A Framework for Motion Recognition with Applications to American Sign Language and Gait Recognition

Christian Vogler, Harold Sun, and Dimitris Metaxas
Vision, Analysis, and Simulation Technologies Laboratory
Department of Computer and Information Science, University of Pennsylvania
200 S. 33rd Street
Philadelphia, PA 19104–6389
{cvogler,hsun}@gradient.cis.upenn.edu, dnm@central.cis.upenn.edu

Abstract

Human motion recognition has many important applications, such as improved human-computer interaction and surveillance. A big problem that plagues this research area is that human movements can be very complex. Manging this complexity is difficult. We turn to American Sign Language (ASL) recognition to identify general methods that reduce the complexity of human motion recognition.

In this paper we present a framework for continuous 3D ASL recognition based on linguistic principles, especially the phonology of ASL. This framework is based on parallel Hidden Markov Models (HMMs), which are able to capture both the sequential and the simultaneous aspects of the language. Each HMM is based on a single phoneme of ASL. Because the phonemes are limited in number, as opposed to the virtually unlimited number of signs that can be composed from them, we expect this framework to scale well to larger applications.

We then demonstrate the general applicability of this framework to other human motion recognition tasks by extending it to gait recognition.

1. Introduction

Human motion recognition is a field with a wide variety of applications. Of particular interest are gesture recognition for new modes of human-computer interaction, and gait recognition for video surveillance systems and intrusion detection. These applications share a common problem: Human movements can be very complex, with many actions taking place both sequentially and simultaneously. As an example of sequential complexity, consider a gesture that consists of a complex series of hand movements. As an example of simultaneous complexity, consider a human handing an object over to another human, while walking at the same time. Likewise, when a human performs a complex gesture, he could use both hands to perform two different actions at the same time.

Because there are so many different combinations of sequential and simultaneous human movement actions, it is impossible to model them all explicitly. We elaborate on this problem in Sec. 3.1 and Sec. 3.2. For this reason, a comprehensive framework for human motion recognition must provide a way to reduce the complexity of the problem. An obvious approach is to break down the actions into smaller primitives that are powerful enough to be combined into any conceivable action. Unfortunately, we have little data on what these primitives are for most human motion recognition applications, because they are relatively unconstrained.

American Sign Language (ASL) recognition yields valuable insights into the problem of managing complexity. Unlike most other motion recognition applications, ASL recognition is highly structured and constrained, thanks to the status of ASL as a language. Furthermore, the linguistics of ASL have been extensively researched (e.g., [13]), which helps us identify the primitives ("phonemes") of ASL. For this reason, it is beneficial to research ASL recognition first before applying the results to other research areas.

In this paper we describe a novel and extensive framework for continuous ASL recognition based on an extension to hidden Markov models (HMMs). The main contributions of this work are (1) modeling each sign in terms of its constituent phonemes, thus handling sequential complexity; (2) reducing simultaneous complexity by modeling signs in terms of independent channels and recognizing them with parallel HMMs, which are essentially regular HMMs applied to several channels simultaneously; and (3) recognizing signed sentences from full-fledged 3D data, which we collect either with a magnetic tracking system, or with 3D computer vision methods [7].

To demonstrate that the ASL recognition framework can be generalized, we discuss its application to gait recognition. Although gait and ASL are two very different areas,

0-7695-0939-8/00 $10.00

we show that many concepts are similar. These similarities allow us to carry over the framework with virtually no modifications.

The rest of the paper is organized as follows: We discuss related work, then provide an overview on the phonological structure of ASL and its sequential and simultaneous aspects. We then describe how to model these aspects with parallel HMMs and provide experiments to verify our approach. We then generalize the framework to gait recognition. In the concluding remarks we discuss briefly what the framework has accomplished and provide an outlook for future work.

2. Related Work

Much previous work has focused on isolated sign language recognition with clear pauses after each sign, although the research focus is slowly shifting to continuous recognition. These pauses make it a much easier problem than continuous recognition without pauses between the individual signs, because explicit segmentation of a continuous input stream into the individual signs is very difficult. For this reason, and because of coarticulation effects, work on isolated recognition often does not generalize easily to continuous recognition.

Some isolated recognition work used neural networks [3, 16]. Other work focused on computationally inexpensive methods [6].

Most work on continuous sign language recognition is based on HMMs, which offer the advantage of being able to segment a data stream into its constituent signs implicitly. It thus bypasses the difficult problem of segmentation entirely.

T. Starner and A. Pentland used a view-based approach with a single camera to extract two-dimensional features as input to HMMs with a 40-word vocabulary and a strongly constrained sentence structure [12]. They assumed that the smallest unit in sign language is the whole sign. This assumption leads to scalability problems, as vocabularies become larger.

H. Hienz and colleagues used HMMs to recognize a corpus of German Sign Language [5] with 2D-based methods. They also experimented with stochastic bigram language models to improve recognition performance. The results of using stochastic grammars largely agreed with our results in [14].

Y. Nam and K. Y. Wohn [10] used three-dimensional data as input to HMMs for continuous recognition of gestures. They introduced the concept of movement primes, which make up sequences of more complex movements. The movement prime approach bears some superficial similarities to the phoneme-based approach in this paper.

R. H. Liang and M. Ouhyoung used HMMs for continuous recognition of Taiwanese Sign Language with a vocabulary between 71 and 250 signs. [8] Unlike other work in this area, they did not use the HMMs to segment the input stream implicitly. Instead, they segmented the data stream explicitly based on discontinuities in the movements. They integrated the handshape, position, orientation, and movement aspects at a higher level than the HMMs.

We used HMMs and 3D computer vision methods to model phonological aspects of ASL with an unconstrained sentence structure [14]. In [15] we extended the conventional HMM framework to capture the parallel aspects of ASL, which ordinarily would make the recognition task too complex.

3. Overview of the Framework

We now discuss the two main aspects of our framework, which are the linguistic modeling of ASL, and the modeling of the sequential and simultaneous aspects of ASL with a novel HMM-based approach. Although taking advantage of research into linguistics to model the signs is specific to signed languages, the principal idea of breaking down larger units into their constituent parts applies to other recognition tasks. Likewise, the HMM framework can be applied to other recognition tasks without alterations. In Sec. 5 we show by example of gait recognition how to extend our framework to other applications.

3.1. ASL Linguistics

ASL is the primary mode of communication for many deaf people in the USA. It is a highly inflected language; that is, many signs can be modified to indicate subject, object, and numeric agreement. They can also be modified to indicate manner (fast, slow, etc.), repetition, and duration [13]. Like all other languages, ASL has structure, which sets it clearly apart from most other human motion recognition problems. It allows us to test ideas in a constrained framework first, before attempting to generalize the results.

In particular, managing the complexity of large data sets is an area where ASL recognition work can yield valuable insights. Managing complexity is already difficult in the constrained field of ASL recognition, because signs can appear in many different forms, both sequentially and simultaneously. Other human motion recognition applications are often much less constrained than ASL, so this problem will only be exacerbated. It is, therefore, important to develop methods that make the complexity of ASL, and, by extension, other human motion recognition problems manageable.

The key idea behind managing complexity is that actions can be broken down into smaller subunits, and that any action can be described in terms of these subunits. In the case of ASL these subunits are called **phonemes**[1]. Formally, a phoneme is defined to be the smallest contrastive unit in a language. In English, examples of phonemes are the sounds */c/*, */a/*, and */t/*. In ASL, examples of phonemes are the

[1] Some people prefer to associate the term "phoneme" with spoken languages only, and use the term "chereme" for sign languages. We follow the terminology of spoken language linguistics, because the underlying concepts are the same.

Figure 1. The sign for "father." The white X indicates contact between the thumb and the forehead after each tap. The location of the hand at the forehead and the tapping movements are examples of phonemes.

movement of the hand toward the chin in the sign for "FATHER," and the starting location of the hand in front of the forehead at the beginning of this sign (Fig. 1).

Phonemes are *limited in number,* as opposed to the *virtually unlimited number* of words or signs that can be constructed from them. In English, there are approximately 40 distinct phonemes, whereas in ASL there are approximately 150–200 distinct phonemes[2]. For this reason, taking advantage of phonology can make an otherwise intractable modeling task feasible. It is practical to provide enough training data for a small set of phonemes, from which every sign can be constructed. Doing the same for signs that are not modeled in terms of phonemes would become impossible with vocabularies larger than a few hundred signs.

Unlike in spoken language linguistics, sign language linguists have not yet agreed on a common phonological model for ASL. Surveying all of the different phonological models is beyond the scope of this paper. We now briefly describe the one that we use in our recognition framework.

The Movement-Hold Model The Movement-Hold model [9] assumes that signs can be broken down into two major types of segments, which are called **movements** and **holds**. Movements are those segments, during which some aspect of the signer's configuration changes, such as a change in handshape, or a hand movement from one location to another. Holds, in contrast, are those segments, during which the hands remain translationally stationary.

Signs are made up of sequences of movements and holds. A very common sequence is *MMMH* (three movements followed by a hold), such as in the sign for "FATHER" (Fig. 1). This sign starts out with a movement toward the forehead, then away from the forehead, toward the forehead again, followed by a hold touching the forehead. Attached to each segment is a bundle of articulatory features, which primarily describe the handshape, orientation, and location of each segment. Fig. 2 shows a schematic example.

For a detailed description of all the existing phonemes in the Movement-Hold model, see [9]. For a detailed description of the phonemes that we have used so far in our framework, see [15].

Simultaneous Aspects of ASL The Movement-Hold model is ideally suited for ASL recognition, because it emphasizes sequential aspects over simultaneous aspects. This emphasis fits HMMs very well, because they are sequential in nature. Yet, despite the emphasis on sequentiality, a lot of phonemes also occur simultaneously. For example, often the handshape changes simultaneously with the hand movement in a sign. Likewise, many signs are two-handed, and both hands move simultaneously. A purely sequential framework cannot capture this kind of simultaneity.

A look at the Movement-Hold model immediately suggests an approach to incorporating simultaneity into the framework by modeling all possible combinations of segments and feature bundles. This approach fails because of the sheer number of possible combinations of phonemes. If we consider both hands, and assume 30 basic handshapes, 8 hand orientations, 8 wrist orientations, and 20 major body locations [9], the total number of phoneme combinations is $(30 \times 8 \times 8 \times 20)^2 \approx 1.5 \times 10^9$. Even if we employ some constraints on the weak hand for two-handed signs, the number is still approximately 2.9×10^8 [15]. It would be impossible to get enough training data for 10^9 models.

This problem is not unique to sign language recognition. Many other motion recognition applications, such as gestures and full human body movement, are even worse off, because they are less constrained than ASL. For this reason, a different approach toward handling simultaneous processes is necessary.

For this reason, we make a major modification to the Movement-Hold model. Instead of attaching the bundles of articulatory features to the movement and hold segments,

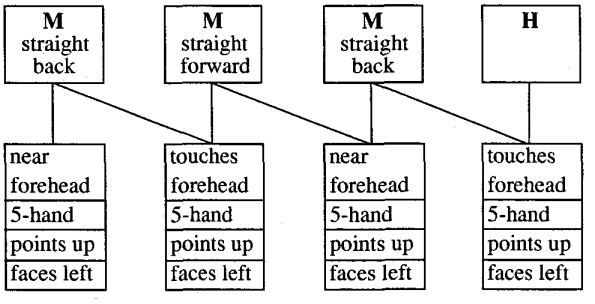

Figure 2. Schematic description of the sign for "FATHER" in the Movement-Hold model. It consists of three movements, followed by a hold (compare Fig. 1).

[2]This number applies to the the Movement-Hold phonological model [9] described in Section 3.1. The numbers for other models vary slightly.

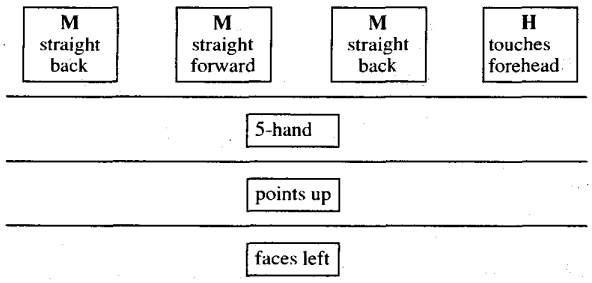

Figure 3. The sign for "father," where the different features are modeled in separate channels. Compare with Fig. 2.

we break them up into **channels** that are independent from one another. One channel consists of movements and hold segments that describe the type of movement and the body locations of the right ("strong") hand. Other channels could consist of the segments in the left ("weak") hand, the handshape, the hand orientation, and the wrist orientation. Figure 3 shows how the sign for "FATHER" is represented with this modification.

By modeling the simultaneous aspects of ASL as independent channels, we gain the ability to model each channel separately, yet combine each channel on the fly during recognition of a signed sentence. The success of this approach depends primarily on how independent the channels are from one another in reality. In the case of ASL, there is some linguistic evidence that the strong and weak hands move independently from each other [9]. Our experiments in Sec. 4 suggest that the independence assumption is at least partially valid.

3.2. Recognition with Hidden Markov Models

One of the main challenges in ASL recognition is to capture the variations in the signing of even a single human. Hidden Markov models (HMMs) are a type of statistical model embedded in a Bayesian framework and thus well suited for capturing these variations. In addition, their state-based nature enables them to describe how a signal changes over time.

An HMM λ consists of a set of N states $S_1, S_2, \ldots, S_N$. At regularly spaced discrete time intervals, the system transitions from state S_i to state S_j with probability a_{ij}. The probability of the system initially starting in state S_i is π_i. Each state S_i generates output $O \in \Omega$, which is distributed according to a probability distribution function $b_i(O) = P\{\text{Output is } O | \text{System is in } S_i\}$. In most recognition applications $b_i(O)$ is a mixture of Gaussian densities.

We use one HMM per phoneme, which are then chained together to form the signs. The individual signs in turn are chained together into a network. Then the recognition problem is reduced to finding the most likely state sequence through the network that could have generated the input signal with the signs to be recognized. From the state sequence the sequence of signs, and hence the recognized sentence, can be recovered.

The Baum-Welch algorithm is used to train HMMs on a set of training data in polynomial time, and the Viterbi algorithm is used to find the most likely state sequence in polynomial time through a network of HMMs during the recognition phase. For details on these algorithms, see [11].

Modeling Simultaneous Aspects Regular HMMs, as we have described them so far, can model the sequential aspects of the Movement-Hold model well, but they are not suitable for modeling the simultaneous aspects. In the past, researchers have suggested using factorial hidden Markov models [4] and coupled hidden Markov models [2]. Although these two approaches are good at capturing simultaneous, coupled processes, they would still require *a priori* knowledge of all the possible phoneme combinations at training time. In other words, they do not address the underlying problem, which is the sheer number of possible phoneme combinations.

Instead, we introduce Parallel HMMs (PaHMMs) as a modification to the HMM framework that directly reflects the decomposition of the simultaneous aspects of ASL into independent channels, as described in Sec. 3.1. We model each channel separately with HMMs and train them separately. At recognition time, PaHMMs combine the probabilities from each channel by multiplying them. That is, PaHMMs are essentially regular HMMs that are used in parallel. We describe the details of this approach and the algorithms for PaHMMs in [15].

Because the channels are independent, the complexity problems with the number of possible phoneme combinations disappear. With PaHMMs we can train the phonemes in each channel separately and put together new, previously unseen combinations of phonemes on the fly. Thus, during the training phase, we need only enough data for a robust estimate of the HMMs' parameters for each phoneme, instead of all combinations of these.

4. Experiments

We ran several continuous recognition experiments with 3D data to test the feasibility of modeling the movements of the left and the right hands with PaHMMs. We used two channels, which modeled the movements and holds of the left and the right hands, respectively. Our database consisted of 400 training sentences and 99 test sentences over a vocabulary of 22 signs. The transcriptions of these signs are listed in [15].

We collect the sentences with an Ascension Technologies MotionStar™ 3D tracking system, and with our vision-based tracking system at 60 frames per second. The latter uses physics-based modeling to track the arms and the hands of the signer, as depicted in Figure 4. The physics-based models are estimated from the images from a subset of three orthogonal cameras. These are selected on a

Figure 4. These images show the 3D tracking of the sign for "father."

Regular HMMs		
Level	Accuracy	Details
sentence	80.81%	H = 80, S = 19, N = 99
sign	93.27%	H=294, D=3, S=15, I=3, N=312

Parallel HMMs		
Level	Accuracy	Details
sentence	84.85%	H = 84, S = 15, N = 99
sign	94.23%	H=297, D=3, S=12, I=3, N=312

Table 1. Results of the recognition experiments. H denotes the number of correct sentences or signs, D the number of deletion errors, S the number of substitution errors, I the number of insertion errors, and N the total number of sentences or signs in the test set.

per-frame basis depending on the occluding contour of the signer's limbs [7].

We used an 8-dimensional feature vector for each hand. Six features consisted of 3D positions and velocities relative to the base of the signer's spine. For the remaining two features, we computed the largest two eigenvalues of the positions' covariance matrices over a window of 15 frames centered on the current frame. In normalized form, these two eigenvalues provide a useful characterization of the global properties of the signal.

In the experiments we compared the recognition accuracy of modeling only the movements and holds of the right hand with regular HMMs and modeling both hands with PaHMMs. The results are given in Table 1 and show that one the sentence level, the difference in recognition accuracy between regular and parallel HMMs is significant. Hence, PaHMMs can make the recognition system more robust.

5. Extensions to Gait Recognition

Most of the framework for ASL recognition readily carries over to gait recognition. To test this hypothesis, we set up an experiment within our framework to discriminate among walking on level terrain, walking upward a slope, and walking downward a slope.

The basic unit in gait recognition is the half-step; that is, the time a leg takes to complete one of the stance or swing phases. A step consists of two half-steps. The first half-step

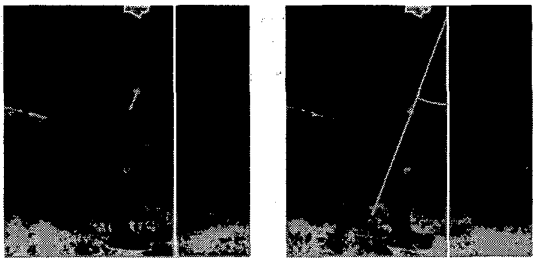

Figure 5. Sagittal elevation angles. We calculate them from the 2D positions of the markers at the sites indicated on the pictures.

models the leg during the stance phase, and the second one models the leg during the swing phase. The type of gait can change any time a half-step has been completed. Thus, concepts from ASL recognition have direct equivalents in gait recognition: A whole signs correpsonds to a step, and a phoneme corresponds to a half-step.

Before describing the experiment, we briefly cover how to represent gait data, which is very different from ASL data.

5.1. Data Representation

Elevation angles measure the orientation of a limb segment with respect to a vertical line in the world. We define the limb segment $\vec{v}$ between two points $\vec{a}$ and $\vec{b}$ on the body: $\vec{v} = \vec{a} - \vec{b}$. Typically $\vec{a}$ and $\vec{b}$ are points at opposite ends of a limb. The **sagittal elevation angles** are obtained by first projecting $\vec{v}$ onto the sagittal plane to form $v^{\vec{sag}}$. The angle between $v^{\vec{sag}}$ and the negative y axis is its sagittal elevation angle, ψ (Fig. 5).

We have followed the definition of elevation angles and placement of markers as used in [1], with the addition of a heel marker. Unlike joint angles and absolute coordinate values of the limbs, elevation angles are invariant with respect to different size humans. In addition, they appear to be invariant across different humans, as long as they perform the same kind of walking activity (e.g., walking on a level plain, walking on a slope) [1]. This property makes elevation angles a compelling choice for recognition features, especially for person-independent gait recognition.

5.2. Experiment

The task of the experiment was to discriminate among walking on level terrain, walking upward on slopes, and walking downward on slopes; as well as to identify the timing of the half-steps correctly. The slopes had different inclinations anywhere between 8 and 15 degrees. The shape of the terrain affects only the elevation angle of the foot, whereas the other angles appear to be unaffected. For this reason, we used the three elevation angles of the lower leg, the upper leg, and the pelvis as the feature vector.

We measured the elevation angles from a walking subject with the help of markers, as shown in Fig. 5. Future work could use our framework for tracking 3D body models [7], instead, to measure the elevation angles from any perspective. For the training set we used a set of ten measurements from a single person for each of level terrain, a 15 degree upward slope, and a 15 degree downward slope. Hence, we used a total of six HMMs — two for each type of step — chained together into a network. The sampling rate was 60 frames per second.

The test set contained the elevation angles of a person walking across uneven terrain. The recognizer was able to identify all half-steps in the test set correctly. The recognition of the timing of the steps worked well, as long as the type of step did not change. At transitions from one type of step to another, the recognizer often identified the end of the half step up to seven frames too early or too late. One possible explanation is that the elevation angles behave differently during a transition. In this case, modeling the transitions explicitly with HMMs, similar to modeling transitions between signs in sign language recognition [14], might improve the results.

6. Conclusions

We have developed a framework for human motion recognition. Although we initially applied it to ASL recognition, we have shown by example of gait recognition that it can be generalized to other recognition tasks. This makes our framework a promising contribution to the areas of human-computer interaction and video surveillance tasks.

Future work in ASL recognition should model other channels, such as handshape and orientation, and incorporate facial expressions, which constitute a large part of the grammar of ASL. It should also verify the framework with larger vocabularies. However, a prerequisite to experimenting with large vocabularies is a standardized corpus of ASL sentences. No such corpus exists at present.

Future work in gait recognition should model the transitions between different types of steps, incorporate more different types of steps (e.g., climbing a ladder or a stair), and model the differences between walking and running. It should also use 3D human body tracking, instead of measuring the sagittal elevation angles from the side with the help of markers.

Acknowledgments

This work was supported in part by an NSF Career Award NSF-9624604, ONR Young Investigator Proposal, NSF IRI-97-01803, AFOSR F49620-98-1-0434, and NSF EIA-98-09209.

References

[1] A. Borghese, L. Bianchi, and F. Lacquaniti. Kinematic determinants of human locomotion. *J. Physiology*, (494):863–879, 1996.

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[3] R. Erenshteyn and P. Laskov. A multi-stage approach to fingerspelling and gesture recognition. Proceedings of the Workshop on the Integration of Gesture in Language and Speech, Wilmington, DE, USA, 1996.

[4] Z. Ghahramani and M. I. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–275, 1997.

[5] H. Hienz, K.-F. Kraiss, and B. Bauer. Continuous sign language recognition using hidden Markov models. In Y. Tang, editor, *ICMI'99*, pages IV10–IV15, Hong Kong, 1999.

[6] M. W. Kadous. Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pages 165–174, Wilmington, DE, USA, 1996.

[7] I. Kakadiaris, D. Metaxas, and R. Bajcsy. Model based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *Proceedings of the CVPR*, pages 81–87, 1996.

[8] R.-H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, pages 558–565, Nara, Japan, 1998.

[9] S. K. Liddell and R. E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–277, 1989.

[10] Y. Nam and K. Y. Wohn. Recognition and modeling of hand gestures using colored petri nets. To appear in IEEE transactions on Systems, Man and Cybernetics (A), 1999.

[11] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[12] T. Starner, J. Weaver, and A. Pentland. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[13] C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington DC, 1995.

[14] C. Vogler and D. Metaxas. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 156–161, Orlando, FL, 1997.

[15] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of American Sign Language. To appear in *Computer Vision and Image Understanding*, 2001.

[16] M. B. Waldron and S. Kim. Isolated ASL sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–71, September 1995.