# Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language

Mohammed Waleed Kadous (waleed@cse.unsw.edu.au)
Computer Science & Engineering, University of New South Wales

**Abstract**

Instrumented gloves use a variety of sensors to provide information about the user's hand. They can be used for recognition of gestures; especially well-defined gesture sets such as sign languages. However, recognising gestures is a difficult task, due to intrapersonal and interpersonal variations in performing them. One approach to solving this problem is to use machine learning. In this case, samples of 95 discrete Australian Sign Language (Auslan) signs were collected using a PowerGlove. Two machine learning techniques were applied – instance-based learning (IBL) and decision-tree learning – to the data after some simple features were extracted. Accuracy of approximately 80 per cent was achieved using IBL, despite the severe limitations of the glove.

## Introduction

Sign language recognition is interesting for a number reasons; it represents an interesting domain in itself with obvious real-world applications, but it also makes a good starting point for gesture recognition in general because sign language has a strong research foundation, the signs are well-defined, and the signs have well-defined meanings.

## Auslan and other sign languages

Auslan (Johnston, 1989) is the language used by the Australian Deaf community. It has strong similarities to British Sign Language (BSL) and marginal similarities to American Sign Language (ASL). The language contains approximately four thousand well-defined signs. It has its own grammar; the main difference from the grammar of spoken languages such as English being that it is less order-dependent, less tense-sensitive and more concise (on average one sign represents two English words). In addition, signers sometimes mime to describe an object or situation. They also use places in space in the same way non-signers use pronouns in spoken language – that is as a temporary place-holder for some entity. Signers also use finger-spelling (where there is a sign for each letter of the English alphabet) to communicate concepts not easily expressed in Auslan. Unlike ASL, Auslan has a two-handed finger-spelling alphabet.

Much like spoken language, there is usually no pause between consecutive signs. This poses serious problems for recognition, and the sign segmentation problem is one that remains difficult, since deciding when one sign finishes and the next starts is not easy.

In these experiments, the segmentation problem, mime aspects of sign language, finger-spelling and spatial pronouns are not handled. We focus on recognising isolated signs.

## Instrumented gloves

Instrumented gloves (Sturman & Zeltzer, 1994; Sturman, 1992) have been used extensively recently, mostly for direct manipulation for virtual environments (Rheingold, 1991). But they may also be used for gesture and sign language recognition. They have several advantages (and some disadvantages) when compared to video methods. Their advantages are:

- Glove systems compare well in terms of cost to video systems, especially with current improvements in glove technology.
- The processing power and bandwidth requirements for real-time video processing are high. The data extracted from a glove are concise and accurate compared with the information from a video camera.
- Certain data are very difficult to extract from visual images – such as hand orientation, forward/backward motion and finger position information (due to occlusion).
- Gloves can potentially be used without modification of the environment (although the current generation do), while cameras need to be set up in the environment.

On the other hand, gloves are an encumbrance to the user. There are also several technical problems which need to be resolved, such as automatic calibration of the gloves and the handling of noise.

In this experiment, a PowerGlove, originally designed for use with the Nintendo Entertainment System, was used. The PowerGlove is at the bottom of the instrumented glove market, and although no longer produced, can still be obtained for approximately US$50 or less. It provides the following data (Student Chapter of the ACM, 1994):

- x, y and z position, relative to the point at which the glove is synchronised. Each can have one of 255 possible values. It tends to be very noisy, as this is implemented by way of ultrasonic sensors, which are only approximately linear.
- Wrist roll in 30-degree increments.
- Finger bend for each of the first 4 fingers. Each finger can have 4 possible values. This can also be noisy, since it depends on how well the glove fits.

## Machine Learning

Machine learning (ML), as the name suggests, is about making computers learn. One of the traditional ML formalisms is categorisation or classification; where we are given objects which share similar properties (or attributes) and we know what type (or class) of object they belong to. Our goal is to find a way of classifying a new object of unknown class.

Many techniques have been formulated for this; including instance-based learning, neural networks, rule-learning systems, decision tree building systems, genetic algorithms and inductive logic programming. Each derives and expresses its classification scheme in a different way.

Clearly, the quality of the recognition is closely related to the attributes we provide the classifier. Most objects of interest have many possible attributes, few of which are useful for classification. Attribute selection is therefore important.

In this experiment we considered two ML techniques that have not been extensively used in gesture recognition: Instance-based learning and decision-tree building.

Instance-based learning (Aha, Kibler & Albert, 1990; Cover, 1968; Cover & Hart, 1967), also known as 1-nearest neighbour, works by storing all the training instances in "attribute space". Given a test instance, it finds the closest instance in the attribute space and classifies the test instance according to this "nearest neighbour". There are many variations on this focussed mainly on (a) limiting the instances kept (b) adjusting what is meant by "closest" and (c) exactly how the the classification is made (for example, we might look at the five nearest instances and use a "vote" technique).

Decision tree building (Quinlan, 1993) works by building a hierarchy of decisions based on attribute values. For instance, we might want to learn when to play and when not to play golf, given weather conditions. We might think that the important attributes are the rain level (qualitatively), the wind level and the temperature. We would give examples to the decision tree builder of what we would do for a given situation, and the decision tree builder might produce a tree as shown in figure 1. For these experiments, we used C4.5 (Quinlan, 1993) as the decision tree builder.

## Previous work

Murakami and Taguchi (Murakami & Taguchi, 1991) tried to recognise ten signs using instrumented gloves and a very large (403-node) recurrent neural net and achieved an accuracy of 96 per cent on a random sample. Charayaphan and Marble (Charayaphan & Marble, 1992) tried 31 ASL signs using a video camera, but sampled each sign once and simulated the variation and consistently got 27 out of the 31 correct, with the remaining four sometimes correctly classified. Starner (Starner, 1995; Starner & Pentland, 1995) tried to recognise brief sentences of ASL (with a vocabulary of 40 signs) using video cameras and coloured gloves; and obtained accuracies of 91.3 per cent on raw signs and 99.2 per

```
if Rain = Heavy then
  if WindLevel = Low then play-golf.
  else
    if TemperatureC < 10 then dont-play-golf.
    else play-golf.
else
  if TemperatureC > 35 then dont-play-golf.
  else play-golf.
```

Figure 1: A simple decision tree (represented here as if-then-else statements) for deciding whether or not to play golf.

cent by using a very strict grammar for sentences. Starner also tried to recognise signs without the use of gloves (Starner & Pentland, 1996), with a corresponding decrease in accuracy (91.9 per cent with strict grammar, 74.5 per cent without).

Some work has also been done on the segmentation problem. Using an instrumented glove, Ohira et al. (Ohira, Sagawa & Sakiyama, 1995) built a system for segmentation, based on rests in motion and velocity envelopes. This appeared to be accurate.

## Experimental setup

95 signs found in Auslan were selected on the basis of frequency of occurrence, coverage of handshapes and complexity. Some pairs of signs were deliberately chosen due to their similarity. Some information on the signs selected is shown in table 1.

(a) *Hand usage*

| Item | Number of signs |
|---|---|
| One-handed | 59 |
| Double-handed | 27 |
| Two-handed | 9 |

(b) *Reasons for selection*

| Item | Number of signs |
|---|---|
| Common sign | 66 |
| Similarity | 11 |
| Rare Handshape | 9 |
| Sign Complexity | 5 |
| Other | 4 |

Table 1: Some important statistics on the signs used.

To collect these signs, a single (right-handed) PowerGlove was attached to an SGI Iris 4D/35, despite there being several two-handed and double-handed signs in the list[1].

[1]Two-handed signs have different handshapes in each hand, while double-handed signs have the same handshape.

Between 8 and 20 samples were then obtained from each of five signers for each of the 95 signs. The order of the signs were randomly permuted[2] to avoid fatigue affecting results, and the signer was asked to make a sign. Each sign was made discretely, beginning and ending at a well-defined location in space. In this way a total of 6 650 signs were collected.

The data were put through a simple "glitch" filter to remove data that are not plausible, caused by ultrasonic noise. No filtering was performed on the learning instances; all were fed to the learning algorithms.

Algorithms were tested using 5-fold cross-validation. This means that the data collected was split into 5 equally sized sets. Each time, one set was used as the test set and the remainder as training. This was done with each of the sets and the results averaged. This strikes a medium between computational efficiency, and rigorousness (where we train on every instance except one and test on the single instance – known as leave-one-out cross-validation).

## Feature Extraction

As mentioned previously, feature extraction is critical to the success of the recognition process. Thus a key part of the recognition is finding a set of attributes which accurately describe a sign. The accuracy of each feature set was determined individually. The following features were tested:

### Distance, Energy and Time

Clearly the distance covered in making a sign can be a good discriminant. Furthermore, even though signs can cover similar distances, sometimes the gestures may be more energetic, such as those involving making small circles with the hand. Simple techniques were used to give an approximation of the distance and energy of each sign. Also, some signs take longer to make than others, so this was thought to be also a potentially useful attribute.

This did not turn out to be the case. It appears that the noise generated by the glove dominates the measurement of distance and energy, and that the length of time required to make the sign is not a good discriminant. The accuracy obtained using these three attributes was approximately 8 per cent.

### Bounding Box

The bounding box of a sign is the box in space in which the sign fits. The bounding box can be represented as 2 vectors: the coordinates of the bottom left-hand near corner of the

[2]Except in the case of the first person as a control to see the effect that it would have.

box $(x_{min}, y_{min}, z_{min})$; and the coordinates of the upper right-hand far corner of the box $(x_{max}, y_{max}, z_{max})$.

The results of using the bounding boxes are good. They provide accuracy of approximately 30 per cent for both C4.5 and IBL. Bounding boxes may work well because of their insensitivity to random noise. However, they are still sensitive to "glitch" noise, since one "glitch" can easily throw out the whole bounding box.

### Histograms

A number of histograms were derived from the data. The application of histograms to gesture recognition is novel to the investigator's knowledge. Histograms work by segmenting a range of values into sub-regions and then working out the relative amount of time spent in that sub-region. For example, we might find that the x-position was between 0 and 0.5 60 per cent of the time and between 0.5 and 1.0 40 per cent of the time.

A complicating issue with histograms for ML is the optimum number of divisions – whether we should divide into two divisions as above, or five, which would give us ranges of 0 to 0.2, 0.2 to 0.4 and so on. If we have too many divisions, noise will interfere with values, and it will be liable to too much sensitivity. If we have too few divisions, the sign will not be sufficiently characterised to aid in its recognition.

Histograms were calculated on the following pieces of information:

**x, y and z position:** These proved to be good discriminants. With IBL, approximately 25 per cent accuracy was obtained. With C4.5 they did worse, with approximately 15 per cent accuracy.

**Wrist rotation and finger bend:** These too appeared to be good discriminants, achieving accuracies of 40 per cent with IBL and 30 per cent with C4.5.

**Distance and energy:** These form an approximation of the energy spectrum of a sign. The results were not as promising and seem to be plagued by the same problems as the global distance and energy measurements, obtaining accuracies of approximately 4 per cent.

It was found that in the case of the position histograms that 6 divisions worked best. For the rotation and finger bend, accuracy was so low that division was not an issue.

### Time division

Another technique is to segment the sign into a fixed number of equally sized segments and then calculate the average values of x, y and z position, wrist rotation and finger bend for each segment. These can again be used as the basis for comparison. The question

arises, as before, as to the optimal number of segments. Too many, and they will be extremely sensitive to variation in time and noise. Too few and it will not characterise signs sufficiently to be useful.

With IBL, accuracy was approximately 65 per cent and with C4.5, the accurracy was approximately 40 per cent. It was empirically found that five segments led to the best results.

## Synthesis

The best of the above attributes were selected: the x, y, z position histograms; rotation and finger bend histograms and segment averages. Note that this is a simplistic approach; just because the attributes work well individually does not mean that attributes will work well when combined; and conversely for poor attributes.

Using the best attributes to an accuracy of approximately 80 per cent for the three large samples collections using IBL (11 samples/sign = 80.6 per cent, 13 samples/sign = 81.4 per cent, 16 samples/sign = 83.0 per cent). For the control case, with no fatigue and 6 training samples, the accuracy was 87.4 per cent. For the other person with 6 training samples (and fatigue) accuracy was 58.5 per cent. Performance with the decision tree builder was significantly worse than with instance-based learning (performance between 35 per cent and 55 per cent). Considering the capabilities of the glove and the size of the lexicon, the investigator believes the results are very promising.

## Error Analysis

To better understand the errors made by such a system, an attempt was made to visualise the data collected from the glove, and to analyse the types of error made. There were many reasons for the errors – among them:

- While one hand was usually sufficient to tell signs apart, it was not always. Some signs differed only in the position of the non-dominant hand. Some signs therefore looked very similar for the data captured (for example, money compared to buy).
- The glove provided erroneous positional data when the transmitters (on the glove) were pointing away from the receivers (on the monitor), and there was no way to tell where the hand was pointing in the lateral plane (for example, the signs me and you could not be told apart very well).
- The little finger can be very important in sign languages, since it is usually used to indicate "bad", which tend to be compounded into other signs (e.g. sick, wrong, headache all use an extended fifth finger). The PowerGlove does not provide this data.

- Several signers complained that the glove was limiting their ability to sign naturally. The glove limits hand motion and limits the degree to which the fingers can be moved separately.
- The features extracted were too coarse to pick up some of the slight differences in repeated motions. For example, who and what are both in the same place, but what is a to-fro motion, whereas who is a circular motion.

# System behaviour

Accuracy is not the only consideration in sign language recognition. It is also important to see what effect the number of samples for each sign has on the error rate, since we would expect that more samples will reduce the error rate. For each of the three large sample sets (16 or more samples per sign), the error rate for a given number of samples was calculated from 2 to 14. The results are shown in figure 2.

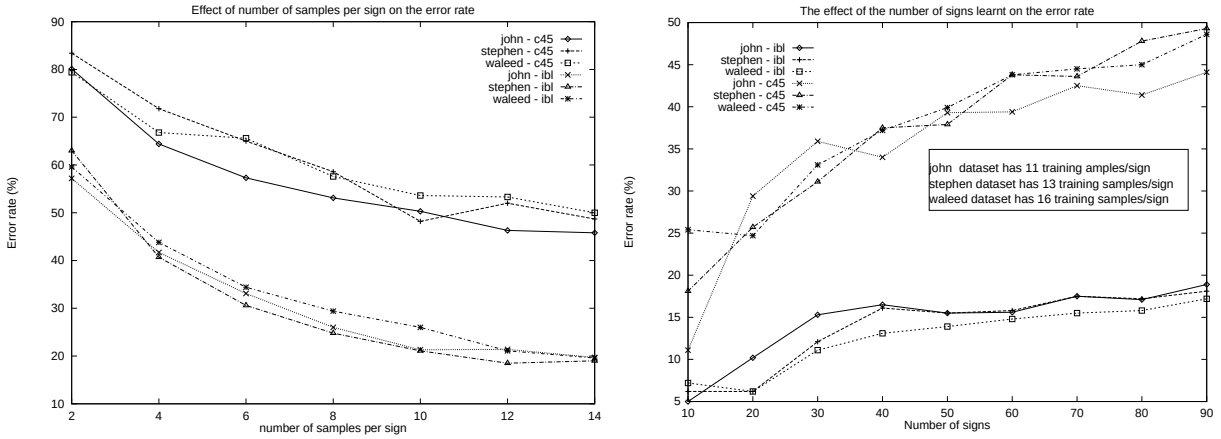


Figure 2: The effects of the number of signs and number of samples for each sign on the accuracy of the system.

As can be seen, an increasing number of samples results in better performance at a decreasing rate. Thus the recognition would appear to improve with use.

The impact of expanding the lexicon on the error rate is also an important consideration. Thus, limited numbers of randomly selected signs were tested to see how the number of signs learnt affect the error rate. The results are shown in figure 2.

As can be seen, there is – as expected – an increase in the error rate with the number of signs. But its effects seem to taper off at a less than linear rate. This seems to indicate the potential expansion to large lexicons is possible.

### Time Performance

The calculations performed were simple and could easily be accomplished in real time, at least with 95 signs. As is, a real-time system built out of UNIX pipes and perl scripts was assembled on an SGI 4D/35[3]. Furthermore, the algorithms used were shown to be approximately $O(n)$ in learning and $O(\log n)$ for recognising. This is a good indication for future real-time system development.

### Inter-signer Learning

As an aside, inter-signer learning was investigated, that is: how well does the system recognise the signs from people on whom it has not been trained? To test this, the system was trained on four people and tested on the fifth, for each of the five users. The results were not promising, with accuracies of approximately 12 to 15 per cent. However, it must be noted that no attempt was made to calibrate the data that comes from different people. Further efforts to "match" people to some more accurate model of motion may be effective.

## Conclusion

A large set of isolated signs from a real sign language can be recognised with some success using a low-end instrumented glove, some simple feature extraction and machine learning. The main limitation appears to be the glove itself, and not the techniques, as the types of error that occur indicate. As glove technology improves, we can expect better accuracy.

Furthermore, these techniques appear to be generalisable to larger lexicon systems, with the error rate behaving in a better than linear manner. Similarly, the more instances are used for learning, the better the accuracy.

Of course, there are many problems that remain to be solved – most importantly the segmentation problem – before such a system can be of any practical use.

## Acknowledgements

The author would like to thank The Creator for giving him the ability to do this research. Also, Dr Andrew Taylor for supervising this work and offering many useful suggestions and Professor Claude Sammut for his help in proofreading. For a more complete list of who helped with this research, please consult the thesis at the above address.

[3]This is a MIPS R3000 33 MHz machine whose CPU specifications are now exceeded by most modern desktop PCs.

# References


Aha, D. W., Kibler, D. & Albert, M. K. (1990). Instance-based learning algorithms. *Draft submission to Machine Learning.*

Charayaphan, C. & Marble, A. (1992). Image Processing system for interpreting motion in American Sign Language. *Journal of Biomedical Engineering, 14*, 419–425.

Cover, T. M. (1968). Estimation by the nearest neighbour rule. *IEEE Transactions on Information Theory, IT-14(1)*, 50–55.

Cover, T. M. & Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory, IT-13(1)*, 21–27.

Johnston, T. (1989). *Auslan Dictionary: a Dictionary of the Sign Language of the Australian Deaf Community.* Deafness Resources Australia Ltd.

Murakami, K. & Taguchi, H. (1991). Gesture recognition using recurrent neural networks. In *CHI '91 Conference Proceedings* (pp. 237–242). ACM.

Ohira, E., Sagawa, H. & Sakiyama, T. (1995). A Segmentation Method for Sign Language Recognition. *IEICE Transactions on Information and Systems, E78-D(1)*, 49–57.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Rheingold, H. (1991). *Virtual Reality.* Touchstone Books.

Starner, T. (1995). Visual recognition of American Sign Language using Hidden Markov Models. Master's thesis, MIT Media Lab. URL: ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-316.ps.Z.

Starner, T. & Pentland, A. (1995). Visual recognition of American Sign Language using Hidden Markov Models. Technical Report TR-306, Media Lab, MIT. URL: ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-306.ps.Z.

Starner, T. & Pentland, A. (1996). Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. Technical Report TR-375, Media Lab, MIT. URL: ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-375.ps.Z.

Student Chapter of the ACM (1994). *Power Glove Serial Interface* (2.0 Ed.). Student Chapter of the ACM, UIUC.

Sturman, D. J. (1992). *Whole Hand Input.* PhD thesis, MIT. URL: ftp://media.mit.edu/pub/sturman/WholeHandInput/*.

Sturman, D. J. & Zeltzer, D. (1994). A survey of glove-based input. *IEEE Computer Graphics and Applications, 14(1)*, 30–39.