Rapid Signer Adaptation for Isolated Sign Language Recognition

Ulrich von Agris, Daniel Schneider, Jörg Zieren, and Karl-Friedrich Kraiss
Institute of Man-Machine Interaction, RWTH Aachen University, Germany
{vonagris,schneider,zieren,kraiss}@mmi.rwth-aachen.de

Abstract

Research in the field of sign language recognition has not yet addressed the problem of interpersonal variance in large vocabulary on the classification level. Current recognition systems are designed for signer-dependent operation. 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. This paper describes a vision-based recognition system that quickly adapts to unknown signers. A combination of Maximum Likelihood Linear Regression and Maximum A Posteriori estimation was implemented and modified to consider the specifics of sign languages, such as one-handed signs. An extensive evaluation was performed in supervised and unsupervised mode on a vocabulary of 153 isolated signs. The proposed adaptation approach significantly increases accuracy even with a small amount of adaptation data. Supervised adaptation with 80 adaptation sequences yields a recognition accuracy of 78.6%, which is a relative improvement of 41.6% compared to the signer-independent baseline.

1. Introduction

The development of automatic recognition systems for sign language has made significant advances in recent years. Research efforts were mainly focused on the robust extraction of manual and non-manual features from the signer's articulation. Additional attention was paid to classification methods. First implementations proved that using subunit models has advantages over word models for each whole sign when recognizing large vocabularies.

The present achievements are the basis for future applications with the objective of supporting the integration of deaf people into hearing society. Translation systems, user interfaces, and automatic indexing of signed videos are just some examples. Further applications arise in the field of human-computer interaction. Multimodal user interfaces and the control of human avatars in virtual environments could be realized via gesture and mimic recognition.

All mentioned applications 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.

Although signer-independence is an essential precondition for future applications, only little investigations have been made in this field so far. This unexplored gap in sign language recognition is the subject of this paper.

Sign language Deaf and hearing impaired people use sign language for everyday communication. Information is conveyed through manual and non-manual means such as the signer's hands and facial expressions. The set of signs can be subdivided into one-handed and two-handed signs. The hand used for one-handed signs is called the dominant hand.

Interpersonal variability The performance drop in case of signer-independent recognition results from the broad interpersonal variability in production of sign languages. Even within the same dialect, considerable variations are commonly present. Fig. 1 shows different articulations of an exemplary sign in British Sign Language.

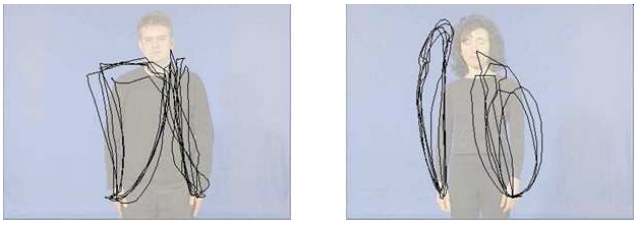

Figure 1. The sign "tennis" performed five times by two different native signers using the same dialect. Position of the hands are visualized as motion traces for comparison.

Analysis of the hand motion reveals that the variation between different signers are much higher than within one signer. Other manual features such as hand shape, posture, and location exhibit analogue variability.

2. Previous work

This section gives a short overview of existing sign language recognition (SLR) systems. An in-depth introduction to gesture and sign language recognition is found in [6]. Since there is no standardized benchmark the below recognition rates cannot be compared directly.

Vision-based systems face the problem of noisy and ambiguous input data. Most work focuses on this challenge and does not consider interpersonal variance. For isolated signs, recognition rates reach e.g. 98.9% on 229 signs [13] or 92.5% on 439 signs [12]. Datagloves yield more reliable and descriptive features, recognizing vocabularies as large as 5119 signs with 92.8% accuracy [11].

Recognition of continuous signing poses the additional difficulties of temporal segmentation and transition/coarticulation effects, but allows to support classification through the use of language models. Recognition rates of 93.2% (97 signs) were published for a vision-based system [1], while the glove-based method described in [2] attains 92.1% with 208 signs.

All above values refer to a signer-dependent recognition task. Only the following three publications specify signer-independent performance.

[2] uses datagloves and reports 85.0% recognition rate for sentences comprised of 208 signs. The authors employ a language model; however, they do not specify the degree of similarity between training and test sentences. Isolated signs from the same vocabulary are classified with 88.2% accuracy. No information is given on the composition of the vocabulary. The fact that signer-dependent performance is only 7.1% higher suggests low interpersonal variance.

[13] describes a vision-based system that achieves a maximum performance of 44.1% on 221 signs. The accuracy varies with the constellation of the training/test signers.

In [9] a vocabulary of 20 conceived gestures is recognized on the basis of visual features that reflect aspects of sign language grammar. Bayesian networks and HMMs are used for classification.

Supervised adaptation to one unknown signer with a set of all 20 gestures yields 88.5% accuracy.

Feature normalization for signer-independence is only described in [2, 13]. Except for [9], no publication listed above or in recent reviews [10] addresses the problem of interpersonal variance on the classification level.

3. System design

Fig. 2 shows a schematic of the vision-based adaptive SLR system described in this paper. The feature extraction stage builds on [13] and is designed to process real-world images. It uses a generic skin color model [5] to detect hands and face. The segmentation threshold is automatically chosen so that the resulting face candidate best matches the average face shape. For each pixel, the median color computed from all input images (which are buffered for this purpose) yields a reliable and parameter-free background model. This allows to eliminate static distractors.

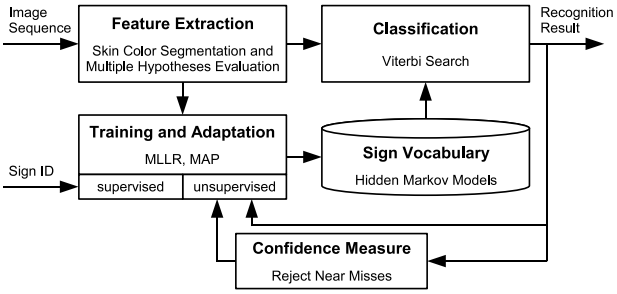

Figure 2. Schematic of the adaptive SLR system.

The remaining hand candidates still allow many interpretations. Therefore, multiple tracking hypotheses are pursued in parallel. The winner hypothesis is determined only at the end of the sign, using high level knowledge of the human body and the signing process to compute the likelihood of all hypothesized configurations per frame and all transitions between successive frames. This approach exploits all available information for the computation of the final tracking result, yielding robustness and facilitating retrospective error correction.

Features are computed from the hand candidate border as shown in Fig. 3. During periods of overlap, template matching is performed to accurately determine the center coordinates x, y using preceding or subsequent unoverlapped views. All other features are linearly interpolated.

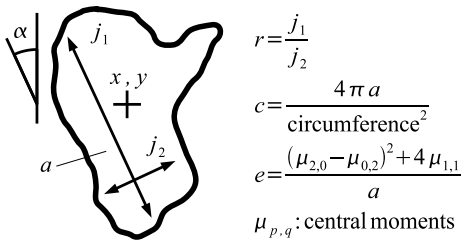

Figure 3. Shape-based features computed for each hand.

Hand center coordinates x, y are specified relative to the corresponding shoulder position, which is estimated from the width w_F and position of the face. In addition, x, y are normalized by w_F, and a by w_F^2. Since $\alpha \in [-90°, 90°)$, it is split into $o_1 = \sin 2\alpha$ and $o_2 = \cos \alpha$ to ensure stability at the interval borders. r, c and e describe the shape's axis ratio, compactness, and eccentricity. The derivatives $\dot{x}, \dot{y}, \dot{a}$ complete the 22-dimensional feature vector

$$x_t = [\ \underbrace{x\ \dot{x}\ y\ \dot{y}\ a\ \dot{a}\ o_1\ o_2\ r\ c\ e}_{\text{left hand}}\ \underbrace{x\ \dot{x}\ y\ \dot{y}\ \ldots}_{\text{right hand}}\ ] \quad (1)$$

If the hand is not visible or remains static throughout the sign, its features are set to zero.

The classification stage uses HMMs with an average of 41 states in Bakis topology for the representation of each sign. Emission probabilities are represented by Gaussian mixture models. Training and classification apply the Viterbi algorithm.

Corpus and baseline The test corpus consists of 153 isolated signs from British Sign Language, performed by four native signers five times each, totaling 3060 video clips of approx. 50 frames. Resolution is 384 x 288 pixels at 25 fps. The vocabulary comprises news items and navigation commands and was not selected for discernability. To reduce the amount of noise in the manual features, recordings were conducted in a controlled environment with diffuse lighting and a homogeneous background. The signers wear black clothes with long sleeves.

Signer-dependent recognition rates average 97.9%. Most information is carried by x, y and their derivatives, which account for approx. 92%. Signer-independent performance in a leaving-one-out test with no adaptation sets a baseline of 55.5%.

4. Signer adaptation

Selected adaptation methods from speech recognition are modified for the use in sign language recognition tasks to improve the performance of the signer-independent recognizer.

A set of adaptation data consisting of isolated signs is collected from the unknown signer, either supervised with known transcription or unsupervised. In the latter case, the signer-independent recognizer estimates a transcription, using a confidence measure to assess the quality of the recognition result as shown in Fig. 2.

Based on the adaptation data, the adaptation process reduces the mismatch between signer-independent models and observations from the unknown signer.

4.1. Choice of adaptation methods

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.

While *feature-based* methods such as Vocal Tract Length Normalization require knowledge from the speech production domain, *model-based* approaches are well suited for adapting the recognition system.

Model-based adaptation alters the parameters of the underlying HMMs based on the given adaptation data. Two methods are evaluated: Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) estimation. Both are employed in current speech recognition systems and have proven to perform excellent in the speech domain.

The evaluated approaches are introduced below, along with necessary modifications for signer adaptation.

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 [3].

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 [3], 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. The root contains all mixture components, yielding a global transformation W. The sons of a node form a partition of the father class, so deeper nodes yield more specialized transformations derived from fewer components. As more adaptation sequences become available, deeper transformations can be robustly estimated.

If the direct estimation of a transformation for a certain node is not possible for numerical reasons, computationally expensive techniques as described in [3] can be used. Alternatively, the next node on the path to the root can be chosen, yielding a more general but numerically stable transform.

The approach is adapted to sign language recognition using explicit handling of signs that are only performed with one hand and a method for transforming models that have not been observed the in the adaptation data.

One-hand transformations The corpus contains several signs where only the dominant hand is active during the whole sequence. It is presumed that the right hand is always

dominant, as features from left-handed signers are mirrored. Thus the feature extraction yields a feature vector sequence $[x_1, \ldots, x_T]$, where for single-handed signs the entries of the non-dominant hand of each feature vector $x_t \in \mathbb{R}^{D+D}$ equal zero:

$$x_t = \begin{bmatrix} 0 & \ldots & 0 & x_{t,1} & \ldots & x_{t,D} \end{bmatrix} \tag{4}$$

Here, $x_{t,d}$ is the d-th feature of the dominant hand. If HMMs are trained with such sequences, the mean vectors of the resulting mixture components have the same special form. As the adapted models should be of the same form, dedicated *one-hand transformations* are introduced.

Each class of the Regression Class Tree containing only one-hand mixture components is marked as a one-hand class. The sons of such a class again represent one-hand classes as they form a partition of the father node. Thus each one-hand class defines a *one-hand subtree* containing only one-hand classes.

A sample Regression Class Tree is shown in Fig. 4. The root node contains all components, represented by their mean vectors. These are either collected from one-hand or two-hand models. If a created node contains only one-hand means during tree construction, the whole subtree defined by that node will contain only one-hand classes. Such one-hand subtrees can make up a large part of the whole Regression Class Tree.

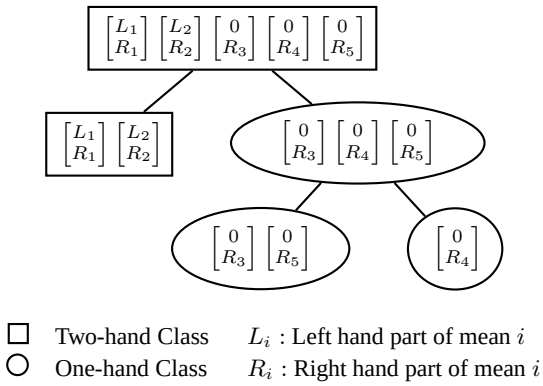

Figure 4. One-hand classes as part of the Regression Class Tree.

The first half of a Gaussian parameter corresponding to a one-hand mixture component contains only zero entries and is therefore ignored during the adaptation process. Transformations for classes that are part of a one-hand subtree are estimated from the one-hand versions of the corresponding Gaussian parameters, consisting only of the second half of mean and variance.

The use of one-hand transformations guarantees that the features for the passive hand remain passive after the transformation. Complexity of the estimation process is halved in the one-hand case due to the dimensionality reduction.

Handling of unseen signs Sign models are called *seen* or *unseen*, depending on whether they are observed in adaptation data or not. The mixture components of an unseen HMM are transformed based on the seen components of the regression class they belong to. Although this works for large and general regression classes near the root of the tree, specialized transformations for small classes towards the tree leaves tend to produce unsatisfying results. As the transformations are highly optimized for the seen components, the unseen components are not adapted well.

Reducing the tree size would result in broader regression classes at the tree leaves and the most special possible transformations would still be applied to a large amount of mixture components. If these general transformations are used even if more adaptation sequences become available, the effect of MLLR saturates after a certain amount of data. Thus a special handling of the unseen components is proposed.

Not updating the unseen components at all degrades the quality of the adapted models in terms of recognition accuracy. After the transformation, the mean parameters of seen components are much closer to the range of the observations from the unknown signer than the parameters of unseen components. Thus the Viterbi score of a model corresponding to a seen sign is likely to be higher than the score of an unseen model, so the recognizer prefers seen models in general.

This can be solved by using general transformations only for unseen components. The seen components are adapted using the most special transformation that can be robustly estimated using the Regression Class Tree, while unseen components are adapted using a global transformation estimated at the root node of the tree.

4.3. 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 the 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 [7], the MAP estimate is

$$\tilde{\mu}_{\text{MAP}} = \frac{\tau}{\tau + N} \cdot \mu_m + \left(1 - \frac{\tau}{\tau + N}\right) \cdot \bar{x}_m \tag{5}$$

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.

5. Experimental results

The adaptation experiments were carried out on the corpus described in section 3. Three signers are used for training the signer-independent model, one signer is used for testing. All results given are average values from the four possible combinations.

Only the Gaussian means were updated by MAP and MLLR, variances and mixture weights remain unchanged as the mean covers most of the variability between the speakers [8]. The results below are derived using Gaussian single densities, experiments with Gaussian mixtures show the same behavior due to the small training population.

Explicit one-hand transformations were used in all MLLR experiments. The conventional full-dimensional approach cannot estimate transformations for one-hand classes without using computationally expensive pseudo-inverses. Moreover, full-dimensional transformation of a one-hand parameter would yield a two-hand mean due to the translation of the whole vector.

5.1. Supervised adaptation

For the supervised experiments, variations 1 to 4 were used for static adaptation with different amounts of adaptation data while variation 0 was reserved for testing.

Fig. 5 illustrates the effect of the proposed methods for handling mixture components from unseen signs. Seen components were adapted with the most special transform from the Regression Class Tree in all three experiments. As described, transforming the unseen components with a global transformation outperforms the conventional approach and is superior to ignoring the unseen components during adaptation. Thus the MLLR approach is suited for rapid signer adaptation using only a small amount of adaptation data.

The combination of modified MLLR and standard MAP as shown in Fig. 6 results in the same effect which has been observed in speech recognition: the rapid adaptation using MLLR is preserved, while its saturation is compensated by MAP.

Table 1 summarizes the supervised adaptation experiments, showing the recognition performance of adapted models using the different methods. MLLR followed by MAP yields the best models, regardless of the number of adaptation sequences. Using class-based MLLR, rapid adaptation to an unknown signer is possible without covering the whole vocabulary during adaptation as described in [9], which only applies MAP adaptation.

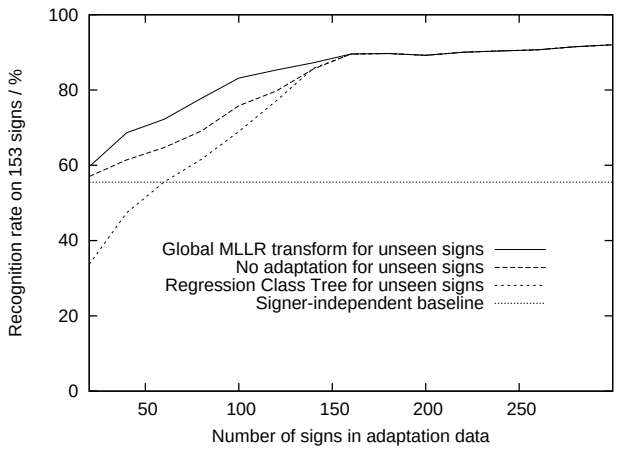

Figure 5. Handling of unseen signs, supervised MLLR adaptation.

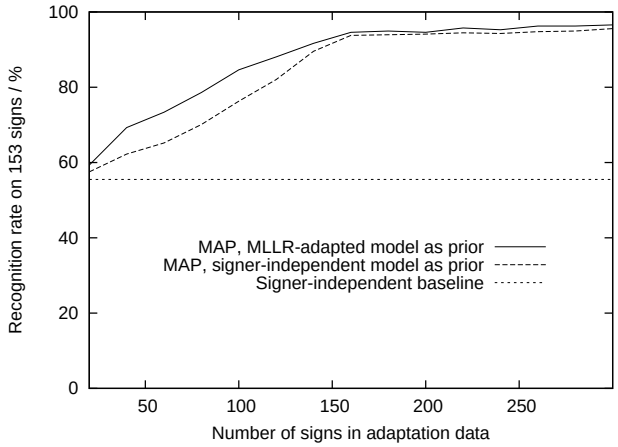

Figure 6. Supervised MAP: Signer-independent vs. MLLR-adapted model as prior.

Method	Recognition rate / % on 153 signs Number of adaptation utterances		
	80	160	320
Signer-Dep.	97.9	97.9	97.9
Signer-Indep.	55.5	55.5	55.5
MAP	70.1	93.8	95.9
MLLR	77.8	89.5	91.7
MLLR→MAP	**78.6**	**94.6**	**96.9**

Table 1. Supervised adaptation.

5.2. Unsupervised adaptation

Unsupervised experiments were carried out using all five variations from the testing signer as both adaptation and testing data. Each recognized transcription was stored for adaptation together with the input feature sequence. The incremental adaptation of the models was then performed at

an interval of 20 collected signs, yielding updated models for the recognition of the next input signs. Table 2 shows the overall performance using different adaptation setups.

Only the correctly recognized signs should contribute to the actual adaptation. The output of the current signer-independent recognizer is wrong in almost half of the cases. Even if we knew exactly which signs were recognized correctly and which were not, the amount of available adaptation sequences would be halved, and some signs will never be recognized correctly. Therefore the recognition rate remains at a level of about 73%, even in the case of an ideal confidence measure.

If no confidence measure is applied at all, then all 765 signs are used for adaptation. Hence a huge amount of signs is supplied with a wrong transcription, disturbing the adaptation process. Fortunately, MLLR can cope with this effect to some extent, as the components of the correct model often fall in the same regression class as the components of the recognized model.

A simple confidence measure for isolated word recognition as described in [4] was applied to reject unlikely recognition results. Using this confidence measure, results improve towards the ideal value.

Method	Confidence measure	Recognition rate / % on 765 signs
Signer-Indep.	N/A	53.5
MLLR→MAP	Ideal	72.9
MLLR→MAP	No	58.3
MLLR→MAP	N-Best-based	61.6

Table 2. Unsupervised adaptation.

6. Conclusion

Applying adaptation methods from speech recognition in a sign language context yields significant performance improvements. The proposed modified MLLR approach allows rapid adaptation of a signer-independent system, preserving the structure of one-hand models. The combination with MAP results in high accuracy for larger sets of adaptation data. Supervised adaptation with 80 adaptation sequences yields a recognition accuracy of 78.6%, which is a relative improvement of 41.6% compared to the signer-independent baseline. In the unsupervised case, a relative improvement of 15.6% is obtained without requiring an explicit enrollment session from each new signer.

The most crucial problem for unsupervised adaptation is the performance of the signer-independent models. More training speakers are required to improve the baseline, and other confidence measures have to be evaluated to robustly reject incorrect recognition results.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation). The video material was kindly provided by the British Deaf Association.

References

[1] B. Bauer, H. Hienz, and K.-F. Kraiss. Video-Based Continuous Sign Language Recognition Using Statistical Methods. In *Proceedings of the 15th IAPR International Conference on Pattern Recognition*, 2000.

[2] G. Fang and W. Gao. A SRN/HMM System for Signer-independent Continuous Sign Language Recognition. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[3] M. Gales and P. Woodland. Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech and Language*, 10:249–264, 1996.

[4] G. Hernandez-Abrego, X. Menendez-Pidal, and L. Olorenshaw. Robust and efficient confidence measure for isolated command recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 449–452, 2001.

[5] M. J. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. Technical Report CRL 98/11, Compaq Cambridge Research Lab, December 1998.

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

[7] 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 Transactions on Acoustics, Speech, and Signal Processing*, 39(4):806–814, April 1991.

[8] C. J. Leggetter. *Improved Acoustic Modelling for HMMs using Linear Transformations*. PhD thesis, Cambridge University, 1995.

[9] S. C. W. Ong and S. Ranganath. Deciphering Gestures with Layered Meanings and Signer Adaptation. In *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[10] S. C. W. Ong and S. Ranganath. Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning. *IEEE TPAMI*, 27(6):873–8914, June 2005.

[11] C. Wang, W. Gao, and S. Shan. An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[12] L.-G. Zhang, Y. Chen, G. Fang, X. Chen, and W. Gao. A vision-based sign language recognition system using tied-mixture density HMM. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 198–204, New York, NY, USA, 2004. ACM Press.

[13] J. Zieren and K.-F. Kraiss. Robust Person-Independent Visual Sign Language Recognition. In *Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis*, volume Lecture Notes in Computer Science, 2005.