A Linguistic Feature Vector for the Visual Interpretation of Sign Language

Richard Bowden[1,2], David Windridge[1], Timor Kadir[2], Andrew Zisserman[2], and Michael Brady[2]

[1] CVSSP, School of EPS,
University of Surrey, Guildford,
Surrey, UK
{r.bowden,d.windridge}@eim.surrey.ac.uk
[2] Department of Engineering Science,
University of Oxford,
Oxford, UK.
{timork,az,jmb}@robots.ox.ac.uk

Abstract. This paper presents a novel approach to sign language recognition that provides extremely high classification rates on minimal training data. Key to this approach is a 2 stage classification procedure where an initial classification stage extracts a high level description of hand shape and motion. This high level description is based upon sign linguistics and describes actions at a conceptual level easily understood by humans. Moreover, such a description broadly generalises temporal activities naturally overcoming variability of people and environments. A second stage of classification is then used to model the temporal transitions of individual signs using a classifier bank of Markov chains combined with Independent Component Analysis. We demonstrate classification rates as high as 97.67% for a lexicon of 43 words using only single instance training outperforming previous approaches where thousands of training examples are required.

1 Introduction

Sign Language is a visual language and consists of 3 major components: fingerspelling – used to spell words letter by letter; word level sign vocabulary – used for the majority of communication; and non manual features – facial expressions and tongue, mouth and body position.

Within the literature there has been extensive work performed on fingerspelling, e.g. [1,7]; but this area is a small subset of the overall problem. For word level sign recognition, the most successful methods to date have used devices such as data-gloves and electromagnetic/optical tracking, rather than monocular image sequences, and have achieved lexical sizes as high as 250 signs or words [3,4,5,6]. However, without using such devices recognition is typically limited to around 50 words and even this has required a heavily constrained artificial grammar on the structure of the sentences [8,11].

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 390–401, 2004.

Our objective is large lexicon (word level) sign recognition from monocular image sequences. Traditionally, sign recognition has been based upon extensive training. However, this limits scalability as the acquisition of labelled training data is expensive and time consuming. Thus our aim is a method with low training requirements. Additionally, we aim for transferable learning, so that signs learnt from one individual enable signs from another individual to be recognized. In a perfect world 'one-shot-training' should be capable of addressing both of these aims, and this work presents an approach capable of achieving high recognition rates across individuals with as little as a single training instance per sign/word.

Previous approaches to word level sign recognition borrow from the area of speech recognition and rely heavily upon tools such as Hidden Markov Models (HMMs) [8,11,12] to represent temporal transitions. In turn this has meant extensive training sets have been required, for example Vogler and Metaxas [12] require 1292 training examples for a 22-word lexicon.

The novelty of the work presented here is that we structure the classification model around a linguistic definition of signed words, rather than a HMM. This enables signs to be learnt reliably from just a handful of training examples, and we have been able to reach the state of the art (49 words) using just this training set.

The classification process is divided into two stages. The first generates a description of hand shape and movement at the level of 'the hand has shape 5 (an open hand) and is over the left shoulder moving right'. This level of feature is based directly upon those used within sign linguistics to document signs. Its broad description aids in generalisation and therefore significantly reduces the requirements of further stages of classification. In the second stage, we apply Independent Component Analysis (ICA) to separate the channels of information from uncorrelated noise. Final classification uses a bank of Markov models to recognise the temporal transitions of individual words/signs.

In the following section, we describe the system architecture and its motivation. Section 3 then presents the vision elements used to track the user; in Section 4, the results of tracking are classified in terms of a linguistically inspired description of activity. Section 5 describes the second stage of classification, in which the temporal aspects of signs are both learnt and recognised. Finally, Section 6 presents results and a discussion of the techniques used and Section 7 highlights our future work.

2 Overview

A graphical overview of the system is given in Figure 1. Our approach is based upon a novel two stage classification:

Classification stage I: Raw image sequences are segmented in order to extract the shapes and trajectories of the hands in the monocular image sequence. The initial classification stage converts these into a "viseme" representation (the visual equivalent of a phoneme) taken from sign linguistics [2]:

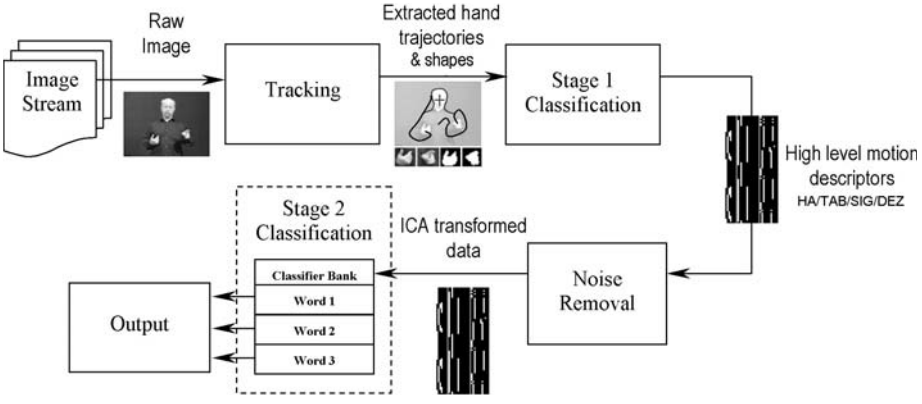


Fig. 1. Block diagram showing a high level overview of the stages of classification.

HA Position of the hands relative to each other
TAB Position of hands relative to key body locations
SIG Relative movement of the hands
DEZ The shape of the hand(s)

This HA/TAB/SIG/DEZ notation provides a high-level feature descriptor that broadly specifies events in terms such as *hands move apart, hands touch or right hand on left shoulder*. This description of scene content naturally generalises temporal events, hence reduces training requirements. This is described in more detail in Section 4.

Classification stage II: Each sign is modelled as a 1st order Markov chain in which each state in the chain represents a particular set of feature vectors (denoted *symbols* below) from the stage I classification. The Markov chain encodes temporal transitions of the signer's hands. During classification, the chain which produces the highest probability of describing the observation sequence is deemed to be the recognised word. In the training stage, these Markov chains may be learnt from a single training example.

Robust Symbol Selection: An appropriate mapping from stage I feature vectors to symbols (representing the states in the Markov chains) must be selected. *If* signs were produced by signers without any variability, or if the stage I classification was perfect, then (aside from any computational concerns) one could simply use a one-to-one mapping; that is, each unique feature vector that occurs in the course of a sign is assigned a corresponding state in the chain. However, the HA/TAB/SIG/DEZ representation we employ is binary and signers do not exhibit perfect repeatability. Minor variations over sign instances appear as perturbations in the feature vector degrading classification performance.

For example the BSL sign for 'Television', 'Computer' or 'Picture' all involve an iconic drawing of a square with both hands in front of the signer. The hands move apart (for the top of the square) and then down (for the side) etc. Ideally, a HMM could be learnt to represent the appropriate sequence of

HA/TAB/SIG/DEZ representations for these motions. However the exact position/size of the square and velocity of the hands vary between individual signers as does the context in which they are using the sign. This results in subtle variations in any feature vector however successfully it attempts to generalise the motion.

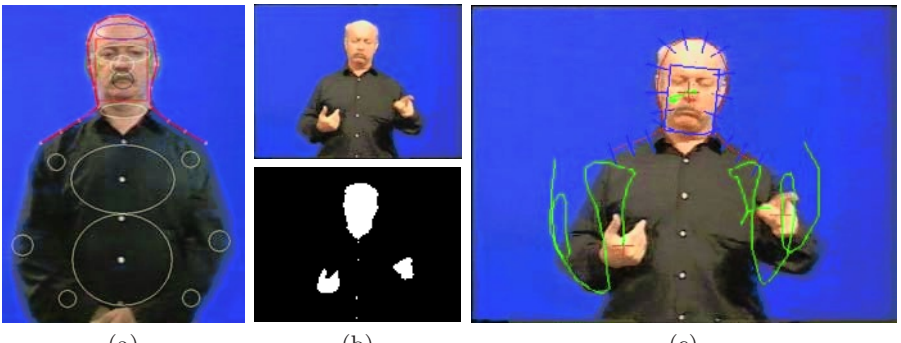
To achieve an optimal feature-to-symbol mapping we apply Independent Component Analysis (ICA). Termed feature selection, this takes advantage of the separation of correlated features and noise in an ICA transformed space and removes those dimensions that correspond to noise.

3 Visual Tracking

For completeness, we briefly describe the head and hand tracking, though we do not consider it to be the innovative part of our work.

Signers generally face the viewer as directly as possible to ease understanding and remove ambiguities and occlusions that occur at more oblique angles. The system uses a probabilistic labelling of skin to roughly locate the face of a signer. This, coupled with a contour model of the head and shoulders, provides a body-centred co-ordinate system in which to describe the position and motion of the hands. The 2D contour is a coarse approximation to the shape of the shoulders and head and consists of 18 connected points as shown in Figure 2a. The contour is a mathematical mean shape taken from a number of sample images of signers. The contour is then fitted to the image by estimating the similarity transform which minimises the contour's distance to local image features.

Estimates for key body locations, as indicated in Figure 2a, are placed relative to the location of the head contour. This means that as the contour is



(a) (b) (c)

Fig. 2. (a) The 2D contour of the head and shoulders fitted to an image with positions and approximate variances for key body locations. (b) Tracking the hands using colour: top — an original frame from the signing video sequence; bottom — the binary skin map (white pixels denote a Mahalanobis distance < 3). (c) The result of contour tracking and trajectory of the hands over time.

transformed to fit the location of the user within the video stream, so the approximate locations of the key body components are also transformed. Figure 2c shows the system tracking the upper torso and hands of an active signer, the trails from the hands show the path taken over the last 50 frames.

4 Stage I Classification

The 5 HA, 13 TAB, 10 SIG and 6 DEZ states we currently use are listed in Table 1 and are computed as follows:

HA: the positions of the hands relative to each other is derived directly from deterministic rules on the relative x and y co-ordinates of the centroids of the hands and their approximate area (in pixels).

TAB: the position of the hands is categorised in terms of their proximity to key body locations (shown in Figure 2) using the Mahalanobis distance computed from the approximate variance of those body parts.

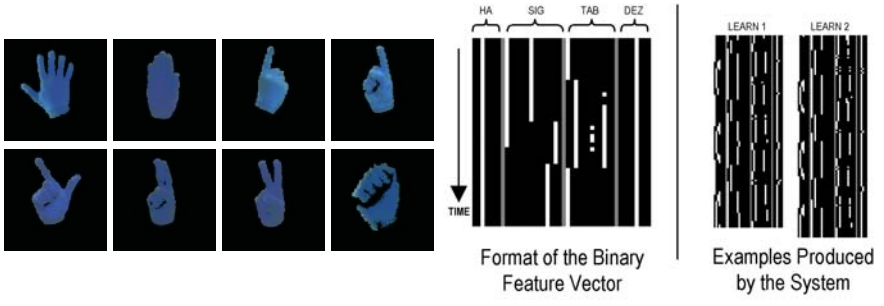
SIG: the movement of the hands is determined using the approximate size of the hand as a threshold to discard ambient movement and noise. The motion is then coarsely quantised into the 10 categories listed in Table 1.

DEZ: British Sign Language has 57 unique hand-shapes (excluding fingerspelling) which may be further organised into 22 main groups [2]. A visual exemplar approach is used to classify the hand shape into six (of the 22) groups. This is described in detail below.

Table 1. The high level features after stage I classification.

HA	TAB	SIG	DEZ
1. Right hand high	1. The neutral space	1. Hand makes no movement	1. 5
2. Left hand high	2. Face	2. Hand moves up	2. A
3. Hands side by side	3. Left Side of face	3. Hand moves down	3. B
4. Hands are in contact	4. Right Side of face	6. Hand moves left	4. G
5. Hands are crossed	5. Chin	7. Hand moves right	5. H
	6. R Shoulder	8. Hands moves apart	6. V
	7. L Shoulder	9. Hands move together	
	8. Chest	10. Hands move in unison	
	9. Stomach		
	10. Right Hip		
	11. Left Hip		
	12. Right elbow		
	13. Left elbow		

Figure 3b shows the features generated by the system over time. The horizontal binary vector shows HA, SIG, TAB and DEZ in that order delineated by grey bands. The consistency in features produced can clearly be seen between examples of the same word. It is also possible to decode the vectors back into a textual description of the sign in the same way one would with a dictionary. The feature vector naturally generalises the motion without loss in descriptive ability.



(a) (b)

Fig. 3. (a) Examples of hand-shapes used in British Sign Language – from top-left clockwise '5', 'B', 'G', 'G', 'A', 'V', 'H', 'G'. (b) Graphical representation of feature vectors for different occurrences of signs demonstrating the consistency of feature description produced and two instances of the sign 'learn'.

Linguistic evidence [10] suggests that sign recognition is based primarily on the dominant hand (which conveys the majority of information). For this reason, we currently discard the non dominant hand and concatenate the HA, TAB, SIG and DEZ features to produce a 34 dimensional binary vector which describes the shape and motion in a single frame of video.

4.1 Hand Shape Classification

Within the lexicon of 49 words used to date in this work, six main sign groups appear, denoted '5', 'A', 'B', 'G', 'H' and 'V' following the definition in [2]. Typical examples of these are shown in Figure 3. Our objective is to classify hand-shapes into one of these six groups.

The visual appearance of a hand is a function of several factors which a hand shape classifier must take into account. These include: pose, lighting, occlusions and intra/inter-signer variations. To deal with such variations, we adopt an exemplar based approach, where many visual exemplars correspond to the same sign group.

Segmentations of the hands are first obtained from the tracking stage discussed in Section 3. Hand-shape is represented as a binary mask corresponding to the silhouette of the hand and these masks are normalised for scale and orientation using their first and second moments. Learning proceeds by generating a set of normalised masks from training data (see Section 6) and clustering these to form an exemplar set. We use a greedy clusterer with a normalised correlation distance metric. A threshold on this distance controls the degree of grouping.

Novel hand-shapes are then classified by matching their normalised binary masks to the nearest one in the exemplar set using normalised correlation as a distance metric. Similar approaches have been used by several previous authors, for example [7,9].

This exemplar approach has a number of attractive properties. Different hand-shapes in the same group, and variations in appearance of the same hand-

shape due to different poses or signer variation, may be represented by separate exemplars assigned to the same group label.

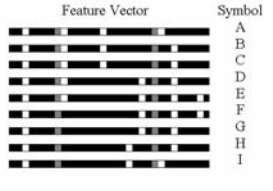
While it is clear that this basic hand classifier cannot distinguish between all poses and shapes, we demonstrate that it complements the HA, TAB and SIG features, hence is *sufficient* to discriminate between a fixed lexicon of 49 words. For much larger lexicons the representation may be augmented in various ways, for example through the use of internal features to capture information about the positions of the fingers in closed hand-shapes, or by allowing discrimination within hand-shape groups. Set at an operating point chosen to give 1770 exemplars, the hand-shape classifier achieves an average correct classification rate of 75%. The results presented in Section 6 use this operating point.

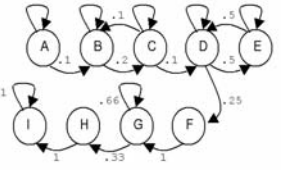
5 Stage II Classification

5.1 Training

In order to represent the temporal transitions which are indicative of a sign, we make a 1st order assumption and construct a 1st order Markov chain for each word in the lexicon. However, this assumption requires that an ergodic model be constructed. With a 34 dimensional binary feature vector, this would result in a chain with $2^{28} \times 6$ states (5 HA + 13 TAB + 10 SIG multiplied by the 6 mutually exclusive DEZ features) and over 2.6×10^{17} possible transitions requiring a prohibitive amount of a storage. However, as can be seen in Figure 3b, the number of transitions in each word is typically small and governed by the duration of the sign and the capture rate (in this case 25Hz).

It is also evident that out of the $2^{28} \times 6$ possible states only a small subset ever occur and that there are even fewer transitions than combinatorially possible, due to the physical limitations of the human body. Therefore, we build only as much of the ergodic model as is needed. This is done by adding new feature states to the state transition matrix as they occur during training. The result is a sparse state transition matrix, $P_w(s_t|s_{t-1})$, for each word w giving a classification bank of Markov chains.





(a)

	A	B	C	D	E	F	G	H	I	J
A	.9	.1	0	0	0	0	0	0	0	. . .
B	0	.8	.2	0	0	0	0	0	0	. . .
C	0	.1	.8	.1	0	0	0	0	0	. . .
D	0	0	0	.25	.5	.25	0	0	0	. . .
E	0	0	0	.5	.5	0	0	0	0	. . .
F	0	0	0	0	0	1	0	0	0	. . .
G	0	0	0	0	0	0	.66	.33	0	. . .
H	0	0	0	0	0	0	0	0	1	. . .
I	0	0	0	0	0	0	0	0	1	. . .
J	.	.	.	.	.	.	.	.	.	. . .

(b)

Fig. 4. The process of classifier chain construction ($P_w(s_t|s_{t-1})$): (a) first, a look up table is built mapping binary feature vectors to symbols and, (b) then a sparse ergodic state transition matrix representing the Markov chain can be generated.

5.2 Feature to Symbol Mapping

Mapping feature vectors directly onto states in the Markov chain as described above results in unsatisfactory classification performance. Minor variations across instances of the same sign and the rule-based stage I classification produces perturbations, or noise, in the binary feature vector resulting in misclassification.

One way to mitigate these effects is to use clustering to group 'similar' events together. In such a scheme, the mapping is no longer one-to-one; sets of stage I feature vectors map onto single symbols used for the Markov chain. However, in our binary feature space the concept of similarity, or distance, cannot be assumed to be Euclidean. Instead, we should *learn* this distance metric from the training data.

Our approach is to apply Independent Component Analysis (ICA) to the training data, the result of which is a transformation matrix, which, when applied, transforms the binary feature space into a space where an Euclidean metric can be used. Intuitively, the effect of the ICA is to move features that co-occur in the training data closer together and conversely those that are independent, further apart.

ICA attempts to separate the correlated features from uncorrelated noise; each incoming feature vector is transformed into the ICA space and an exhaustive feature selection process performed to decide which of the ICA transformed features are important to classification and which constitute noise. At each step a single feature, i.e. a single component of the transformed vector, is discarded. The feature to be removed is selected such that the overall performance in terms of classification of the reduced feature vector is maximised.

Once an ICA transformation matrix has been learnt, a look-up table (LUT) is generated from the training data to map the ICA transformed features to symbols for use in Markov chains.

5.3 Classification

During classification, the model bank is applied to incoming data in a fashion similar to HMMs. A sliding temporal window T is used to check each of the classifiers in turn against the incoming feature vectors. The objective is to calculate that chain which best describes the incoming data i.e. has the highest probability that it produced the observation sequence s. The probability of a model matching the observation sequence is calculated as $P(w|s) = \pi \prod_{t=1}^{l} P_w(s_t|s_{t-1})$, where $1 < l < T$ and $P_w(s_t|s_{t-1}) > 0, \forall t$, π is the prior probability of a chain starting in any one of its states. However, setting $\pi = 1$ and the stopping criteria $P_w(s_t|s_{t-1}) > 0, \forall t$, provides some robustness to the problems of sign co-articulation.

Of course for continuous recognition, this introduces a bias for short signs as $P(w|s)$ decreases as l increases. It is therefore necessary to use both the model match probability $P(w|s)$ and the number of observations the model describes, l, to find the sequence of words that best explain the observations over time. This

is done by maximising the overall probability using a Viterbi algorithm. The use of the Viterbi algorithm adds the problem that our approach has no model for silence. To overcome this, a running average of model match probability is used to synthesis a silence model within the Viterbi decoder. For isolated word recognition the chain which describes the largest consecutive segment of the observation sequence (l) can be used which overcomes the problems of word length bias.

6 Performance Results

The training and test data consists of 49 randomly selected words (listed in the appendix). Unlike previous approaches, signs were not selected to be visually distinct but merely to represent a suitable cross section of signs to allow a short conversation to take place.

Video sequences were collated for a single person performing the 49 signs. Each sign was repeated numerous times within the dataset resulting in a total of 249 individual signs, averaging 5 repetitions for each sign. The video was hand labelled for ground truth. A single instance of each sign was selected for training and the remaining 200 signs retained as an unseen test set. A classifier bank was learnt consisting of 49 Markov chains, one for each individual sign. The unseen data was then presented to the classifier bank and the most probable word determined using the approach described in Section 5. The output of the classifier bank was then compared to the ground truth to determine if a successful classification had been made. The results are presented in Figure 5.

Figure 5 shows the classification performance as a function of the number of features selected by ICA feature selection. The ICA transformed data results in a performance boost from 73% up to 84% percent classification rate beyond the 6 feature mark.

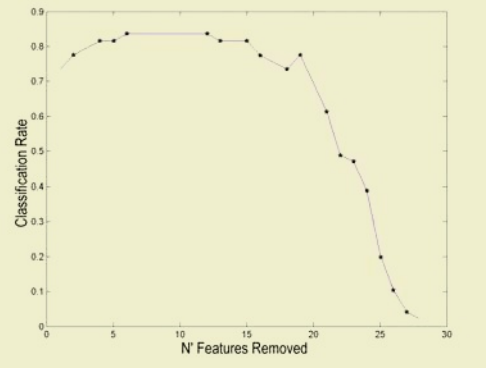


Fig. 5. Classification performance as a function of number of features removed in the ICA feature selection.

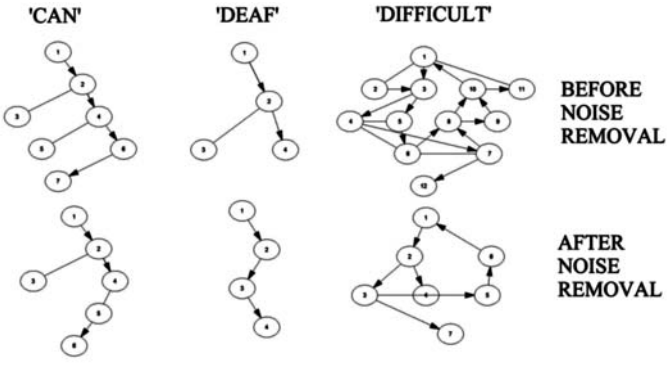


Fig. 6. Chains representing temporal transitions for both raw and ICA transformed motion descriptors for the words 'can', 'deaf' and 'difficult'.

By selecting a lexical subset which does not contain ambiguous signs, for example by removing those signs for which facial gesture or contextual grammar are a significant cue, the results improve to 97.67% classification rate on a lexicon of 43 words. This compares well with previous approaches. It is important to note that these high classification results have been obtained without constraints on grammar and furthermore with only a single training instance per sign. This can be compared to other viseme level approaches based upon HMM's where thousands of training examples are required to achieve similar levels of accuracy [8,11,12].

Figure 6 shows the results of ICA feature selection upon the transitions made by the feature vector. It can clearly be seen that the ICA transformed data produces more compact models. This is particularly evident for the word 'difficult' where the complex transitions between 12 states has been simplified through the removal of noise down to 7 states. More interesting to note is that the word 'deaf' is simplified to a left-right transition model almost identical to that which would be assumed for a HMM approach, however, many other signs do not fit this simplistic assumption. For example, the sign 'difficult' repeats 2 or more times as the thumb of the right hand taps against the left open palm. The resulting chain, shown on the right in Figure 6, is clearly cyclic with 2 possible end states. Such a chain intuitively fits with this type of repeating sign. It is not clear how such signs can be forced to map to a left-right transition model without some loss of information. This serves to illustrate that the assumptions on which speech recognition are based do not naturally extend to sign.

7 Conclusions

Our current demonstrator runs at 25fps on a 1GHz laptop and is capable of operating in an unconstrained indoor environment with background clutter. Due to the generalisation of features, and therefore the simplification in training,

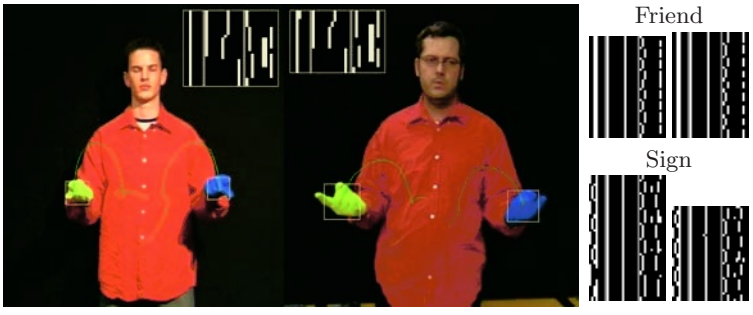


Fig. 7. Generalisation of feature vectors across different individuals: (a) two people signing the sign 'different'; (b) binary feature vectors for two people for the signs 'friend' and 'sign'. The signs are correctly classified.

chains can be trained with new signs on the fly with immediate classification. This is something that would be difficult to achieve with traditional HMM approaches. However, the real power of this approach lies in its ability to produce high classification results on 'one shot' training and can demonstrate real time training on one individual with successful classification performed on a different individual performing the same signs. Figure 7 shows the word 'different' being performed by two different people along with the binary feature vector produced. The similarity is clear, and the signed words are correctly classified.

Acknowledgements. This work is funded by EC Project CogViSys.

References

1. R. Bowden and M. Sarhadi. A non-linear model of shape and motion for tracking finger spelt american sign language. *Image and Vision Computing*, 20(9-10):597–607, 2002.
2. D. B. (ed). *Dictionary of British Sign Language*. British Deaf Association UK, Faber and Faber, ISBN: 0571143466, 1992.
3. S. S. Fels and G. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesiser. *IEEE Trans. on Neural Networks*, 4(1):2–8, 1993.
4. M. W. Kadous. Machine recognition of auslan signs using powergloves: towards large lexicon recognition of sign language. In *Proc. Workshop on the Integration of Gesture in Language and Speech*, pages 165–174, 1996.
5. J. Kim, W. Jang, and Z. Bien. A dynamic gesture recognition system for the korean sign language (ksl). *IEEE Trans. Systems, Man and Cybernetics*, 26(2):354–359, 1996.
6. R. Liang and M. Ouhyoung. A real time continuous gesture recognition system for sign language. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 558–565, 1998.
7. R. Lockton and A. W. Fitzgibbon. Real-time gesture recognition using deterministic boosting. In *Proc. British Machine Vision Conf.*, 2002.

8. T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 189–194, 1995.
9. B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *Proc. Intl. Conf. on Computer Vision*, volume II, pages 1063–1070, 2003.
10. R. Sutton-Spence and B. Woll. *The Linguistics of British Sign Language, An Introduction*. Cambridge University Press, 1999.
11. C. Vogler and D. Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Proc. Intl. Conf. on Computer Vision*, pages 363–369, 1998.
12. C. Vogler and D. Metaxas. Towards scalability in asl recognition: Breaking down signs into phonemes. In *Gesture Workshop*, pages 17–99, 1999.

Appendix: Visual Lexicon Data

Listed here are the 49 signs used in the experiments reported in this paper:

‘I_me’, ‘america’, ‘and’, ‘baby’, ‘because’, ‘british’, ‘but’, ‘can’, ‘computer’, ‘deaf’, ‘different’, ‘difficult’, ‘easy’, ‘english’, ‘exam_change’, ‘fast’, ‘fingerspelling’, ‘have’, ‘hello’, ‘know’, ‘language’, ‘last’, ‘learn’, ’level’, ‘many’, ‘meet’, ‘name’, ‘nice’, ‘people’, ‘recognise’ ‘research’, ‘rich’, ‘same’, ‘say’, ‘sign’, ‘start’, ‘summer’, ‘teach’, ‘this’ ‘to’, ‘translate’, ‘try’, ‘understand’, ‘want’, ‘we’, ‘what’, ‘wife’, ‘with’, ‘yes’.