Stereovision-Based Head Tracking Using Color and Ellipse Fitting in a Particle Filter

Bogdan Kwolek

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

Abstract. This paper proposes the use of a particle filter combined with color, depth information, gradient and shape features as an efficient and effective way of dealing with tracking of a head on the basis of image stream coming from a mobile stereovision camera. The head is modeled in the 2D image domain by an ellipse. A weighting function is used to include spatial information in color histogram representing the interior of the ellipse. The lengths of the ellipse's minor axis are determined on the basis of depth information. The dissimilarity between the current model of the tracked object and target candidates is indicated by a metric based on Bhattacharyya coefficient. Variations of the color representation as a consequence of ellipse's size change are handled by taking advantage of the scale invariance of the similarity measure. The color histogram and parameters of the ellipse are dynamically updated over time to discriminate in the next iteration between the candidate and actual head representation. This makes possible to track not only a face profile which has been shot during initialization of the tracker but in addition different profiles of the face as well as the head can be tracked. Experimental results which were obtained on long image sequences in a typical office environment show the feasibility of our approach to perform tracking of a head undergoing complex changes of shape and appearance against a varying background. The resulting system runs in real-time on a standard laptop computer installed on a real mobile agent.

1 Introduction

Visual tracking of objects in video sequences is becoming an important task in a wide range of applications utilizing computer vision interfaces, including human action recognition, teleconferencing, robot teleoperation as well as human-computer interaction. Many different trackers for various tasks have been developed in recent years and particular interests and research activities have increased significantly in vision-based methods. One of the purposes of visual tracking is to estimate the states of objects of interest from an image sequence. However, cluttered backgrounds, unknown and changing lighting conditions and multiple moving objects make the vision-based tracking tasks challenging. Some vision-based systems allow a determination of a body position and real-time tracking of head and hands. Pfinder [1] uses a multi-class statistical model of color and shape to obtain a blob representation of the tracked silhouette in a

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 192–204, 2004.

wide spectrum of viewing conditions. In the techniques known as CamShift [2] and MeanShift [3] the current frame is searched for a region in a variable-size window, whose color content matches best a reference model. The searching process proceeds iteratively starting from the final location in the previous frame. The new object location is calculated based on the mean shift vector as an estimation of the gradient of the Bhattacharyya function. This method requires that the new target center lies within the kernel centered on the previous location of the target. The original application of the particle filter in computer vision was for object tracking in an image sequence [4]. Particle filtering is now a popular solution to problems relying on visual tracking. In the work of [5] a fixed ellipse is used to approximate the head outline during 2D tracking on the basis of the particle filter. A system developed recently by Chen *et al.* [6] uses a causal 1D contour model in dynamic programming to find the best contour with respect to a given initial one. A five dimensional ellipse is used to represent the head contour in multiple hypothesis framework. Nummiaro *et al.* [7] used an ellipse with fixed orientation to model a head and to extract the color distribution of the ellipse's interior. The likelihood is calculated on the basis of weighted histogram representing both color and shape of the head. Global color reference models and Bhattacharyya coefficient as a similarity measure between the color distribution of the model and target candidates have been used in a Monte Carlo tracker [8]. A histogram representation of the region of interest has been extracted in a rectangular window. Recently, the laser range fingers have been used to track people in populated environments for interactive robot applications [9].

In this paper, we focus our attention on tracking human head/face, one of the most important features in tasks consisting in people tracking and action recognition. The main objective of the research is to detect and track the head to perform person following with a real mobile agent which is equipped with an on-board camera. The initial position of the head to be tracked is determined by means of face detection. We consider scenarios where a stereo camera is mounted on a mobile agent and our aim is tracking the head which can undergo complex changes of shape and appearance. The appearance of the object of interest changes continuously due to non-rigid human motion and a change in viewpoints. There are many other difficulties in extracting features distinguishing the target and challenge lies in the fact that a background may not be static. We consider the problem of head tracking by taking advantage of gradient, color together with shape as well as depth information which are combined with the particle filter. One of the problems of tracking on the basis of color is that lighting conditions may have influence on perceived color of the target. Even in the case of constant lighting conditions, the seeming color of the target may change over a frame sequence, since the target can be shadowed by other objects. The color distributions representing the target in image sequences are therefore not stationary.

The main goal of the tracker is to find the most probable sample distribution. The particles representing the candidate ellipses are verified in respect of intensity gradient near the edge of the ellipse and matching score of the color histograms representing the interior of an ellipse surrounding the tracked object

and currently analyzed one. During samples weighting stage in which candidate ellipses are considered one after another, the projected ellipse size into image is dependent on the depth information. The color histogram and parameters of the ellipse are dynamically updated over time to discriminate in the next iteration between the candidate and actual head representation.

The contribution of our work lies in the use of particle filters combined with mentioned above cues to robustly solve a difficult and a useful problem of head tracking in color images. The tracker has been evaluated in experiments consisting in face tracking with a stereovision camera mounted on a real mobile agent. A version of the tracker which utilizes gradient, color as well as shape information combined with particle filters has been evaluated using the PETS-ICVS 2003 video data set which is provided to conduct experiments relating to smart meeting room.

The rest of the paper is organized as follows. In the next section we briefly describe particle filtering. The usage of color cue, gradient, shape information and stereovision in a particle filter is explained in section 3. In sections 4 and 5 we report results which were obtained in experiments. Finally, some conclusions follow in the last section.

2 Particle Filtering

In this section we formulate the visual tracking problem in a probabilistic framework. Among the tracking methods, the ones based on particle filters have attracted much attention recently and have proved as robust solutions to reduce the computational cost by searching only those regions of the image where the object is predicted to be. The key idea underlying all particle filters is to approximate the probability distribution by a weighted sample collection.

The state of the tracked object at time t is denoted $\mathbf{x}_t$ and its history is $X_t = \{\mathbf{x}_1, ..., \mathbf{x}_t\}$. Similarly the set of image features at time t is $\mathbf{z}_t$ with history $Z_t = \{\mathbf{z}_1, ..., \mathbf{z}_t\}$. The evolution of the state forms a temporal Markov chain so that the new state is conditioned directly on the immediately preceding state and independent of the earlier state, $p(\mathbf{x}_t \mid X_{t-1}) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. Observations $\mathbf{z}_t$ are assumed to be independent, both mutually and with respect to the dynamical process, $p(Z_{t-1}, \mathbf{x}_t \mid X_{t-1}) = p(\mathbf{x}_t \mid X_{t-1}) \prod_{i=1}^{t-1} p(\mathbf{z}_i \mid \mathbf{x}_i)$. The observation process is defined by the conditional density $p(\mathbf{z}_t \mid \mathbf{x}_t)$. Given a continuous-valued Markov chain with independent observations, the conditional state density $p(\mathbf{x}_t \mid Z_t)$ represents all information about the state at time t that is deducible from the entire data-stream up to that time.

We can use Bayes' rule to determine the *a posteriori* density $p(\mathbf{x}_t \mid Z_t) = p(\mathbf{x}_t \mid \mathbf{z}_t, Z_{t-1})$ from the *a priori* density $p(\mathbf{x}_t \mid Z_{t-1})$ in the following manner

$$p(\mathbf{x}_t \mid Z_t) = \frac{p(\mathbf{z}_t \mid \mathbf{x}_t, Z_{t-1}) p(\mathbf{x}_t \mid Z_{t-1})}{p(\mathbf{z}_t \mid Z_{t-1})} = k_t p(\mathbf{z}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid Z_{t-1}) \tag{1}$$

where k_t is a normalization factor that is independent of $\mathbf{x}$ and

$$p(\mathbf{x}_t \mid Z_{t-1}) = \int_{\mathbf{X}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid Z_{t-1}) dx_{t-1} \tag{2}$$

This equation is used to propagate the probability distribution via the transition density $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. The density function $p(\mathbf{x}_t \mid Z_{t-1})$ depends on the immediately preceding distribution $p(\mathbf{x}_{t-1} \mid Z_{t-1})$, but not on any function prior to $t-1$, so it describes a Markov process. Multiplication by the observation density $p(\mathbf{z}_t \mid \mathbf{x}_t)$ in the equation for *a priori* density $p(\mathbf{x}_t \mid Z_{t-1})$ applies the reactive effect expected from observations. The observation density $p(\mathbf{z}_t \mid \mathbf{x}_t)$ defines the likelihood that a state $\mathbf{x}_t$ causes the measurement $\mathbf{z}_t$. The complete tracking scheme, known as the recursive Bayesian filter first calculates the *a priori* density $p(\mathbf{x}_t \mid Z_{t-1})$ using the system model and then evaluates *a posteriori* density $p(\mathbf{x}_t \mid Z_t)$ given the new measurement, $p(\mathbf{x}_{t-1} \mid Z_{t-1}) \xrightarrow{dynamics} p(\mathbf{x}_t \mid Z_{t-1}) \xrightarrow{measurement} p(\mathbf{x}_t \mid Z_t)$.

The density $p(\mathbf{x}_t \mid Z_t)$ can be very complicated in form and can have multiple peaks. The need to track more than one of these peaks results from the fact that the largest peak for any given frame may not always correspond to the right peak. The random search which is known as particle filtering has proven useful in such considerable algorithmic difficulties and allows us to extract one or another expectation. One of the attractions of sampled representations of probability distributions is that some calculations can be easily realized.

Taking a sample representation of $p(\mathbf{x}_t \mid Z_t)$, we have at each step t a set $S_t = \left\{(\mathbf{s}_t^{(n)}, \pi_t^{(n)}) \mid n = 1...N\right\}$ of N possibly distinct samples, each with associated weight. The sample weight represents the likelihood of a particular sample being the true location of the target and is calculated by determining on the basis of depth information the ellipse's minor axis and then by computing the gradient along ellipse's boundary as well as matching score of histograms representing the interior of ellipses which bound (i) the tracked object and (ii) currently considered one. Such a sample set composes a discrete approximation of the probability distribution. The prediction step of Bayesian filtering is realized by drawing with replacement N samples from the set computed in the previous iteration, using the weights $\pi_{t-1}^{(n)}$ as the probability of drawing a sample, and by propagating their state forward in time according to the prediction model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. This corresponds to sampling from the transition density [10]. The new set would predominantly consist of samples that appeared in previous iteration with large weights. In the correction step, a measurement density $p(\mathbf{z}_t \mid \mathbf{x}_t)$ is used to weight the samples obtained in the prediction step, $\pi_t^{(n)} = p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}^{(n)})$. The complete scheme of the sampling procedure outlined above can be summarized in the following pseudo-code:

$S_t = \emptyset$
`for` $n = 0$ `to` N `do`
  `select` k `with probability` $\pi_{t-1}^{(n)} / \sum_{i=1}^{N} \pi_{t-1}^{(i)}$
  `propagate` $\mathbf{s}_t^{(n)} = A\mathbf{s}_{t-1}^{(k)} + w$
  `calculate non-normalized weight` $\pi_t^{(n)} = p(\mathbf{z}_t \mid \mathbf{s}_t^{(n)})$
  `add` $\mathbf{s}_t^{(n)}$ `to` S_t
`endfor`

The component A in the propagation model is deterministic and w is a multivariate Gaussian random variable. As the number of samples increases, the precision with which the samples approximate the pdf increases. The mean state can be estimated at each time step as $E[S_t] = \sum_{n=1}^{N} \mathbf{s}_t^{(n)} \pi_t^{(n)}$, where $\pi_t^{(n)}$ are normalized to sum to 1.

3 Representation of the Target Appearance

The shape of the head is one of the most easily recognizable human parts and can be reasonably well approximated by an ellipse. In work [11] a vertically oriented ellipse has been used to model the projection of a head in the image plane. The intensity gradient near the edge of the ellipse and a color histogram representing the interior were used to handle the parameters of the ellipse over time. Additionally, this method assumes that all pixels in the search area are equally important. The discussed tracking method does not work when the object being tracked temporarily disappears from the camera view or changes shape significantly between frames. In the method proposed here, an ellipse-based head likelihood model, consisting of gradient along the head boundary as well as a matching score between color histograms as a representation of the interior of (i) an ellipse surrounding the tracked object and (ii) a currently considered ellipse, together with depth information is utilized to find the weights of particles during tracking. Particle locations where the weights have large values are then considered to be the most likely locations of the object of interest. The particle set improves consistency of tracking by handling multiple peaks representing hypotheses in the distribution.

Although the use of color discrimination is connected with some fundamental problems such as the lack of robustness in varying illumination conditions, color is perceived as a very useful discrimination cue because of its computational efficiency and robustness against changes in target orientations. The human skin color filtering has proven to be effective in several settings and has been successfully applied in most of the face trackers relying primarily on color [12],[13],[14],[15] or on color in conjunction with other relevant information [16]. Color information is particularly useful to support a detection of faces in image sequences because of robustness towards changes in orientation and scaling of an appearance of object being in movement. The efficiency of color segmentation techniques is especially worth to emphasize when a considered object is occluded during tracking or is in shadow.

In our approach we use color histogram matching techniques to obtain information about possible location of the tracked target. The main idea of such an approach is to compute a color distribution in form of the color histogram from the ellipse's interior and to compare it with the computed in the same manner histogram representing the tracked object in the previous iteration. The better a histogram representing the ellipse's interior at specific particle position matches the reference histogram from previous iteration, the higher the probability that the tracked target at considered candidate position is. The outcome

of the histogram matching that is combined with gradient information is used to provide information about expected target location and is utilized during weighting particles.

In the context of head tracking on the basis of images from a mobile camera the features which are invariant under head orientations are particularly useful. In general, histograms are invariant to translation and rotation of the object and they vary slowly with the change of angle of view and with the change in scale. The histogram is constructed with a function $h : R^2 \rightarrow \{1...K\}$ which associates the color at location $\mathbf{y}$ to the corresponding bin. A histogram representation can be obtained in a simple way by quantizing the ellipse's interior colors into K bins and counting the number of times each discrete color occurs. Due to the statistical nature, a color histogram can only reflect the content of images in a limited way and thus the contents of the interior of the ellipses taken at small distances apart are strongly correlated. If the number of bins K is to high, the histogram is noisy. If K is too low, density structure of the image representing the ellipse's interior is smoothed. Histogram-based techniques are effective only when K can be kept relatively low and where sufficient data amounts are available. The reduction of bins makes a comparison between the histogram representing the tracked head and the histogram of candidate head faster. Additionally, such a compact representation is tolerant to noise that can result from imperfect ellipse-approximation of a highly deformable structure and curved surface of a face causing significant variations of the observed colors. The particle filter works well when the conditional densities $p(\mathbf{z}_t \mid \mathbf{s}_t)$ are reasonably flat.

It can be demonstrated that with a change of lighting conditions the major translation of skin color distribution is along the lightness axis of the RGB color space. Skin colors acquired from a static person tend to form tight clusters in several color spaces while color acquired from moving ones form widen clusters due to different reflecting surfaces. To make the histogram representation of the tracked head less sensitive to lighting conditions the HSV color space has been chosen and the V component has been represented by 4 bins while the HS components obtained the 8-bins representation.

The histogram intersection technique [17] is a popular measure between two distributions represented by a pair of histograms I and M, each containing L values. The intersection of the histograms is defined as follows: $H = \sum_{u=1}^{K} min(I^{(u)}, M^{(u)})$, where the terms $I^{(u)}$, $M^{(u)}$ represent the number of pixels inside the u-th bucket of the candidate histogram in the current frame and the histogram representing the tracked head in the previous frame, respectively, whereas K the total number of buckets. The result of the intersection of two histograms is the number of pixels that have the same color in both histograms. To obtain a match value between zero and one the intersection is normalized and the match value is determined as follows: $H_{\cap} = H / \sum_{u=1}^{K} I^{(u)}$. The work [3] demonstrated that the metric $\sqrt{1 - \rho(I, M)}$ derived from Bhattacharyya coefficient ρ is invariant to the scale of the target and therefore is superior to other measures such as histogram intersection or Kullback divergence. Considering

discrete densities the considered coefficient is defined as follows

$$\rho(I, M) = \sum_{u=1}^{K} \sqrt{I^{(u)} M^{(u)}} \tag{3}$$

Given the center of the target, a feature distribution including spatial information in color histogram can be calculated using a 2-dimensional kernel centered on the target center [18]. The kernel is used to provide the weight for a particular feature according to its distance from the center of the kernel. In order to assign smaller weights to the pixels that are further away from the region center a nonnegative and monotonic decreasing function $k : [0, \infty) \rightarrow R$ can be used [18]. The probability of particular histogram bin u at location $\mathbf{y}$ is calculated as

$$d_{\mathbf{y}}^{(u)} = C_a \sum_{i=1}^{L} k\left(\left\|\frac{\mathbf{y} - \mathbf{y}_i}{a}\right\|^2\right) \delta\left[h(\mathbf{y}_i) - u\right] \tag{4}$$

where $\mathbf{y}_i$ are pixel locations of the face candidate, L is the number of pixels in the region, δ is the Kronecker delta function and constant a is the radius of the kernel. The normalization factor

$$C_a = \frac{1}{\sum\limits_{i=1}^{L} k\left(\left\|\frac{\mathbf{y}-\mathbf{y}_i}{a}\right\|^2\right)}$$

ensures that $\sum_{u=1}^{K} d_{\mathbf{y}}^{(u)} = 1$. This normalization constant can be precalculated [3] for the utilized kernel and assumed values of a. The 2-dimensional kernels have been prepared offline and then stored in lookup tables for the future use.

The length of the minor axis of a considered ellipse is determined on the basis of depth information. Taking into account the length of the minor axis resulting from the depth information we also considered smaller and larger projection scale of the ellipse and therefore a larger as well as smaller minor axis about one pixel have been taken into account as well. The length of the minor axis has been maintained by performing the local search to maximize the goodness of the following match: $w^* = \arg\max_{w_i \in W}\{G(w_i) H_S(w_i)\}$, where G and H_S are normalized scores based on intensity gradients and color histogram similarity. In order to favor head candidates whose color distributions are similar to the target color distribution we utilized Gaussian weighting with σ variance [7]

$$H_S = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1-\rho}{2\sigma^2}} \tag{5}$$

where small Bhattacharyya distances correspond to large matching scores. The search space W comprises the ellipse's length obtained on the basis of depth information as well as smaller/larger minor axes about one pixel.

The samples are propagated on the basis of a dynamic model $\mathbf{s}_t = A\mathbf{s}_{t-1} + w_t$, where A denotes a deterministic component describing a constant velocity movement and w_t is a multivariate Gaussian random variable. The diffusion component represents uncertainty in prediction and therefore provides a way of performing a local search about a state. The weight of each hypothetical head region

$\pi_t^{(n)}$ is dependent on normalized intensity gradients and color histogram similarity which were obtained for the length of minor axis w^*.

The elliptical upright outlines with an assumed fixed aspect ratio equal to 1.4 have been prepared and stored for the future use in the construction phase. For each possible length of the minor axis we prepared off-line an elliptical outline to compute gradient and kernel lookup table to include spatial information in color histograms. Expanding the algorithm about non-upright ellipses is straightforward.

The histogram representing the tracked head has been adapted over time. This makes possible to track not only a face profile which has been shot during initialization of the tracker but in addition different profiles of the face as well as the head can be tracked. The actualization of the histogram has been realized on the basis of the equation $M_t^{(u)} = (1-\alpha)M_{t-1}^{(u)} + \alpha I_t^{(u)}$, where α is accommodation rate, I_t represents the histogram of the interior of the mean state ellipse, M_t the histogram of the target from previous frame, whereas $u = 1...K$.

4 Tracking on the Basis of Moving Camera

A kind of human-machine interaction which is useful in practice and can be very serviceable in testing a robustness of a tracking algorithm is person following with a mobile robot. In work [19] the condensation-based algorithm is utilized to keep track of multiple objects with a moving robot. The tracking experiments described in this section were carried out with a mobile robot Pioneer 2 DX [20] equipped with commercial binocular Megapixel Stereo Head. The dense stereo maps are extracted in that system thanks to small area correspondences between image pairs [21] and therefore poor results in regions of little texture are often provided. The depth map covering a face region is usually dense because a human face is rich in details and texture, see Fig. 1. Thanks to such a property this stereovision system provides a separate source of information and considerably supports the process of approximating the tracked head with an ellipse.

A typical laptop computer equipped with 2 GHz Pentium IV is utilized to run the prepared visual tracker operating at 320x240 images. The position of the tracked face in the image plane as well as person's distance to the camera are written asynchronously in block of common memory which can be easily

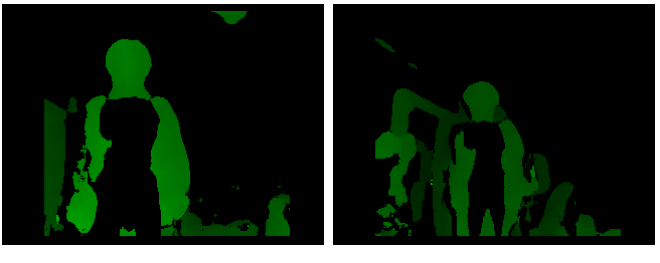

Fig. 1. Depth images (frame 1 and frame 600)

accessed by Saphira client. Saphira is an integrated sensing and control system architecture based on a client server-model whereby the robot supplies a set of basic functions that can be used to interact with it [20]. Every 100 milliseconds the robot server sends a message packet containing information on the velocity of the vehicle as well as sensor readings to the client. During tracking, the control module keeps the user face within the camera field of view by coordinating the rotation of the robot with the location of the tracked face in the image plane. The aim of the robot orientation controller is to keep the position of the tracked face at specific position in the image. The linear velocity has been dependent on person's distance to the camera. In experiments consisting in person following a distance 1.6 m has been assumed as the reference value that the linear velocity controller should maintain. To eliminate needless robot rotations as well as forward and backward movements we have applied a simple logic providing necessary insensitivity zone. The PD controllers have been implemented in the Saphira-interpreted Colbert language [20]. The tracking algorithm was implemented in C/C++ and runs at a frame rate about 10 Hz depending on image complexity.

We have undertaken experiments consisting in following a person facing the camera within walking distance without the tracked face loss. Experiments consisting in realization of only a rotation of mobile robot which can be seen as analogous to experiments with a pan-camera have also been conducted. In such experiments a user moved about a room, walked back and forth as well as around the mobile robot. The aim of such a scenario was to evaluate the quality of ellipse scaling in response of varying distance between the camera and the user during person following. Our experiment findings show that thanks to stereovision the ellipse is properly scaled and therefore because of appropriate head approximation, sudden changes of the minor axis length as well as ellipse's jumps are considerably eliminated. Figure 2 indicates selected frames from the discussed scenario, see also Fig. 1. The color of the door is very similar to that of human face and it can cause great difficulty to color-based tracking algorithms, see also image from frame 390 in Fig. 2. The region cue reflected by weighted color histogram varies slowly with slow translation of the target but does not express appropriately the content of the image with reduced scale, see image from frame 600 in Fig. 2. The likelihood model combining gradient information with a weighted histogram of the ellipse's interior demonstrated abilities to localize target correctly in case of reduced scale. The gradient modality complement the color modality when the object is moving because color information may become unreliable due to changes in the object pose and illumination, whereas strong localization cues may be obtained from the gradient information. The gradient information can therefore improve the accommodation of the color model over time. In particular, the depth information allows us to set accommodation rate α to zero when face is localized above an assumed distance to the camera.

The depth map covering the face region is usually dense and this together with skin-color and symmetry information as well as eyes-template assorted with the depth has allowed us to apply the eigenfaces method [22] and to detect the presence of the vertical and frontal-view faces in the scene very reliably and

thus to initialize the tracker automatically. Thanks to the head position it is possible to recognize some static commands on the basis of geometrical relations of the face and hands and to interact with mobile robot during person following. Using the discussed system we have realized experiments in which the robot has followed a person at distances which beyond 100 m without the person loss. By dealing with multiple hypotheses this approach can track a head reliably in cases of temporal occlusions and varying illumination conditions.

Fig. 2. Face tracking relying only upon a rotation of the moving camera (frames 1,35,390,600)

5 Evaluation Using PETS-ICVS Data Sets

The experiments described in this section have been realized on the basis of PETS-ICVS data set which has been prepared in smart meeting room. The aim of the experiments was to track the meeting participants based on static color camera. The images of size 576x720 have been converted to size of 320x240 by subsampling (consisting in selecting odd pixels in only odd lines) and bicubic based image scaling. Initialization of the tracker has been performed by searching for an elliptical object in determined in advance head-entry and head-exit zones. A simple background subtraction procedure which was executed in mentioned above boxes has proven to be sufficient in person entry/exit detection. In this version of the tracker a sample in distribution represents an ellipse described by $s = \{\mathbf{y}, \dot{\mathbf{y}}, l, \dot{l}\}$, where $\mathbf{y}$ denotes the location in the xy-image plane, $\dot{\mathbf{y}}$ motion, l the length of the minor axis and $\dot{l}$ corresponding scale change.

Figure 3 depicts example frames from a typical experiment of Scenario C which was viewed from Camera 1. The frame 13667 demonstrates a behavior

of the tracker in case of non-upright head orientation. Because of only vertical orientation of the ellipses which has been assumed in advance, the tracker fitted an ellipse in a search region in the proximity of the true location. Such a tug work of the tracker has been observed at 15 succeeded frames and after that the algorithm continued a smooth tracking of the head. In frames 13765, 13917, 14140 we can perceive a poor approximation of the head of the third person by an ellipse. But such undesirable effect has been observed occasionally during processing of PETS data sets. The number of poor misfits can be greatly reduced by utilizing the nearly constant distance of the tracked person to the camera and thus by operating with smaller range of lengths of the ellipse's minor axis. The experiments described in this section have been conducted using a relatively large range of the axis lengths which were needed during person following, namely from 6 to 30. A typical length of the ellipse's axis for the presented in the Fig. 3 frame range is about 11. Another method of improving the robustness of the tracker in situations where misfits have been observed is to combine it with fast and robust algorithm for detecting faces with out-of-plan rotation [23].

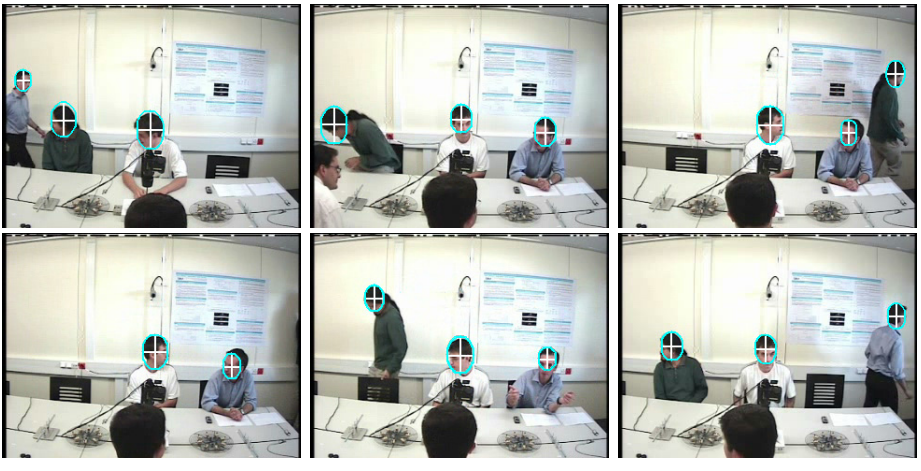

Fig. 3. Frames 11224, 13667, 13765, 13917, 14140, 14842 of Scenario C

Figure 4 illustrates example frames of tracking on the basis of the CamShift algorithm [2]. The tracker has been initialized in frame 10952, see the left frame in Fig. 4, with number of bins equals 30, Smin=40 and Vmin=60.

6 Conclusion

We have presented a vision module that robustly tracks and detects a human face. By employing shape, color, stereovision as well as elliptical shape features the proposed method can track a head in case of dynamic background. The combination of above-mentioned cues and particle filter seems to have a considerable

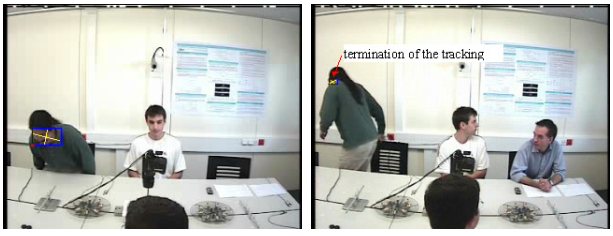

Fig. 4. Face tracking using CamShift

perspective of applications in robotics and surveillance. The algorithm is robust to sensor noise and uncertainty in localization. The resulting particle filter was able to track the head reliably, even during varying lighting conditions. Moreover, the particle filtering performs satisfactory even in the presence of partial occlusions. To show the correct work of the system, we have conducted several experiments in naturally occurring in laboratory circumstances. In particular, the tracking module enables the robot to follow a person. Thanks to the real-time robot control, the moving camera provides a considerably large searching area for a vision system. Face tracking can be used not only for directing the vision system's attention to a user/intruder but also as a prerequisite stage for face recognition and human action understanding. One of the future research directions of the presented approach is to explore the unscented particle filter [7],[24]. One difficulty of utilizing of gradient along the head boundary is the high nonlinearity of the observation likelihood and even small difference in parameters of the ellipse could involve large changes in likelihood. The unscented particle filter places the limited particles in an effective way in comparable computational overhead over the conventional particle filtering scheme.

Acknowledgment. This work has been supported by KBN within the project 4T11C01224

References

1. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body, IEEE Trans. on PAMI **19**(7) (1997) 780–785
2. Bradski, G.R.: Computer Vision Face Tracking as a Component of a Perceptual User Interface, In Proc. IEEE Workshop on Applications of Comp. Vision, Princeton (1998) 214–219
3. Recognition Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of Non-Rigid Objects Using Mean Shift, In Proc. of IEEE Conf. on Comp. Vision and Pat. Rec. (2000) 142–149
4. Isard, M., Blake, A.: Contour Tracking by Stochastic Propagation of Conditional Density, European Conf. on Computer Vision, Cambridge (1996) 343–356
5. Rui, Y., Chen, Y.: Better Proposal Distributions: Object Tracking Using Unscented Particle Filter, In Proc. IEEE Conf. on Comp. Vision and Pat. Rec. (2001) 786–793

6. Chen, Y., Rui, Y., Huang, T.: Mode-based Multi-Hypothesis Head Tracking Using Parametric Contours, In Proc. IEEE Int. Conf. on Aut. Face and Gesture Rec. (2002) 112–117
7. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An Adaptive Color-Based Particle Filter, Image and Vision Computing **21**(1) (2003) 99–110
8. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-Based Probabilistic Tracking, European Conf. on Computer Vision (2002) 661–675
9. Schulz, D., Burgard, W., Fox, D., Cremers A.B.: Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association, In Proc. of the IEEE Int. Conf. on Robotics and Automation (2001) 1665–1670
10. Isard, M., Blake, A.: A Mixed-State Condensation Tracker with Automatic Model-Switching, In Proc. of IEEE Int. Conf. on Comp. Vision, Mumbai (1998) 107–112
11. Recognition, Santa Birchfield, S.: Elliptical Head Tracking Using Intensity Gradients and Color Histograms, In Proc. of IEEE Conf. on Comp. Vision and Pat. Rec., Santa Barbara (1998) 232–237
12. Hunke, M., Waibel, A.: Face Locating and Tracking for Human-Computer Interaction, In Proc. of the 28th Asilomar Conf. on Signals, Systems and Computers (1994) 1277–1281
13. Fieguth, P., Terzopoulos, D.: Color-Based Tracking of Heads and Other Mobile Objects at Video Frame Rates, In Proc. of the IEEE Conf. on Comp. Vision Pat. Rec., Hilton Head Island (1997) 21–27
14. Schwerdt, K., Crowley, J.L.: Robust Face Tracking Using Color, In Proc. of the Int. Conf. on Automatic Face and Gesture Rec. (2000) 90–95
15. Sobottka, K., Pitas, I.: Segmentation and Tracking of Faces in Color Images, In Proc. of the Second Int. Conf. on Automatic Face and Gesture Rec. (1996) 236–241
16. Darrell, T., Gordon, G., Harville, M., Woodfill, J.: Integrated Person Tracking Using Stereo, Color, and Pattern Detection, Proc. of IEEE Conf. on Computer Vision and Pat. Rec., Santa Barbara (1998) 601–609
17. Swain, M.J., Ballard, D.H.: Color Indexing, Int. Journal of Computer Vision **7**(1) (1991) 11–32
18. Cheng, Y.: Mean Shift, Mode Seeking, and Clustering, IEEE Trans. on PAMI **17**(8) (1995) 790–799
19. Meier, E.B., Ade, F.: Using the Condensation Algorithm to Implement Tracking for Mobile Robots, In Proc. of the Third European Workshop on Advanced Mobile Robots (1999) 73–80
20. ActivMedia Robotics, Pioneer 2 mobile robots (2001)
21. Konolige, K.: Small Vision System: Hardware and Implementation, Proc. of Int. Symp. on Robotics Research, Hayama (1997) 111–116
22. Turk, M.A., Pentland, A.P.: Face Recognition Using eigenfaces, Proc. of IEEE Conf. on Comp. Vision and Pat. Rec. (1991) 586–591
23. Schneiderman, H., Kanade, T.: A Histogram-Based Method for Detection of Faces and Cars, In Proc. of the 2000 Int. Conf. on Image Processing (2000) 504–507
24. Merwe, R., Doucet, A., Freitas, N., Wan, E.: The Unscented Particle Filter, Advances in Neural Information Processing Systems (2000) 584–590