2007－CVIM－160 (20)
2007／9／3

Automatic Human Action Analysis for Human Robot Interaction

Seong-Whan Lee*
Department of Computer Science and Engineering, Korea University
Anam-dong, Seongbuk-gu, Seoul 136-713, Korea
swlee@image.korea.ac.kr

Abstract

As recent interests in robotics move to intelligent robot and humanoid, human-robot interaction (HRI) technologies have become more essential and important. In case of the human being, they interact by recognizing various cues such as hand sign, face, gesture, speech, and so on. HRI can be accomplished by mimicking this way. Automatic recognition of human action which uses whole body action is required for HRI to communicate naturally and comfortably. This presents challenging problems, because detecting and tracking 3D human body components, and describing and modeling human action patterns from the body motion is a complex task. A human subject is firstly described by a set of features encoding the angular relations between a dozen body parts. Then, the extracted features are analyzed by HMM. For verifying the proposed method, we make several experiments using Korea University Gesture Database and apply technologies to interesting applications for HRI. The results and demonstration show that the proposed method can be effective in HRI, for automatic recognition of human action from motion sequences.

1. Introduction

Robotics research is currently supported in a dynamic environment. Traditional robots were used in factories for the purpose of manufacturing, transportation, and so on. Recently, a new generation "service robots" has begun to emerge [2, 10]. The United Nations (UN), in their recent robotics survey, divided robotics into 3 main categories: industrial, professional service and personal service robotics [10]. Many personal service robots are operated by non-expert users. Therefore, the robot is required to be able to interact naturally with humans, as close as possible to the way human-human interaction takes place.

There are many gesture recognition systems for HRI [2, 4, 5, 7, 8, 12, 13]. However, automatic recognition of gestures from whole body motion sequence for HRI is rare. Most previous approaches for HRI only recognize static arm poses, sign language, or command gestures, and cannot recognize gestures defined through specific motion gestures, such as waving a hand, bowing, and so on.

Waldherr et al. [12] introduced a hand command gesture interface for the control of a mobile robot equipped with a manipulator. A camera was used to track a person and recognize hand gestures involving arm motion. The developed algorithm is integrated in the mobile robot AMELA, which is equipped with a color camera mounted on a pan-tilt unit.

Gesture segmentation using continuous video was explicitly attempted by Lee and Kim [5]. Lee and Kim proposed explicit use of a threshold model corresponding to connecting patterns between gestures. Later, Barbic et al. [1] focused only on the segmentation problem, and proposed three methods, based on Principal Component Analysis (PCA), probabilistic PCA, and the Gaussian mixture model.

More recently Kahol et al. [4] attempted segmentation of complex human motion (e.g. dancing) sequence. The HMM was used for individual gesture patterns to spot dance sequences.

Human motion sequences are typically analyzed by segmenting them into shorter motion sequences, called gestures [13]. Gestures are most commonly used for communication among humans, reducing the chances of misclassifying static poses, by using continuous information. Gestures can be divided into two gestures, a communicative

* To whom all correspondence should be addressed.

gesture (a key gesture or a meaningful gesture) and a non-communicative gesture (a garbage gesture or a transition gesture) [9]. A good gesture recognizer attempts to process this transition motion in a systematic manner. The goal of this paper is to model transition gestures explicitly. Fig. 1 shows a sample video sequence containing several atomic gestures.

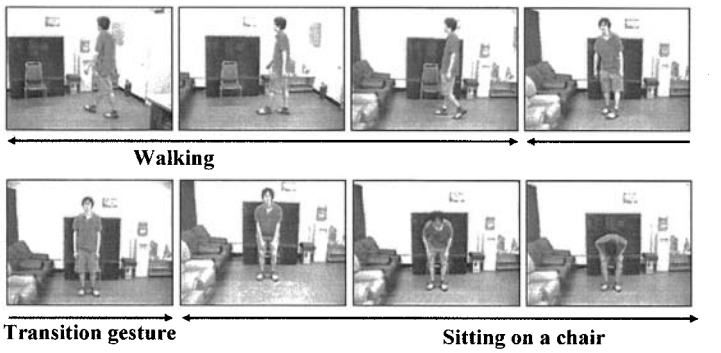


Figure 1. A motion example consisting of a sequence of key gestures and transition gestures

Among a variety of modeling tools, the Hidden Markov Model (HMM), is chosen. HMM is well-known for its capability in modeling spatiotemporal variability [5, 6, 7]. In this method a HMM is trained to model the variability of a target pattern. However, in the case of spotting tasks, the problem of treating non-target patterns remains. It is not straightforward to create a model for unspecified, unclassified, unlabeled patterns occurring between gestures. A new systematic method of building a model for transition gestures is proposed in this paper.

2. Estimation of 3D Human Body Pose

Linear combinations of prototypes based approach is used to reconstruct 3D human body pose from continuous depth images. If a sufficiently large number of pairs of a depth, and its 3D body model [14] as are used as prototypes of the human gesture, an input 2D depth image is reconstructed by a linear combination of 2D depth image prototypes. The reconstructed 3D body model can be obtained by applying the estimated coefficients to the corresponding 3D body model of the prototypes as demonstrated in Fig. 2. The goal is to find an optimal parameter set α which best estimates the 3D human body pose from a given depth image. To make various prototypes of 2D depth images and their 3D body models, data is generated using the 3D human model [14].

The depth image is represented by a vector $d_i = (d'_1,...,d'_n)^T$ where n is the number of pixels in the image and d' is a value of a pixel in the depth image. The 3D body model is represented by a vector $p_i = ((x_1, y_1, z_1),...,(x_q, y_q, z_q))^T$, where x, y and z are the position of body joint in the 3D world. Eq. (1) explains training data.

$$D = (d_1,...,d_m), P = (p_1,...,p_m), S = (s_1,...,s_m) \tag{1}$$

where m is the number of prototypes and $s_i = (s'_1,...,s'_n)^T$ is a silhouette image, s' is a value of a pixel in the silhouette image.

A 2D depth image is represented by a linear combination of a number of prototypes of 2D depth images, and its 3D body model is represented by estimated coefficients of the corresponding 3D body model of prototypes by Eq. (2).

$$\tilde{D}_i = \sum_{k=1}^{m} \alpha_k d_k \,,\ \tilde{P}_i = \sum_{k=1}^{m} \alpha_k p_k \tag{2}$$

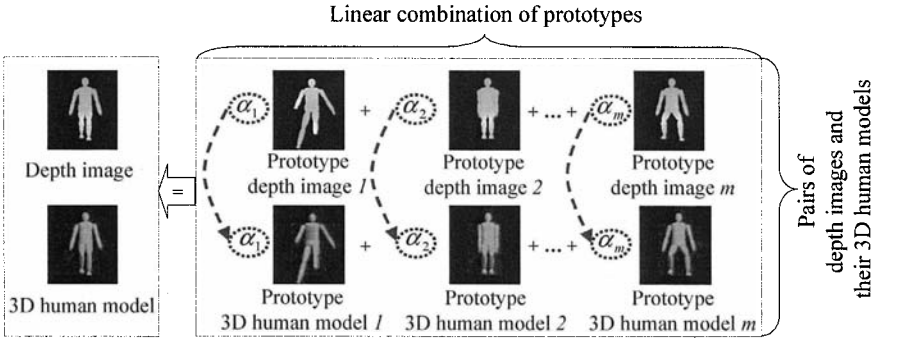


Figure 2. Gesture representation

In order to cluster the prototypes, the algorithm is constructed hierarchically. Given a set of silhouette images, the depth images and their 3D body models are used for training, these are classified into several clusters. A set of cluster is built in which each has similar shape in 2D silhouette image space. Then, each cluster is recursively divided into several sub-clusters. To divide training data into sub-clusters, the *k*-means algorithm is applied. The hierarchical model has four-levels as presented in Fig. 3.

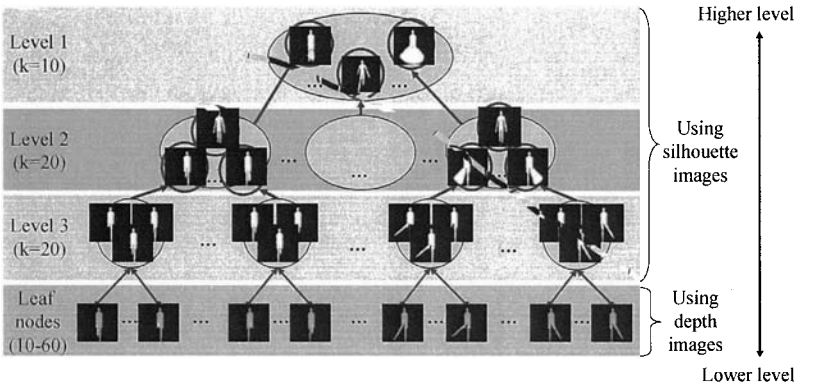


Figure 3. Building a hierarchical human body model database

3. Gesture Feature Representation

3.1. Feature Extraction

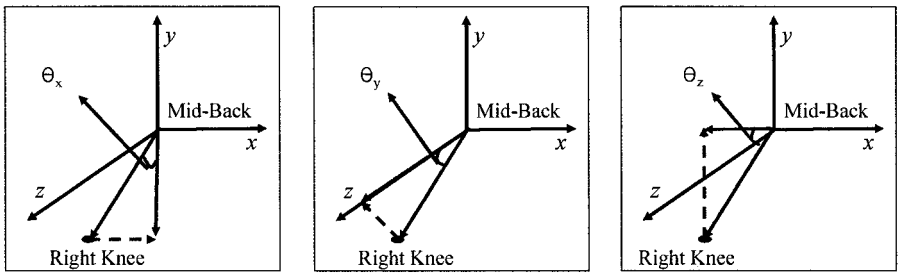
The information about body components in 3D allows us to locate various structural feature points around the body. Among them are the thirteen feature points are selected.

The angle from the vertical axis measured at the center of Mid-Back to each of the feature points are selected as features. The coordinates of the each body components are projected into *x*, *y* and *z* plane respectively to extract the features as shown in Fig. 4. The feature vector corresponding to the frame at time *t* is represented as follows:

$$X_t = [F_{L\_shoulder}, F_{L\_elbow}, F_{L\_Wrist}, \ldots, F_{R\_kee}, F_{R\_ankle}]^T, \quad X_t \in \Re^{36}, \quad t > 0$$

$$F_k = [\theta_x, \theta_y, \theta_z]$$

where F_k is the three angle values of the 3D human body component at k and the selected body points are left wrist, left elbow, left shoulder, right wrist, right elbow, right shoulder, left hip, left knees, left ankles, right hip, right knees, right ankles.



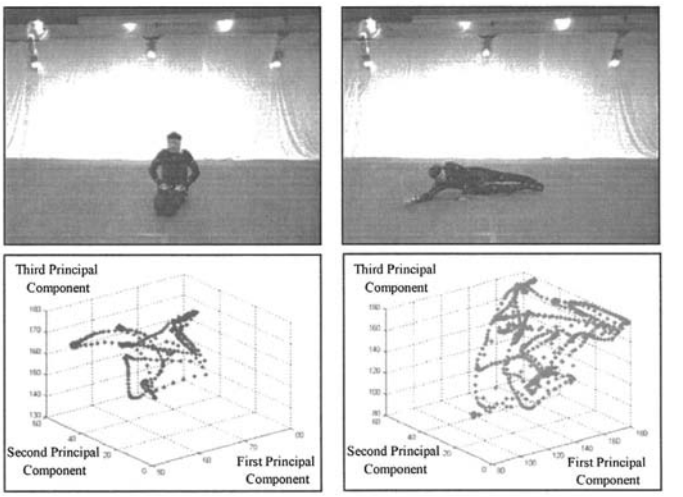
yz plane *xz* plane *xy* plane

Figure 4. Thirteen feature points extracted from each body component and the definition of angle features

3.2. Feature Clustering

Human motion including gestures can be represented as a sequence of feature vectors. The sequence of feature vectors constitutes a complex spatiotemporal trajectory in multi-dimensional space. Here the motion trajectory is considered as a sequence of vectors describing meaningful key gestures and meaningless inter-gesture motion.

Let us write $\mathbf{x}_t \in \Re^n$ to be a feature. Then a whole trajectory can be represented as a sequence of feature vectors as $X = \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3 \cdots \mathbf{x}_T$. Fig. 5 shows sample trajectories of two gestures in low three-dimensional subspace; PCA was done for visualization.



Sitting on the floor **Lying down on the floor**

Figure 5. Feature trajectories of two gestures in a reduced-dimensional subspace

The first step of feature processing to the gesture analysis is partitioning the feature space. To achieve this goal we divide a set feature vectors into a set of clusters. This allows us to model the trajectory in the feature space by

one in the cluster space. Different gestures have different cluster trajectories, even though different simple motion is in the same cluster in particular time. The technique of EM-based Gaussian mixture model (GMM) is employed as a means of clustering feature vectors.

4. Gesture Spotting and Recognition

4.1. Gesture and Transition Garbage Model

Since there are strong temporal constraints in gestures, we use left-right models rather than ergodic models for gesture HMM models [7]. The underlying state sequence associated with the left-right models has the property that as time increases the state index increases or stays the same, i.e., the states proceed from left to right. Clearly, the left-right type of HMM has the desirable property that it can readily model signals whose properties changes over time.

Conversely, the ergodic model for the transition gesture model is used, because it has to be able to represent all motions. The ergodic model is a fully connected HMM, that is, each state of the model can reach to all other states in a single transition. However, as the number of states of the transition gesture model increases, the structure of the transition gesture model will be more complex. This topological structure is simplified by introducing two null states, ST and ET that have no observations.

4.2. Gesture Spotting Model

In continuous human motion, gestures appear intermittently with meaningless connecting motion. There is no specific order among different gestures and any knowing when any gesture starts to appear and ends. We have defined the meaningless inter-gesture pattern as garbage. Then one way to define the alternating sequence of gestures and garbage is to construct a cascade connection of gesture HMMs and a garbage HMM repeatedly. A more effective structure is a circular interconnection of HMMs: gesture HMMs and then one or more garbage HMMs which are then connected to the start of the gesture HMMs. In this research, we designed the network shown in Fig. 6. We can easily expand the vocabulary by adding new key gesture HMM model and rebuilding transition gesture model. The detail model is described in [13].

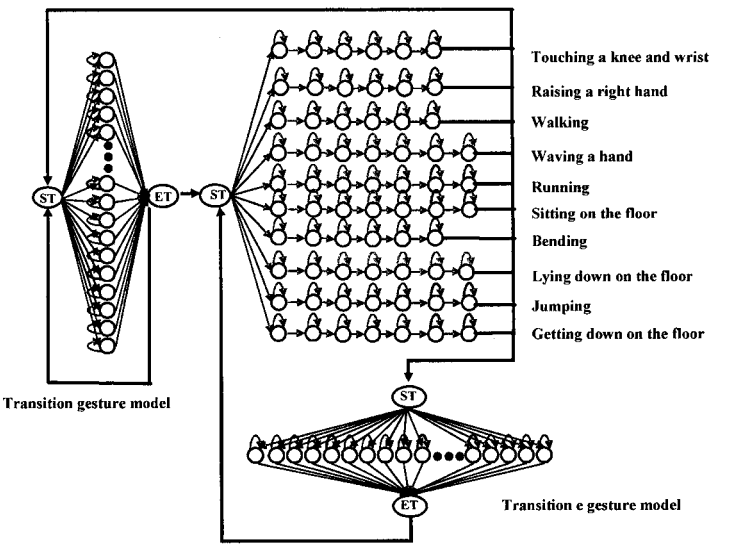


Figure 6. Key gesture spotting model

The transition gesture model is used as a confidence measures for gesture spotting. The confidence measure can be calculated using the transition gesture model as an approximation of $P(X)$.

$$P(X) = P(X \mid \lambda_G) \tag{3}$$

where X present an input observation sequences, and λ_G represents the HMM parameters of the garbage gesture model.

The transition gesture model emits the meaningless gestures. A gesture is spotted only if the likelihood of the best gesture model is higher than that of the garbage gesture model, represented by following equation. Let Λ be a set of gesture HMMs.

$$\forall g : p(X \mid \lambda_g) < p(X \mid \lambda_G), \quad g \in \Lambda \tag{4}$$

where λ_g represents the HMM parameters of the gesture model.

Therefore, the output likelihood of the transition gesture model can be used as an adaptive confidence measure for spotting. With the gesture spotting network, the start point and the end points of any gestures are found embedded in the input stream.

To retrieve the single best state sequence, $O_{t_1 t_2} = o_{t_1} o_{t_1+1} o_{t_1+2} \ldots o_{t_2}$, the Viterbi likelihood $p(O_{t_1 t_2}, Q^g_{t_1 t_2} \mid \lambda_g)$ with $Q^g_{t_1 t_2} = q_{t_1} q_{t_2+1} \ldots q_{t_2}$ being the 'best' state sequence in each HMM λ^g can be computed using the following relation:

$$\delta^g_t(j) = \max_{t_1} \{\Delta_t \pi_j b^g_j(o_t), \delta^g_{t_1, t-1}(j) a^g_{jj} b^g_j(o_t)\} \tag{5}$$

with the highest probability along a single path arriving at si at time t and accounting for the first observation as following induction.

$$\delta^g_t(j) = \max_i \delta^g_{t-1}(i) a^g_{ij} b^g_j(o_t) \tag{6}$$

For the backtracking information, $\psi^g_t(j)$ is used to keep the argument maximizing it for each t and j.

$$\psi^g_t(j) = \arg\max_i \delta^g_{t-1}(i) a^g_{ij}, \qquad \forall j, \ 2 \le t \le T \tag{7}$$

Finally, to uncover the most likely state sequence after the preceding computation, we must trace back to the initial state by following the Viterbi path [11].

5. Experimental Results

5.1. Experimental Data

For training the proposed gesture recognition method, KU Gesture database [3] is used. However, we need more data to test the proposed method so that we generated the gestures for adequate variation. The generated gestures were based on captured data and characterized by sufficient variation using eigengesture.

5.2. Robot Platform

The robot used in the proposed research, T-Rot, is personal service robot. T-Rot's aim is to support old men [9, 13]. Old men are not expert at operating robots; therefore, T-Rot is required to be able to interact naturally with old men, similar to the way human-human interaction takes place. Therefore, T-Rot has various interaction methods to provide natural interaction between a robot and its users.

As shown in Fig. 7, T-Rot is equipped with two stereo cameras, Videre STH-MDCS2, mounted on a pan-tilt unit. The cameras are located on T-Rot's head. The first has a 6 *mm* focal length and the second has a 12 *mm* focal length. The stereo cameras both have a resolution of 320×240. The second camera, with 6 *mm* focal length, is used to recognize gestures and the first camera, with 6 *mm* focal length, is used to recognize a face or object located near T-Rot. The height of the lens is approximately 1.3 *m* from the ground. T-Rot does not move when the gesture recognition module is running, so its body does not tremble. As a result, the captured image from the camera in T-Rot is adequate for recognizing gestures. The optimum distance for recognizing gestures is approximately 2~3 *m* from the subject.

In order to evaluate the proposed whole body gesture recognition in the real world, the proposed method has been integrated into the T-Rot.

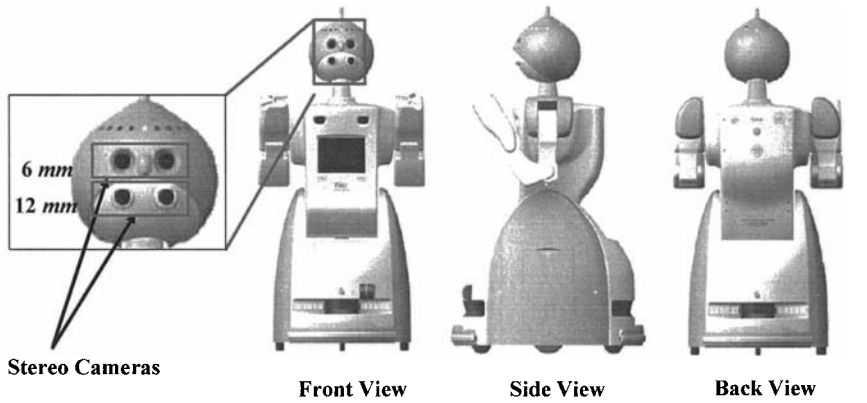


Figure 7. T-Rot, the robot used in the proposed experiments

5.3. Experimental Results

In general, most spotting tasks involve three types of errors, namely, substitution, insertion, and deletion errors. An insertion error occurs when the spotter reports a non-existent gesture. A deletion error occurs when the spotter fails to detect a gesture existing in the input stream. A substitution error occurs when an input gesture is classified into a wrong category. Following the convention, we measured the system performance in terms of those errors and the reliability. The overall performance is defined as:

$$reliability = \frac{\#\ of\ correctly\ recognized\ gestures}{\#\ of\ input\ gestures + \#\ of\ insertion\ errors} \times 100\%. \quad (8)$$

TABLE I
Key gesture spotting results

Gestures	N	N_{Hit}	N_{DE}	N_{SE}	N_{IE}	$R(\%)$
Walking	58	55	2	1	2	91.6
Running	62	59	1	2	3	90.7
Bending	54	54	0	0	0	100.0
Jumping	62	61	0	1	1	96.8
Lying down on the floor	61	58	1	2	2	92.0
Waving a hand	60	59	0	1	1	96.7
Sitting on the floor	62	58	2	2	3	89.2
Raising a right hand	62	62	0	0	0	100.0
Getting down on the floor	61	58	1	2	2	92.0
Touching a knee and wrist	60	60	0	0	0	100.0
Total	602	584	7	11	14	94.9

N : Number of input gestures
N_{Hit} : Number of correctly recognized gestures
N_{DE} : Number of deletion errors
N_{SE} : Number of substitution errors
N_{IE} : Number of insertion errors
R : Reliability

Table 1 shows the detailed result of the spotting test. Note that most of the errors are substitution and insertion errors. The substitution errors imply incorrect classifications, and the insertion errors imply incorrect segmentation and incorrect modeling of gesture patterns. The overall reliability with equal prior is 94.9 % as shown in the bottom row.

6. Conclusion and Further Research

This paper proposed an HMM-based method of spotting and recognizing gestures embedded in continuous whole body motion for human-robot interaction. The proposed method employs GMM clustering in feature space, producing efficient transition gesture models. Feature space clustering and the transition gesture HMM state reduction together form a highly efficient recognition network. The method of merging two states based on relative entropy and data dependent weighting allows the model to be more effective at capturing the variability in inter-gesture patterns. In fact, when compared with a recently proposed method, operating without explicit transition gesture modeling, a definite advantage was seen. In effect, the proposed transition gesture modeling is believed to be an excellent mechanism for recognizing gestures, as opposed to transition gesture, and rejecting these transition gestures.

This paper demonstrated that the proposed gesture recognition interface transcends to a much broader range of personal service robots. Near-term future work includes extending the proposed method for spotting and recognition of command gestures for HRI.

Acknowledgements

The author would like to thank OMRON Corp. for sponsorship of this talk.

This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

References

[1] J. Barbic, N.S. Pollard, J.K. Hodgins, C. Faloutsos, J.Y. Pan, and A. Safonova, "Segmenting Motion Capture Data into Distinct Behaviors," Proc. of Int'l Conf. on Graphics Interface, Ontario, Canada, 2004, pp. 17-19.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially. Interactive Robots," Robotics and Autonomous Systems, Vol. 42, 2003, pp. 143-166.

[3] B.-W. Hwang, S. Kim, and S.-W. Lee, "2D and 3D Full-Body Gesture Database for Analyzing Daily Human Gestures," Advances in Intelligent Computing, Lecture Notes in Computer Science, Vol. 3644, 2005, pp. 611-620, The KU Gesture Database, http://gesturedb.korea.ac.kr/.

[4] K. Kahol, P. Tripath, and S. Panchanthan, "Automated Gesture Segmentation From Dance Sequences," Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition, Seoul, Korea, 2004, pp. 883-888.

[5] H.-K. Lee and J.H. Kim, "An HMM-based Threshold Model Approach for Gesture Recognition," IEEE Trans. on Pattern Analysis and Machine Recognition, Vol. 21, No. 10, 1999, pp.961-973.

[6] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. of IEEE, Vol. 77, 1989, pp. 257-286.

[7] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, 1998, pp. 1371-1375.

[8] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural Human-Robot Interaction using Speech, Gaze and Gestures," Proc. of Int'l Conf. on Intelligent Robots and Systems, Sendai, Japan, 2004, pp. 2422-2427.

[9] T-Rot: Thinking Robot, Center for Intelligent Robotics, KIST, http://www.irobotics.re.kr.

[10] U.N. and I.F.F.R, "United Nations and the International federation of Robotics: World Robotics 2002," United Nations, New York and Geneva, 2002.

[11] A.J. Viterbi, "Error Bounds for Convolution Codes and an Asymptotically Optimum Decoding Algorithm," IEEE Trans. on Information Theory, Vol. 13, 1967, pp. 260-269.

[12] S. Waldherr, R. Romero, and S. Thrun, "A Gesture Based Interface for Human-Robot Interaction," Autonomous Robots, Vol. 9, No. 2, pp.151-173.

[13] H.-D. Yang, A.-Y. Park, and S.-W. Lee, "Gesture Spotting and Recognition for Human-Robot Interaction," IEEE Trans. on Robotics, Vol. 23, No. 2, 2007, pp. 256-270.

[14] H.-D. Yang and S.-W. Lee, "Reconstruction of 3D Human Body Pose from Stereo Image Sequences based on Top-down Learning," Pattern Recognition, Vol. 40, No. 11, 2007, pp. 3120-3131.