Sign Language Analysis and Recognition: A Preliminary Investigation

Shujjat Khan, Gourab Sen Gupta, Donald Bailey
School of Engineering and Advanced Technologies
Massey University
Palmerston North, New Zealand
shujjaatkhan@gmail.com

Serge Demidenko, Chris Messom
Monash University, Sunway Campus, Bandar Sunway
Petaling Jaya, Malaysia

Abstract—: In this review paper, we analyse the basic components of sign language and examine several techniques which are helpful to design a large vocabulary recognition system for a sign language. The main focus of this research is to highlight the significance of unaddressed issues, their associated challenges and possible solutions over a wide technology spectrum.

Keywords-component; Sign language, manual, Non manual sign, detection, sensors.

I. INTRODUCTION

Sign language (SL) is a subset of gestural communication used in deaf-muted community, in which postures and gestures have assigned meanings with a proper grammar. Like any other verbal language, its discourse comprises of well structured rendering and reception of non-verbal signals according to the context rules of the complex grammar. Postures are the basic units of a sign language, and when collected together over a time axis and arranged according to the grammar rules, they reflect a concrete meaning. According to Rung-Hui et al [1], a posture is a static gesticulation of an articulator (hand, eyes, lips, body) while a gesture is a sequence of postures having defined meaning in a particular SL. For example, finger spelling in any sign language is communicated by making postures (static signs) for each letter ("a", "b", "c" etc) but in the case of continuous discourse, most signs comprise of gestures (dynamic signs). Apart from the temporal classification of signs, they can also be categorized according to their major articulators. Manual signs are performed using hands while non-manual signs (NMS) mainly include facial expressions, body movement and torso orientation. Although manual signs constitute a large proportion of sign language vocabulary, NMS also own a significant share to convey the whole context. Hence it differs from a spoken language in a way that a spoken language structure uses the words in a sequential manner but the SL structure allows manual and non-manual components to be performed in parallel. Another unique feature of SL over any spoken language is its capability to convey multiple ideas at a single instant of time. For example a signer can tell about an incoming person with the help of his left hand while his right hand can be used in parallel to report the sitting people [NZSL course Book]. Therefore an ideal SL recognition system not only requires simultaneous observation of manual and NMS but it also needs to understand the situations shown by individual articulators (multi-modal approach). In the following sections, we briefly discuss not only the existing techniques but some of the modern methods will also be explored. In the concluding section, we will propose some possible solutions on the basis of critical analysis of existing systems.

II. CURRENT APPROACHES

With a rapid growth of research in gesture recognition, new areas are being addressed in all stages of SL i.e. analysis, recognition, translation and synthesis. Ong and Ranganath [2] discussed some of the important issues related to automatic recognition of SL, rendered with both manual and NMS. Most of the existing recognition systems work with sequentially performed manual actions, due to their excessive use in gesticulation and ease of system development.

A. Non-Imaging methods and their challenges

In a non-image method, there is no image processing involved. All detection and recognition is performed on a set of data incoming to a processor unit from multiple sensor streams. All information related to signing articulators is captured by sensors or trackers. Sensors are worn by the signer and they may include displacement sensors, positional sensors or trackers. When a signer performs signing, articulator's data is captured on a specific rate and fed to the recognition stage. Unlike vision based methods, these schemes are efficient and robust due to smaller vocabulary set. On the other hand, they severely affect the user independence due to a dense mesh of installed sensors.

In the following discussion of a few non-imaging methods, we will highlight their important aspects.

1) Embedded sensors based gloves

Linguistic research in sign language has revealed that manual signs mainly consist of four major components: posture, position, orientation and motion [3]. A *DataGlove* [4], with multiple electronic sensors (installed on the finger joints, wrist and palm) is shown in Fig. 1 [7, 8], which feeds

978-1-4244-4698-8/09/$25.00 ©2009 IEEE

these measurements in real time to a processing unit [2, 5]. The processing unit compares the set of static sign samples with existing templates and generates output.

Static signs are easy to incorporate due to their defined boundaries but in the case of continuous signing, a 2D motion trajectory is formed and matched with an existing template or a learnt model. Most of these researches involve recognition of a small set of words, mainly static postures or finger spellings. In another approach, a glove fitted with 7 sensors is used for sign detection [6]. Out of 7 sensors, 5 were dedicated for finger joints and 2 for tilting and rotation of hand. Each sensor returns a discrete value from 0 to 4095, where 0 indicates fully open and 4095 for the fully bent state. The sample rate was 4 Hz and the minimum sign hold duration was 750 mSec.

The proposed design was restricted to only hand static postures and there were two extra punctuations (for word spacing and full stop) which are not available in the sign language.

The performance and effectiveness of these methods is heavily dependent on the sensor density; for example, more sensors can be added to measure the elbow bends. Rung Hui et al [1], proposed a more sophisticated embedded sensor based solution for continuous signing in real time situations. A more complex hand glove, with 10 finger sensors for one hand along with a 3D tracker for orientation, is used for detecting continuous hand gestures by their motion trajectories. In this method, the motion trajectory of a gesture is divided into 10 vectors whose relative cosines, turning points and orientations are matched with stored templates. Use of accelerometers is a great idea for providing complementary information about the hand/wrist rotation and orientation [7]. Use of more sensors in a single design may increase the burden on the processor but the proposed accelerometers work in a master-slave topology with increased local computation and reduced load on the central processor.

Wearable sensor based methods are suitable for smaller vocabulary recognition systems for static signs but they incurs an unavoidable burden of system's weight and interfaces which may affect the natural way of signing.

Figure 1. DataGloves

2) Laser based method

Perrin [8] proposed a low cost and short vocabulary single laser-based gesture recognition system in which a laser beam is transmitted and the amount of reflected energy is correlated with a reference signal to measure the displacement from the point of contact. Electronically controlled micro mirrors are installed to direct the maximum amount of reflected energy towards the receiver. In other words, these mirrors align themselves in the direction of signing hand. The current hand position is extrapolated using the velocity information and the current state of the micro-mirrors from the previous known states. In continuous signing, the 2D hand position is estimated with the help of the mirror's direction while the depth information is calculated through the approximation of displacement. For recognition, these positional parameters are matched with existing templates or fed to other classifiers. This scheme doesn't need any sensor to be worn and is suitable exclusively for those manual signs whose meanings are mainly reflected by hand or arm motion and velocity, not by finger flexion or orientation.

3) Integration of Non-Manual Signs (NMS)

Non manual signs (NMS) are those visual signals which are not conducted by hand or arms but they are shown by facial expressions and body movements.

They are an important asset of any sign language hence a complete sign language system must have provision to recognize NMS in parallel with manual signs.

Fig. 2 best describes the importance of NMS in which the left three images show a gestural sign for an American Sign Language (ASL) verb "Clean" is being performed only by hands. In next three images, the signer horizontally sweeps her head repetitively while performing the same action ("Clean") and due to this non-manual movement, this is now translated as "Very Clean". Similarly some NMS may include negligible facial expression but they may result in a major difference in context understanding.

Figure 2. Sign for "Clean" and "Very clean"

Figure 3. Facial expressions

For example, an ASL sign for "You Study Very Hard" is shown in Fig. 3 but if the recognizing person or system is unable to notice the NMS (Facial expression of signer), they may take this sign as "You Study". Many researchers are

trying to incorporate these NMS for a complete recognition system. For example some muscular movement based NMS (smile, disgust, and fear) and head movement can be detected by installing sophisticated sensors in a cap or eye glasses [9, 10].

Embedded sensor based gloves (DataGloves) are quite efficient for the detection and recognition of static signs of a short vocabulary language but they severely affect the natural way of signing by imposing the cumbersome electronics to be worn. Moreover these arrangements are not suitable to incorporate the non-manual movements of facial articulators (eyes, eye-brows etc). The same is true for laser based tracking methods.

B. Vision based methods and their challenges

In order to increase signer independence, researchers have started to think of vision based methods for interpretation, in which a signer needs to perform in front of a camera and software then interprets and translates to other spoken languages. This arrangement is termed as the second person view. Some other researchers have worked on a first person view [10], in which, the signer needs to wear the camera (in their hat, or eye glasses) covering the person's natural signing space. The camera arrangements and images can be seen in Fig. 4 [13]. We will mainly focussing on second person view.

1) Skin colour based method

Akmeliawati et al [11] proposed a short vocabulary SL interpreting system which recognises manual signs. Hands are detected using their natural skin colours along with various other features (position, velocity etc) and matched with an existing set of templates. Other important methods of recognition include hidden Markov model (HMM) [9, 12, 13] and artificial neural network (ANN) classifiers [14]. Alvi et al [5] use a stochastic language model for sentence formulation by rearranging individually recognized static signs. Davis and Shah [15] use Finite State machine to recognise a smaller set of dynamic gestures. Skin colour detection systems are severely affected by varying illumination, complex background, the signer's ethnicity (skin colour), and articulator occlusion. To eliminate the lighting problem, instead of RGB, fixed threshold ranges for Cb and Cr (YCbCr space) were utilized in a controlled environment [6].

This system may seem suitable for laboratory experimentation due to tightly controlled lighting but it will result in a large error rate if installed at public places like hospitals, courts and shops.

2) Coloured gloves method

To avoid the false positives in skin colour recognition, colour coded gloves were introduced. There were different colours on different parts of gesticulated hand (palm, fingers and back) [16] so that each part can be identified independently. These sorts of schemes, although they restrict signer's independence, are more robust as compared to skin colour based methods.

Figure 4. Hat mounted camera

3) Occlusion

Irrespective of skin colour or coloured gloves, occlusion in hand gesticulation is a normal phenomenon in which articulated parts (hands, face, and body) combine to form a posture.

It also requires a recognition system to keep track of articulators even when they are occluded (one hand with or behind the other hand, or face). In case of inter-hand occlusion, instead of individual hand detection, classical skin colour based approaches yield a single larger blob. To solve this problem, Qiang et al [17] developed an adaptive skin model which classifies the targeted skin pixels out of a larger set of skin similar pixels. Skin similar pixels are those pixels which can belong to a wide range of skin colours. In an input image, true skin colour is detected by Gaussian modelling. Two separable Gaussians model both the classes with the prominent Gaussian for skin pixels and the weaker one for false skin pixels in skin similar space. In another approach, 2D *Hand Shape Model* [18], a *Shape Transition Network* (STN) is composed of 2D hand models interconnected by links which represent the transition from one shape to other.

A simple STN is shown in Fig. 5. In case of occlusion, the hand gesticulation is predicted by the previous hand positions and the transitions (learnt in the training stage). In training, either different signers sign in front of the system or annotated sign language databases [3] are used. These video databases contain a number of sign language sentences with different signers, lighting conditions and backgrounds. These benchmarks are recorded and annotated for training, development and testing of SL recognition systems.

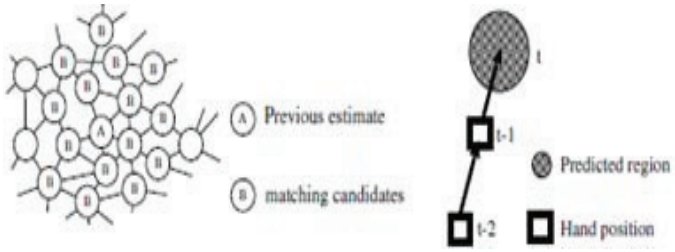

Figure 5. Shape transition network

4) Lack of depth information

Apart from occlusion, some manual signs (e.g. "Ask") involve the hands motion towards the observer which requires depth information to be incorporated. Similarly temporal information (i.e. past, future) of a sign can also be incorporated by performing it in spatially different location. This is a very challenging situation especially for vision systems because 2D images are unable to present 3D scenes.

5) Stereo vision

To overcome this issue, a stereo-vision based method was devised to incorporate depth information associated with a sign. Stereo imaging uses multiple images of the same scene taken from different camera locations. Disparity is defined as the relative movement of an object between two or more views, with the disparity being a function of depth. A dense disparity map is computed between both acquired images by applying an affine transform on corresponding points in images [19]. Objects closer to the camera have a greater disparity of movement between two images, and this is used to calculate the distance to the objects [20]. If same techniques are applied in a SL recognition system, a large number of signs can be recognized by analyzing the 3D trajectory instead of 2D. Because of computationally expensive nature of stereo vision, the direction of research has shifted to alternative range finding methods.

6) Time of flight camera

Latest technological development has resulted in the advent of the state-of-the-art *Time of Flight* (ToF) cameras [21] which can acquire not only intensity image but also a depth image of the scene with precision of a few millimetres. Currently a few ToF camera based solutions are proposed which accurately spot the movement and orientation of articulators in 3D [22, 23]. Although a ToF camera can produce intensity image, it normally has low lateral resolution which is unsuitable for processing fine details.

Vision based methods have the added potential of incorporating NMS along with the hand gesticulations due to on-going research on visual gesture recognition. Although many researchers work on the recognition of facial expressions and head movement classification, very few have focussed on the integration of NMS with existing sign language interpreting systems [2]. The main challenge is due to the lack of local information to capture NMS along with hand/fingers flexion. With a standard VGA resolution, we can detect global movement (hand, arm, head and body) but detection of local movements (eye, eyebrow, forehead muscles) and finger gesticulations using lower local resolution may lead to undesired results. A possible solution is to use a higher resolution camera.

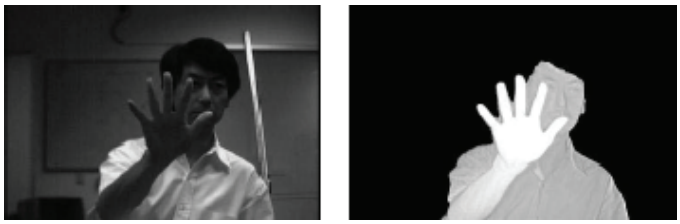

Figure 6. Intensity image and depth image

7) Foveal vision

Gesture recognition in a higher resolution image takes more computational time and result in a noticeable drop in video frame rate. The situation becomes worse when complex recognition or iterative learning algorithms are implemented. To cope with this situation, a biologically inspired foveal vision [24] technique can be adapted. In this technique, an image is sub-sampled around a fixation point in such a way that resolution decreased for the area away from the fixation point. At one or more fixation points, image resolution is sufficiently high to detect minor NMS related to facial components and expression (eyes, mouth, cheeks and forehead) [25]. These techniques have been successfully deployed in face recognition and facial expressions and have potential to work with SL recognition.

C. Grammar Models

In a SL, continuous signs are detected and re-arranged in an appropriate order for a discreet context. Grammar models are very important referential models for complete syntax and semantics of a formulated sentence.

1) Dynamic gestures

In the real world, all signers sign in a natural and continuous way with a specific signing frequency and transition delay. Most of the existing systems work fine with static postures but due to insufficient training, their error-rates are higher for dynamic gestures. Dynamic signs can be sub-divided into a set of basic movements which can be analysed in higher recognition layers with the support of a powerful grammar model or Markov chain [26]. Due to division of a gesture into basic units, it is possible to represent a huge range of sign gestures. Another advantage of this scheme is that some undetected or erroneous units can simply be skipped. However the requirement of a powerful grammar model will result in a computationally expensive solution.

2) Pointing gestures

Current recognition systems are unable to understand pointing gestures (deictic and anaphora). These are important signs (called classifiers) used in, for example, telling a story or to refer to a person, object or place which is available or absent from signing space [2]. In order to incorporate the signals, an intelligent grammar model needs to be applied in algorithms that can differentiate these classifiers from other manual signs.

III. Summary and Conclusion

The study of sign language recognition uncovers the importance of a complete system that can interpret a large vocabulary of manual and non-manual signs. Existing techniques mainly focus on the detection of only static manual postures (without NMS) or those dynamic gestures which do not involve any motion in the third dimension. The advent of ToF range camera can be beneficial to aid the practical development of such a system but because of its lower resolution intensity images, fine details of signers are unavailable e.g. the camera can detect a moving hand in 3D but local details of NMS (eye-movements, eyebrows etc) are not captured with high resolution.

A number of other hybrid schemes may be used to augment the existing systems e.g. a few simple embedded sensors can be used with vision based approaches which can aid the system to spot the articulators in a complex or cluttered background. Similarly a combination of 3D arrangement and a powerful grammar model can

significantly augment the vocabulary of SL recognition systems. With such systems, a signer can express himself with more freedom of discourse and with fewer restrictions.

REFERENCES

[1] Rung-Huei, and M. Ouhyoung, ""A real-time Continuous Gesture Recognition System for Sign Language"," *IEEE International Conference on Automatic Face and Gesture Recognition*, Japan, pp. 558-567, 14-16 April, 1998

[2] S. C. W. Ong, and S. Ranganath., "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, 2005, pp. 873-891.

[3] Z. Morteza, D. Philippe, R. David, D. Thomas, B. Jan, and N. Hermann, "Continuous Sign Language Recognition –Approaches from Speech Recognition and Available Data Resources," *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios (2006)*, Genoa, Italy, pp. 21-24, 24-26 May, 2006

[4] W. jiangqin, G. wen, S. yibo, L. wei, and P. bo, "A simple sign language recognition system based on data glove," *4th International Conference on Signal Processing (ICSP 98)*, Beijing, China, pp. 1257-1260, Oct 12-16, 1998

[5] A. K. Alvi, M. Y. B. Azhar, M. Usman, S. Mumtaz, S. Rafiq, R. U. Rehman, and I. Ahmed, "Pakistan Sign Language Recognition Using Statistical Template Matching," *International Conference On Information Technology, (ICIT 2004)*, Istanbul, Turkey, pp. 108-111, 17-19 December, 2005

[6] S. Mehdi, and Y. Khan, "Sign language recognition using sensor gloves," *Proceedings of the 9th International Conference on Neural Information Processing ICONIP '02*, Vol 5, Orchid Country Club, Singapore, pp. 2204-2206, 18-22 November, 2002

[7] H. B. Thad, and T. Starner, "Using multiple sensors for mobile sign language recognition," *Seventh IEEE International Symposium on Wearable Computers*, New York, USA, pp. 45-52, 21-23 October, 2003

[8] S. Perrin, A. Cassinelli, and M. Ishikawa, "Gesture recognition using laser-based tracking system," *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, pp. 541-546, 17-19 May, 2004

[9] T. Starner, and A. Pentland, "Real Time American Sign Language Recognition from Video using Hidden Markov Model," *International Symposium on Computer Vision*, Florida, USA, pp. 265-270, 19-21 November, 1995

[10] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, 1998, pp. 1371-1375.

[11] R. Akmeliawati, M. P. Leen Ooi, and Y. C. Kuang, "Real-Time Malaysian Sign Language Translation using Colour Segmentation and Neural Network," *Instrumentation and Measurement Technology Conference (IMTC2007)*, Warsaw, Poland, pp. 1-6, 1-3 May, 2007

[12] G. L. Fang, X. J. Gao, W. Gao, and Y. Q. Chen, "A novel approach to automatically extracting basic units from Chinese sign language," *17th International Conference on Pattern Recognition (ICPR)*, Cambridge, England, pp. 454-457, Aug 23-26, 2004

[13] W. Gao, G. L. Fang, D. B. Zhao, and Y. Q. Chen, "A Chinese sign language recognition system based on SOFM/SRN/HMM," *Pattern Recognition*, vol. 37, no. 12, 2004, pp. 2389-2402.

[14] W. Gao, J. Y. Ma, J. Q. Wu, and C. L. Wang, "Sign language recognition based on HMM/ANN/DP," *2nd International Conference on Multimodal Interface (ICMI 99)*, Hong Kong, pp. 587-602, 5-7 January, 1999

[15] J. Davis, and M. Shah, "Visual gesture recognition," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 141, no. 2, 1994, pp. 101-106.

[16] W. Ken, I. Yoshio, Y. Yasushi, and Y. Masahiko, "Recognition of sign language alphabet using colored gloves," *Systems and Computers in Japan*, vol. 30, no. 4, 1999, pp. 51-61.

[17] Z. Qiang, C. Kwang-Ting, W. Ching-Tung, and W. Yi-Leh, "Adaptive learning of an accurate skin-color model," *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, pp. 37-42, 17-19 May, 2004

[18] Y. Hamada, N. Shimada, and Y. Shirai, "Hand Shape Estimation under Complex Backgrounds for Sign Language Recognition," *IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, pp. 589-595, 17-17 May, 2004

[19] P. Dreuw, P. Steingrube, T. Deselaers, and H. Ney, "Smoothed Disparity Maps for Continuous American Sign Language Recognition," *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, Vol 5524, Póvoa de Varzim, Portugal, pp. 24-31, 10-12 June, 2009

[20] J. R. Seal, D. G. Bailey, and G. Sen Gupta, "Depth perception with a single camera," *International Conference on Sensing Technology*, Palmerston North, New Zealand, pp. 96-101, 21-23 November, 2005

[21] A. Kolb, E. Barth, and R. Koch, "ToF-sensors: New dimensions for realism and interactivity," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '08)*, Alaska, USA, pp. 1-6, 24-26 June, 2008

[22] M. B. Holte, T. B. Moeslund, and P. Fihl, "Gesture Recognition using the CSEM SwissRanger SR-2 Camera," *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3/4, 2008, pp. 295-303

[23] L. Xia, and K. Fujimura, "Hand gesture recognition using depth data," *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* Seoul, Korea, pp. 529-534, 17-17 May, 2004

[24] M. Silviu, M. Sridhar, H. John M, and D. Fred, *Biologically Motivated Computer Vision*, Springer Berlin / Heidelberg, 2000.

[25] A. Holmes, M. Kiss, and M. Eimer, "Attention modulates the processing of emotional expression triggered by foveal faces," *Neuroscience Letters*, vol. 394, no. 1, 2006, pp. 48-52.

[26] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, 2009, pp. 1685-1699.