CogVisSys

Vision based Interpretation of Natural Sign Languages

Richard Bowden[12], Andrew Zisserman[2], Dave Windridge[1], Timor Kadir[2], Mike Brady[2]
[1] CVSSP, School of EPS, University of Surrey, Guildford, Surrey, GU2 7XH, UK
r.bowden@eim.surrey.ac.uk
http://www.ee.surrey.ac.uk/Personal/R.Bowden
[2] Dept. Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK.

Abstract. This poster outlines our current demonstration system for translating visual Sign to text. The system is based around a broad description of scene activity that naturally generalizes, reducing training requirements and allowing the knowledge base to be explicitly stated. This allows the same system to be used for different sign languages requiring only a change of the knowledge base.

Introduction

Sign Language is a visual language and consists of 3 major components:

1) Fingerspelling - used to spell words letter by letter
2) Word level sign vocabulary - used for the majority of communication
3) Non manual features - Facial expressions, tongue/mouth/body position.

Within the literature the majority of work has been in area 1, which is a small subset of the overall problem and to a lesser extent area 2. Typically, this is within a constrained problem domain of a limited lexicon (<50 words) and a heavily constrained artificial grammar.

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) and dynamic programming. The HMM relies on the assumption that within a complex signal there is a simpler underlying process which can describe the event. These underlying/hidden processes cannot be observed directly and are therefore learnt in an optimal fashion. However, to produce accurate results that generalise well, extremely large training sets are required. Lexicon size is limited for this reason, as training requirements grow exponentially with the number of words.

Feature Description

Our current system is based around describing the visemes of sign in a manner similar to that used in sign dictionaries. Using a HA/TAB/SIG notation:

HA – Hand Arrangement, describes the position relative to each other
TAB – Position of hands relative to key body locations
SIG – Relative movement of the hands

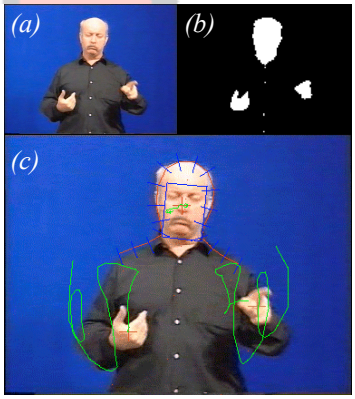
This provides a high-level feature descriptor that specifies temporal events in broad terms such as *hands move apart*, *hands touch* or *right hand on left shoulder*. This broad description of scene content naturally generalises temporal events and hence reduces training requirements.

HA	TAB	SIG
1. Right hand high	1. The neutral space	1. Hand makes no movement
2. Left hand high	2. Face	2. Hand moves up
3. Hands side by side	3. Left Side of face	3. Hand moves down
4. Hands are in contact	4. Right Side of face	6. Hand moves left
5. Hands are crossed	5. Chin	7. Hand moves right
	6. R Shoulder	8. Hands moves apart
	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	

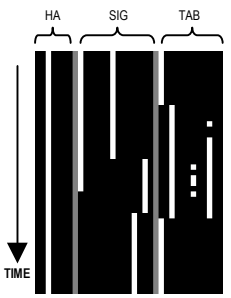
Generalised Motion Descriptors

Visual Tracking

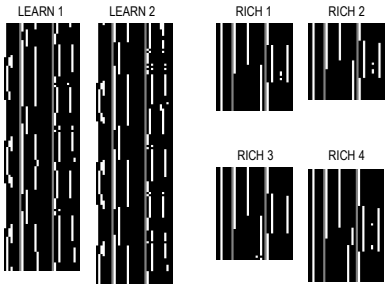
The system uses a probabilistic labelling of skin to roughly locate the face of a signer. This coupled with a contour model of the head and shoulders provides a body centered co-ordinate system in which to describe the position and motion of the hands. The hands are tracked using either skin tone or coloured gloves and their location in terms of key body parts described using their Malhalanobis distance to produce the TAB notation. Relative hand location and motion is then used to produce HA and SIG receptively. Recognition is performed using markov chains to explain the temporal sequence of events at a word level. Unlike HMMs, the chains can be built from as little as a single training example or alternatively a hand coded description of the sign.



Tracking the hands using colour
a) Original image b) Segmented skin tones
c) Tracking in operation



Format of the Binary Feature Vector



Examples Produced by the System

Results & Future Work

Due to the natural generalisation in the feature description we can achieve recognition rates as high as 100% for a lexicon of 21 words with as little as 2 or 3 training examples per word. This can be compared to other viseme level approaches based upon HMM's where over 1000 training examples are required to achieve similar levels of accuracy. As we increase the lexicon size this accuracy can begin to drop drastically but this is due to the natural ambiguity in signs. Our current feature vector does not contain orientation or hand shape classification and therefore many words form the same visual pattern. By incorporating further descriptors in the feature vector (e.g. orientation and hand shape) we expect to be able to attempt a vocabulary of several hundred words (required for minimal communication) without any grammatical constraints. It is the low training requirements that will facilitates this. Further, by removing the need for an underlying hidden process, classification becomes transparent and we hope that models can be constructed from a dictionary with refinement from video footage to further reduce training requirements. This allows the knowledge base to remain explicit, providing a mechanism to apply the same framework to different sign languages without lengthy data collation and training. Immediate future work involves

1. the development of a more complex deformable body model that will provide a more accurate description of TAB
2. the introduction of exemplar based hand shape classification and facial expression recognition to address a more extensive vocabulary
3. further generalisation using ICA to provide a distance metric within our feature space.

University of Surrey