Gesture Recognition Using Shape and Depth Information of Body for Human-Robot Interaction

Jae-Yong Oh and Chil-Woo Lee
Dept. of Electronics and Computer Eng., Chonnam Nat'l Univ., Gwangju, Korea
ojyong@image.chonnam.ac.kr, leecw@chonnam.ac.kr

Abstract

In this paper, we describe an algorithm which can automatically recognize human gesture for Human-Robot interaction by utilizing 3 dimensional features extracted from bodily region of the images. In the algorithm, we first extract a feature vector from stereo image. Efficient method is required for representing the 3D gesture features, so we proposed the method that can represent 3D and 2D information of gesture simultaneously. The next step is constructing a gesture space by analyzing the statistical information of training images with PCA. And then, input images are compared to the model and individually symbolized to one portion of the model space. In the last step, the symbolized images are recognized with HMM as one of model gestures. The experimental results indicate that the proposed algorithm is efficient on gesture recognition, and it is very convenient to apply to real world situation.

Keywords: Gesture recognition, Principle Component Analysis, Hidden Markov Model

1. Introduction

Recently, various applications of robot system become more popular accordance with rapid development of computer hardware/software, artificial intelligence, and automatic control technology. So far, robot mainly have been used for industrial field, however, nowadays it is said that the robot will do an important role in home service application in the near future. To make the robot more useful, we require further researches on natural communication method between human and the robot system, and autonomous behavior generation. The gesture recognition technique is one of the most convenient methods for natural human-robot interaction, so it is to be solved for implementation of intelligent robot system.

Briefly speaking, that we understand the human behavior and gesture is to track the temporal sequence of movement of human body automatically, and then to estimate some meanings of the temporal movement from previously trained and arranged motion data[1]. That means, for understanding human behavior in a sequence of image, we have to construct a model space for human behaviors previously and analyze temporal changes of human body. However, it is very difficult to analyze temporal changes, namely historical meaning of bodily motion automatically because a human body is a three-dimensional object with very complicated structure and flexibility.

In early works, many researchers tried to measure the configuration of the human body with sensors attached to the joints of limbs and to recognize gesture by analyzing the variation of joint angles. However, this requires the user to put on irritating devices such as a data glove and a data suit, and it usually needs long cables connection from the devices to computers. So, it hinders the easy usage and embarrasses the user with unnatural dullness. Any awkwardness of wearing devices can be dissolved by using video-based *non-contact* recognition techniques. One approach adopts a set of video cameras and computer vision techniques to interpret gestures. We call the method "appearance-based gesture recognition" since only the visual appearance is used in the recognition process. Appearance-based gesture recognition approaches are different depending on whether they use 3D model or 2D model of the human body. The 3D model method has difficulty in modeling and matching process because the joints connecting the bones naturally have different degrees of freedom (DoF). Furthermore, the 2D model method requires complex calculations since the appearances become different according to the position or direction of the cameras.

As knowing from the facts mentioned above, most algorithms are work under many restricted conditions. In this paper, we describe the algorithm which can automatically recognize human gesture without such constraints by utilizing three-dimensional features extracted from stereo images.

The paper structured as follows: Section 2 presents an overview of the recognition system. Section 2.1 and 2.2 describe how the bodily region and features can be extracted from the stereo images. And in section 2.3, we explain how the gesture model space is constructed by

1) This research has been supported by research funding of "Center for High-Quality Electric Components and Systems", Chonnam National University, Korea

2) This research has been supported by research funding of "Development of humanoid technology based on network", KIST, Korea

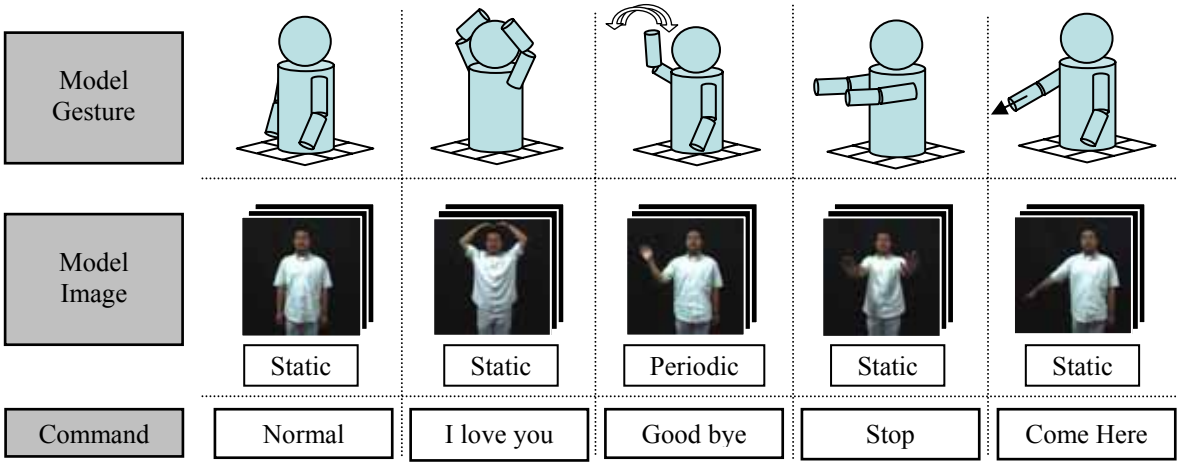


Fig. 1 Basic gesture definition for Human-Robot interaction

analyzing the statistical information of training images with the Principle Component Analysis (PCA) method. And then, section 3.4 presents how the sequential poses are classified into model gestures with Hidden Markov Model (HMM) algorithm. Section 4 shows the efficiency of the proposed method with experimental results. Finally the paper is closed with mentioning the conclusion and further works.

2. The Recognition System

2.1 Gesture Definition

Human use many kinds of gestures in daily life. But it is very difficult to implement the recognition algorithm which can understand all the gestures. For controlling the robot system, we can assume several prominent gestures. Therefore, here we define five different gestures for human-robot interaction as shown in Fig.1. These gestures are mutually isolated from each other in meaning, and are classified into static or periodic type.

Fig.2 shows an overview of the recognition system and it contains three processing procedures: 1) Preprocessing, 2) Feature Extraction, and 3) Statistical Classification. The preprocessing procedure extracts the foreground image; bodily region, and feature extraction procedure calculates a vector for each frame of foreground image. Finally, recognition module using PCA and HMM classifies the gestures.

2.2 Preprocessing

In the first step of gesture recognition, we need to extract the foreground image having bodily motion. In general, the scene includes a lot of objects in the background that may disturb gesture recognition. In the case of fixed background, foreground is easily estimated by subtracting simple spatial or temporal background model from input image. But, this method is not suitable for mobile robot vision system because the background is so changeable.

In the paper, we use a simple foreground extraction

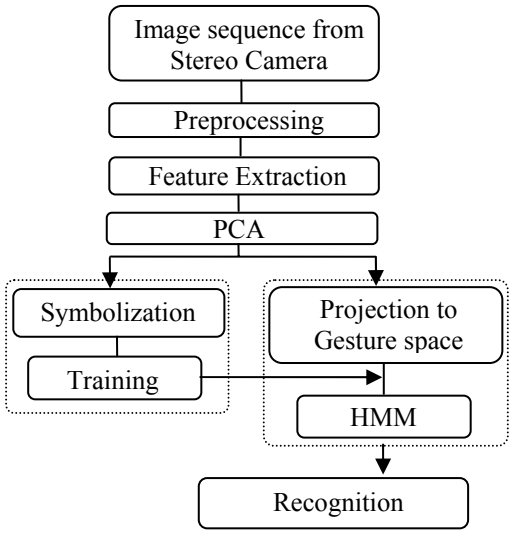


Fig. 2 Overview of the Recognition system

method using depth information. If we concern the situation of visual interaction with a robot, we can assume that human body is relatively closer to the robot than other background objects. The procedure is as follows; first, we detect a face region of the frontal person[2], and then estimate the distance from the robot camera to the face by calculating stereo geometry. Finally, we can obtain the closer region of input image as the foreground. It is shown in Equation (1).

$$F(x,y) = \begin{cases} 0 & : \quad D(x,y) > D_f + c \\ F(x,y) & : \quad D(x,y) \le D_f + c \end{cases} \qquad (1)$$

In the equation, $F(x,y)$ is gray-level value of input image at position (x,y) of the corresponding disparity image, $D(x,y)$ is the distance from robot camera for each pixel, D_f is distance from camera to the face, and c is a constant value considering thickness of the body. For the face detection algorithm, we adopt CMU's face detection algorithm which uses Gaussian Mixture Model (GMM) of color distribution of the face region and the

background region[2]. And, for calculating the stereo geometry, we use a commercialized vision system; Bumblebee Camera set of Point Grey Research. This system estimate correspondence between stereo images with Sum of Absolute Differences[3].

This algorithm works very well if there are no obstacles between the person and the robot. Considering more complicate situation; there are some obstacles, we have to improve the algorithm for the case by introducing mixture of color information and depth information.

2.3 Feature Extraction

Once bodily region is extracted, we need to find the features of the body configuration that can efficiently represent the posture and motion of the body.

One of the famous algorithms using global motion feature information is *Motion History Image* (MHI) method. In this method, the accumulation of object silhouettes is used to analyze the human motion[4][5]. The limitation of this method is that we cannot recognize the motion which is occurring inside of body region and it is incongruent with variation of motion speed.

Another method is Ross Cutler's method[6]. In the method, firstly optical flow is obtained, and the result is segmented into several blobs. Gestures are recognized by using a rule-based technique based on characteristic analysis of the motion blobs such as relative motion and size. The disadvantage of this technique is that the occlusion of motion blobs might result in wrong recognition.

In this paper, we propose a efficient model, *Active Plane Model* (APM), for representing the posture and motion information simultaneously. The APM is characterized by that it can represent three-dimensional depth information and two-dimensional shape information at the same time. And the APM is less sensitive to noise or appearance variation since it uses the standardized model as shown in Fig. 3.

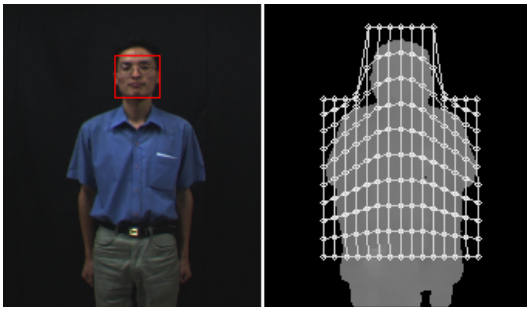


Fig. 3 Standardization Model of APM
(Red rectangle of left side is the face region)

APM consist of several nodes connected to each other in grid shape and it is deformed by the shape and depth information. For the first step to deform the APM, we find the contour of foreground region. The contour is searched from the minimal rectangular boundary surrounding foreground region into center of rectangle. If we find the contour, outer nodes of APM are fixed, then, inner nodes are moved to new location in order to keep the same distance between neighbor nodes. That means the inner nodes are rearrange by calculating the average position between the most outer nodes as shown in equation (2).

$$N_i^{t+1} = \text{Average potion of neighbor nodes of } N_i^t \quad (2)$$

$$N_i^{t+1} - N_i^t \le T \ , (t \ge 0) \quad (3)$$

In equation (3), T expresses a suitable threshold distance. By adopting the criterion of equation (3), we can stop the deformation quickly. Fig. 4 shows the deformation shape of APM for several gestures.

From this deformed APM, we can calculate a displacement of nodes of grid.

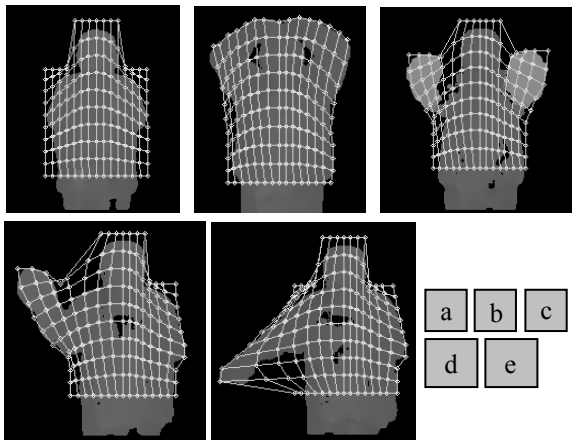


Fig. 4 Deformed APM examples for several gestures.
(a) Normal, (b) I love you, (c) Stop, (d) Good bye,
(e) Come here

Therefore, multi-dimensional features are extracted for every frame by using the equation (4),

$$F_t = \{N_1, N_2, ..., N_n\} \quad (4)$$

$$N_i = \{x_i, y_i, z_i\} \quad (5)$$

$$(0 \le t \le T, \quad 1 \le i \le n, \quad n \ge 4)$$

where, F_t is a feature vector set at time t, N_i is i th nodal position of APM, and n is the total number of APM nodes.

Also, direction vector of pointing gesture; for the case of 'Come here', can be calculated using APM. The most outlier point in an APM can be assumed as pointing point, and then direction vector is obtained by connecting the point to the face region.

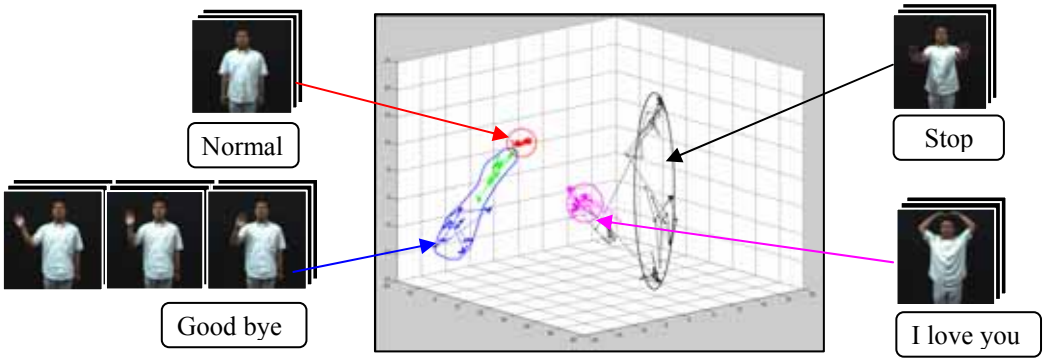


Fig. 5 Projection of model gesture sequence into the gesture space

2.4 PCA of APM deformation

A common method for linear dimension reduction is *Principal Components Analysis* (PCA). The method is performed by applying eigen-value decomposition on the covariance matrix of feature data. A small subset of resulting eigenvectors, covering a desired amount of the data variance, forms a new linear subspace for representing the data set[7].

PCA is clearly capable of reducing the dimensionality of data sets with structure, but not necessarily suitable for finding spatio-temporal structure for several reasons. The method assumes that the underlying structure of the data is linear. Unless the underlying structure is linear, the resulting subspaces will overestimate the reduced dimensionality and yield wrong principal components and reduced data that do not allow for the discovery of structure. The features extracted from an image sequence are multi-dimensional and we can not solve the multi-dimension problem easily.

By subtracting the average vector, ***c,*** of the all features, as in equation (7), the new feature matrix ***X*** is obtained. The covariance matrix, ***Q***, of gesture features can be obtained from equation (8). Then, the PCA is straight-forward requiring only the calculation of the eigenvectors satisfying equation (9).

$$c = (1/N)\sum_{i=1}^{N} x_i \qquad (6)$$

$$X \overset{\Delta}{=} [x_1 - c, x_2 - c, ..., x_N - c]^T \qquad (7)$$

$$Q \overset{\Delta}{=} X \cdot X^T \qquad (8)$$

$$\lambda_i \cdot e_i = Q \cdot e_i \qquad (9)$$

There are many numerical methods for calculating the eigenvector. We utilize the SVD (*Singular Value Decomposition*) algorithm. The SVD provides a series of Eigenvalue $\lambda_i (i = 1,2,..,N)$ (in decreasing order of size) and eigenvectors $\boldsymbol{e_i}$, which are orthogonal to each other.

It should be noted that the magnitude of an eigenvalue corresponds to the weight of that vector in the eigenspace. All N eigenvectors are needed to represent the feature sets accurately in a gesture space, however, a small number, ***k (k « N)***, of eigenvectors is generally sufficient for capturing the primary appearance characteristics of the gestures. From equation (10), a small number of eigenvectors can be chosen which span the whole space without any faults including much error. The ***k*** is selected such that the front eigenvectors of ***Q*** in decreasing order capture the important appearance variations in the feature sets like the following equation.

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq T_1 \qquad (10)$$

Where, the threshold T_1 is close to, but less than, unity.

Consequently, by using equation (10), an ***N***-dimensional vector ***X*** can be projected to a low ***k***-dimensional Eigenspace. An input feature, set ***x***, is subtracted from an average vector ***c***, and projected into the eigenspace as in equation (11).

$$m_i = [e_1, e_2, ..., e_k]^T (x_i - c) \qquad (11)$$

Using the *Principal Component Analysis*, a small number of component vectors are chosen which can represent the whole space, and we call it dimension reduction. Fig. 5 shows an example of the reduced space, and we call it *Gesture Space*. And from the figure, we can notice that image sequences are well classified by few component vectors.

2.5 Symbolic Gesture Recognition

HMM is a stochastic process and also a probabilistic network with *hidden* and *observable* states[8]. A time domain process demonstrates a Markov property if the

conditional probability density of the current event, even though all present and past events are given, depends only on the *j*-th most recent events. If the current event depends solely on the most recent past event, then the process is a first order Markov process. The initial topology for an HMM can be determined by estimating how many different states are involved in specifying a sign. Fine tuning for this topology can be performed empirically. HMM λ is represented by the several parameters. The parameter a_{ij} indicates state transition probability of HMM changing from state *i* to state *j*. The parameter $b_{ij}(y)$ indicates that probability of that the output symbol *y* will be observable in the transition of state *j* from state *i*. And another parameter π_i presents the probability of the initial state. Learning of HMM is equal to estimating the parameters $\{ \pi, A, B \}$ of a sequential symbolic data.

$$\xi_t(i,j) = \frac{P(s_t = i, s_{t+1} = j, Y \mid \lambda)}{P(Y \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_{ij}(y_{t+1}) \beta_{t+2}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_{ij}(y_{t+1}) \beta_{t+2}(j)} \quad (12)$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad (13)$$

In the equation (12), $\xi_t(i,j)$ is the probability of being in state s_i at time *t* and in state s_j at time *t+1*, and $\gamma_t(i)$ is the probability of being state S_i at time *t*. Using equation (12) and (13), the gesture model can be estimated.

$$\overline{\pi_i} = \gamma_1(i) \quad (14)$$

$$\overline{\alpha_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (15)$$

$$\overline{b_{ij}}(k) = \frac{\sum_{t=1, s.t. y_t = k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad (16)$$

Given the observation sequence [Y], the HMM model λ can be calculated using the forward variable $\alpha_t(i)$ and backward variable $\beta_t(i)$ with the following equation (17). This means that a symbol chain, namely an input image is recognized as a model which has the maximum value of equation (17).

$$P(Y \mid \lambda_i) = \sum_i \sum_j \alpha_t(i) a_{ij} b_{ij}(y_{t+1}) \beta_{t+1}(j) \quad (17)$$

3. Experimental Result

3.1 Gesture Database

For the training and evaluation of the algorithm, we construct an image database about five different gestures. This gesture database has the images captured by each stereo camera respectively. The resolution of the image is 320x240 in pixels and capturing speed is about 10 frames per a second. We recorded five kinds of gesture image for ten different persons. Training images are recorded in the controlled background with black screen, and include only the upper part of whole body. Test images are recorded while doing the five gestures continuously in the same environment.

3.2 Gesture recognition results

We made an experiment for recognizing gestures with test images. In the experiment, an APM of 15x10 in resolution was utilized, so that 450-dimensional feature vector was extracted for every frame. We used only five eigenvectors, and this low-dimensional vector can represent the whole gesture space well. Figure 6 shows the fact since only five eigenvectors with larger value occupy most part of the gesture space. We used five different HMMs for the five different gestures. The HMM parameters were estimated with the Forward-backward algorithm[9] under assuming that each HMM had 15 states.

The algorithm was implemented on a Pentium-IV PC with two CPUs of one GHz clock cycle. It spent 100ms for a frame on the average. The frame grabbing and disparity calculating time occupied almost 70% of total computation time.

Table 1 shows the recognition result for each gesture. “Good bye” and “Stop” gestures have recognition rates that are lower than the average rate respectively. We suppose that the gestures contain little motion or 3D information comparing to other gestures. Fig. 7 shows the failed cases of recognition. Fig. 8 shows the trajectories for the 4 points on the floor.

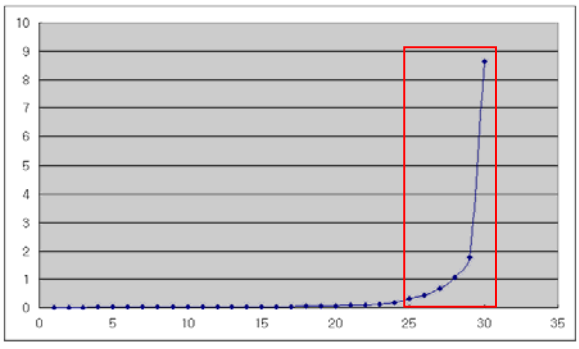
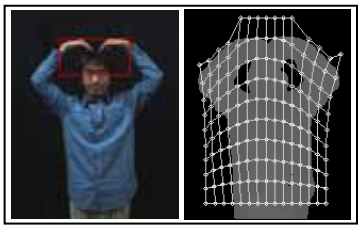


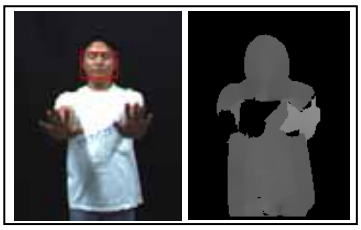
Fig 6. Amount of Eigenvalue for 30 Eigenvectors in increasing order. The last five eigenvector occupy most of Eigenvalue.

Table 1. Recognition rate for the test image

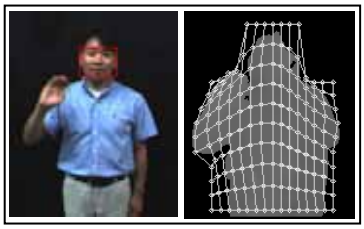
	Correct (%)	Incorrect (%)
Normal	98	Come here (2)
I love you	95	Normal (5)
Good bye	85	Normal (10), Come here (5)
Stop	80	Come here (20)
Come here	100	-
Average	91.6	-



(a) Failure to detect face region



(b) Failure to calculate disparity



(c) Too little motion

Fig.7 Examples of failed gesture recognition

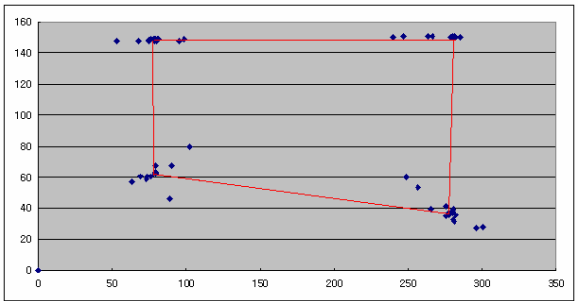


Fig.8 Point gesture and Trajectories for the 4 points

4. Conclusion

In this paper, one novel gesture recognition method using 2D shape and 3D depth information simultaneously is proposed. The algorithm works very well since it is robust to shape and depth variation. There are many gesture recognition algorithms using visual information. However, most algorithms work under many restricted conditions. To eliminate such constraints, we have proposed Active Plane Model (APM) to recognize human gestures. This model comes from structured model approach instead of geometric feature model approach using edge or corners. Therefore, we can make a general model for recognition of human gesture in real world. Also we can estimate the direction vector of pointing gesture easily by using the APM. However, there are some problems to be solved. One problem is that it is difficult to classify all gestures with several models obtained from training image. That means the algorithm become unstable under little information of ambiguous 2D shape and 3D depth. So, in order to improve the method for real world application, we must gather all kind of gesture images and analyze them statistically. Also our further works includes analyzing gestures of multiple persons concurrently.

References

1. Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang,"Visual Interpretation of Hand Gestures for Human-Computer Interpretation: A Review", IEEE Transaction on PAMI, Vol. 19, No. 7, July 1997

2. Xia Liu, Kikuo Fujimura, "Hand gesture Recognition using depth data", Automatic Face and Gesture Recognition, 2004. Proceedings. sixth IEEE International Conference on, May 2004

3. Point Grey Inc. (http://www.ptgrey.com)

4. James Davis, "Recognizing Movement using Motion Histograms", MIT Media Lab. Technical Report No. 487, March 1999

5. Davis, J.W.; Bobick, A.F.; “The representation and recognition of human movement using temporal templates” Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on , 17-19 June 1997

6. Ross Cutler, Matthew Turk, “View-based Interpr-etation of Real-time Optical Flow for Gesture Recognition”, Third IEEE International Conf. on Automatic Face and Gesture Recognition, 1998.

7. Chil-Woo Lee, Hyun-Ju Lee, Sung H. Yoon, and Jung H. Kim, “Gesture Recognition in Video Image with Combination of Partial and Global Information”, in Proc. of VCIP 2003, Lugano, July, 2003

8. L.R. Rabiner and B. H. Juang, “An introduction to hidden Markov Models”, IEEE ASSP Mag., pp 4-16, Jun. 1986.

9. Thad Starner, Alex Pentland, "Real-Time Americ- an Sign Language Recognition from Video using Hidden Markov Models", ISCV, 1995