# Recognition of sign language gestures using neural networks

Peter Vamplew

Department of Computer Science, University of Tasmania
GPO Box 252C, Hobart, Tasmania 7001, Australia

*vamplew@cs.utas.edu.au*

## ABSTRACT

This paper describes the structure and performance of the SLARTI sign language recognition system developed at the University of Tasmania. SLARTI uses a modular architecture consisting of multiple feature-recognition neural networks and a nearest-neighbour classifier to recognise Australian sign language (Auslan) hand gestures.

**Keywords:** sign language, hand gestures, communication aid

## 1. INTRODUCTION

Sign languages such as BSL and Auslan (Australian sign language) are the primary form of communication between members of the Deaf community. However these languages are not widely known outside of these communities, and hence a communications barrier can exist between Deaf and hearing people. The hand tracking technologies developed for VR enable the possibility of creating portable devices which can convert sign language to speech, as an approach to overcoming these difficulties. The SLARTI system is a prototype of such a device, based on Auslan.

## 2. SYSTEM DESIGN

### *2.1 Input Hardware*

In computer recognition of spoken language, speech data is captured using a microphone connected to an analog-to-digital converter. Similarly a data-capturing device is also required in order to recognise sign language; in this case measuring the position and movement of the signer's hands. Two broad categories of input hardware have been used for recognition of hand gestures – glove-based devices such as those used by Kramer et al (1989) and Fels et al (1993), and camera-based systems as used by Holden (1993). The latter approach has some benefits, particularly as it does not require specialised hardware, but this is offset by the complexity of the computer vision problems faced in extracting the necessary data about the hands from a visual image. Therefore for this research glove-based input was used, as this allowed the research effort to be focused on the area of sign recognition rather than that of data capturing.

The specific input devices used in developing SLARTI were a CyberGlove, and a Polhemus IsoTrak. The CyberGlove measures the degree of flexing of the various joints of the hand and wrist. The version of the CyberGlove used for this research provides 18 sensors. The Polhemus allows tracking of the spatial position and orientation of the hand with respect to a fixed electro-magnetic source. Using only a single glove restricts the system to the recognition of one-handed signs (and hence eliminates the possibility of recognising the Auslan manual alphabet which is two-handed), but it is envisaged that the techniques used in developing the system could be extended to two-handed signs if appropriate input hardware was available.

### *2.2 System Architecture*

Linguistic analysis of sign language has revealed that signs can be described in terms of four basic manual features, which may be modified in meaning by more subtle factors such as body language and facial expression (see for example Johnston 1989). The handshape defines the configuration of the joints of the hand. Orientation specifies the direction the hand and fingers are pointing, whilst the place of articulation is the location of the hand relative to the body. The most complex feature is motion, which consists of a change over time of any combination of the other three features (although for this research only changes in location have been considered).

The task of transforming a stream of input data directly into a classification of the sign being performed is an extremely difficult one. Instead the approach taken within SLARTI was to initially process the input data so as to

produce a description of this sequence in terms of the four features discussed above. The sign can then be classified on the basis of this feature vector. The SLARTI system consists of four separate feature-extraction neural networks, each trained specifically to recognise one of the features of the sign. The feature vector produced by these networks is then used to perform the overall classification of the input sequence, as shown in Figure 1.

This approach of decomposing the problem and applying a modular structure of networks has a number of benefits. First, as demonstrated on the task of speech-recognition by Waibel et al (1989), it allows the use of several smaller networks rather than one massive network and thereby reduces the amount of training time and data required. It may also result in superior classification accuracy. Second, it produces a system which can more easily be extended. The features recognised by the feature-extraction networks are expressive enough to describe an extremely large number of signs, not all of which may be recognised by the final classifier. If the vocabulary of the system is to be extended then only the final classifier will require modification. This greatly reduces the costs involved in performing such expansion of the system, and makes it practical to tailor the vocabulary of the system to a particular user.

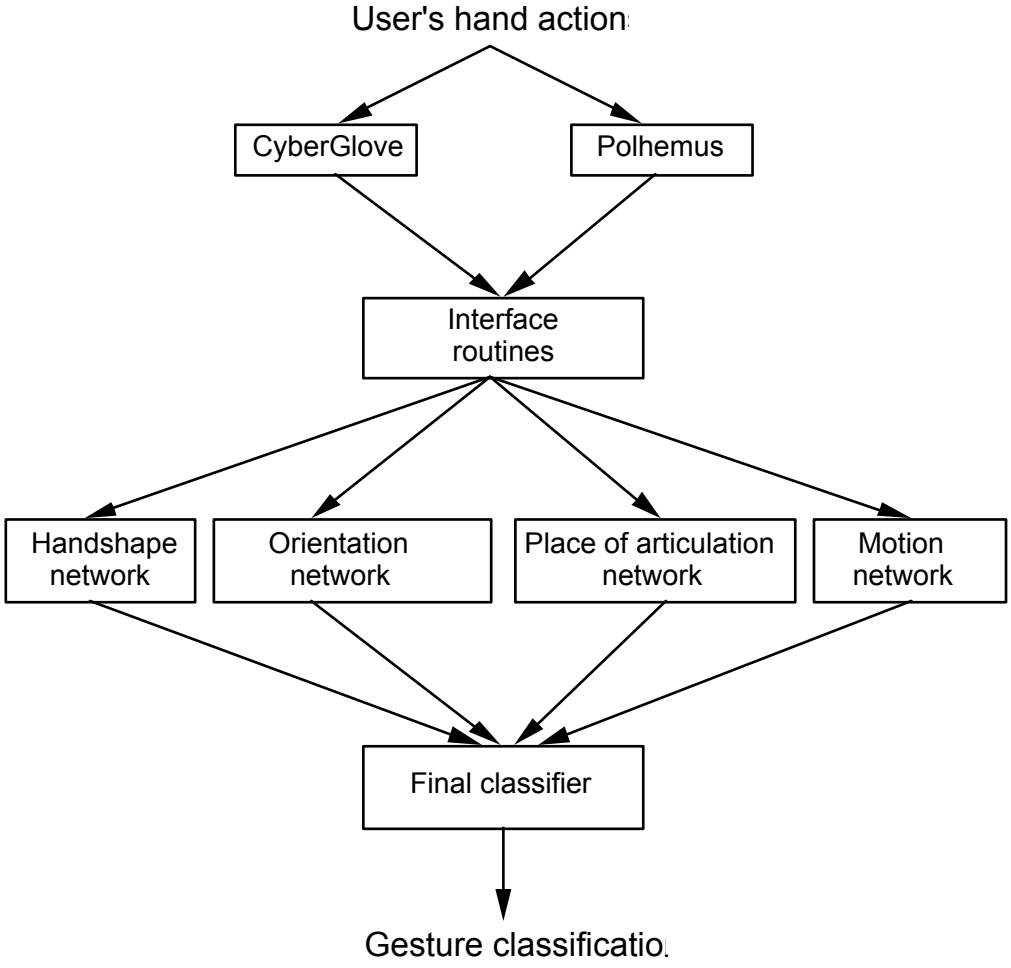


**Figure 1.** *The modular architecture of the SLARTI system*

## 3. FEATURE EXTRACTION NETWORKS

*3.1 Data Gathering and Training Methodology*

All of the feature-extraction networks were trained on examples gathered from 7 signers (which will be referred to as the registered signers), and tested on both fresh examples from the same signers and examples from 3 other signers (the unregistered signers), so as to assess the possibility of creating a signer-independent system. A fully-connected feed-forward architecture with a single hidden layer was used for all four networks and backpropagation without momentum was used as the training algorithm. All input data were scaled to lie in the range -1 to 1. The results reported are the average of results over 25 trials from different starting weights.

*3.2 Handshape Recognition Network*


Proc. 1st Euro. Conf. Disability, Virtual Reality & Assoc. Tech., Maidenhead, UK, 1996

Johnston (1989) identified 30 different primary handshapes used in Auslan, which can be further subdivided into 61 variant handshapes, although for the purposes of classifying signs it is only necessary to be able to distinguish between the primary handshapes. For each of the registered signers, 4 examples of each of the 61 variant handshapes was gathered for use in a training set. A further example of each handshape was gathered from all 10 users to constitute the 2 test sets. Prior to gathering the handshapes each user was asked to perform a simple calibration routine consisting of several handshapes chosen to measure the range of movement of the user's joints. By calibrating the handshape data relative to these extremities it was hoped to improve the network's ability to generalise to the unregistered users.

Networks were trained on both the calibrated and uncalibrated data. In both cases the networks had 18 inputs, 40 hidden nodes and 30 output nodes (this will be denoted as an 18:40:30 architecture), and were trained for 1,000,000 pattern presentations with a learning rate of 0.2 The results reported in Table 1, show that although the calibration process slightly reduced performance on the registered test set, it had a larger beneficial effect on the unregistered test set, and therefore this calibration was incorporated into the final system.

**Table 1.** *Mean classification accuracy of networks trained using raw and calibrated versions of the handshape data sets*

| | Training set | Registered test set | Unreg. test set |
|---|---|---|---|
| Raw data | 97.9 | 96.6 | 87.9 |
| Calibrated data | 98.0 | 96.2 | 89.9 |

*3.2 Orientation Recognition Network*

The orientation of the hand can be described in terms of two orthogonal directions – the facing of the palm, and the direction in which the hand is pointing. If we consider only six possible directions (up, down, left, right, towards the signer, away from the signer) then there are 15 different orientations used in Auslan (in fact some signs involve directions such as 'left and up', but such small distinctions are never the sole difference between signs).

The input to the network consisted of the 3 orientation values from the Polhemus sensor, and also calibrated values for the 2 wrist sensors on the CyberGlove. The latter was required as the positioning of the Polhemus mount on the CyberGlove was above the wrist for some users, meaning that the orientation values were affected by the degree to which the wrist was flexed (early trials conducted without any wrist data performed poorly). The orientation values returned by the Polhemus were cyclical in nature (ranging from 0 to 255, and then back to 0). To avoid the problems caused by this discontinuity in the input data the network was presented with the sine and cosine of the orientation values rather than the raw values. Therefore the networks had an 8:14:15 topology.

The results of training these networks are reported in Table 2. These show that the overall accuracy is only moderate. However if these mistakes are broken down in terms of the misclassification of the component directions of the orientation, then it can be seen that the majority of errors consist of confusing adjacent directions. These mistakes are less likely to be important in distinguishing between signs than if the network were to confuse opposite directions.

**Table 2.** *Mean percentage accuracy obtained by different encodings on the hand orientation data, broken down by error of the two component directions*

| Directions | Training set | Reg. test set | Unreg. test set |
|---|---|---|---|
| Both correct | 94.8 | 90.4 | 89.1 |
| One correct, one adjacent | 2.5 | 5.0 | 5.9 |
| Both adjacent | 2.7 | 4.5 | 4.9 |
| One correct, one opposite | 0.0 | 0.0 | 0.0 |
| One adjacent, one opposite | 0.0 | 0.1 | 0.0 |
| Both opposite | 0.0 | 0.0 | 0.0 |

*3.3 Location Recognition Network*

Auslan signs can occur in three main groups of locations – neutral space (the space in front of the signer's body), primary locations (those on or near the body or head) and secondary locations (on or near the hands). In order to recognise secondary locations it is necessary to track the position of both hands, which would require a second Polhemus sensor. Therefore the SLART system considers only the 16 primary locations as well as neutral space. Any signs made using the subordinate hand as a base were regarded as being performed in neutral space.

The location and orientation features affect each other, and so the training and test data for the location network was gathered at the same time as the orientation data. For the same reason it was necessary to provide the network with the orientation and wrist flex values as inputs, in addition to the 3 Polhemus location values. Hence the networks developed had an 11:19:19 architecture. They were trained for 1,000,000 pattern presentations at a learning rate of 0.1. As with the handshape data, a calibration routine was used to measure the extremes of the input values for each user. In this case each signer made 5 gestures which measured the maximum extent of movement of their hand in each direction. Networks were trained using both the calibrated and uncalibrated data.

**Table 3.** *Mean classification accuracy of networks trained on the raw and calibrated versions of the hand location data*

| | Training set | Registered test set | Unreg. test set |
|---|---|---|---|
| Raw data | 71.7 | 67.7 | 64.5 |
| Calibrated data | 80.5 | 74.7 | 68.4 |

Two conclusions can be drawn from the results in Table 3. First the calibration routine was extremely beneficial, increasing performance on both the registered and unregistered test sets. Second, the location network achieved a much lower level of accuracy than any of the other feature-extraction networks. This is due primarily to the tracking technology used. The Polhemus measures the position of the glove relative to a fixed source, which for these experiments was placed on a wooden desk behind and to the left of the signer. Ideally the input to the system would be the position of the hand relative to the signer's body, rather than relative to the Polhemus source. In the current system any change in the positioning of the body relative to the source will affect the accuracy of the system, particularly with regards to the closely-spaced locations on the signer's head. This is one area in which a visually-based tracking system would have an advantage as it would allow more direct measurement of the hand position relative to the body.

*3.4 Motion Recognition Network*

Motion is the feature for which it is most difficult to enumerate the complete range of possible categories used within Auslan, as many signs involve 'tracing' motions which indicate the shape of an object, and hence are unique to that sign. For this research only the 13 most commonly used motions were classified, consisting of simple movement of the hand in either direction along the three primary spatial axes, back-and-forth motions along the same axes, circling motions aligned with the axes and stationary.

Motion differs from the other features in that it is inherently temporal in nature. Two approaches were taken to dealing with this aspect of the problem. The first was to use a recurrent network with 3 inputs per time frame, feeding into a layer of 30 recurrently-interconnected nodes (13 of these were output nodes, the remainder served to store the network's internal state). The input values were the difference in location from the previous time-step. This recurrent network was trained using the backpropagation-through-time algorithm with a learning rate of 0.05.

The second approach was to pre-process the input sequence to extract features for presentation to a standard feed-forward network. For each of the three location values the change at each time step was summed over the entire sequence (giving a measure of the difference between the initial and final position), as was the absolute value of the changes. The sum of the absolute value of the change in velocity at each time step was also presented to the network, as was the length of the input sequence, giving a total of 8 inputs. The network architecture used was 8:8:13, and this was trained for 750,000 pattern presentations at a learning rate of 0.05.

Table 4 compares the results obtained by the two network architectures. It can be seen that the non-recurrent network fared much better, slightly outperforming the recurrent network on the training data but giving a significant improvement in generalisation to the test sets. Therefore a non-recurrent network was used in the final system.

**Table 4.** *Mean classification accuracy of recurrent and non-recurrent networks on the hand motion data*

| | Training set | Registered test set | Unreg. test set |
|---|---|---|---|
| Recurrent net | 89.7 | 78.6 | 63.4 |
| Non-recurrent net | 93.5 | 91.6 | 75.7 |

## 4. CLASSIFICATION OF SIGNS

Once all of the feature-extraction networks had been trained, the best network for each feature was selected for inclusion in the final system (as determined by performance on the registered test set). Table 5 summarises the performance of these networks.

**Table 5.** *Summary of the performance of the best network for each feature on the training set and test set for the registered and unregistered signers*

| | Training set | Registered test set | Unreg. test set |
|---|---|---|---|
| Handshape | 98.0 | 97.4 | 89.5 |
| Orientation | 94.5 | 91.6 | 89.2 |
| Location | 80.9 | 76.4 | 69.0 |
| Motion | 93.7 | 92.3 | 76.9 |

Each signer was asked to perform 52 signs selected from Auslan to form SLARTI's initial vocabulary. Unfortunately due to age-related failure of the CyberGlove it was only possible to gather test sets from 4 of the 7 registered signers, although training sets were gathered from all 7. Test sets were also gathered from the 3 unregistered signers.

The 52 signs were randomly divided into 13 sequences of 4 signs which were performed by each signer, manually indicating the start and end of each sequence via a switch held in the non-signing hand. The signs were segmented at these points, and the input sequence was processed by the feature-extraction nets. The handshape, orientation and location features were found for both the start and end of the sequence, whilst the motion feature was extracted for the entire sequence. Hence each sign was described by a vector of 7 features which were then used to perform the final classification. A neural network was not used for this final classifier for two reasons. First the size of the resultant network (139 inputs, 52 outputs) would require an extremely large number of training examples in order to achieve a suitable level of generalisability. Second, this approach would mean retraining this large network any time that changes were made to the system vocabulary. For this reason other pattern classification techniques were preferred.

The first method used was the nearest-neighbour lookup algorithm. Four variants of this simple algorithm were used. One difference was in the nature of the examples considered by the lookup – in one version the examples from the training sets were used, whilst the second version used instead the definitions of the signs as derived from the Auslan dictionary. The second difference was in the nature of the distance measure used. In the simple distance measure (SDM) all categories of a feature were considered equidistant from each other. A heuristic distance measure (HDM) was also tested, which was derived by examination of the confusion matrices of the feature-extraction networks on the training examples. This heuristic aimed to account for the systematic errors introduced by the feature networks, by weighting these errors less heavily.

The results of these variants of the nearest neighbour lookup for each signer are reported in Table 6.

From Table 6 it can be seen that using the simple distance measure the lookup algorithm using the training examples easily outperforms that using the sign definitions. However the heuristic distance measure successfully captures the extra information present in the training examples, as it enables equal or better performance to be obtained using only the sign definitions. This is extremely useful as it allows the vocabulary to be extended without the need to gather examples of the new signs.

**Table 6.** *Classification accuracy of the nearest neighbour lookup algorithm on complete signs from each signer*

| Signer | Definitions (SDM) | Definitions (HDM) | Training set (SDM) | Training set (HDM) |
|---|---|---|---|---|
| A | 88.5 | 94.2 | 92.3 | 94.2 |
| C | 71.2 | 92.3 | 100.0 | 100.0 |
| E | 71.2 | 96.2 | 67.3 | 90.4 |
| F | 86.5 | 94.2 | 86.5 | 88.5 |
| Reg. signers (mean) | 79.4 | 94.2 | 86.5 | 93.3 |
| D | 67.3 | 82.7 | 75.0 | 86.5 |
| H | 65.4 | 88.5 | 76.9 | 75.0 |
| K | 71.2 | 84.6 | 84.6 | 82.7 |
| Unreg. signers (mean) | 68.0 | 85.3 | 78.8 | 81.4 |

The second classification algorithm trialed was the C4.5 inductive learning system developed by Quinlan (1992). C4.5 builds a decision tree on the basis of training examples, which can subsequently be pruned to obtain a smaller tree. The process of generating the decision tree is extremely fast in comparison to neural networks, meaning that creating a new decision tree every time the vocabulary was extended is a viable proposition.

**Table 7.** *Classification accuracy of the C4.5 algorithm on complete signs from each signer*

| Signer | Standard unpruned | Standard pruned | Subset unpruned | Subset pruned |
|---|---|---|---|---|
| Tree size | 649 | 397 | 140 | 133 |
| Training examples | 92.3 | 88.2 | 96.2 | 95.9 |
| A | 86.5 | 84.6 | 90.4 | 90.4 |
| C | 96.8 | 92.3 | 98.1 | 98.1 |
| E | 73.1 | 71.2 | 55.8 | 55.8 |
| F | 78.8 | 78.8 | 76.5 | 76.5 |
| Reg. signers (mean) | 83.8 | 81.7 | 80.2 | 80.2 |
| D | 63.5 | 63.5 | 65.4 | 67.3 |
| H | 63.5 | 61.5 | 65.4 | 65.4 |
| K | 71.2 | 69.2 | 78.8 | 78.8 |
| Unreg. signers (mean) | 66.1 | 64.7 | 69.9 | 70.5 |

Table 7 reports results for C4.5 using both the pruned and unpruned versions of the tree, and bot with and without the subsetting option (this option allows each node in the decision tree to incorporate multiple values of an attribute). The results obtained by C4.5 are generally below those obtained by applying the nearest neighbours lookup algorithm to the same training examples, even if only the simple distance measure is used. In particular the nearest neighbours technique generalises much better to the unregistered signers.

## 5. CONCLUSION

SLARTI is capable of classifying Auslan signs with an accuracy of around 94% on the signers used in training, and about 85% for other signers. The modular design of the system allows for future enhancement of the system both in terms of expanding its vocabulary, and in improving the recognition accuracy. The major area in which accuracy could be improved is in the classification of sign location where the performance could be improved by the addition of extra position tracking sensors.


Proc. 1st Euro. Conf. Disability, Virtual Reality & Assoc. Tech., Maidenhead, UK, 1996

Currently the hardware used is not portable enough to be used in the real-world as a communications device, but it could be applied as a teaching aid for people learning Auslan. The techniques developed are not specific to Auslan, and so the system could easily be adapted to other sign languages or for other gesture recognition systems (for example, as part of a VR interface or for robotic control).

## 6. REFERENCES


S Fels and G Hinton (1993), Glove-Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesiser, *IEEE Transactions on Neural Networks*, **4**, *1*, pp. 2-8

E Holden (1993), Current Status of the Sign Motion Understanding System, Technical Report 93/7, Department of Computer Science, University of Western Australia,

T Johnston (1989), Auslan: The Sign Language of the Australian Deaf Community, PhD thesis, Department of Linguistics, University of Sydney

J Kramer and L Leifer (1989), The Talking Glove: A Speaking Aid for Nonvocal Deaf and Deaf-Blind Individuals, *RESNA 12th Annual Conference*, New Orleans, Louisiana

J Quinlan (1992), *C4.5: Programs for Machine Learning*, Morgan Kaufmann

A Waibel, H Sawai and K Shikano (1989), Modularity and Scaling in Large Phonemic Neural Networks, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **37**,*12*