

ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis

Christian Vogler and Dimitris Metaxas

Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

cvogler@gradient.cis.upenn.edu, dnm@central.cis.upenn.edu

Abstract

We present a framework for recognizing isolated and continuous American Sign Language (ASL) sentences from three-dimensional data. The data are obtained by using physics-based three-dimensional tracking methods and then presented as input to Hidden Markov Models (HMMs) for recognition. To improve recognition performance, we model context-dependent HMMs and present a novel method of coupling three-dimensional computer vision methods and HMMs by temporally segmenting the data stream with vision methods. We then use the geometric properties of the segments to constrain the HMM framework for recognition. We show in experiments with a 53 sign vocabulary that three-dimensional features outperform two-dimensional features in recognition performance. Furthermore, we demonstrate that context-dependent modeling and the coupling of vision methods and HMMs improve the accuracy of continuous ASL recognition.

1 Introduction

American Sign Language (ASL) is the primary mode of communication for many deaf people in the USA. It is a highly inflected language with sophisticated grammatical properties, which constrain strongly the order and appearance of signs. Because of the constraints, it provides an appealing test bed for understanding more general principles governing human motion and gesturing, including human-computer gesture interfaces. Such interfaces are essential in virtual reality applications, where the user must be able to manipulate virtual objects by gesturing. A working ASL recognition system could also facilitate interaction of deaf people with their surroundings.

To date, most attempts at ASL recognition have either used only two-dimensional computer vision methods, or they have used other input devices, such as datagloves, instead of computer vision, to collect input from the signer [9, 1, 14]. In this paper we present a new approach to ASL recognition. First, we use computer vision methods to extract the three-dimensional parameters of a signer's arm motions. We then use Hidden Markov Models (HMMs) to recognize isolated and continuous ASL utterances from the three-dimensional input. We develop context-dependent modeling of HMMs and methods for coupling the application of HMMs and the application of three-dimensional computer vision methods to improve continuous recognition performance. Our approach at-

tempts to overcome some of the limitations of the previous approaches that use two-dimensional visual input, do not use context-dependent modeling, or do not couple computer vision methods with HMMs [9, 1, 8, 6].

Three-dimensional image-based shape and motion tracking of a human's arm and hand is difficult because of the complexity of the motions and occlusion effects. Recently, a methodology has been developed [4, 5] that allows three-dimensional tracking of human motion from multiple images. In this paper we augment this methodology to track the three-dimensional motion of a subject's arms and hands from multiple images. This method is based on the use of deformable models, whose shape and motion fits the given image sequences based on occluding contour information and theorems from projective geometry. The output of this method consists of the three-dimensional motion parameters of the subject's arms. For efficiency reasons, and because arm movements already carry much of the information needed for recognizing ASL signs, we do not use the hand information in this paper.

Apart from obtaining accurate data, ASL recognition is difficult, because there are always statistical variations in the way humans perform motions, even with identical meaning. In addition, in continuous utterances, there are no clear boundaries between individual signs. HMMs provide a framework for capturing statistical variations in both position and duration of the movement, as well as implicit segmentation of the input stream. Furthermore, continuous recognition is complicated by coarticulation effects, that is, the pronunciation¹ of a sign is influenced by the preceding and following signs. Coarticulation effects can be partly alleviated by training context-dependent HMMs.

The theory behind HMMs makes several assumptions that are often not valid in practice. For this reason, we develop a new approach that couples computer vision methods with HMM modeling. It is based on a temporal segmentation process that operates by extracting geometric properties of the three-dimensional computer vision parameters. These properties are obtained independently from the HMM algorithms and are used to impose additional constraints on HMM-based recognition.

To test our algorithms and assumptions, we performed a series of experiments based on a vocabulary consisting of 53 different signs that make extensive use of space. We ex-

¹By "pronunciation" we mean motion. We follow the terminology of spoken language linguistics where applicable.

perimented with both isolated and continuous ASL recognition for both three-dimensional and two-dimensional data. As HMMs require large amounts of training data and the computer vision process is computationally expensive, we used data from an Ascension Technologies Flock of Birds and computer vision processes interchangeably.

Our goal is to discover and analyze a usable framework for both isolated and particularly continuous ASL recognition. We do not address more general gesture recognition topics and signer independence in this paper. Neither do we address the involved aspects of ASL linguistics [10] at this point, but obviously, a viable future ASL recognition system should be able to handle them.

In the following sections, we discuss related work and give an overview on the theory behind the vision methods and HMMs. Afterward, we address the use of HMMs for continuous ASL recognition, and coupling computer vision processes with the HMM algorithms. Finally, we outline data collection and provide experimentation results for isolated and continuous recognition and the coupling of computer vision and HMMs.

2 Previous Work

Previous work on sign language recognition focuses primarily on fingerspelling recognition and isolated sign recognition. Some work uses neural networks [1, 13]. For this work to apply to continuous ASL recognition, the problem of explicit temporal segmentation must be solved, which is a limitation that HMM-based recognition does not have. Mohammed Waleed Kadous [14] uses Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy, with an emphasis on computationally inexpensive methods. There is very little previous work on continuous ASL recognition. Thad Starner and Alex Pentland [9] use a view-based approach to extract two-dimensional features as input to HMMs with a 40 word vocabulary. Yanghee Nam and Kwang Yoen Wahn [6] use three-dimensional data as input to HMMs for continuous recognition of a very small set of gestures.

3 Model-based Tracking of a Human’s Arms

In this section we give a brief overview of our formulation that allows the three-dimensional arm shape and motion estimation from multiple images [2, 3, 4, 5].

Our approach consists of two parts. The first part [2, 3] consists of an active, integrated approach that identifies reliably the parts of a moving articulated object and estimates their shape and motion from a *controlled set* of motions that reveal the object’s structure. We use the algorithm developed in [2, 3], which segments the apparent body contour of a moving human into the constituent parts. Initially, a single deformable model is used in order to fit the image data. As the model deforms to fit the deformed (due to the motion of the human) subsequent image contours, a novel *Human Body Part Identification Algorithm* (HBPIA) is developed to identify all the body parts. By applying the

HBPIA iteratively over the subsequent frames, all the moving parts are identified. In addition, we have extended this algorithm to allow the estimation of the three-dimensional shape of a subject’s body parts, based on the integration of images taken from three orthogonally placed cameras. We used this methodology to estimate the three-dimensional shape of the subject’s arms shown in the examples in Section 7. It is worth noting that we have recovered the lower arm and the hand as one part, since in our ASL recognition experiments we did not use the motion of the lower arm and the hand relative to each other.

The second part of the algorithm consists of using the extracted three-dimensional shape of the arm to track the three-dimensional position and orientation of a subject’s body parts [4]. To alleviate difficulties arising from occlusion and degenerate views during the unconstrained movement of the arm, we use three calibrated cameras placed in a mutually orthogonal configuration. At every image frame and for each body part, we derive a subset of the cameras that provide the most informative views for tracking. This *active* and time varying selection is based on the visibility of a part and the observability of its predicted motion from a certain camera. Once a set of cameras has been selected to track each part, we use concepts from projective geometry to relate points on the occluding contour to points on the three-dimensional shape model. Using a physics-based modeling approach, we transform this correspondence, in addition to two-dimensional forces arising from the discrepancy between the model’s occluding contour and the image data, into generalized forces that are applied to the model to estimate the model’s translational and rotational degrees of freedom. To improve the tracking results further, the dynamic system is embedded within an extended Kalman filter framework, and we use the *predicted* motion of the model at each frame to establish point correspondences between occluding contours and the three-dimensional model.

We used this two step approach to track the motion of the subject’s arms performing the ASL gestures, as shown in Section 7. The output of the system is a set of rotation, \mathbf{q}_θ , and translation, \mathbf{q}_c , parameters that we use as input to the HMMs and the vision-based segmentation algorithm presented in the following sections.

4 Hidden Markov Models

Hidden Markov Models (HMMs) are a type of statistical model. They have been used successfully in speech recognition, and recently in handwriting, gesture, and sign language recognition. We now give a summary of the basic theory behind HMMs, which is covered in detail in [7].

4.1 Definition of HMMs

An HMM consists of a number N of states S_1, S_2, \dots, S_N , together with transitions between states. The system is in one of the HMM’s states at any given time. At regularly spaced discrete time intervals, the system takes an outgoing transition from its current state to a new state.

Each transition from S_i to S_j has an associated probability a_{ij} of being taken. Hence, $\sum_i a_{ij} = 1$. Each state S_i also has an initial probability π_i of the system starting in S_i . In addition, each state S_i generates output $k \in \Omega$, which is distributed according to a probability distribution function $b_i(k) = P\{\text{Output is } k | \text{System is in } S_i\}$.

4.2 The Three Fundamental HMM Problems

There are three fundamental problems in HMM theory:

- (1) For a sequence of observations $O = O_1, \dots, O_T$, $O_i \in \Omega$, compute the probability $P(O|\lambda)$ that an HMM λ generated O .
- (2) For some O and an HMM λ , recover the most likely state sequence S_1, \dots, S_T that generated O .
- (3) Adjust the parameters of an HMM λ such that they maximize $P(O|\lambda)$ for some O .

The first problem corresponds to recognizing an unknown data sequence with a set of HMMs, each of which corresponds to a sign. For each HMM, the probability $P(O|\lambda)$ is computed that it generated the unknown sequence, and then the HMM with the highest probability is selected as the recognized sign. For computing $P(O|\lambda)$, let $Q = Q_1, Q_2, \dots, Q_T$ be a state sequence in λ :

$$\alpha_t(i) = P(O_1, O_2, \dots, O_t, Q_t = S_i | \lambda) \quad 1 \leq i \leq N, \quad (1)$$

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i), \quad (2)$$

$$\alpha_1(i) = \pi_i b_i(O_1), \quad (3)$$

$$\alpha_{t+1}(i) = b_i(O_{t+1}) \sum_{j=1}^N \alpha_t(j) a_{ji} \quad 1 \leq t \leq T-1 \quad (4)$$

These equations assume that the O_i are independent, and they make the Markov assumption that a transition depends only on the current state, a fundamental limitation of HMMs. This method is called the forward-backward algorithm and computes $P(O|\lambda)$ in $O(N^2T)$ time.

The second problem corresponds to finding the most likely path through an HMM λ , given an observation sequence O , and is equivalent to maximizing $P(Q, O|\lambda)$. Let

$$\delta_t(i) = \max_{Q_1, \dots, Q_{t-1}} P(Q_1 Q_2 \dots Q_t = S_i, O|\lambda), \quad (5)$$

$$\delta_{t+1}(i) = b_i(O_{t+1}) \cdot \max_{1 \leq j \leq N} \{\delta_t(j) a_{ji}\}, \quad (6)$$

$$P(Q, O|\lambda) = \max_{1 \leq i \leq N} \{\delta_T(i)\}. \quad (7)$$

$\delta_t(i)$ corresponds to the maximum probability of all state sequences that end up in S_i at time t . Equations 6 and 7 follow from Equation 5 by induction on t . The Viterbi algorithm is a dynamic programming algorithm that, using Equation 7, computes both the maximum probability $P(Q, O|\lambda)$ and the state sequence Q in $O(N^2T)$ time.

The recovery of the state sequence makes the Viterbi algorithm invaluable for continuous recognition, since it bypasses the difficult problem of segmenting the utterances into its individual parts. Instead, a sequence of HMMs corresponding to individual signs is concatenated into a network, as schematically depicted in Figure 1. Thus, the most likely state sequence recovers the sequence of signs.

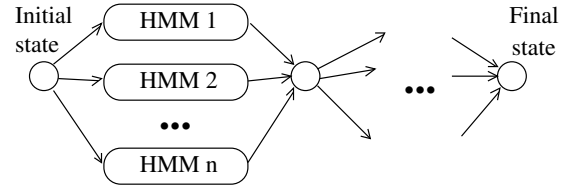


Figure 1: Concatenation of HMMs into a network

The third problem corresponds to training the HMMs with data, such that they are able to recognize previously unseen data correctly after the training phase. There exists no analytical solution for maximizing $P(O|\lambda)$ for given observation sequences, but an iterative procedure, called the Baum-Welch procedure, maximizes $P(O|\lambda)$ locally. In the case of continuous density output probabilities, the reestimation process works as follows.

Define $b_j(O)$ as $b_j(O) = \sum_{m=1}^M c_{jm} G(O, \mu_{jm}, U_{jm})$, where M describes the number of mixtures, j is the state number, c describes the weight of mixture m in state j , and G is a Gaussian density with mean μ , and covariance matrix U . Define the backward variable β as

$$\beta_t(i) = P(O_{t+1} O_{t+2}, \dots, O_T | Q_t = S_i, \lambda), \quad (8)$$

$$\beta_T(i) = 1, \quad (9)$$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad (10)$$

$$1 \leq i \leq N, \quad 1 \leq t \leq T-1. \quad (11)$$

Furthermore, define ξ and γ as

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}, \quad (12)$$

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j). \quad (13)$$

$\sum_t \xi_t(i, j)$ can be interpreted as the expected number of transitions from S_i to S_j ; likewise $\sum_t \gamma_t(i)$ can be interpreted as the expected number of transitions taken from S_i . With these interpretations, the reestimation formulae for the transitions and output probabilities are

$$\bar{\pi}_i = \gamma_1(i), \quad (14)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad (15)$$

$$\bar{c}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m)}{\sum_{t=1}^T \sum_{k=1}^M \gamma_t(j, k)}, \quad (16)$$

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) O_t}{\sum_{t=1}^T \gamma_t(j, m)}, \quad (17)$$

$$\bar{U}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) (O_t - \mu_{jm})(O_t - \mu_{jm})^T}{\sum_{t=1}^T \gamma_t(j, m)}. \quad (18)$$

Repeated use of this procedure converges to a maximum probability [7], typically after 5–10 iterations.

5 Use of HMMs for ASL Recognition

In the previous section we reviewed the extraction of three-dimensional features from computer vision and the HMM theory. We now discuss how they fit in the framework of ASL recognition.

HMMs are an attractive choice for processing three-dimensional sign data, because their state-based nature enables them to describe how a sign changes over time and to capture variations in the duration of signs.

Isolated sign recognition, with silence before and after each sign, is a comparatively straightforward process, once reliable features have been extracted. The presence of silence makes it easy to spot the boundaries between signs. Each sign can be extracted and presented to the trained HMMs individually. For a complete discussion of the aspects of isolated recognition in our framework, see [12].

Continuous sign recognition, on the other hand, is much harder than isolated sign recognition. There is no silence between the signs, so the straightforward method of using silence to distinguish boundaries fails. Here HMMs offer the compelling advantage of being able to segment the streams of signs automatically with the Viterbi algorithm.

Coarticulation effects further complicate continuous recognition. Speech recognizers handle them by training phoneme context-dependent HMMs. The same idea applies to sign language recognition, and we performed some experiments to verify the applicability, see Section 8.1. Coarticulation in sign language, however, is more complicated to handle compared to speech recognition, because of the insertion of a wide range of different movements between signs that are not present in the signs' lexical forms.²

A sign in our data collected at natural signing speeds was between 10 and 45 frames long, not counting the frames needed for the transition between signs. The HMM topology should be flexible enough to accommodate the variations in the length of different signs and duration. These considerations led us to using the left-right model shown in Figure 2. We determined the optimal number of

²This phenomenon is called movement epenthesis. An alternative to context-dependent modeling, which models epenthesis, is discussed in [11].

states for our recognition problem experimentally. For the output probabilities, we chose a single Gaussian density with diagonal covariance, as we had insufficient training data for estimating full-rank covariance matrices.

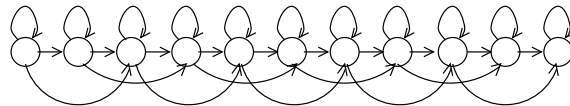


Figure 2: Topology of the context-dependent model. The arcs that skip states allow the modeling of variabilities in the duration of different signs.

The entire training session followed this scheme: First, we estimated the global means and variances of all training examples — as labeling sign boundaries by hand is notoriously difficult — and assigned them as the initial HMM output probabilities. Then we made several embedded [15] training passes for context-independent models. From these we cloned context-dependent models. We also tied the transition matrices of signs occurring in similar contexts, in order to decrease the total number of distinct HMMs. This was done before we performed several more embedded training runs on the context-dependent models, until the parameters converged.

6 Coupling of Vision and HMMs

In the preceding section we reviewed how HMMs can be used for ASL recognition. The use of HMMs alone, however, imposes some limitations, one of which is insufficiency of training data, especially while training context-dependent models. Furthermore, the probability theory assumptions underlying the HMM theory, as described in Section 4.2, are often not valid.

Another problem is that the HMM theory does not provide for any dynamic weighting of features depending on a sign's context. For example, the invariant features for some signs, such as "I," are the endpoints of their movements with respect to a body part, and the movements are unimportant. For other signs, only the movements are invariant. The parts of the feature set that should be examined and ignored for each class of signs are mutually exclusive.

To alleviate these limitations, we investigated the coupling of the HMM recognition process with an independent computer vision-based motion analysis that temporally segments the signal and extracts its geometric properties. The presence of three-dimensional information is crucial for the coupling to work. In the past, geometric fitting of planes has already been used for rough segmentation [6], but not for providing additional information about the nature of the fits to the HMM recognition process.

6.1 Segmentation of the Signal

To extract the geometric properties of the continuous signal estimated with our computer vision methods, it must first be segmented temporally into its parts. Any change of

the type of arm movement is likely to be accompanied by a dip in the velocity. Thus, minima in the absolute values of the velocity vector provide strong hints at segmentation boundaries. After performing initial segmentation based on velocities, our algorithm attempts to fit geometric primitives to the individual segments. These currently consist of lines, planes, and holds at a position in space.

The fit of a hold is determined by computing the covariance matrix over the segment’s position data. If there is little movement, the eigenvalues of the matrix in every direction are small, and consequently its trace is small.

The fit of a line is governed by

$$\sum_i e_i = \sum_i |\mathbf{p}_i - (\mathbf{d} \cdot \mathbf{p}_i) \mathbf{d}|^2, \quad (19)$$

where e_i is the distance of \mathbf{p}_i to the line, and \mathbf{d} is the line’s unit direction vector. Let \mathbf{P} be a matrix containing the points \mathbf{p}_i in the segments as its row vectors. Minimizing Equation 19 with respect to \mathbf{d} corresponds to maximizing $\mathbf{d}^T \mathbf{P}^T \mathbf{P} \mathbf{d}$. By Raleigh’s principle, the maximal-eigenvalue eigenvector of $\mathbf{P}^T \mathbf{P}$ maximizes this equation, which is equivalent to the maximal-eigenvalue eigenvector of the points’ covariance matrix. This eigenvector is the line’s direction vector. The other two eigenvalues indicate the goodness of fit.

The fit of a plane is found by minimizing $\mathbf{d}^T \mathbf{P}^T \mathbf{P} \mathbf{d}$ with respect to \mathbf{d} . By the same argument as above, this minimization is equivalent to finding the minimal-eigenvalue eigenvector of the points’ covariance matrix. This eigenvector is the plane’s normal vector. The largest two eigenvalues indicate the goodness of fit.

After the initial fit, the algorithm pools the primitives into a directed acyclic graph, schematically depicted in Figure 3. If the algorithm fails to fit any geometric primitives to some segment, it inserts the segment into the DAG as a “wild card,” which is defined conservatively to match any kind of geometric primitive. It then attempts to merge adjacent segments if they fit the same kind of primitive, or if a primitive could be part of another after the merge.

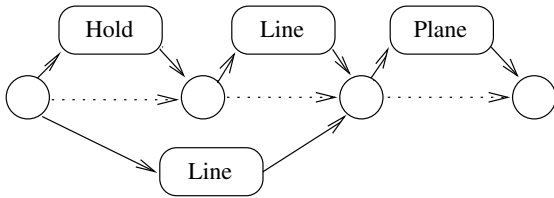


Figure 3: Geometric primitives pooled into a DAG. Circles denote segmentation boundaries. Dotted arcs denote null transitions; they are necessary to overcome spurious fits.

6.2 Using the Motion Analysis with HMMs

Each sign in the vocabulary has associated one or more templates that comprise the sign’s geometric primitives

with weights of each feature’s relative importance. These primitives are matched against those in the DAG. Assuming that the segmentation process yields correct results, any series of geometric matches that is correct for the signal must form a path through the DAG. Thus, by successively matching the templates for the signs to the DAG, it is possible to determine whether some particular sequence of signs could have occurred in the signal.

Determining whether a sequence of signs matches a signal can serve as a backup check for the output of the Viterbi algorithm. A possible approach is rejecting hypotheses from the Viterbi algorithm, if they do not match the geometric properties of the signal.

The HMM recognition algorithm and the vision matching algorithm complement each other. The advantages of the former method are automatic segmentation during both training and recognition, and a fully formalized training procedure. The disadvantages are poor performance in the presence of insufficient training data, no formal way to weight features dynamically, and violations of the stochastic independence assumptions. The advantages of the latter method are the possibility of weighting the relative importance of features dynamically, and independence from insufficient training data. The disadvantages are that estimating the geometric properties for the sign templates requires manual labeling and analysis of the data, that segmentation must be done explicitly, and that coarticulation sometimes changes the geometric properties of the signal.

7 Data Collection

For our experiments we collected data, using both our computer vision system, and an Ascension Technologies Flock of Birds. The reason for using the latter was that it is faster at this point than the computer vision system, and hence more suitable for prototyping.

The computer vision system yields rotation, \mathbf{q}_θ , and translation, \mathbf{q}_c , of each segment of the arm, as described in Section 3. Figure 4 gives an example of the computer vision tracking process. The images show the high accuracy of the computer vision system; in fact, it is comparable to the accuracy achieved by the Flock of Birds system.

The Flock of Birds system consists of a magnet and six sensors that detect their rotation, $\hat{\mathbf{q}}_\theta$, and translation, $\hat{\mathbf{q}}_c$, with respect to the magnet at 25 frames per second. We used the data from both systems interchangeably with a simple alignment of coordinate systems. The coordinate system was right-handed, with the origin at the base of the signer’s spine.

We used the 53-sign vocabulary listed in Table 1. Their pronunciations followed the ASL dialect used in the Philadelphia, PA, area. The goals in choosing the vocabulary were to be able to express sentences that could have occurred in a natural conversation, and to make intensive use of the signing space, so as to demonstrate the advantages of three-dimensional data over two-dimensional data. We collected 486 continuous ASL sentences, each between

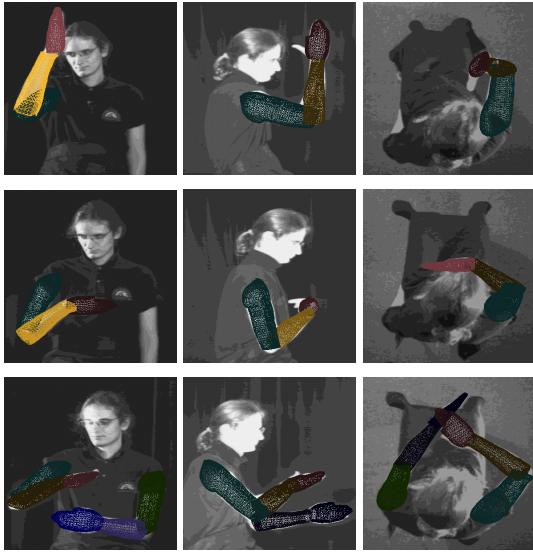


Figure 4: Fitting the three-dimensional models to the signer’s arms. From top to bottom, the signs for “FATHER,” “I,” and “MAIL” are displayed. From left to right, the front, side, and top views are displayed.

2 and 12 signs long, with a total of 2345 signs. The only constraints on the order and occurrence of signs were those dictated by the grammar of ASL [10].

Category	Signs used
Nouns	America, Christian, Christmas, book, brother, chair, college, family, father, friend, interpreter, language, mail, mother, name, paper, president, school, sign, sister, teacher
Pronouns	I, my, you, your, how, what, where, why
Verbs	act, can, give, have, interpret, like, make, read, sit, teach, try, visit, want, will, win
Adjectives	deaf, good, happy, relieved, sad
Other	if, from, for, hi

Table 1: The complete 53 sign vocabulary

8 Experiments

We performed isolated, continuous, and vision-HMM coupled ASL recognition experiments. The goal of the isolated recognition experiments was to discover a set of features that maximizes HMM recognition performance. The two main results of these were that Cartesian and polar position coordinates are better features than velocities, and that it is necessary to perform a large number of distinct experiments, in the range of several hundred, to determine the true merits of one feature set versus another. The best features regularly recognized more than 98 percent of the signs correctly. The detailed results and a discussion are

provided in [12].

We now describe our experiments with continuous HMM-based recognition and HMM-vision coupling. We used Entropic’s Hidden Markov Model Toolkit (HTK) Version 2.02 for training and testing in all of our experiments.

8.1 Continuous recognition experiments

We split the 486 sentences randomly into a training set with 389 examples and a test set with 97 examples (containing 456 signs). Each sign in the vocabulary occurred at least once in the test set. The training and test sets were the same throughout all experiments, and no portion of the test set was used for training in any way. We ran three-dimensional experiments with and without context-dependent HMMs, and two-dimensional experiments (by projecting the data on planes; the results given are the best that we found). We chose a mix of Cartesian and polar position coordinates, velocities, and wrist orientation angles as the feature vector, and a word loop as the task grammar.

Table 2 shows the experimental results. We use word accuracy as our evaluation criterion. It is computed by subtracting the number of insertion errors from the number of correctly spotted signs. The number of words in the result for two-dimensional data is lower than in the other results, because for one sentence the Viterbi beam-searching optimization pruned all paths through the HMM network.

Type of experiment	Word accuracy	Details
3D context independent	87.71%	H=416, D=8, S=32 I=16, N=456
3D context dependent	89.91%	H=424, D=6, S=26 I=14, N=456
2D context dependent	83.63%	H=394, D=14, S=44 I=16, N=452

Table 2: Results of experiments. H denotes the number of correct 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 signs in the test set.

8.2 Analysis of the recognition experiments

The results are clearly in favor of using three-dimensional data over two-dimensional for continuous recognition. The 6.3 percent difference is large, although, according to our experiences with isolated recognition, one experiment is not enough to estimate the real difference reliably.

Context-dependent models outperformed context-independent models, but the increase in performance was small, probably to a large extent because of insufficient training data — context-dependent modeling requires huge amounts of data to become effective. Also, cross-sign context-dependent modeling for ASL is implausible from a phonological point of view. We discuss an alternative that appears to perform better in [11].

More than half of the substitution errors in each experiment were confusions between “I” and “MY,” and “YOU”

and “YOUR,” which differ only in hand configuration. We expect that adding features describing the hand configuration will improve recognition performance significantly.

Repeating the context-dependent experiment with five-best recognition showed that the absence of a strong grammar for constraining the HMM network degrades recognition performance significantly. Unfortunately, using a strong grammar for a test set as diverse as ours is not practical, because the size of an HMM network grows exponentially with the number of rules present in the grammar.

8.3 Coupling experiments

To investigate the effects of coupling the three-dimensional motion analysis with the HMM framework, we first analyzed the signals of all sentences in the test set from the previous continuous experiments with our motion analysis, so as to estimate the number of sentences that would have been incorrectly rejected if the HMM framework had yielded *perfect* results. There were 10 such sentences out of 97. Five of these 10 incorrect rejects were not recognized correctly by the context-dependent HMMs in the first place. The motion analysis accepted those incorrect recognition hypotheses from the HMMs as correct.

Running the coupling algorithm on the *actual* recognition hypotheses from the context-dependent HMMs also eliminated 10 sentences out of 97. Five of these were correctly rejected, so at the current moment, coupling HMMs with motion analysis breaks even with using the HMM framework by itself. In particular, three of the correct rejects took advantage of the motion analysis’s capability to weight movement and position features dynamically.

As we have used only a small part of the full power of computer vision motion analysis so far, we see these results as evidence that coupling will eventually be able to outperform either method independently. Future extension of motion analysis will consist of recognizing more different geometric properties, fine-tuning the sign templates, and fine-tuning the dynamic weighting of features based on the properties of each sign that is matched to the signal.

9 Conclusion

We have developed a framework for recognizing American Sign Language from three-dimensional data obtained with computer vision techniques. We showed how to collect three-dimensional data from computer vision and use them as input to Hidden Markov Models. We also determined that three-dimensional features are superior over two-dimensional ones. By using context-dependent modeling, we improved recognition performance. Through coupling vision processes with Hidden Markov Models, we took a first step toward overcoming the limitations of either method by itself.

Besides elaborating on context-dependent modeling, modeling of ASL linguistic features, and coupling, future research will need to address the topic of how to counterbalance the impracticability of using strong grammars to constrain large HMM networks. Bigram probabilities offer

a solution to this problem, but for it to succeed, a corpus of labeled real-world ASL conversations must be established first. Presently, no such corpus exists.

Acknowledgments

This work was supported in part by a NSF Career Award NSF-9624604, ONR-DURIP’97 N00014-97-1-0385 and N00014-97-1-0396, ONR Young Investigator Proposal, and NSF IRI-97-01803. Ioannis Kakadiaris helped in obtaining the computer vision samples.

References

- [1] 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, pp. 185–194, Wilmington, DE, 1996.
- [2] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach. CVPR’94, pp. 980–984.
- [3] I. A. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. ICCV’95, pp. 618–623
- [4] I. A. Kakadiaris and D. Metaxas. Model based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. CVPR’96, pp. 81–87.
- [5] D. Metaxas. *Physics-based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging*. Kluwer Academic Publishers, November 1996.
- [6] Y. Nam and K. Y. Wohn. Recognition of space-time hand-gestures using Hidden Markov model. ACM Symposium on Virtual Reality Software and Technology, pp. 51–58, Hong Kong, 1996.
- [7] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
- [8] J. M. Siskind and Q. Morris. A maximum-likelihood approach to visual event classification. ECCV’96.
- [9] T. Starner and A. Pentland. Visual recognition of American Sign Language using Hidden Markov models. International Workshop on Automatic Face and Gesture Recognition, pp. 189–194, Zürich, Switzerland, 1995
- [10] C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington DC, 1995.
- [11] C. Vogler and D. Metaxas. Adapting Hidden Markov models for ASL recognition by using three-dimensional computer vision methods. SMC’97.
- [12] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. CIS Technical Report, Department of Computer and Information Science, University of Pennsylvania, 1997.
- [13] 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.
- [14] M. Waleed Kadous. Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. Proceedings of the Workshop on the Integration of Gesture in Language and Speech, pp. 165–174, Wilmington, DE, 1996.
- [15] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book (for HTK 2.0)*. Cambridge University, 1995.