# Visual System for Tracking and Interpreting Selected Human Actions

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

## ABSTRACT

This paper describes an autonomous vision system for realization of tasks consist of following a person with a mobile robot as well as interpreting some static and dynamic commands signaled by hand. Detection of the person is realized on the basis of color image segmentation combined with stereovision analysis. The elaborated algorithms of face detection and localization improves quality of tracking as well as makes possible to recognize some nonverbal commands using geometrical relations of face and hands and in particular to recognize the pointing arm-posture.

## Keywords

Color image processing, human-machine interface, robot vision.

## 1. INTRODUCTION

Recently, a great attention is paid to interfaces that aid human-machine interaction. A dialog and interaction are one of the fundamental features of any multimedia system. As a consequence, people's detection techniques have become a center of interest as a powerful tool in surveillance, practical application of service robots and effective content-based image retrieval, opening many new directions of research and technology. An ability to understand and interpret human behaviors is particularly important in context of practical applications of service robots. A service robot that does not respond to human voice and does not understand human behavior might be dangerous to environment in a situation that has not been expected by its programmer. The ability to acquire knowledge about a task to an execution without any help of an expert is very useful in a practical applying of service robots. In order to be accepted by its potential users, the service robots have to be adaptable to the needs of each individual person [Tri98a].



***Journal of WSCG, Vol.11, No.1., ISSN 1213-6972***
*WSCG'2003, February 3-7, 2003, Plzen, Czech Republic.*


### Related Work

Color as a cue for detection an object was presented by Swain and Ballard [Swa91a]. There it was shown that the distribution of color information can be used to solve the object extraction problem. In work [Wal00a] 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 system [Kah96a] 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 [Wre97a] 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 [Sid99a] 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 [Mck98a] 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.

### Our Approach

In this work we refer to modern interfaces which aid human-computer interaction. The main result of this work is a visual system interpreting selected human actions. This takes place by analyzing a color image and utilizing a range image of a scene obtained by a stereo vision system. At first, the algorithm extracts skin regions and afterwards a color region growing is applied to extract the shirt of the tracked person. At the initialization stage of the color region growing a number of pixels below the considered skin blob is verified whether their distances to the camera are compared to skin one. If such verification is successful the pixels verified in such a way are considered as a seed region and the color region growing is started. After that some geometrical relations of the skin blob and the detected region are verified. At the next stage each person's candidate is verified in context of the presence of a face. The range information is utilized to set the size of the reflectional symmetry search area of each face candidate. It is used once again at the next step to set the proper size of the eyes-template used in a template matching. Finally, the well-known eigenfaces approach [Tur91a] is applied to detect the presence of a face in the window localized in such a way. If the face detection is unsuccessful the algorithms tries to track the person on the basis of skin and shirt segmentation. The shirt detection allows us to reduce the number of face candidates to verify and also aids the hand detection. The verification of the face presence can be realized without the successful shirt detection. Thanks to the reliable face detection some static commands can be recognized very fast and easily using only geometrical relations between the face as well as the left and the right hand.

## 2. FILTERING DESIRABLE OBJECTS IN COLOR IMAGES

Color image segmentation techniques are used in many practical applications due to the ability to fast detection of desirable objects. Their efficiency is especially worth to emphasize if a considered object is occluded or is in a shadow what can be in general significant practical difficulty using edge-based methods. The detection of a desirable object is realized on the basis of a prepared off-line model of color distribution. A representative color model for wide spectrum of illumination can be obtained on the basis of chrominance distribution. The model of color distribution is then prepared in a two-dimensional chrominance space.

Skin-color segmentation is a key step in many existing real-time systems of the people tracking. Our system uses skin-color segmentation based on the work of [Yan96a] to get an initial set of face candidates. Our experimental results show that the stated before the normalized color space gives repeated results on images which were acquired from an inexpensive on-board CCD camera. The authors of work [Ber00a] came to the conclusion that the normalized color space is the best chrominance representation considering the face detection tasks. Valuable remarks about a performance of some color spaces in color image segmentation can be found in work [Ska94a]. One of the advantage of the normalized color space is the speed of the $\mathrm{RGB} \rightarrow \mathrm{rg}$ conversion. But the speed of the conversion can play an insignificant part in computers equipped with enough amount memory because the conversion of the discussed color space as well as computationally expensive $\mathrm{RGB} \rightarrow \mathrm{HS}$ conversion could be tabelarized in 2x16MB memory block.

The detection of a desirable object is realized using Gaussian approximation of histogram representation of the chrominance distribution. Parameters of such a model: expectation values $\mu_r$, $\mu_g$ and covariances $v_{rr}$, $v_{gg}$, $v_{rg}$ are obtained on the basis of prepared in advance model database (skin, shirt) which has been constructed in typical illumination conditions which appear in our laboratory.

In the on-line phase the chrominance components of each pixel are used in the two dimensional Gaussian equation. The probability picture of the desirable object is the result of such an operation. The conversion to the normalized color space is preceded by filtering with 3x3 Gaussian mask. Convolution is associative and thanks to this property the filtering of each color component is realized using convolution with 3x1 mask and then convolution with 1x3 mask [Gau98a]. The multiplication by weights of binomial masks and division by the sum of the convolution mask weights are realized by the use of bitwise-shift operations. In achromatic areas it is not possible to obtain a reliable chrominance components and therefore they are not included in the described color image segmentation [Tse92a].

## 3. PROBABILITY IMAGE PROCESSING

A connected component labeling leads to extraction of separate blobs which in our approach represent candidates of the desirable object. Thresholding and morphological closing precede the pointed out connected component operation [Jah97a]. The method of the color segmentation which was discussed in the previous section is applied in our approach to coarse detection of desirable objects. Having above on regard the chose of the threshold did have not a considerable influence on the obtained results and therefore a small threshold has been applied in our system. The labeled picture is then used to compute areas as well as centers of coordinates of the extracted candidates.

The method based on segmentation of two relatively homogenous objects that are in proximity on each other (e.g. face and shirt) and analysis of two appropriate probability pictures is quite useful and has been proven as successful in experiments consist of following a person with mobile robot at distances 100 m even with the presence of students in our laboratory. Very simple geometrical relations have been considered to extract the tracked person: face above shirt, face area is smaller than shirt area, the coordinate centers are approximately vertical, face and shirt are next to each other, face is located in an area determined by max/min horizontal shirt coordinates.

The described above method needs the second color model, namely the shirt color model. With aim to reduce onerousness related to a preparation of the second model we have developed a color region growing procedure to extract the shirt of the tracked person without the necessity of preparation the color model. We believe that onerousness related to a preparation of the skin model will be eliminated in a near future and the skin extraction will be done automatic by dedicated cameras. At first, a coarse segmentation stage is begun from a seed region which has been localized below a face candidate. If the distance to the camera of such a seed region is different from the face distance the verified seed is not taken into consideration. The following homogeneity criteria are used in a two-stage region growing

$$\sqrt{(I-\bar{I})^2 + S^2 + \bar{S}^2 - 2S\bar{S}\cos(H-\bar{H})} \le d_1$$

$$\left|D-\bar{D}\right| \le d_2 \qquad (1)$$

where $H$, $S$, $I$ (hue, saturation, intensity) are current values of color components of tested pixel, $D$ is depth, $\bar{H}$, $\bar{S}$, $\bar{I}$, $\bar{D}$ are mean values related to the seed region, $d_1$ and $d_2$ are threshold values. The HSI color space is used at this stage and the conversion is realized on the basis of prepared in advance look-up tables. The HSI criterion considers all three color components and takes into account the cylindrical nature of HS components. For reasons of angular value of hue the values belonging to the seed region are converted into Cartesian coordinates in the following manner:

$$x_i = \cos(H), \; y_i = \sin(H) \qquad (2)$$

Once the first stage is completed the area of extracted region is calculated and then the relation of the extracted area to the distance to the camera is verified. Next, a second stage of the color region growing with the obtained in this manner seed region is conducted if the extracted area with respect to distance to the camera was to small. The average values $\bar{x} = \frac{1}{N}\sum_{i=0}^{N-1} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=0}^{N-1} y_i$ of the seed region were used in the arctan function for computing the mean $\bar{H}$ value. The two remaining values $\bar{S}$, $\bar{I}$ of the detected region were extracted and then in the second stage the HSI criterion with obtained in this manner new seed region and reduced to half $d_1$ value were used. Finally, the pointed out above geometrical relations between the face and the shirt are considered and in consequence some face candidates are rejected.

## 4. REAL-TIME FACE DETECTION

Thus far, we have discussed how the color segmentation techniques combined with the distance that is obtained from a stereovision might be used in performing of person detection. For a service robot that is required to work in a dynamic environment with people, automatic face detection is essential to a successful interpretation of human actions. Face detection is difficult because certain common but significant features, such as glasses or a moustache, can either be present or absent on a considered face candidate. Because of three-dimensional facial structures a change in lighting conditions, a movement of a face or a camera can cast or remove significant shadows from a tracked face. Face detection is a time-consuming operation due to the lack of constraint on the size and the location of tracked face in a sequence of images obtained from an active camera.

The digital images are spatially redundant. The Karhunen-Loeve transformation which is also know under the name Hotelling transform or principal component analysis being based on statistical properties ensures the extraction of data with decreasing statistical significance (smaller and

smaller eigenvalues). In the methods that are based on eigenfaces the feature space is reduced using principal components [Tur91a]. The scaled query windows of the input image are projected into the classification subspace and the closer the distance to the projected training image is, the greater the probability that such window corresponds to the trained human face is. Therefore, in order to apply the eigenfaces approach we must specify a window of proper size which will be examined in terms of the face presence. Thanks to the being in disposal information about the distance of the examined candidate of face to the camera, the reflected symmetry is checked in windows with dimensions covering only the considered face. The searching in the pointed out above windows is initialized from groups of pixels with a suitable probability of skin class membership. Next, the eyes are detected on the basis of a template matching techniques. With the approximate location of the face candidate the simple binary matching technique suitable for real-time computation is used to effectively detect and locate eyes candidates in the frontal view of a human face. One of the prepared off-line templates is chosen according to the distance between the examined face candidate and the camera. It has been proved that human eyes distinguish themselves from the whole rest of a face even if a user wears glasses similarly to the author. We have found that the proper selection of the template size and determination of the windows localization play an important role in frontal view face detection.

The eigenfaces algorithm operates on gray images and a collection of training faces [Tur91]. We have prepared a collection of training faces that consists of 64 images. While preprocessing, the average image for this collection is computed. Next, this image is subtracted from each training image and placed in the matrix. This matrix is used to create the covariance matrix. The eigenvalues and eigenvectors of such a covariance matrix are then determined. The first 16 normalized eigenvectors, sorted by decreasing eigenvalue represent subspace in which classification at run-time phase is performed. The eigenfaces are normalized eigenvectors which are the principal components of face space and they reflect the statistical properties of facial appearance. The first ten eigenfaces related to our training collection are shown in fig. 1.

**Figure 1. The first ten eigenfaces computed from our training collection**

Prior to the run-time phase the set of reference images is read and projected into the classification subspace. In run-time the L1 distance between these images and the projected (and centered) image onto the subspace is computed to determine the closest match. The size of the sub-image to be cropped is scaled according to the distance between the camera and the person. The location of the sub-image is determined on the basis of eyes location. The algorithms based on eigenfaces recognize only a face appearance that has been taken from a narrow angle, most often in the frontal view. Therefore when the face detection at the stage of following a person is ended oneself unsuccessful (due to side-view person orientation, shadows etc.) the robot continues tracking on the basis of pointed out above color and stereo cues combination. Thanks to the obtained in advance localization of face candidates, the presence of a vertical and frontal-view face in the scene is verified in a very fast manner.

## 5. PERSON LAYER EXTRACTION

In our system the distance of the tracked user to the camera is obtained on the basis of a stereovision system. Stereovision provides us information about the real distances to the camera. However, the distance between the robot and its operator is limited to certain range due to stereovision geometry. If there are not enough textures, some parts of the scene are not included in the depth image. In our system the stereo depth information is extracted thanks to small area correspondences between image pairs [Kon97a] and therefore it gives poor results in regions of little texture. But the depth map covering a face region is usually dense because a human face is rich in details.

At first, a disparity histogram is processed to characterize the scene depth and mainly to extract a distance layer containing the tracked person. The median filtering applied to the histogram reduces the noise in the histogram and preserves the information on the peak position in clusters. A peak with the lowest disparity value represents the background and is rejected from the histogram. If a disparity value is smaller than the disparity minimum value representing a distance to the camera equal to 3.5 m it is also considered as noise and ignored.

A threshold-based approach to extraction of distance layers on the basis of histogram analysis does not include proximity information contained in disparity image and in our application the peakness detection is used only to determine seed values in the region-growing labeling. The labeled pictures of face and shirt candidates allow us to connect suitable peaks and to extract easily the distance layer of the tracked person and use it during a robot commands recognition stage as well as following a person.

## 6. STATIC ARM-POSTURES AND DYNAMIC COMMANDS RECOGNITION

The static arm-posture based robot commands which we have developed are very simple and can be described by the spatial, stable for a certain period of time relation between the face and the hands of an instructor who stands in front of the robot, see fig. 2. If the instructor stops its moving for a certain period of time during the following a person, the robot turns into the robot commands recognition mode. The command stop following can be recognized easily because the tracked face and the hands are approximately at the same height to the floor. Due to proximity of the face and the hands we have achieved for them similar lighting conditions and therefore good recognition ratio. This command is tried to recognize even if the face detection is unsuccessful and only using color and stereo cues combination. The system checks whether this command was given also during following a person. The person's normal position which is expressed as the accept command must be performed after all remaining commands. The emergency stop command which consists of placing a hand near the camera or shutting out the camera lens has also been provided for. The mapping between the robot commands to be recognized and the associated actions of the robot is predefined. For example the turn left command is interpreted by the robot as turn about fifteen degrees.

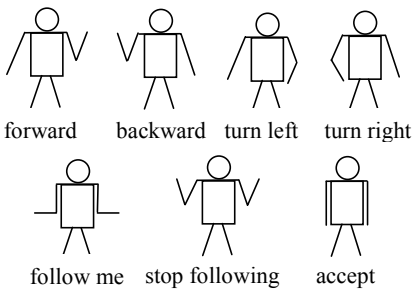


**Figure 2. Geometrical model of static arm-postures**

The stop following command is also provided to signal the start as well as the end of the pointing command. The distances of the face and the hand to the camera that are obtained from the depth image as well as two-dimensional positions of them allows us to obtain direction of pointing. We have assumed that between the start command and the pointing command the user face occupies approximately the same position. The pointing command has been tested in tasks consisting in selecting a desirable object on the laboratory floor.

The dynamic commands we use for navigating the robot are the following: swinging the hand forward and backing to the user (approach an instructor), waving of the hand diagonally to right/left (avoid of an instructor's at right/left side), rotating the hand clockwise/counterclockwise (turn left/right). Since false-positives (the robot recognizes the command which was not given by the user) are in human-robot communication less desirable than false-negatives (the robot fails to recognize the hand-action), our approach rests on the minimum twice execution of each mentioned above dynamic commands, excluding the last two consisting of hand rotating. A dynamic input command represented by estimated hand positions is encoded into a 1-dimensional code sequence used finally by the dynamic commands recognition module. The recognition module uses hidden Markov models [Rab86a], a popular means for representing and recognition of trajectories outlined by hand.

To use discrete HMMs for modeling of dynamic commands we must deliver a sequence of discrete symbols. The segmentation procedure of the input command is very simple and relies on assumption that hand of the operator is still for a short time between dynamic commands. When the hand stops moving for a while, the construction of the vector is ended and recognition procedure is activated.

The vector of features is created on the basis of set of the positions determined in frame sequence. *i*-th element of this vector indicates the direction of transition from the hand position registered in frame *i* to *i+1*. Eight directions coded by successive integers from the range <1, 8> are used for preparing code sequence, see fig. 3. The single direction of movement is determined on the basis of the last three positions of hand. In order to determine a code sequence representing only a phase of execution of a command a small logic is used to analyze the geometrical relationships between the last three positions. The temporal inertness of the hand is coded by simple symbol 9 regardless of the number of measurements of positions of the hand being in such a state.

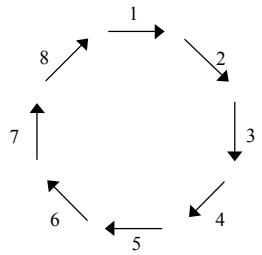


**Figure. 3. The rule of preparing code sequence**

The five HMM were designed and trained for each dynamic command. To find the probability of observation sequence, each HMM is evaluated by the Viterbi algorithm to compute the best path per HMM.

Before a HMM is able to classify a sequence it must be trained by a user. For each of the used five commands, a HMM has been trained separately using the Baum-Welch algorithm. The training vector has been prepared on the basis of minimum 50 commands which were registered by vision system.

## 7. CONTINUOUS TRACKING IN SEQUENCE OF IMAGES

The aim of tracking is generally speaking the exclusion to loss the desirable object and thus the acceptation of an accidental object for the object of interest. After the end of temporal difficulties with the object localization the tracking algorithm should concentrate again on the desirable object. The essential aim of tracking is therefore the exclusion the "jump" from desirable objects to another during processing a sequence of images. Therefore, in order to avoid the mentioned above undesirable effects as well as to decrease the noisy effect in the visual measurements of centroids, we use a recursive estimator, the Kalman filter, which estimates the best value, in a least-square sense, of a state vector from a set of noisy measures [Bro97a]. It can be shown that Kalman filter is the best possible estimator if the noises are Gaussian, and the best linear estimator if the noises are arbitrary distributed. A robust tracking is particularly important in context of adaptation of color model to lighting conditions. After an accidental jump from the tracked object the adaptation module could loose information about desirable distribution of chrominance very fast and in consequence fit the chrominance model of the tracked object to model of accidental object. The adaptation topic is discussed in the next section.

The following approximate model of a moving object is used

$$\xi_k = A\xi_{k-1} + w_k, \quad \eta_k = C\xi_k + v_k \tag{3}$$

where $k$ denotes the sample time ($t_{k+1} = t_k + T$; $T$ is the sample period), $\xi_k = \left[X_k, \dot{X}_k, Y_k, \dot{Y}_k\right]^T$ is the system state $\eta_k = \left[X_k, Y_k\right]^T$ is the measurement, $X_k$ and $Y_k$ indicate the center of the desirable object, $\dot{X}_k$, $\dot{Y}_k$ 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

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4}$$

Our model for system dynamics is a constant velocity model and acceleration is modeled as noise. The recursive equation for the prediction of the face center is given as

$$\hat{\xi}_{k/k-1} = A\hat{\xi}_{k-1/k-1} \tag{5}$$

Thanks to prediction we utilize the zone of expected location in the detection of face. The estimates $\hat{\xi}$ are defined by the Kalman filter algorithm

$$\hat{\xi}_{k/k} = \hat{\xi}_{k/k-1} + K_F\left(\eta_k - C\hat{\xi}_{k/k-1}\right) \tag{6}$$

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 [Koh97a].

## 8. MODEL ADAPTATION

The adaptation of the model of the chrominance distribution over time is essential to cope with varying lighting conditions especially during experiments with a mobile camera. When a camera moves the seeming colors change due to varying position of tracked object in relation to the camera as well as to source light. If illumination conditions cause the apparent skin-color to change then the color model will only partially reflect the actual skin-colors. Therefore we use in our approach a parametric approximation of skin-tone distribution and simple adaptive filter

$$\mu^t = \alpha\mu + (1-\alpha)\mu^{t-1}, \quad \Sigma^t = \alpha\Sigma + (1-\alpha)\Sigma^{t-1} \tag{7}$$

which computes the new parameters of Gaussian at time step $t$ from the old values $\mu^{t-1}$, $\Sigma^{t-1}$ and taken with a weighting factor $\alpha$ the new values $\mu$, $\Sigma$ which were obtained via mentioned in the previous sections techniques for detection of the tracked face.

The improvement of adaptation efficiency and the most important of all the elimination of circumstances leading to the total loss of information about real distribution of tracked chrominance is achieved thanks to (i) determination of skin locus in different and typical to experiments lighting conditions, (ii) determination of possible set of model parameters. Our approach differs from work [Wal00a] since it

uses combination of several visual cues and the tracking technique to detect the proper skin blob, predefined set of chrominance values as well as the predefined set of Gaussian parameters in model adaptation instead of simple windowing techniques. The skin model adaptation is realized only on the basis of face pixels.

## 9. PRACTICAL EXPERIMENTS

The position estimate of the tracked face as well as person distance to the camera is written asynchronously in block of common memory which can be easily accessed by Saphira client. Saphira is an integrated sensing and control system architecture designed to operate with a robot server [Kor98a]. The server sends a message packet to the client every 100 milliseconds, containing information on the robot velocity, the active camera's angles, sensor readings, and etc. Saphira supports high-level functions for robot control and sensor interpretation using an activities control language Colbert which is based on finite state machines semantics. The language is interpreted and the main construction is the activity schema, a procedure that describes a set of coordinated actions that the robot should perform. Activities are invoked synchronously every 100 ms cycle and they are interpreted by the Colbert executive, which supports also concurrent processing.

A mobile robot Pioneer 2 DX [Act01a] that was designed to move across a relatively flat surface was used in experiments and tests of prepared software. The robot was equipped with SRI's MEGA-D Megapixel Stereo Head and Sony's EVI-D31 PTZ (pant/tilt/zoom) camera. A typical laptop computer equipped with 850 MHz Pentium III, an analog frame-grabber as well as IEEE 1394 interface is utilized to run the prepared software. To combine color and stereo cues coming from different camera systems we utilize the well-known Tsai's calibration algorithm [Len88a]. The fixed stereo head which was equipped with wide angle lenses looks ahead of the mobile robot. The active camera is applied to provide a smooth movement of the mobile robot. If the person is located, the vision system keeps his/her face within the camera field of view by coordinating the pan/tilt movement of the camera with the tracked centroid location. In the assumed approach the control module of the active camera keeps the tracked face in predefined position and the robot follows the camera simultaneously trying to adjust its azimuth to camera's azimuth. The camera's pan angle is used as input of the robot orientation controller. The aim of the robot orientation controller is to minimize the angle between the robot and the camera. The linear velocity has been dependent on person's distance to the camera. A distance 1.7 m has been assumed as the reference value that the linear velocity controller should keep. To eliminate an unnecessary forward and backward robot movements we have applied a simple logic. The camera's pan and tilt controllers should keep the person's face in a half of image width and 4/5 of image height. We have implemented all above PD controllers in the Colbert language.

The image processing and recognizing algorithm runs at 320x240 image resolution at frame rates of 4-6 Hz depending of image complexity. It was implemented using C/C++ language. We have realized experiments in which the robot has followed a person at distances which beyond 100 m without the person loss. At the begin of experiments the robot motors are switched off and the active camera tries to locate the person to be tracked. After the person's layer extraction, the detection of the face presence, the extraction of the shirt with suitable area the motors are switched on and system is turned into human action interpreting mode. When in sequence of several frames the person is not found the motors are off and the camera pans and tilts in searching for a tracked person. In such a situation the robots tries to localize skin blobs or dominant areas of movement. To localize areas with dominant movements the camera is still for a period of time needed to acquire three consecutive frames and then the optical-flow field approach [All93a] is applied. Experiments showed that to re-localize the tracked person it is sufficient to demonstrate near the camera the hand of the tracked person or simply wave ones.

## 10. CONCLUSIONS

The presented system is non-intrusive and it enables a single user standing upright in front of the camera to interact with a mobile robot through movement as well as static and dynamic commands. The control strategy which is based on an active camera and a static stereovision system allows us to achieve smooth behaviors of the robot in response to a rapid movement of the tracked person. Stereovision has proved to be a very useful and robust cue in person and his/her face detection. In particular, it allows us to determine the direction of pointing. The detection of the presence of the vertical and frontal-view faces in the scene is fast and reliable because of the depth map covering the face region is usually dense and this together with color cues allows us to utilize symmetry information as well as the eyes-template before the usage of computationally expensive the eigenfaces method. The next step is to elaborate more sophisticated method to follow a person not necessary facing the robot during the following stage. During such a following the robot should recognize dynamic commands responsible for robot acceleration and deceleration. The elaborated method of face detection

in real-time is fast and very useful in a recognition of robotic commands. Particularly, thanks to the face position it is possible to recognize some static commands on the basis of geometrical relations of the face and hands. The method of color adaptation has been proven to have significant influence on obtained results, particularly during the following a person with the mobile robot. The presented system runs at 320x240 image resolution at frame rates of 4-6 Hz on 850 MHz Pentium III laptop which has been installed on Pioneer 2 DX mobile robot. The vision system enables the robot to follow a person at speed up to 30 cm per second. To show the correct work of the system, we have conducted several experiments in naturally occurring in laboratory circumstances and recorded movies for demonstration purposes.

## 11. ACKNOWLEDGMENTS

This work has been supported by the Polish State Committee for Scientific Research (project 7T11C03420)

## 12. REFERENCES

[Act01a] ActivMedia Robotics Pioneer 2 mobile robots, 2001.

[All93a] Allen, P.K, Timcenko, A., Yoshimi, B., and Michelman P. Automated tracking and grasping of a moving object with a robotic hand-eye system. IEEE Trans. on Robotics and Automation 9, No.2, pp. 152-165, 1993.

[Ber00a] Bergasa, L.M., Mazo, M., Gardel, A., Sotelo, M.A., and Boquete, L. Unsupervised and adaptive Gaussian skin-color model, Image and Vision Computing 18, pp. 987-1003, 2000.

[Bro97a] Brown, R.G., and Hwang, P.Y.C. Introduction to random signals and applied Kalman filtering, John Wiley & Sons, 1997.

[Gau98a] 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.

[Jah97a] Jähne, B. Digitale Bildverarbeitung, 4. Auflage, Springer-Verlag, 1997.

[Kah96a] Kahn, R.E., Swain, M.J., Prokopowicz, P.N., and Firby, R.J. Gesture recognition using the Perseus Architecture, In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, pp. 734-741, 1996.

[Koh97a] 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, Sept. 10-12th, pp. 137-146, 1997.

[Kon97a] Konolige, K. Small Vision System: Hardware and implementation, In Proc. of Int. Symposium on Robotics Research, Hayama, Japan, pp. 111-116, 1997.

[Kor98a] Kortenkamp, D., Bonasso, R.P., and Murphy, R. (eds.). Artificial intelligence and mobile robots - Case studies of successful robot systems, The MIT Press, Cambridge, Massachusetts and London, 1998.

[Len88a] Lenz, R.K., Tsai, R.Y. Techniques for calibration of the scale factor and image center for high accuracy 3-D machine vision metrology, IEEE Trans. on Pattern Analysis and Machine Intelligence 10, No.5, pp. 713-720, 1988.

[Mck98a] McKenna, S.J., Gong, S., and Raja, Y. Modelling facial colour and identity with Gaussian mixtures, Pattern Recognition 31, No.12, pp. 1883-1892, 1998.

[Rab86a] Rabiner, L.R., and Juang, B.H. An introduction to hidden Markov models, IEEE ASSP Magazine, pp. 4-16, 1986.

[Sid99a] Sidenbladh, H., Kragić, D., and Christensen, H.I. A person following behaviour for a mobile robot, In Proc. of the IEEE Int. Conf. on Robotics and Automation, Detroit, MI, pp. 670-675, 1999.

[Ska94a] Skarbek, W., and Koschan, A. Colour image segmentation - a survey, technical report 84-32, Technische Universität Berlin, 1994.

[Swa91a] Swain, M.J., and Ballard, D.H. Color indexing, Int. J. of Computer Vision 7, No.1, pp. 11-32, 1991.

[Tri98a] Triesch, J., and von der Malsburg, Ch. A gesture interface for human-robot-interaction, Proc. of the IEEE Conf. on Automatic Face and Gesture Recognition, Nara, Japan, pp. 546-551, 1998.

[Tse92a] Tseng, D.C., and Chang, C.-H. Color segmentation using perceptual attributes, Proc. 11th Int. Conf. on Pattern Recognition, Den Hague, Netherlands, 30 Aug.-3 Sept., Vol. III, pp. 228-231, 1992.

[Tur91a] Turk, M. A., and Pentland, A.P. Face recognition using eigenfaces, In Proc. of Conf. on Computer Vision and Pattern Recognition, pp. 586-591, 1991.

[Yan96a] Yang, J., and Waibel, A. A real-time face tracker, Proc. of 3rd IEEE Workshop on Applications of Computer Vision (WACV'96), Sarasota, Florida, USA, pp. 142-147, 1996.

[Wal00a] Waldherr, S., Romero, S., and Thrun, S. A gesture-based interface for human-robot interaction, Autonomous Robots 9, pp. 151-173, 2000.

[Wre97a] Wren, Ch.R., Azarbayejani, A., Darrell, T., and Pentland, A.P. Pfinder: real-time tracking of the human body, IEEE Trans. on Pattern Analysis and Machine Intelligence 19, No.7, pp. 780-785, 1997.