Color vision based person following with a mobile robot

Bogdan Kwolek

Abstract. ***The work presented in this paper describes a vision based interface that enables the mobile robot to detect and follow a person through office environments. The proposed algorithm works by estimating the separate distributions of colors on a person's face and shirt. Using heuristics on vertical alignment of face and shirt color as well as bordering on each other, we are able to extract the person reliably and to guide the robot to desirable places. Fast and adaptive scheme of color processing has been tested in changing lighting conditions and using very long image sequences. The elaborated algorithm utilizes a linear Kalman filter to perform smooth tracking. The algorithms we have developed have been implemented and tested on a mobile robot equipped with an active camera. The developed system is able to track a person at approximately 10 frames per second.***

***Index terms*. Human Computer Interaction, Color Image Processing, Mobile Robot Control.**

I. Introduction

Interfaces which take advantage of speech and pattern recognition are in center of interest of many companies and research centers connected with design and applying of new devices. Particularly, rapid qualitative progress refers to service robots and interfaces which enable collision free movement in unknown environment as well as programming of task to the realization through commands given by voice and by a way of demonstration. A demonstration is particularly useful at programming of new tasks by non-expert user. It is far easier to point at an object and demonstrate a track which robot should follow during transferring a load than to verbally describe its exact location and possible to accomplish the path of movement. Arm-pestures are an easy way to convey a geometrical description to the robot [16].

Mobile robot applications of vision modules impose several requirements and limitations on the use of known vision systems. First of all, the vision module needs to be small enough to mount on the robot and to derive enough small portion of energy from the battery, that would not take place a significant reduction of working time of the vehicle. Additionally the system must operate at an acceptable speed and may not assume a static background or a fixed location of the camera and particularly constant lighting conditions. Naturally, the system should be user-independent and to ensure safety and a collision free movement for surrounding and the user. Natural way of communication with human requires a rate of 10+ frames per second and several score milliseconds' delays.

This paper describes a vision based low-level interface that has been used to conduct several experiments consist of following a person with autonomous robot in natural office environment. The presented system is being implemented as part of currently realized interactive interface for mobile robot programming by demonstration and instruction.

One of the goals of the work reported here was to investigate the usability of known color processing algorithms in the context of an application of mobile robot equipped with typical laptop computer and analog frame-grabber for realization of tasks consist of following a person. General-purpose algorithms are not robust and usually not algorithmically efficient [10] to implement a vision system of mobile robot.

Finding and tracking people is the core of any vision based gestural commands recognition system. Visual tracking of people has been studied extensively over past few years [8, 12, 15]. However, the majority of existing approaches assumes that camera is mounted at a fixed location. Such approaches typically rest on static background and in these circumstances the human motion can be detected through e.g. image differencing. To realize experiments consist of following with mobile robot equipped with an active camera, we propose an approach to detecting and estimating position of the person facing the robot.

A. Related work

In work [16] a fast and adaptive algorithm for tracking and following a person by robot-mounted camera has been described. After locating the person in front of the robot an initial probabilistic models of the person's shirt and face colors are created. The tracking is realized on the basis of combination of such colors and it has been assumed that they are arranged vertically. The system uses window of fixed size to adapt the face and shirt color models and it is essential that these windows can only contain face and shirt colors. In particularly, the distance between the robot and the person must be approximately fixed. The Perseus [17]

Rzeszów University of Technology, W. Pola 2, 35-959 Rzeszów,
bkwolek@prz.rzeszow.pl

system is capable of finding the object pointed to by a person. The system assumes that people is only moving object in the scene. Perseus uses independent types of information e.g. feature maps and disparity feature maps. The distance between the robot and person may vary. Pfinder [12] uses adaptive background subtraction and pixel classification to track people in a static environment. A people body is modeled as connected sets of Gaussian blobs. These distributions are used to track the various body parts of the person. In other system [18] a behavior for a person following with an active camera mounted on robot has been presented. In that system the head of person is located using skin color detection in the HSV color space and color thresholding techniques. Well established methods of color distribution modeling, such as histograms and Gaussian mixture models [8] have enabled the construction of suitably accurate skin filters. However such techniques are not ideal for use in adaptive real-time applications with moving camera.

B. *Our approach*

Our color-based tracking system has been implemented on 850 MHz Pentium equipped with a PCMCIA analog frame-grabber connected with the Sony EVI-D31 active camera. The robot and camera's panning, tilting, and zooming are controlled by laptop computer via a serial port.

We focus on a person detection method that uses color as the primary feature. The color is a powerful fundamental cue that can be used as a first step in the process of moving object detection in complex scene because color image segmentation yields real-time performance in standard hardware while being relatively robust to changes over illumination [14]. On the other hand, color segmentation methods do not require many parameters. Robustness is achieved by separating the chrominance from the luminance in the original color image.

Our experiences which were attained during evaluation experiments with color tracking have inclined us to confirm the one of the conclusion from work [16] that the color as the only feature is insufficient for reliable tracking people with a mobile robot in office environment with naturally occurring circumstances. Therefore apart from skin color we included into the tracking the second color typically aligned with a person's head: shirt color.

Our algorithm works by estimating two distributions: skin colors on a person's face and colors on shirt. This information is then used to detect which parts of the image belong to the face and shirt and which are the part of the background. Two adaptive color models have been used to perform tracking under varying illumination conditions. It has been found necessary to use at least two pairs of color models, one for interior lighting and one for exterior natural daylight.

By applying two probabilistic detectors of colors to each pixel, we can obtain two probability images, where each pixel is represented by the probability that it belongs to the user's face or shirt, respectively. The probabilities have been then used to segment the candidates of person's face and shirt from the background. We have then included at a thresholding stage only pixels with a great likelihood of belonging to the face or the shirt classes. At the next stage we have applied a connected component analysis. The connected component analysis was used to gather to groups the adjacent pixels whose probability had been over a threshold. The areas, coordinates of centers of gravity and geometrical relations of labeled skin-like regions with shirt-like regions have been then used in detection of the person within an image sequence. This approach guarantees that only one person is tracked at a time. Particularly, the information about similarity of features of extracted face and of shirt from the last frame and the one before has an influence on accommodation rate in continuous adaptation of color models over time.

The output of the visual module is the position of face in image and its area. The control system of the camera uses the position of face to drive the pan and tilt angles. The goal is to keep the face of the tracked person in specific location in the image. The actual pan angle of the camera has been used as input for rotation controller of the robot. This controller should minimize the angle between the axis of the camera and of the robot. The mentioned above control strategy allows us to achieve smooth behaviors of the robot in response to a rapid movement of person. The controller of the linear velocity of the robot maintains the area of face detected in image on desirable level while the person is coming closer to the robot or moving further away.

These techniques, taken together, have proven useful in building an almost real-time version of the robotic system suitable for demonstration purposes, which can process 320x240 color images in 100 to 120 milliseconds, depending on the image complexity. We show how these steps can be performed quickly and robustly, with the result that we can realize experiments in unstructured environments and operate in real-time.

II. FILTERING THE TARGET IN COLOR IMAGES

In this section, the use of color as a cue for detection and tracking is described. Our approach tracks a person on the basis of combination of two colors, face color and shirt color. Face color as a feature for tracking people has been used for example in work described by Yang and Waibel [15]. The advantage of such an approach consists in the speed at which present low-cost personal computers can process color images obtained through a frame-grabber. That aspect is particularly important considering on the one hand the limited computational power of computer being in disposal and on the other hand the necessity of control with rate enables the mobile robot to achieve smooth movement.

The most common space used in computer displays and CCD cameras is the RGB color space. Each pixel is represented by a combination of R (red), G (green), B (blue) which are usually called three primary colors. Since the brightness is not useful for an extraction of a target, we can represent color of target in the chromatic color space. Chromatic colors, also known as pure colors are defined in the following manner [9, 13]

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B} \tag{1}$$

This two dimensional intensity-normalized chromaticity color space is relatively robust to the change of the illumination. But an obvious shortcoming is that the pure colors are very noisy if they are under low intensities. Tseng and Chang [11] suggest splitting up the color image into chromatic and achromatic areas. The criteria for achromatic areas were measured by experimental observation of human eyes and they are defined by two zones

zone I: (intensity $>$ 95) or (intensity $\leq$ 25)

zone II: (81 $<$ intensity $\leq$ 95) and (saturation $<$ 18)

(61 $<$ intensity $\leq$ 81) and (saturation $<$ 20)

(51 $<$ intensity $\leq$ 61) and (saturation $<$ 30)

(41 $<$ intensity $\leq$ 51) and (saturation $<$ 40)

(25 $<$ intensity $\leq$ 41) and (saturation $<$ 60)

where the intensities are in range [1,100] and the saturation in range [1,180]. In our approach we have used the following criterion: (intensity $\geq$ 95) or (intensity $\leq$ 10) or (saturation $\leq$ 10). The achromatic areas have not been considered in detection of person and have been marked as background. To shorten the conversion time of image from the RGB color space to *rg* color space we applied lookup tables. The criterion for achromatic areas has been mapped to *rg* color space and then reflected in the conversion tables.

Filtering on the basis of the following mask

1	2	1
2	4	2
1	2	1

that weights correspond to the binomial coefficients preceded the normalization of colors [3]. Convolution is associative thus the convolution has been realized twice with the following three element masks

1
2
1

and

1	2	1

Since the weights in the mentioned masks are powers of two, bitwise shift operation has been used instead of multiplication as well as division by the normalization coefficient [3].

We decided to solve our task of human following on the basis of person's face position to the center of the image. The normalized color space is sufficient to extract a face from the background when the background color distribution differs significantly from the face color distribution. When in normalized color space the background color distribution is similar to the face color distribution, the effect is poor even if background color and face color differ each other in the RGB color space. The reason is that some information is lost when the image is converted from 3-dimensional color space to low-dimensional space as well as because of noise, quantization errors.

The usage of two colors for tracking of a person improves of smoothness of movement and makes navigation problem easier because of reduction of an identification of an accidental object for the tracked person and thus lower amount of standstills of the robot in searching for the tracked person. The second color used in our experiments is typically aligned vertically with a person face. We assumed that the tracked face is always located above a person's shirt and has a common border with it. The combination of two colors and mentioned above geometrical relations are sufficient to detect the specific person during following. We decided to use very simple model of colors taken into account during a tracking and therefore the second color should be relatively homogenous. If such a color is relatively homogenous, its distribution can be directly characterized by using a single Gaussian model $M = (\mu_r, \mu_g, \Sigma)$ [15] with the means and variances determined in the following manner

$$\mu_r = \frac{1}{N}\sum_{i=1}^{N} r_i, \quad \mu_g = \frac{1}{N}\sum_{i=1}^{N} g_i, \quad \Sigma = \begin{pmatrix} v_{rr} & v_{rg} \\ v_{gr} & v_{gg} \end{pmatrix}$$

$$v_{rr} = \frac{1}{N}\sum_{i=1}^{N}(r_i - \mu_r)^2, \quad v_{gg} = \frac{1}{N}\sum_{i=1}^{N}(g_i - \mu_g)^2 \tag{2}$$

$$v_{rg} = v_{gr} = \frac{1}{N}\sum_{i=1}^{N}\left((r_i - \mu_r)(g_i - \mu_g)\right)$$

The parameters of the models have been obtained on the basis a set of training images. Each image with user outlined regions of interest has undergone a color space transformation. The aim of such an operation was to obtain a representative sample of typical colors of objects and thus it was not essential to outline the whole target.

One attraction of bimodal normal distribution is that it can be used to generalize on small amounts of training data. The extraction process of the colors of target is analogous to the single hypothesis classifier described in [2]. Single hypothesis classifier deals with problems in which one class is well defined while others are not. Let $\zeta = [r\ g]^T$ denote the feature vector consists of color components of a pixel and ω_s denote the skin or shirt class, respectively. Thus the probability that pixel belongs to class ω_s can be expressed as

$$p(\zeta|\omega_s) = \frac{1}{2\pi\sqrt{\det(\Sigma_s)}} e^{-\frac{1}{2}(\varsigma-\mu_s)^T \Sigma_s^{-1}(\varsigma-\mu_s)} \tag{3}$$

where μ_s is the mean color and Σ_s is the covariance matrix of considered class. The quantity d in $d^2 = (\zeta - \mu_s)^T \Sigma_s^{-1} (\zeta - \mu_s)$ is the Mahalanobis distance. A small value of d indicates a high pixel probability and vice-versa.

As it was mentioned above, the extraction of target relies on two distinctive colors: face color and shirt color. To extract the person's face and shirt, each pixel is examined by two models. The better the pixel matches color model, the higher the probability and response of such a color filter are, see fig. 1 c, d.

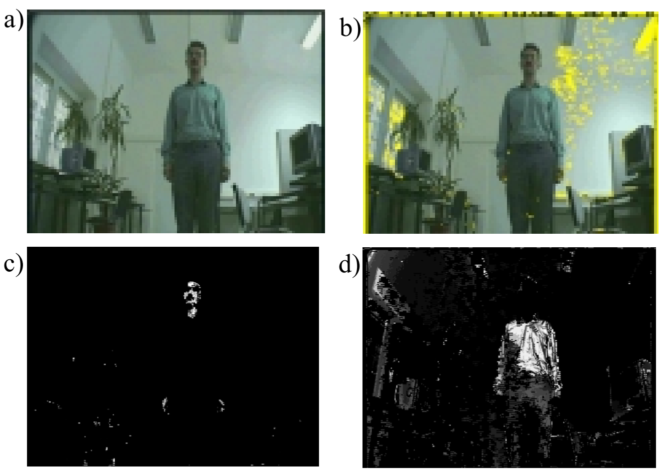

Fig. 1. Detecting a person: a) the original image shown as graylevel image, b) the original image with marked achromatic regions, c) face probability picture, d) shirt probability picture

III. Model adaptation over time

When the robot moves and pans during the following a person, the apparent appearance of target changes due to lighting fluctuation, shadows, occlusions, image noise an so on. The seeming color of face as well as shirt change as the relative positions among robot, person and light vary. During experiments we noticed that the model adaptation is essential to cope with different lighting conditions. The simple way to deals with this important aspect is to use a linear combination of the known parameters to predict or approximate new parameters. In our approach we used a simple adaptive filter [16]

$$\mu^t = \alpha\mu + (1-\alpha)\mu^{t-1}, \quad \Sigma^t = \alpha\Sigma + (1-\alpha)\Sigma^{t-1} \tag{4}$$

which computes the new parameters of Gaussian at time step t from the old values μ^{t-1}, Σ^{t-1} and the measured new values μ, Σ with a weighting factor α. Once initialized from models outlined by hand, the means and variances are estimated recursively using the previous values and the new means and variances of the extracted face and shirt. The accommodation rate α has been set to 0.05, 0.025 or 0 with according to similarity of areas of extracted face and of shirt from the last frame and the one before. The aim of this simple heuristics was to change the rate of accommodation in the case of a rapid change of lighting conditions. The accommodation rate has been set to zero when the detected area was comparatively little to the area set up during the building of the color model.

IV. Recognition Module

Two probability pictures have been obtained on the basis of face and shirt filters. A pixel was identified as a candidate of face or shirt if the corresponding probability was greater than a threshold. After an image has been thresholded, morphological closing [5] has been performed. The aim of such an operation was to fill small holes and to smooth border. The intention of the next stage was to find a common pixel or even better a common area between two regions coming from last two dilated pictures. In practice this operation has been performed through an overlapping of the dilated pictures and a comparison of values of pixels from equal coordinates. Thanks to dilation, the face and shirt have meet on some area and thus their common neighborhood was determined more reliable. The area that was lying just below the chin was usually in shadow and color filter has classified such a region as background and therefore occasionally it was necessary to realize a dilation of pictures several times until an achievement of the common pixel.

Having the common pixel of the face and the shirt which was obtained in above manner, the connected component analysis has been performed. The objective of this stage was to get individual scene objects which we could analyze further.

Connected component labeling scans an image and groups its pixels into components basing on sharing the similar intensity values by pixels [4,5]. It works by scanning an image, pixel-by-pixel, from top to bottom and left to right in order to extract pixels and regions which share the appropriate set of intensity values. In our approach connected component algorithm works on closed images and is based on 8-connectivity [4,5]. At each point the algorithm examines pixel p from closed image and if that pixel is set to foreground its four neighbors from labeled image which have already been processed (i.e. the neighbor to the left of actually analyzed pixel, above it, upper left and right). Having obtained values of pixels one of the following can occur:

if all four neighbors are 0, assign a new label to pixel p, else
if only one neighbors has a label, assign its label to pixel p, else
if one or more of the neighbor have two different labels, assign to p a label which is smaller and update the glue table for equivalencies.

An each label is replaced in the second pass by a unique label which had been assigned to each region and stored in the glue table in the previous pass.

The regions extracted in such a way were used to calculate areas and gravity centers [5] of detected candidates of face and shirt. Then after the examination of areas and geometrical relation of gravity centers, the extraction of face was finalized or a next pair of candidates has been analyzed. Finally, the accommodation rate α has been set to one of the mentioned in previous section values.

V. TRACKING OF TARGET IN COLOR IMAGE SEQUENCE

To smooth the temporal trajectories of the face center position as well as to avoid a "jump" from the tracked face to another we have included in our system the Discrete Kalman Filter [1]. The Kalman filter is a recursive, linear optimal filter for state estimation in dynamic systems corrupted by noise. A system state can be a single variable or a vector that describes the characteristics of the system. In our approach the Kalman filter has also been used to predict the face position. To decrease the amount of computation per image the prediction was used in determination of the position of bounding window centered around the tracked individual. Prediction of the position was computed using the following approximate model of a moving object

$$\xi_k = A\xi_{k-1} + w_k, \ \ \eta_k = C\xi_k + v_k \tag{5}$$

where k denotes the sample time ($t_{k+1} = t_k + T$; T is the sample period), $\xi_k = [X_k, \dot{X}_k, Y_k, \dot{Y}_k]^T$ is the system state $\eta_k = [X_k, Y_k]^T$ is the measurement, X_k and Y_k indicate the center of the face, $\dot{X}_k$, $\dot{Y}_k$ are the velocities, w_k and v_k are disturbance noises assumed to be described by zero mean, Gaussian mutually independent noises with covariances Q and R, respectively.

Matrixes of state A and of measurements C in the accepted model have a form resulting from assumed constant speed in sampling period

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \ C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{6}$$

The recursive equation for the prediction of the window center is given as

$$\hat{\xi}_{k/k-1} = A\hat{\xi}_{k-1/k-1} \tag{7}$$

The estimates $\hat{\xi}$ are defined by the Kalman filter algorithm

$$\hat{\xi}_{k/k} = \hat{\xi}_{k/k-1} + K_F\left(\eta_k - C\hat{\xi}_{k/k-1}\right) \tag{8}$$

where the Kalman gain K_F can be computed off-line. The proper selection of the input dynamic disturbance noise covariance matrix Q and the measurement noise covariance matrix R is very important. The covariances are usually determined by experiments. The initialization problem of the Kalman filter of vision based systems for human motion tracking is widely discussed in [6]. To make sure that the tracked person completely lies within the predicted bounding window, this window was increased by a certain factor. The estimates of face position have been used by control module of the active camera.

VI. CONTROL STRATEGY

The output of elaborated visual module is the position in the image of face to be tracked as well as its area. The mentioned values are written asynchronously by vision module VM in block of common memory, which can be easily accessed from Saphira client [7], see fig. 2. Saphira supports a packet-based communications protocol for sending commands to the robot server receiving information back from the robot.

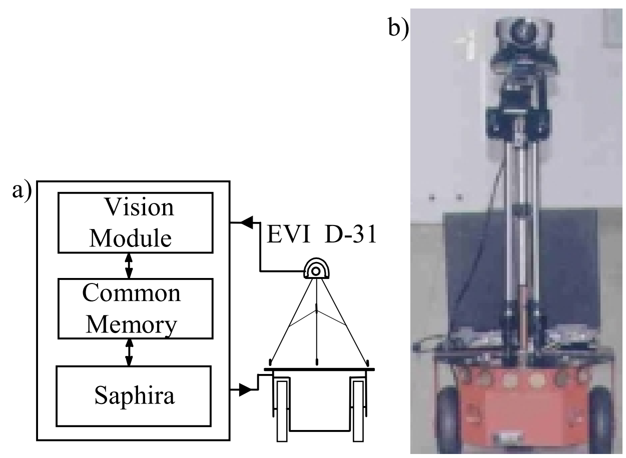

Fig. 2. Scheme of system architecture (a), robot Pioneer II (b)

Control of the camera has been based on the position of the tracked face in the image. The controller should keep the central point of the face on the horizontal position located in the half of picture and vertical position located in 4/5 of height of picture. To achieve this a PD controller of the following form has been used

$$y_t = K_P x_{w,t} + K_D \frac{x_{w,t} - x_{w,t-1}}{T} \tag{9}$$

where $x_{w,t}$ is the reference signal at time t. K_P, K_D are control coefficients, T is the sampling period. The actual pan angle of the camera has been used as input for the rotation controller of the robot. The rotation controller should minimize the angle between the axis of the camera and of the robot.

Since we have been using only one camera, it is difficult to obtain accurate estimates of the absolute distance of the person to the camera and thereby the robot. However, single camera is sufficient to compute the relative depth (i.e. the depth relative to some reference location). With this information it is possible to tell whether the person is coming closer or moving further away. One way to compute this is to simply determine the total area of the labeled face. As the face moves closer to the camera it will occupy a larger section of the image and the total area will increase. In our system the linear velocity of the robot has been dependent on measured area of the face. The area that has been set up during the building of the color model has been treated as the reference value that robot should keep. The controller of the linear velocity of the robot maintains the area of the face on desirable level while the person is coming closer to the robot or moving further away.

VII. CONCLUSIONS

We have presented a system which is capable of following a person using an autonomous mobile robot equipped with an active color CCD camera. It has been used to successfully navigate in laboratory environment. The system is quite stable and robust, particularly the position of the tracked face in the image plane is maintained very good. Our system has achieved a frame rate 10 frames per second using a laptop equipped with typical low cost frame-grabber. The algorithm enables the robot to track and follow a person at speed up to 25 cm per second. There are essentially two parts to the system: the color segmentation module and the control strategy. To show the correct work of the system, we conducted several experiments in naturally occurring in laboratory circumstances. An experiment was classified as successful if the robot had followed the person for minimum 50 meters, but experiments during which the system processed over 10 000 images had also taken place. Experiments have shown that the robot successfully deals with varying lighting conditions. The system is being implemented as part of currently realized interactive interface for robot programming-by-demonstration. Initial tests have shown that the mentioned above speed is satisfactory for tasks consist of following with simultaneous instruction.

ACKNOWLEDGES

This work was supported by the Polish State Committee for Scientific Research (KBN) within the project 7T11C03420

REFERENCES

1. Brown R. G., Hwang P. Y. C.: Introduction to random signals and applied Kalman filtering, John Wiley & Sons, New York, 1997
2. Fukunaga K.: Introduction to Statistical Pattern Recognition, sec. ed., Acad. Press, 1990
3. Gauch J. M.: Noise removal and contrast enhancement; in Sangwine S. J., Horne R. E. N. (eds.), The colour image processing handbook; Chapman & Hall, London, 1998
4. Horn B. K. P.: Robot vision, The MIT Press, 1986
5. Jähne B.: Digitale Bildverarbeitung, 4. Auflage, Springer-Verlag, 1997
6. Kohler M. R. J.: System architecture and techniques for gesture recognition in unconstraint environments, Proc. of the 1997 Int. Conf. on Virtual Systems and MultiMedia (VSMM'97), Geneva, September 10-12th, 1997, pp. 137-146
7. Kortenkamp D., Bonasso R. P., Murphy R. (Ed.): Artificial intelligence and mobile robots-Case studies of successful robot systems, The MIT Press, Cambridge, Massachusetts and London, 1998
8. Mckenna S. J., Gong S., Raja Y.: Modelling facial colour and identity with Gaussian mixtures, Pattern Recognition, vol. 31, no. 12, 1998, pp.1883-1892,
9. Palus H.: Representations of colour images in different colour spaces; in Sangwine S. J., Horne R. E. N. (eds.), The colour image processing handbook; Chapman & Hall, London, 1998
10. Skarbek W., Koschan A.: Colour image segmentation - a survey, technical report 84-32, Technische Universität Berlin, 1994
11. Tseng D.-C., Chang C.-H.: Color segmentation using perceptual attributes, Proc. 11th Int. Conf. on Pattern Recognition, Den Hague, Netherlands, 30 Aug.-3 Sept., 1992, vol. III, pp. 228-231
12. Wren Ch. R., Azarbayejani A., Darrell T., Pentland A. P.: Pfinder: real-time tracking of the human body, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997, pp. 780-785
13. Wyszecki G., Stiles W. S.: Color science: Concepts and methods, quantitative data and formulae, sec. ed. John Wiley, New York, 1982
14. Swain M. J., Ballard D. H.: Color indexing, Int. J. of Computer Vision, vol. 7, no. 1, 1991, pp. 11-32
15. Yang J., Waibel A.: A real-time face tracker, Proc. of 3rd IEEE Workshop on Applications of Computer Vision (WACV'96), Sarasota, Florida, USA, 1996, pp. 142-147
16. Waldherr S., Romero S., Thrun S.: A gesture-based interface for human-robot interaction, Autonomous Robots, vol. 9, 2000, pp. 151-173
17. Kahn R. E., Swain M. J., Prokopowicz P. N., Firby R. J.: Gesture recognition using the Perseus Architecture, In Proc. of the IEEE Conf. on Computer Vision and Pattern Rec., San Francisco, CA, 1996, pp. 734-741
18. Sidenbladh H., Kragić D., Christensen H. I.: A person following behaviour for a mobile robot, In Proc. of the IEEE Int. Conf. on Robotics and Automation, Detroit, MI, 1999, pp. 670-675