Real-Time American Sign Language Recognition from Video Using Hidden Markov Models

Thad Starner and Alex Pentland
Room E15-383, The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge MA 02139
thad,sandy@media.mit.edu

Abstract

Hidden Markov models (HMM's) have been used prominently and successfully in speech recognition and, more recently, in handwriting recognition. Consequently, they seem ideal for visual recognition of complex, structured hand gestures such as are found in sign language. We describe two experiments that demonstrate a real-time HMM-based system for recognizing sentence level American Sign Language (ASL) without explicitly modeling the fingers. The first experiment tracks hands wearing colored gloves and attains a word accuracy of 99%. The second experiment tracks hands without gloves and attains a word accuracy of 92%. Both experiments have a 40 word lexicon.

Introduction

While there are many different types of gestures, the most structured sets belong to the sign languages. In sign language, each gesture already has assigned meaning, and strong rules of context and grammar may be applied to make recognition tractable.

To date, most work on sign language recognition has employed expensive wired "datagloves" which the user must wear (Takahashi & Kishino 1991). In addition, these systems have mostly concentrated on finger signing, in which the user spells each word with finger signs corresponding to the letters of the alphabet (Dorner 1993). However, most signing does not involve finger spelling but instead, gestures which represent whole words, allowing signed conversations to proceed at about the pace of spoken conversation.

In this paper, we describe an extensible system which uses one color camera to track hands in real time and interprets American Sign Language (ASL) using Hidden Markov Models (HMM's). The hand tracking stage of the system does not attempt a fine description of hand shape; studies of human sign readers have shown that such detailed information is not necessary for humans to interpret sign language (Poizner, Bellugi, & Lutes-Driscoll 1981; Sperling *et al.* 1985). Instead, the tracking process produces only a coarse description of hand shape, orientation, and trajectory. The hands are tracked by their color: in the first experiment via solidly colored gloves and in the second, via their natural skin tone. In both cases the resultant shape, orientation, and trajectory information is input to a HMM for recognition of the signed words.

Hidden Markov models have intrinsic properties which make them very attractive for sign language recognition. Explicit segmentation on the word level is not necessary for either training or recognition (Starner *et al.* 1994). Language and context models can be applied on several different levels, and much related development of this technology has already been done by the speech recognition community (Huang, Ariki, & Jack 1990). Consequently, sign language recognition seems an ideal machine vision application of HMM technology, offering the benefits of problem scalability, well defined meanings, a pre-determined language model, a large base of users, and immediate applications for a recognizer.

American Sign Language (ASL) is the language of choice for most deaf people in the United States. ASL's grammar allows more flexibility in word order than English and sometimes uses redundancy for emphasis. Another variant, English Sign Language, has more in common with spoken English but is not in widespread use in America. ASL uses approximately 6000 gestures for common words and finger spelling for communication of obscure words or proper nouns.

Conversants in ASL may describe a person, place, or thing and then point to a place in space to store that object temporarily for later reference (Sperling *et al.* 1985). For the purposes of this experiment, this aspect of ASL will be ignored. Furthermore, in ASL the eyebrows are raised for a question, relaxed for a statement, and furrowed for a directive. While we have also built systems that track facial features (Essa, Darrell, & Pentland 1994), this source of information will not be used to aid recognition in the task addressed here.

While the scope of this work is not to create a user independent, full lexicon system for recognizing ASL, the system should be extensible toward this goal. Another goal is real-time recognition which allows easier exper-

imentation, demonstrates the possibility of a commercial product in the future, and simplifies archiving of test data. "Continuous" sign language recognition of full sentences is necessary to demonstrate the feasibility of recognizing complicated series of gestures. Of course, a low error rate is also a high priority. For

Table 1: ASL Test Lexicon

part of speech	*vocabulary*
pronoun	I, you, he, we, you(pl), they
verb	want, like, lose, dontwant, dontlike, love, pack, hit, loan
noun	box, car, book, table, paper, pants, bicycle, bottle, can, wristwatch, umbrella, coat, pencil, shoes, food, magazine, fish, mouse, pill, bowl
adjective	red, brown, black, gray, yellow

this recognition system, sentences of the form "personal pronoun, verb, noun, adjective, (the same) personal pronoun" are to be recognized. This sentence structure emphasizes the need for a distinct grammar for ASL recognition and allows a large variety of meaningful sentences to be generated randomly using words from each class. Table 1 shows the words chosen for each class. Six personal pronouns, nine verbs, twenty nouns, and five adjectives are included making a total lexicon of forty words. The words were chosen by paging through (Humphries, Padden, & O'Rourke 1990) and selecting those which would generate coherent sentences when chosen randomly for each part of speech.

Machine Sign Language Recognition

Attempts at machine sign language recognition have begun to appear in the literature over the past five years. However, these systems have generally concentrated on isolated signs and small training and test sets. Tamura and Kawasaki demonstrate an early image processing system which recognizes 20 Japanese signs based on matching cheremes (Tamura & Kawasaki 1988). (Charayaphan & Marble 1992) demonstrate a feature set that distinguishes between the 31 isolated ASL signs in their training set (which also acts as the test set). More recently, (Cui & Weng 1995) have shown an image-based system with 96% accuracy on 28 isolated gestures.

(Takahashi & Kishino 1991) discuss a user dependent Dataglove-based system that recognizes 34 of the 46 Japanese kana alphabet gestures using a joint angle and hand orientation coding technique. The test user makes each of the 46 gestures 10 times to provide data for principle component and cluster analysis. A separate test set is created from five iterations of the alphabet by the user, with each gesture well separated in time. (Murakami & Taguchi 1991) describe a similar Dataglove system using recurrent neural networks. However, in this experiment a 42 static-pose finger alphabet is used, and the system achieves up to 98% recognition for trainers of the system and 77% for users not in the training set. This study also demonstrates a separate 10 word gesture lexicon with user dependent accuracies up to 96% in constrained situations.

Use of Hidden Markov Models in Gesture Recognition

While the continuous speech recognition community adopted HMM's many years ago, these techniques are just now accepted by the vision community. An early effort by (Yamato, Ohya, & Ishii 1992) uses discrete HMM's to recognize image sequences of six different tennis strokes among three subjects. This experiment is significant because it uses a 25x25 pixel quantized subsampled camera image as a feature vector. Even with such low-level information, the model can learn the set of motions and recognize them with respectable accuracy. (Darrell & Pentland 1993) dynamic time warping, a technique similar to HMM's, to match the interpolated responses of several learned image templates. (Schlenzig, Hunter, & Jain 1994) use hidden Markov models to recognize "hello," "good-bye," and "rotate." While Baum-Welch re-estimation was not implemented, this study shows the continuous gesture recognition capabilities of HMM's by recognizing gesture sequences. Recently, (Wilson & Bobick 1995) explore incorporating multiple representations in HMM frameworks.

Hidden Markov Modeling

While a substantial body of literature exists on HMM technology (Baum 1972; Huang, Ariki, & Jack 1990; Rabiner & Juang 1986; Young 1993), this section briefly outlines a traditional discussion of the algorithms. After outlining the fundamental theory in training and testing a discrete HMM, this result is then generalized to the continuous density case used in the experiments. For broader discussion of the topic, (Huang, Ariki, & Jack 1990; Starner 1995) are recommended.

A time domain process demonstrates a Markov property if the conditional probability density of the current event, given all present and past events, depends only on the jth most recent events. If the current event depends solely on the most recent past event, then the process is a first order Markov process. While the order of words in American Sign Language is not truly a first order Markov process, it is a useful assumption when considering the positions and orientations of the hands of the signer through time.

The initial topology for an HMM can be determined by estimating how many different states are involved in specifying a sign. Fine tuning this topology can be performed empirically. While different topologies can

be specified for each sign, a four state HMM with one skip transition was determined to be sufficient for this task (Figure 1).

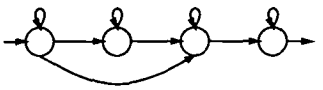

Figure 1: The four state HMM used for recognition.

There are three key problems in HMM use: evaluation, estimation, and decoding. The evaluation problem is that given an observation sequence and a model, what is the probability that the observed sequence was generated by the model ($Pr(\mathbf{O}|\lambda)$) (notational style from (Huang, Ariki, & Jack 1990))? If this can be evaluated for all competing models for an observation sequence, then the model with the highest probability can be chosen for recognition.

$Pr(\mathbf{O}|\lambda)$ can be calculated several ways. The naive way is to sum the probability over all the possible state sequences in a model for the observation sequence:

$$Pr(\mathbf{O}|\lambda) = \sum_{allS} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(O_t)$$

However, this method is exponential in time, so the more efficient forward-backward algorithm is used in practice. The following algorithm defines the forward variable α and uses it to generate $Pr(\mathbf{O}|\lambda)$ (π are the initial state probabilities, a are the state transition probabilities, and b are the output probabilites).

- $\alpha_1(i) = \pi_i b_i(O_1)$, for all states i (if $i \epsilon S_I, \pi_i = \frac{1}{n_I}$; otherwise $\pi_i = 0$)
- Calculating $\alpha()$ along the time axis, for $t = 2, ..., T$, and all states j, compute

$$\alpha_t(j) = [\sum_i \alpha_{t-1}(i) a_{ij}] b_j(O_t)$$

- Final probability is given by

$$Pr(\mathbf{O}|\lambda) = \sum_{i \epsilon S_F} \alpha_T(i)$$

The first step initializes the forward variable with the initial probability for all states, while the second step inductively steps the forward variable through time. The final step gives the desired result $Pr(\mathbf{O}|\lambda)$, and it can be shown by constructing a lattice of states and transitions through time that the computation is only order $O(N^2T)$. The backward algorithm, using a process similar to the above, can also be used to compute $Pr(\mathbf{O}|\lambda)$ and defines the convenience variable β.

The estimation problem concerns how to adjust λ to maximize $Pr(\mathbf{O}|\lambda)$ given an observation sequence $\mathbf{O}$. Given an initial model, which can have flat probabilities, the forward-backward algorithm allows us to evaluate this probability. All that remains is to find a method to improve the initial model. Unfortunately, an analytical solution is not known, but an iterative technique can be employed.

Using the actual evidence from the training data, a new estimate for the respective output probability can be assigned:

$$\bar{b}_j(k) = \frac{\sum_{t \epsilon O_t = v_k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

where $\gamma_t(i)$ is defined as the posterior probability of being in state i at time t given the observation sequence and the model. Similarly, the evidence can be used to develop a new estimate of the probability of a state transition ($\bar{a}_{ij}$) and initial state probabilities ($\bar{\pi}_i$).

Thus all the components of model (λ) can be re-estimated. Since either the forward or backward algorithm can be used to evaluate $Pr(\mathbf{O}|\bar{\lambda})$ versus the previous estimation, the above technique can be used iteratively to converge the model to some limit. While the technique described only handles a single observation sequence, it is easy to extend to a set of observation sequences. A more formal discussion can be found in (Baum 1972; Huang, Ariki, & Jack 1990; Young 1993).

While the estimation and evaluation processes described above are sufficient for the development of an HMM system, the Viterbi algorithm provides a quick means of evaluating a set of HMM's in practice as well as providing a solution for the decoding problem. In decoding, the goal is to recover the state sequence given an observation sequence. The Viterbi algorithm can be viewed as a special form of the forward-backward algorithm where only the maximum path at each time step is taken instead of all paths. This optimization reduces computational load and allows the recovery of the most likely state sequence. The steps to the Viterbi are

- Initialization. For all states i, $\delta_1(i) = \pi_i b_i(O_1)$; $\psi_i(i) = 0$
- Recursion. From $t = 2$ to T and for all states j, $\delta_t(j) = Max_i[\delta_{t-1}(i)a_{ij}]b_j(O_t)$; $\psi_t(j) = argmax_i[\delta_{t-1}(i)a_{ij}]$
- Termination. $P = Max_{s \epsilon S_F}[\delta_T(s)]$; $s_T = argmax_{s \epsilon S_F}[\delta_T(s)]$
- Recovering the state sequence. From $t = T-1$ to 1, $s_t = \psi_{t+1}(s_{t+1})$

In many HMM system implementations, the Viterbi algorithm is used for evaluation at recognition time. Note that since Viterbi only guarantees the *maximum* of $Pr(\mathbf{O}, S|\lambda)$ over all state sequences S (as a result of the first order Markov assumption) instead of the *sum* over all possible state sequences, the resultant scores are only an approximation. However, (Rabiner & Juang 1986) shows that this is often sufficient.

So far the discussion has assumed some method of quantization of feature vectors into classes. However, instead of using vector quantization, the actual probability densities for the features may be used. Baum-Welch, Viterbi, and the forward-backward algorithms can be modified to handle a variety of characteristic densities (Juang 1985). In this context, however, the densities will be assumed to be Gaussian. Specifically,

$$b_j(O_t) = \frac{1}{\sqrt{(2\pi)^n |\sigma_j|}} e^{\frac{1}{2}(O_t - \mu_j)' \sigma_j^{-1} (O_t - \mu_j)}$$

Initial estimations of μ and σ may be calculated by dividing the evidence evenly among the states of the model and calculating the mean and variance in the normal way. Whereas flat densities were used for the initialization step before, the evidence is used here. Now all that is needed is a way to provide new estimates for the output probability. We wish to weight the influence of a particular observation for each state based on the likelihood of that observation occurring in that state. Adapting the solution from the discrete case yields

$$\bar{\mu_j} = \frac{\sum_{t=1}^{T} \gamma_t(j) O_t}{\sum_{t=1}^{T} \gamma_t(j)}$$

and

$$\bar{\sigma_j} = \frac{\sum_{t=1}^{T} \gamma_t(j)(O_t - \bar{\mu_j})(O_t - \bar{\mu_j})^t}{\sum_{t=1}^{T} \gamma_t(j)}$$

For convenience, μ_j is used to calculate $\bar{\sigma_j}$ instead of the re-estimated $\bar{\mu_j}$. While this is not strictly proper, the values are approximately equal in contiguous iterations (Huang, Ariki, & Jack 1990) and seem not to make an empirical difference (Young 1993). Since only one stream of data is being used and only one mixture (Gaussian density) is being assumed, the algorithms above can proceed normally, incorporating these changes for the continuous density case.

Tracking Hands in Video

Previous systems have shown that, given some constraints, relatively detailed models of the hands can be recovered from video images (Dorner 1993; Rehg & Kanade 1993). However, many of these constraints conflict with recognizing ASL in a natural context, either by requiring simple, unchanging backgrounds (unlike clothing); not allowing occlusion; requiring carefully labelled gloves; or being difficult to run in real time.

In this project we have tried two methods of hand tracking: one, using solidly-colored cloth gloves (thus simplifing the segmentation problem), and two, tracking the hands directly without aid of gloves or markings. Figure 2 shows the view from the camera's perspective in the no-gloves case. In both cases color NTSC composite video is captured and analyzed at 320 by 243 pixel resolution. On a Silicon Graphics Indigo 2 with Galileo video board we can achieve a constant 5 frames per second, while using a Silicon Graphics 200MHz Indy workstation we were able to track the hands at 10 frames per second.

In the first method, the subject wears distinctly colored cloth gloves on each hand (a yellow glove for the right hand and an orange glove for the left) and sits in a chair facing the camera. To find each hand initially, the algorithm scans the image until it finds a pixel of the appropriate color. Given this pixel as a seed, the region is grown by checking the eight nearest neighbors for the appropriate color. Each pixel checked is considered part of the hand. This, in effect, performs a simple morphological dilation upon the resultant image that helps to prevent edge and lighting aberrations. The centroid is calculated as a by-product of the growing step and is stored as the seed for the next frame. Given the resultant bitmap and centroid, second moment analysis is performed as described in the following section.

Figure 2: View from the tracking camera.

In the second method, the the hands were tracked based on skin tone. We have found that all human hands have approximately the same hue and saturation, and vary primarily in their brightness. Using this information we can build an *a priori* model of skin color and use this model to track the hands much as was done in the gloved case. Since the hands have the same skin tone, "left" and "right" are simply assigned to whichever hand is leftmost and rightmost. Processing proceeds normally except for simple rules to handle hand and face ambiguity described in the next section.

Feature Extraction and Hand Ambiguity

Psychophysical studies of human sign readers have shown that detailed information about hand shape is not necessary for humans to interpret sign language (Poizner, Bellugi, & Lutes-Driscoll 1981; Sperling *et al.* 1985). Consequently, we began by considering only very simple hand shape features, and evolved a more complete feature set as testing progressed (Starner *et al.* 1994).

Since finger spelling is not allowed and there are few ambiguities in the test vocabulary based on individual finger motion, a relatively coarse tracking system may

be used. Based on previous work, it was assumed that a system could be designed to separate the hands from the rest of the scene. Traditional vision algorithms could then be applied to the binarized result. Aside from the position of the hands, some concept of the shape of the hand and the angle of the hand relative to horizontal seemed necessary. Thus, an eight element feature vector consisting of each hand's x and y position, angle of axis of least inertia, and eccentricity of bounding ellipse was chosen. The eccentricity of the bounding ellipse was found by determining the ratio of the square roots of the eigenvalues that correspond to the matrix

$$\begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$$

where a, b, and c are defined as

$$a = \int\int_{I'} (x')^2 dx'dy'$$

$$b = \int\int_{I'} x'y' dx'dy'$$

$$c = \int\int_{I'} (y')^2 dx'dy'$$

(x' and y' are the x and y coordinates normalized to the centroid)

The axis of least inertia is then determined by the major axis of the bounding ellipse, which corresponds to the primary eigenvector of the matrix (Horn 1986). Note that this leaves a 180 degree ambiguity in the angle of the ellipses. To address this problem, the angles were only allowed to range from -90 to +90 degrees.

When tracking skin tones, the above analysis helps to model situations of hand ambiguity implicitly. When a hand occludes either the other hand or the face, color tracking alone can not resolve the ambiguity. Since the face remains in the same area of the frame, its position can be determined and discounted. However, the hands move rapidly and occlude each other often. When occlusion occurs, the hands appear to the above system as a single blob of larger than normal mass with significantly different moments than either of the two hands in the previous frame. In this implementation, each of the two hands is assigned the moment and position information of the single blob whenever occlusion occurs. While not as informative as tracking each hand separately, this method still retains a surprising amount of discriminating information. The occlusion event is implicitly modeled, and the combined position and moment information are retained. This method, combined with the time context provided by hidden Markov models, is sufficient to distinguish between many different signs where hand occlusion occurs.

Training an HMM network

When using HMM's to recognize strings of data such as continuous speech, cursive handwriting, or ASL sentences, several methods can be used to bring context to bear in training and recognition. A simple context modeling method is embedded training. While initial training of the models might rely on manual segmentation or, in this case, evenly dividing the evidence among the models, embedded training trains the models *in situ* and allows model boundaries to shift through a probabilistic entry into the initial states of each model (Young 1993).

Generally, a sign can be affected by both the sign in front of it and the sign behind it. For phonemes in speech, this is called "co-articulation." While this can confuse systems trying to recognize isolated signs, the context information can be used to aid recognition. For example, if two signs are often seen together, recognizing the two signs as one group may be beneficial.

A final use of context is on the word or phrase level. Statistical grammars relating the probability of the co-occurrence of two or more words can be used to weight the recognition process. Grammars that associate two words are called bigrams, whereas grammars that associate three words are called trigrams. Rule-based grammars can also be used to aid recognition.

Experimentation

Since we could not exactly recreate the signing conditions between the first and second experiments, direct comparison of the gloved and no-glove experiments is impossible. However, a sense of the increase in error due to removal of the gloves can be obtained since the same vocabulary and sentences were used in both experiments.

Experiment 1: Gloved-hand tracking

The glove-based handtracking system described earlier worked well. Occasionally tracking would be lost (generating error values of 0) due to lighting effects, but recovery was fast enough (within a frame) that this was not a problem. A 5 frame/sec rate was maintained within a tolerance of a few milliseconds. However, frames were deleted where tracking of one or both hands was lost. Thus, a constant data rate was not guaranteed. This hand tracking process produced an eight-element feature vector (each hand's x and y position, angle of axis of least inertia, and eccentricity of bounding ellipse) that was used for subsequent modeling and recognition.

Of the 500 sentences collected, six were eliminated due to subject error or outlier signs. In general, each sign was 1 to 3 seconds long. No intentional pauses were placed between signs within a sentence, but the sentences themselves were distinct.

Initial estimates for the means and variances of the output probabilities were provided by iteratively using Viterbi alignment on the training data (after initially dividing the evidence equally amoung the words in the sentence) and then recomputing the means and variances by pooling the vectors in each segment. Entropic's Hidden Markov Model ToolKit (HTK) is used

Table 2: Word accuracy of glove-based system

experiment	*training set*	*independent test set*
grammar	99.5% (99.5%)	99.2% (99.2%)
no grammar	92.0% (97%) (D=9, S=67, I=121, N=2470)	91.3% (96%) (D=1, S=16, I=26, N=495)

as a basis for this step and all other HMM modeling and training tasks. The results from the initial alignment program are fed into a Baum-Welch re-estimator, whose estimates are, in turn, refined in embedded training which ignores any initial segmentation. For recognition, HTK's Viterbi recognizer is used both with and without a strong grammar based on the known form of the sentences. Contexts are not used, since a similar effect could be achieved with the strong grammar given this data set. Recognition occurs five times faster than real time.

Word recognition accuracy results are shown in Table 2; the percentage of words correctly recognized is shown in parentheses next to the accuracy rates. When testing on training, all 494 sentences were used for both the test and train sets. For the fair test, the sentences were divided into a set of 395 training sentences and a set of 99 independent test sentences. The 99 test sentences were not used for any portion of the training. Given the strong grammar (pronoun, verb, noun, adjective, pronoun), insertion and deletion errors were not possible since the number and class of words allowed is known. Thus, all errors are vocabulary substitutions when the grammar is used (and accuracy is equivalent to percent correct). However, without the grammar, the recognizer is allowed to match the observation vectors with any number of the 40 vocabulary words in any order. Thus, deletion (D), insertion (I), and substitution (S) errors are possible. The absolute number of errors of each type are listed in Table 2. The accuracy measure is calculated by subtracting the number of insertion errors from the number of correct labels and dividing by the total number of signs. Note that, since all errors are accounted against the accuracy rate, it is possible to get large negative accuracies (and corresponding error rates of over 100%). Most insertion errors correspond to signs with repetitive motion.

Analysis

The 0.8% error rate of the independent test set shows that the HMM topologies are sound and that the models generalize well. With such low error rates, little can be learned by analyzing the remaining errors.

However, the remaining 8.7% error rate (based on accuracy) of the "no grammar" experiment better indicates where problems may occur when extending the system. Without the grammar, signs with repetitive or long gestures were often inserted twice for each actual occurrence. In fact, insertions caused more errors than substitutions. Thus, the sign "shoes" might be recognized as "shoes shoes," which is a viable hypothesis without a language model. However, a practical solution to this problem is the use of context training and a statistical grammar instead of the rule-based grammar.

Using context modeling as described before may significantly improve recognition accuracy in a more general implementation as shown by the speech and handwriting recognition communities (Starner *et al.* 1994). While a rule-based grammar explicitly constrains the word order, statistical context modeling would have a similar effect while generalizing to allow different sentence structures. In the speech community, such modeling occurs at the "triphone" level, where groups of three phonemes are recognized as one unit. The equivalent in ASL would be to recognize "trisines" (groups of three signs) corresponding to three words, or three letters in the case of finger spelling. In speech recognition, statistics are gathered on word co-occurence to create "bigram" and "trigram" grammars which can be used to weight the liklihood of a word. In ASL, this might be applied on the phrase level. For example, the random sentence construction used in the experiments allowed "they like pill yellow they," which would probably not occur in natural, everyday conversation. As such, context modeling would tend to suppress this sentence in recognition, perhaps preferring "they like food yellow they," except when the evidence is particularly strong for the previous hypothesis.

Further examination of the errors made without the grammar shows the importance of finger position information. Signs like "pack," "car," and "gray" have very similar motions. In fact, the main difference between "pack" and "car" is that the fingers are pointed down for the former and clenched in the latter. Since this information is not available in the model, confusion occurs. While recovering specific finger positions is difficult with the current testing apparatus, simple palm orientation might be sufficient to resolve these ambiguities.

Since the raw screen coordinates of the hands were used, the system was trained to expect certain gestures in certain locations. When this varied due to subject seating position or arm placement, the system could become confused. A possible solution is to use position deltas in the feature vector, as was done in the second experiment.

Experiment 2: Natural skin tracking

The natural hand color tracking method maintained a 10 frame per second rate at 320x240 pixel resolution on a 200MHz SGI Indy. Higher resolution tracking suffered from video interlace effects. While tracking was somewhat noisier due to confusions with similarly-colored background elements, lighting seemed to play

a lesser role due to the lower specularity of skin compared to the gloves. Since only one hand "blob" might be expected at a given time, no frames were rejected due to lack of tracking of a hand. Due to the subject's increased familiarity with ASL from the first experiment, the 500 sentences were obtained in a much shorter span of time and in fewer sessions.

During review of this data set and comparison with the earlier set of sentences, it was found that subject error and variability increased. In particular, there was increased variability in imaged hand size (due to changes in depth under perspective) and increased variability in body rotation relative to the camera. Ignoring these unintentional complications, 478 of the sentences were correctly signed; 384 were used for training, and 94 were reserved for testing.

Table 3: Word accuracy of natural skin system

experiment	*training set*	*independent test set*
original	87.9% (87.9%)	84.7% (84.7%)
+ area	92.1% (92.1%)	89.2% (89.2%)
Δ + area	89.6% (89.6%)	87.2% (87.2%)
full	94.1% (94.1%)	91.9% (91.9%)
full-no grammar	81.0% (87%) (D=31, S=287, I=137, N=2390)	74.5% (83%) (D=3, S=76, I=41, N=470)

Word accuracies; percent correct in parentheses. The first test uses the original feature set from the first experiment. The second adds the area of the imaged hand. The change in position of the hands replaces the absolute position in the third test, and the final test uses the full set of features: x, y, Δx, Δy, angle, eccentricity, area, and length of the major eigenvector. All tests use the grammar except for the last result which shows "no grammar" for completeness.

In the first experiment an eight-element feature vector (each hand's x and y position, angle of axis of least inertia, and eccentricity of bounding ellipse) was found to be sufficient. However this feature vector does not include all the information derived from the hand tracking process. In particular, the hand area, the length of the major axis of the first eigenvector, and the change in x and y positions of the hand were not used. In this second experiment these features were added to help resolve the ambiguity when the hands cross. In addition, the use of combinations of these feature elements were also explored to gain an understanding of the their information content. Training and recognition occurred as described previously. The word accuracy results are summarized in Table 3; the percentage of words correctly recognized is shown in parentheses next to the accuracy rates.

Analysis

A higher error rate was expected for the gloveless system, and indeed this was the case. The 8 element feature set (x, y, angle, and eccentricity for each hand) was not sufficient for the task due to loss of information when the hands crossed. However, the simple addition of the area of each hand improved the accuracy significantly.

The subject's variability in body rotation and position was known to be a problem with this data set. Thus, signs that are distinguished by the hands' positions in relation to the body were confused since only the absolute positions of the hands in screen coordinates were measured. To minimize this type of error, the absolute positions of the hands can be replaced by their relative motion between frames (Δx and Δy). While this replacement causes the error rate to increase slightly, it demonstrates the feasibility of allowing the subject to vary his location in the room while signing, removing another constraint from the system.

By combining the relative motion and absolute position information with the angle, eccentricity, area, and length of the major eigenvector, the highest fair test accuracy, 91.9%, was reached. Without the context information provided by the grammar, accuracy dropped considerably; reviewing the errors showed considerable insertions and substitutions at points where the hands crossed.

Discussion and Conclusion

We have shown an unencumbered, vision-based method of recognizing American Sign Language (ASL). Through use of hidden Markov models, low error rates were achieved on both the training set and an independent test set without invoking complex models of the hands.

With a larger training set and context modeling, lower error rates are expected and generalization to a freer, user independent ASL recognition system should be attainable. To progress toward this goal, the following improvements seem most important:

- Measure hand position relative to each respective shoulder or a fixed point on the body.
- Add finger and palm tracking information. This may be as simple as counting how many fingers are visible along the contour of the hand and whether the palm is facing up or down.
- Use a two camera vision system to help disambiguate the hands in 2D and/or track the hands in 3D.
- Collect appropriate domain or task-oriented data and perform context modeling both on the trisine level as well as the grammar/phrase level.
- Integrate explicit face tracking and facial gestures into the feature set.

These improvements do not address the user independence issue. Just as in speech, making a system

which can understand different subjects with their own variations of the language involves collecting data from many subjects. Until such a system is tried, it is hard to estimate the number of subjects and the amount of data that would comprise a suitable training database. Independent recognition often places new requirements on the feature set as well. While the modifications mentioned above may be initially sufficient, the development process is highly empirical.

So far, finger spelling has been ignored. However, incorporating finger spelling into the recognition system is a very interesting problem. Of course, changing the feature vector to address finger information is vital to the problem, but adjusting the context modeling is also of importance. With finger spelling, a closer parallel can be made to speech recognition. Trisine context occurs at the sub-word level while grammar modeling occurs at the word level. However, this is at odds with context across word signs. Can trisine context be used across finger spelling and signing? Is it beneficial to switch to a separate mode for finger spelling recognition? Can natural language techniques be applied, and if so, can they also be used to address the spatial positioning issues in ASL? The answers to these questions may be key to creating an unconstrained sign language recognition system.

Acknowledgements

The authors would like to thank Tavenner Hall for her help editing and proofing this document.

References

Baum, L. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities* 3:1–8.

Charayaphan, C., and Marble, A. 1992. Image processing system for interpreting motion in American Sign Language. *Journal of Biomedical Engineering* 14:419–425.

Cui, Y., and Weng, J. 1995. Learning-based hand sign recognition. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition.*

Darrell, T., and Pentland, A. 1993. Space-time gestures. *Proc. Comp. Vis. and Pattern Rec.* 335–340.

Dorner, B. 1993. Hand shape identification and tracking for sign language interpretation. In *IJCAI Workshop on Looking at People.*

Essa, I.; Darrell, T.; and Pentland, A. 1994. Tracking facial motion. In *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects.*

Horn, B. K. P. 1986. *Robot Vision.* Cambridge, MA: MIT Press.

Huang, X.; Ariki, Y.; and Jack, M. A. 1990. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press.

Humphries, T.; Padden, C.; and O'Rourke, T. 1990. *A Basic Course in American Sign Language.* Silver Spring, MD: T. J. Publ., Inc.

Juang, B. 1985. Maximum likelihood estimation for mixture multivariate observations of Markov chains. *AT&T Tech. J.* 64:1235–1249.

Murakami, K., and Taguchi, H. 1991. Gesture recognition using recurrent neural networks. In *CHI '91 Conference Proceedings*, 237–241.

Poizner, H.; Bellugi, U.; and Lutes-Driscoll, V. 1981. Perception of American Sign Language in dynamic point-light displays. *J. Exp. Pyschol.: Human Perform.* 7:430–440.

Rabiner, L. R., and Juang, B. H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 4–16.

Rehg, J. M., and Kanade, T. 1993. DigitEyes: vision-based human hand tracking. School of Computer Science Technical Report CMU-CS-93-220, Carnegie Mellon University.

Schlenzig, J.; Hunter, E.; and Jain, R. 1994. Recursive identification of gesture inputs using hidden Markov models. *Proc. Second Annual Conference on Applications of Computer Vision* 187–194.

Sperling, G.; Landy, M.; Cohen, Y.; and Pavel, M. 1985. Intelligible encoding of ASL image sequences at extremely low information rates. *Comp. Vis., Graph., and Img. Proc.* 31:335–391.

Starner, T.; Makhoul, J.; Schwartz, R.; and Chou, G. 1994. On-line cursive handwriting recognition using speech recognition methods. In *ICASSP*, 125–128.

Starner, T. 1995. Visual recognition of American Sign Language using hidden Markov models. Master's thesis, MIT, Media Laboratory.

Takahashi, T., and Kishino, F. 1991. Hand gesture coding based on experiments using a hand gesture interface device. *SIGCHI Bulletin* 23(2):67–73.

Tamura, S., and Kawasaki, S. 1988. Recognition of sign language motion images. *Pattern Recognition* 21:343–353.

Wilson, A. D., and Bobick, A. F. 1995. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*

Yamato, J.; Ohya, J.; and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. *Proc. Comp. Vis. and Pattern Rec.* 379–385.

Young, S. 1993. *HTK: Hidden Markov Model Toolkit V1.5.* Washington DC: Cambridge Univ. Eng. Dept. Speech Group and Entropic Research Lab. Inc.