

Human Sign Recognition for Robot Manipulation

Leonardo Saldivar-Piñon, Mario I. Chacon-Murguia,
Rafael Sandoval-Rodriguez, and Javier Vega-Pineda

Visual Perception Applications on Robotic Lab,
Chihuahua Institute of Technology, Mexico
{Lsaldivar, mchacon, rsandova, jvega}@itchihuahua.edu.mx

Abstract. This paper addresses the problem of recognizing signs generated by a person to guide a robot. The proposed method is based on video color analysis of a moving person making signs. The analysis consists of segmentation of the middle body, arm and forearm location and recognition of the arm and forearm positions. The proposed method was experimentally tested on videos with different target colors and illumination conditions. Quantitative evaluations indicate 97.76% of correct detection of the signs in 1807 frames.

Keywords: sign recognition, robot manipulation, video segmentation.

1 Introduction

Human – Robot interaction (HRI) is an ongoing area that has become increasingly important because robots have expanded their presence on human activities. HRI can be considered as the interdisciplinary study of the dynamic interaction among humans and robots [1]. HRI involve researches and specialists from different areas like engineering, computer science, social and human sciences, among other [1]. From the human-robot relation point of view, the human user may be classified in several levels. The higher level corresponds to the specialist followed by skilled workers, unskilled workers, handicap, general public and children [2]. Rogers [3] proposes another classification for human-robot relation based on three taxonomies: numeric relation, spatial relation and authority relation. Still another type of classification is based on the human responsibility of the robot functionality [4]. This classification involves the following categories: manual control, manual control with intelligent assistance, negotiated control and supervised control.

The HRI relations aforementioned are found in areas like; person-operator interaction, automation, science fiction [3], cooperative work, cognitive science [5], ethology, emotion and personality. The person-operator relation involves two types of classification; man-centered [6] and machine-centered [7]. Cognitive science in robotic is related to the study of how and individual (person/ robot) may acquire information or abilities from other individual. A highly inspired work environment based on ethology, psychology and cognitive development was built to design motivational systems for autonomous robots [8].

Some tools and techniques used for HRI in these days are; keyboards, touchscreens, written symbols, sound commands, helmets, gloves, pressure sensors, EEG

signals, voice commands, body signs, etc. Most of these techniques have been made possible thanks to the evolution of personal computers [9]. From the previous techniques body signs recognition using vision systems is a research area with an increasing interest because it offers a natural and robust interaction. Besides, it provides a more human-like communication form. Another advantage of using body signs recognition by vision systems is that it avoids the use of artificial devices. Some works following the previous have been reported in the literature. Horain and Bomb [3] present a vision system to detect gestures. Khan et al. [4] developed a stereo vision system for face identification and arm gesture detection. Salti et al. [9], proposed an algorithm to detect the arms using color skin analysis. Motion detection based on human body descriptors is reported by Hsuan et al. [6] and Siddiqui and Medioni [7] presented an algorithm to detect and track arms, hands and faces based on color face detection.

Research on HRI is a paramount issue considering that currently there not exists a robot with a complete autonomous capacity. Therefore, advances on HRI area rely on the development of new investigation. The work reported in this paper deals with the recognition of human signs generated with arm positions to guide a robot. The work is based on video color analysis to detect a person, segmentation of the middle body and recognition of arm and forearm positions. After an initialization stage, the output elements of the proposed method are used to guide a robot in real time activating its hardware mechanisms. The remainder of this paper is structured as follows. Section 2 establishes the experimental framework. Section 3 describes the middle body detection and sign recognition methods. Finally Section 4 and 5 analyze the results and comments the conclusions of this research.

2 Experimental Framework

The data base for experimental analysis consists of 9 videos with persons in front of the camera at a distance between 1.5 and 3 meters and using a resolution of 240x320 pixels. Six videos were acquired with the Sony Cybershot camera in an environment with poor illumination. The other 3 videos were acquired with the webcam Microsoft VX-600 in an office environment with good illumination. Figure 1 shows representative frames of some videos as well as the number of frames in it. It is assumed that in each video a person is continuously doing signs. Therefore, there is not absence of the person in any video frame. Also, a transition state can be detected during the change of one sign to another. This state was considered by default as a stop sign.

3 Middle Body Detection and Sign Recognition

The proposed method consists of 4 stages, image preprocessing, middle body segmentation, feature extraction and sign recognition. The first step consists on the location of the person's middle body using color information and Mahalanobis distance [3]. Once the middle body is located, the arms and shoulders are detected based on

modifications of conditions reported in [4] and [10]. Once the previous body parts are detected their relative positions are measured and the sign is determined based on these positions.




Video	Sample	Video	Sample	Video	Sample
Video 1 210 frames		Video 3 233 frames		Video 9 196 frames	

Fig. 1. Video data base samples

3.1 Image Preprocessing

This initialization stage includes a training step which corresponds to manually indicate to the algorithm a sample of the color region, R , which will be used in the segmentation step. Based on R a binary mask $M(x,y)$ of the same size as the video resolution is derived from this region

$$M(x, y) = \begin{cases} 1, & p(x, y) \in R \\ 0, & p(x, y) \notin R \end{cases} \quad (1)$$

where $p(x,y)$ denotes a pixel. Figure 2 illustrates this process. Before executing the next steps a space color transformation, RGB to HSV, is achieved.

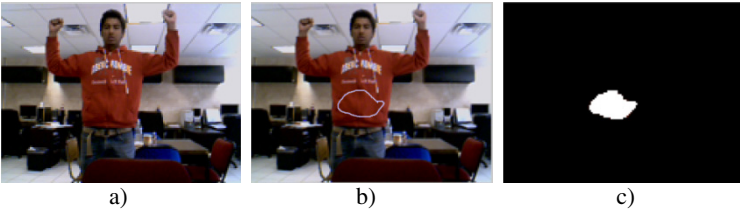


Fig. 2. Mask generation, a) Original frame, b) Region sample R , c) $M(x,y)$

The color pixel values of the HSV planes are extracted from R and stored as follows

$$\begin{aligned} M_H &= H(x, y) \text{ if } (x, y) \in M(x, y) \\ M_S &= S(x, y) \text{ if } (x, y) \in M(x, y) \\ M_V &= V(x, y) \text{ if } (x, y) \in M(x, y) \end{aligned} \quad (2)$$

The mean value for each of the HVS planes is computed

$$\mu_{M_H} = \sum_{k=0}^{n_m} M_H(k) / n_m \quad \mu_{M_S} = \sum_{k=0}^{n_m} M_S(k) / n_m \quad \mu_{M_V} = \sum_{k=0}^{n_m} M_V(k) / n_m \quad (3)$$

where n_m is the number of pixels in $M(x,y)$. Next, the Mahalanobis distance between the frame pixels and the mean of R is computed by

$$D_{Mh}(x, y) = \sqrt{\left(p_i(x, y) - \mu_j \right)^T C^{-1} \left(p_i(x, y) - \mu_j \right)} \text{ for } i = \{H, S, V\}, j = \{M_H, M_S, M_V\} \quad (4)$$

where C^{-1} is the covariance matrix of the pixel in R . D_{mh} tends to decrease as the image pixels become more alike to the training region R and vice versa. In the distance image $I_{DMh}(x, y)$ similar color pixels are mapped to black.

In order to segment the image $I_{DMh}(x, y)$ must first be binarized. The threshold τ used for the binarization is determined as the right minimum value of the Mahalanobis histogram distances of the pixels inside R . Applying this threshold to $I_{DMh}(x, y)$ the binary image $I_B(x, y)$, (5) is obtained, Figure 3.

$$I_B(x, y) = \begin{cases} 1 & I_{DMh}(x, y) \leq \tau \\ 0 & I_{DMh}(x, y) > \tau \end{cases} \quad (5)$$

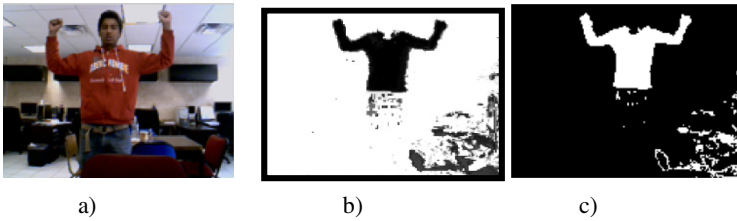


Fig. 3. a) Original frame, b) Mahalanobis distance, c) Binarization of $I_{DMh}(x, y)$

3.2 Middle Body Segmentation

The segmentation process works with the image $I_B(x, y)$ obtained in the preprocessing stage. The first operation in this part is to clean up $I_B(x, y)$ of possible noisy regions not corresponding to the middle body, and to fill hole regions. Noisy regions are eliminated using a threshold area of 1250 pixels. This threshold is determined according to the visual field of the camera and the distance of the person to the camera as explained in Section 3.1. This process is shown in Figure 4.

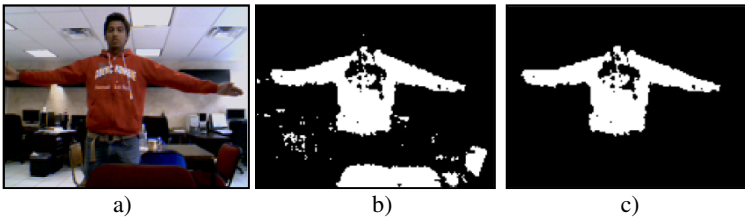


Fig. 4. a) Original frame, b) Binary image, c) Area thresholding

The resultant binary image of the threshold operation is next processed by the following three morphologic operations: dilation by a disk of radius 3, hole filling and opening. Despite these three operations some videos still showed unwanted regions, therefore a new additional process is applied. This process consisted in keeping as desirable only the region closer to the center of the image. This was accomplished by

measuring the distances of the regions centroids to the center of the view field. Finally, in order to obtain only the region of the middle body a new opening operation is performed with a rectangular element of 40×20 . The complete process is illustrated in Figure 5.

3.3 Arm and Forearm Detection

Arm detection starts with the shoulder location based on the middle body center of mass. The shoulders are considered to be in a specific orientation with respect to the middle body center of mass and their distance is proportional to the area of the middle body.

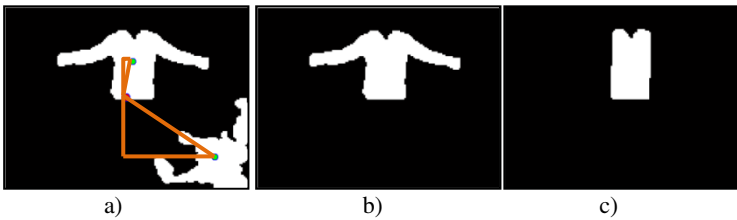


Fig. 5. a) Distance to the field of view center calculation, c) Extra region eliminated, d) Detection of the middle body

Main body points used for body analysis are the ones described in [11]. In this work, only the body upper points are used, shoulders, elbow and wrist. The information of distance used to determine relative positions of shoulder – elbow and elbow – wrist were defined from analysis of 20 frames in each of the 9 videos. In Table 1 a summary of the previous distances with respect to the middle body area is presented. The shoulder locations using the information of Table 1 is shown in Figure 6. The arm is located by a modification of the method proposed in [4]. Instead the proposed blocks used in [4], the arm is located by tracing a set of lines with the length defined in Table 1. The first step to approximate the arm position is using a line $L1$ of length $D2$ as shown in Figure 7a. $L1$ starts in the shoulder oriented to the right to find out the right elbow and oriented to the left to find out the left elbow. The search process is done by tracing $L1$ in a range of different angles ϕ_2 , from -70° to 90° in steps of 5° , this is shown in Figure 7b.

Table 1. Shoulder, elbow and wrist distances with respect to the middle body area

Middle body area	Distance		
	Shoulders centroid	Shoulder -elbow	Elbow- wrist
>2700	26	25	21
>2400	24	22	19
>1500	22	20	17
<1501	19	17	14

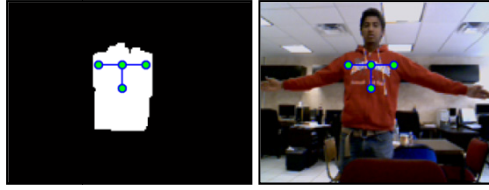


Fig. 6. Location of the center of mass and shoulders

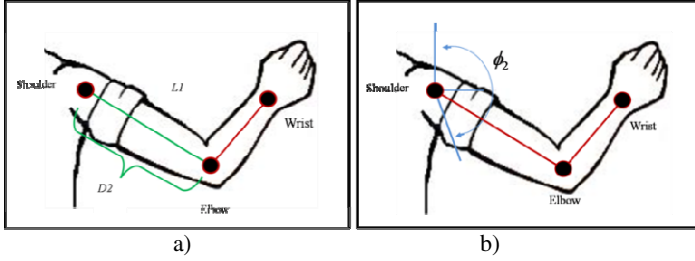


Fig. 7. a) Line search LI , b) Limits of the search ϕ_2

Line LI is traced from the starting point (x_h, y_h) with angle ϕ_2 to the ending point (x_{f1}, y_{f1}) using:

$$x_{f1} = x_h - \cos(\phi_2) * D2 \tag{6}$$

$$y_{f1} = y_h - \text{sen}(\phi_2) * D2 \tag{7}$$

The number of points in LI is 10 testing points α_a , Figure 8, with a distance

$$D3 = D2 / 10 \tag{8}$$

Each testing point α_a in each line LI contains the pixel value of the binary image. Therefore if the sum of the testing points α_a is 10 the traced line is inside the arm. Since several traced lines may be inside the arm, its estimated orientation, ϕ_{pa} , is computed by the average value of ϕ_2 including all the corresponding LI lines inside the arm, A .

$$\phi_{pa} = \sum \phi_{va} / \|A\| \tag{9}$$

where $\phi_{va} = \{ \phi_2 \in A \}$ and $\| \cdot \|$ stands for cardinality.

Once the arm orientation, ϕ_{pa} , is determined, Figure 8, the elbow location, (x_e, y_e) , is computed by:

$$x_e = x_h - \cos(\phi_{pa}) * D2 \tag{10}$$

$$y_e = y_h - \text{sen}(\phi_{pa}) * D2 \tag{11}$$

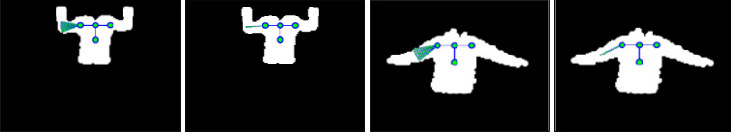


Fig. 8. Arm and elbow location

The wrist position is computed from the elbow position. The method is similar to the one used to compute the arm and elbow positions but using the tracing line $L3$ with length $D4$ and the following equations

$$x_{f3} = x_e - \cos(\phi_2 + \phi_1) * D4 \quad y_{f3} = y_e - \sin(\phi_2 + \phi_1) * D4 \quad (12)$$

The orientations of the tracing lines go from from -10° to 100° with steps of 5° as shown in Figure 9. The number of points in $L3$ is 10 testing points α_w . To verify if a line $L3$ is inside or outside the forearm and its final position, we used the same scheme used in the arm case. Since several traced lines may be inside of the forearm, their estimated orientation, ϕ_{pw} , is computed by the average value of ϕ_{vw} including all the corresponding $L3$ lines inside the forearm, B .

$$\phi_{pw} = \sum \phi_{vw} / \|B\| \quad (13)$$

where $\phi_{vw} = \{ \phi_2 \in B \}$, B is the set of $L3$ lines that hold the condition over α_w .

Once the orientation of the forearm, ϕ_{pw} , is determined, Figure 10, the wrist location, (x_w, y_w) , is obtained by

$$x_w = x_e - \cos(\phi_{pw}) * D4 \quad y_w = y_e - \sin(\phi_{pw}) * D4 \quad (14)$$

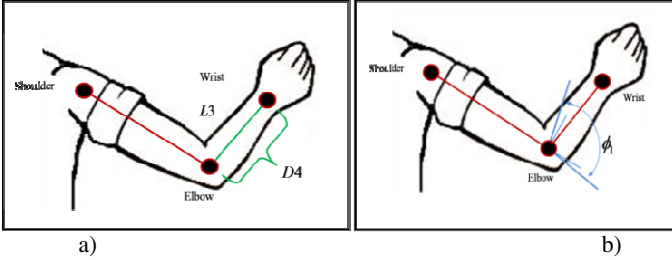


Fig. 9. a) Line search $L3$, b) Limits of the search ϕ_l

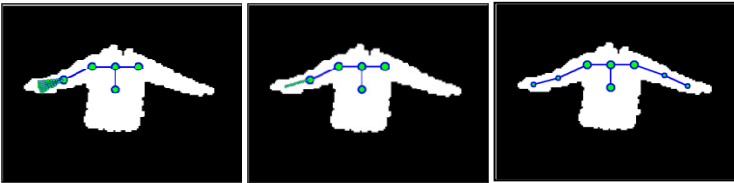







Fig. 10. Forearm and wrist location

3.4 Benchmark Signs for Recognition

Five different signs, taken from the works reported in [8] and [12], were considered for recognition: forward, back, stop, right and left. The definitions of the five signs are presented in Table 2. The signs are determined by the angles of the arm and forearm, where ϕ_{2d} , ϕ_{2i} , ϕ_{1d} and ϕ_{1i} are the angles with respect the horizontal of the right arm, left arm, right forearm and left forearm respectively.

Table 2. Sign definition

Sign	Position	Limits	Sign	Position	Limits
Forward		$25^\circ < \phi_{2d} < -25^\circ$ $25^\circ < \phi_{1d} < -25^\circ$ $25^\circ < \phi_{2i} < -25^\circ$ $25^\circ < \phi_{1i} < -25^\circ$	Back		$25^\circ < \phi_{2d} < -25^\circ$ $100^\circ < \phi_{1d} < 60^\circ$ $25^\circ < \phi_{2i} < -25^\circ$ $100^\circ < \phi_{1i} < 60^\circ$
Stop		$-40^\circ < \phi_{2d} < -75^\circ$ $-50^\circ < \phi_{1d} < -90^\circ$ $-40^\circ < \phi_{2i} < -75^\circ$ $-50^\circ < \phi_{1i} < -90^\circ$	Right		$25^\circ < \phi_{2d} < -25^\circ$ $25^\circ < \phi_{1d} < -25^\circ$ $-40^\circ < \phi_{2i} < -75^\circ$ $-50^\circ < \phi_{1i} < -90^\circ$
Left		$-40^\circ < \phi_{2d} < -75^\circ$ $-50^\circ < \phi_{1d} < -90^\circ$ $25^\circ < \phi_{2i} < -25^\circ$ $25^\circ < \phi_{1i} < -25^\circ$			

4 Experimental Results

The proposed method was implemented in MATLAB running in a personal computer with CORE 2 DUO E4 400 processor at 2 GHz and Windows XP 32 bits. Figure 11 illustrates several cases including the different steps of the method: segmentation, centroid and shoulder location, forearm location and sign recognition. The performance of the method was computed by the analysis of 1807 frames. Also, in Figure 11 some results with respect the arms and forearms detection as well as the sign recognition is shown. It is considered that there is a sign present in each frame. The percentage value corresponds to hits conditions (good recognition). The complement of the percentage value is the miss rate (failure recognition). When a frame contains a transition state, that is, a body position in transition from one sign to another is also considered as a stop sign. Results shown in Table 3 correspond to the numerical performance of the method; they indicate a good sign recognition performance considering the wide variety of scene conditions present in the video. Regarding the processing time, the average time was 0.0645 seconds per frame with a maximum of 0.1 seconds and a minimum of 0.058 seconds. Considering the average time, the algorithm can do the processing of 15 frames per second which allows real time processing in some applications.

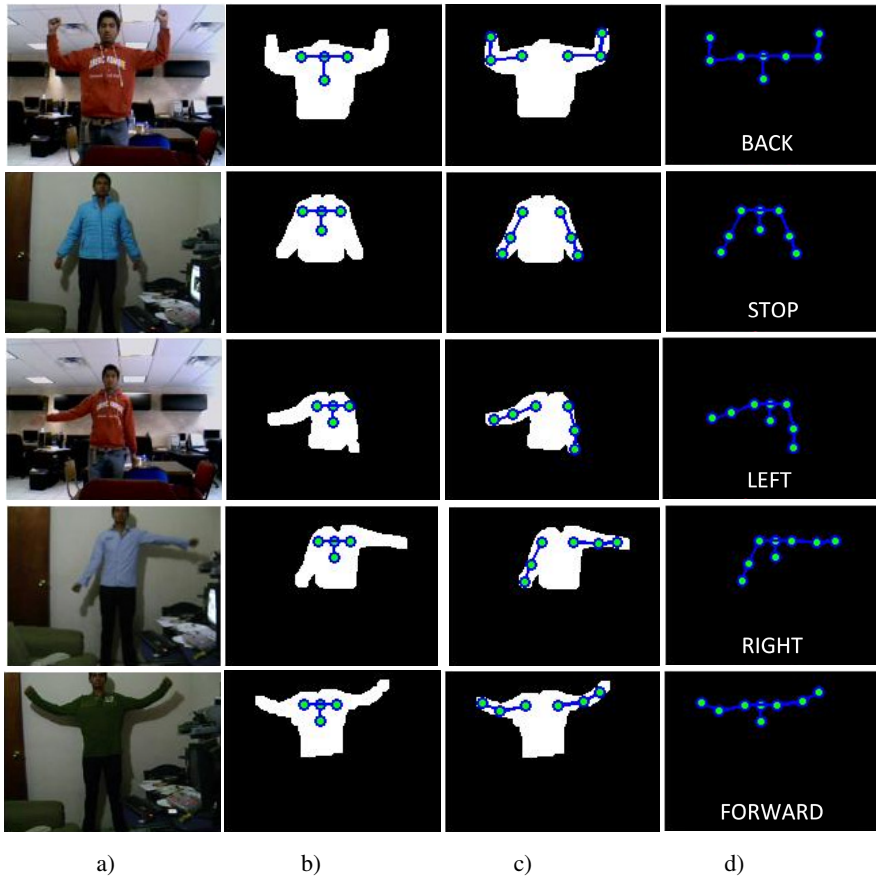


Fig. 11. a) Original frame, b) Centroid and arm location, c) Forearm location, d) Sign recognition

Table 3. Performance of the proposed method

	Frames	Detection rate				
		Left arm	Left forearm	Right arm	Right forearm	Sign
Video 1	210	100%	99.52%	100%	99.05%	99.05%
Video 2	238	100%	100%	100%	100%	100%
Video 3	233	100%	100%	100%	100%	100%
Video 4	125	100%	100%	100%	100%	100%
Video 5	125	100%	100%	100%	100%	100%
Video 6	246	100%	100%	100%	91.46%	91.46%
Video 7	212	100%	100%	100%	100%	100%
Video 8	198	100%	98.48%	100%	95.96%	95.96%
Video 9	196	100%	96.94%	98.4%	96.43%	93.37%
Average		100%	99.44%	99.83%	98.10%	97.76%

5 Conclusions

The paper presented a new approach for human sign recognition based on color video analysis and arm and forearm location. Accordingly to the presented results, the method turned to be efficient with respect to the computation time performance as well as with its capability for recognition. The proposed method was implemented to guide a mobile robot with acceptable results in real time.

Acknowledgements. The authors thanks to Fondo Mixto de Fomento a la Investigación Científica y Tecnológica CONACYT-Gobierno del Estado de Chihuahua, by the support of this research under grant CHIH-2009-C02-125358.

References

1. Feil-Seifer, D.J., Mataric, M.: Human-Robot Interaction. In: Encyclopedia of Complexity and Systems Science, pp. 4643–4659. Springer, New York (2009)
2. Khamis, A.M.: Interacción Remota Con Robots Móviles Basada En Internet, Universidad Carlos III de Madrid, Madrid, España, Doctoral Thesis (2003)
3. Roger, C.: Asimov's Laws of Robotics: Implications for Information Technology Part 1. IEEE Computer Society 22(12), 53–61 (1993)
4. Draper, V.: Environmental Restoration and Waste Management Program Teleoperator Hand Controllers: Contextual Human Factors Assessment, OAK Ridge National Laboratory, Departamento de Energia de los Estados Unidos, Reporte (1994)
5. Hee-Deok, Y., A-Yeon, P., Seong-Whan, L.: Gesture Spotting and Recognition for Human-Robot Interaction. IEEE Transaction on Robotics 23(2), 256–270 (2003)
6. Khan, I.R., Miyamoto, H.: Face and Arm-Posture Recognition for Secure Human-Machine Interaction. In: Systems, Man and Cybernetics, International Conference, pp. 411–417 (2008)
7. Salti, S., Schreer, O., Stefano, D.: Real-time 3D Arm Pose Estimation from Monocular Video for Enhanced HCI. In: Proceeding of the 1st ACM Workshop on Vision Networks for Behavior Analysis, Vancouver, Canada, pp. 1–8 (2008)
8. Siddiqui, M., Medioni, G.: Robust Real-Time Upper Body Limb Detection and Tracking. In: 4th ACM International Workshop on Video Surveillance & Sensor Networks, Santa Barbara, California, USA (2006)
9. Patrick, H., Mayank, B.: 3D Model Based Gesture Acquisition Using a Single Camera. In: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, Orlando, Florida, pp. 158–164 (2002)
10. Shaker, S., Saade, J., Asmar, D.: Fuzzy Inference-Based Person-Following Robot. International Journal of Systems Applications, Engineering and Development 2(1), 29–34 (2008)
11. Chen, H., Chen, T., Chen, Y., Lee, S.: Human Action Recognition Using Star Skeleton. In: 4th ACM International Workshop on Video Surveillance & Sensor Networks, Santa Barbara, California, USA (2006)
12. Tarokh, M., Kuo, J.: Vision Based Person Tracking and Following in Unstructured Environments. Department of Computer Science, San Diego State University, San Diego, California, U.S.A.