Phonetic Model for Automatic Recognition of Hand Gestures

Jose L. Hernandez-Rebollar
The George Washington University, Electrical and Computer Engineering
725 23rd St. NW. Lab 302, Washington DC 20052
jreboll@gwu.edu

Abstract

This paper discusses a phonetic model of hand gestures that leads to automatic recognition of isolated gestures of the American Sign Language by means of an electronic instrument. The instrumented part of the system combines an AcceleGlove and a two-link arm skeleton. The model brakes down hand gestures into unique sequences of phonemes called Poses and Movements. Recognition system was trained and tested on volunteers with different hand sizes and signing skills. The overall recognition rate reached 95% on a lexicon of 176 one-handed signs. The phonetic model combined with the recognition algorithm allows recognition of new signs without retraining.

1. Introduction

The development of automatic means to study sign languages is destined to have enormous impact on economy, society and science. Costello [1999] estimates that American Sign Language (ASL) is the fourth most used language in the United States with 13 million people, including members of both the hearing and deaf community. Some 300,000 to 500,000 of them are ASL native-speakers, which means that their full integration to society depends on their ability to overcome the language barrier by using all means at their disposal. William Stokoe [1995] was probably the first linguist to involve engineers, not only educators, in solving the challenge of better communication, he wrote: "Looking back, it appears that linguistics was made possible by the invention of writing. Looking ahead, it appears that a science of language and communication, both optic (gestures) and acoustic (speech), will be enabled, in all probability, not by refinements in notational systems, but by increasing sophistication in techniques of recording, analyzing, and manipulating visible and auditory events electronically."

It is ironic that even though humans learned how to communicate through gestures before learning how to speak, methodologies for analyzing speech and spoken languages are far better understood than the methodologies for analyzing and, in consequence, recognizing gestures and sign languages.

Engineers found a way to capture speech in 1915 with the invention of the carbon microphone. This transducer produces an electrical signal corresponding to change in air pressure produced by sound waves, which contains all the information required to record and reproduce speech through a speaker. Sign language, in turn, combines hand movements, hand shapes, body posture, eye gaze, and facial expression that are not easy to capture by using only one type of sensor. Approaches that use arrays of video cameras to capture signing struggle to find an adequate way of reproducing tri-dimensional images. The high resolution needed to capture hand shape and eye gaze results in a reduced field of view unable to fit hand movement or body posture, and a high bandwidth connection (processor) is required to transmit (analyze) the data stream and reproduce the video at acceptable speed.

An alternative is the combination of angular sensors of different types mounted directly on the signer's joints of interest. Although bulkier, cumbersome and more obtrusive, these instrumented approaches have been more successful in capturing hand postures [Grimes1983, Kramer1998] than the approaches based on video alone [Uras1994].

In this work the combination of a phonetic model of hand gestures and a novel instrumentation to capture and recognize the hand gestures in American Sign Language, is discussed. Non-manual components such as facial expression, eye gaze and body posture are not considered here.

2. Review of previous approaches.

The first and most important step in the recognition process is to extract, from a given gesture, all the necessary features that allow the recognition system to classify it as member of one and only one class. Two things are needed to achieve that step: a model that describes gestures in terms of necessary and sufficient features, and a capturing system suitable to detect such features. It is imperative for the resulting set of features (*pattern*) to be different for each gesture, and it is desirable for the resulting pattern to have a constant number of features (*fix dimensionality*) and as few as possible (*reduced dimensionality*).

The model proposed in this work is based on the assumption that any hand gesture can be analyzed as a sequence of simultaneous events, and each sequence is unique per gesture. Those events are referred in this work as phonemes. As straightforward as this scheme could sound, it could be cause of debate among many signers and teachers who conceive signs as indivisible entities. The following is a review of different phonemes and structures that have been proposed to model hand gestures.

2.1. Phonetic structure

By using traditional methods of linguistics to isolate segments of ASL, Stokoe found that signs could be broken down into three fundamental constituent parts: the hand shape (*dez*), hand location with respect to the body (*tab*), and the movement of the hand with respect to the body (*sig*), so these phonemes happen simultaneously. Lidell [1989] proposed a model of movements and holds, Sandler [1986] proposed movements and locations, and Perlmutter

[1988] proposed movements and positions, all of them happening sequentially.

Some automatic systems have followed models similar to Stokoe [Bauer, 2000; Vamplew, 1996] and Lidell [Vogler, 1999]. By using Stokoe's model, patterns are of reduced and fix dimensionality but similar for gestures that are only different in their final posture (such as GOOD and BAD). Patterns that result from Liddell's model eliminate this problem by considering the initial, final, and intermediate states of the hand and the movements that happen in between. Still, the model produces ambiguous patterns with variable dimensionality. As an example, when signing FATHER, tapping the thumb of a 'five' hand shape against the forehead, the sequence can be described as a Movement followed by a Hold followed by a Movement and finished by a Hold (MHMH) or as a HMHMH if the hand is considered to start from a static position, or as a simple Hold, as many signers do not make long movements when tapping. Closely linked to these models are the recognition methods suitable to recognize the resulting patterns. Hidden Markov Models (HMM) and Neural Networks (NN) have been used to recognize complete sentences [Starner, 1998], isolated words [Waldron, 1995], or phonemes [Vamplew, 1996], but none of those approaches has been able to integrate hand gesture and finger spelling in one recognition system.

2.2. The Pose-Movement model

Under the sequential models previously explained, ASL resembles the linear structure of spoken languages: phonemes make up words, and words in turn make up sentences. Phonemes in these models are, in some degree, the three simultaneous components of Stokoe, so the execution of ASL gestures can be seen as a sequential combination of simultaneous phonemes. Specifically, two types of phonemes: one static and one dynamic.

Definition 1: A ***pose*** is a static phoneme composed of three simultaneous and inseparable components represented by vector **P** = [hand shape, palm orientation, hand location]. The static phoneme occurs at the beginning and at the end of a gesture.

Definition 2: A ***posture*** is a vector of features **Ps** = [hand shape, palm orientation]. Twenty-four out of the 26 letters of the ASL alphabet are postures that keep their meaning regardless of location. Letters J and Z are not considered postures because they have movement.

Definition 3: ***Movement*** is a dynamic phoneme composed by the shape and direction of the trajectory described by hands when traveling between successive poses. **M**=[direction, trajectory].

Definition 4: A ***manual gesture*** is a sequence of poses and movements, P-M-P.

Definition 5: **L**, the set of purely manual gestures that convey meaning in ASL is called the ***lexicon***.

Definition 6: A manual gesture **s** is called a ***sign*** if s belongs to L.

Definition 7: ***Signing space*** refers to the physical location where signs take place. This space is located in front of the signer and is limited by a cube bounding the head, back, shoulders and waist.

By following definitions 1 to 7, icons, letters, initialized, and non-initialized signs, are modeled by PMP of fixed dimensionality, while compound, pantomimic, classifiers, and lexicalized finger spelled words, are modeled as sequences of variable length. These patterns are listed in Table 1.

Sign	Model
Two handed icons	PMP, PMP one sequence per hand
Finger spelled words	PMP per letter
Lexicalized finger spelled, *compound signs, **pantomimic	sequence of 2n-1 phonemes n= number of letters n=number or signs* n=number of pauses**

Table 1. Signs and their respective sequences of phonemes

As a proof of concept, a Lexicon of one-handed signs from two dictionaries [Costelo,1999; IDRT, 2001] with patterns of the form PMP were targeted for recognition. Since any sign is merely a new combination of the same phonemes, the recognition system is composed by small subsystems that capture a finite number of phonemes complemented by a search engine, which compares captured sequences against stored sequences.

3. Instrumentation

The instrument designed to capture all the phonemes found in the resulting sequences (53 postures, including six orientations; twelve movements and eleven locations) comprises an Acceleglove [Hernandez, 2002] to capture hand postures, and a two-link skeleton attached to the arm to capture hand location (with respect to the shoulder) and hand movement. Data is sent serially to a laptop Tthinkpad running windows 98 on a Pentium III. The sign recognizer is based on a search algorithm.

3.1 Training and testing

Posture, location and movement were recognized independently; trained and tested with help of 17 volunteers of different skill levels, from novice to native signer. That selection allowed covering a range of accents and deviations with respect to the citation form. The search algorithm was tested with 30 one-hand gestures first, and 176 later to test scalability. The complete list of signs is found in [Website].

3.2. Postures

The posture module starts recognizing any of six palm orientations: vertical, horizontal, vertical up-side down, horizontal tilted, horizontal palm up, and horizontal tilted counter clockwise.

Afterwards, the posture recognizer progressively discriminates postures by the position of fingers. Decision trees are generated as follows [Hernandez, 2002b].

-**For** all trees, **start** decision nodes evaluating the position of the pinky finger and base the subsequent node's decision on the next finger (ring, middle, index, thumb).

-**If** postures are not discriminated by finger flexion, **then** continue with finger abduction.

-**If** postures are not different by individual finger flexions or abductions, **then** base classification on the overall finger flexion and overall finger roll.

To train the orientation nodes, all seventeen signers were asked to hold the initial pose of FATHER, NICE, PROUD, PLEASE, THING and ASIDE. In average, the orientation module accurately recognized 94.8% of the samples. The worst recognition rate corresponded to horizontal postures where the threshold is blurred by the deviations introduced by signers' accents, since they were asked to hold their poses, not to hold their hand in a certain position.

3.2.1. Aliases

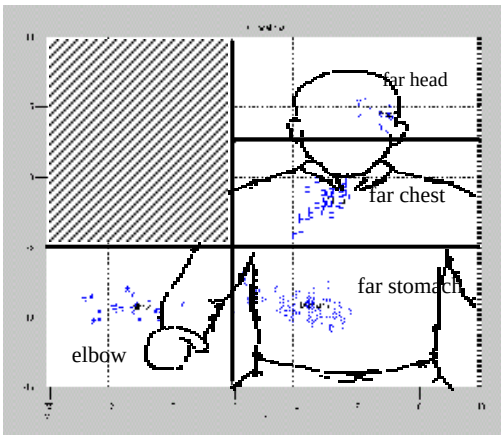
Since accelerometers do not detect angular positions around the gravity vector, 10 postures were impossible to discriminate based on finger bending or abduction around the gravity vector. These postures are called *aliases*. This aliasing reduced the number of recognizable postures from 53 to 43. The highest accuracy (100%) corresponded to the vertical palm with knuckles pointing down used to sign PROUD, the worst accuracy rate corresponded to postures C and E, with 68%, for a recognition average of 84%.

3.3. Locations

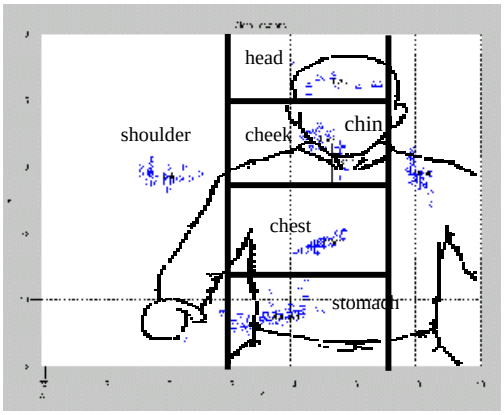
By looking at the initial and final position of the hand during the execution of each sign in the lexicon, eleven regions in the signing space were identified: head, cheek, chin, right shoulder, chest, left shoulder, stomach, elbow, far head, far chest and far stomach. To train the recognizer, four signers were asked to locate their hand at the initial poses of several signs that start or finish at those regions: FATHER, KNOW, TOMORROW, WINE, THANK YOU, NOTHING, WHERE, TOILET, PLEASE, SORRY, KING, QUEEN, COFFEE, PROUD, DRINK, GOD, YOU, FRENCH FRIES and THING. Volunteers were chosen based on their heights so they cover the full range of height among the group of volunteers.

Figure 1 shows the initial and final locations captured with the two-link skeleton as executed by the middle height signer (1.70 mts). Figure 1a corresponds to locations close to the body and Figure 1b corresponds to locations away from the body. A human silhouette is superimposed on the plane to show locations related to signer's body. The plane *y*-*z* is parallel to the signer's chest, with positive values of *y* running from the right shoulder to the left shoulder and positive values of *z* above the right shoulder.

Similar to orientations and postures, locations are solved using a decision tree, thresholds on *y* and *z* boundaries are set at least 4σ around the mean, and 3σ on *x* due limitations imposed by the instrumentation.



(a)



(b)

Figure 1. a) Far locations. b) Close locations.

The overall accuracy rate was 98.1% : head 98%, cheek 95.5%, chin 97.5%, shoulder 96.5%, chest 99.5%, left shoulder 98.5%, far chest 99.5%, elbow 94.5 %, stomach, far head and far stomach 100%. The skeleton system does not need an external reference source, and it is immune to ambient noise; that makes it a better choice for a portable instrument that infrared and magnetic trackers.

3.4. Movements

Movements of the one-handed signs considered in this work are described by means of two movement primitives: curviness [Bevilaqua2001] and direction. Both metrics are orientation and scale independent. As with the case of hand postures and locations, the exact movement varies from signer to signer and from trial to trial. Six directions (up, down, right, left, towards, and away) and two levels of curviness (straight and circular) were identified in the Lexicon that gave a total of twelve different movements. Same four signers were asked to perform the six basic movements along the main axes and the two curves ten times each. Directions ‘left’ and ‘right’ were classified with less than 100% (77% and 75%) reducing overall accuracy to 92%. A curviness greater than 4 discriminated circles from straight lines with 100% accuracy, but only signs with straight movements were implemented in the recognition algorithm.

4. Search Engine.

A variation of template matching called *conditional template matching* was used to classify complete signs. Conditional template matching compares the incoming vector of phonemes (captured with the instrument) against a pre-stored file of patterns, component by component, and stops the comparison when a condition is met:

-**For** all patterns in the lexicon, extract a list of signs matching the initial **posture** captured by the Acceleglove. This is the first list of candidate signs.

-**For** all patterns in the list of candidates, select the signs matching the **initial location** captured by the two-link skeleton. This is the new list of candidate signs.

Repeat the matching and creation of new lists of candidates by using movement, final posture and final location.

Stop when all components have been used **OR** when there is only one sign on the list after matching the initial location. That sign on the list is called 'the most likely'.

The search algorithm can be seen as a decision tree with a variable number of nodes. The expected probability of finding a given sign is inversely proportional to the depth of the tree. In other words, it is more likely to recognize a sign if it is the only one in the lexicon performed with certain initial pose (such as PROUD), and it is less likely to recognize two signs when only the final pose makes them different (such as GOOD and BAD).

4.1. Evaluation.

An initial evaluation used only 30 signs taken from Starner (1998), Vogler (1999), and Waldron (1995): BEAUTIFUL, BLACK, BROWN, DINNER, DON'T LIKE, FATHER, FOOD, GOOD, HE, HUNGRY, I, LIE, LIKE, LOOK, MAN, MOTHER, PILL, RED, SEE, SORRY, STUPID, TAKE, TELEPHONE, THANK YOU, THEY, WATER, WE, WOMAN, YELLOW, and YOU. The PMP sequences reflect the citation forms as found in Costello [1999] and in the Ultimate ASL Dictionary [IDRT2001]. The overall recognition rate was 98% since almost all of them have different initial poses.

4.2. Scalability

Since any new sign is a combination of the same phonemes, the lexicon can be expanded without retraining the search algorithm. When tested on 176 one handed signs performed by one signer the overall recognition rate reached 95%.

5. Conclusions and Future Work

The model, instrumentation and recognition algorithm explained in this work represent a framework for a more complex system where a larger lexicon can be recognized by extending the patterns to include non-manual gestures when the required instrumentation to detect them becomes available.

Work in the immediate future will incorporate a second PMP sequence for the non-dominant hand, and migrate the recognition program to a wearable computer for a truly portable electronic translator. The long-term objective shall include a grammar correction module to rearrange the sequence of translated glosses and correct for tenses, gender, and number as needed by the spoken language.

6. References

- Bauer, B., Hienz, H., and Kraiss, K., 2000. Video-Based Continuous Sign Language Recognition Using Statistical Methods. IEEE 2000, pp 463-466.
- Bevilacqua F., Naugle L., and Valverde I., 2001. Virtual Dance and Music Environment Using Motion Capture. Proc. of the IEEE-Multimedia Technology and Applications Conference, Irvine, CA.
- Costelo, Eleine 1999. Random House Webster"s Concise American Sign Language Dictionary. Random House Inc. NY.
- Fels, Sidney S., and Hinton, Geoffrey E., 1993. Glove Talk -A Neural-Network Interface Between a Data-Glove and a Speech Synthesizer. IEEE Transactions on Neural Networks, vol. 4, No. 1. January.
- Grimes, G. 1983. US Patent 4,414,537. November.
- Hernandez Jose L., Kyriakopoulos, N., Lindeman, R.. The AcceleGlove a Hole-Hand Input Device for Virtual Reality. ACM SIGGRAPH Conference Abstracts and Applications 2002. pp 259.
- Hernandez, Jose L., Kyriakopoulos, N., Lindeman R. A Multi-Class Pattern Recognition of Practical Finger Spelling Translation, IEEE International Conference on Multimodal Interfaces ICMI'02. October 2002, pp 185-190.
- IDRT 2001. The Ultimate American Sign Language Dictionary. The Institute for Disabilities Research and Training Inc. Copyright 2001.
- Kramer, J., and Leifer, L., 1988. The Talking Glove: An Expressive and Receptive Verbal Communication Aid for the Deaf, Deaf-Blind, and Nonvocal, SIGCAPH 39, pp.12-15 (spring 1988).
- Lidell S., and Johnson, R., 1989. American Sign Language: The phonological base. Sign Language Studies, 64: 195-277.
- Perlmutter, D., 1988. A moisac theory of American Sign Language syllable structure, paper presented at Second Conference on Theoretical Issues in Sign Language Research. Gallaudet University, Washington, DC.
- Starner, T., Weaver, J., and Pentland, A., 1998. A Wearable Computer Based American Sign Language Recognizer, MIT Media Lab. Technical Report 425.
- Stokoe, William C., Armstrong, David F., Wilcox, Sherman E. Gesture and the Nature of Language. Cambridge University Press, 1995
- Uras, C., and Verri, A. 1994, 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 Vol 2,1994. pp 334-338.
- Vamplew, P., 1996, Recognition of Sign Language Using Neural Networks, Ph.D. Thesis, Department of Computer Science, University of Tasmania.
- Vogler, C. and Metaxas, D., 1999, Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes. Gesture Workshop"99, Gif-sur-Yvette, France, March 17-19.
- Waldron Majula B. 1995, Isolated ASL Sign Recognition System for Deaf Persons, IEEE Trans. On Rehabilitation Engineering vol 3 No. 3 September.
- Sandler, W. 1986, The Spreading hand autosegment of American Sign Language. Sign Language Studies 50: 1-28.
- website: http://home.gwu.edu/~jreboll/signlist.txt