A SIGN LANGUAGE RECOGNITION APPROACH FOR HUMAN-ROBOT SYMBIOSIS

Afzal Hossian[1], Shahrin Chowdhury[2], and Asma-ull-Hosna[3]

[1]IIT (Institute of Information Technology), University of Dhaka, Bangladesh
[2]Chalmers University of Technology, Gothenburg, Sweden
[3]Sookmyung Women's University, Seoul, South-Korea

ABSTRACT

This paper introduces a new concept for the establishment of human-robot symbiotic relationship. The system is based on the implementation of knowledge-based image processing methodologies for model based vision and intelligent task scheduling for an autonomous social robot. This paper aims to develop an automatic translation of static gestures of alphabets and signs in American Sign Language (ASL), using neural network with backpropagation algorithm. System deals with images of bare hands to achieve the recognition task. For each individual sign 10 sample images have been considered, which means in total300 samples have been processed. In order to compare between the training set of signs and the considered sample images, are converted into feature vectors. Experimental results reveal that this can recognize selected ASL signs (accuracy of 92.00%). Finally, the system has been implemented issuing hand gesture commands for ASL to a robot car, named "Moto-robo".

KEYWORDS

American Sign Language, Histogram equalization, Human-robot symbiosis, Moto-Robo, Skin colour segmentation.

1. INTRODUCTION

In the dictionary of American Cultural Heritage the word "symbiosis" is defined as following: "A close, prolonged association among two or more different organisms of different species that may but does not necessarily benefit each member" [1]. In recent times this biological term has been used to define similar relations among wider collection of entities. In this research, the main purpose is to establish a symbiotic relationship between robots and human beings for their coexistence and co-operative work and consolidate their relationship for the benefit of each other. Image understanding concerns the issues of finding interpretations of images. These interpretations would explain the meaning of the contents of the images. In order to establish a human-robot symbiotic society, different kinds of objects are being interpreted using the visual, geometrical and knowledge-based approaches. When the robots are working cooperatively with human beings, it is necessary to share and exchange their ideas and thoughts. Human hand gesture is, therefore, immerging tremendous interest in the advancement of human-robot interface since it provides a natural and efficient way of exploring expressions.

The sign language is the fundamental communication method between people who suffer from hearing defects. In order for an ordinary person to communicate with deaf people, a translator is usually needed to translate sign language into natural language and vice versa [2]. As a primary

component of many sign languages and in particular the American Sign Language (ASL), hand gestures and finger-spelling language plays an important role in deaf learning and their communication. Therefore, sign language can be considered as a collection of gestures, movements, postures, and facial expressions corresponding to letters and words in natural languages.

American Sign Language (ASL) is considered to be a complete language which includes signs using hands, other gesture with the support of facial expression and postures of the body [2]. ASL follows different grammar pattern compare to any other normal languages. Near about 6000 gestures of common words are represented using finger spelling by ASL. 26 individual alphabets are signified by 26 different gesture with the use of single hand. These 26 alphabets of ASL are presented in Fig. 1.

Charayaphan and Marble [3] investigated a way using image processing to understand ASL. Out of 31 ASL symbols 27 can correctly recognize by their suggested system. Fels and Hinton [4] have developed a system. VPL DataGlove Mark II along with a Polhemus tracker was used as input devices in their developed system. For categorized hand gestures neural network method was applied. For the input of HMMs, two-dimensional features of a solo camera along with view-based approach were applied by Starner and Pentland [5]. Using HMMs and considering 262-sign vocabulary, 91.3% accuracy was achieved for recognized isolated signs by Grobel and Assan [6]. While collecting sample features from the video recordings of users, they were using colour gloves. Bowden and Sarhadi [7] developed a non-linear model of shape and motion for tracking finger spelt American Sign Language. This approach is similar to HMM where ASL's are projected into shape space to guess the models and also to follow them.

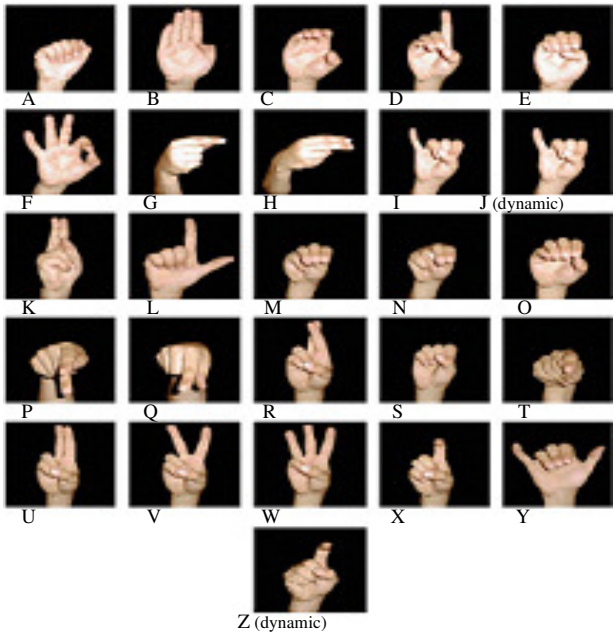

Figure 1. Alphabets of American Sign Language

This system is capable of visually detecting all static signs of the American Sign Language (ASL): alphabets, numerical digits as well as general words for example: like, love, not agreed

etc. can also be represent using ASL. Fortunately the users can interact using his/her fingers only; there is no need to use additional gloves or any other devices. Still, variation of hand shapes and operational habit leads to recognition difficulties. Therefore, we realized the necessity to investigate the signer independent sign language recognition to improve the system robustness and practicability. Since the system is based on Affine transformation, our method relies on presenting the gesture as a feature vector that is translation, scale and rotation invariant.

2. SYSTEM DESIGN

The ASL recognition system has two phases: the feature extraction phase and the classification phase, as shown in Fig. 2.

The image samples are resized and then converted from RGB to YIQ colour model. Afterwards the images are segmented to detect and digitize the sign image.

In the classification stage, a 3-layer, feed-forward backpropagation neural network is constructed. It consists

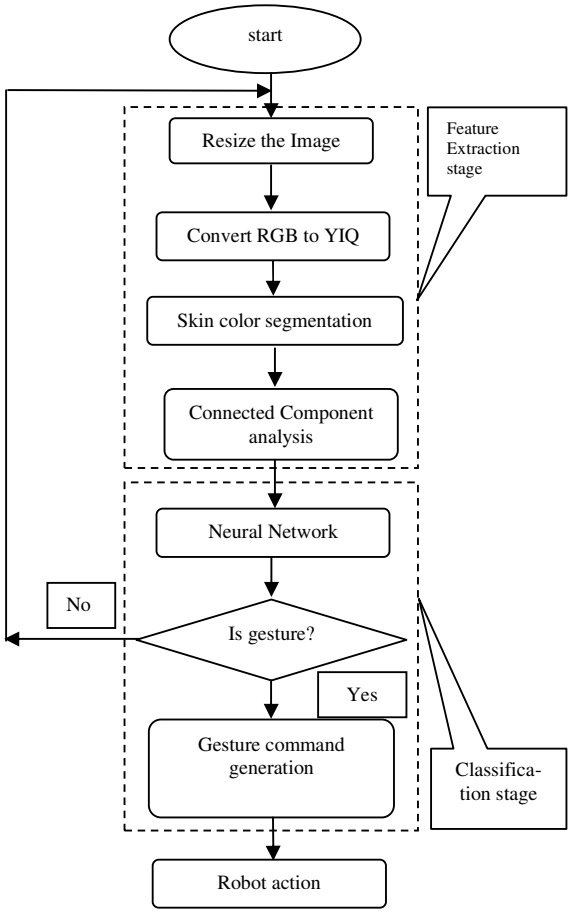

Figure 2. System overview

of (40×30) neurons in the input layer, 768 (70% of input) neurons for the hidden layer, and 30 (total number of ASL image for the classification network) for the neurons in the output layer.

2.1 Features to analyse images

Normalization of sample images, equalization of histogram, image filtering, and skin colour segmentation are highlighted in this phase.

2.1.1 Normalization of sample images

A low pass filter is used in order to reduce aliasing the nearest neighbour interpolation method, to find out the values of pixels in the output image where images are resized to 160 by 120.

2.1.2. Equalization of Histogram

Equalization of Histogram is used to improve the lighting conditions and the contrast of image as the hand images contrast depends on the lighting condition. Let the histogram $h(r_i) = \frac{p_i}{n}$ of a digital hand image consists of the colour bins in the range $[0, C-1]$, where r_i is the *I* th colour bin, p_i is the number of pixels in the image with that colour bin and *n* is the total number of pixels in the image.

Some scaling constant are calculated using the cumulative sum of bins for any interval [0,1] of r [8]. For individual pixel value *r* in the original images of level s and $0 \le T(r) \le 1 \ for \ 0 \le r \le 1.$ is used to yield the mapping to perform the function $s = T(r),$ of transforming by allowing the range. The histogram equalization process is illustrated in Fig. 3.

2.1.3. Image Filtering

Prewitt filter provides the advantage of suppressing the noise which collected from various sources without erasing some of the image details like low-pass filter.

2.1.4. Skin colour segmentation

Skin colour segmentation is based on visual information of the human skin colours from the image sequences in YIQ colour space. The image samples are converted from RGB to YIQ colour model. To check, the amount of skin colour value to identify the specific colour that have dominance over the image by searching in YIQ space.

In the following matrix, luminance channel and two chrominance channels are represented with Y and (I,Q) respectively where linear transformation of RGB is produced from YIQ.
Luminance, hue and saturation these three attributes are described using YIQ colour model [8]:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.25 & 0.587 & 0.49 \\ 0.45 & -0.384 & -0.320 \\ 0.212 & -0.639 & 0.79 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad (1)$$

Here red, green, and blue component values are denoted with *R*, *G*, *B between the range of* [0,255].

Since the human skin colours are clustered in colour space and differ from person to person and of races, so in order to detect the hand parts in an image, the skin pixels are thresholded empirically [9],[10].

The threshold value is calculated using following equation:

$$(60 < Y < 200) \; and \; (20 < I < 50) \qquad (2)$$

The detection of hand region boundaries by such a YIQ segmentation process is illustrated in Fig. 4.

The exact location of the hand is then determined from the image with largest connected region of skin-coloured pixels. For uneven segment image detection of connected components, the Region-growing algorithm is applied.

In this experiment, 8-pixel neighbourhood connectivity is employed. In order to remove the false regions from the isolated blocks, smaller connected regions are assigned by the values of the background pixels

2.2 Classification phase

The classification phase includes neural network training for the recognition of binary image patterns of the hand. In the neural networks the result will be not perfect. Sometimes practice represents the best solution. Decision in this field is very difficult; we had to examine different architectures and decide according to their results

Figure 3. Histogram equalization

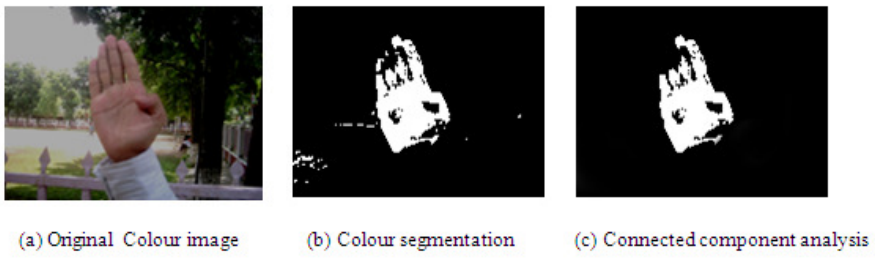

Figure 4. Skin colour segmentation

Therefore, after several experiments, it has been decided that the proposed system should be based on supervised learning in which the learning rule is provided with the set of examples (the training set). When the parameters, weights and biases of the network are initialized, the network is ready for training. The multi-layer perceptron, as shown in Fig. 5, with backpropagation algorithm has been employed for this research.

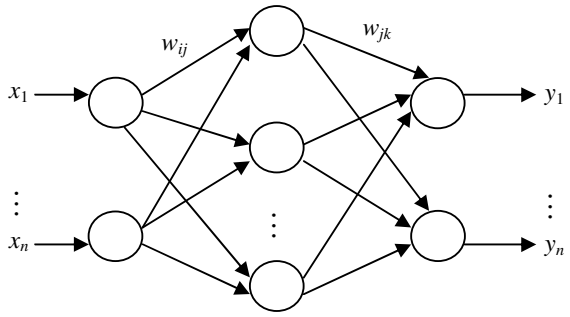

Figure 5. Network architecture of BPNN

The number of epochs was 10,000 and the goal was 0.0001. The back-propagation training algorithm is given below.

Step 1 Initial Phase

To ensure the uniform distribution, the random numbers are generated using weight and threshold from network levels

$\left[-\frac{2.4}{F_i},+\frac{2.4}{F_i}\right]$, where F_i is the total number of inputs of neurons I in the network.

Step 2 Active Phase

Back-propagation neural network is activated to get desire yields $y_{d,1}(t), y_{d,2}(t),...,y_{d,n}(t)$. by applying inputs $x_1(t), x_2(t),...,x_n(t)$.

(a) The hidden layer of authentic output of neurons, is calculated using below function:

$$y_j(t) = sigmoid\left[\sum_{i=1}^{n} x_i(t) \times w_{ij}(t) - \theta_j\right], \qquad (3)$$

where n is the number of inputs of neuron j in the hidden layer, and sigmoid is the sigmoid activation function.

(b) The output layer of authentic outputs of the neurons, is calculated using below function:

$$y_k(t) = sigmoid\left[\sum_{j=1}^{m} y_j(t) \times w_{jk}(t) - \theta_k\right], \qquad (4)$$

where m is the number of inputs of neuron k in the hidden layer.

Step 3 Training of Weight

The following equation is used to propagate errors related with output neurons to update the weights in the back-propagation network:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p), \quad (5)$$

where $\Delta w_{ij}(p) = \alpha y_j(p)\delta_k(p)$, (6)

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p), \quad (7)$$

where $\Delta w_{ij}(p) = \alpha x_j(p)\delta_j(p)$, (8)

where error, $e_k(p) = y_{dk}(p) - y_k(p)$, (9)

error gradient for neuron in the output layer is:

$$\delta_k(p) = y_k(p)[1 - y_k(p)]e_k(p), \quad (10)$$

and $\delta_j(p) = y_j(p)[1 - y_j(p)] \sum_{k=1}^{l} \delta_k(p) w_{jk}(p)$. (11)

Step 4 Iteration
Increase iteration *t* by one, go back to **Step 2** and repeat the process until the error value reduces to the desired level. A complete flow chart of our proposed network is shown in Fig. 6.

3. EXPERIMENTS RESULTS & PERFORMANCE

The performance and effectiveness of the system has been justified using different hand expressions and issuing commands to a robot named "Moto-Robo". The computer configuration for this experiment was Pentium IV 1.2 GHz PC along with 512 MB RAM. Visual C++ was used as the programming language to implement the algorithm.

3.1. Interfacing the robot

The communication link between the computer and the robot has been established by means of parallel communication port. The parallel port is a 25 pin D-shaped female (DB25) connector equipped in the back of the computer. The pin configuration of DB25 connector is shown in the Fig. 7. The lines in DB25 connector are divided in to three groups: Status lines, Data lines and Control lines. As the name refers, data is transferred over data lines, Control lines are used to control the peripheral and of course, the peripheral returns status signals back computer through Status lines.

In order to access parallel ports by the programmers some library functions are used.

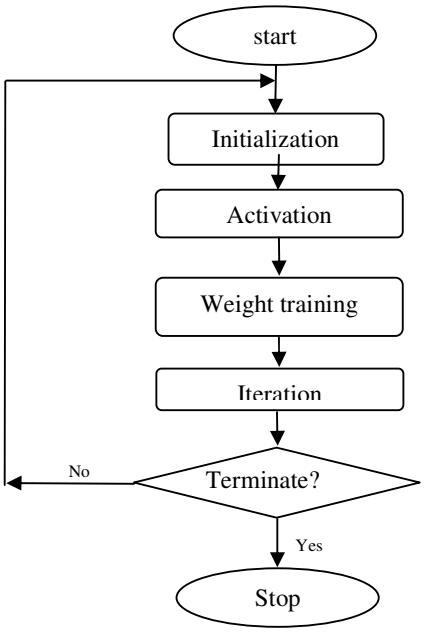

Figure 6. Flow chart of BPNN

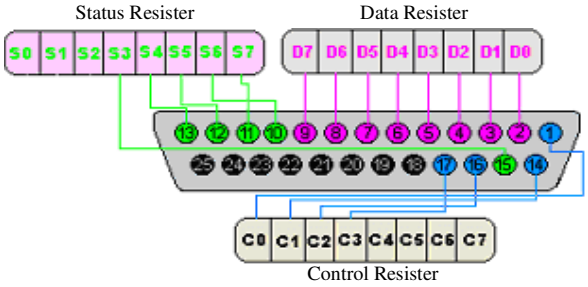

Figure 7. Pin configuration of DB25 port

Visual C++ provides two functions to access IO mapped peripherals, 'inp' for reading data from the port and 'outp' for writing data into the port.

3.2 Analysis of Experiments

At first to determine the testing ability of the recognition system of hand gesture, to classify signs for both training and testing set of data where the quantity of inputs influences the neural network. Few of the signs have resemblances between them which lead to create some problems in the performance.

In this experiment, the binary images are used to recognize the system using training and testing data set. Also 10 samples for each sign were taken from 10 different volunteers. For each sign, 5 out of 10 samples were used for training purpose, while the remaining 5 signs were used for testing. Various orientations and distances is considered while collecting sample images using

digital camera. This way, we were able to obtain a data set with cases that had different sizes and orientations, so we could examine the capabilities of our feature extraction scheme.
Performance evaluation of the system depends on its capability of correctly categorizing samples related to their classes. The ratio of correctly categorizing samples and total amount of sample is denoted as recognition ratio, i.e.

$$\text{Recognition rate} = \frac{\text{Number of correctly categorized sign}}{\text{Total amount of signs}} \times 100\% \quad (12)$$

In backpropagation learning algorithm modification in weights considered for a number of periods results to continuous decrement in Training curve is represented in Fig. 8

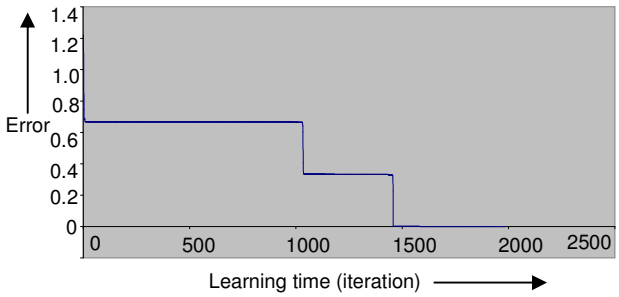

Figure 8. Error versus iteration for training the BPNN

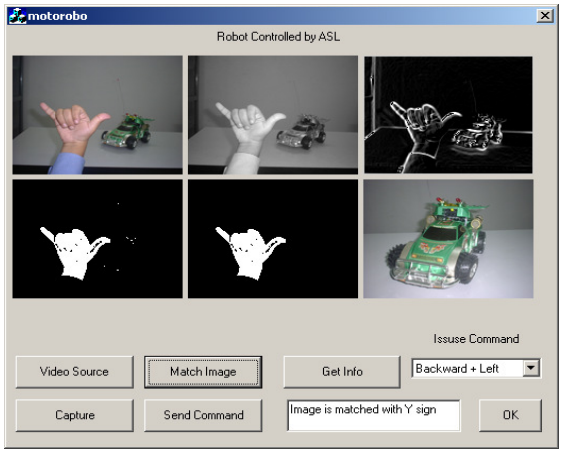

Figure 9. Program interface for robot control

ASL Sign	Action
F	Forward
B	Backward
R	Turn Right
L	Turn Left
O	Forward + right
D	Forward + left
V	Backward + right
Y	Backward + left
S	Stop
M	Load or Hide Missile Chamber
C	Fire Missile
W	Turn on Light
E	Turn of Light

Table 1 Command to control Moto-Robo

3.3 Implementation

A remote control car (Moto-Robo), connected to the pc through the parallel port, has been controlled by means of commands directed by the hand gesture of the user. The car has several movements, such as: Forward, Backward, Turn right, Turn Left, Turn Light on, Turn Light off and so on depending on the sign languages F, B, R, L, W, E, respectively. Some of the ASL employed for controlling the robot is listed in Table 1.

The system was tested with (300) images, (ten images for each sign) untrained images; previously unseen for the testing phase. In order to determining the yields in a suitable way a GUI has been created by us. An example is shown in the Fig 9, where one of the actions of the robot as a result of hand gesture recognition process is shown.

4. CONCLUSION

This research presents the development of a system for the recognition of American Sign Language. On recognition of different hand gestures, a real time robot interaction system has been implemented. For individual image pattern related to the set of training a set of input data for accomplishing the work. Without the need of any gloves, images for different signs were captured by digital camera. Deviation in position, direction, size and gesture are proved to be easily adapted by the developed system. This is because the extracted features method used Affine transformation to make the system translation, scaling and rotation invariant. The recognition rate for training data and testing data are 92.0% and 80% respectively for the future system.

The work presented in this research deals with static signs of ASL only. Adaptation of dynamic signs can be an interesting thing to watch in future. There is a limitation in the existing system that it only deals with images that have a non-skin colour background, overcoming this limitation can make the system more compatible in real life. Beside hand images other types of images for example eye tracking, facial expression, head gesture etc. can also be considered as sample images for the network to analyse. The goal is to create a symbiotic environment in order to give the opportunity to the robots to exchange their ideas with human beings which will definitely bring benefits for both and also have an positive impact on the society.

REFERENCE

[1] M. A. Bhuiyan and H. Ueno,(2003) "Image Understanding for Human -robot Symbiosis", 6th ICCIT, Dhaka, Bangladesh, Vol. 2, pp. 782-787.

[2] International Bibliography of Sign Language, (2005),[online]. Available: http://www.signlang.unihamburg.de/bibweb/FJournals.html

[3] C. Charayaphan and A. Marble, (1992) "Image processing system for interpreting motion in American sign language", Journal of Biomedical Engineering, Vol. 14, pp. 419–425.

[4] S. Fels and G. Hinton, (1993) "GloveTalk: a neural network interface between a DataGlove and a speech synthesizer", IEEE Transactions on Neural Networks, Vol. 4, pp. 2–8.

[5] T. Starner and A. Pentland,(1995) "Visual recognition of American sign language using hidden Markov models", International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 189–194.

[6] K. Grobel and M. Assan, (1996) "Isolated sign language recognition using hidden Markov models. In Proceedings of the international conference of system, man and cybernetics", pp. 162–167.

[7] R. Bowden and M. Sarhadi, (2002) "A non-linear model of shape and motion for tracking finger spelt American sign language", Image and Vision Computing, Vol. 9–10, pp. 597–607.

[8] R. C. Gonzalez and R. E. Woods, (2003) "Digital Image Processing", Pearson Education Inc., 2nd Edition, Delhi.

[9] M. A. Bhuiyan, V. Ampornaramveth, S. Muto, and H. Ueno,(2003) "Face Detection and Facial Feature Localization for Human-machine Interface", NII Journal, Vol. 5, No. 1, pp. 26-39.

[10] M. A. Bhuiyan, V. Ampornaramveth, S. Muto, and H. Ueno,(2004) " On Tracking of Eye for Human-Robot Interface", International Journal of Robotics and Automation, Vol. 19, No. 1, pp. 42-54.