Large Lexicon Detection of Sign Language

Helen Cooper and Richard Bowden

CVSSP, SEPS, University of Surrey, Guildford, UK
{H.M.Cooper,R.Bowden}@Surrey.ac.uk

Abstract. This paper presents an approach to large lexicon sign recognition that does not require tracking. This overcomes the issues of how to accurately track the hands through self occlusion in unconstrained video, instead opting to take a detection strategy, where patterns of motion are identified. It is demonstrated that detection can be achieved with only minor loss of accuracy compared to a perfectly tracked sequence using coloured gloves. The approach uses two levels of classification. In the first, a set of viseme classifiers detects the presence of sub-Sign units of activity. The second level then assembles visemes into word level Sign using Markov chains. The system is able to cope with a large lexicon and is more expandable than traditional word level approaches. Using as few as 5 training examples the proposed system has classification rates as high as 74.3% on a randomly selected 164 sign vocabulary performing at a comparable level to other tracking based systems.

1 Introduction

The objective of this research is to produce a non-tracking/detection based system for recognising sign language. Sign Language, being as complex as any spoken language, has many thousands of signs each differing from the next by minor changes in hand motion, shape or position. Its grammar includes the modification of signs to indicate an adverb modifying a verb and the concept of placement where objects or people are given a spatial position and then referred to later. This, coupled with the intra-signer differences make true Sign Language Recognition (SLR) an intricate challenge.

Most state of the art approaches track the hands and then classify the path that they take. This causes difficulties as the hands move quickly in sign (introducing blur and interlacing effects), have high degrees of freedom (and therefore vary in appearance) and often occlude each other. In addition, tracking often employs skin tone which means that the face and hands can be easily confused and the clothing worn by the signer must be of a contrasting colour and have sleeves which cover the arms. All of these issues are limiting factors to the success of tracking approaches. Furthermore, by combining the output of tracking with further sign classification, there are two systems which can fail, reducing overall performance. To date, relatively little work has been done on using detection for gestures or actions [1][2][3][4] and it has been limited to extremely small lexicons of around 5-10 classes. To the authors knowledge, no work to-date has

M. Lew et al. (Eds.): HCI 2007, LNCS 4796, pp. 88–97, 2007.

addressed the scalability needed for a detection approach to tackle large lexicon recognition in sign. To allow direct comparison of our work with a tracking based approach, the dataset of Kadir et al[5] is used. The proposed detection approach can tackle large lexicon sign recognition with only a small drop in performance when compared to perfectly tracked data.

2 Background

Many of the solutions to SLR that have achieved large lexicon recognition use data gloves to acquire an accurate 3D position and trajectory of the hands [6] which, while facilitating a large vocabulary are cumbersome to the user. The majority of vision approaches are tracking based solutions with relatively small lexicons. Staner and Pentland [7] used colour to segment the hands for ease of tracking and reported classification results on a 40 sign lexicon. More recently, scalability has been addressed by turning to sign linguistics to aid classification. Vogler and Metaxas' [8] initial work operated on a lexicon of 53 signs but later reported a scalable solution using parallel HMMs on both hand shape and motion to recognise a 22 sign lexicon. Kadir et al [5] took this further by combining head, hand and torso position as well as hand shape to create a system that could be trained on five or fewer examples on a large lexicon of 164 signs. It is this work that we will make a direct comparison with as the dataset is available and allows our detection approach to be compared with the results of tracking.

Detection/non-tracking based approaches have recently begun to emerge, Zahedi et al [1] apply skin segmentation combined with 5 types of differencing to each frame in a sequence which are then down sampled to get features. Wong and Cippola [2] use PCA on motion gradient images of a sequence to obtain their features. Blank et al used space-time correlation to identify activity [3] while YanKe [4] employed boosted volumetric features in space-time to detect behaviour. All of these approaches are designed for gesture or behaviour recognition and typically only address a small number of gestures (<10). It is not obvious how these approaches could be extended to larger lexicons in a scalable way.

3 Methodology

Sign language can be broken down into visemes in much the same way that speech can be broken down into phonemes. These visemes can be separated into 5 main categories [9] based on hand; shape(s) (*dez*), placement (*tab*), movement (*sig*), orientation(s) (*ori*) and arrangement (*ha*). This work concentrates on the *tab*, *sig* and *ha* visemes shown in table 1.

Figure 1 shows an overview of the approach. Signs are recognised by a two stage process. In the second stage a high level classifier bank made up of 1st order Markov chains recognises the temporal order of visemes as they are produced. The visemes are *detected* by three different types of viseme level classifiers.

For *tab* visemes there needs to be correlation between where the motion is happening and where the person is; to this end spatial grid features centred

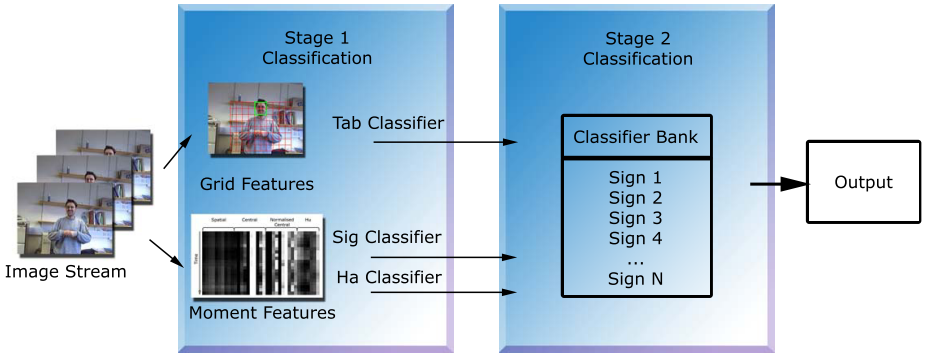

Fig. 1. Block diagram showing a high level overview of the stages of classification

around the face of the signer are used. For *sig* visemes we are interested in what type of motion is occurring, often regardless of its position, orientation or size, this is approached by extracting moment features and using local binary pattern (LBP) and additive classifiers based on their changes over time. *Ha* visemes look at where the hands are in relation to each other so these are only relevant for bi-manual signs, this is done using the same moment features as for *sig* but this time over a single frame since there is no temporal context required.

All of these viseme level classifiers are learnt using boosting which provides a way of building a strong classifier that performs well through a simple selection process. An iterative algorithm, boosting first selects the best weak classifier from a set compiled of all available features (each with an optimum response threshold). It then applies a weighting to each training example. Reducing the weighting of examples classified in the last pass and increasing the weighting of those not classified boosts the importance of examples which prove challenging to classify. This encourages the next iteration to concentrate more on the challenging examples with the heaviest weightings. More specifically in this paper AdaBoost [10] is used.

The next section discusses in detail the approaches to spatial and moment based feature extraction along with the classifiers applied to them. Section 5 then discusses how signs are recognised from the detected viseme sequences and section 6 provides comparative results. Finally conclusions are drawn.

4 Stage 1 - Viseme Detection

4.1 Skin Segmentation

In order to perform viseme detection the video is first preprocessed to find candidate hand regions. This is done by first finding the face of the user using the Viola Jones Face detector [10] included in the OpenCV library [11]. From this face region a Gaussian skin colour model is learnt using the process outlined

Table 1. The viseme level classifiers that are built

Tab	***Sig* (Both Hands)**	***Sig* (Right Hand Only)**	***Ha***
Upper Face	Apart	Circle (Type 4)	Left Higher
Nose	Together	Up	Right Higher
Ear	Together (Bend Wrist)	Up & Left	Side by Side
Eyes	Circle (Type 1)	Up & Away	Interlinking
Whole Face	Circle (Type 2)	left	Contacting
Cheek	Alt Circle (Type 3)	Left & Down	Right Near
Mouth & Lips	Up	Right	Left Near
Lower Face/Chin	Right	Right & Away	
Under Chin	Wiggle	Wiggle	
Front of Shoulders	Alt Away & Towards	Palm Down	
On Right Shoulder	Up & Down	Away & Towards	
Chest	Alt Up & Down	Away	
Right of Chest	Tap	Away & Down	
Left of Chest	Down	Spiral Away	
Upper Arm		Towards	
Lower Arm		Towards & Up	
Neutral Space		Down	
		Down & Away	
		Away & Towards(Twist Wrist)	
		Tap	
		Side to Side	

in [12]. Then the background is modelled using a normalised histogram (PDF). A threshold applied to the likelihood ratio of *face* to *background* for each pixel gives a binary, skin segmented frame. Morphological opening is used to clean up any noise and the result is shown in figure 2. Although this provides candidate hand regions it also segments the face, however, as this is consistent across both negative and positive training examples the viseme detectors will ignore its presence. Likewise, any noise in the segmented image can also be ignored as it will be inconsistent across positive training examples.

4.2 *Tab* - Spatial Features

In order that the motion can be localised in relation to the signer, a grid is applied to the image dependant upon the position and scale of the face detection. Each rectangle is a quarter of the face size and the grid is 10 rectangles wide by 8 deep, as shown in figure 3 (a). The skin segmented frame is then quantised into this grid and a rectangle is considered to be firing if over 50% of its pixels are made up of skin. For each of the *tab* visemes a classifier can then be built via boosting to show which rectangles fire for that particular viseme, examples of these classifiers are shown in figure 3 (b).

4.3 *Sig* and *Ha* - Moment Feature Vectors

There are several different types of moments which can be calculated over a segmented image, each of them displaying different properties. Four of the basic

Fig. 2. Skin segmented frame showing hands and face

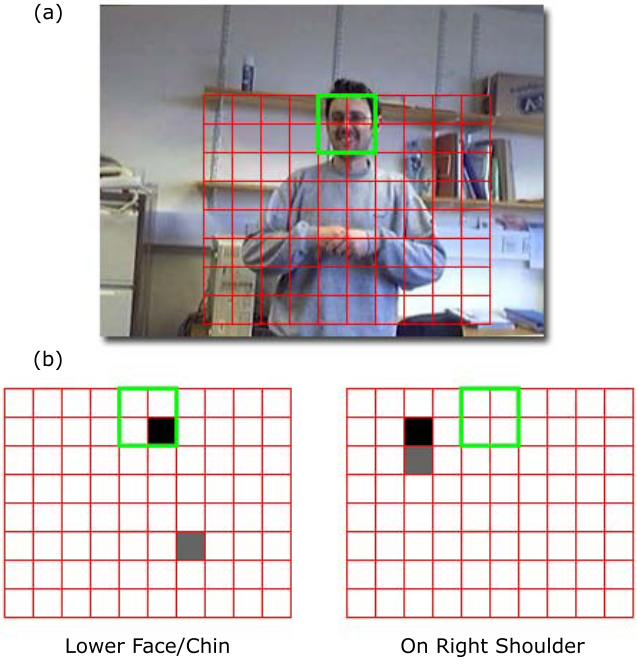

Fig. 3. (a) An example of the grid produced from the face dimensions. and (b) Grid features chosen by boosting for two of the 17 *tab* visemes. The thick central box shows the face location and the first and second chosen features are shown in black and grey respectively.

types were chosen to form a feature vector: spatial, central, normalised central and the Hu set of invariant moments. The central moments are invariant to position, the normalised central moments invariant to size and position and

the Hu moments offer rotational and skew invariance. Taking up to the 3rd order from each of these types gives a vector of 37 different parameters with a wide range of different properties. Since spatial moments are not invariant to translation and scale there needs to be a common point of origin and similar scale across examples. To this end, the spatial moments are treated in a similar way to the spatial features in 4.2 by centring and scaling about the face of the signer. For training *Ha*, this vector is used to boost a set of thresholds for individual moments, but for *Sig*, temporal information needs to be included. So the video clips are described by a stack of these vectors, like a series of 2D arrays, as shown in figure 4 (a) and temporal features employed (see next section).

4.4 *Sig* - Local Binary Patterns and Additive Classifiers

Boosting chooses from two different types of classifiers which act upon the 2D feature array; local binary patterns (LBPs) and additive classifiers. LBPs work on the gradient of a feature over time, they vary in size from 2 bits to 5 bits and there are therefore 60 different classifier patterns ($2^2 + 2^3 + 2^4 + 2^5$). We run the LBPs parallel with the time axis so that they are always operating on one type of value. In essence, the LBPs encode whether a moment is increasing or decreasing with time. For an LBP to return a 1 every gradient must match its corresponding value in the patten, 1 for an increase or 0 for a decrease or no change as can be seen in figure 4 (b).

The additive classifiers sum the values across a single moment type for a given number of frames, they can be as small as a single value or as large as the maximum classifier size allowed (tests were run of classifiers up to 26 frames long). They therefore contain information about the magnitude of values across a given viseme which complements the LBPs gradient data.

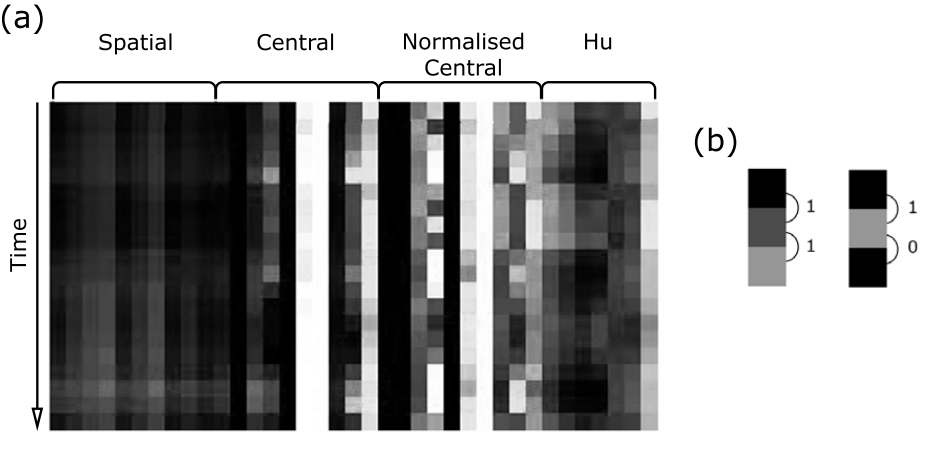

Fig. 4. (a) A pictorial description of moment vectors (normalised along each moment type for a selection of examples). (b) Local Binary Patterns, an increase in gradient is depicted by a 1 and a decrease or no change by a 0.

5 Stage II - Word Level Learning

The boosted viseme classifiers are combined to create a binary feature vector which is fed into a second stage classifier similar to that used in Kadir et al's work [5]. In order to represent the temporal transitions which are indicative of a sign, a 1st order assumption is made and a Markov chain is constructed for each word in the lexicon. An ergodic model is used and a Look Up Table (LUT) used to maintain as little of the chain as is required. Code entries not contained within the LUT are assigned a nominal probability. This is done to avoid otherwise correct chains being assigned zero probabilities. 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.

During classification, the model bank is applied to incoming data in a similar fashion to HMMs. The objective is to calculate the chain which best describes the incoming data i.e. has the highest probability that it produced the observation sequence s. Symbols are found in the symbol LUT using an L1 distance on the binary vectors. The probability of a model matching the observation sequence is calculated as $P(w|s) = \upsilon \prod_{t=1}^{l} P_w(s_t|s_{t-1})$, where l is the length of the word in the test sequence and υ is the prior probability of a chain starting in any one of its states, as in [5] this is set to $\upsilon = 1$.

6 Experimental results

6.1 Data Set

The data set used is the same 164 sign set as used by Kadir et al [5] and therefore a direct comparison can be made between their tracking based system and this detection based approach. The data set consists of 1640 examples (10 of each sign). Signs were chosen randomly rather than picking specific signs which are known to be easy to separate. The viseme classifiers are built using only data from the first 4 of the 10 repetitions of each sign and the word level classifier is then trained on up to 5 examples (including the 4 previously seen) leaving 5 completely unseen examples for testing purposes. Furthermore, only visemes from the first 91 signs are used in the viseme detector learning.

6.2 Stage 1 Classification Results

Since the time taken for a viseme differs between *Sig* types, several different length classifiers were boosted starting at 6 frames long, increasing in steps of 2 and finishing at 26 frames long. Training classification results were then found for each viseme and the best length chosen to create a final set of classifiers of various lengths as shown in table 2. As can be seen there is a large disparity between optimum classifier lengths for different visemes, while short motions like the *wiggle* visemes result in short classifiers, only 6 or 10 frames, others like *together* and *spiral away* benefit from a longer classifier.

A breakdown of viseme classifiers combined with the second stage classifier is shown in table 3. Unfortunately this data is not available for comparison in the Kadir et al paper [5]. As can be seen, each of the first stage classifiers can achieve around 30% accuracy when combined individually with the second stage classifier. This is relatively poor performance as it is not possible to distinguish 164 signs on something as simple as hands moving appart, however, through their combination impressive results can be achieved as will be seen in the next section.

Table 2. Classifier lengths for a given viseme

***Sig* (Both Hands)**	**length**	***Sig* (Right Hand Only)**	**length**
Apart	10	Circle (Type 4)	24
Together	26	Up	8
Together (Bend Wrist)	20	Up & Left	20
Circle (Type 1)	18	Up & Away	18
Circle (Type 2)	6	left	26
Alt - Circle (Type 3)	6	Left & Down	20
Up	12	Right	6
Right	22	Right & Away	12
Wiggle	6	Wiggle	10
Alt - Away & Towards	14	Palm Down	6
Up & Down	26	Away & Towards	24
Alt - Up & Down	22	Away	14
Tap	16	Away & Down	22
Down	14	Spiral Away	22
		Towards	14
		Towards & Up	12
		Down	26
		Down & Away	24
		Towards & Away (Twist Wrist)	18
		Tap	8
		Side to Side	6

Table 3. Classification performance using *Ha Tab Sig* classifiers individually with Stage 2 Classification trained on 5 examples

Stage 1 Classifier	***Ha***	***Tab***	***Sig***
Mean	33.2	31.7	29.4
Minimum	31.6	30.7	28.2
Maximum	35.0	32.2	30.5
Std. Deviation	0.9	0.4	0.6

6.3 Stage 2 Classification Results

Tests were performed on the 5 unseen examples of each of the 164 signs using a random selection of training 1 to 5 examples. The results from these runs are shown in table 4 along with the results from Kadir et al [5]. As can be seen, the detection based method is only 6.6% less accurate than the tracking used in their paper for 5 training examples.

Since the grid used for *tab* classification can produce a binary feature vector of 80 values it was tried in place of the 17 *tab* classifiers (see table 5), while it offered a minor increase when training on 5 examples it was less able to generalise with fewer training examples and consistently performed worse. In addition, this increases the size of the combined viseme vector to 122 in place of 59, more than doubling it which drastically increases the possible states and transitions in the second stage classifier.

Table 4. Classification performance compared with Kadir et al [5] trained on 5 examples using *Ha, Tab, Sig* classifiers together with Stage 2 Classification

No. Training Examples	**1**	**2**	**3**	**4**	**5**	**Kadir et al [5]**
Mean	35.5	50.2	58.6	64.6	**72.6**	**79.2**
Minimum	35.1	49.5	57.6	63.2	**68.7**	**76.1**
Maximum	35.7	50.7	59.1	65.6	**74.3**	**82.4**
Std. Deviation	0.2	0.4	0.4	0.7	**1.5**	**2.1**

Table 5. Classification performance trained on 5 examples using *Ha, Sig* classifiers and using the vector output from the grid in place of the trained *tab* classifiers together with Stage 2 Classification

No. Training Examples	**1**	**2**	**3**	**4**	**5**
Mean	31.7	44.0	54.7	63.7	**74.3**
Minimum	31.0	42.1	53.3	61.8	**69.8**
Maximum	32.2	44.8	55.5	64.6	**77.2**
Std. Deviation	0.3	0.8	0.7	0.8	**2.2**

7 Conclusions

This paper has shown that near equivalence with tracking can be achieved using solely detection in sign language recognition. This has also been done over a large lexicon database with few training examples. It demonstrates the power of combining viseme level classifiers to create word level classifiers in order to reduce the complexity as the vocabulary of the system increases. Kadir et al [5] noted a 10% increase when a *dez* classifier was included so a logical extention of this work would be to include a non-tracking based classifier for hand shapes/orientations which should afford a similar boost to the stated results.

References

1. Zahedi, M., Keysers, D., Ney, H.: Appearance-based recognition of words in american sign language. In: Second Iberian Conference in Pattern Recognition and Image Analysis, vol. 1, pp. 511–519 (June 2005)
2. Wong, S.F., Cipolla, R.: Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images. In: Proceedings of the British Machine Vision Conference, Oxford, UK, vol. 1, pp. 379–388 (September 2005)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE International Conference on Computer Vision(ICCV), Beijing, China (October 2005)
4. Ke, Y., Sukthankar, R., Hebert, M.: Efficient Visual Event Detection Using Volumetric Features. In: International Conference on Computer Vision (2005)
5. Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: Proceedings of the British Machine Vision Conference, vol. 2, pp. 939–948 (2004)
6. Fang, G., Gao, W., Ma, J.: Signer-independent sign language recognition based on sofm/hmm. In: RATFG-RTS 2001. Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Washington, DC, USA, p. 90. IEEE Computer Society, Los Alamitos (2001)
7. Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: ISCV 1995. Proceedings of the International Symposium on Computer Vision, Washington, DC, USA, p. 265 (1995)
8. Vogler, C., Metaxas, D.N.: Handshapes and movements: Multiple-channel american sign language recognition. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 299–308. Springer, Heidelberg (2004)
9. British-Deaf-Association: Dictionary of British Sign Language/English. Faber and Faber (1992)
10. Viola, P., Jones, M.: Robust Real-time Object Detection. Second International Workshop on Statistical and Computational Theories Of Vision Modelling, Learning, Computing, and Sampling (2001)
11. OpenCV-User-Group: OpenCV Library (2007), `http://opencvlibrary.sourceforge.net`
12. Micilotta, A.: Bowden, R.: View-based location and tracking of body parts for visual interaction. In: British Machine Vision Conference (BMVC), Kingston, UK (September 2004)