A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation

José L. Hernández-Rebollar
Department of ECE
The George Washington Univ.
jreboll@seas.gwu.edu

Robert W. Lindeman
Department of CS
The George Washington Univ.
gogo@seas.gwu.edu

Nicholas Kyriakopoulos
Department of ECE
The George Washington Univ.
kyriak@seas.gwu.edu

Abstract

This paper presents a portable system and method for recognizing the 26 hand shapes of the American Sign Language alphabet, using a novel glove-like device. Two additional signs, 'space', and 'enter' are added to the alphabet to allow the user to form words or phrases and send them to a speech synthesizer. Since the hand shape for a letter varies from one signer to another, this is a 28-class pattern recognition system. A three-level hierarchical classifier divides the problem into "dispatchers" and "recognizers." After reducing pattern dimension from ten to three, the projection of class distributions onto horizontal planes makes it possible to apply simple linear discrimination in 2D, and Bayes' Rule in those cases where classes had features with overlapped distributions. Twenty-one out of 26 letters were recognized with 100% accuracy; the worst case, letter U, achieved 78%.

1. Introduction

Hand shape and gesture recognition has been an active area of investigation during the past decade. Beyond the quest for a more 'natural' interaction between humans and computers, there are many interesting application in robotics, virtual reality, tele-manipulation, tele-presence, and sign language translation. According to American Sign Language (ASL) linguist William Stokoe [13] and the ASL Dictionary [1], a sign is described in terms of four components: hand shape, location in relation to the body, movement of the hands, and orientation of the palms. Hand shape (position of the fingers with respect to the palm), the static component of the sign, along with the orientation of the palm, forms what is known as ***posture***. A set of 26 unique distinguishable postures makes up the alphabet in ASL used to spell names or uncommon words that are not well defined in the dictionary.

While some applications, like image manipulation and virtual reality, allow the researcher to select a convenient set of postures which are easy to differentiate, such as **point**, **rotate**, **track,** [22] **fist**, **index**, **victory**, [19] or the "NASA Postures" [2], the well-established ASL alphabet contains some signs which are very similar to each other. For example, the letters 'A', 'M', 'N', 'S', and 'T' are signed with a closed fist (see [1]). The amount of finger occlusion is high and, at first glance, these five letters can appear to be the same posture. This makes it very hard to use vision-based systems in the recognition task. Nevertheless Uras and Verri [15] tried to recognize the shapes using the "size function" concept on a Sun Sparc Station with some success. Lamar [7] achieved a 93% recognition rate in the easiest (most recognizable letter), and a 70% recognition rate in the most-difficult case (the letter 'C'), using colored gloves and neural networks. Starner, Weaver, and Pentland [12] implemented a successful gesture recognizer with as high as 98% accuracy, but they abandoned the idea of recognizing hand postures and captured only hand area, instead.

Despite instrumented gloves being described as 'cumbersome,' 'restrictive', and 'unnatural' for those who prefer vision-based systems, they have been more successful recognizing postures. In 1983, Grimes [4] was granted a patent of his Data Entry Glove, which he used to enter ASCII characters to a computer using switches and other sensors sewn to the glove. Kramer [6] used his patented CyberGlove, along with a look-up table, to recognize the 26 letters of the alphabet. Erenshteyn [3] also used the CyberGlove and a method involving coded output, such as Hamming, Golay, and other hybrid codes. The VPL Data Glove invented by Zimmerman [21] has been used to recognize postures in different sign languages. For example, Liang [8] solved a set of 51 basic postures of Taiwanese Sign Language using probability models, and Waldron [16] was able to recognize 36 ASL postures by solving a two-stage neural network.

In a search of more-affordable options (the CyberGlove costs between US$ 9,800 and US$ 14,500, depending on the number of sensors), Kadous [17] proposed a system for Australian Sign Language based on Mattel's Power Glove, but the glove could not be used to recognize the alphabet hand shapes because of a lack of sensors on the pinky finger. Pister [9] used accelerometers at the fingertips to implement a tracking system for pointing purposes. This glove has not been applied to fingerspelling. Thorough reviews of the use of gloves for

gesture recognition can be found in papers by Sturman [14] and Watson [18].

The system proposed in this paper addresses the issues of portability and affordability. We use an inexpensive micro controller and a set of five MEMS (Micro Electro Mechanical System) dual-axis accelerometers manufactured by Analog Devices, without compromising accuracy. Our device, called The Accele Glove [23], is different from previous approaches in that it does not require an external tracker system to identify hand orientation, and is therefore capable of identifying postures that others cannot. First, data are collected and analyzed 'off line' on a PC. The features we extract, after applying a simple transformation on raw data, allow easy visualization of classes as vectors in the posture space. After the algorithm is tested, a three level hierarchical classifier, implemented as a sequence of 'if-then-else' statements in the micro controller, executes the algorithm in real time.

2. The System

The key component of the system is the Accele Glove, which provides a measurement of finger position with respect to the gravitational vector. Figure 1 shows how all components are interconnected.

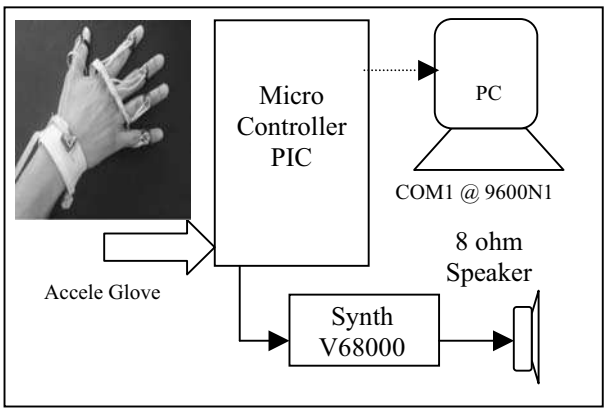


Figure 1. Block Diagram. The micro controller reads the Accele Glove and sends the letter's ASCII code to the speech synth. The PC was used to analyze data off-line.

By using the digital output of the MEMS dual-axis accelerometers attached to fingers, no A/D converter is necessary, and a single-chip micro controller can be used . The PC is used for data analysis and algorithm training, and disconnected after that. When programmed with the trained algorithm, the micro controller feeds the voice synthesizer (V68000) with the ASCII of the recognized letter, so the signer actually 'speaks out' words and short sentences.

2.1 Sensors Location.

The human hand has 17 active joints above the wrist: three on each of the index, middle, ring, and pinky fingers; three on the thumb; and the pitch and yaw on the wrist (rolling is generated in the elbow). The number of joints needed to distinguish between signs is a crucial factor; if the system, vision-based or instrumented, fails in acquiring enough information, ambiguity (the same set of signals for different postures) will reduce the recognition rate to unacceptable levels. Attaching five dual-axis sensors on the proximal inter phalangeal (PIP, or middle) joints of the fingers and the Inter phalangeal (distal) joint of the thumb, eliminates ambiguity for the 26 postures of the ASL alphabet. Location of the axes is shown in Figure2

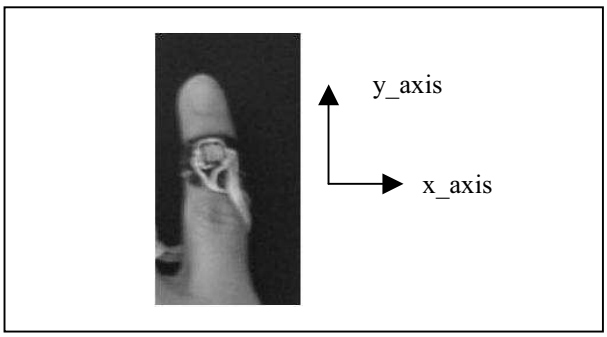


Figure 2: Detailed view of sensors placed on the Accele Glove.

2.2 Accelerometers.

The accelerometer is composed of a small mass suspended by springs. Capacitive sensors distributed along two orthogonal axes (X and Y) provide a measurement proportional to the displacement of the mass with respect to its rest position. Because the mass is displaced from the center, either due to acceleration or due to an inclination with respect to the gravitational vector **g**, the sensor can be used to measure absolute angular position.

The Y-axis points toward the fingertip, providing a measure of joint flexion. The X-axis can be used to extract information of hand roll or yaw, or individual finger abduction. The total number of signals being measured is 10. Although palm orientation and translation (relative to the wrist) can be viewed as affecting all fingers simultaneously, and would therefore allow us to eliminate three X measurements, it has been observed that this is not true in the general case. The main difference between the postures for 'U' and 'V' (see [1]) is precisely the individual orientation of the index and middle fingers in the X direction (the abduction, or finger spread).

2.3 Signals.

Position is read measuring the duty cycle of a train of pulses of 1kHz. When a sensor is in its horizontal position, the duty cycle is 50%. When it is tilted from +90° to –90° the duty cycle varies from 37.5% (.375 msec) to 62.5% (.625 msec), respectively. The micro controller monitors the output, and measures how long the output remains high (pulse width), using a 10 microsecond clock counter, meaning a range from (375/10) = 37 counts for 90° to a maximum of (625 /10) = 62 counts for -90°, a span of 25 counts. Nonlinearity and saturation, two characteristics of this mechanical device, reduce the usable range to ±80°. Therefore, the resolution is (160° / 25 counts) = 6.5° per count. The error of any measure was found to be ±1 bit, or ±6.5°, which is slightly larger than the error of 5° reported by Wise [20], and smaller than the mean of 11° reported by Quam [10] for the VPL DataGlove. The CyberGlove can be adjusted to measure different ranges of flexion using 8 bits, but the accuracy is reduced, due to problems of repeatability, to 15° according to Kessler *et al*. [5].

2.4 Data Collection.

Ten pulse widths are read sequentially by the micro controller, beginning with the X-axis followed by the Y-axis, thumb first. It takes 10 milliseconds to gather all finger positions. During the analysis stage of the project, position is sent as a package of 10 bytes through the serial port to a PC and saved as a row in a text file.

Five volunteers (three females, two males between 26 and 36 years old) were asked to wear the prototype shown in Figure 1 and to sign every letter of the alphabet ten times. Letters 'J' and 'Z' are sampled only at their final position. This allowed us to capture the differences and similarities among signers in twenty-six files with 10*x*10 matrices.

3. Feature Selection and Extraction.

The set of 10 measurements, two axes per finger, represents a vector of raw data. As we indicated in the introduction, different representations have been tested in the literature: Uras represented postures as contours, Lamar extracted the eigenvalues and eigenvectors from images. Kramer did not try feature extraction, but worked with the vector of raw data, as did Shahabi [11]. In his paper, he claims to be the first to use a decision tree for the analysis of haptic data (taken from a CyberGlove). He also claims to be the first to use Bayesian Classifier for raw haptic data analysis, and to use and compare Back Propagation Neural Networks with Bayesian Decision Tree for the recognition of static data. Contrary to his results, where he concludes that Decision trees are not suited to the task of sign recognition, we state and prove the contrary.

The goal is to extract a set of features that represents a posture without ambiguity in "posture space". If Stokoe [13] and Costello [1] already defined posture as being composed of hand shape and orientation, and that these two features are sufficient to represent any letter without ambiguity, then the feature extraction process has to recover these two components from raw data. Our Accele Glove is different from all other devices we have found, in that it is able to measure not only finger flexion (hand shape), but hand orientation (with respect to the gravitational vector) without the need for any other external sensor like a magnetic tracker. or Mattel's ultrasonic trackers [8], [16], [17].

We define a vector p in posture space,

p = [hand shape, palm orientation]

as the product of raw data vector D by a transformation matrix T:

$$p = D * T = [Xg\ Yg\ y_i] \quad (1)$$

where the raw data vector is

$$D = [x_t\ y_t\ x_i\ y_i\ x_m\ y_m\ x_r\ y_r\ x_p\ y_p] \quad (2)$$

t= thumb, i= index, m= medium, r= ring, p= pinky.

Hand shape is sub-divided in two components, Xg and Yg. The first component measures abduction (finger spread) or rolling with respect to the wrist; the second component measures finger bending. The other component of posture space, palm orientation y_i, classifies postures into three broad subclasses: closed, horizontal, and vertical (or open). Index y-position is used for this purpose. Transformation matrix

$$T = \begin{vmatrix} 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \end{vmatrix}^T \quad (3)$$

extracts posture components as

$$Xg = \Sigma\, x_n,\ n \in \{t,i,m,r,p\}; \quad (4)$$

$$Yg = \Sigma\, y_n,\ n \in \{t,i,m,r,p\}; \quad (5)$$

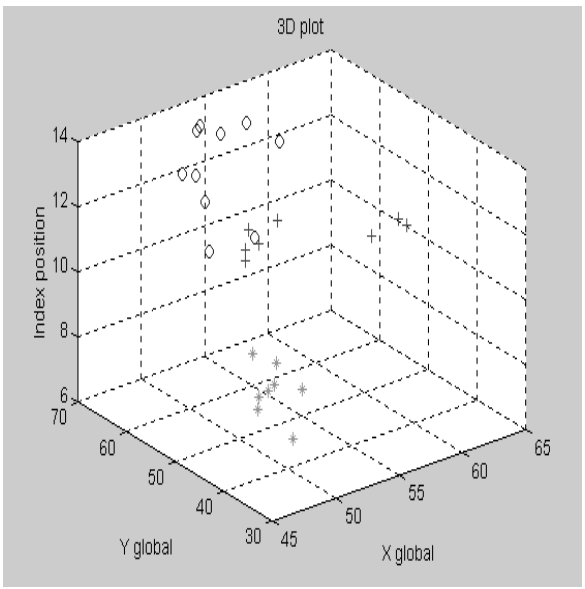
$$y_i = Data[4] \quad (6)$$

called: X-global position (Xg), Y-global position (Yg), and orientation y_i (subclass).

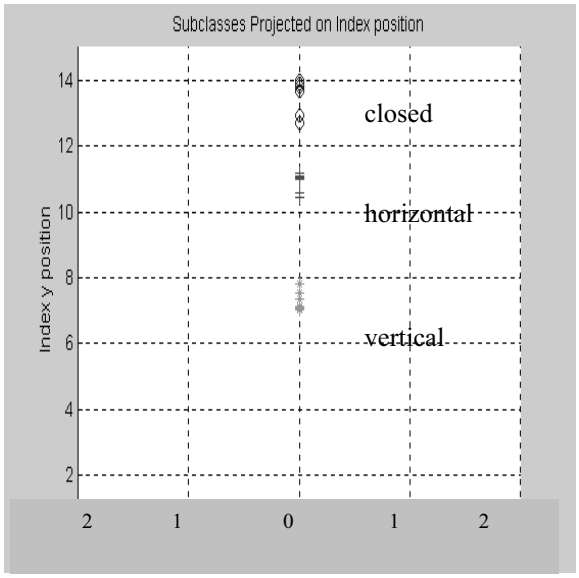
The first component, Xg, given by the sum of the x signals on the raw vector, tells us about the orientation of the palm; it measures hand roll when in a horizontal position and measures finger abduction when the fingers are vertical. The second component, Yg, fairly describes hand shape as the total number of fingers bent, obtained by adding all five y-signals. The fourth component of the 10-D vector is y_i. With these new features, patterns are easier to visualize as points in a 3-dimensional feature space.

4. Classification

Our sample space consists of 50 posture vectors *p* for each of the 26 classes (letters). Figure 4a shows the mean value over the 50 samples of every one of the 26 letters in posture space. Figure 4b shows the mean values projected onto the vertical axis, given by the y-position of the index finger y_i.



(a)



(b)

Figure 4: (a) 26 classes plotted as 3-dimensional vectors. Each point represents the mean over all patterns in the class.
(b) Projection of all classes onto index y axis. 'o'=closed, '+'=horizontal, '*'=vertical

In principle, it is possible to solve any multi-class recognition problem with a single classifier. However the complexity of such a classifier would be high, as was found in previous work. Instead, we divide the space using two planes, creating three major subclasses clearly grouped in Figure 4b: "Vertical," "Horizontal," and "Closed".

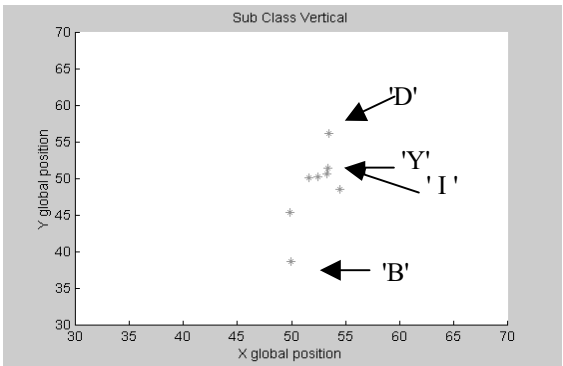
After this first hierarchical discrimination, the classifiers are of reduced complexity, and easy to implement using a single-chip PIC16f8XX micro controller. The process for recognizing a given posture can thus be described as follows:

Step 1: Subclasses are discriminated using two planes defined by the index finger position in Figure 4b.

Step 2: Members of each subclass are projected onto the plane that defines the subclass. In other words, only Xg and Yg are taken into account later in the classification process. Figure 5(a), (b) and (c) show this new reduction in dimension.

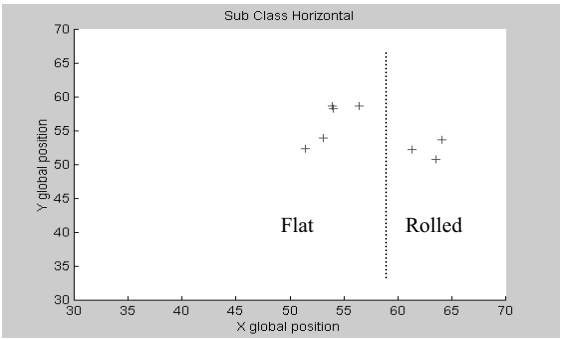
Step 3: If the new representation is not enough to discriminate a letter using a simple rule (Bayes' Rule, linear function), a new subclass is dispatched to another classifier that looks for particular differences from raw data.

Subclass Vertical is defined by all classes below the plane defined by $y_i = 9$. The letters 'B,' 'D,' 'K,' 'L,' 'R,' 'U,' 'V,' and 'W' belong to this subclass. Subclass Horizontal is placed between the planes defined by $y_i = 9$ and the plane defined by $y_i = 12$. The letters 'C,' 'E,' 'G,' 'H,' 'J,' 'P,' 'T,' 'X,' and 'Z' belong to this subclass. All classes with $y_i > 12$ belong to subclass Closed. That is the case for letters 'A,' 'F,' 'I,' 'M,' 'N,' 'O,' 'Q,' 'S,' and 'Y.'

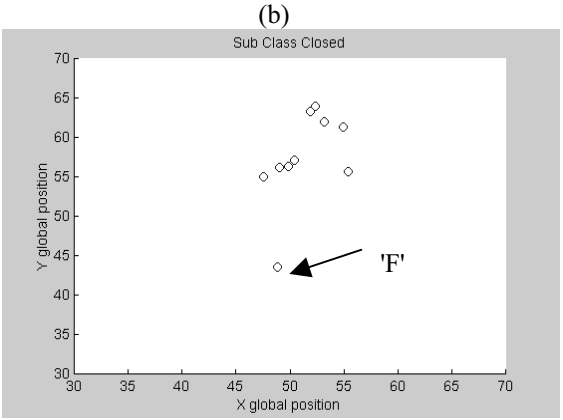


(a)

Figure 5 (a). Vertical subclass projected onto the plane defined by $y_i = 0$.



(b)



(c)

Figure 5: (b) Horizontal subclass projected onto the plane defined by $y_i = 8$. (c) Closed subclass projection onto the plane defined by $y_i = 12$. Letter 'F' is pointed by an arrow.

Notice that, even though the 'F' sign does not look like it uses a closed fist (see [1]), and that 'D' does not look like an open vertical hand, they were classified as such because the discrimination was made based on the state of the index finger only.

5. Hierarchical Structure

The decision tree uses the notation followed by Erenshteyn [3]. The first block in the hierarchy refers to feature extraction where Xg, Yg, and y_i are calculated using equations (1), (2) and (3), three-dimensional patterns are formed. The first dispatcher (level 1) performs a simple comparison on y_i to differentiate the three subclasses shown in Figure 4b. The second level of dispatchers will differentiate further into another set of subclasses based on similarities in their features, like 'flat' and 'rolled' postures in Figure 5b. A recognizer will look at attributes in the feature space to finally recognize letters out from the set of subclasses. As an example, in figure 5c, the letter 'F' is easily recognized from the rest by drawing a line at Yg=45. Like 'F', many letters are recognized in two steps while the most difficult (more alike to each other) go up to the bottom of the tree.

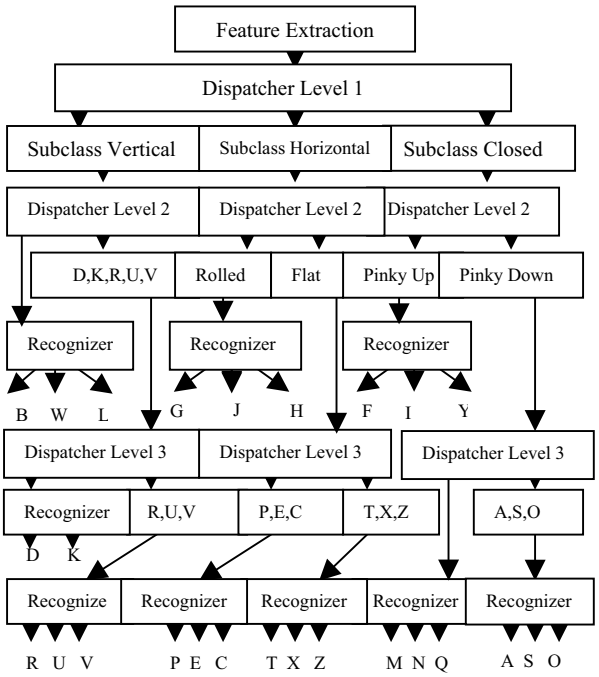


Figure 6. Classifier's Hierarchy.

6. Results

The classification method was implemented as a series of 'if-then-else' sentences using Matlab 5.0, and was tested using a total of 1,300 samples to measure the recognition rate, before the micro controller was programmed with the algorithm. Twenty-one out of the 26 letters reached a 100% recognition rate using re substitution method. Letters 'I' and 'Y' overlapped, as shown in Figure 5(a). To recognize them, Bayes' Rule was applied as follows: sample x is assigned to class wI if:

$$p(x|wI)\, p(wI) > p(x|wY)\, p(wY) \qquad (7)$$

where p(wI) is the prior probability of the letter 'I' and p(wY) is the prior probability of the letter 'Y.' To estimate conditional probabilities p(x|wI) and p(x|wY), we used the histograms with the class distribution over Xg.

The misclassification rate introduced by this Bayes' Rule, based on histograms, is estimated as the percentage of classified samples. In particular, 2 samples of Y are miss classified as members of class I, the error rate is: E(Y) = 2 / 50 = 4%.

The letters 'R,' 'U,' and 'V' represented the worst cases, as their class distributions overlap significantly, making it impossible for a linear function to discriminate them without errors; two bins of class 'U,' and one from classes 'V' and 'R' are misclassified dropping the accuracy to

90%, 78%, and 96%, respectively. The feature used to assemble the histogram in this case is the index X position, or finger abduction.

7. Conclusions and Future Work

A portable, battery-powered finger spelling recognizer has been built using state-of-the-art, but affordable, MEMS accelerometers. To our knowledge, the use of accelerometers at the PIP joints in a self-contained system to recognize the ASL alphabet is novel. The representation of signs as a set of 3-D patterns and the subsequent projection onto planes that allow the reduction of the problem dimensions seems also to be a novel approach in hand shape recognition. The system proved to be flexible and accurate for recognizing signs from different users even though they were not experimented signers. Unlike the CyberGlove and DataGlove, accuracy is not user-dependent, does not depend on hand size, and does not require user-specific calibration. The main problem of using accelerometers as angular position sensors is that they respond to change with respect to the gravitational vector ***g,*** but are insensitive to rotations around it. Then, is not possible to use the Accele Glove with horizontal postures, as are the NASA postures. An advantage of the projection method is that after the projection of a subclass onto its corresponding plane, it is possible to find empty spaces where a complete new sign can be introduced. That is the case with the 'space' ('B' posture with extended thumb) that was added to the ASL alphabet to indicate separation between words, and 'enter' ('thumbs up') which indicates that the word or sentence is complete and has to be sent to the synthesizer. These two commands really transformed the system into a Finger Spelling to Speech Translator.

Future work may include investigating the impact of changing weights in the transformation matrix T on the recognition rate of the most difficult letters.

References:

[1] E. Costello, "*Random House Webster's Concise American Sign Language Dictionary*". Random House N.Y., 1999

[2] S. Bryson. "*Implementing Virtual Reality*," number 43 in SIGGRAPH course Notes,. August 1993.

[3] R. Erenshteyn et. al. *"Distributed Output Encoding for Multi-Class Pattern Recognition".* Proceedings of the Intl. Conf. On Image Analysis and Processing 1999. 229-234.

[4] G. Grimes. *US Patent No. 4414537*, November 1983

[5] Kessler, Hodges, and Walker, "*Evaulation of the CyberGlove as a Whole Hand Input Device".* ACM Trans. On Computer-Human Interactions, 2 (4) pp. 263-283. 1995

[6] J. Kramer and L. Leifer *"The Talking Glove: An Expressive and Receptive "Verbal" Communication Aid for the Deaf, Deaf-Blind, and Nonvocal".* SIGCAPH 39, pp.12-15 (spring 1988).

[7] M.V. Lamart, M.S. Bhuiyant *"Hand Alphabet Recognition Using Morphological PCA and Neural Networks".* Proceedings of the International Joint conference on Neural Networks, Vol. 4, pp 2839-2844, Washington, USA, 1999.

[8] R. Liang, M. Ouhyoung.*"A Real-time Continuous Gesture Recognition System for Sign Language".* Proceeding of the Third IEEE Int. Conf. On Automatic Face and Gesture Recognition 1998. pp 558-567

[9] K. Pister and J. Peng, UC of Berkeley. "Big Promise in Thinking Small". *Los Angeles Times*, Science File Section by Kendall S. Powell, August 17th 2000.

[10] D. Quam, W. Williams, J. Agnew, and P. Browne *"An experimental determination of human hand accuracy with a DataGlov"'.* Proceedings of the Human Factors Society 33rd Annual Meeting, pp. 315-319

[**11**]. C. Shahabi, L. Kaghazian, S. Mehta, A. Ghoting, G. Hanbhag, M. McLaughlin. "Analysis of Haptic Data for Sign Language Recognition", *Volume 3 of the Proceedings of HCI International 2001, 1st International Conference on Universal Acces in HCI.* Edited by Constantine Stephanidis. August 5-10, 2001 New Orleans, Louisiana, USA, pp 441-445

[12] T. Starner, J. Weaver and A. Pentland *"A Wearable Computer Based American Sign Language Recognizer".* MIT Media Lab. Technical Report 425. 1998

[13] William Stokoe. "*Sign Language Structure: an Outline of the Visual communication system of the American Deaf*". Studies in Linguistics: Occasional Papers 8. Linstok Press, Silver Spring MD, 1960. Revised 1978.

[14] D.J. Sturman and D. Zeltser *"A Survey of Glove-Based Input".* IEEE Computer Graphics & Applications. 1994, pp 30-39.

[15] C. Uras and A. Verri.*"On the Recognition of The Alphabet of the Sign Language through Size Functions".* Dipartimento di Fisica, Universitá di Genova. Proceedings of the 12th IAPR Int. Conf. On Pattern Recognition. Conference B: Computer Vision and Image Processing 1994. Vol 2. 334-338.

[16] M.B. Waldron et al. *"Isolated ASL Sign Recognition System for Deaf Persons".* IEEE Trans. On Rehabilitation Engineering vol 3 No. 3 September 1995.

[17] M.W. Kadous *"Grasp: Recognition of Australian sign language using instrumented gloves".* Master's thesis, University of New South Wales, Oct. 1995

[18] R. Watson. *"A Survey of Gesture Recognition Techniques".* Technical Report TCD-CS- 93-11. Department of Computer Science, Trinity College, Dublin. July 1993.

[19] J. Weissman. *"Gesture recognition for Virtual Reality Applications Using Data Gloves and Neural Networks".* IJCNN 99. Intl. Conf. On Neural Networks, 1999, Vol 3. 2043-2046.

[20] S. Wise, W. Gardner, E. Sabelman. *"Evaluation of a fiber optic glove for semi-automated goniometric measurements".* Journal of Rehabilitation Research and Development, 27(4), pp 411-424, 1990.

[21] T. Zimmerman *"Optical Flex Sensor".* US Patent No. 4,542,291. 1987

[22] M. Zhao *"RIEVL: Recursive Induction Learning in Hand Gesture Recognition".* IEEE Trans. On Pattern Analysis and Machine Intelligence. Vol 20, No.11 Nov. 98.

[23] J. Hernandez, N. Kyriakopoulos, R. Lindeman. *"The AcceleGlove a Whole Hand Input for Virtual Reality".* SIGGRAPH 2002. San Antonio Texas July 21-26, 2002.