IMTC 2007 – Instrumentation and Measurement
Technology Conference
Warsaw, Poland, 1-3 May 2007

Real-Time Malaysian Sign Language Translation using Colour Segmentation and Neural Network

Rini Akmeliawati[1], Melanie Po-Leen Ooi[2], Ye Chow Kuang[3]
Monash University,
2 Jalan Universiti, Bandar Sunway, 46150 Petaling Jaya, Malaysia
[1]rini.akmeliawati@eng.monash.edu.my [2]melanie.ooi@eng.monash.edu.my [3]kuang.ye.chow@eng.monash.edu.my

Abstract *– In this paper we present an automatic visual-based sign language translation system. Our proposed automatic sign-language translator provides a real-time English translation of the Malaysia SL. To date, there have been studies on sign language recognition based on visual approach (video camera). However, the emphasis on these works is limited to a small lexicon of sign language or solely focuses on fingerspelling, which takes different approaches respectively. In practical sense, fingerspelling is used if a word cannot be expressed via sign language. Our sign language translator can recognise both fingerspelling and sign gestures that involve static and motion signs. Trained neural networks are used to identify the signs to translate into English.*

Keywords *– Image processing, sign language, neural network*

I. INTRODUCTION

Sign language is a highly structured non-verbal language utilizing both manual communication and non-manual signals. Manual gestures and communication consists of movement and orientation of hand/arm that conveys symbolic meaning while non-manual signals are mainly facial expression, head movement, posture and orientation of body and torso [1]. Contrary to popular belief, although sign language is used mainly deaf communities globally, it does not have a universal language for all deaf communities around the world. Rather, sign languages have systematic and complex grammars, which differ from country to country and may even have much variation in its local dialect depending on the level of dependency on local culture, for example [2]-[4].

For many individuals in Malaysia who were either born deaf or became hard of hearing at an early age, the Malaysian Sign Language, better known as *Bahasa Isyarat Malaysia* (BIM) is used as their first language. English and Malay languages are learnt only as a second language. As a result, their reading and writing skills for both English and Malay are often below average as they mostly opt to converse in BIM. Although some can read, many others are disadvantaged in cases where reading is needed. Examples include accessing government websites whereby no video clip for deaf and mute is available or filling out forms online whereby no interpreter may be present to help.

At present, automatic sign language translation systems generally use two approaches; data-glove [5] and visual-based [6][7],[11] approaches. The data-glove approach uses a specially built electronic glove, worn by a signer, which has in-built sensors that work to detect and transmit information on the hand posture. Most commercial sign language translation systems use the data-glove method, as it is easy to obtain information about the degree of finger flexing and 3D position of the hand using the gloves. Thus, this system requires less computational power, and real-time translation is much easier to achieve.

There are, however, some downsides to this approach. Firstly, these data-gloves can be quite costly. The VPL data-glove, for instance, costs over US$9,000 [5]. While it is possible to utilize cheaper data-gloves, they are much more susceptible to noise, and the reduced amount of sensors causes loss of important information about the hands. This results in the loss of accuracy in sign translation. Furthermore, the data-glove somehow is less comfortable to the signers.

With recent advancement in computer and information technology, there has been an increased attention to vision-based approach. A camera is used to capture images of the signer and some image and video processing is carried out to perform recognition of the sign language. The major advantage of this approach compared to the data-glove approach, is the flexibility of the system. It can be developed to include non-manual signals such as recognition of facial expressions and head movements as well as perform lip-reading. Vision-based sign language translator (SLT) systems in general utilises colour information, and can be divided into two major techniques: those that use custom-made colour gloves [6],[11] and those that are based on skin-colour recognition.

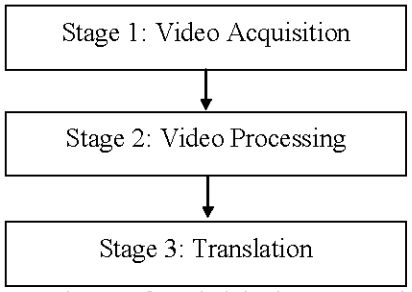

Fig. 1 The general stages of a typical sign language translator system

Fig. 1 shows a typical vision-based SLT system that consists of three stages: image capturing, pre-translation

image processing and performing actual translation, which includes outputting the English translation of the sign(s).

To date, there has been much research poured into stage 3 of the SLT system (refer to Fig 1). Holden [6] attempts to extract the 3D model of the hand using a hand-motion understanding unit and a fuzzy recognition system to recognize a set of 22 signs. Binh et al. [7] uses a variation of Hidden Markov models for trained gesture recognition systems. Lamar [8] and Symeonidis [9] make use of neural networks for gesture recognition. A comprehensive survey on the work done in the field can be found in Ong [1].

Much of the previous work done on sign language recognition focuses exclusively on fingerspelling. However, fingerspelling solely is not enough to construct an SLT system that is of use to the deaf communities. Furthermore, most systems rely on focusing the camera on only the hand, with the aim of obtaining larger, clearer images of the hand. A more practical system should ideally be capable of recognizing signs based on hand position, posture, and movements in addition to fingerspelling. The camera should also capture more than just the hands of the signer (for manual communication), it ideally should also include the face and upper torso for the signer (for non-manual communication). The tradeoff of this system is that at this distance, the size of the hand on the image can be quite small, and the large amount of objects in the background will add to noise to the image.

This paper proposes an SLT system is able to translate alphabets, numbers and a few word from the Malaysian Sign Language or Bahasa Isyarat Malaysia (BIM) in real-time using minimal hardware. It is developed and tested for different background environments corresponding to the distances and conditions expected for a practical and commercial sign language translator. *Minimal hardware* is achieved using only a basic webcam and simple glove to reduce the overall cost of the system. By providing real-time translation, the system will perform image acquisition and processing as well as provide translation of sign language all within a short time frame from the instant it is signed. This is in an attempt to overcome the huge computational requirements that are the most common barrier in creating an automatic real-time sign language translator.

II. SIGN LANGUAGE TRANSLATOR SPECIFICATIONS

As was mentioned in the introduction, BIM is the main sign language used in Malaysia. In the event that a word could not be recognized by the translator due to its limited lexicon, finger spelling of the alphabet one-by-one of a word could be performed. The proposed automatic SL Translator has the following criteria:

1) real-time: The translator is sufficiently fast to capture images of signer, process the images and display the sign translation on the computer screen.
2) Vision-based: The input to the translator is sets of images, which are captured by camera.
3) Automatic: The system should be able to perform the translation without continuous prompting such as inputs from computer keyboard.
4) Translation: The translator should be able to recognise particular sets of BIM signs (including fingerspelling) as shown in Table 1 and translate them into English with accuracy above 95%. At this early study, we neglect the grammar. We will only aim at word-to-word translation of meaningful sentence-like signs.

Table 1 BIM Vocabulary used

Numbers:	1 to 9
Alphabet:	A to Z
Words:	Beautiful, close, driving, he, his, house, I, is/am/are, mine, telephone, you

The set of words chosen are commonly used, and can form meaningful short sentence to demonstrate the feasibility of recognizing complicated series of signs. In addition, similar signs that share the same movement are selected to exhibit that the translator design is capable of differentiating signs even with minor difference involved. The above BIM vocabulary consists of 36 static signs and 10 signs with motion.

A more detailed structure of the SLT system (based on Fig. 1) is shown below:

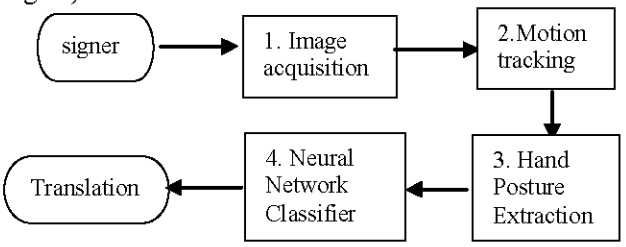

Fig. 2 Vision-based Automatic SLT

The rest of this paper will follow the order of the process in Fig. 2. In Section III the image acquisition process will be discussed. Section IV the hand-motion tracking will be presented, followed by Neural Network Classifier in Section V. Finally, the paper is summarized in Section VI.

III. IMAGE ACQUISITION

A. Equipment

a. Camera: Images from a signer is obtained by using Logitech webcam in the form of still images and video streams in RGB (red-green-blue) colour with a screen bit depth of 24-bits and a resolution of 352 X 288 pixels.

Webcam is used as it is widely available and easily obtained. The above resolution is chosen to satisfy the execution time constraint. Higher resolution causes considerable delay in the execution of the acquisition process and longer processing time.

b. Custom-made gloves: The signer is equipped with colour-coded gloves as shown in Fig. 3a and b. The colours will aid the extraction of data from the sign images through colour segmentation. These custom-made gloves are ordinary pairs of gloves with specific-colour patches on the palm, and each fingertip. Unlike the electronic data-gloves, they are cheap and non-restrictive towards motion.

i) Colours for the right-hand glove were chosen to contrast strongly with each other as they are far apart from each other in the RGB colour space (each colour lies on separate corners of the RGB colour cube) and easily distinguished from common background, skin and clothing colours. The thumb tip is purple, the index finger and the ring finger tips are red, the middle finger and the little finger tips are dark blue. The rest of the glove is left as white. See Fig. 3a.

ii) The left-hand glove is red coloured in the metacarpus part of the hand (palm and the back of the hand) as shown in Fig. 3b. The left hand has no role in finger-spelling and in most signs is either not used[1], or has a posture very similar to that of the right hand. Thus, for the present study, the hand posture of the left hand is ignored; however, its movements are used in the translation process.

(a) (b)

Fig. 3 Gloves for the signer

c. environmental set-up: Image acquisition process is subjected to many environmental concerns such as the position of the camera, lighting sensitivity and background condition. The video camera is placed to focus on an area that can capture the maximum possible movement of the hand and take into account the difference in height of individual signers. Thus, the proposed position of the web camera will be about 2 meters from the floor and 1.5 meters away from the person. In this way, the movement of the hand towards or away from the camera can be detected as well. Sufficient lighting is required to ensure that the colour of the gloves is bright enough to be seen and analysed.

[1] The signer is assumed to be right-handed, so the dominant hand in all word or fingerspelling gestures will be the right hand.

B. Image Acquisition Process

The objective of this stage of work is to collect frames of images at suitable intervals so that the frames are sufficient to detect the movement of the hand. Capturing five frames per second (*fps*) is found to be sufficient. Higher fps will only lead to higher computation time of the computer as more input data to be processed. Similarly, for the purpose of the image acquisition, frame resolution of 352x288 pixels is deemed reasonable, as higher resolution will take up more memory space and computation time. As the acquisition process runs at real time, this part of the process has to be efficient. Thus, previous frame that has been processed will be automatically deleted to free the limited memory space in the buffer.

Image acquisition that involves signs with motion uses the Sum of Absolute Difference (SAD) algorithm, which is expressed by (1). Colour segmentation of yellow colour, which represents the glove will be carried out before the SAD algorithm is applied to eliminate the background, which might involve moving objects.

$$D_{m-1}(t) = \frac{1}{N}\left|I(t_i) - I(t_{i+1})\right| \quad , \qquad (1)$$

where N is the number of pixels in the image used as the scaling factor, $I(t_i)$ is the image I at time i , $I(t_{i+1})$ is the first image after image I at time i, $D(t)$ is the normalized sum of absolute difference between two frames and m is the number of frames captured. In an ideal case when there is no motion, $I(t_i) = I(t_{i+1})$ and $D_{m-1}(t)=0$.

When the SAD value is less than a predefined threshold value for a few consecutive frames of images, then there are no movements of the glove between these frames. Thus, the last frame will be collected for further processing in the next stage. All other previous frames that show almost identical images will be discarded.

IV. HAND TRACKING MOTION

Fig. 4 shows the simplified process flow of acquiring the image and performing pre-processing. Upon launching of the SLT program, frames are read in and color segmentation of yellow colour regions will be performed to remove any influence from moving objects in the background. This is because only the movements of the hands are of interest at this point.

The centroid of the color segmented region is found for hand tracking purposes. This is done to reduce scan area and

improve computation time and is vital for interpretation of continuous signing and signs that involve motion.

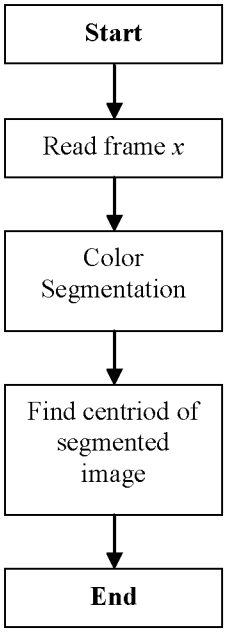

Fig. 4 Flowchart of image acquisition and pre-processing

A. Image Processing and Segmentation

Due to the vast quantity of unwanted data that the captured image inevitably contains, some preprocessing work is necessary to filter out unwanted parts. The preprocessing task consists of the following operations:

- Colour space transformation – The image is transformed into the YCbCr colour space using the following operations:

$$\begin{aligned} Y &= 0.257R + 0.504G + 0.098B + 16, \\ Cb &= -0.148R - 0.291G + 0.438B + 128, \\ Cr &= 0.439R - 0.368G - 0.071B + 128, \end{aligned} \qquad (2)$$

where *R, G, B* are the intensity values of red, green and blue colours, whereas *Y* is the luminence, and *Cb* and *Cr* are the chromiance of the pixel. The results of this process are shown in Fig. 5.

- Colour segmentation – Portions of the image of the same colour tones as those specified for the gloves are extracted from the image. This is done by building a binary mask over the image based on threshold values of the desired colours. Thresholds are set based on chrominance (*Cb*, *Cr*), and chrominance difference (*Cb-Cr* or *Cr-Cb*). Results of the colour segmentation are shown in Fig. 6.
- Morphological opening and closing – Opening removes any stray noise pixels from the image that managed to slip past the colour segmentation filtering. Closing is then used to smoothen the image contours.
- Fill holes – A fill operation is used to close up any small holes within the finger/hand portions of the glove in the image.

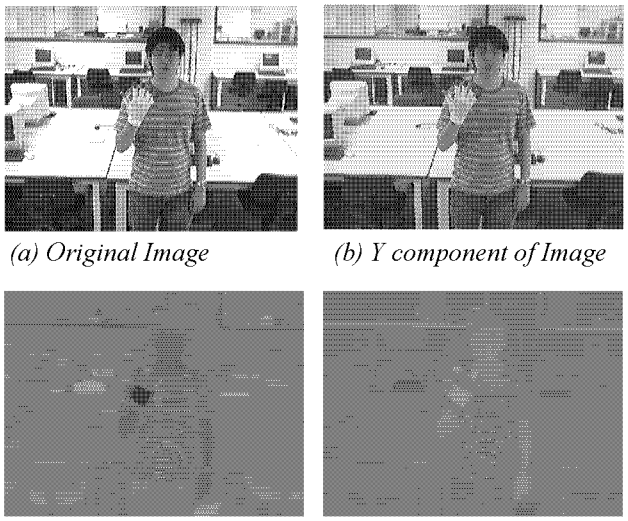

Fig. 5 Results of Colour Space Transformation

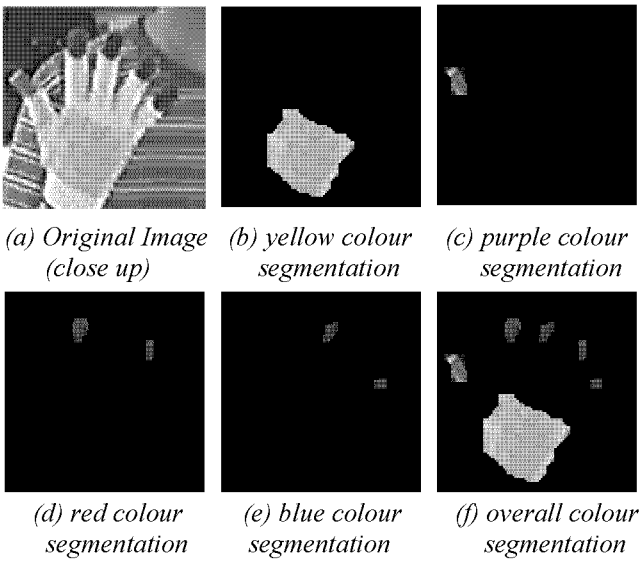

Fig. 6 Results of Colour Segmentation

B. Tracking of Hand Movement

For the detection of motion-based sign language, the details of movement of each finger were omitted. This is because most signing does not involve individual finger movement as used in finger spelling; instead it only involves gestures, which represent whole words [6]. This allows signed conventions to proceed at about the pace of a verbal conversation.

At this stage of work, the location of the centroid of the glove, which is the center of mass is calculated for the yellow region of the image. This yellow region is determined via colour segmentation as mentioned earlier. By repeating the process for consecutive frames of images, a set of changing centroid location is obtained. This process is shown in Fig. 7 where three consecutive centroids is plotted on the final image. Thus, initial information about the hand position (x and y position) and movement can be extracted from here.

In order to provide further useful feature selection, the eccentricity of the ellipse that has the same movement region of the hand is calculated. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. This is useful in providing information whether the hand has moved in a straight line or in a circle.

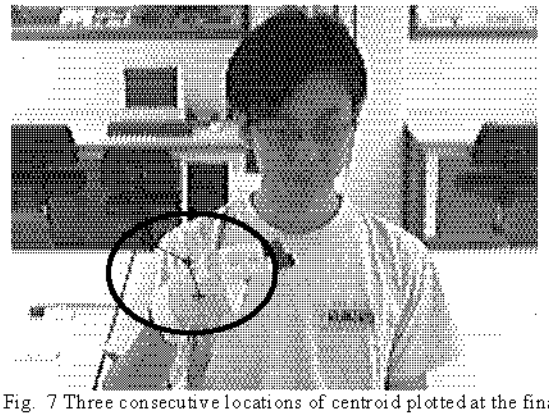

Fig. 7 Three consecutive locations of centroid plotted at the final image

C. Hand Posture Extraction

For finger spelling, the exact position of each finger is vital, as signing one alphabet may differ from another by the flexing of just a single finger, such as in the case of 'I' and 'S'. Information of the finger position is extracted from the image in the form of vectors from the centroid of the hand. As hand posture is the main identifier for fingerspelling signs, it needs to be sensitive enough to detect the flexing of a single finger.

The use of 3 different colours for the glove fingertips: purple for the thumb and alternating red and blue colours for the other 4 fingers allows good estimation about the flexing of each finger. The location of each finger is determined as a finger vector of the finger centroid relative to the centroid of the hand:

$$xyvector = [\,(x_{c,hand} - x_{c,finger}),\ (y_{c,hand} - y_{c,finger})\,] \qquad (3)$$

where xc = x-coordinate of the hand/finger centroid

yc= y-coordinate of the hand/finger centroid

Each finger of the hand produces 1 xyvector, so the hand posture is determined as a set of 5 xyvectors. The data is then presented to the neural network as a set of 10 finger_vectors, calculated by the following equation:

$$finger\_vector = [xyvector_i]T\ ,\ i = 1,2,\ldots,5 \qquad (4)$$

The results of the finger_vector extraction are shown in Fig. 8.

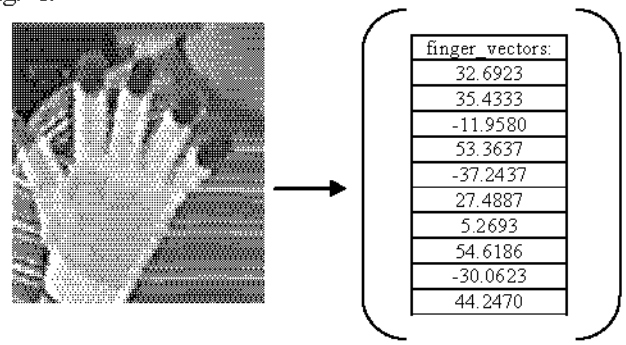

Fig. 8 Extraction of finger vectors from the image

V. ARTIFICIAL NEURAL NETWORK (ANN) CLASSIFICATION

For the sign recognition, we use two-layer feedforward neural network. Neural networks are a powerful tool for image processing and sign recognition, due to their fast speed, nonlinear mapping, ability to learn from experience, and ability to generalize from examples to correctly respond to new data. The input and output layers use the tan-sigmoid and the log-sigmoid transfer functions, respectively. Three neural networks are used to recognise alphabet, number and word signs separately. For training the sign database we use the resilient backpropagation rule as it provides the fast rate of training with considerably small memory requirement.

The decision process on recognising the signs using ANN is shown in Fig. 9.

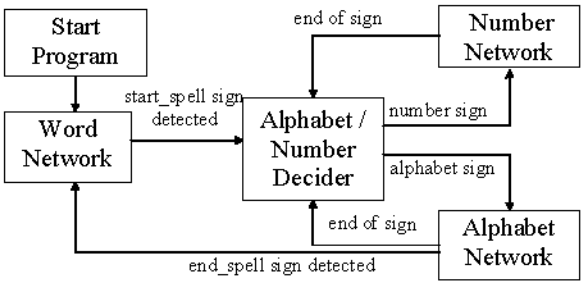

Fig. 9 Translation Process of a Sign

For a single neuron, the output a is obtained as a result of the transfer function f, with input represented by a summation of the weighted input w_p, and the bias b. The transfer function is typically a step function, such as the hard-limit transfer function, or a sigmoid function such as log-sigmoid.

$$a = f(w_p + b) \qquad (5)$$

The neuron can also take in a vector of inputs, rather than a single input, each with their own weights. The training of the neural network trains the weights of the input vector for the output to match the desired output.

The finger vectors extracted earlier are used as inputs to train and test the neural network. Most of the sign recognition projects utilizing neural networks rely on a single network to successfully identify and recognize the different signs. However, when larger sets of signs are used as a database, the number of neurons needed to recognize the new database becomes larger. This causes the whole system to slow down as huge networks require a vast amount of complex computation. Furthermore, the accuracy of the recognition also drops with larger databases.

For this sign language translator, it is proposed that only a few vectors containing the important information of the hand be used, thus reducing computational requirements while maintaining accuracy. Furthermore, the database will be classified first, and split into a number of neural networks based on certain characteristics. For example, the region of signing can be divided into multiple sections, and different neural networks called to recognize gestures in different signing regions. By doing this, we also save time on training the neural network, as only 1 network needs to be retrained when a new sign is added to the database.

VI. RESULTS

Our proposed automatic sign language translator (ASLT) recognizes all 49 signs in the BIM vocabulary shown in Table 1. The ASLT was tested on a test database and results of the system are as tabulated below in Table 2.

Table 2 Performance of ASLT

Category:	Number of training	Training time	Accuracy of Detection
Numbers	55 epochs	< 10 seconds	99.333%
Alphabets	105 epochs	½-1 minute	95.673%
Words	32 epochs	< 10 seconds	95.000%

The training of the neural network indicated in Table 2 is in order to achieve the desire performance criteria of 0.001 sum-squared error.

The accuracy of the ASLT deteriorates when the signs are not properly conducted. Additionally, signs that are similar in posture and gesture to another sign can be misinterpreted, resulting in a decrease in accuracy of the system. To increase the performance and accuracy of the ASLT, the quality of the training database used to train the ANN should be enhanced to ensure that the ANN picks up correct and significant characteristics in each individual sign.

VII. CONCLUSION

The automatic visual-based sign language translator proposed in our paper can translate Bahasa Isyarat Malaysia (BIM) or Malaysian Sign Language, in real-time. Our system achieved the recognition rate of over 90%. This system is the first step in developing a low-cost ASLT for commercial use.

REFERENCES

[1] Ong, S.C.W, Ranganath, S., "Automatic Sign Language Analysis: A Survery and the Future beyond Lexical Meaning", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 16-19, Jun. 2005.

[2] Werapan, W.;,Chotikakamthorn, N., "Improved dynamic gesture segmentation for Thai sign language translation", *Proc. of Int'l Conference on Signal Processing (ICSP 2004)*, vol. 2, pp. 1463-1466, Sept. 2004.

[3] Ohene-Djan, J., Naqvi, S.,, "An adaptive WWW-based system to teach British sign language", *Proc. of Int'l Conference on Advanced Learning Technologies (ICALT 2005)*, pp. 127-129, Jul. 2005.

[4] Wang, C., Chen, X., Gao, W., "Expanding Training Set for Chinese Sign Language Recognition", Proc. of Int'l Conference on Automatic Face and Gesture Recognition, (FGR 2006), pp. 323-328, Apr. 2006.

[5] Waleed M.K., Whole Hand Input Devices [Online] [accessed 2005 April]. Available from URL: http://www.cse.unsw.edu.au/~waleed/thesis.

[6] Holden, E., "Visual Recognition of Hand Motion", PhD Thesis, *Department of Computer Science*, University of Western Australia, 1997.

[7] Binh, N.D., Shuichi E. and Toshiaki E., "Real-Time Hand Tracking and Gesture Recognition using Pseudo 2-D Hidden Markov models", *Proc. of Int'l Conference on Robotics, Vision, Information and Signal Processing (ROVISP 2005)*, pp.403-407, July 2005.

[8] Lamar, M.V., "Hand Gesture Recognition Using T-CombNET", PhD thesis, *Faculty of the Graduate Division*, Nagoya Institute of Technology, 2001.

[9] Symeonidis, K., "Hand Gesture Recognition Using Neural Networks", Master's Thesis, *School of Electronics and Electrical Engineering*, UniS, 2000.

[10] http://en.wikipedia.org/wiki/Malaysian_Sign_Language.

[11] Starner, T., and Pentland, A., "Visual Recognition of American Sign Language Using Hidden Markov Models", *Proc. Int. Work. on Auto. Face and Gesture Recognition*, Zurich, 1995

[12] Gonzalez RC, Woods RE. "Digital image processing using MATLAB," New Jersey: *Pearson Prentice Hall*; 2004