

Hand Gesture Recognition: Sign to Voice System (S2V)

Oi Mean Foong, Tan Jung Low, and Satrio Wibowo

Abstract—Hand gesture is one of the typical methods used in sign language for non-verbal communication. It is most commonly used by people who have hearing or speech problems to communicate among themselves or with normal people. Various sign language systems have been developed by manufacturers around the globe but they are neither flexible nor cost-effective for the end users. This paper presents a system prototype that is able to automatically recognize sign language to help normal people to communicate more effectively with the hearing or speech impaired people. The Sign to Voice system prototype, S2V, was developed using Feed Forward Neural Network for two-sequence signs detection. Different sets of universal hand gestures were captured from video camera and utilized to train the neural network for classification purpose. The experimental results have shown that neural network has achieved satisfactory result for sign-to-voice translation.

Keywords—Hand gesture detection, neural network, sign language, sequence detection.

I. INTRODUCTION

THIS system was inspired by the special group of people who have difficulties communicate in verbal form. It is designed with the ease of use for human-machine interface in mind for the deaf or hearing impaired people. The objective of this research is to develop a system prototype that automatically helps to recognize two-sequence sign languages of the signer and translate them into voice in real time.

Generally there are two ways to collect gesture data for recognition. Device based measurement which measures hand gestures with equipment such as data gloves which can archive the accurate positions of hand gestures as its positions are directly measured. Secondly, vision-based technique which can cover both face and hands signer in which signer does not need to wear data gloves device. All processing tasks

Manuscript received June 30, 2008. Oi Mean Foong is the corresponding and is attached to the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Tronoh 31750, Perak, Malaysia (e-mail: foongoimean@petronas.com.my).

Tan Jung Low is a Senior Lecturer with Universiti Teknologi PETRONAS. He is now with the Department of Computer and Information Sciences, Bandar Seri Iskandar, Tronoh 31750, Perak, Malaysia (e-mail: lowtanjung@petronas.com.my)

Satrio Wibowo graduated with Bachelor of Technology (Hons.) Degree in Information and Communication Technology, Universiti Teknologi PETRONAS. Currently, he is working as System Analyst in Netherland.

are solved by using computer vision techniques which are more flexible and useful than the first method [1].

During the last half of the century, sign languages are now accepted as minority languages which coexist with majority languages [2] and they are the native languages for many deaf people. The proposed system prototype is designed to help normal people to communicate with deaf or mute people more effectively. This paper presents a prototype system known as Sign to Voice (S2V) which is capable of recognizing hand gestures by transforming digitized images of hand sign language to voice using Neural Network approach.

The rest of the paper is organized as follows: Section II surveys the previous work on image recognition of hand gestures. Section III proposes the system architecture of SV2 prototype. Section IV discusses the experimental set up and its results, and lastly Section V draws conclusions and suggests for future work.

II. RELATED WORK

Attempts on machine vision-based sign language recognition have been published only recently with relevant literature several years ago. Most attempts to detect hand gestures/signs from video place restrictions on the environment. For examples, skin colour is surprisingly uniform so colour-based hand detection is possible [3]. However, this by itself is not a reliable modality.

Hands have to be distinguished from other skin-coloured objects and these are cases of sufficient lighting conditions, such as colored light or grey-level images. Motion flow information is another modality that can fill this gap under certain conditions [4], but for non-stationary cameras this approach becomes increasingly difficult and less reliable.

Eng-Jon Ong and Bowden [5] presented a novel, unsupervised approach to train an efficient and robust detector which applicable of not only detecting the presence of human hands within an image but classifying the hand shape too. Their approach is to detect the location of the hands using a boosted cascade of classifiers to detect shape alone in grey scale image.

A database of hand images was clustered into sets of similar looking images using the k-mediod clustering algorithm that incorporating a distance metric based on shape context [5]. A tree of boosted hand detectors was then formed, consisting of two layers, the top layer for general hand detection, whilst branches in the second layer specialize in classifying the sets of hand shapes resulting from the unsupervised clustering method.

The Korean Manual Alphabet (KMA) by Chau-Su Lee et al [6] presented a vision-based recognition system of Korean manual alphabet which is a subset of Korean Sign Language. KMA can recognize skin-coloured human hands by implementing fuzzy min-max neural network algorithm using Matrox Genesis imaging board and PULNIX TMC-7 RGB camera.

III. SYSTEM ARCHITECTURE

Fig. 1 shows the system architecture of the proposed S2V system prototype. Image acquisition for hand detection is implemented using the image processing toolbox in MATLAB. This is to develop functions to capture input from signer and detect the hand-region area. The limitation here is the background of the image can only be in black color. Therefore several problems were encountered in capturing and processing image in RGB value. Thus, we need to find an approach to detect the hand-region to produce satisfactory results.

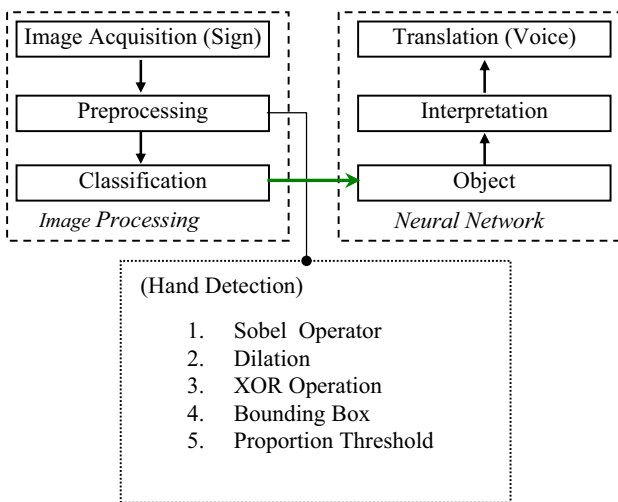


Fig. 1 S2V System Architecture

A. Image Recognition

The input images are captured by a webcam placed on a table. The system is demonstrated on a conventional PC Laptop computer running on Intel Pentium III Processor with 256 MB of RAM. Each image has a spatial resolution of 32 x 32 pixels and a grayscale resolution of 8 bit. The system developed can process hand gestures at an acceptable speed. Given a variety of available image processing techniques and recognition algorithms, we have designed our preliminary process on detecting the image as part of our image processing. Hand detection preprocessing workflow is showed in Fig. 1.

The system starts by capturing a hand image from signer with a webcam setup towards certain angle with black background. The next process will convert the RGB image into grey scale with either black (0) or white (1). The edge of each object is then computed against the black background. The object can then be segmented and differs greatly in contrast to the background images.

B. Preprocessing

Changes in contrast can be detected by operators that calculate the gradient of an image. One way to calculate the gradient of an image is the Sobel operator [7], [8], [9], which creates a binary mask using a user-specified threshold value.

The binary gradient mask shows lines of high contrast in the image. These lines do not quite delineate the outline of the object of interest. Compared to the original image, the gaps in the lines surrounding the object in the gradient mask can be seen.

These linear gaps will disappear if the Sobel image is dilated using linear structuring elements, which applied by strel function. After finding the holes, the 'imfill' function is applied to the image to fill up the holes. Finally, in order to make the segmented object looks natural, smoothing process of the object is applied twice with a diamond structuring element. The diamond structured element is created by using the strel function. Then, bitwise XOR operation sets the resulting bit to 1 if the corresponding bit in binary image or result from dilate image is a 1. Bitwise manipulation enhanced the wanted image of the hand region. Further processing with bounding box approaches may be used to clean up the segmented hand region.

After having a bounding box, the proportion of the white image as compared to the black image inside the bounding box is calculated. If the proportion of white image is changed over the threshold, then the second image will be captured. Currently, the prototype uses only two-sequence sign language.

C. Classification

Feed Forward Neural network, as shown in Fig. 2, is used in the classification process to recognize the various hand gestures. It consists of three layers: input layer, hidden layer and output layer. Input to the system includes various types of two-sequence sign language which have been converted to a column vector by neural network toolbox in MATLAB 7.0. The input signals are propagated in a forward direction on a layer-by-layer basis. Initialize weight is assigned to each neuron. The neuron computes the weighted sum of the input signals and compares the result with a threshold value, θ . If the net input is less than the threshold, the neuron output is a value of -1, otherwise the neuron becomes activated and its output attains a value of +1 instead. Thus, the actual output of the neuron with sigmoid activation function can be represented as

$$Y = \text{sigmoid} \left[\sum_{i=1}^n x_i w_i - \theta \right] \quad (1)$$

In the case of supervise learning, the network is presented with both the input data and the target data called the training set. The network is adjusted based on the comparison of the output and the target values until the outputs almost match the targets, i.e. the error between them are negligible.

However, the dimension of the input for neural network is large and highly correlated (redundant). It can slow down the processing speed. Thus, we used Principle Component Analysis to reduce the dimension of the input vectors.

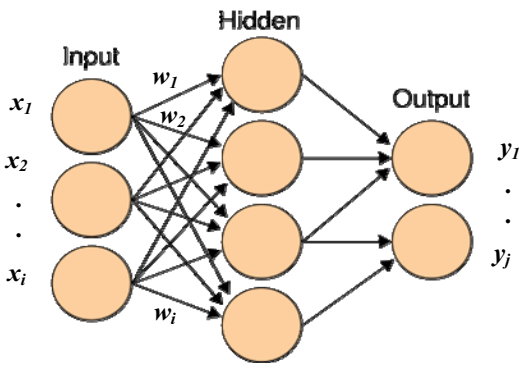


Fig. 2 Feed Forward Neural Network (FFNN)

In MATLAB, an effective procedure for performing this operation is the principal component analysis (PCA). This technique has three effects: the first one is to orthogonalizes the components of the input vectors (so that they are uncorrelated with each other); second it orders the resulting orthogonal components (principal components) so that those with the largest variation come first; and finally it eliminates those components that contribute the least to the variation in the data set. By using this technique, the learning rate of training the neural network is increased.

As a result, the prototype has successfully detected the hand region as shown in Fig. 3.

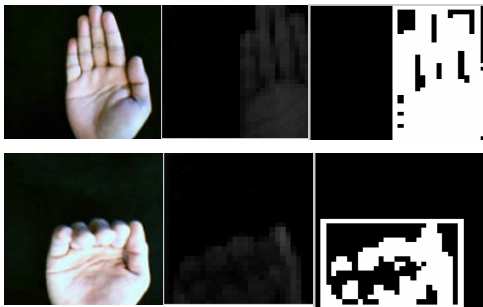


Fig. 3 Hand Image Detection

IV. EXPERIMENTAL RESULTS

A. Neural Network Training Technique

Fig. 4 shows the comparisons of FFNN training without PCA and that of Fig. 5 FFNN training with PCA implemented.

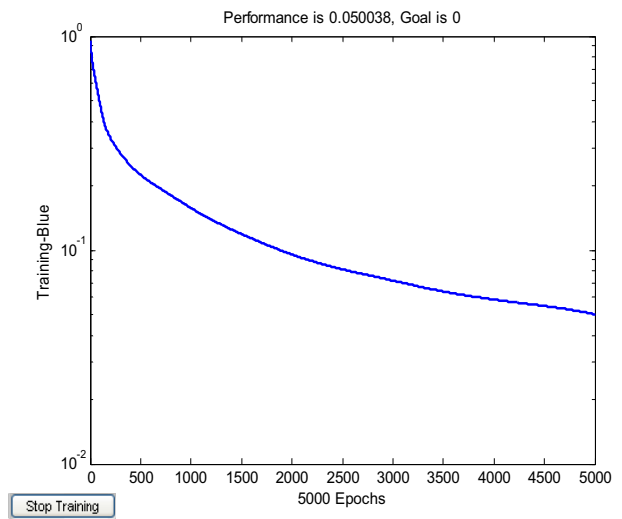


Fig. 4 FFNN Training without PCA

The parameter of neural network training are 0.01 in learning rate, epochs in 5000 and momentum coefficient is 0.8.

Fig. 4 shows the difference-curve of the error rate with difference technique before conduct the training to the neural network. It showed that principle component analysis will increase the learning rate of the neural network.

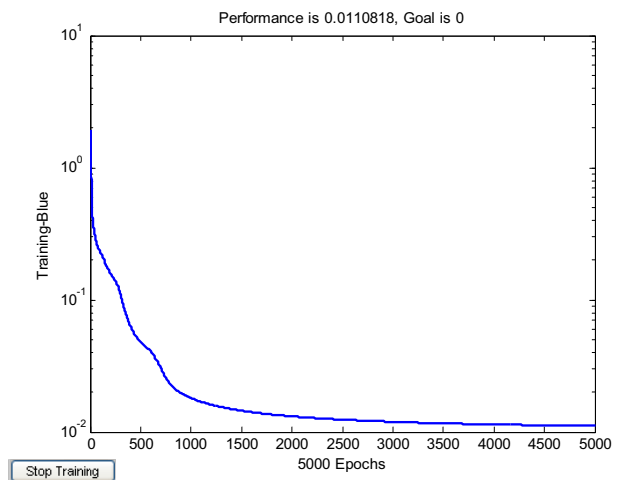


Fig. 5 FFNN Training with PCA
















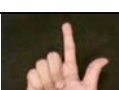
The result of the NN training without PCA is MSE 0.050038/0, Gradient 0.114071/1e-010 whereas NN training with PCA (Fig. 5) is MSE 0.0110818/0, Gradient 0.00490473/1e-010.

B. Sequence Detection Testing

Two types of testing were conducted i.e. positive testing and negative testing. The positive testing is to prove the sequence of sign language that can be recognized by the system. The negative testing is to prove that every time the sign language is not move, the systems will not response

anything to the signer. Table I shows the results of these sequence detection testing.

TABLE I
RESULT OF SEQUENCE DETECTION TESTING

Sequence 1	Sequence 2	(+) Test	(-) Test	Result
		Yes	No	True
		Yes	No	True
		No	Yes	False
		No	Yes	False
		Yes	No	True
		Yes	No	True
		Yes	No	True
		Yes	No	True

The proposed solution is to implement S2V for real-time processing. The system is able to detect the sequence of sign symbols with additional functions that has to be automated to calculate the proportion of the black and white images and compare with threshold value specified by the program. The difficulties that faced here were to recognize a little/small difference in portion of the images which were not detected by the threshold value (and even in recognition part). However, for this prototype, we manage to get the output by implementing the proposed technique to detect the sequence of sign symbols.

C. Recognition Rate

70 set of two-sequence hand gestures were captured in real-time from signers using video camera in which 20 were used as training set and the remaining 10 were used as test set. The recognition rate of the sign languages is calculated as follows:

$$\text{Recognition rate} = \frac{\text{No. of correctly classified signs}}{\text{Total No. of signs}} \times 100\% \quad (2)$$

The overall results of the system prototype were tabulated in Table II below:

TABLE II
S2V SYSTEM RECOGNITION RATE

Data	No. of Samples	Recognized Samples	Recognition Rate (%)
Training	50	40	80.0
Testing	20	15	75.0
Total	70	55	78.6 (Average)

The results of segmentation and feature detection are performed as explained above. Experimental results of the 70 samples of hand images with different positions gave consistent outcomes.

Based on the above experiments, the two-sequence sign language or hand gestures have been tested with an average recognition rate of 78.6%.

V. CONCLUSION

Hand gestures detection and recognition technique for international sign language has been proposed and neural network was employed as a knowledge base for sign language. Recognition of the RGB image and longer dynamic sign sequences is one of the challenges to be look into this technique. The experimental results show that the system prototype S2V has produced satisfactory recognition rate in the automatic translation of sign language to voice. For future research, we propose Hidden Markov Model (HMM) to detect longer sequences in large sign vocabularies and shall integrate this technique into a sign-to-voice system, or vice-versa, to help normal people to communicate more effectively with mute or hearing impaired people.

REFERENCES

- [1] Noor Saliza Mohd Salleh, Jamilin Jais, Lucyantje Mazalan, Roslan Ismail, Salman Yussof, Azhana Ahmad, Adzly Anuar, and Dzulkifli Mohamad, "Sign Language to Voice Recognition: Hand Detection Techniques for Vision-Based Approach," Current Developments in Technology-Assisted Education, FORMATEX 2006, vol. 2, pp.967-972.
- [2] C. Neider, J. Kegel, D. MacLaughlin, B. Bahan, and R.G. Lee, The syntax of American sign language. Cambridge: The MIT Press, 2000.
- [3] M. J. Jones, and J. M. Rehg, "Statistical Color Models with Application to skin Detection," International Journal of Computer Vision, Jan. 2002, vol. 46, no.1, pp. 81-96.
- [4] D. Saxe, and R. Foulds, "Automatic Face and Gesture Recognition," IEEE International Conference on Automatic Face and Gesture Recognition, Sept. 1996, pp. 379-384.
- [5] E.J. Ong, and R. Bowden, "A Boosted Classifier Tree for Hand Shape Detection," Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004), IEEE Computer Society, 2004, pp. 889-894.
- [6] C.S. Lee, J.S. Kim, G. T. Park, W. Jang, and Z.N. Bien, "Implementation of Real-time Recognition System for Continuous Korean Sign Language (KSL) mixed with Korean Manual Alphabet (KMA)," Journal of the Korea Institute of Telematics and Electronics, 1998, vol. 35, no.6, pp. 76-87.

- [7] Gonzalez R., and R. Woods, Digital Image Processing. Addison Wesley, 1992.
- [8] Boyle R., and R. Thomas, Computer Vision: A First Course. Blackwell Scientific Publications, 1988.
- [9] Davies E., Machine Vision: Theory, Algorithms and Practicalities. Academic Press, 1990.
- [10] X.L. Teng, B. Wu, W.Yu, and C.Q. Liu, "A Hand Gesture Recognition System Based on Local Linear Embedding," Journal of Visual Languages and Computing, vol. 16, Elsevier Ltd., 2005, pp. 442 – 454.
- [11] W.W. Kong, and S.Ranganath, "Signing Exact English (SEE): Modeling and Recognition," The Journal of Pattern Recognition Society, vol. 41, Elsevier Ltd., 2008, pp. 1638 -1652.
- [12] Y.H.Lee, and C.Y. Tsai, "Taiwan Sign Language (TSL) Recognition based on 3D Data and Neural Networks," Expert Systems with Applications, Elsevier Ltd., 2007, pp. 1-6.