Real-time sign language recognition based on neural network architecture

Conference Paper · April 2011

DOI: 10.1109/SSST.2011.5753805 · Source: IEEE Xplore

4 authors, including:

A. Davari
West Virginia University
77 PUBLICATIONS **911** CITATIONS

SEE PROFILE

Real-time Sign Language Recognition based on Neural Network Architecture

Priyanka Mekala[1], Ying Gao[2] , Jeffrey Fan[1] , Asad Davari[3]
[1]Dept. of Electrical and Computer Eng., Florida International University, FL, U.S.A
[2]Electrical Eng. Dept., University of Wisconsin, Platteville, WI, U.S.A
[3]Dept. of Electrical and Computer Eng., West Virginia University Institute of Tech, WV, U.S.A

Abstract- **In real-time, it is highly essential to have an autonomous translator that can process the images and recognize the signs very fast at the speed of streaming images. In this paper, architecture is being proposed using the neural networks identification and tracking to translate the sign language to a voice/text format. Introduction of Point of Interest (POI) and track point provides novelty and reduces the storage memory requirement.**

I. INTRODUCTION

Since ages communication has served as a medium to build relationships, know people, understand technology and allow rapid growth and development on a global basis. Normal people can communicate their thoughts and ideas to others through speech. One important means of communication method for the hearing impaired community is the use of sign language, as in [1]. 500,000 and 2,000,000 people use Sign Language as their major daily communication tool. These numbers may deviate from other different sources but it is surprisingly popular as mentioned in Trudy Suggs book: American Sign Language is the 3rd most-used language in the United States. It seems that 3.68% of the total population is found to be hard of hearing and 0.3% of the total population is functionally Deaf, out of a total population of about 268,000,000 (2005) in the US. In Canada and the U.S.A., American Sign Language (ASL) is generally preferred as the vehicle of communication for the hard of hearing and the deaf alike.

Several methods has been proposed in the past to translate the signs using the gestures and features of the signer. Primarily, Ko and yang developed a finger mouse that enables a user to specify commands with the fingers as in [2]. Other novel approaches include the colored glove based method, skin color segmentation, video sequence appearance modeling and Hidden Markov Model (HMM) systems as in [3]-[5].

The contribution of this paper can be summarized as the following: (1) Introduction of concerns and issues affection the real-time processing of Sign Language Recognition (SLR), (2) Proposed novel Point of Interest (POI) features for extraction and use of angle chart to predict the motion vector sequences.

The rest of the paper is organized as follows: Section II presents the real-time sign language architecture on a system level and each of the subsystems in details. Section III presents the experimental results and discussions. Finally, Section IV concludes the paper.

1. What is ASLR?

Signing takes place in a 3D space, called signing space close to the trunk and the head, as in [7]. Signs are either one-handed or two-handed. For one-handed signs the so called "dominant hand" performs the sign, whereas for two handed signs the second hand, the non-dominant hand, is also needed. Many researches began in late 90s to deal with recognition of sign language automatically in various languages. In the works that have been carried out previously on various sign language recognition techniques, initially with gesture recognition using electromechanical devices affects the signer's signing ability. The other category deals with use of colored gloves in order to provide color segmentation and extract features to recognize the signs. It uses instrumented glove-based data collection for training and testing. Lately evolved the stream where no use of any electromechanical devices or colored gloves is required.

Neural networks, HMMs, distance-based, skin color based, and other statistical methods have successfully solved the sign language recognition when considered for word/sentence recognition, as in [6]-[11]. It only requires video-based data collection and hence leads to a better natural interface for the user. Sign language when compared to the spoken language has different grammar. In spoken language, the speech is group of sentences where words in the sentence are linear (i.e. one word followed by another) whereas in sign language, a simultaneous structure exists with a parallel temporal and spatial configuration.

2. Concerns and Issues affecting Real-time SLR

- A video image acquisition process is subjected to many environmental concerns such as the position of the video camera, environmental conditions like lighting sensitivity, background condition and number of cameras used.
- Occlusion plays a crucial role factor in real time as while signing; some fingers or even a whole hand can be occluded, as in [12].
- Sign boundaries have to be detected automatically. The start and end of a sign are required to be detected automatically from the captured video streaming data.
- A sign is affected by the preceding and the subsequent sign (co articulation).

978-1-4244-9592-4/11/$26.00 ©2011 IEEE

- The position of the signer in front of the camera may vary. This results to unwanted temporal and spatial change of the co-ordinate axis under consideration. Movements of the signer, like shifting in one direction or rotating around the body axis, must be considered.
- Each sign varies in time and space. The signing speed differs significantly. Even if one person performs the same sign twice, small changes of speed and position of the hands will occur.
- The projection of the 3D scene on a 2D plane results in loss of depth information. The reconstruction of the 3D-trajectory of the hand in space is not always possible [7].
- The processing of a large amount of image data is time consuming, so real-time recognition is difficult.
- Higher resolution causes considerable delay in the execution of the acquisition process and longer processing time.
- Real-time processing: The translator is sufficiently fast to capture images of signer, process the images and display the sign translation on the computer screen.

II. REAL-TIME SIGN LANGUAGE ARCHITECTURE

System Design at Module level

The system design of the English Sign language translator

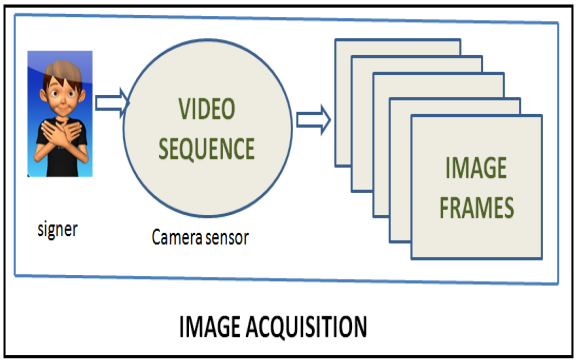


Fig.1. Image Acquisition block of the Automatic Sign Language Recognition (ASLR) system design

[13] is as shown in Figure 1. It consists of many block modules described in detail below. This design describes the top level architecture of the ASLR. A combinational neural network model is developed for the recognition of gestures using the features computed from the video stream. The recognized signs are connected to the specific audio signals using MATLAB software for the communication between the ordinary and deaf people.

1. *Image Acquisition*

The video sequence of the signer, i.e. the person conveying in the sign language, can be obtained by using a camera. In this work, we consider that the camera faces towards the signer in order to capture the front view of the hand gestures of the signer. The initiation of the acquisition is being done manually. Figure 1 shows the description of the image acquisition block of the ASLR system design. A camera sensor is needed in order to capture the features/ gestures of the signer.

2. *Preprocessing*

Local changes due to noise and digitization errors should not radically alter the image scene and information. In order to satisfy the memory requirements and the environmental scene conditions, preprocessing of the raw video content is highly important [14]. Various factors like illumination,

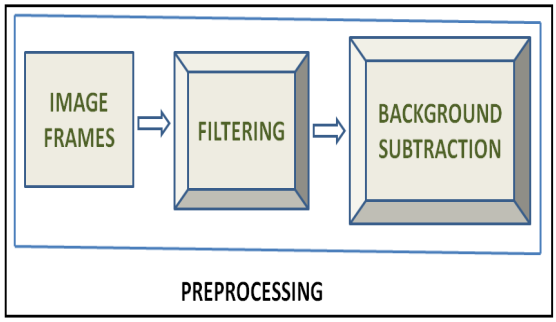


Fig.2. Preprocessing block of the ASLR system design

background, camera parameters, and viewpoint or camera location are used to address the scene complexity. These scene conditions affect images of the same object dramatically. The first most step of preprocessing block is filtering as shown in Figure 2. A moving average or median filter is used to remove the unwanted noise from the image scenes. Background subtraction forms the next major step in the preprocessing block. Running Gaussian average method [17] is used in order to obtain the background subtraction as it is very fast and consumes low memory when compared to other methods. The probability distribution function (pdf) of the background is given by

Where α is the learning rate, *i* refers to the current frame index, *B* refers to the background frame, *F* refers to the actual image frame. This takes into consideration of the illumination changes like lightning, camera motion changes, high frequency background objects, such as the tree leaves and branches.

3. *Feature Extraction*

Under different scene conditions, the performance of different feature detectors will be significantly different. The nature of the background, existence of other objects (occlusion), and illumination must be considered to determine what kind of features can be efficiently and reliably detected [15]. Usually the hand shape and the movement are of major concern in order to guess the word/sentence. The feature vector is a single row column matrix of N elements. The feature vector computation involves time and memory. In general, as the output classes are increased and the non-linearity of differentiation in the classes increases, more

feature vectors are necessary for the object recognition technique.

Features are divided as two sub-categories: hand shape and hand movement. The state of the hand gestures are given by the attributes called Point of Interest (POI) of the hands. We consider two POIs in order to represent the 'shape' and 'direction of movement'. It is shown in Figure 3 both the POIs and their description.

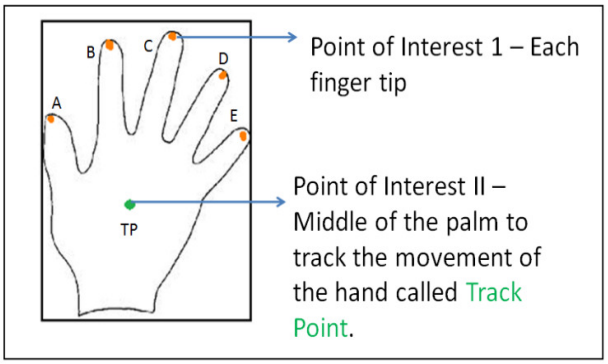
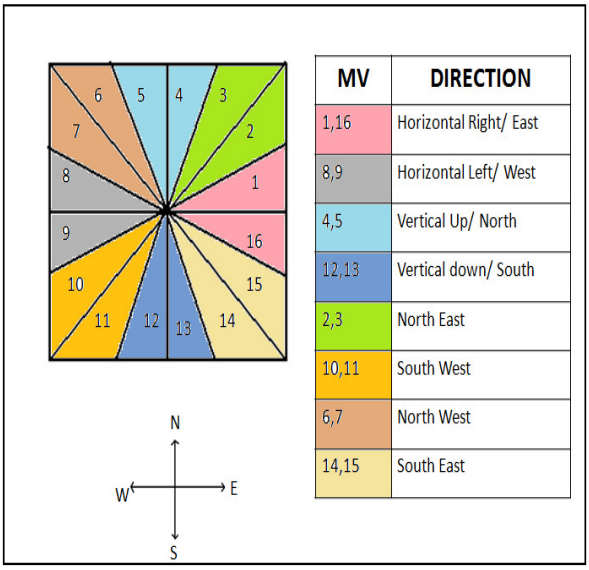


Fig.3. Point of Interest I (Each fingertip of the hand) and II (midpoint of the palm) - POIs.

The finger tips are referred to as A, B, C, D, E points and the midpoint of the hand as Track-point (TP) as shown in Figure 3. The motion vector of the TP indicates the direction of motion as an angle lying in one of the cycles of the angle chart as shown in Figure 4.



MV	DIRECTION
1,16	Horizontal Right/ East
8,9	Horizontal Left/ West
4,5	Vertical Up/ North
12,13	Vertical down/ South
2,3	North East
10,11	South West
6,7	North West
14,15	South East

Fig.4.Angle Chart – Motion Vector related to direction.

Table I describes the relation between the Motion Vector (MV) direction and the angle chart. Since the angle chart is divided into 16 slots, the motion vector is categorized as one of them. Four bits are used to represent the MV since 16 intervals.

TABLE I
MOTION VECTOR DIRECTION

Vector Direction from Angle Chart	Angle	Quadrant
1,2,3,4	0° – 90°	I
5,6,7,8	90° – 180°	II
9,10,11,12	180° – 270°	III
13,14,15,16	270° – 360°	IV

The feature vector consists of 55 features. These include 5-finger tip elements (A, B, C, D, E, 1 - present, 0- absent), 4-motion vector elements; 6 elements are the MV sequence; and the remaining 40 are from the wavelet transform of the Fourier transformed image of a gesture[19-21]. The Fourier descriptors are used to outline the shape seen as a closed curve, i.e. sequence of points on the boundary curve. Fourier descriptors are the Fourier transformation coefficients of the object's boundary curve. It is the frequency domain analysis result of the boundary curve, and is one of the method to describe the curve which unaffected by transformation and revolving of the origin of coordinates.

The DFT (Discrete Fourier Transform) coefficients are calculated as in [25]:

$$z(k) = \sum_{k=0}^{n-1} p(k) exp\ (j\frac{2\pi lk}{n}) \quad (l = 0,1,\dots n-1)$$

where z is the Fourier transformation of p, and it is also the expression of a point sequence in the frequency domain. $p(k)$ is the representation of the object's boundary in a one-dimensional space as:

$$p(l) = x(l) + jy(l) \quad (l = 0,1,\dots n-1), \quad j = \sqrt{-1}$$

The MV sequence of each gesture that is trained is unique. Every gesture has a corresponding stored model in the neural network model. Recognition of gesture by gesture is based on a stage where it reaches minimum energy at the end of a gesture. The number of key frames in one gesture varies from man to man due to difference in their speed of action. But the motion vector acts the same way for a unique gesture.

4. Neural Networks System

Training and generalizing are the most basic and important properties of the neural networks. The neural network architecture consists of three layers - an input layer, one hidden layer and an output layer. In the gesture classification stage, a simple neural network model is developed for the recognition of gestures signs using the features computed from the video captured. The features can then be extracted from the video captured using any of the following system. Different network models exist for training the neural net of which the best chosen are Back propagation (most widely used), 3D Hopfield, simple recurrent networks (SRN- Elman and Jordan networks), and self organizing maps (SOM) [24]. Depending on the feature vectors, the best neural net training method is chosen. Recurrent neural networks for many reasons provide significant advantages over feed forward

networks especially for applications in areas which require temporal processing where time counts.

Assumptions are made that there is no overlap between the two hands to avoid occlusions caused due to over lapping of two hands in normal signs.

Sign language recognition using neural networks is based on the learning of the gestures using a database set of signs. In the case where the database is small, a sequential exhaustive search can solve the problem [23]. There is necessity of universal database as the applications grow and in that case such sequential search algorithms fail to meet the timing and memory constraints. In order to reduce the size of search space, a new search algorithm called combinational neural networks is being used. This architecture of CNN is based on the cache search memory concept of a CPU.

In the proposed approach we will use a new scheme called combinational neural networks [22]. A parallelism exists between the feature extraction layer and the neural network layer as shown in Figure 5. A dual bus is provided in order to allow the flow of data in both directions. The feature vector computation involves time and memory. In general, as the output classes are increased and the non-linearity of differentiation in the classes increases, more feature vectors are necessary for the object recognition technique.

A three layer network called back propagation is used to build the CNN. The transfer function used for the back propagation layers is typically a step function, such as the hard-limit transfer function, or a sigmoid function such as log-sigmoid.

The network layer consists of 3 stages: stage 1, stage 2, and stage 3 as in [22]. Each stage acts as a back propagation neural network layer that takes the elements of feature vector as input and outputs the class of object. Each stage receives their input elements from the feature extraction layer.

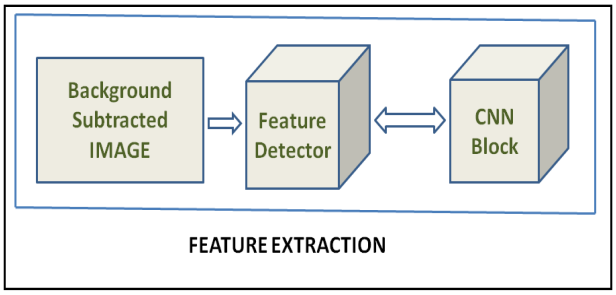


Fig. 5. Feature extraction and Combinational Neural Network blocks of ASLR system design.

III. Results and Analysis

The system is designed to recognize simple gestures or signs. The design is very simple and does not require any kind of gloves to be worn. Also the system is applicable to different backgrounds. This sign language recognition approach requires a computer with at least 1GHz processor and at least 256 MB of free RAM. The training set consists of all alphabets A to Z (26 patterns). Figure 6 shows the sign language representation of English alphabets. The letters J and Z involve motion and, hence, the motion vector is needed to recognize them. All the other letters have 0 motion vectors sequence before they encounter the lowest energy frame of the gesture. The Figure 7 represents the output table of the algorithm performed using the MATLAB. The algorithm is able to detect all the alphabets from A to Z with 100% recognition rate. In case of noise immunity, the experimental results that are shown in Figure 8 reveal that the algorithm is error free up to an average of 48%. It implies that in the case of noise corruption which results in loss of the feature vector elements of the input image, the gesture is detected error free. Noise corruption may include noises due to environment or loss of information due to fading, blur or damaged. The next phase involves applying the algorithm to simple words of sign language gestures. The output of the algorithm can be connected to a 5x3 LED pixel array to indicate the letters. Here in the figure it can be observed that we have used a 5x3 array to represent the output.

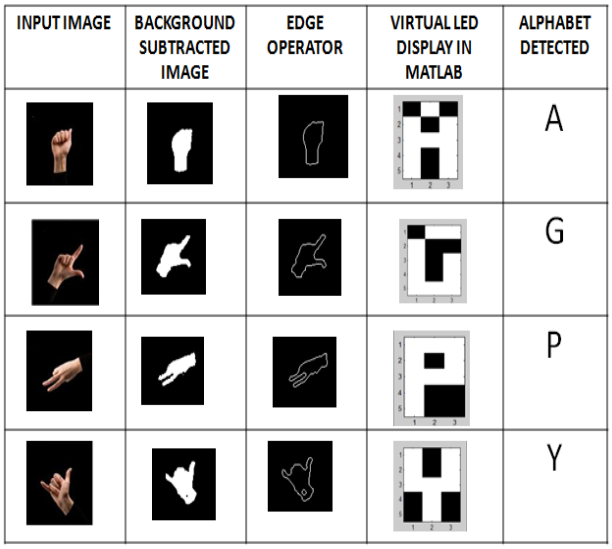


Fig. 7. Matlab results of the input image and letter recognition algorithm.

TABLE I
NOISE IMMUNITY OF SIGNS

Alphabet	Noise Immunity	Average Noise Immunity	Standard Deviation
A	48%	-	-
G	30%	-	-
P	37%	-	-
Y	52%	-	-
Results		48%	10.0789

IV. Conclusion

The signs for all the alphabets from A to Z are being recognized using the combinational neural networks architecture. The advantage of using the algorithm is high processing speed which can produce results in real-time manner. The speed of processing is increased due to the

neural network architecture. It is also advantageous as even noise corrupted almost up to 48%, the signs can still be retrieved. For future extensions processing of words and sentence gestures can be included. In that case the grammar and syntax plays an important role in deciding the efficiency and speed of the structure.

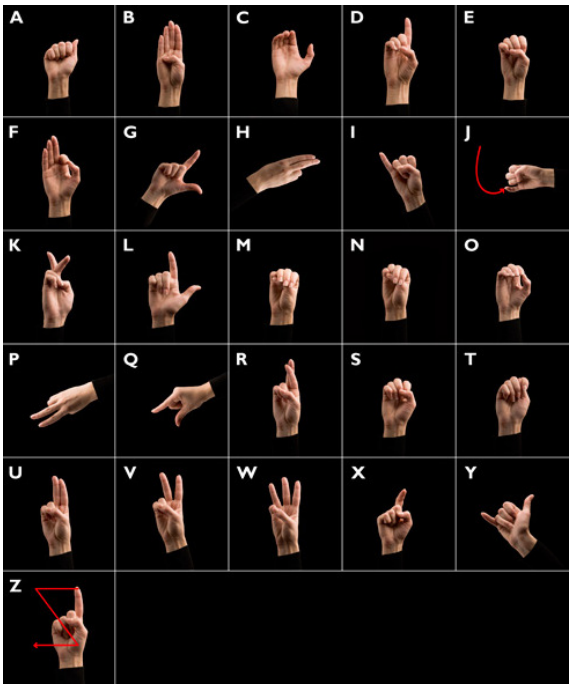


Fig. 6. Sign Language English alphabets representation [18].

REFERENCES

[1] Paulraj M P, Sazali Yaacob, Mohd Shuhanaz bin Zanar Azalan, Rajkumar Palaniappan, "A Phoneme based sign language recognition system using skin color segmentation", Signal Processing and Its Applications (CSPA), pp: 1 – 5, 2010.

[2] K. K. Byong and H. S. Yang, ''Finger mouse and gesture recognition system as a new human computer interface," in Computer & Graphics, Vol. 21, No. 5, PP. 555 - 561, 1997.

[3] Manar Maraqa, Dr. Raed Abu-Zaiter, "Recognition of Arabic sign language (ArSL) using recurrent neural networks", Applications of Digital Information and Web Technologies, pp: 478 – 481, 2008.

[4] Yang quan, "Chinese Sign Language Recognition Based On Video Sequence Appearance Modeling", Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on , pp: 1537 – 1542.

[5] K.Kawahigasi, Y.Shirai, J.Miura, N.Shimada "Automatic Synthesis of training Data for Sign Language Recognition Using HMM", PROC.ICCHP, pp.623-626, 2006.

[6] R. Liang and M. Ouhyoung (1998), "Real-time continuous gesture recognition system for sign language", Proc Third. IEEE International Conf: on Automatic Face and Gesture Recognition, pp. 558-567.

[7] Bauer, B.; Hienz, H., "Relevant features for video-based continuous sign language recognition", Automatic Face and Gesture Recognition, 2000. Proceedings, pp: 440 – 445, 2000.

[8] L. Rabiner and B. Juang. An introduction to hidden markov models. IEEE ASSP Magazine, pages 4–16, January 1986.

[9] K. Grobel. Videobasierte Geb¨ardenspracherkennung mit Hidden– Markov–Modellen. Ph. D. Thesis, Aachen University of Technology, VDI-Verlag, D¨usseldorf, 1999.

[10] Maebatake, M.; Suzuki, I.; Nishida, M.; Horiuchi, Y.; Kuroiwa, S., "Sign Language Recognition Based on Position and Movement Using Multi-Stream HMM", Universal Communication, 2008. ISUC '08. Second International Symposium, pp: 478 – 481, 2008.

[11] E.-J.Holden, G.Lee, R.Owens, Australian Sign Language Recognition, Machine Vision and Applications, Vol.16, pp. 312-320, 2005.

[12] P. Mekala, R. Salmeron, Jeffrey Fan, A. Davari, J. Tan, "Occlusion Detection Using Motion-Position Analysis", IEEE 42nd Southeastern Symposium on System Theory (SSST'10), Tyler, TX, pp. 197-201, March 7-9, 2010.

[13] H.Brashear, K.-H.Park, S.Lee, V.Henderson, H.Hamilton, T.Starner, "American Sign Language Recognition in Game Development for Deaf Children" Proc.Assets, pp.79-86, 2006.

[14] Brian L. Pulito, Raju Damarla, Sunil Nariani, " 2-D Shift Invariant image Classification Neural Network, which overcomes Stability, Plasticity Dilemma", Vol 2, International Joint Conference on Neural Network, San Deigo, June 17-21,1990.

[15] Hong Pan and Liang-Zheng Xia, "Efficient Object Recognition Using Boundary Representation and Wavelet Neural Network", Neural Networks, IEEE Transactions, vol-19, pp.2132 - 2149.

[16] P. Mekala, S. Erdogan, Jeffrey Fan, "Automatic object recognition using combinational neural networks in surveillance networks", to appear on IEEE 3rd International Conference on Computer and Electrical Engineering (ICCEE'10), Chengdu, China, November 16-18, 2010.

[17] Jong Bae Kim,Hye Sun Park,Min Ho Park,Massimo Piccardi,'Background subtraction techniques: a review',Systems, Man and Cybernetics,vol.4,2004,IEEE International Conference,pp:3099-3104.

[18] http://www.colourlovers.com/blog/2008/01/22/color-by-hand-the-american-sign-language-spectrum/

[19] C. H. Lin and C. H. Wang, "Adaptive wavelet networks for powerquality detection and discrimination in a power system," IEEE Trans. Power Delivery, vol. 21, no. 3, pp. 1106–1113, Jul. 2006.

[20] Y. C. Huang and C. H. Huang, "Evolving wavelet network for power transformer condition monitoring," IEEE Trans. Power Delivery, vol. 17, no. 2, pp. 412–416, Apr. 2002.

[21] K. H. Roberto, Galvao, and T. Yoneyama, "A competitive wavelet network for signal clustering," IEEE Trans. Syst. Man Cybern. B, Cybern., vol. 34, no. 2, pp. 1282–1288, Apr. 2004.

[22] P. Mekala, S. Erdogan, Jeffrey Fan, "Automatic object recognition using combinational neural networks in surveillance networks", IEEE 3rd International Conference on Computer and Electrical Engineering (ICCEE'10), Chengdu, China, Vol. 8, pp. 387-391, November 16-18, 2010.

[23] Tzay Y. Young, King-Sun Fu,"Handbook of Pattern Recognition and Image Processing", Academic Press Inc, 1986.

[24] Lubin, J.; Jones, K.; Kornhauser, A.; "Using back-propagation networks to assess several image representation schemes for object recognition",Neural Networks, 1989. IJCNN., International Joint Conference,1989.

[25] Yang quan, "Chinese Sign Language Recognition Based On Video Sequence Appearance Modeling", Industrial Electronics and Applications (ICIEA), pp: 1537 – 1542, 2010.