A Two-step Approach based on PaHMM for the Recognition of ASL

Jiangwen Deng and H.T. Tsui
Dept. of Electronic Engineering, The Chinese University of Hong Kong
{jgdeng,httsui}@ee.cuhk.edu.hk

Abstract

A main problem of American Sign Language (ASL) recognition is how to have good performance on a large vocabulary size with a limited set of training data. In this paper, the problem is tackled by a phoneme representation for hand movements and a 3D skeletal model for hand spellings. The overall recognition scheme consists of two steps. The first step is to select a few signs with similar movements and locations by using a small set of movement phonemes modeled by HMM (Hidden Markov Model). In the second step, a factored sampling process is used to approximate the posterior probability for posture recognition. Combining this information and that of the step 1 in a three-channel Parallel-HMM system, the candidate with the highest matching score is selected. Real experiments on 192 ASL signs show that our phoneme-based approach can achieve better performance than the original HMM approach in a smaller training set. The main advantage of our method is its real potential for application on a large vocabulary.

1. Introduction

A hand gesture or sign is composed of some global features, such as hand movements and locations, and local features, such as hand shapes and orientations. A hand shape at a given orientation is called a hand posture. A sign language recognition scheme has to deal with these spatial-temporal features [1]. There exist two approaches to acquire these features: instrumented glove-based and view-based. Obviously, the view-based approach is much more convenient for human.

Using statistical characterization of the signal, Hidden Markov Model (HMM) has been proved a successful tool to represent dynamic patterns in speech or on-line character recognition. Adopting it for gesture recognition has become popular in recent years. Starner and Pentland [2] used HMMs for ASL recognition in a 40-sign vocabulary with a strong grammar to constrain the solution. With the help of a color glove, Bauer and Hienz [5] introduced a system for continuous German Sign Language recognition on a 97-sign vocabulary.

One important issue of HMMs for gesture recognition is the scalability. The ASL vocabulary contains thousands of signs. Training and testing thousands of HMMs is very difficult and not practical. Vogler and Metaxas [3] made use of parallel HMMs, assuming that each hand movement is independent. They also try a phoneme approach for ASL recognition [6] with a 22-sign vocabulary. Hand postures have not been used in their approaches.

On the other hand, although global features may have a more important role in gesture recognition, local features can provide additional information for better performance. Most HMM-based approaches take the local features directly as the observation in HMMs, no matter the feature space is large [5] or not [2]. It means that Gaussian or multi-Gaussian distributions are used to model posture features. However, a gesture always contains variance or noise spatially and temporally. It becomes more serious after camera projection. It is difficult to model this kind of variants, especially when the training data are limited. The appearance-based scheme, such as the one by Cui and Weng [11], is to classify a posture to a number of clusters without hand-spelling interpretation. To explicitly analyze postures, a model-based approach is another way for the task. Mapping an image to a posture of the model is a kinematics problem. Rehg and Kanade [7] used a kinematic model to predict occlusions and windowed templates to track partially occluded objects. Wu and Huang [8] introduced a two-step iterative optimization to capture hand motions. Since the highly articulated human hand motions often have rotation, translation and self-occlusion, the optimization is often trapped in one of the local minima. Up to now, the use of hand model for large-scale ASL recognition has not yet been reported in the literatures.

Moreover, different channels of observation may provide different discriminative capabilities. The well-known Baum-Welch training is Maximum Likelihood (ML) based. It means that the HMM model parameters are tuned to fit the training data well, but not for the ability to discriminate them in general. Vogler and Metaxas [3] label the strong hand to the weak hand and manually assign different weights to them. Automatically determining and weighting the most discriminating states will improve the recognition result.

In this paper, a color-coded glove is used for feature extraction. There are three HMM channels: one for the global features of the right hand, one for those of the left hand and one for the local features (hand shape and hand orientation) of the right hand. Nine movement phonemes

are defined. They are modeled by HMMs. The N-best paths [13] instead of only the best Viterbi path are used to capture the movement variance. A two-step scheme is proposed for ASL recognition. In the first step, signs with the same phoneme list as one of the N-best paths matched to the observation are selected as the candidates. Then the start and end locations of each candidate are compared with those detected from the images. The mismatched signs are rejected, leaving a set of several candidate signs for further selection in step 2.

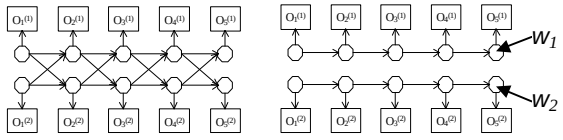
In step 2, posture information is evaluated for the remaining candidates. Instead of using a hand model with large degrees of freedom (DoF), a skeletal hand model is built for each hand spelling (hand shape). With fixed finger flexion, there are only 3 DoFs in each model. A 2-state Markov chain is used to model the posture changes for a given sign. Each state stores the corresponding hand spelling and its orientation distribution. Then the Bayesian technique of factored sampling [16] is applied to approximate the posterior density. Finally the evaluation scores of these three observation channels are weighted and summed according their discriminating power.

Section 2 of this paper introduces the application of Parallel Hidden Markov Models (PaHMMs) to model multi-process in ASL recognition. In section 3, hand skeletal models are built for hand spellings and the factored sampling is described for posture recognition. Section 4 gives the overview of the system architecture. In section 5, we evaluate our algorithm on a 192-sign vocabulary by real experiments. A conclusion is given in section 6.

2. HMM for ASL recognition

2.1. Modeling multi-process in HMMs

An ASL sign involves the motions of both hands and their posture sequences. These multi-processes take place simultaneously and may be modeled in several HMM channels. Coupled HMMs [9] model these channels dependently. On the other hand, parallel HMMs assume the channels to be independent [3]. The comparison of these two HMMs is given in Figure 1.



a. CHMMs: The states affect each other

b. PaHMMs: The states are independent

(w_i is the corresponding weight for each channel)

Figure 1. Comparison of CHMMs and PaHMMs

Vogler and Metaxas [3] show PaHMM is potentially more scalable than other HMM extensions. They model the motions of left and right hands and manually weigh the strong hand to the weak one. Following their interpretation, the goal is to find

$$\max_{Q^{(1)},...,Q^{(C)}} \{\log P(Q^{(1)},...,Q^{(C)},\ O^{(1)},...,O^{(C)} \mid l^{(1)},...,l^{(C)})\}, \quad (1)$$

where $Q^{(i)}$ is the state sequence of channel i with output sequence $O^{(i)}$ through the HMM $l^{(i)}$. The main characters of PaHMMs are given as following:

- Spatial independence:

$$\max_{Q^{(1)},...,Q^{(C)}} \{\log P(Q^{(1)},...,Q^{(C)},\ O^{(1)},...,O^{(C)} \mid l^{(1)},...,\ l^{(C)})\} =$$

$$\max_{Q^{(1)},...,Q^{(C)}} \{\sum_{i=1}^{C} \log P(Q^{(i)},O^{(i)} \mid l^{(i)})\}. \quad (2)$$

It means that, for PaHMMs, each channel can be evaluated separately. In our algorithm, we evaluate hand motions first, and the posture information can be recognized in the second stage. Assuming all channel outputs independent and summarizing them for the joint probability is valid, only on the condition that the models are accurate and the training data is sufficient. Neither case is hold in gesture recognition. So one channel may hold a lot more information than the others for the identity of a sign. So the idea here is to give a larger weight to a more discriminating channel, vice versa. Then Eq. (2) is rewrote as:

$$P^{w} = \sum_{i=1}^{C} w_i \log P(Q^{(i)},O^{(i)} \mid l^{(i)}) \cdot \quad (3)$$

- Temporal independence:

$$\max_{Q^{(1)},...,Q^{(C)}} \{\sum_{i=1}^{C} \log P(Q^{(i)},O^{(i)} \mid l^{(i)})\} =$$

$$\max_{Q^{(1)},...,Q^{(C)}} \{\sum_{i=1}^{C} \sum_{t} \log P(Q_t^{(i)},O_t^{(i)} \mid l_t^{(i)})\} \quad (4)$$

It means that each channel in PaHMMs can be further divided into any sub-segments. Each of these segments represents a movement phoneme in this paper.

2.2. Movement phonemes

It is well known that HMMs do not perform well without sufficient training data. Alternatively, phoneme-based approaches have been tested by some researchers [6] for ASL recognition. These systems are based on the magnetic tracking. In a view-based approach, a 3D motion is projected into a 2D image and great ambiguities arise. This is particularly serious, even with a frontal view, when the motion is in the direction of the optical axis of the camera. So if only considering the optimum path for movement from Viterbi, many possible cases will be discarded. However, our algorithm is divided into two steps since the posture recognition is heavy

computational. Finding the optimum path with all information will be out of the real-time range.

To tackle this problem, the most general strategy is the N-best paradigm, which uses a subset of the knowledge sources (KSs) to generate the *N* most likely hypotheses for the given observation. Then, each hypothesis can be further evaluated using the remaining KSs, and the most likely alternative is chosen finally [13].

We discretize a hand motion into eight directions and a hold phoneme is also built. The total motion-phoneme number is 9. Each of these phonemes is represented by a 3-state HMM. In fact, the movements of some signs are difficult for explicitly labeling by this simple set of phonemes. Only the signs without any ambiguous motion are used for training this set of HMMs. When these 9 HMMs are trained, the phoneme list for each sign is acquired from the optimum path for the given training data. It means that a sign may have more than one motion-phoneme list, which is determined from the training output.

In testing, the N-best paths are selected for further evaluation with other KSs, such as the start or end positions, those of the left hands and the posture information.

3. Posture recognition

Although global features may perform a more important role in gesture recognition, local features, posture, supplement the analysis, especially on a larger-scale vocabulary [5].

There are two popular approaches for posture recognition: the model-based approach and the appearance-based approach. Until now, most view-based ASL recognition schemes are appearance-based. Gaussian or multi-Gaussian distribution is used to model posture features. However, gesture always contains variance or noise spatially and temporally. It becomes more serious after camera projection. Let x denote the features of a given posture, Q be the hand model parameters, and F describe the projection of the hand model to feature space. The mapping, $x=F(Q)$, is not continuous. It means that a little variance in Q may cause great difference in the feature domain. If the viewing angel varies a lot, more training data are needed [12].

3.1. Skeletal hand model

Compared with the volumetric model, a skeletal hand model has the advantage of representing a hand with reduced degrees of freedom (DoF). Skeletal models are composed of joint angle parameters together with segment lengths. Then each part of the hand is represented by a stick with the corresponding kinematic relationship between them. We are currently using the skeletal models in Dorner and Hagen [10] with 23 degrees of freedom.

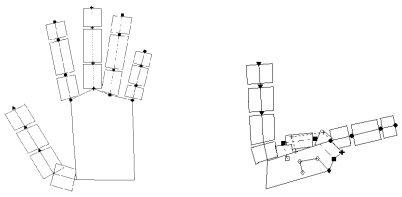
3.2. Projection of hand model

Orthographic projection of our hand model is assumed. The joint coordinate, X, in the world is projected into images:

$$x = t + s\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{bmatrix} \bullet X \qquad (5)$$

where x is the corresponding 2D-feature point, t and s are the camera parameters and they are assumed to be known as a priori.

The other difficulty is due to the self-conclusion of highly articulated hand. To deal with this problem, as shown in Figure 2, we introduce a 2D rectangle to represent each finger segment in the image.



a. '5' handshape b. 'L' handshape

Figure 2. Projection of finger segments
(Solid points are visible and hollow ones are invisible)

Just as shown in Figure 2, each segment is represented by a rectangle with 1cm width, irrelevant to the view angle. A palm plane is described by a polygon linking the wrist, the thumb's CMC joint and the MCP joints [10] of all other fingers. Let p represent a joint, z_p denote its depth. P is the set of the projection planes of finger segments that overlap p on image. So p may be occluded by them. Then

$$p \text{ is visible, iff } z_p \pounds z_i, \ \forall i \ni \Pi,$$

where z_i means the depth of the i^{th} finger segment.

3.3. Posture verification

Finding the parameters of hand involves inverse kinematics. Unfortunately, the human hand is a highly articulated object with rotation, translation and self-occlusion. The solution may not be unique and some constraints must be added. The optimization often gets trapped in local minima [8]. Moreover, for ASL recognition, the camera should zoom at the whole upper body area. Due to the low resolution of current video cameras, there are heavy errors in finding the joint position even when a color glove is used. To find the hand parameters under these conditions, a good initial posture

is always needed and the rotation of wrist should be limited.

What we concern here is whether the current input matches a given ASL sign. It means that the prior posture distribution for a given sign is known. Let Q denote a posture state vector, and O be the positions of color marks. The Bayesian factored sampling [16] is a random-sampling method to approximate the observation distribution $p(O|Q)$, if $p(O|Q)$ is too complicated to sample directly. Since the prior $p(Q)$ is known and can be sampled for each sign, the posterior density $p(Q|O)$ can be evaluated. Factored sampling generates a set of N samples $\{s^n\}$ according to the prior $p(Q)$, and then assigns to each sample a weight $p^n = p(O|Q=s^n)$ corresponding to the measurement density. Then the weighted set $\{s^n, p^n\}$ represents an approximation $\hat{P}(\Theta \mid O)$ to the desired posterior $p(Q|O)$, where a sample is drawn from $\hat{P}(\Theta \mid O)$ by choosing one of the s^n with probability p^n. As *N*®¥, $p(Q|O)$ could be arbitrarily closely approximated from $\hat{P}(\Theta \mid O)$.

With the help of a color-coded glove [10], the joint positions in image are assumed to be known. Every marker is a colored ring around the corresponding joint. Only one color is used for each finger.

Let S be a similarity measurement between the observed joint positions $O=\{o^{pi}\}$ and the projection $x=\{x^{pj}\}$ of the hand model X with 3 rotation parameters Q, where o^{pi} and x^{pj} individually mean the color marker position or the projection position of the i^{th} or j^{th} joint of the p^{th} finger. Then the observation probability is given by:

$$p(O \mid \Theta) = S(x, O) \, . \qquad (6)$$

We define

$$S(x,O) = \{\sum_k \exp(-d_k^2 / 2s^2) - an - bm\} / N \, ,$$

where n is the number of unmatched markers in the observation, m is the number of unmatched joints in the projected hand-skeletal model, N is the total number of joints including the visible joints in the skeletal model and the observation makers in the image, and a, b, s are some positive constants. d_k is the distance between the matched joints. So the observation probability is given as following:

- Let m=0, n=0 and k=0
- For each finger p.
  1. If the number of observation makers> the number of visible joints
     n=n+(the number of observation makers - the number of visible joints);

     Else
     m=m+(the number of visible joints - the number of observation makers);
  2. For each pair of joints between the observation markers and the visible joints, calculate their 2D Euclidian distances $D=\{d_{ij}\}$.
  3. While D<>NULL
     a. $k=k+1$; Let $d_k=min(D)$,
     b. Denote i, $j=argmin(D)$. Then delete the items in D with visible joint i or with observation maker j.

     End while.

  End for.
- Calculate the observation probability as Eq. (6).

For simplicity, the posture information of a sign are only stored in two Markov states for the start and end moments. It is not so-called hidden here. Now the prior rotation distribution of the given skeletal model in a state is assumed to be a mixture normal density:

$$\Theta \sim N(\overline{\Theta}, P) \, .$$

So, in testing, a serial of samples $\{s_u^n\}$ are generated from the prior density of the state u. Their corresponding $\{p_u^n\}$ are measured by Eq. (6). Then the probabilities of being this state u is given by:

$$p(u \mid O) = \sum_n p_u^n \, . \qquad (7)$$

In training, the prior density $p(Q)$ is a wide range average distribution. And hand parameters are estimated by MAP (Maximum A Posteriori):

$$\overline{\Theta} = \arg\max_{\Theta} \; p(\Theta \mid O) \, . \qquad (8)$$

4. System overview

As described before, our approach is a two-step algorithm. The first step is to select a few candidates with similar movements and locations. Then, within the candidate signs from step 1, pick out the one with a highest matching score by combining with the hand posture information and the information in step 1. A sign model is described in Figure 3:

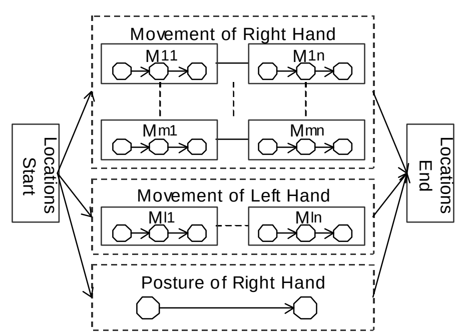


Figure 3. A sign model.

A gesture is modeled by PaHMMs. Each channel is represented by a box in the figure and it models the corresponding channel of features. A 'M' in the figure is a movement phoneme. Nine 3-state HMMs are built for these phonemes as described in section 2.2. The posture sequence of a gesture is also modeled by a 2-state Markov chain. In this paper, we assume that the right hand is the strong hand which plays a more important role in recognition. So only the right hand posture information is evaluated by a 2-state Markov chain.

Different channels of observation may provide different discriminative capabilities. The well-known Baum-Welch training is Maximum Likelihood (ML) based. It means that the model parameters are tuned to fit the training data well, but not for the ability to discriminate them in general. A simple multiplication of probability for each channel is not so suitable if the model is not so accurate or the trained data are not sufficient. So we make use of Fisher's discriminant analysis [17] to automatically determine and weigh the most discriminating states, and the recognition result will be improved in this way.

Word-Dependent N-best algorithm [13] is used to find the possible paths for the right-hand movement. According to this sequence of movement units, a list of gestures is selected. This list of gestures is further verified by their left hand movements and the hand positions at the start and end moments. The remaining candidates are ready for posture verification in step two. Finally, all the evaluation scores in each channel of PaHMMs are weighted, and the most likelihood one is pick out as the solution.

5. Experiments and discussion

The system runs on a Petium-III 800 PC. Following the textbook [15] on ASL, the vocabulary consists of 192 isolated signs. All the training and testing data are provided by two non-native signers, who learned the ASL from the textbook. Each sign is performed two times per person. Two of these are for training and the other two for testing. With the help of a color-coded glove, the right-hand features are assumed to be known. A fixed frontal view is used in capturing the data.

We first make a comparison between the phoneme-level and the word-level HMM approaches. The movements of both hands are considered. The word-level HMMs consist of 3 hidden states for each channel of each sign. The recognition result is given in Table 1. Although the word-level approach gives a very good recognition result in the training data set, it cannot catch the variance of the testing data even in a fixed frontal view because of the shortage of training data. Actually, the word-level HMMs are overfitting in this case. Even using 2-state HMMs cannot relief this problem.

Next, for posture recognition, two examples of the MAP estimation are shown in Figure 4. The skeletal hand model is not accurately fit to a given hand. Moreover, due to the low resolution and bias in joint location, there are errors between the observation and the estimation. Especially in case 2, the distortion in thumb is quite large, since the CMC joint of a thumb has only 2 DOF and cannot model the real thumb well. The result shows that our algorithm can still give an optimum solution. The factored sampling number N is 100 in the experiment. The posture recognition result is given in Table 2. In this case, the factored sampling is heavily computational and the recognition is not real-time. For comparison, we also evaluate the performance of an appearance-based approach, which scales the posture image into 25×25 and makes use of principal component analysis (PCA) [17]. Again the appearance-base approach is overfitting. On the other hand, the recognition result of the model-based approach in testing does not deteriorate compared with that in training.

Finally, all the evaluation scores in each channel of PaHMMs are added with a weigh vector according to their discriminative abilities. The recognition result is shown in Table 3. In the first step of recognition, about 13.5% of the signs cannot be detected by matching movements and locations. This is mainly due to the loss of depth information. The average number of candidate signs after step 1 is 2. On the average, it takes about 1 second to recognize one gesture. Without any program code optimization, it runs close to real time. The computation time increases only linearly with the vocabulary size. Thus, it is suitable for large vocabulary recognition even in a random search.

Table 1. Comparison of recognition rate between phoneme-level and word-level HMMs
(For right-hand movement only)

	Training data	Testing data
Phoneme-level	79.0%	77.0%
Word-level	91.0%	67.2%

Table 2. Appearance-based and model-based approach for posture recognition

	Training data	Testing data
Appearance-based	91.8%	66.2%
Model-based	45.6%	44.6%

Table 3 . Recognition rate of isolated signs

	Error in the 1st step	Avg. Candidate No. after 1st step	Recogniti on Result
Training	0.2%	1.85	98.7%
Testing	4.6%	1.78	93.3%

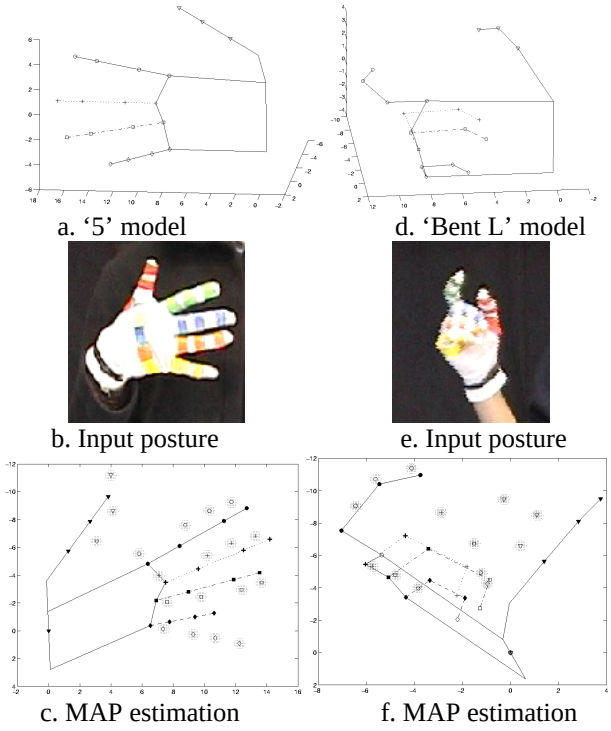


Figure 4. MAP estimation result.
(Solid points are visible, hollow ones are invisible, and shadow ones are observation feature points)

6. Conclusion

In this paper, we introduce a novel two-step model-based approach for ASL recognition. This is mainly for isolated signs at this stage. There are three main contributions of this paper. First, a set of 9 phonemes modeled by HMMs is proposed. This reduces the number of HMMs (for coding movement) to be trained from over several hundreds for word modeling to 9 for phoneme modeling. The experiments show that this approach performs better than the explicitly modeling of a sign when the set of training data is limited. Second, a three-channel PaHMM system is proposed to model the changes of the right and left hand positions and the changes of the postures of the right hand. They are balanced with a weigh vector such that a more discriminative channel plays a more important role for recognition. Third, a skeletal hand model is designed for each handspelling (hand posture). So the solution space is limited by the 3-DoF rotations of the hand. Then it is further constrained by the selected candidates in step 1. Then the optimum one is given by factored sampling. Although a color glove is required at the moment, our model-based approach may be adapted to other feature extraction scheme. With these properties, our algorithm is more suitable for a much larger vocabulary as computation time and training data size do not grow excessively with the vocabulary size as in most previous methods. The proposed method has been verified by comprehensive real experiments with the testing data provided by a number of different signers during a period of several weeks.

7. Acknowledgement:

This research is partially supported by the RGC grant CRC 4/98.

8. References

[1] V.I. Pavlovic, R. Sharma and T.S. Huang,*"Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review,"* IEEE PAMI, Vol. 19, No. 7, July 1997.

[2] T. Starner and A. Pentland, *"Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video,"* IEEE PAMI, Vol. 20, No. 12, Dec. 1998.

[3] C. Vogler and D. Metaxas, *"Parallel hidden Markov models for American sign language recognition," "*, Proc. IEEE Int'l Conf. Computer Vision , Vol. 1, pp. 116, 1999.

[4] L.R. Rabiner, *"A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,"* Proceedings of the IEEE, Vol.77, No. 2, Feb. 1989.

[5] B. Bauer and H. Hienz, *"Relevant features for video-based continuous sign language recognition"* Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 440-445, 2000.

[6] C. Vogler and D. Metaxas, *"Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes"*, gesture workshop, 1999.

[7] J.M. Rehg and T. Kanade, *"Model-Based Tracking of Self-Occluding Articulated Objects"* Proc. IEEE Int'l Conf. Computer Vision, pp. 612-617, 1997.

[8] Y. Wu and T.S. Huang, *"Capturing Articulated Human Hand Motion: A Divide-and-Conquer Approach"* Proc. IEEE Int'l Conf. Computer Vision, 1999.

[9] M. Brand, N. Oliver, and A. Pentland. *"Coupled hidden markov models for complex action recognition"*, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 1997.

[10] B. Dorner and E. Hagen, *"Towards an American Sign Language Interface"*, Artificial Intelligence Review, 8:235-253, 1994.

[11] Y. Cui and J. Weng, *"Appearance-Based Hand Sign Recognition from Intensity Image Sequences"*, Computer Vision and Image Understanding, vol. 78, pp. 157-176, 2000.

[12] Y. Wu, T.S. Huang, *"View-independent Recognition of Hand Postures"*, CVPR, 2000.

[13] C.H. Lee, F. K. Soong and K.K. Paliwal, *"Automatic Speech and Speaker Recognition"*, Kluwer Academic Publishers, 429-456.

[14] Lawrence Rabiner, Biing-Hwang Juang, *"Fundamentals of Speech Recognition"*, Prentice Hall, 1993.

[15] J. C. Hafer and R. M. Wilson, *"Come Sign With Us: Sign Language Language Activities for Children"*, Gallaudet University Press.

[16] B. D. Ripley, *"Stochastic simulation"*, New York: Wiley, 1987.

[17] S.S. Wilks, *"Mathematical Statistics"*, Wiley, New York, 1963.