PERSON FOLLOWING AND MOBILE ROBOT VIA ARM-POSTURE DRIVING USING COLOR AND STEREOVISION

Bogdan Kwolek

Rzeszów University of Technology,
W. Pola 2, 35-959 Rzeszów, Poland

Abstract: This paper describes an approach to arm-posture recognizing and interpreting by mobile robot while following a person. We combine color and stereo cues to detect and track a person in dynamic indoor environments in real-time. The arm-postures are very simple and can be described by a spatial relation between a person's face and hands. Color image processing and labeling techniques are used to detect candidates of face, hands and shirt. The localization and recognition take place by analyzing the labeled images and utilizing a range histogram. Using heuristics for the size and spatial relations we can locate the tracked person as well as the person's face and hands. Experiments conducted with a Pioneer 2 DX mobile robot equipped with an active color camera and a stereovision system have shown the presented approach to be repeatable and robust to noise.

Keywords: Vision systems, man-machine interface, navigation of mobile robots

1. INTRODUCTION

Visual tracking of people has been studied extensively over past few years (McKenna *et al.*, 1998). However, the majority of existing approaches assumes that camera is mounted at a fixed location. Most of the tracking techniques rely on simple cues for detecting and segmenting the body parts from the background such as motion detection or skin-tone color. Common industrial robots usually perform repeating actions in an exactly predefined environment. In contrast to them service robots are designed for supporting jobs for people in their life environment and therefore the mobile cameras are more useful. These intelligent machines should operate in dynamic and unstructured environment and provide services while interaction with people who are not specially skilled in a robot communication. A human-robot interaction is imposed by several constraints like dynamic and complex background, shadows, reflections, occlusions and varying lighting conditions, a deformable shape of silhouette, hands, etc. (Triesch and von der Malsburg, 1998). Such a system should be a user-independent and to ensure safety and a collision free movement for surrounding and the user. 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 (Waldherr *et al.*, 2000).

A kind of human-machine interaction which is very interesting and has some practical use is following a person by a mobile robot. This behavior can be useful in several applications including robot programming by demonstration and instruction which in particular can contain tasks consisting in a guidance a robot to specific place where the user can point to object of interest. A demonstration is particularly useful at programming of new tasks by non-expert users. It is far easier to point at an object and demonstrate a track which robot should follow than to verbally describe its exact location and possible to accomplish the path of movement (Waldherr *et al.*, 2000).

In this paper we present a simple, fast and robust vision based low-level interface that has been used to conduct several experiments consisting in recognizing arm-postures while following a person with autonomous robot in natural laboratory environment. In previous work we have presented a system for following a person which was based on an active color camera (Kwolek, 2002). The central idea of estimation a distance between the robot and the user relies on fact that the size of the object image increases proportionally to the inverse of the distance to the object. The experiments showed that the estimated distance is very reasonable, even if the robot is not moving in the direction of the person. A film presented at MMAR'02 (Kwolek, 2002) showed that the system enables the robot to track and follow a person in a number of times round about laboratory. In the discussed approach we have utilized a stereo system to measure a distance between the user and the robot. Additionally we have extended the system about an arm-posture recognition module.

1.1 RELATED WORK

Color as a cue for detection an object was presented by Swain and Ballard (Swain and Ballard, 1991). There it was shown that the distribution of color can be used to solve the object extraction problem. In work (Waldherr *et al.*, 2000) a fast and adaptive algorithm for tracking and following a person by a robot-mounted camera has been described. After locating the person in front of the robot 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 (Kahn *et al.*, 1996) 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 (Wren *et al.*, 1997) uses adaptive background subtraction and pixel classification to track people in 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. A combination of color based with a robust contour based object detection is presented in (Schlegel *et al.*, 1998). The color is used for determining regions with likely suitable contour candidates. The similarity between the model and an edge candidate is estimated on the basis of the generalized Hausdorff-distance. While following the person the color model and contour model are continuously updated. The contour based approach requires significant contrast between the tracked individual and the background. In other system (Sidenbladh *et al.*, 1999) behavior for a person following with an active camera mounted on a 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 (McKenna *et al.*, 1998) 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.

1.2 OUR APPROACH

Our algorithm works by estimating two distributions: skin colors and colors on shirt. 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 skin or shirt, respectively. The probabilities are then used to segment the shirt as well as skin region candidates from the background. We include then at a thresholding stage only pixels with a great likelihood of belonging to the skin or the shirt classes, respectively. At the next stage we apply a connected component analysis (Jähne, 1997). The connected component analysis is used to gather to groups the adjacent pixels whose probability is over a threshold. The areas, coordinates of centers of gravity, distance to the camera and geometrical relations of labeled skin-like regions with shirt-like regions are then used in detection of the person within an image sequence. This approach guarantees that only user of the robot is tracked at a time. If the operator stops his moving for a while, the robot turns into the command recognition mode. First of all the algorithm locates the user's face and hands and afterwards the arm-postures are recognized. This takes place by analyzing the labeled images and utilizing the range image of the scene obtained by the stereo vision system.

The output of the tracking module is the position of face in an image. The control system of the camera uses the position of the 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 is used as input for rotation controller of the robot. This controller should minimize the angle between the axis of the camera and the robot. The mentioned above control strategy allows us to achieve smooth behaviors of the robot in response to a rapid movement of the tracked person. The controller of the linear velocity of the robot maintains the distance to the camera of detected face on desirable level while the person is coming closer to the robot or moving further away.

2. COLOR IMAGE SEGMENTATION

Skin-tone filtering is an important key in many real-time tracking systems of people. The advantage of such an approach relies on the speed at which present low-cost personal computers can extract the object of interest from color images. That aspect is particularly important considering on the one hand the limited computational power of an on-board computer and on the other hand the necessity of work at rate which enables to achieve smooth movement of a mobile robot. Using color, a decision can be made on the basis of a separate pixel. In contrast, motion estimates are usually computed on a pixel block basis. We use skin-tone filtering to get an initial set of skin as well as shirt candidates. Yang and Weibel (Yang and Waibel, 1996) have shown that human skin colors cluster in a small region in the normalized color space. Conversion to the two-dimensional intensity-normalized color space rg, where $r = R/(R+G+B)$, $g = G/(R+G+B)$ reduces brightness dependence.

Because the skin locus occupies a contiguous region in the normalized color space it can be approximated by 2-D Gaussian model $G = (\mu_r, \mu_g, \Sigma)$ obtained from cut-out skin region of the face. Using of only two dimensional color space reduces the number of parameters to be estimated, which is important because only a limited number of training pixels is usually in disposal. For the generation of a skin color model we manually segment a set of images containing skin regions. Each image with user outlined regions of interest is undergone the color space transformation. The aim of such an operation is to obtain a representative sample of typical skin colors and thus it is not essential to outline the whole face. Next we compute the histograms of such regions. A color histogram is constructed by counting the pixels of a given value in an image and can be used as non-parametric skin model. A parametric model for skin color (e.g. Gaussian fitting) provides a generalization about small amounts of training data and tends to smoothen a training set distribution.

When a robot moves and pans during the following a person, the apparent appearance of tracked individual 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. If illumination conditions cause the apparent skin color to change then the model will only partially reflect the actual skin colors. The simple way to adapt model over time is to use a linear combination of the known parameters to predict or approximate new parameters. Therefore we use in our approach a parametric approximation of skin-tone distribution and simple adaptive filter

$$\begin{aligned} \mu^t &= \alpha\mu + (1-\alpha)\mu^{t-1} \\ \Sigma^t &= \alpha\Sigma + (1-\alpha)\Sigma^{t-1} \end{aligned} \tag{1}$$

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. The knowledge of possible skin locus that was obtained by the robot camera in different illumination conditions is used to select the appropriate pixels for updating the color model.

The frame to be segmented is transformed into rg color space and each pixel p_i with chromaticity (r_i, g_i) is assigned the value of Gaussian function at (r_i, g_i). The better the pixel matches color model, the higher the probability and response of such a color filter are.

The extraction process of skin-tone is analogous to the single hypothesis classifier described in (Fukunaga, 1990). Single hypothesis classifier deals with problems in which one class is well defined while others are not. Let $\zeta = [r_i, g_i]^T$ denote the feature vector consists of color components of a pixel and ω_s denote the skin class. Thus the probability that pixel belongs to class ω_s can be expressed as

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

where μ_s is the mean color vector and Σ_s is the covariance matrix of considered class.

The outlined above color segmentation provides accurate results of skin regions extraction if there is a good contrast between skin colors and those of the background. The usage of two colors for extraction of the person prevents of an identification of an accidental object for a tracked individual. This is important particularly in relation to adaptation system which in consequence of incorrect decision of the recognition module would cause the loss of information about color distribution of the tracked object.

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 a very simple model of colors taken into account during a tracking and therefore the person to be recognized should wear a single solid color shirt. If such a color is relatively homogenous, its distribution can be directly characterized by using a single Gaussian. The range information makes person tracking more reliable and allows us to detect not only face but also hands and thus to recognize simple arm-postures.

3. PERSON LOCALIZATION

Using prepared in advance color models we apply the outlined above color filter to each pixel in the image and threshold the result. As a result two probability pictures are obtained and a pixel is identified as a candidate of a face (shirt) if the corresponding probability is greater than a threshold. The threshold is set up to a low value. Because we are interested in coarse extraction of face and shirt candidates, the choice of threshold has not a considerable influence on obtained results. After an image was thresholded, a morphological closing operator is performed. The connected component analysis is utilized to gather neighborhood pixels into larger regions. In that operation each object receives a unique label which can be used to distinguish regions during further processing. The regions extracted in such a way are used to calculate areas and gravity centers of the detected candidates of the face and the shirt. Next, the information about the distance of the user to the camera allows us to apply simple heuristics to check if a face-shirt relationship which can be characterized by distances, areas, vertical alignment, etc., is physically possible.

The distance of the user to the camera is obtained on the basis of a stereovision system. Stereovision gives in particular information about the real distance between object and robot. Moreover, partially occluded objects can be detected, which is a major problem for other techniques. 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 (Konolige, 1997) and therefore it gives poor results in regions of little texture. But the depth map covering a face region is usually dense because the human face is rich in details.

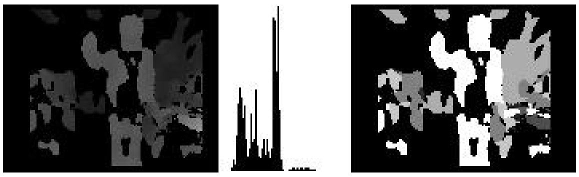

Fig. 1. Depth image, depth histogram, depth layers

A sample disparity map that was obtained by the stereovision system is shown in fig. 1. A disparity histogram is then processed to characterize the scene depth and mainly to extract a distance layer containing the person. 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 histogram. If the value of separate peak or a peek group is smaller than a predefined minimum value it is also considered as noise and ignored.

The detection of peeks and valleys is in general a nontrivial problem. For an automatic thresholding scheme we should have a measure of the peakness and valleyness of clusters in the histogram. But such a threshold-based approach to extraction of distance layers does not include proximity information contained in disparity image and in our application a peakness detection was used only to determine start values in region-growing labeling. Our aim is to extract the person and therefore the labeled pictures of skin and shirt candidates allow us to connect suitable peaks and easily extract the person distance layer.

The person distance layer which is obtained in this manner is then used in spatial test with aim to check the geometrical relations as well as distances between skin-tone and shirt candidates and to extract physically possible configuration. In order to estimate person's width, distance to the camera and especially distances between objects in the image we use real world lengths. Then after the examination of distances, areas, vertical alignment of face and shirt candidates as well as bordering on each other, the face was finally extracted. Finally, after the confirmation of attachment of the detected face (shirt) pixels to skin (shirt) locus the actual Gaussian parameters have been extracted and then used in appropriate model adaptation.

4. CONTINUOUS PERSON TRACKING

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 (Brown and Hwang, 1997). 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. The following approximate model of a moving object is used

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

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, $\eta = [X_k, Y_k]$ 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

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

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

$$\hat{\xi}_{k|k-1} = A\hat{\xi}_{k-1|k-1} \quad (5)$$

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

$$\hat{\xi}_{k|k} = \hat{\xi}_{k|k-1} + K_F(\eta_k - C\hat{\xi}_{k|k-1}) \quad (6)$$

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 (Kohler, 1997).

5. ARM-POSTURE ROCOGNITION

If the operator stops its moving for a while, the robot turns into the arm-posture recognition mode. Once we have found the position of the person in the scene we can try to recognize some of his arm-postures. As we know the position of the head, we try to localize the hands on the left and respectively on the right side of the head. For complexity reasons, we have predefined and have assumed only arm-poses, which are stable for a certain period of time. The arm-postures are very simple and can be described by the spatial relation between person's face and hands, see fig. 2. A small discrete set of arm-postures which has been defined for the basic control of a robot includes stop, move forward/backward, follow me, turn left/right commands. The mapping between the gestures to be recognized and the associated actions of the robot is predefined. The commands allow the user to activate or deactivate the specific behaviors of the robot. The arm-posture "accept" was used to enter one of the commands: forward, backward, turn or follow me. The command "stop following" can also be recognized during a following.

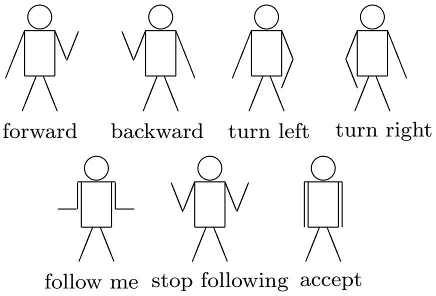

Fig. 2. Geometrical model of arm-postures

6. EXPERIMENTS

The estimate of face position in image is written asynchronously by the vision module in block of a common memory, which can be easily accessed from Saphira client. Saphira supports a packet-based communications protocol for sending commands to the robot server receiving information back from the robot controller (*Pioneer 2*, 2001). A Pioneer 2 DX mobile 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 Pentium III 850 MHz processor services the robot.

Once a person is located, the vision system keeps the person within the camera field of view by coordinating the pan/tilt movement of the camera with the location of the detected face. The active camera 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 two PD controllers have been used.

The actual pan angle of the camera has been used as input for the orientation controller of the robot. A PD 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 the tracked person and guarantees to keep the person in the field of view of the camera during the robot movement. A PD controller of the linear velocity of the robot maintains the distance to the camera of detected face on desirable level while the person is coming closer to the robot or moving further away.

The presented approach enables the robot to track and follow a person at speed up to 25 cm per second. To show the correct work of the system, we conducted several experiments in naturally occurring in laboratory circumstances. The system enables the robot to track and follow a person in a number of times round about laboratory. The recognition rate of the robot commands is over 95%. During a realization of given command the robot executes action and the active camera is keeping the user in desirable image position. The stereovision system uses wide-angle cameras and therefore the user remains in the field of view for a large scale robot orientation.

7. CONCLUSIONS

Following a person with a mobile robot is a much more challenging task than ones with a fixed camera because of both motion of the camera and the user. Experiments conducted with a Pioneer 2 DX mobile robot equipped with an active color camera and a stereovision system have shown the presented algorithms to be repeatable and robust to noise. The system is non-intrusive and it enables a single user standing upright in front of the camera to interact with mobile robot through movement and arm-postures. The tracking currently runs at a maximal speed of 8 Hz on a standard Pentium without any special image processing hardware.

The control strategy with 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 method to follow the user with a mobile robot. The combination of skin-tone color segmentation and stereovision seems to have a very large application area in robotics and the chances are that in a short period of time will be accessible cameras that will be oriented to automatic skin-tone detection. The next step is to elaborate more sophisticated method to combine stereo and color cues. Another interesting direction to be investigated is the application of double thresholding region-growing method in a skin-tone blobs extraction. In particular, the robustness of method using only skin-tone segmentation and stereovision should be verified.

8. ACKNOWLEDGEMENT

This work has been supported by the Polish State Committee for Scientific Research (project 4T 11C 01224)

REFERENCES

Brown, R. G. and P. Y. C. Hwang (1997). *Introduction to Random Signals and Applied Kalman Filtering.* John Wiley & Sons.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Acad. Press.

Jähne, B. (1997). *Digitale Bildverarbeitung.* Springer-Verlag.

Kahn, R. E., M. J. Swain, P. N. Prokopowicz and R. J. Firby (1996). Gesture recognition using the perseus architecture. In: *Proc. of the IEEE Conf. on CVPR.* pp. 734–741.

Kohler, M. (1997). System architecture and techniques for gesture recognition in unconstraint environments. In: *Proc. of Int. Conf. on Virtual Systems and Multimedia.* pp. 137–146.

Konolige, K. (1997). Small vision system: Hardware and implementation. In: *Proc. Int. Symp. on Robotics Research.* pp. 111–116.

Kwolek, B. (2002). Person following, obstacle detection and collision-free path planning on autonomous mobile robot with color monocular vision. In: *Proc. of the 8th IEEE Int. Conf. MMAR.* pp. 971–978.

McKenna, S. J., S. Gong and Y. Raja (1998). Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition* **31(12)**, 1883–1892.

Pioneer 2 (2001). ActivMedia Robotics.

Schlegel, Ch., I. Jërg, H. Jaberg, M. Schuster and R. Wörz (1998). Vision based person tracking with a mobile robot. In: *Ninth BMVC.* pp. 418–427.

Sidenbladh, H., D. Kragi and H. I. Christensen (1999). A person following behaviour for a mobile robot. In: *Proc. of the IEEE Int. Conf. on Robotics and Automation.* pp. 670–675.

Swain, M. J. and D. H. Ballard (1991). Color indexing. *Int. J. of Comp. Vision* **7**, 11–32.

Triesch, J. and Ch. von der Malsburg (1998). A gesture interface for human-robot-interaction. In: *Proc. of the IEEE Conf. on Automatic Face and Gesture Recogn.* pp. 546–551.

Waldherr, S., S. Romero and S. Thrun (2000). A gesture-based interface for human-robot interaction. *Autonomous Robots* **9**, 151–173.

Wren, Ch. R., A. Azarbayejani, T. Darrell and A. P. Pentland (1997). Pfinder: real-time tracking of the human body. *IEEE Trans. on PAMI* **19(7)**, 780–785.

Yang, J. and A. Waibel (1996). A real-time face tracker. In: *Proc. of 3rd IEEE Workshop on Applications of Comp. Vision.* pp. 142–147.