Multi-Layer Gesture Recognition: An Experimental Evaluation*

Jacob Eisenstein, Shahram Ghandeharizadeh, Leana Golubchik,
Cyrus Shahabi, Donghui Yan, Roger Zimmermann
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA

Abstract

Gesture recognition techniques often suffer from being highly device-dependent and hard to extend. If a system is trained using data from a specific glove input device, that system is typically unusable with any other input device. The set of gestures that a system is trained to recognize is typically not extensible, without retraining the entire system. We propose a novel gesture recognition framework to address these problems. This framework is based on a multi-layered view of gesture recognition. Only the lowest layer is device dependent; it converts raw sensor values produced by the glove to a glove-independent semantic description of the hand. The higher layers of our framework can be reused across gloves, and are easily extensible to include new gestures. We have experimentally evaluated our framework and found that it yields at least as good performance as conventional techniques, while substantiating our claims of device independence and extensibility.

1 Introduction

Gesture recognition offers a new medium for human-computer interaction that can be both efficient and highly intuitive. However, gesture recognition software is still in its infancy. While many researchers have documented methods for recognizing complex gestures from instrumented gloves at high levels of accuracy [6, 9, 12, 14], these systems suffer from two notable limitations: device dependence and lack of extensibility.

Conventional approaches to gesture recognition typically involve training a machine learning system to classify gestures based on sensor data. A variety of machine learning techniques have been applied, including hidden Markov models [6, 9, 15], feedforward neural networks [14], and recurrent neural networks [6, 11]. These different approaches have a common feature: they all treat gesture recognition as a one-step, *single-layer* process, moving directly from the sensor values to the detected gesture. Consequently, the properties of the specific input device used in training are built into the system. For example, a system that was trained using a 22 sensor CyberGlove would almost certainly be of no use with a 10 sensor

*This research is supported in part by NSF grants EEC-9529152 (IMSC ERC), and IIS-0091843 (SGER).

DataGlove. The system expects 22 inputs, and would be unable to produce meaningful results with the 10 inputs offered by the DataGlove.

Ideally, a gesture recognition system should be able to work with a variety of input devices. As the number of sensors is reduced, the performance might degrade, but it should degrade gracefully. We call this property *device independence*, and it is the first significant advantage of our approach.

In order to achieve device independence, we have reconceptualized gesture recognition in terms of a *multi-layer* framework. This involves generating a high-level, device-independent description of the sensed object – in this case, the human hand. Gesture recognition then proceeds from this description, independent of the characteristics of any given input device.

Because our approach allows the gesture recognition system to use a clean, semantic description of the hand, rather than noisy sensor values, much simpler techniques can be employed. It is not necessary to use anything as complicated as a neural network; rather, simple template matching is sufficient. Template matching provides a key advantage over more complex approaches: it is easily extensible, simply by adding to the list of templates. To recognize a new gesture with a conventional system, the entire set of gestures must be relearned. But we will show experimentally that with our approach, it is possible to add new gestures without relearning old ones. These new gestures are recognized with nearly the same accuracy as those in the original training set. Thus, *extensibility* is the second main advantage of our approach.

In Section 2, we describe our multi-layer framework in more detail. Section 3 presents our implementation for the task of ASL fingerspelling recognition, which we believe will generalize to other virtual reality applications. The results of an experimental evaluation of this implementation are given in Section 4. Section 5 surveys related work, and Section 6 presents brief conclusions and future research directions.

2 A Multi-Layer Framework

Our proposed framework is based on a multi-level representation of sensor data. It consists of the following four levels:

1. Raw data: This is the lowest layer and contains continuous streams of data emanating from a set of sensors. We have addressed the analysis of this general class of data in [4]. In this paper, we are specifically concerned with data generated by the sensors on a glove input device. A detailed description of the sensors included with the CyberGlove haptic device are provided in Table 1. The data at this level is highly device-dependent; each device may have a unique number of sensors, and the sensors may range over a unique set of values. However, raw sensor data is *application* independent; no assumptions are made about how the data will be used, or even what it describes. Conventional approaches to dealing with streaming sensor data typically operate at exactly this level. Consequently, these approaches are usually very difficult to adapt to new devices, and they fail to take advantage of human knowledge about the problem domain.

2. Postural Predicates: This level contains a set of predicates that describe the posture of the hand.

Sensor number	Sensor description	Sensor number	Sensor description
1	thumb roll sensor	12	ring inner joint
2	thumb inner joint	13	ring middle joint
3	thumb outer joint	14	ring outer joint
4	thumb-index abduction	15	ring-middle abduction
5	index inner joint	16	pinky inner joint
6	index middle joint	17	pinky middle joint
7	index outer joint	18	pinky outer joint
8	middle inner joint	19	pinky-ring abduction
9	middle middle joint	20	palm arch
10	middle outer joint	21	wrist flexion
11	middle-index abduction	22	wrist abduction

Table 1: CyberGrasp Sensors

Table 2 provides a list of the predicates to represent the hand postures for ASL fingerspelling. A vector of these predicates consisting of 37 boolean values describes a general hand posture. For example, a *pointing* posture is described by noting that the index finger is open, Open (index_finger), while every other finger is closed, Closed (thumb), Closed (middle_finger), etc. Each predicate describes a single features of the overall posture – e.g., Closed (index_finger). In our pointing posture, five vector values (corresponding to the aformentioned predicates evaluate as *true*), while the remaining ones evaluate as *false*.

While we do not claim that Table 2 presents a comprehensive list that captures all possible postures, we do show that these predicates can describe the ASL alphabet. Preliminary research on other fingerspelling systems suggests that this set of predicates is general, and we plan to investigate its generality in support of non-fingerspelling applications. To this end, we plan to apply this set of predicates to a virtual reality application in future research.

The postural predicates are derived directly from the lower level of raw sensor data. This process is described in more detail later in Section 3.2. The derivation of postural predicates from sensor data is, by necessity, device-dependent. However, it is application-independent, if our postural predicates are indeed general over a range of applications.

Once postural predicates are extracted from the sensor data, device-independent applications can be implemented using this higher level of representation. Thus, our multi-layered approach provides for the sound software engineering practice of modular reuse. Some modules can be reused across multiple devices, while others can be reused across multiple applications.

3. Temporal Predicates: Our work thus far has mainly focused on postural predicates. Here, we assume the ASL alphabet as our working data set where most signs are $static$ and do not require any hand motion (and hence have no temporal aspect). The extension of the framework to temporal signs is part of our future work.

4. Gestural Templates: This layer contains a set of templates, each of which corresponds to a whole hand gesture. Postures contain no temporal information; a gesture may contain temporal information,

Name	Definition	Applicability	Number of predicates
Open (X)	Indicates that finger X is extended parallel to the palm.	Any finger, and the thumb	5
Closed (X)	Indicates that finger X is closed with the palm. Note that the *open* and *closed* predicates are mutually exclusive, but they are not complete. A finger may neither entirely open, nor closed.	Any finger, and the thumb	5
Touching-thumb (X)	Indicates that finger X is touching the thumb.	Any finger other than the thumb	4
Grasping (X)	Indicates that finger is grasping something with the thumb.	Any finger other than the thumb	4
Split (X, Y)	Indicates that adjacent fingers X and Y are spread apart from each other.	Any adjacent pair of fingers, and the index finger and the thumb.	4
Crossing (X, Y)	Indicates that finger X is crossing over finger Y, with Y closer to the palm than X.	Applies to any two fingers, but because of the limited flexibility of the human hand, it is assumed that the index finger cannot cross with the ring finger or the pinky, and that the middle finger cannot cross with the pinky.	14
Palm-facing-in ()	Indicates that the palm is facing the signer, rather than the recipient of the signs.	Applies to the whole hand.	1

Table 2: Postural Predicates

Alphabet	Set of predicates (corresponding to a template)
A	Closed (F1, F2, F3, F4)
B	Open (F1, F2, F3, F4), Closed (T)
C	Grasping F1, F2, F3, F4 (1in, 0 degrees)
D	Open (F1), Touching-thumb (F2, F3)
E	Closed (T), Crossing (F1, T), Crossing (F2, T) Crossing (F3, T)
F	Open (F2, F3, F4), Touching-thumb (F1), Split (F2, F3), Split (F3, F4)
G	Open (F1), Closed (F2, F3, F4), Crossing (T, F2), Palm-facing-in()
H	Open (F1, F2), Closed (F3, F4), Crossing (T, F3), Crossing (T, F4), Palm-facing-in()
I	Closed (F2, F3, F4), Open (F4), Crossing (T, F1), Crossing (T, F2), Crossing (T, F3)
K	Open (F1, F2), Closed (F3, F4)
L	Closed (F2, F3, F4), Open (F1, T), Split (F1, T)
M	Closed (T, F4) Crossing (F1, T), Crossing (F2, T), Crossing (F3, T), Crossing (T, F4)
N	Closed (T, F3, F4), Crossing (F1, T), Crossing (F2, T), Crossing (T, F3)
O	Touching-thumb (F1, F2, F3, F4)
P	Closed (T, F3, F4), Open (F1, F2)
Q	Open (T, F1) Closed (F2, F3, F4)
R	Closed (F3, F4), Open (F1, F2), Crossing (T, F3), Crossing (T, F4), Crossing (F2, F1)
S	Closed (F1, F2, F3, F4), Crossing (T, F1), Crossing (T, F2)
T	Closed (F2, F3, F4), Crossing (T, F2), Crossing (F1, T)
U	Closed (F3, F4), Open (F1, F2), Crossing (T, F3), Crossing (T, F4)
V	Closed (F3, F4), Open (F1, F2), Crossing (T, F3), Crossing (T, F4), Split (F1, F2)
W	Closed (F4), Open (F1, F2, F3), Crossing (T, F4), Split (F1, F2), Split (F3, F4)
X	Closed (F2, F3, F4), Crossing (T, F2) Crossing (T, F3) Crossing (T, F4)
Y	Closed (F1, F2, F3), Open (T, F4)

Table 3: ASL Fingerspelling Templates

although this is not required. A gesture is a description of the changing posture and position of the hand over time. An example of a gesture is a hand with a pointing index finger moving in a circular trajectory with a repetitive cycle.

Gestures are described as *templates* because they are represented as a vector of postural and temporal predicates. In order to classify an observed hand motion as a given gesture, the postural and temporal predicates should match the gestural template. This might be an approximate match; in our ASL application, we simply choose the gestural template that is the closest match to the observed data (see Section 3.1.1 for details).

3 Implementation

Figure 1 shows the two key modules of our implementation: 1) a set of predicate recognizers, and 2) a template matcher. The predicate recognizers convert raw sensor data to a vector of predicates. The template matcher then identifies the nearest gestural template. The template matcher is assisted by two other components: a *confidence* vector, and a probabilistic *context*. We will first describe how these components work together to detect ASL signs. Next, Section 3.2 describes how the system is trained.

3.1 ASL Sign Detection

This section explains how the system moves from sensor data to the vector of predicates. We then describe the basic template matching technique, and show how it can be augmented with context and the confidence vector.

3.1.1 Predicate Recognizers

The predicate recognizers use traditional gesture recognition methods to evaluate each postural predicate from a subset of the sensor data. In this case, we implemented the predicate recognizers as feedforward neural networks; we have explored other approaches in the past [3]. Each predicate recognizer need not consider data from the entire set of sensors. Rather, the sensors are mapped as input to the predicate recognizers manually, using common knowledge of which sensors are likely to be relevant to each predicate. For example, the predicate Crossing (T, F1) receives input only from the sensors on the thumb and index finger. By mapping only those sensors that are relevant to each predicate recognizer, human knowledge can be brought to bear to dramatically improve both the efficiency and accuracy of training. Table 2 shows the six predicate types and the thirty-seven predicates required to describe a single handshape.

To perform recognition of these thirty-seven postural predicates, we employ thirty-seven individual neural networks (see Figure 1). Each neural net consumes between 4 and 10 sensor values, includes ten hidden nodes, and produces either a zero or a one as output, denoting the logical valence of its predicate. The outputted predicates are then collated together into a vector, which is fed as input to a template matcher.

3.1.2 Template Matching

The gesture recognizers for a specific application, such as ASL, are realized using these postural predicates. Since these gesture recognizers manipulate high-level *semantic* data rather than low-level sensor values, it becomes possible to employ simpler and more extensible approaches. Our system performs gesture recognition by simple template matching on the detected vector of postural predicates. Template matching can be extended by simply adding to the set of pre-existing templates. In addition, this template matching component can be used across many different gloves (see Section 4).

The template matcher works by computing the Euclidean distance between the observed predicate vector and every known template. The template that is found to be the shortest distance from the observed predicate vector is selected. Mathematically, for a perceived predicate vector v, we want to find the gesture template i that minimizes $d_{i,v}$, which is the Euclidean distance between the two vectors.

$$d_{i,v} = \sum_{0 \leq p < P} |\mathbf{i}[p] - \mathbf{v}[p]| \tag{1}$$

P is equal to the total number of predicates; in our application $P = 37$. Sections 3.1.2 and 3.1.3 will augment this equation with context and confidence to improve performance.

Table 3 shows the true predicates in the template for each static sign in the ASL fingerspelling alphabet. They were determined by consulting an ASL textbook. Since the meaning of each predicates is entirely straightforward, constructing the templates is not expected to be a significant bottleneck to the usage or extension of our system. We believe that manually specifying new templates is far easier than finding additional training data and retraining a neural network.

3.1.3 Confidence

Ideally, each predicate recognizer performs perfectly; failing that, we would at least like to see all of the predicate recognizers perform equally well. In reality, this is not the case. For example, the Closed(T) predicate recognizer might correctly classify every test case, while the Crossing(T, F1) predicate recognizer might produce results no better than chance. To treat these two predicate recognizers equally would be a mistake; we should apply more weight to the data from the Closed(T) recognizer, and ignore data from the Crossing(T, F1) recognizer. To achieve this, we construct a vector $\mathbf{k}$, with *confidence* ratings between 0 and 1 for every predicate recognizer. The template matcher uses the confidence vector by factoring it into the distance metric:

$$d_{i,v} = \sum_{0 \leq p < P} \mathbf{k}[p] |\mathbf{i}[p] - \mathbf{v}[p]| \tag{2}$$

For example, suppose the predicate recognizers return the vector “0110”, and the two closest templates are “1110” and “0111”. Suppose that it is known that the first predicate recognizer (first bit) is only as good as chance, with a confidence rating of 0.1, but the 4th predicate recognizer (fourth bit) is rated with a confidence of 0.9. In this case, it is very likely that the fourth bit is indeed zero, while the value of the first bit is uncertain. Equation 2 uses this information, and selects “1110” as the appropriate template.

3.1.4 Context

Language users typically follow known patterns, rather than producing letters or words at random. Using context, the system constrains the space of possible utterances to improve performance. Context also reduces the effect of noise on the data, and can act as a tie-breaker between two templates that are equally close to the observed predicate vector.

We incorporate context into our application using an n-gram letter model [10]. N-gram based matching has had success in dealing with noisy data in other problem domains, e.g., text retrieval [2], natural language processing [17]. In the domain of English word spellings, an n-gram is an n-character contiguous sequence of a longer string. For example, the ASL spelling of word “fish” would be composed of the following n-grams:

bi-grams: _f, fi, is, sh, h_
tri-grams: _fi, fis, ish, sh_

quad-grams: _fis, fish, ish_

In our case, a blank, i.e., a "_", denotes the start of an ASL word spelling. In general, a word consisting of k ASL characters will have $k+1$ bi-grams, k tri-grams, and $k-1$ quad-grams. Each n-gram is assigned a probability, based on how often it occurs in the dictionary. For example, the probability of the bigram "th" is relatively high, whereas the probability of "tx" is relatively low.

We can use n-grams to help recognize gestures. Suppose we are trigrams, and we know that the current context is "_q." We can then examine the probabilities of all trigrams starting with "_q." Given this context, it is of course extremely probable that the next letter is "u." Even if noisy data obscures the recognition, the context component can help us to make the correct identification.

Context is represented by a vector $\mathbf{c}$, which includes the recent history of signs produced by the user. The size of $\mathbf{c}$ is equal to the size of the n-gram minus one; if trigrams are used, then $\mathbf{c}$ stores a two-element history. We then use the n-gram probabilities to find the *conditional probability* that each possible letter follows the letters in the context. The conditional probability of some letter i, given the context $\mathbf{c}$, is denoted by $P(i|\mathbf{c})$.

We now have to factor this probability into our original equation. Recall that originally we chose the gesture template i that minimizes $d_{i,v}$ for an observed predicate vector v. This is the same thing as choosing the gesture template i that maximizes $1/d_{i,v}$. We can include the conditional probability of each gesture in the equation by instead maximizing:

$$P(i|\mathbf{c}) \left(\frac{1}{d_{i,v}}\right)^n \tag{3}$$

By varying n, we can control the extent to which context influences the gesture recognition. As n approaches zero, the $(1/d_{i,v})^n$ term approaches 1, and the conditional probability becomes more important. As n goes to infinity, the $(1/d_{i,v})^n$ term becomes more significant, overwhelming the conditional probability and dominating the calculation. This tradeoff is quantified in detail in Section 4.

3.2 Training

The training phase impacts two components of the system: 1) each of the 37 predicate recognizers, and 2) the confidence vector. We describe each in turn.

3.2.1 Training the Gesture Recognizers

Figure 2 shows how the predicate recognizers are trained. During training, the system is presented with a set of example inputs, and the corresponding expected output templates. While the sensor values are passed along to the appropriate predicate recognizers, the expected output is passed to the list of gesture templates. The appropriate template is selected; for example, if the expected output is a "G," then the template "0110...1" is selected. The output of the ith predicate recognizer must match the ith element in

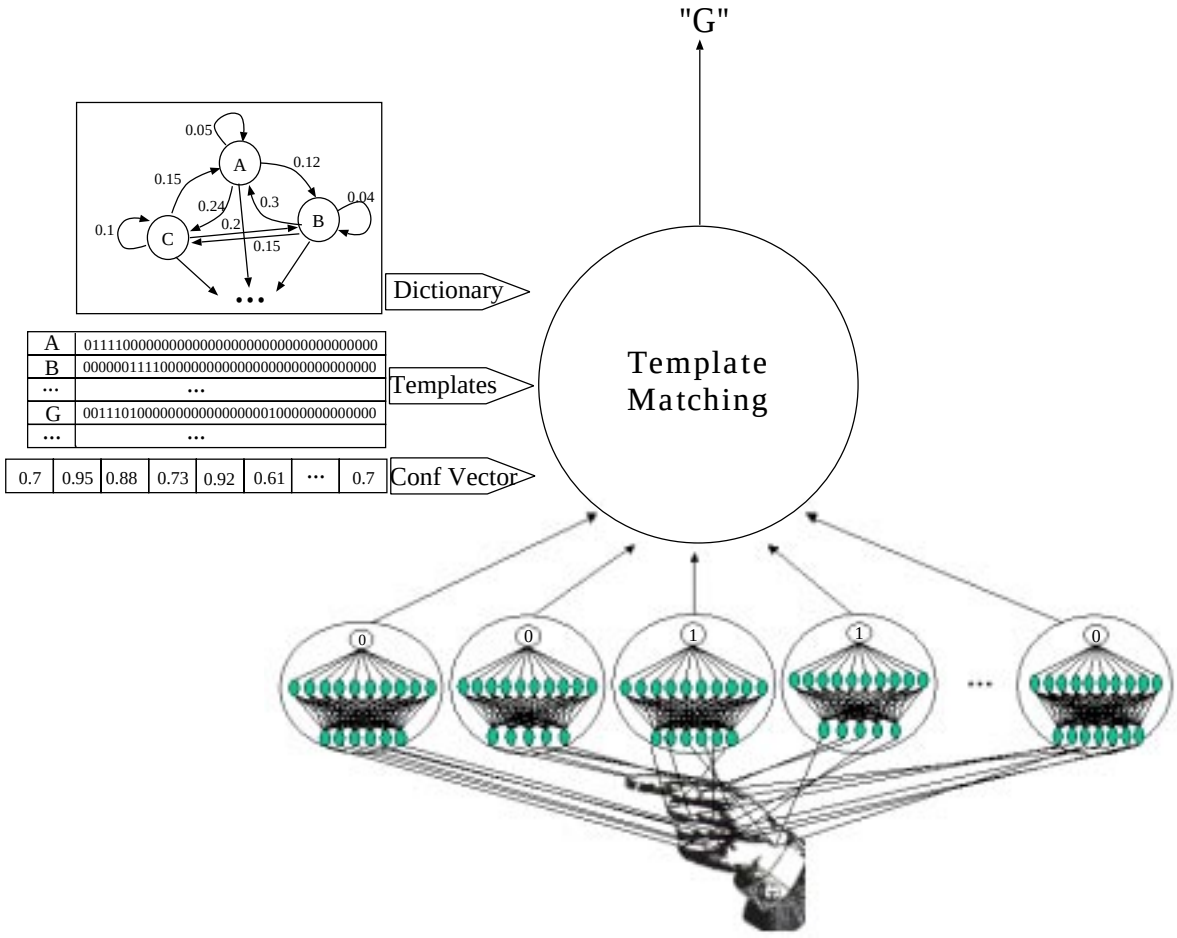

Figure 1: A Multi-Layer Framework to Detect Static ASL Signs

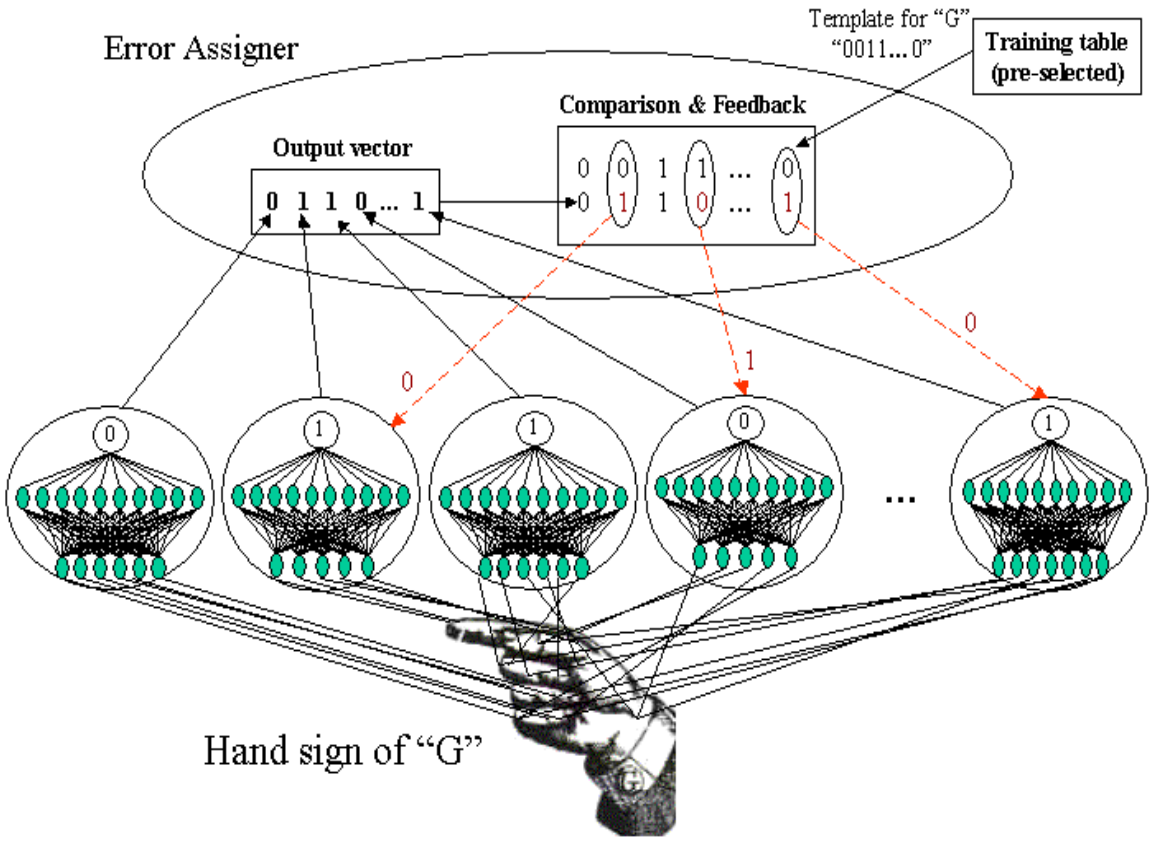

Figure 2: Training of Spatial Sign Detector

the template vector. In our example, the output of the 2^{nd} predicate must be a "1." If they do not match then the expected value is returned to the predicate recognizer and it is trained using a conventional error backpropagation method for feedforward neural networks [1, 13].

While the use of thirty-seven individual neural networks might appear computationally expensive, the small size of each network makes the cost of this approach similar to that of traditional techniques. The average predicate recognizer has six inputs, ten hidden nodes, and one output, for a total of $6*10+10*1 = 70$ weights that must be updated during training. With 37 networks, a total of $37 * 70 = 2590$ weights must be updated. A traditional approach, employing a single large neural network to recognize ASL signs directly from raw sensor data, would require 22 inputs (one for each sensor), and 24 outputs (one for each static sign). In our experiments, we found that 25 hidden nodes yielded the best performance for the traditional single-network approach. Thus, this approach still requires that $22 * 25 + 25 * 24 = 1150$ weights be updated. Even though we use thirty-seven networks instead of one, our approach is only a little more than two times as costly as the conventional technique. We believe that the benefits of device-independence and extensibility, which will be quantified later, more than justify this additional cost. Anywhere from 500 to 3000 iterations are required for each gesture recognizer. The entire process requires less than twenty minutes on a 1 GHz Pentium 4, running Linux, with 512 MB of memory. Our software is written entirely in Java, and performance would probably improve if ported to C.

3.2.2 Setting the Confidence Vector

As described in Section 3.1.2, the confidence vector maintains a set of ratings on performance of each predicate recognizer. These ratings range between 0 and 1, and they control the extent to which each predicate recognizer influences the final outcome. In order to determine the ratings in the confidence vector, 20% of the training data is withheld from training. This data is called the "tuning set", and it is used to compute the confidence vector.

As in training, we evaluate each predicate recognizer based on the sensor data and the expected output. However, instead of performing any additional training, we simply record its performance. It is critical that we use data not present in the training set, since this gives an indication of whether the predicate recognizers can generalize to examples outside of the training data. In this way, the confidence vector helps us account for and deal with *overfitting*, a serious problem in many machine learning applications.

Let $\mathbf{a}[i]$ represent the accuracy of the i^{th} predicate recognizer on the tuning set. All predicate recognizers evaluate to either 1 or 0; consequently, if $\mathbf{a}[i] = 0.5$, then the predicate recognizer is only performing at chance. In this case, our confidence rating should be zero, since the predicate recognizer is no better than flipping a coin. On the other hand, if $\mathbf{a}[i] = 1$, then the predicate recognizer is performing perfectly, and a confidence rating of one should be assigned. All of this can be quantified in the following equation:

$$\mathbf{k}[i] = 2(\mathbf{a}[i] - 0.5)^{+} \tag{4}$$

The superscript "+" indicates that the term in parentheses can never go below zero; if it does, it is set to exactly zero, i.e. $(\mathbf{a}[i] - 0.5)^+ = \max(\mathbf{a}[i] - 0.5, 0)$. A coefficient of two is applied in order to normalize confidence rating to a range between 0 and 1.

4 Experimental Results

We have evaluated our approach in the domain of ASL fingerspelling. Specifically, we focus on static signs and use the twenty-four letters which do not contain temporal characteristics (i.e., Z and J are omitted). Extending our system to handle spatio-temporal gestures is the subject of future work. Our evaluation proceeds along three dimensions: performance, device-independence, and extensibility. We will show that our approach achieves at least comparable performance to a conventional approach, while providing a level of device-independence and extensibility that are well beyond the capabilities of any known conventional system.

4.1 Performance

As a baseline for comparison, we use a conventional feedforward neural network, which consumes all twenty-two raw sensor values as input, includes twenty-five hidden nodes, and has twenty-four output nodes (see Figure 3. The letter corresponding to the maximally activated output node is considered to be the output. This baseline approach is very similar to the one used in [14].

Many experiments used the same individual signers in the test set as those who were used to train the system[1] ([6] is an exception). Our testing methodology is substantially more rigorous, because we have attempted to achieve *signer independence*. Specifically, out of sixteen signers in our dataset, twelve were used in training, and four were used in testing. Moreover, we performed only the most cursory calibration, taking less than thirty seconds for each signer. To achieve confidence in our results, we performed ten separate experiments, with different, randomly chosen test and training sets in each. The results reported below are the averages over the ten experiments. In a commercial application, developers would be free to choose a training set that yielded maximum performance, but we constructed our training sets randomly to ensure the integrity of our experiments.

We use only a bigram model for context; this improved accuracy by roughly 10%. We also experimented with a trigram model, but found that it yielded only a marginal improvement beyond the bigram, and required significantly more time to evaluate the system. A more thorough discussion of the role of context and the parameter n from Equation 3 can be found in Section 4.4.

Since the baseline approach could not take advantage of context, we compared our approach both with and without context against the baseline. We tested all systems by simulating each user signing every word in the English dictionary. The results are shown in the first line of Table 4. With the help of a bigram context,

[1] Our experience is that this makes a significant difference in the performance of the recognition system. Hence, it would not be fair to compare our results with those from systems that do not attempt to achieve signer independence.

our system strongly outperformed the baseline. Without the bigram context, our system was slightly better than the baseline. This validates our claim that our approach performs as well as the baseline.

4.2 Device Independence

To show that our framework supports device independence, we tested our system on six different glove input devices. The predicate recognizers and confidence vector were retrained for each device; the template matcher and context modules were reused. The first glove, *CyberGlove 22*, is a real glove used in our lab and the other five gloves are real gloves that are simulated by removing sensors from the data files produced by the first glove. The details of all six gloves are as follows.

1. *CyberGlove 22.* A 22 sensor glove, with a sensor for every joint in the human hand [8].
2. *CyberGlove 18.* Omits the distal (outermost) joints in each finger (the thumb has no distal joint). In Table 1, the omitted sensors are 7, 10, 14, and 18 [8].
3. *DataGlove 16.* Omits the distal joints (7, 10, 14, and 18), and the wrist flexion (21) and abduction (22) sensors [7].
4. *DataGlove 10.* Two sensors per finger and thumb (2, 3, 5, 6, 8, 9, 12, 13, 16, and 17). Omits distal joints, wrist flexion and abduction, palm arch (20), thumb roll (1), and abduction between fingers (4, 11, 15, 19) [19].
5. *TCAS Glove 8.* One sensor per finger and thumb, plus thumb roll and palm arch. Included sensors: 1, 2, 3, 5, 8, 12, 16, 20. [18].
6. *DataGlove 5.* One sensor per finger and thumb (2, 5, 8, 12, 16) [7].

Whereas the baseline neural network had to be entirely retrained for each glove, our system only retrained the low-level gesture recognizers and the confidence vector. Even under these conditions, our system – with context – outperforms the baseline on every glove except the DataGlove 10, where the difference is insignificant. Without context, we outperform the baseline on both CyberGloves, and on the TCAS 8.

The sharp discontinuity between our system's performance on the DataGlove 16 and the DataGlove 10 can be partly explained by the lack of abduction sensors on the latter glove. This eliminates the critical sensors for four of our predicate recognizers, rendering those recognizers useless. On the baseline system, there is a sharp reduction in performance between the DataGlove 10 and the TCAS 8, while our system experienced almost no change. The biggest difference between these two gloves is that the TCAS only includes one sensor per finger, whereas the DataGlove 10 has two. The joint sensors that were removed were not critical to any predicate recognizer; this might help to explain why our system weathered this change so smoothly.

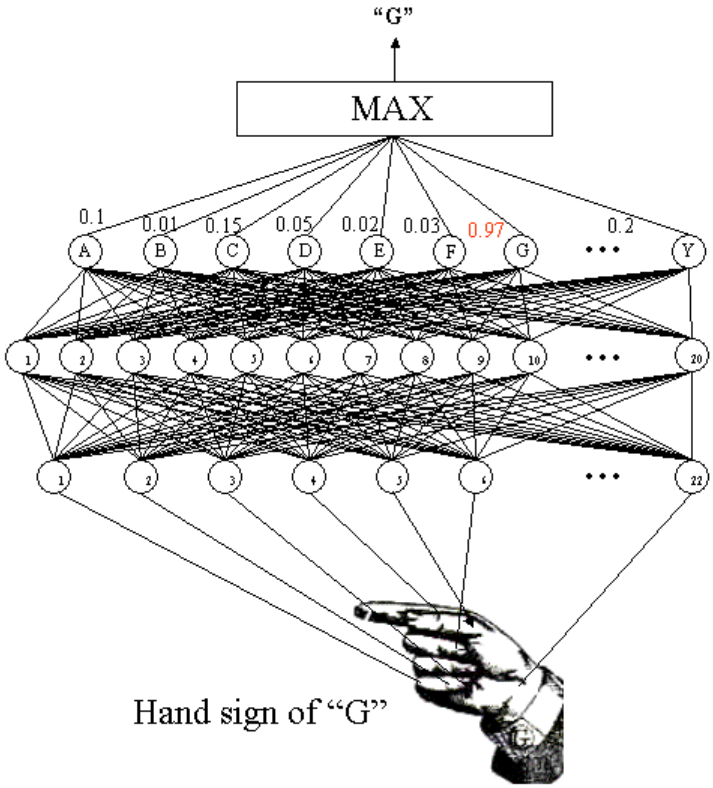

Figure 3: The Baseline Approach.

4.3 Extensibility

To demonstrate the extensibility of our system, we conducted the following experiment. We removed letters from the training data when the gesture recognizers were being trained, and then later added their templates to the library. The goal is to determine whether our system could be extended to recognize these new templates, even though they were not in the training set of the gesture recognizers.

With the complete training set, our approach achieved 73% accuracy with context, and 62% without, as shown on the first line of Table 4. We successively removed every letter from the training set, one at a time, and found that the average accuracy across all twenty-four letters was 66% with context, and 57% without. This shows that performance degrades only marginally on new templates that were not part of the training set, and suggests that it is indeed possible to extend the system with new signs and maintain adequate performance.

4.4 The Role of Context

The role of context can be explored by examining the change in performance as a function of the parameter n from Equation 3. Recall that as we reduce the value of n to zero, the sensor values of the glove are ignored and only context is considered. As we increase n toward infinity, context is ignored, and only the sensor

System	Baseline	Multilayered	Multilayered
Context	no	yes	no
CyberGlove 22	58.9%	72.9%	62.2%
CyberGlove 18	59.2%	70.0%	60.1%
DataGlove 16	57.8%	62.4%	47.3%
DataGlove 10	45.2%	42.7%	29.0%
TCAS 8	24.1%	41.4%	28.4%
DataGlove 5	20.1%	26.1%	10.7%

Table 4: Accuracy Results for the Six Gloves with $n = 2$.

values are considered. When the system is presented with a random sequence of letters, the use of context should actually detract from performance, since our context system operates on the assumption of normal ASL usage. Consequently, the performance of our system on a random sequence of letters should increase with n, as the context becomes less important. However, we expect that the performance on the simulated dictionary should increase with n to a certain point, which represents the optimal balance of context and sensor values, and then decrease.

This expected behavior was born out in our experiments, and is depicted in Figure 4, which has n increasing exponentially on the horizontal axis. Accuracy for the random sequence of letters data source starts at chance, which is approximately 4% ($1/24$). It then increases monotonically, leveling off at around 62.5%. Accuracy for the Dictionary data source starts at around 25%, which is the chance of guessing the next letter correctly given that you only know the previous letter, and nothing about what the user has actually signed. Accuracy increases with n, until it reaches a peak between 2.0 and 4.0, and then decreases very slowly, asymptotically approaching the performance with the Random Letters data source. As n approaches infinity, context becomes irrelevant; that is why the accuracy for the Dictionary and Random Letters data sources converge at high values of n.

Our graph is plotted on an exponential axis; it illustrates that the performance of our system is not unduly sensitive to the parameter n. In this case, any value between 2.0 and 8.0 produces similar performance.

5 Related Work

Glove-based gesture recognition has been explored in a number of studies. In an early study, Murakami and Taguchi [11] used recurrent neural networks to detect signs from Japanese Sign Language. Newby used a "sum of squares" template matching approach to recognizing ASL signs, which is very similar to our template matching component [12]. This system does seem to have met the criterion of extensibility, although this is not mentioned explicitly; the method was chosen because of its fast execution time. More recent studies in gesture recognition has focused on hidden Markov models, which have produced highly accurate systems capable of handling dynamic gestures [6, 9, 15].

The idea of device independence in virtual reality applications has been addressed in a number of studies.

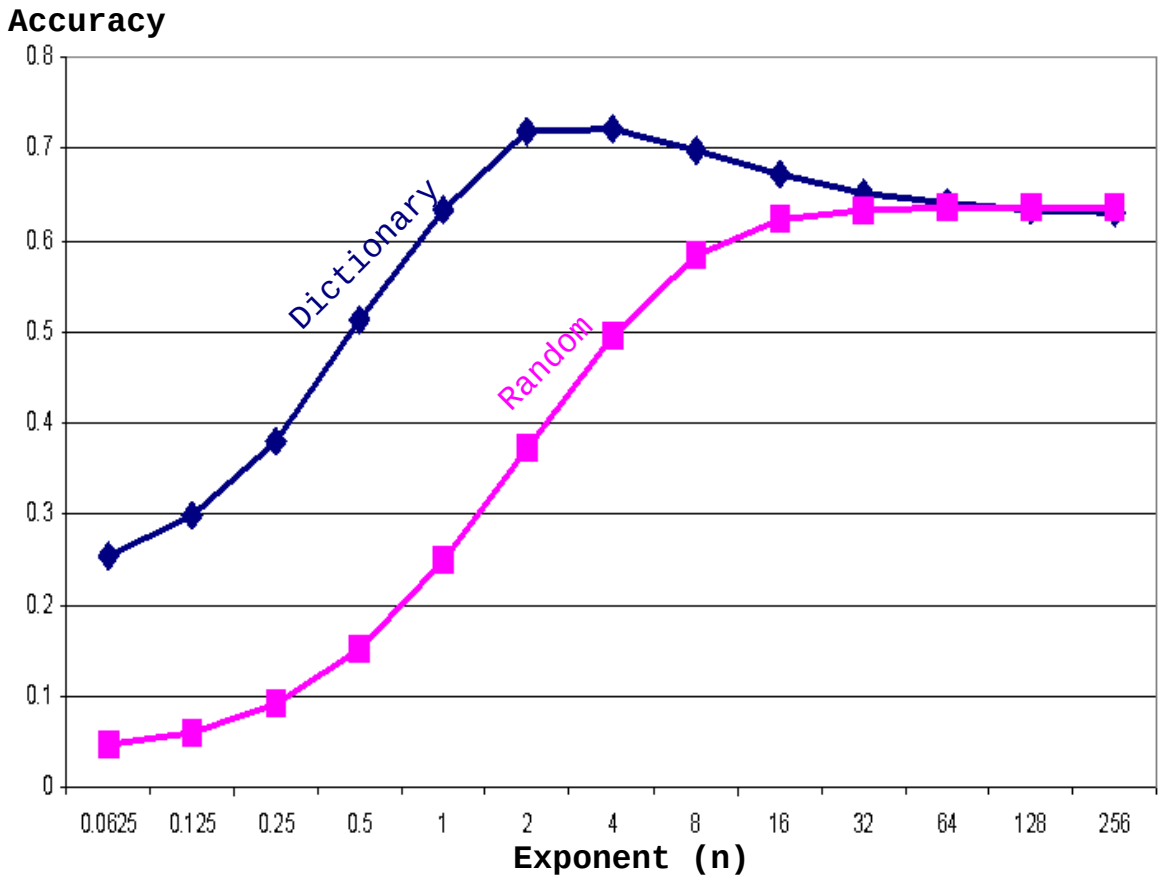

Figure 4: Performance as a Function of n.

One such study, by Faisstnauer et al., describes the Mapper [5], which eases the integration of new devices into virtual reality applications. This study does not tackle the specific issue of gesture recognition, but instead provides a high level software engineering framework for handling heterogeneous devices.

More closely related to our own work is that of Su and Furuta [16]. They propose a "logical hand device" that is in fact a semantic representation of hand posture, similar to our own set of postural predicates. This was done with the express purpose of achieving device independence, but to our knowledge it was never implemented. Our research can be viewed as an implementation and experimental validation of these ideas.

6 Conclusion and Future Research Directions

The experimental results described here are preliminary evidence that our system does indeed achieve our two goals of device independence and extensibility. We would like to make further experiments along these lines. Regarding the claim of device independence, we would like to test our system with other real gloves, rather than merely simulating them. As for the extensibility claim, we would like to show that our predicate recognizers, trained on ASL data, can actually support entirely different sign languages, such as Japanese or Russian Sign Language. Ultimately, we would like to go even further, and show that we can support

applications outside of sign language altogether, such as virtual reality.

We also plan to make a deeper exploration of the role of context. In particular, we are interested in creating buffers for input data, in order to support delayed decisions. The idea is to maintain multiple hypothetical outputs, and only choose a given hypothesis later on. For example, suppose the letter "t" was detected, followed by an uncertain letter, "?". Suppose that we delay evaluation, and then find that the next letter is "e," yielding a string of "t?e". In this case, the uncertain letter is likely to be "h," yielding the string "the." But suppose that instead we delay evaluation and find that we have the string "t?x". Now the uncertain letter is probably "a," yielding the string "tax." By delaying evaluation until both the past and future context is known, a much greater degree of accuracy may be obtained.

Most important of all is the problem of extending our system to include gestures that have a temporal characteristic. Since we have focused only on postural predicates, our predicate recognizers use feedforward neural networks, which we believe are well-suited to this task. We hope that the same basic framework can be applied to temporal predicates, perhaps using recurrent neural networks or hidden Markov models for the predicate recognizers. The questions of how to build these temporal predicate recognizers, and how to incorporate them into our existing framework are clearly crucial next steps towards creating a useful system.

Acknowledgements

We thank Yong Zeng for preparing the figures, and for his assistance with the experimental evaluation of the system.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.

[2] W. B. Cavnar. N-Gram based Text Filtering for TREC-2. In *Proceedings of the 2nd Text Retrieval Conference*, NIST, Gaithersburg, Maryland, 1993.

[3] J. Eisenstein, S. Ghandeharizadeh, L. Huang, C. Shahabi, G. Shanbhag, and R. Zimmermann. Analysis of Clustering Techniques to Detect Hand Signs. In *Proceedings of the 2001 International Symposium on Intelligent Mulimedia, Video and Speech Processing*, Hong Kong, May 2001.

[4] J. Eisenstein, S. Ghandeharizadeh, C. Shahabi, G. Shanbhag, and R. Zimmermann. Alternative Representations and Abstractions for Moving Sensors Databases. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM)*, Atlanta, GA, November 5-10, 2001.

[5] C. Faisstnauer, D. Schmalstieg, and Z. Szalavári. Device-Independent Navigation and Interaction in Virtual Environments. In *Proceedings of the VRST Adjunct Workshop on Computer Graphics*, Taipei, Taiwan, November 5-6, 1998.

[6] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma. Signer-independent Continuous Sign Language Recognition Based on SRN/HMM. In *Proceedings of the IEEE ICCV Workshop on Recognition,*

Analysis, and Tracking of Faces and Gestures in Real-Time, pages 90–95, Vancouver, BC, Canada, July 2001.

[7] Fifth Dimensions Technologies, Santa Clara, CA. *5DT Data Glove Series Data Sheet*, 2000.

[8] Immersion Corporation, San Jose, CA. *CyberGlove Reference Manual*, 1998.

[9] R.-H. Liang and M. Ouhyoung. A Sign Language Recognition System Using Hidden Markov Model and Context Sensitive Search. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST'96)*, pages 59–66, Hong Kong, June 1996.

[10] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 2000.

[11] K. Murakami and H. Taguchi. Gesture Recognition Using Recurrent Neural Networks. In *Proceedings of the Conference on Human Factors and Computing Systems (CHI'91)*, pages 237–242, New Orleans, Louisiana, 1991.

[12] G. B. Newby. Gesture Recognition Based upon Statistical Similarity. *Presence*, 3(3):236–243, 1994.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1:318–362, 1986.

[14] R. Salomon and J. Weissmann. Gesture Recognition for Virtual Reality Applications Using Data Glove and Neural Networks. In *Proceedings of IEEE International Joint Conference on Neural Networks*, Washington, DC, 1999.

[15] T. Starner and A. Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, Zürich, 1995.

[16] S. A. Su and R. Furuta. A Logical Hand Device in Virtual Environments. In *Proceedings of the ACM Conference on Virtual Reality Software and Technology (VRST'94)*, pages 33–42, Singapore, August 23-26, 1994.

[17] C. Y. Suen. N-Gram Statistics for Natural Language Understanding and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):164–172, April 1979.

[18] C. Youngblut, R. E. Johnson, S. H. Nash, R. A. Wienclaw, and C. A. Will. Review of Virtual Environment Interface Technology. Technical Report IDA Paper P-3186, Log: H96-001239, Institute for Defense Analysis, 1996.

[19] T. G. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill. A Hand Gesture Interface Device. In *Proceedings of ACM CHI+GI: Human Factors in Computing Systems and Graphics Interface*, pages 189–192, 1987.