Recognition of ASL for Human-robot Interaction

Md. Al-Amin Bhuiyan

College of Computer Sciences & Information Technology, King Faisal University, Al Ahsa, Saudi Arabia.

Summary

This article addresses on the development of a framework of American Sign Language for human-robot interaction. The approach is based on the computer vision strategies and intelligent task scheduling for an entertainment robot. The system is organized with the identification of signs in American Sign Language (ASL) using adaptive resonance theory neural network. Experimental results indicate that the method is capable of identifying the ASL signs with an accuracy of more than 90% under different illumination conditions.

Key words:

*American Sign Language, Histogram equalization, Human-robot interaction, Aibo, ART Neural network, Skin color segmentation***.**

1. Introduction

ASL is a visual gesture language that works as the principal way of expressing thoughts of deaf communities. Over the last few decided, ASL has occupied a position of overwhelming dominance in research in the fields of image processing, computer vision, pattern recognition and so on. Since it offers a natural and effective way of exploring expressions, ASL is, therefore, providing remarkable interest in the advancement of human-robot interaction. ASL involves about 6000 gestures of common words with finger spelling used to communicate indistinct words or proper nouns. Finger spelling uses one hand and 26 gestures to communicate the 26 letters of the alphabet. The 26 alphabets of ASL are shown in **Fig. 1**.

Numerous research papers have been published on ASL [1-5]. Vogler and Metaxas trained context-dependent Hidden Marcov Models and modeling transient movements between signs inspired by the characteristics of ASL phonology. They employed computer vision techniques for 3D object shape and motion parameter extraction to achieve accurate 3D movement parameters of ASL sentences [6]. Fok et al [7] developed a framework of a real-time multi-sensor recognition system for ASL. They collected data from leap motion sensors and fused them for recognition employing hidden Markov models. Charayaphan and Marble [8] proposed vision techniques to recognize ASL. Fels and Hinton [9] proposed a neural network interface between a VPL data-glove and a speech synthesizer. Their method used backpropagation neural network for mapping hand movements. Starner and Pentland [10] employed a Hidden Marcov model to extract the features for recognizing isolated signs of ASL. Grobel and Assan [11] employed Hidden Marcov model and was able to identify ASL signs with 91.3% accuracy. They pulled out the features from video recordings of signers wearing colored gloves. Bowden and Sarhadi [12] used a nonlinear point distribution model, allowing a much richer description of hand shape, and augment this with a Markov model for American Sign Language. Their method is based on one-state transitions of the English Language, which are projected into shape space for tracking. Munib et al [13] employed Hough transform and neural networks to recognize ASL signs. Zhang et al. [14] propoed an ASL recognition method depending on action recognition strategies to identify the ASL signs. They designed a corpus for analyzing the red, green and blue components of color model and aggregated multiple signal modalities for depth videos of different sentences of ASL.

This system is organized to identify all the signs of the ASL. Users do not require to wear any gloves to interact with the system. The signers vary their hand shapes only. The system is based on Affine transformation [15], i.e., scaling, translation and rotation on presenting the gesture.

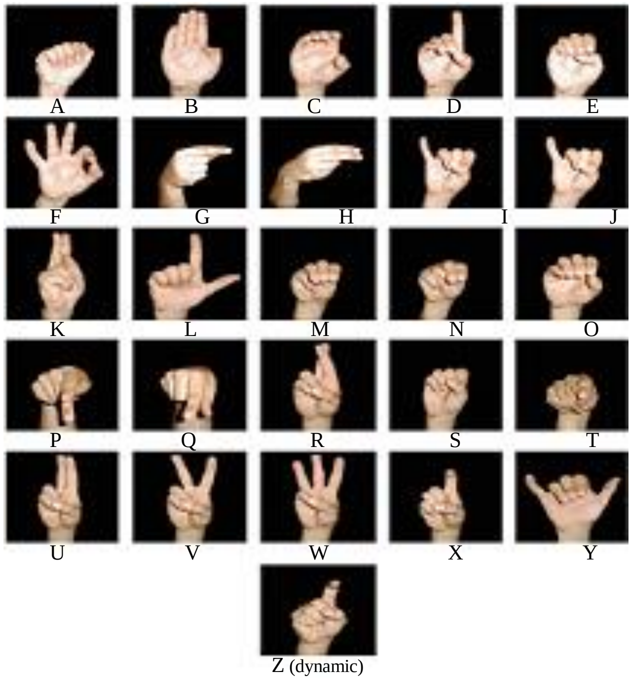


Fig. 1 American Sign Language of alphabets.

Manuscript received May 17, 2017

Manuscript revised June 29, 2017

The remaining part of the article is organized as follows. Section 2 describes the System design and architecture. The necessary image processing and image segmentation process is explained in details in different subsections. Section 3 describes the experimental results and performance of the system. Finally, Section 4 draws the overall conclusions of this research and proposes the future directions.

2. System Design and Architecture

The ASL recognition system comprises with two segments: (i) the feature extraction and (ii) identification of signs, as illustrated in **Fig. 2**. The feature extraction process is initiated with an image processing procedure, which involves an algorithm to detect and segment various desired segments of the sign. For this, each color image is resized and converted from RGB to HSV color space. After applying the skin color segmentation process a binary image is generated from which the largest connected component is being analyzed, converted into a feature vector that is compared to the feature vectors of a training set of signs. The feature extraction process consists of image standardization, image enhancement, filtering, and skin color segmentation.

2.1 Image Standardization

Images of signs were resized to 30×24, by default uses nearest neighbor interpolation to determine the values of pixels in the output image. This research employs a lowpass filter before interpolation to reduce aliasing.

2.2 Fuzzy Histogram Equalization

The hand images may be of poor contrast because of the limitations of the illumination conditions. Therefore, fuzzy histogram equalization is used to reimburse for the illumination conditions and recover the contrast of the image, as shown in **Fig. 3**.

Fuzzy histogram equalization technique is based on the fuzzy normalized histogram of the image. The equalized fuzzy histogram is defined:

$$fhe(r_k) = \sum_{j=0}^{k} fp_r(r_j) = \sum_{j=0}^{k} \frac{\left| j^{th} \text{ gray level} \right|}{\left| \text{gray level} \right|}, \quad (1)$$

where $k = 0,1,2,..L-1$.

The algorithm to improve the contrast of the image using fuzzy histogram equalization technique is as follows:

FHISEQ()

This algorithm increase contrast of image using classical histogram equalization technique. Let the maximum gray level is L, r_k is k th gray level over [0,L-1] and s_k is the transformed gray level of r_k.

Step1. Construct the fuzzy image of the given image by using the following membership function:

$$\mu_k = {r_k}/{L} \quad (2)$$

Step2. Calculate the fuzzy histogram of the constructed fuzzy image in step 1.

Step3. Calculate the fuzzy equalized histogram.

Step4. Transform the gray level of each pixel as follows:

$$s_k = L * fhe(r_k) \quad (3)$$

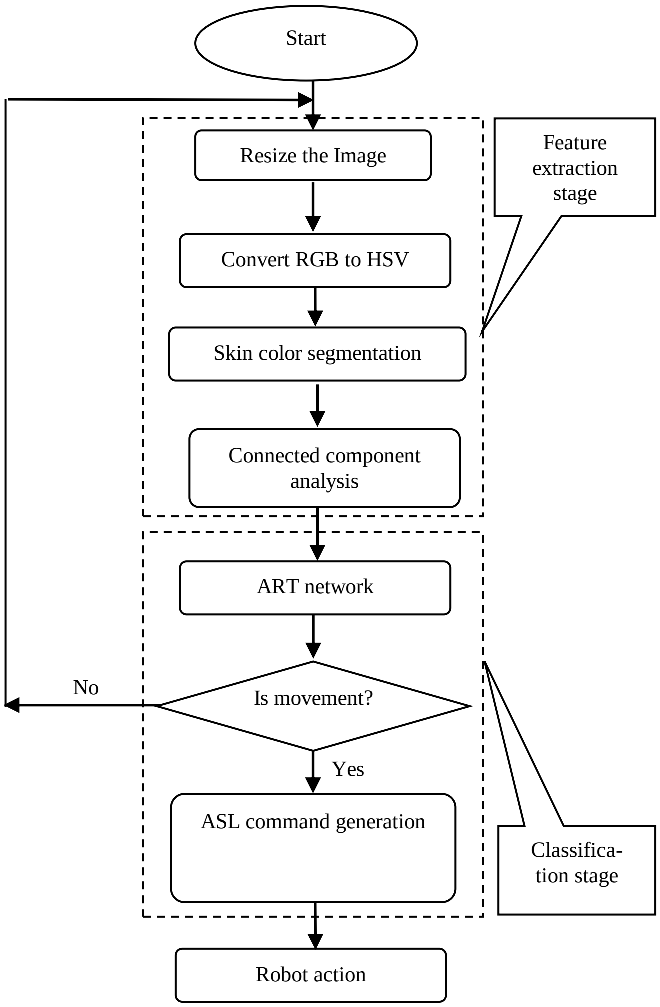


Fig. 2 System Architecture.

The fuzzy expectation of a discrete random variable x having values $x_1, x_2, x_3, ... x_n$ with respective fuzzy probabilities $P(x_1), P(x_2), P(x_3), ... P(x_n)$ is defined by

$$FE(x) = \sum_{i=0}^{n} x_i P(x_i), \quad \text{where} \quad \sum_{i=0}^{n} P(x_i) = 1. \tag{4}$$

2.3 Filtering

The ASL images are sometimes corrupted by numerous sources of noise. Therefore, Prewitt filtering is used to suppress the noise.

2.4 Skin color segmentation

Skin color segmentation is organized with visual information of the hand skin colors extracted from different images. This research uses HSV color space for skin color segmentation.

In the HSV color space a color is described by three attributes: hue, saturation and value. Hue is the visual attribute of color sensation linked with the dominant colors, saturation implies the relative purity of the color content and value measures the brightness of a color. The transformation from RGB to HSV is given by the equations [16-20]:

$$\mathrm{H}_1 = \cos^{-1}\left\{ \frac{\frac{1}{2}\left[(\mathrm{R}-\mathrm{G}) + (\mathrm{R}-\mathrm{B})\right]}{\sqrt{(\mathrm{R}-\mathrm{G})^2 + (\mathrm{R}-\mathrm{B})(\mathrm{G}-\mathrm{B})}} \right\} \tag{5}$$

ranging $[0, 2\pi]$, where $H = H_1$ if $B \le G$; otherwise $H = 360^\circ - H_1$;

$$\mathrm{S} = \frac{\max(\mathrm{R},\mathrm{G},\mathrm{B}) - \min(\mathrm{R},\mathrm{G},\mathrm{B})}{\max(\mathrm{R},\mathrm{G},\mathrm{B})}, \qquad \mathrm{V} = \frac{\max(\mathrm{R},\mathrm{G},\mathrm{B})}{255}, \tag{6}$$

where *R,G,B* are the red, green, and blue component values which exist in the range [0,255].

Since the human skin colors are clustered in color 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 [21-23]. In this research, the hue values are chosen *h*= [0, 40].

The detection of hand region boundaries by such an HSV 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-colored pixels.

2.5 Classification phase

The classification phase involves neural network training for the recognition of binary image patterns of the hand. In the classification stage, an Adaptive Resonance Theory (ART) neural network is employed. The ART contains 30×24 neurons in the input layer, 604 (70% of input) neurons in the hidden layer, and 26 neurons in the output layer. The algorithm for the ART network is given bellow.

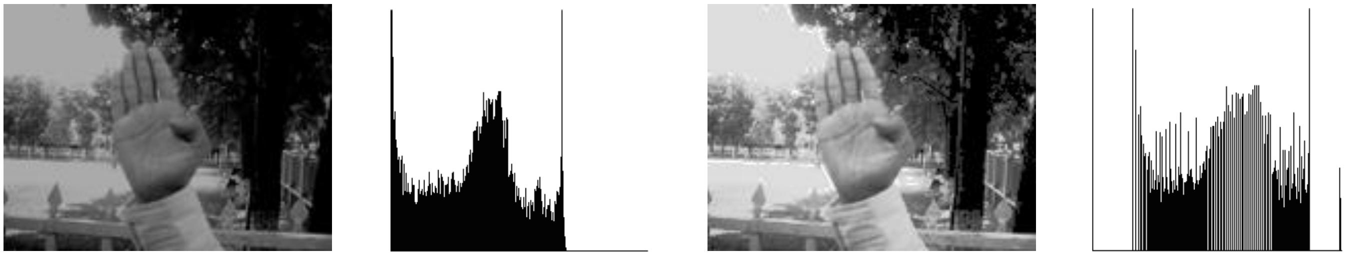
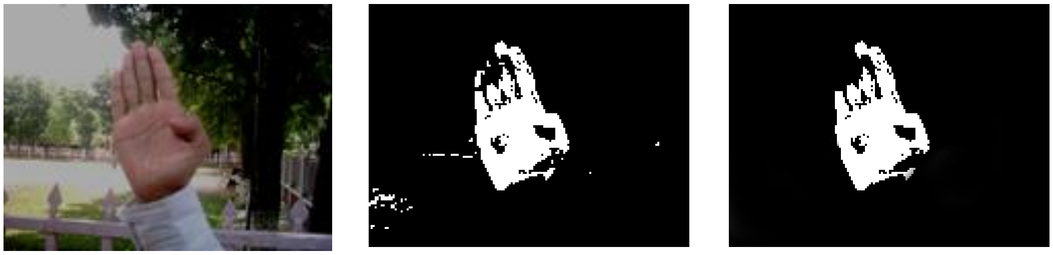


Fig. 3 Fuzzy Histogram equalization.



(a) Original Color image (b) Color segmentation (c) Connected component analysis

Fig. 4 Skin color segmentation.

ART1 Algorithm

```
Step 0 : Initialize parameters and weights [16]:
    $L > 1, 0 < \rho \leq 1, 0 < bij(0) < \frac{L}{L-1+n}, tji(0) = 1.$
Step 1: While stopping condition is false do Steps 2-13
Step 2: For each training input, do steps 3-12
Step 3: Set activations of all F2 units to zero.
        Set activations of F1(a) units to input vectors.
Step 4: Compute the norm of s: $\|s\| = \sum_i s_i$
Step 5: Send input signal from F1(a) to the F1(b) layer:
        $x_i = s_i$ .
Step 6: For each F2 node that is not inhibited:
        If $y_j \neq -1$ then $y_j = \sum_i b_{ij} x_i$ .
Step 7: While reset is true, do Steps 8-11.
Step 8: Find J such that $y_J > y_j$ for all nodes j. If $y_J$
        then all nodes are inhibited and this pattern cannot
        be clustered.
Step 9: Recompute activation x of F1(b): $x_i = s_i t_{Ji}$
Step 10: Compute the norm of vector x: $\|x\| = \sum_i x_i$
Step 11: Test for reset: if $\frac{\|x\|}{\|s\|} < \rho$ then $y_J = -1$ (inhibit
        node J): if $\frac{\|x\|}{\|s\|} \geq$ then proceed to Step 12.
Step 12: Update the weight for node J:
        $bij(new) = \frac{Lxi}{L-1+\|x\|}, tJi(new) = xi$
Step 13: Test for stopping condition.
```

3. Experimental Results and Performance

The effectiveness of the system has been justified using different hand movements and issuing commands to an entertainment robot named “Aibo”. Experiments were carried out with an IntelI Core™ i5-CPU@2.70 GHz with 4 GB RAM. The algorithm has been implemented using Visual C++. The data set employed for training and testing the ASL sign recognition system consists of color images. Five samples for each ASL sign were taken from five different users. For each sign, three out of five samples were employed for training purpose, while the remaining two signs were used for testing. The samples were captured from different distances by digital camera, and with different illumination conditions and alignments.

The performance of the system has been evaluated with two parameters like precision and recall. Precision is the fraction of the correct retrieved ASL images among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over total relevant instances in the image. Both precision and recall are therefore based on an understanding and measure of relevance. The precision and recall are expressed by the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (7)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8)$$

where TP is the true positive, FP the false positive, FN is the false negative, respectively. The precision and recall results for different ASL signs are shown in **Fig. 5**. Obviously, the ASL for the characters C, E, M, N, O, P, Q and S performed poor results due to their confusing recognition characteristics.

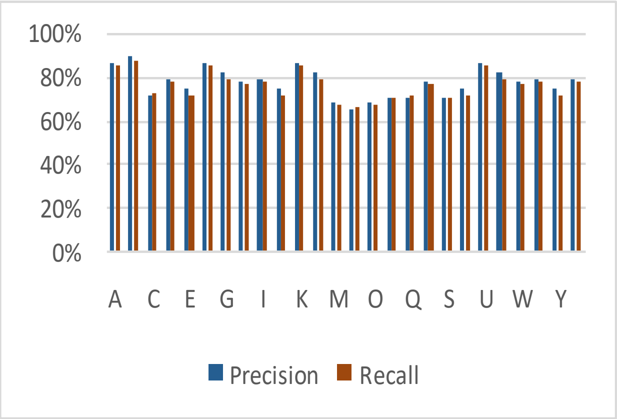


Fig. 5 Precision and recall results for different ASL signs.

The error versus iteration graph for the ART training process is shown in **Fig. 6**. Obviously, as the weights are updated the error is reducing towards zero. Finally, an entertainment robot “Aibo” is being controlled by means of commands directed by the ASL signs. The snapshot for the interface between ASL and “AIBO” robot is shown in **Fig. 7**. The popular command corresponding to ASL y, that is move backward and left is illustrated in this figure. The robot has several movements, such as: Move Forward, Move Backward, Turn left, Turn Right, Light on, Light off and so on depending on the ASL sign languages F, B, L, R, W, E, respectively. Some of the ASL signs employed for controlling the robot are listed in Table 1. The entertainment robot, “AIBO” could follow the instructions successfully.

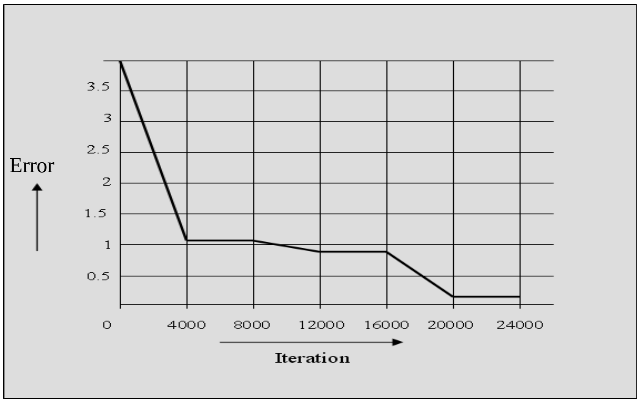
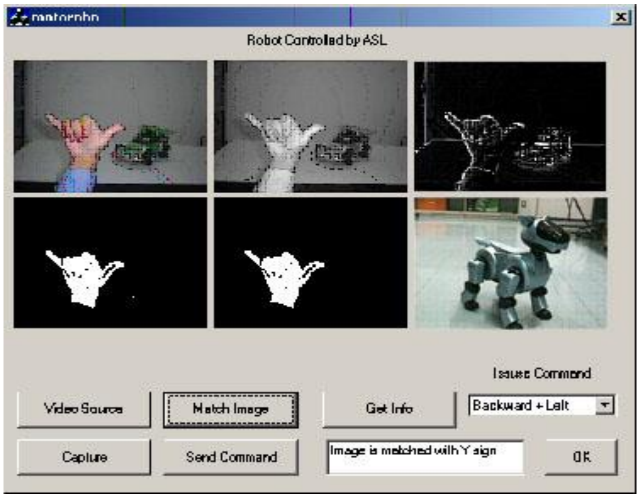


Fig. 6 Error versus iteration for training the ART.



Issue Command: Backward + Left

Image is matched with Y sign

Fig. 7 Snapshot for ASL controlled Robot movement.

Table 1: Commands to control “AIBO”

ASL Sign	*Robot Action*
F	Move Forward
B	Move Backward
L	Turn Left
R	Turn Right
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 off Light

4. Conclusions

This article presents implementation of ASL sign recognition to control an entertainment robot. The work has been accomplished by training a number of hand images with different expressions for ASL. The system is capable of performing the identification of pattern without the need of any hand gloves. Images were captured at different illumination conditions and the proposed fuzzy histogram equalization algorithm was capable of equalizing the images properly. The developed system proved to be robust against changes in gestures, position, size and direction. This is because the extracted features method using Affine transformation to make the system translation, scaling and rotation invariant. The proposed system was able to reach a recognition rate of more than 92.0% for training data and about 85% for testing data. Although this research was involved with static signs of ASL, but nevertheless, the dynamic signs can also be detected for consecutive image sequences through proper mapping of the feature vectors. Our future plan is to include images with non-skin color background and make the system suitable for real life applications, such as eye tracking, gaze direction, facial expression, and gesture recognition.

References

[1] 1. C. Chuan, E. Regina, and C. Guardino, " American Sign Language Recognition Using Leap Motion Sensor”, 13th International Conference on Machine Learning and Applications (ICMLA), Detroit, MI, USA, pp. 541-544, 2014.

[2] C. Savur and F. Sahin, "Real-Time American Sign Language Recognition System by Using Surface EMG Signal", 14th IEEE International Conference on Machine Learning and Applications, pp. 497-502, 2015.

[3] P. Pandey and V. Jain, "Hand Gesture Recognition for Sign Language Recognition: A Review", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, No. 3, pp. 464-470, 2015.

[4] N. Sarawate, M. Chan, and C. OZ, "A real-time American Sign Language word recognition system based on neural networks and a probabilistic model", Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 23, pp. 2107-2123, 2015.

[5] Md. Mohiminul Islam ; Sarah Siddiqua ; Jawata Afnan, "Real time Hand Gesture Recognition using different algorithms based on American Sign Language", IEEE International Conference on Imaging, Vision & Pattern Recognition (ICIVPR), pp. 1-6, 2017.

[6] C. Vogler and D. Metaxas, “Parallel hidden Markov models for American sign language recognition”, Proceedings of the Seventh IEEE International Conference on Computer Vision, pp. 116-122, 1999.

[7] K. Fok, N. Ganganath. C. Cheng and C. Tse, “A Real-

Time ASL Recognition System Using Leap Motion Sensors", International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 411 – 414, 2015.

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

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

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

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

[12] R. Bowden and M. Sarhadi, "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, 2002.

[13] Q. Munib, , M. Habeeb, B. Takruri, H.A. Malik, "American sign language (ASL) recognition based on Hough transform and neural networks", Expert Systems with Applications, Volume 32, Issue 1, pp. 24–37, 2007.

[14] C. Zhang, Y. Tian and M. Huenerfauth, "Multi-modality American Sign Language recognition, IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 2881-2885, 2016.

[15] M.A. Bhuiyan and H. Hama, "Identification of Actors Drawn in Ukiyoe Pictures", Pattern Recognition, Vol. 35, No. 1, pp. 93-102, 2002.

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

[17] M.A. Bhuiyan, "Toward Face Recognition using Eigneface", International Journal of Advanced Computer Science & Applications, Vol. 7, No. 1, pp. 25-31, 2016.

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

[19] A. Khatun and A. Bhuiyan, "Neural Network based Face Recognition with Gabor Filters", International Journal of Computer Science and Network Security, Vol. 11, No. 1, pp. 71-76, 2008.

[20] M.A. Bhuiyan and C.H. Liu, "Intelligent Vision system for Human-Robot Interface", Proc. Of World Academy of Science, Engineering and Technology, pp. 56-63, 2007.

[21] M.A. Bhuiyan, "On Gesture Recognition for Human-Robot Symbiosis", The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2006, pp. 541-545, 2006.

[22] M.E. Islam, N. Begam, and M.A. Bhuiyan, "Vision system for human-robot interface", 11th International Conference on Computer and Information Technology (ICCIT), pp. 1-6, 2008.

[23] S. Sharma and M. Varshney, "An efficient approach for Web-log mining using ART", International Conference on Education, Management and Technology (ICEMT), 2010, pp. 196-199,

Dr. Md. Al-Amin Bhuiyan received both his Bachelor and Master degree from University of Dhaka, Bangladesh, in 1987 and 1988, respectively. He got his PhD degree from Osaka City University, Japan in 2001. Currently he is a an Associate Professor at the Department of Computer Engineering, King Faisal University, Saudi Arabia (under lien leave from Jahangirnagar University, Bangladesh where is employed as a Professor). Prior to this, Dr. Bhuiyan lent his teaching and research experiences at several Universities in Japan, Bangladesh and UK. His research interests include image processing, computer graphics, pattern recognition, artificial intelligence, neural networks, robotic vision, and so on. He has published numerous articles in International refereed journals and conference proceedings.