Vision based Interpretation of Natural Sign Languages

Richard Bowden[12], Andrew Zisserman[2], 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 manuscript outlines our current demonstration system for translating visual Sign to written 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.

1 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 [1], which is a small subset of the overall problem and to a lesser extent area 2 [4][5]. 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 are simpler underlying models which can describe the events. These underlying/hidden models 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.

2 Method

Our current system is based around describing the visemes of sign (aka phase representation or constituent motions) in a manner similar to that used in sign dictionaries. Using a HA/TAB/SIG notation: HA - hand arrangement relative to each other; TAB - position relative to key body parts; SIG - hand motion relative to each other [2]. 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 [3]. The system uses a probabilistic labeling 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 locations described using the malhalanobis distance to produce the TAB notation. Relative hand location and motion is then used to produce HA and SIG receptively. Recognition is then 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.

3 Demonstration and Future Work

The exhibit demonstrates the real time recognition of a lexicon of 55 words from British Sign Language within a video stream. It is our goal to attempt a vocabulary of several hundred words (required for minimal communication) without any grammatical constraints. It is the low training requirements that facilitates this. Further, by removing the need for underlying hidden processes, classification becomes transparent and we hope that models can be constructed from dictionary models 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 the development of a more complex deformable body model that will provide a more accurate description of TAB and the introduction of exemplar based hand shape classification and facial expression recognition to address a more extensive vocabulary.

References

1. Bowden, R., Sarhadi, M.: A non-linear Model of Shape and Motion for Tracking Finger Spelt American Sign Language. Image and Vision Computing, Vol. 20/9-10. Elsevier Science Ltd, (2002) 597-607
2. Dictionary of British Sign Language, British Deaf Association UK. Brien, D (ed). Faber and Faber, ISBN: 0571143466 (1992)
3. Lyons, D., E.: A Qualitative Approach to Computer Sign Language Recognition, MPhil Thesis, Dept Engineering Science, University of Oxford (2002)
4. Starner, T., Pentland, A.: Visual Recognition of American Sign Language using Hidden Markov Models. In Int. Workshop on Automatic Face and Gesture Recognition. Zurich, Switzerland (1995) 189-194
5. Vogler, C., Metaxas, D.: ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis, In Proc. International Conference on Computer Vision. Mumbai, India. (1998) 363-369