3D Model-Based Tracking of the Human Body in Monocular Gray-Level Images

Bogdan Kwolek

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

Abstract. This paper presents a model-based approach to monocular tracking of human body using a non-calibrated camera. The tracking in monocular images is realized using a particle filter and an articulated 3D model with a cylinder-based representation of the body. In modeling the visual appearance of the person we employ appearance-adaptive models. The predominant orientation of the gradient combined with ridge cues provides strong orientation responses in the observation model of the particle filter. The phase that is measured using the Gabor filter contributes towards strong localization of the body limbs. The potential of our approach is demonstrated by tracking of the human body on real videos.

1 Introduction

Accurate and reliable tracking of three-dimensional human body is an important problem. It is particular substantial for service robots that should understand and predict the behavior of humans' beings in their vicinity. Although there exist several methods to perform 3D body tracking [1][2][3], there is still a need to improve the accuracy and reliability of such systems.

The method we utilize in our work relies on a 3D-2D matching between 3D features of a generic human model and corresponding 2D features extracted in a monocular image sequence. Given a 3D object model, the pose estimation problem can be defined as recovering and tracking the model parameters that include translation, rotation, and joint angles, so that the back-projected 3D model primitives match the 2D image features which have been extracted through an image analysis. Extracting 3D body configurations from monocular and uncalibrated video sequences is coupled with complex modeling as well as difficulties related to feature extraction. The process of extraction of the features in a cluttered scene and their matching to a self-occluding body model is an inherently complex task. The non-observable states are due to motions of body segments towards or away from the camera. The ambiguity, non-linearity, and non-observability make the posterior likelihoods over human pose space multi-modal and ill-conditioned. Efficient locating the body features for the 3D-2D matching during estimating the pose of the human body is therefore an important problem.

This work is motivated by necessity of equipment of mobile robot with basic capabilities of understanding and predicting some human behaviors. A service robot that does not understand people behavior might be dangerous to people and environment in a situation that has not been forecasted by a designer. Such understanding capability is therefore a basic prerequisite for service robots.

In this work, we focus on real-time estimating the pose of the upper body using a monocular and uncalibrated camera and a 3D model of human body. The tracking of the 3D model is realized using a kernel particle filter. By using such a filter we avoid the need for a huge number of particles to represent the probability distributions in high dimensional state space [4]. To accomplish strong localization of the body parts we utilize phase that is measured on the basis of Gabor filter responses. By using it our intent is to provide an alternative to edge and ridge cues when they are unreliable, and to supplement them if they are. In order to distinguish the person from the cluttered background effectively we employ appearance-adaptive models. The ridge cue that is combined with predominant orientation of the gradient in a window surrounding the feature provides strong orientation responses in the observation model. In this context, the novelty of our approach lies in the introduction of phase measured via the Gabor filter, predominant orientation as well as adaptive appearance models which results in an observation model with better localization of the limbs.

The paper is organized as follows. Next Section contains an overview of related work. Section 3. is devoted to a short presentation of the bases of probabilistic tracking. The image processing is described in Section 4. We then describe the components and details of the system. We present and discuss experimental results in Section 6. The paper concludes with a summary in the last Section.

2 Related Work

One of the first applications to track a human body in real-time using a single camera is the PFINDER system [5]. The applied body model is rather coarse and tracking provides only information about the position of head, hands and feet. Tracking of a human in 3D with limited computational resources on a mobile robot was described by Kortenkamp et al. [1]. This approach used depth information from a stereo camera to track a 3D body model. The work [6] employs an articulated 3D body model built on truncated cones and a cost function-based on edge and silhouette information. The images have been acquired using 3 calibrated cameras in scenarios with a black background. The experiments have been conducted with 4000 particles using an annealing particle filter. The computation times of this system are far from real-time. Similarly to mentioned work the multiple cameras are often employed to cope with body self-occlusions [7][8][9]. Some of the mentioned above approaches construct a probabilistic model of the body [5][7], whereas other approaches are based on a body model [6][9]. Using a body model only few authors have addressed the problem of 3D body tracking on the basis of uncalibrated monocular cameras [4][6][10][11][12].

3 Particle filtering

Particle filter is an inference technique for estimating the *posterior* distribution $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ for the object state $\mathbf{x}_t$ at time t given a sequence of observations $\mathbf{z}_{1:t}$. For nonlinear models, multi-modal, non-Gaussian or any combination of these models the particle filter provides a Monte Carlo solution to the recursive filtering equation $p(\mathbf{x}_t \mid \mathbf{z}_{1:t}) \propto p(\mathbf{z}_t \mid \mathbf{x}_t) \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1})\ p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1}$. With this recursion we calculate the *posterior* distribution, given a dynamic model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ describing the state evolution and an observation model $p(\mathbf{z}_t \,|\, \mathbf{x}_t)$ describing the likelihood that a state $\mathbf{x}_t$ causes the measurement $\mathbf{z}_t$.

The main idea of particle filtering is to represent the probability distribution by a set of weighted particles $S = \{(\mathbf{x}_t^{(n)}, \pi_t^{(n)}) \mid n = 1, ..., N\}$ evolving over time on the basis of simulation-updating scheme. The resampling selects with higher probability particles that have a high likelihood associated with them, while preserving the asymptotic approximation of the particle-based posterior representation. Sequential Importance Sampling (SIR) is the basic algorithm with resampling applied at every time step. The prior $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}^{(n)})$ is used as the importance density for drawing samples. Within the SIR scheme the weighting equation takes the form $\pi_t^{(n)} \propto p(\mathbf{z}_t \mid \mathbf{x}_t^{(n)})$. This simplification gives a variant of a well-known particle filter in computer vision, CONDENSATION [13]. Since the optimal importance density relies on both the present observation $\mathbf{z}_t$ and previous state $\mathbf{x}_{t-1}$ which are not considered in such a scheme, the SIR-based sampling is not too effective.

4 Image Processing

At the beginning of this section, we show how low-level features for determining the limb orientation are computed. Gabor filter will be presented as the second topic. A description of appearance modeling ends this section.

4.1 Low-level features

The edge is employed in comparison of the gradient angle with the limb angle β resulting from the 3D model. The gradient is computed at the position in the image plane, where the edge of the considered limb is projected. The orientation of the whole limb can be determined by averaging over the feature orientations laying on the projected cylinder. The model is generated in a variety of possible configurations and overlaid on the image to find the true body pose. The ridge detection is based on the LoG filter. The variance of the Gaussian can be chosen such that the feature of interest is highlighted. Therefore the ridge cues are employed to find in the image the elongated structures of a specified thickness. The response of the ridge cue depends on the size of the limb at the image and thus it strongly depends on the distance of the limb to the camera. On the basis of the estimated distance of the limb to the camera the appropriate level in the Gaussian pyramid is chosen to obtain the suitable responses of the ridge filter.

The steered response is calculated by applying an interpolation formula to the second partial derivatives d:

$$f_R(\beta) = \left|\sin^2\beta\, d_{xx} + \cos^2\beta\, d_{yy} - 2\sin\beta\,\cos\beta\, d_{xy}\right| - \left|\cos^2\beta\, d_{xx} + \sin^2\beta\, d_{yy} + 2\sin\beta\,\cos\beta\, d_{xy}\right| . \quad (1)$$

4.2 Predominant orientation of the feature

The orientation of a feature is assumed as the predominant orientation of the gradient in a window around the considered feature. The predominant orientation is calculated as the quadratically interpolated maximum of the histogram of the gradient orientations within a window around the feature. The histogram is weighted both by the magnitude of the gradient and a Gaussian window centered on the feature. Before determining the maximum the histogram is smoothed by a moving average filter. In addition, each local maximum with a value above 0.8% of the global maxium is retained.

4.3 Gabor filter

The choice of Gabor filter responses is biologically motivated since they model the response of human visual cortical cells [14]. Gabor filters extract the orientation-dependent frequency contents, i.e. edge like features. The main advantage of Gabor wavelets is that they allow analysis of signals at different scales, or resolution, and further they accommodate frequency and position simultaneously. Gabor filters remove most of variation in lighting and contrast. They are also robust against shifts and small object deformations. The Gabor wavelet is essentially a sinewave modulated by a Gaussian envelope. The 2-D kernel of Gabor filter is defined in the following manner [15]:

$$f(x, y, \theta_k, \lambda) = \exp\left[-\frac{1}{2}\left\{\frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2}\right\}\right] \exp\left\{i\frac{2\pi R_1}{\lambda}\right\} \quad (2)$$

where $R_1 = xcos\theta_k + ysin\theta_k$ and $R_2 = -xsin\theta_k + ycos\theta_k$, σ_x and σ_y are the standard deviations of the Gaussian envelope along the x and y dimensions, λ and θ_k are the wavelength and orientation of the sinusoidal plane wave, respectively. The spread of the Gaussian envelope is defined in terms of the wavelength λ. θ_k is defined by $\theta_k = \frac{\pi(k-1)}{n}$, $k = 1, 2, ..., n$, where n denotes the number of orientations that are taken into account. For example, when $n = 4$, four values of orientation θ_k are used: 0^o, 45^o, 90^o, and 135^o.

A Gabor filter response is achieved by convolving the filter kernel given by (2) with an image. The response of the filter for sampling point (x, y) is as follows:

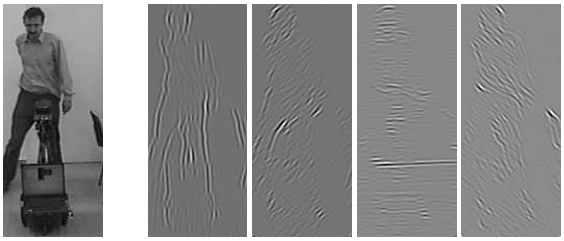
$$g(x, y, \theta_k, \lambda) = \sum_{u=-(N-x)}^{N-x-1} \sum_{v=-(N-y)}^{N-y-1} I(x+u, y+v) f(u, v, \theta_k, \lambda) \quad (3)$$

where $I(x, y)$ denotes a $N \times N$ grayscale image.

In this work four different orientations and four different wavelengths have been utilized, see Fig. 1. The Gabor filter responses were used to locally measure the phase. In contrast to work [16] we are not interested in determining the highest score for the features being in the correspondence, but in determining a weighting factor to express degree of similarity between potential matches. The simplest way to achieve this goal is to use a Gabor filter with orientation θ and scale λ to extract the phase $\phi_{\theta,\lambda}$ of features i and j and then to compare the considered features according to: $\exp(- \mid \phi_{\theta,\lambda}(i) - \phi_{\theta,\lambda}(j) \mid)$. Using the phase of all filters we obtain the following correspondence measure:

$$G_{ij} = c \prod_{\theta,\lambda} \exp(- \mid \phi_{\theta,\lambda}(i) - \phi_{\theta,\lambda}(j) \mid) \tag{4}$$

where c is a normalization constant ensuring that G_{ij} varies between 0 and 1.



a) b)

Fig. 1. Test image of size 100×240 (a). Gabor decomposition of test image at four different orientations and for real channels (b).

Figure 2 demonstrates pairs of images with some phase-based matching results. Left images in each pair depict the locations at the corners of the projected model that have been employed in matching, whereas the right images depict coherence probability between pixels at marked locations and image pixels from Fig. 1a. The images illustrate the Gabor wavelet's capability to match coherent image structures in subsequent frames during human model tracking.

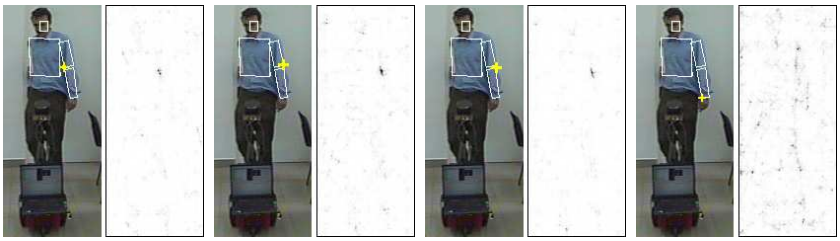


Fig. 2. Searching for pixel coherence using Gabor filter responses. Left images in the pair demonstrate the locations of pixels undergoing matching, the right ones are probability images expressing coherence between the marked pixel and image at Fig. 1a.

4.4 Visual appearance modeling using adaptive models

Our intensity-based appearance model was inspired by work [17] which proposed the model $\mathcal{WSL}$ consisting of three components. The Wandering-Stable-Lost components are employed to track an object while adapting to slowly changing appearance and providing robustness to partial occlusions. During model adaptation each $\mathcal{WSL}$ component votes according to its level of stability. The method has been shown to yield reliable tracking of human faces.

By the usage of 3D human model in this method we can provide and additional support for handling self-occlusions and therefore restrict the adaptation of parametric model to visible pixels. In our approach the model that is assigned to a single data observation d_t consists of three components, namely, the W-component expressing the two-frame variations, the S-component characterizing the stable structure within all previous observations and C-component representing a constant object template. The model represents thus the appearances existing in all observations up to time $t-1$. It is a mixture of Gaussians with centers $\{\mu_{i,t} \mid i = w, s, c\}$ and their corresponding variances $\{\sigma^2_{i,t} \mid i = w, s, c\}$ and mixing probabilities $\mathbf{m}_t = \{m_{i,t} \mid i = w, s, c\}$. The mixture probability density for a new data d_t conditioned on the past observations can be expressed by

$$p(d_t|d_{t-1}, \mu_{s,t-1}, \sigma^2_{s,t-1}, \mathbf{m}_{t-1}) = m_{w,t-1}p_w(d_t|\mu_{w,t-1}, \sigma^2_w) + m_{s,t-1}p_s(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1}) + m_{c,t-1}p_c(d_t|\mu_{c,0}, \sigma^2_c). \quad (5)$$

In wandering term the mean is the observation d_{t-1} from the previous frame and the variance is fixed at σ^2_w. The stable component $p_s(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1})$ is intended to express the appearance properties that are relatively stable in time. A Gaussian density function with slowly accommodated parameters $\mu_{s,t-1}, \sigma^2_{s,t-1}$ captures the behavior of such temporally stable image observations. The fixed component accounts for data expressing the similarity with initial object appearance. The mean is the observation d_0 from the initial frame and the variance is fixed at σ^2_c

Similarly to [17] we assume that with respect to the contributions to current appearance the previous data observations are forgotten according to exponential function. The update of the current appearance model A_{t-1} to A_t is done using the on-line EM algorithm. The posterior ownership probabilities $\{o_{i,t} \mid i = w, s, c\}$ forming a distribution (with $\sum_i o_{i,t}(d_t) = 1$) are computed in E-step for each data d_t as follows:

$$o_{i,t}(d_t) = m_{i,t-1}p_i(d_t|\mu_{i,t-1}, \sigma^2_{i,t-1}) \mid i = w, s, c, \quad (6)$$

Then the mixing probabilities are updated using an accommodation factor α according to:

$$m_{i,t} = \alpha o_{i,t} + (1-\alpha)m_{i,t-1} \mid i = w, s, c. \quad (7)$$

The higher the α value, the faster the model adapts to the new data. During the M-step the ML estimates of the mean and variance are computed using the

moments of the past observations. The first and the second-moment images are computed recursively in the following manner:

$$M_t^{(p)} = \alpha d_t^j o_{s,t}(d_t) + (1-\alpha) M_{t-1}^{(p)} \mid p = 1, 2. \tag{8}$$

In the last step the mixture centers and the variances are updated as follows:

$$\begin{aligned} &\mu_{s,t} = \frac{M_t^{(1)}}{m_{s,t}}, \quad \mu_{w,t} = d_{t-1}, \quad \mu_{c,t} = d_0, \\ &\sigma_s^2 = \frac{M_t^{(2)}}{m_{s,t}} - \mu_{s,t}^2, \quad \sigma_{w,t}^2 = \sigma_{w,0}^2, \quad \sigma_{c,t}^2 = \sigma_{c,0}^2. \end{aligned} \tag{9}$$

In order to initialize the model the initial moment images are set using the following formulas: $M_0^{(1)} = m_{s,0}\, d_0$ and $M_0^{(2)} = m_{s,0}(\sigma_{s,0}^2 + d_0^2)$.

To demonstrate the usefulness of the adaptive appearance models in tracking we performed various experiments on freely available test sequences. Figure 3 depicts some tracking results that were obtained on PETS-ICVS 2003 test sequence. The original images 768 pixels wide and 576 high have been converted to size of 320×240 by subsampling (consisting in selecting odd pixels in only odd lines) and bicubic based image scaling. Inference was performed using 50 particles and CONDENSATION algorithm. The observation likelihood was calculated according to the following equation:

$$p(\mathbf{z}_t \mid \mathbf{x}_t) = \prod_{j=1}^{M} \sum_{i=w,s,c} \frac{m_{i,t}(j)}{\sqrt{2\pi\sigma_{i,t}^2(j)}} \exp\left[-\frac{d_t(j) - \mu_{i,t}(j)}{2\sigma_{i,t}^2(j)} \right], \tag{10}$$

where d_t denotes the value of gray pixel, M is the number of pixels, and $i = w, s, c$. The samples are propagated on the basis of a dynamic model $\mathbf{x}_t = \mathbf{x}_{t-1} + w_t$, where $\mathbf{x}_t = \{x, y\}$ and w_t is a multivariate Gaussian random variable. The size of face pattern in this sequence is comparable to the size of face pattern from Fig. 1a. We can observe that despite the fixed size of the object template the method enables reliable tracking of the face. The cycle time of the tracking at P IV 2.4 GHz is approximately 0.014 sec.

Fig. 3. Face tracking using adaptive appearance models (top row). The tracking starts at frame #11000 and the images from left to right depict some tracking results in frames #11010, #10050, etc. Next rows present the evolution of the mean of stable (middle row) and wandering (bottom row) components during tracking.

Figure 4. illustrates how the mixing probabilities at two locations (x=7, y=10 and x=9, y=10) in the window of the tracker evolve over time. The static size

template is not subject to drifting and algorithm adapts to changing appearances of the face. The mixing probabilities change something more in the second part of the sequence, i.e. starting at frame #11000 + 200, see also Fig. 3.

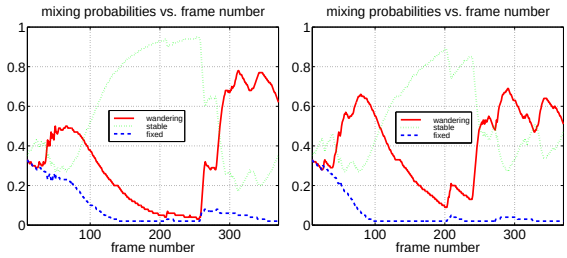


Fig. 4. Mixing probabilities versus frame number.

5 3D body tracking

Modeling the appearance of humans will be presented as the first topic in this section. The observation and motion models will be presented afterwards.

5.1 Modeling the visual appearance of humans

We use an articulated 3D body model composed of cylinders with ellipsoid cross sections [10][4]. Such representation gives best outcomes when the cylinder is observed by a camera from the side. The more the cylinder axis is parallel to the optical axis of the camera, e.g. arm pointing directly into the camera, the more the pose estimation is inaccurate. A given pose configuration is determined by the relative joint angles between connected limbs and the position and orientation of the torso. The kinematic structure is completed by individual joint angle limits, which model the physical constraints of the human body. A typical model of an upper body with two arms has 13 degrees of freedom. The 3D body model is back-projected into the image plane through a pinhole-camera model. This yields an approximate representation of the 3D body model in the image plane consisting of 2D polygons for each limb. The model is generated in a variety of possible configurations and overlaid on the image plane to find the true body pose.

Given three coordinates and three angles determining the pose of the torso we employed a simple cylindrical model depicted in Fig. 5a to re-render the object texture into the requested object pose. Using such re-rendered images and the adaptive appearance models we can compute the observation likelihoods. In our approach the re-rendering is only applied to torso, see Fig. 5c, but for demonstrational purposes we utilize face images to show the usefulness of re-rendering in template matching, see Fig. 5b. The images from bottom row in

Fig. 5b depict some re-rendered faces to frontal pose. These images demonstrate that the usage of such simple cylindrical model consisting of only 12 triangles can lead to better object representation for object matching based on a model containing an initial template. In particular, we can observe that the re-rendered face from the fifth bottom image in Fig. 5b would give a high similarity score in case of comparison with the frontal face template.

a) b) c)

Fig. 5. 3D cylindrical face/torso model (a). The cylindrical model overlaid on face images (b) and torso (c). Bottom images in the column (b) depict re-rendered images.

5.2 Motion and observation model

One way to model the transition of the state is using a random walk which can be described by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta, \tag{11}$$

where $\eta \sim N(0, \nu^2)$, and ν^2 is typically learned from training sequences. The reason for not using a specific motion model in the particle filter is that we want to track general human motions.

Assuming that the cues and limbs are independent the observation model takes the form:

$$p(\mathbf{z}_t \,|\, \mathbf{x}_t) = \prod_{i=1}^{L} \prod_{k=r,a,g,d} p(i,k), \tag{12}$$

where L denotes number of cylinders in the 3D model, r represents ridge, a stands for adaptive appearance model, g accounts for phase, d is dominant orientation and $p(i,k)$ denotes the likelihood of the whole limb i for cue k. The likelihood is calculated on the basis of the following Gaussian weighting: $p(\mathbf{z}^C \,|\, \mathbf{x}) = (\sqrt{2\pi}\sigma)^{-1} e^{-\frac{1-\rho}{2\sigma^2}}$, where ρ denotes the response of the filter for the considered cue. The filter responses are calculated on the basis of techniques described in Section 4.

To exploit the structure of the probability density distribution and to reduce the number of particles an iterative mode-seeking via the mean-shift is employed. The density distribution is estimated through placing a kernel function on each particle and then shifting the particles to high weight areas [4].

6 Experiments

To test the proposed method of 3D body tracking in monocular images we performed several experiments on real images. Experiments demonstrated that through the tracking of the human legs the robot is able to detect collisions. We have also looked at how the system tracks the upper body. The lower arms are one of the hardest body parts to track because in comparison with legs they are smaller and move faster. The modeling their motion is harder. In contrast, the face has more features and textures than the legs and arms and can be tracked easily.

During computation of likelihoods we utilize various combinations of cues. The ridge cues as well as dominant orientations are not utilized in computation of likelihoods of torso, face and hands. The torso and face are tracked in conjuction. The phase is computed using the fast method [18] at four points of the lower arm cylinders. The dominant orientation is extracted at 20 points, whereas the ridge at 30 points of each cylinder.

In order to show the stability of our approach to images with complex background, we have performed experiments in a typical home/office environment. Figure 6 shows some tracking results. The person depicted at this figure is wearing a green shirt and is standing in front of plants. The color of the wall in the background is similar to color of human skin. The mentioned above figure illustrates the behavior of the algorithm in such conditions. The tracking has been done using the kernel particle filter doing 3 mean-shift iterations and built on 200 particles. The adaptive appearance models were employed in computation of the likelihoods of torso, face and lower arms. Other experiments demonstrated that this algorithm allows the tracking of the human body in cluttered environments if no large self-occlusions occur.

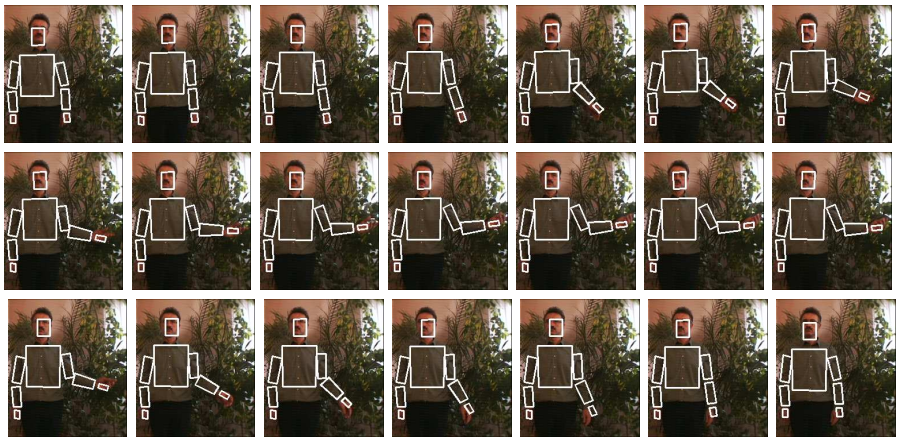


Fig. 6. Tracking of the upper body using gray and monocular images. Every 10^{th} frame from the sequence is depicted.

The algorithm can estimate a direction of pointing towards an object using only 110 particles and doing 2-3 mean shift iterations. Using such settings the algorithm calculates the estimate of 3D hand position at a frame rate about 8 Hz. A configuration of the algorithm with 500 and even 250 particles gives repeatable results. The algorithm was tested on images of size 320x240 acquired by EVI-D31 color camera. A typical laptop computer equipped with 2.4 GHz Pentium IV is utilized to run the C/C++ software under Windows control.

Initial experiments demonstrate that a proposal distribution in form of a Gaussian that is generated on the basis of the face position estimated in advance by a particle filter can lead to shorter computation time. In particular, this particle filter can perform the tracking of the face using smaller number of particles in comparison with the 3D tracker. Such configuration of the system can be used in many scenarios with a person facing the camera. Comparing the appearance-based cues with color we found that the former provides better results. The dominant orientation gives better estimates in comparison to edges. The overall performance of the system built on adaptive appearance models and dominant orientation is better in comparison to a configuration of system based on color and edges. We expect a prior model of body motion can likely improve robustness of the system further. Color images would provide additional information in stabilization of the tracker. But determining which image features are most informative requires more research. The initialization of the tracking algorithm is done manually before starting the tracking. An occlusion test is performed, particularly with respect to pixels for which the Gabor filter responses are calculated.

7 Conclusions

We have presented a model-based approach for monocular tracking of the human body using a non-calibrated camera. By employing ridge, dominant orientation, phase and appearance the proposed method can estimate the pose of the upper body using a monocular and uncalibrated camera. The proposed combination of cues leads to higher performance of the system and better quality of tracking. Once the human body is being tracked, the appearance model adapts according to changes in appearance and therefore improves tracking performance. Experimental results, which were obtained in a typical home/office environment show the feasibility of our approach to estimate the pose of the upper body against a complex background. The resulting system runs in real-time on a standard laptop computer installed on a real mobile agent.

Acknowledgment

This work has been supported by Polish Ministry of Education and Science (MNSzW) within the projects 3 T11C 057 30 and N206 019 31/2664.

References

1. Kortenkamp, D., Huber, E., Bonasso, R.P.: Recognizing and interpreting gestures on a mobile robot. In: Proc. Nat. Conf. on Artificial Intelligence. (1996) 915–921
2. Cham, T., Rehg, J.: A multiple hypothesis approach to figure tracking. In: Int. Conf. on Computer Vision and Patt. Recognition. (1999) 239–245
3. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. Int. J. Comput. Vision **61** (2005) 185–205
4. Fritsch, J., Schmidt, J., Kwolek, B.: Kernel particle filter for real-time 3d body tracking in monocular color images. In: IEEE Int. Conf. on Face and Gesture Rec., Southampton, UK, IEEE Computer Society Press (2006) 567–572
5. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 780–785
6. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: IEEE Int. Conf. on Pattern Recognition. (2000) 126–133
7. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M.: Tracking loose-limbed people. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition. (2004) vol. 1, 421–428
8. Rosales, R., Siddiqui, M., Alon, J., Sclaroff, S.: Estimating 3d body pose using uncalibrated cameras. In: Int. Conf. on Computer Vision and Pattern Recognition. (2001) 821–827
9. Kehl, R., Bray, M., Gool, L.V.: Markerless full body tracking by integrating multiple cues. In: ICCV Workshop on Modeling People and Human Interaction, Beijing, China (2005)
10. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In: European Conference on Computer Vision. (2000) 702–718
11. Sminchisescu, C., Triggs, B.: Mapping minima and transitions of visual models. In: European Conference on Computer Vision, Copenhagen (2002)
12. Urtasum, R., Fleet, D.J., Fua, P.: Monocular 3-d tracking of the golf swing. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition. (2005) vol. 2, 932–938
13. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. of Computer Vision **29** (1998) 5–28
14. Jones, J., Palemer, L.: An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology **58** (1987) 1233–1258
15. Nixon, M., Aguado, A.: Feature extraction and image processing. Newnes, Oxford, Boston (2002)
16. Fleet, D.: Disparity from local weighted phase-correlation. In: Proc. IEEE Int. Conf. on System Man and Cybernetics (SMC). (1994) 46–48
17. Jepson, A.D., Fleet, D.J., El-Maraghi, T.: Robust on-line appearance models for visual tracking. PAMI **25** (2003) 1296–1311
18. Nestares, O., Navarro, R., Portilla, J., Tabernero, A.: Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. J. Electronic Imaging **7** (1998) 166–173