

A Wearable Camera System for Pointing Gesture Recognition and Detecting Indicated Objects

Tomohiro Mashita, Yoshio Iwai, Masahiko Yachida
Graduate School of Engineering Science
Osaka University

1-3, Machikaneyama, Toyonaka, Osaka 560-8531, JAPAN

Abstract

We propose a system for pointing gesture recognition and detecting indicated objects by using a vision sensor. By using random sampling and importance sampling, our method can track hands and estimate hand positions in real-time. By using the concepts of a cognitive origin and a reference plane, our system can also detect a direction to an indicated object. We use an omnidirectional vision sensor in order to cover the wide range of hand operations and movement of indicated objects. The camera is mounted on the head, which enables the system to be tolerant of the occlusion problem. The method for detecting an indicated object uses a linear model with the concepts of a cognitive origin and a reference plane.

1 Introduction

A human being can generate a new idea from others' suggestions by communicating with them. Everyone has had the experience of solving a problem which cannot be addressed alone but only by communicating with others. It is a fact that communication with others is an effective way to generate a new idea or activate creativity. In such a case, what means and information we use for communication with others is important. What place we exchange communications in is also important.

General means of communication are facial expression, gesture, and speaking. These are performed face to face. Only text and voice are used for communication when people are apart from each other. If a system is developed which can send and receive realistic facial expressions and gestures, we will be able to communicate with each other more effectively. The important components in such a system are a natural interface and a real-time process. The pointing gesture is one of the natural interfaces for man-machine interaction. We think that a system becomes more effective in communication if the system can find an indicated object.

A vision system is suitable for such a natural interface since this involves passive sensing and gestures of the user can be estimated without any discomfort. Moreover a wearable camera system has some advantages. One of them is that the user wearing the system can freely move anywhere because the camera moves with the user if the camera has a sufficient field of view. In the case of a fixed camera, the user must ensure that his/her position is in the fixed camera's field of view. Another advantage is that it is possible for the wearable camera to have the same field of view as the user. Because of these advantages, a head-mounted omnidirectional camera is suitable for our system in order to activate communications. We therefore propose a wearable camera system for pointing gesture recognition and estimating orientation of an indicated object in real-time.

Our proposed system can also avoid occlusion problems by using a head-mounted camera and by capturing images from the top. A 3D hand position is estimated from the hand regions by giving the lengths of arms in advance. The hand is detected and tracked quickly by random sampling and importance sampling. The pointing gesture is also recognized in real time. We use a camera with an omnidirectional sensor (HyperOmni Vision)[4] in order to cover the wide range of hand movement and indicated objects. The method for detecting an indicated object uses a linear model and the concepts of a cognitive origin and a reference plane.

1.1 Related work

Many methods for human posture and gesture estimation have been proposed. These methods are divided into two classes: one uses a monocular camera and the other uses multiple cameras.

A method using a monocular camera[19, 20] has the advantage that the system is simple, but has the disadvantage that the occlusion problem occurs frequently. Absolute depth information cannot be estimated, so

some posture estimation methods require known parameters such as link length, initial position, and so on. There is another posture estimation method[1] which uses constraints and evaluation functions without known parameters, but this method takes much time for computation.

A method using multiple cameras[2, 3] is tolerant of the occlusion problem because it has a wide field of view and absolute depth information is estimated by feature matching. By feature matching, a 3D position can also be estimated. However, a system using multiple cameras is more complex and takes much time for feature matching.

Many tracking methods using vision sensors have been proposed. Especially, the color model is important in the case of tracking skin regions such as face, hand, and body. The distribution of skin color has been studied by Yang et al.[5]. They concluded that human skin colors cluster in a small region in a color space, that human skin colors differ more in intensity than in color, and that under a certain lighting condition, a skin color distribution can be characterized by a multivariate normal distribution in the normalized color space. Therefore, many methods for visual tracking using a target’s color have been proposed[6, 7, 8, 9, 11, 2]. For more robust tracking, a target’s shape is also used with performing edge detection[6, 10].

MCMC (Monte Carlo Markov Chain)[12] is also used for visual tracking[13]. MCMC has the advantage that an object moving randomly can be tracked because MCMC keeps the distribution of a moving object. By using factor sampling or importance sampling, tracking can be processed quickly. The method in this paper also uses importance sampling and also has the above advantage.

The pointing and spatial cognition have been studied in the field of neuroscience and cognitive science for a long time [14, 15]. In these years, the pointing gesture has also been studied in the field of human interface and robotics because it is an effective interface for man-machine communication [16, 17, 18]. We think that pointing gesture is one of a natural interface for man-machine interaction too. A method proposed by Kahn and Swain [16] uses positions of head and pointing hand to detect an indicated object. When a human being points an object with looking at the object, the pointing hand doesn’t overlap with the object in his/her view. It is easy to understand that the indicated object isn’t on the line from the eyes or head through the hand. At the present time, the knowledge of pointing in reaching distance increases,

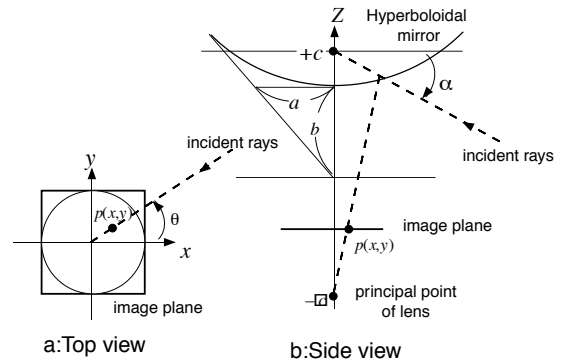


Figure 1: HyperOmni Vision

but the knowledge in walking distance doesn’t increase sufficiently, because the relation between an indicated object in walking distance and the pointing gesture is not clear yet.

In our method for indicated object detection, we use the concepts of the cognitive origin and the reference plane. The estimation of a cognitive origin decide the direction for indicated object. The reference plane limits the cognitive origin’s movable dimension under our assumption that a cognitive origin moves in the reference plane. We defined the relation of a position of cognitive origin in the reference plane and an arm posture by linear. Our method can detect the direction for the indicated object by the model.

We explain a pointing gesture model for detecting an indicated object, a camera model, and a color model in section 2, the algorithms for tracking and posture estimation in section 3 and the system which implements the proposed algorithm in section 4. We conduct experiments to evaluate our proposed method in section 5, and conclude in section 6.

2 Models

2.1 Camera model

We use an omnidirectional image sensor for observation of pointing gestures in our system. This sensor covers the wide range of hand operations and has the same optical characteristics of a common camera, so we can easily estimate the object position. We briefly explain the sensor below.

The sensor consists of a camera fixed upward and a hyperboloidal mirror fixed downward. A hyperboloidal plane of the mirror is expressed by equation 1. A hyperboloidal plane has two focal points: $(0, 0, +c)$ and $(0, 0, -c)$. The hyperboloidal mirror is fixed at the upper focal point, $(0, 0, +c)$, and a focal point of the camera is fixed at the lower focal point, $(0, 0, -c)$,

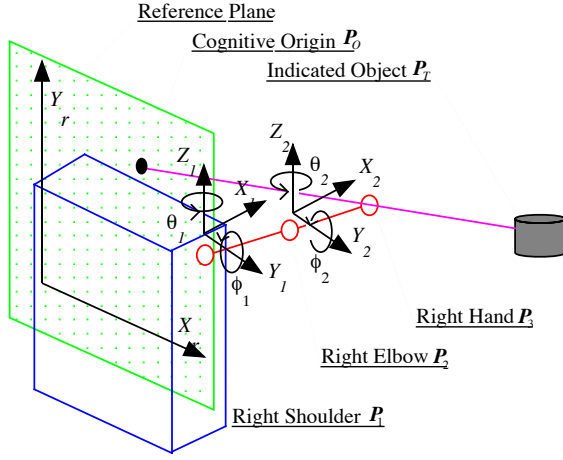


Figure 2: The pointing gesture model

as shown in figure 1. The image plane, xy , is parallel to the XY -plane, and the image plane is fixed at $(0, 0, f - c)$. f is a focal length of the camera.

$$\frac{X^2 + Y^2}{a^2} - \frac{Z^2}{b^2} = -1, \quad (1)$$

$$c = \sqrt{a^2 + b^2},$$

$$Y/X = y/x, \quad (2)$$

$$Z = \sqrt{X^2 + Y^2} \tan \alpha + c, \quad (3)$$

$$\alpha = \tan^{-1} \frac{(b^2 + c^2) \sin \beta - 2bc}{(b^2 - c^2) \cos \beta}, \quad (4)$$

$$\beta = \tan^{-1} \frac{f}{\sqrt{x^2 + y^2}}, \quad (5)$$

where a, b are parameters of a hyperboloidal plane, α is a depression angle, and β is an angle between the optical axis and projected point (x, y) .

2.2 Color model

There are many methods for visual tracking using an object color model or using an object contour model. Most color models are static, C_μ, C_Σ , or even supports temporal change of color, $C_\mu(t), C_\Sigma(t)$.

Our system evaluates the sampling points by using a color model assumed to be a normal distribution. In our system, the method for color evaluation uses RGB color space and static C_μ and C_Σ of each target.

2.3 Pointing gesture model

The pointing gesture model is shown in figure 2. The cognitive origin, P_O , is a point in 3D space, and we assume that the indicated object, P_T , is on the line extending from the cognitive origin to the end of the pointing arm, P_3 . Parameters θ_1 and ϕ_1 are the directions of the vector from P_1 to P_2 , and θ_2

and ϕ_2 are the directions of the vector from P_2 to P_3 . We define that the cognitive origin, P_O , lies in the reference plane. The coordinates of the cognitive origin in the reference plane, (x_r, y_r) , are determined from the posture of the pointing arm. We define a pointing gesture as that in which the positions of a user's right shoulder P_1 , right elbow P_2 and right hand P_3 are collinear; in other words, the case in which $\theta_2 = 0$ and $\phi_2 = 0$. θ_1 and ϕ_1 are calculated from a vector from shoulder P_1 to hand P_3 . We assume that (θ_1, ϕ_1) and (x_r, y_r) are linear, as expressed by the following equation:

$$\begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} \theta_1 & \phi_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_1 & \phi_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix}, \quad (6)$$

where $\alpha_1, \dots, \alpha_6$ are coefficients, θ_1 and ϕ_1 are observed parameters estimated from the arm posture.

3 Algorithms

3.1 Tracking

The proposed method uses a centroid of colored region for depth estimation in order to determine the vector, V_i , from focal point of the mirror, $(0, 0, +c)$, to P_i . The region of interest (ROI) should be small in order to track the region in real time, but the region should be large enough in order to detect it robustly. The size of the ROI is the cost-time trade-off. While many studies for tracking have been done, it is accepted that random sampling reduces computation time more than processing a whole image. We therefore use random sampling and importance sampling for tracking regions, and also use the color information of the sampling points for object detection.

For random sampling, sampling points are uniformly selected in an input image. The color information of the sampling point is compared with the object color model. Random sampling is used for detection and estimation of hand regions when hand regions have not yet been detected.

For importance sampling, sampling points are uniformly selected in the ROI of an input image. Importance sampling is used for estimation of a centroid of the hand region more accurately. The probability that a sampling point is on the tracked object can be defined as follows:

$$P = \left\{ 1 - \left(\frac{R - T_a}{R} \right)^N \right\} P_r(R), \quad (7)$$

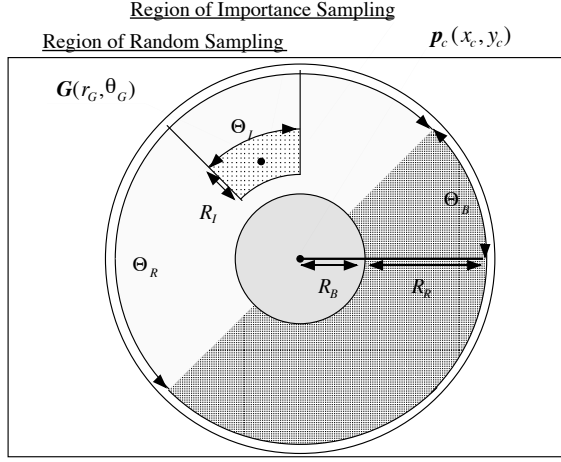


Figure 3: Region of sampling

where T_a is the area of the target, N is a number of sampling points, R is the area of ROI, and $P_r(R)$ is the probability that the target exists in the ROI. In the case where the area of the target, T_a , cannot be enlarged, N should be increased or R should be reduced or $P_r(R)$ should be enlarged in order to gain the detection probability P . Much computation time, however, is required when N is increased. R should therefore be reduced while keeping the probability $P_r(R)$ high. In order to reduce the area of R , we predict a position of the hand region from the previous position, and then we sample data around the predicted point.

As we use a hyperboloidal mirror, sampling points are not simply picked up uniformly in the xy -image coordinate system, but picked up uniformly in the $r\theta$ -polar coordinate system. One of the problems of using a hyperboloidal mirror is that the object becomes small when the object gets close to the center of image. The area of the object is dependent on not only the distance between the camera and the object, but also the depression angle. XY -uniform random sampling on such a sensor causes an estimation bias which drifts an estimate outward. We addressed this problem by performing random sampling uniformly in the $r\theta$ -polar coordinate system. Sampling points therefore become sparse in proportion to r . Figure 3 shows an example of sampling regions. In figure 3, $G(r_G, \theta_G)$ is a position of a target in the previous frame, Θ_I and R_I are parameters of a region for importance sampling, Θ_B and R_B are bias parameters, Θ_R and R_R are parameters of the region for random sampling, and

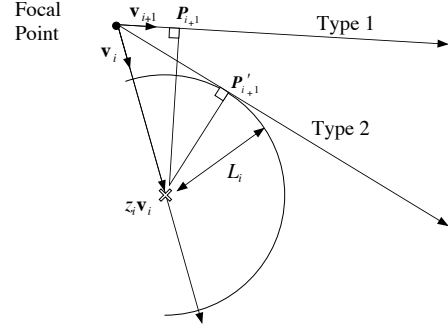


Figure 4: Solution adjustment in case of imaginary number

$p(x_c, y_c)$ is the center of an omnidirectional image.

3.2 Position estimation of targets

A target is tracked by comparing the color models, C_μ, C_Σ , with the color of a sampled point, $C(x_i)$, determined by random and importance sampling. The dissimilarity, $D(x_i)$, is defined by the following equation:

$$D(x_i) \equiv (C(x_i) - C_\mu)^T C_\Sigma (C(x_i) - C_\mu). \quad (8)$$

Sampling point, x_i , is a target if the dissimilarity is less than a certain threshold value, K . The area, S , and the centroid, G , of the target are estimated by the following equation:

$$d(x_i) = \begin{cases} 1 & D(x_i) < K \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

$$S = \sum_{i=0}^N d(x_i), \quad (10)$$

$$G = \frac{\sum_{i=0}^N d(x_i) x_i}{S}, \quad (11)$$

where N is the sum of sampling points. In the case of $S = 0$, the target cannot be found and importance sampling is stopped. The positions of tracked centroids are used for depth estimation of the arm.

3.3 Depth estimation

The human arm model is shown in figure 2. In figure 4. $z_i \mathbf{V}_i \equiv (X_i, Y_i, Z_i)$ expresses the 3D position of each joint, P_i . \mathbf{V}_i is the unit vector from the focal point of the mirror $(0, 0, +c)$ to the joint. z_i is the depth of the joint. L_i is the length of the arm.

We can estimate the depth of each joint and then calculate the 3D position of the joint from equation 2 and equation 3 by giving the lengths of arms, L_i , and the vectors to the joints, \mathbf{V}_i . In the case of HyperOmni

Vision, \mathbf{V}_i is calculated from $\mathbf{p}(x_i, y_i)$ by the following equation:

$$\mathbf{V}_i = \begin{bmatrix} \cos \theta \cos(-\alpha) \\ \sin \theta \cos(-\alpha) \\ \sin(-\alpha) \end{bmatrix}, \quad (12)$$

where, α and θ are obtained from equations 2 and 3.

The following constraint is established when the arm length, L_i , is given:

$$\|z_i \mathbf{V}_i - z_{i-1} \mathbf{V}_{i-1}\| = L_i, \quad (13)$$

where \mathbf{V}_i is a view line from the focal point of the mirror to joint i . The above equation can be explicitly solved as follows:

$$z_i = (\mathbf{V}_i, \mathbf{V}_{i-1})z_{i-1} \pm \sqrt{[(\mathbf{V}_i, \mathbf{V}_{i-1})^2 - 1]z_{i-1}^2 + L_i^2}. \quad (14)$$

All the positions of links are determined relative to z_1 from the above equation. The solution of equation 14 is real when the following condition is satisfied:

$$[(\mathbf{V}_i, \mathbf{V}_{i-1})^2 - 1] z_i^2 + L_i^2 \geq 0. \quad (15)$$

$(\mathbf{V}_i, \mathbf{V}_{i-1})^2 \leq 1$ and let $-\zeta^2 \equiv (\mathbf{V}_i, \mathbf{V}_{i-1})^2 - 1, \zeta > 0$, then we obtain

$$0 < z_i \leq L_i/\zeta. \quad (16)$$

When the solution is an imaginary number, we calculate the solution again by using the following equation:

$$z_i = \sqrt{L_i^2 - z_{i-1}^2}. \quad (17)$$

The geometrical explanation is shown in figure 4.

4 Implementation

The outline of our system is shown in figure 5. The sum of random sampling points and importance sampling points is fixed at constant N , so the computation is performed at a constant rate. The importance sampling points are picked up around the region centering a centroid of a tracking target.

The position of shoulder and the length of arms are given in advance. At this moment in time, the system has up to 4 solutions because of sign ambiguity. The correct posture is decided at the part of the pointing gesture recognition. For learning $\alpha_1, \dots, \alpha_6$, test subjects pointed 75 targets with sitting on a chair in advance. The color parameters, $C_\mu(t)$ and $C_\Sigma(t)$, are also learned in advance.

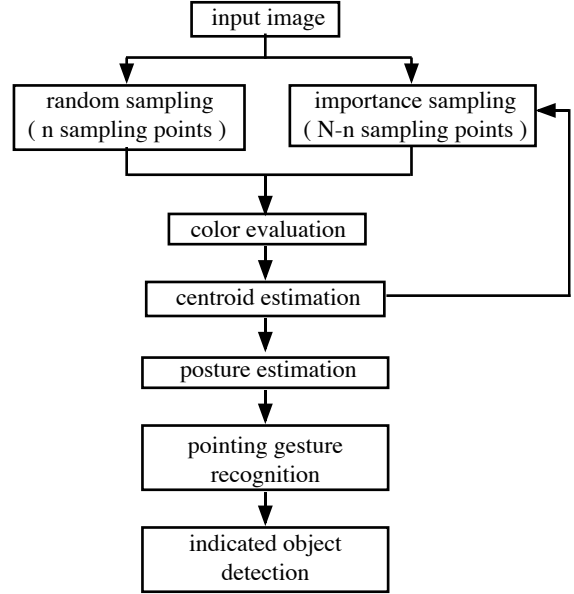


Figure 5: System flow

5 Experimental Results

We conduct experiments on real images by using an SGI workstation (Onyx2, MIPS R12K, 400 MHz) with an omnidirectional sensor (TOM-02-0013, Suekage). The size of an image is 720×243 pixels. The lengths of all arms, L_i , are given in advance, and the root depth, z_1 , is also given.

Figure 6 shows the marker tracking and posture estimation using importance and random sampling. Figure 7 shows the estimated line from the hand to the indicated object. The line projected onto an the omnidirectional image is shown in figure 7. In this figure, the indicated object is a cardboard box. Though the line maintains a slight distance from the target, the distance between the target and the line is acceptable because the system searches for the indicated objects around the line.

6 Summary

We proposed a wearable camera system using the head-mounted omnidirectional camera for pointing gesture recognition and detecting an indicated object. Our system can estimate arm posture and detect an orientation of an indicated object. The advantages of our system are that the system runs in real-time and the camera can shoot an arm and an indicated object. Another advantage is that the user wearing the system can freely move anywhere. We think that it is acceptable to find an indicated object by searching for it around the line estimated from a pointing gesture

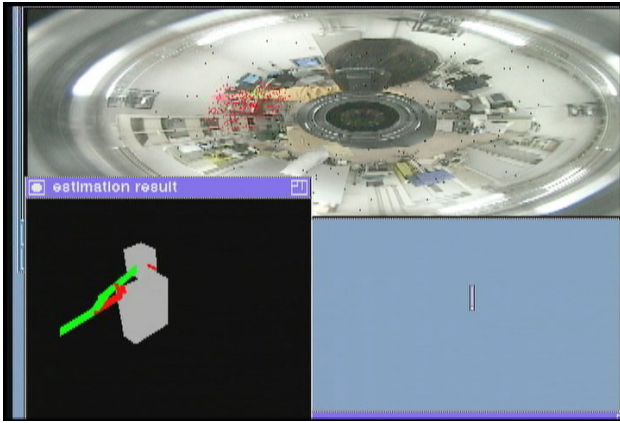


Figure 6: Examples of tracking and posture estimation

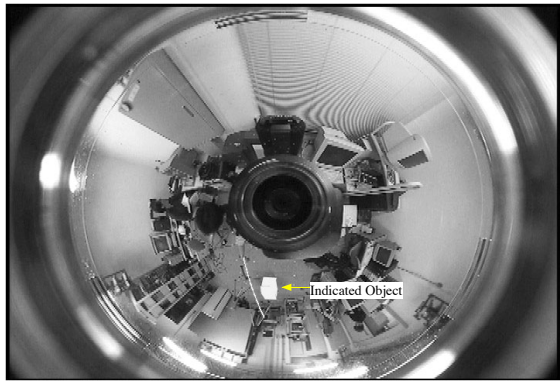


Figure 7: Omnidirectional image and an estimated direction

in an omnidirectional image.

References

- [1] C. Pan and S. Ma, "3D motion estimation of human by genetic algorithm," *ICPR* Vol. 1, pp. 159-163, 2000.
- [2] C. Wren, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *PAMI*, Vol. 19, No. 7, pp. 780-785, July 1997.
- [3] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *ICPR*, pp. 601-608, California, June 1998.
- [4] K. Yamazawa, Y. Yagi, and M. Yachida, "Omnidirectional imaging with hyperboloidal projection," *IROS*, Vol. 2, pp. 1029-1034, 1993.
- [5] J. Yang, L. Weier, and A. Waibel, "Skin-color modeling and adaptation," *ACCV*, pp. 687-694, 1998.
- [6] Y. Wu and T. S. Huang, "A Co-inference approach to robust visual tracking," *ICCV*, Vol. 2, pp. 26-33, 2001.
- [7] L. Sigal, S. Sclaroff, and V. Athitsos, "Estimation and prediction of evolving color distributions for skin segmentation under varying illumination," *CVPR*, Vol. 2, pp. 152-159, 2000.
- [8] Y. Raja, S. J. McKenna, and S. Gong, "Tracking and segmenting people in varying lighting conditions using colour," *AFGR*, pp. 228-233, 1998.
- [9] N. Oliver, A. Pentland, and F. Bérard, "LAFTER: Lips and face real time tracker," *CVPR*, pp. 123-129, 1997.
- [10] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," *CVPR*, pp. 232-237, 1998.
- [11] P. Feiguith and D. Terzopoulos, "Color-based tracking of heads and other mobile objects at video frame rates," *CVPR*, pp. 21-22, 1997.
- [12] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, New York, 1996.
- [13] M. Isard and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," *ECCV*, Vol. 1, pp. 893-908, 1998.
- [14] Brain W. Russell, "Visual disorientation with spatial reference to lesions of the right cerebral hemisphere," *Brain*, Vol.64, pp. 244-272, 1941.
- [15] J. McInTyre, F. Stratta, and F. Lacquaniti, "Viewer-centered frame of reference for pointing to memorized targets in three-dimensional space," *J. NeuroPhysiol.*, Vol. 78, pp. 1601-1618, 1998.
- [16] R.E. Kahn, M.J. Swain, P.N. Prokopowicz, and R.J. Firby, "Gesture Recognition Using Perseus Architecture," *CVPR*, pp. 734-741, 1996.
- [17] S. Waldherr, R. Romero, and S. Thrun, "A Gesture Based Interface for Human-Robot Interaction," *Autonomous Robots*, Vol. 9, No. 2, pp. 151-173, 2000.
- [18] D. Kortenkamp, E. Huber, and R. P. Bonasso, "Recognizing and Interpreting Gestures on a Mobile Robot," *AAAI*, Vol. 2, pp. 915-921, 1996.
- [19] J. O'Rourke and N. I. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *PAMI*, Vol. 2, No. 2, pp. 522-536, 1980.
- [20] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP*, Vol. 59, No. 1, pp. 94-115, 1994.