, 177 for testing). Each sentence was performed once by 25 native signers of different sexes and ages. The articulations of the reference signer were recorded even three times, serving for evaluation of the signer-dependent recognition rates. The whole database will be made available soon for interested researchers.

Signer-dependent recognition rates average 86.6%, while signer-independent performance in a leaving-one-out test with no adaptation sets a baseline of 65.3%.

Results The following supervised adaptation experiments in continuous sign language recognition were conducted on the new EV+MLLR+MAP approach, on our former MLLR+MAP approach as well as on the three methods EV, MLLR, and MAP respectively. The training set was used for static adaptation with different amounts of adaptation data while the test set served for evaluation of recognition performance.

The results below were derived employing Gaussian single densities. Experiments with Gaussian mixtures show the same behavior because of the small training population. Only the means were updated by the adaptation methods. Variances and mixture weights remain unchanged as the mean covers most of the variability between the speakers. Whenever MLLR was applied, seen components were adapted with the most special transform from the regression class tree while unseen components were updated with a global transformation.

Figure 2 summarizes the experiments, showing the recognition performance of the adapted models using the different methods. When using only a small amount of adaptation data, conventional EV outperforms the two methods MLLR and MAP as well as the combined MLLR+MAP approach by up to 10.5%. As expected, the EV approach is also suited for rapid adaptation in the domain of sign language recognition.

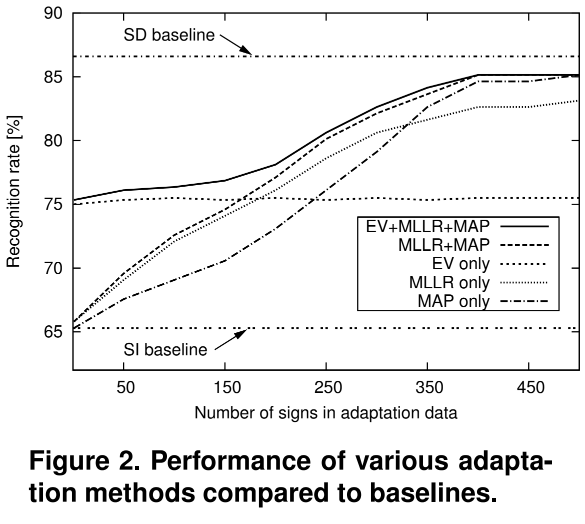

Figure 2. Performance of various adaptation methods compared to baselines.

The proposed combination of this EV method and our former MLLR+MAP approach results in the desired effect: the rapid adaptation using EV is preserved, while its saturation is compensated by MLLR+MAP. The new EV+MLLR+MAP approach generally yields the best models, regardless of the number of adaptation utterances. Thus rapid signer adaptation is possible without covering the whole vocabulary during adaptation as described in [5], which only applies MAP adaptation.

6 Conclusion

Applying adaptation methods from automatic speech recognition in a sign language context yields significant performance improvements. The experimental results prove that the proposed EV+MLLR+MAP approach is superior to all other investigated adaptation strategies, regardless of the amount of adaptation data. The EV approach allows rapid adaptation of a signer-independent system, even when only a few adaptation utterances are available. The combination with MLLR+MAP retards performance saturation, thus resulting in high accuracy for large sets of adaptation data. Supervised adaptation with only 10 adaptation utterances yields a recognition accuracy of 75.8%, which is a 30.2% relative error rate reduction compared to the signer-independent baseline.

References

[1] M. Gales and P. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10:249–264, 1996.

[2] K.-F. Kraiss, editor. *Advanced Man-Machine Interaction*. Springer, 2006.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. on Speech and Audio Processing*, 8(6):695–707, November 2000.

[4] C.-H. Lee, C.-H. Lin, and B.-H. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 39(4):806–814, 1991.

[5] S. C. W. Ong and S. Ranganath. Deciphering gestures with layered meanings and signer adaptation. In *Proc. of the 6th IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2004.

[6] S. C. W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):873–891, June 2005.

[7] U. von Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. In *Proc. of the 7th Intl. Workshop on Gesture in Human-Computer Interaction and Simulation*, 2007.

[8] U. von Agris, D. Schneider, J. Zieren, and K.-F. Kraiss. Rapid signer adaptation for isolated sign language recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, 2006.