Full Body Motion Tracking in Monocular Images Using Particle Swarm Optimization

Boguslaw Rymut, Tomasz Krzeszowski, and Bogdan Kwolek

Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warszawa, Poland
bytom@pjwstk.edu.pl

Abstract. The estimation of full body pose in monocular images is a very difficult problem. In 3D-model based motion tracking the challenges arise as at least one-third of degrees of freedom of the human pose that needs to be recovered is nearly unobservable in any given monocular image. In this paper, we deal with high dimensionality of the search space through estimating the pose in a hierarchical manner using Particle Swarm Optimization. Our method fits the projected body parts of an articulated model to detected body parts at color images with support of edge distance transform. The algorithm was evaluated quantitatively through the use of the motion capture data as ground truth.

1 Introduction

At present human behavior understanding is becoming one of the most active and extensive research topics of artificial intelligence and cognitive sciences. The strong interest is driven by broad spectrum of applications in several areas such as visual surveillance, human-machine-interaction and augmented reality. Tracking of human behavior inherently involves localization of body parts and estimation of the body pose [11]. Pose estimation can be approached with different ways depending on the image sensor configuration and the scenarios. The approaches can be categorized as either model-based and model-free ones [9]. In [13], an example-based approach for view-invariant estimation of 3D pose of upper body using single image has been proposed. In model-based approach, which uses a priori model of the subject to guide the pose estimation, the markerless motion tracking is typically more robust and accurate. In such an approach, the pose estimation is usually formulated as an optimization problem aiming at seeking the pose parameters, which minimize the errors between the projected 3D body segments and the image observations. One of the major difficulties in recovering human pose from 2D images is the high number of degrees-of-freedom (DOF) in the body's movement that has to be estimated. Generally, a human body consists of no less than 10 large body parts, equating to more than 20 DOF that are needed for describing realistic human movements.

Reconstructing 3D human poses from monocular images is considerably more difficult than 3D pose estimate from multiple views. The challenges to be addressed in single camera-based pose estimation are depth and observation ambiguities, self-occlusions, and last but not least the matching imperfect and very

flexible model to cluttered images. Observation ambiguities take place since any image observation can be mapped to several 3D human poses. Besides the difficulties mentioned above, for any realistic human model at least one-third of DOFs are almost unobservable in any given monocular image. In consequence, without depth information it is challenging to reconstruct skeleton in 3D. A successful approach to recovering 3D human body pose from monocular images is presented in [1], which consists in the use of direct nonlinear regression of join angles against histogram-of-shape-context silhouette shape descriptors. The most successful algorithm to date is based on propagating a mixture of Gaussians, which approximate the probability density functions representing the probable 3D poses [14]. The key contribution is efficient and exhaustive searching of the cost surface relating the candidate body configurations to image features. However, it is unclear if without explicit mechanism for re-initialization the propagation of multimodal distribution over longer period of time remains reliable.

The typical framework to human pose estimation is to fit the geometrical models to the image features by the use a deterministic or stochastic strategy. In 3D model based estimation of the human pose in monocular image sequences the particle filters are widely used. Particle filters [3] are recursive Bayesian filters that are based on Monte Carlo simulations. They approximate a posterior distribution for the configuration of a human body given a series of observations. The high dimensionality of articulated body motion requires huge number of particles to represent well the posterior probability of the states. In such spaces, sample impoverishment may prevent particle filters from maintaining multimodal distribution for long periods of time. Therefore, many efforts have been spent in developing methods for confining the search space to promising regions with true body pose. In [12], Schmidt *et al.* proposed a kernel particle filter to effectively explore the probability distributions and achieved reliable real-time tracking of the upper-body in monocular image sequences. Another possibility to constrain the configuration space is to use hierarchical search. In such an approach, a part of the articulated model is localized independently in advance, and then its location is used to constrain the search for the remaining limbs. In [4], an approach called search space decomposition is proposed, where on the basis of color cues the torso is localized first and then it is used to confine the search for the limbs. Recently, Particle Swarm Optimization (PSO) algorithm was proposed to achieve full body motion tracking using single [8] and multiple cameras [15][5]. PSO is a population based stochastic optimization technique [6], which shares many similarities with evolutionary computation techniques. It has been shown to perform well on many nonlinear and multimodal optimization problems.

In this paper, we present an approach for 3D model based reconstructing the 3-dimensional motions of human figure in monocularly-viewed image sequences. Full body pose estimation is performed in a hierarchical manner using PSO. At the beginning of each frame we determine the pose of the torso and afterwards the pose of the remaining limbs. To obtain reliable motion tracking we segment of the person's silhouette into torso and limbs. In order to obtain better orientation of the torso we take into account the direction of person's walking.

2 PSO for Dynamic Optimization

PSO maintains a swarm of particles, where each one represents a candidate solution. Every particle determines its own position, moves with its own velocity in the multidimensional search space and determines its fitness using an objective function $f(x)$. At the beginning each individual is initialized with a random position and velocity. During searching for the best fitness each particle is attracted towards the position that is affected by the best position p_i found so far by itself and the global best position g found by the whole swarm. The i-th particle's velocity and position are updated according to the following two equations:

$$v_i^{k+1} = \omega v_i^k + c_1 r_1 (p_i - x_i^k) + c_2 r_2 (g - x_i^k) \tag{1}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{2}$$

where the constants c_1 and c_2 are used to balance the influence of the individual's knowledge and that of the group, respectively, r_1 and r_2 are uniformly distributed random numbers, x_i is position of the i-th particle, p_i is the local best position of particle i, whereas g stands for the global best position, and ω is an inertia constant. The swarm stops the updating when a termination criterion is met. Because the pose tracking is a dynamic optimization problem, in order to cover possible pose changes the particles are propagated according to weak transition model when a new image becomes available.

3 3D Body Model and Cost Function

3.1 Human Body Model

The articulated human body model is represented as a kinematic tree consisting of 11 segments. It is made of truncated cones that model the pelvis, torso/head, upper and lower arm and legs. Its 3D pose is defined by 26 DOF and it is determined by position and orientation of the pelvis in the global coordinate system and the relative angles between the connected limbs. The perspective projection is used in mapping the model onto 2D image plane. In this way we attain the image of the 3D model in a given configuration, which can then be matched to the person extracted through image analysis. The aim of the tracking is to estimate the pose of the pelvis and the joint angles and this is achieved by maximizing the fitting cost.

3.2 Body Part Detection

Our approach to full body motion tracking in monocular images is motivated by findings from other previous work, which stresses the importance of good 2D features to achieve reliable human pose estimation, cf. [10]. However, detection of body parts, such as torso and the limbs in color images is difficult due to variations caused by varying shape, appearance, clothing, etc.

In first stage of our algorithm the background subtraction is performed using algorithm [2]. The binary foreground image is then employed in determining the silhouette-overlap degree. The silhouette features extracted via background subtraction are complemented by image edges, which contribute towards more precise aligning the body parts. At this stage a cost of fitting the projected model edges to the image edges is determined. The most common approach to edge detection is based on image gradient, which shares many properties with optical flow. In particular, the gradient features are independent from background subtraction. Gradient angle is invariant to global changes of image intensities. In contrast to optical flow, gradients features are discriminative for both moving and non-moving body parts. In our approach, the gradient magnitude is masked by the closed image of the foreground. In this fashion we obtain edges belonging only to the person undergoing tracking. They are then employed to generate the edge distance map, see also Fig. 1. The distance map assigns each pixel a value that is the distance between that pixel and the nearest nonzero edge pixel. In our implementation we employ chessboard distance and limit the number of iterations on the chain propagation to three. A color histogram in HSI color space, quantized into $8 \times 8 \times 8$ bins was used to approximate the distribution of the skin color. The skin color areas were detected via histogram backprojection and then refined using a skin-locus [7]. Owing to skin-locus it is possible to successfully delineate the skin areas even in front of wooden planking, see also images in first row at Fig. 1. The torso has been detected using histogram-based model of color distribution in HSI color space. The remaining part of the foreground blob was segmented as legs.

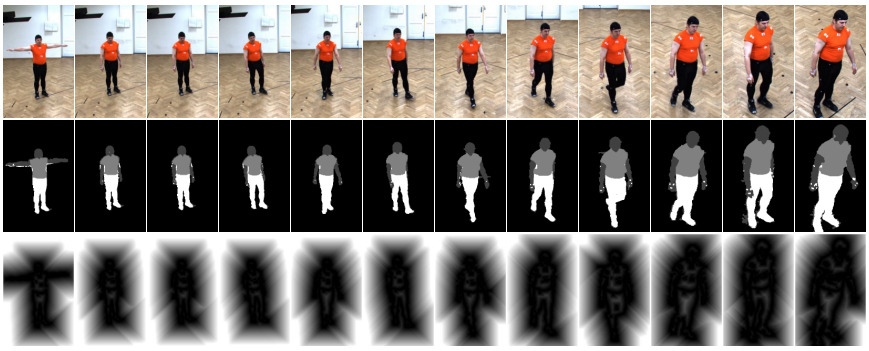

Fig. 1. Input images (upper row), segmented body parts (middle row), edge distance map (bottom row).

3.3 Hierarchical Optimization and Objective Function

In hierarchical fitting the pose of the body parts, which are the most predictable should be estimated first. Therefore, at the beginning of each time step we estimate the position of the torso. The location is determined with regard to the torso area delineated in the image. During determining the orientation of the torso we take into account the direction of motion of the walking person. With the help of camera calibration we determine the contact point with the floor for both legs or single leg. Given such contact point(s), we determine the pose of the legs. Finally, using the segmented forearms and/or arms we estimate the pose of both hands. In hierarchical PSO we used the following fitness function: $f(x) = o_z(x)^{\alpha_1} \times e_z(x)^{1-\alpha_1} \mid z \in \{\text{torso,legs,skin}\}$, where o_z denotes the silhouette overlap term, whereas e_z stands for the edge distance-based fitness. In ordinary PSO we utilized the following fitness function: $f(x) = o(x)^{\alpha_1} \times e(x)^{1-\alpha_1}$, where $o(x) = \alpha T(x) + \beta L(x) + \gamma S(x)$, where $\alpha + \beta + \gamma = 1$ and $T(x), L(x), S(x)$ stand for silhouette overlap term for torso, legs and skin, respectively.

4 Experimental Results

The PSO-based algorithms for full body motion tracking were compared by analyses carried out both through qualitative visual evaluations as well as quantitatively through the use of the motion capture data as ground truth. A GigE vision camera was used to acquire the color images of size 1920×1080 at 25 fps. Human motion tracking was performed on the cropped images with spatial resolution 740×800 pixels, see Fig. 2. An average silhouette height was approximately 300 pixels and varied from 250 pixels to 425 pixels. The swarm was initialized around the default initial pose, see the most left image at Fig. 1.

Fig. 2. Scene view with overlaid person, shot in frames #64, 128, 160,192 and 240.

At Fig. 3 are shown some tracking results that were obtained by ordinary and hierarchical PSO. The overlap of the projected 3D model on the subject undergoing tracking is shown to illustrate the quality of tracking. In the experiments presented below we focused on analyses of motion of walking people with bared

and freely swinging arms. The analysis of the human way of walking, termed gait analysis, has attracted considerable attention in recent years and can be utilized in several applications ranging from medical applications to surveillance.

Fig. 3. Full body motion tracking in monocular images.

We evaluated the accuracy of the PSO-based algorithm for motion tracking on a number of image sequences of a walking person taken from a fixed viewpoint. In Tab. 1 are depicted some quantitative results, which are averages over ten runs of the motion tracker with unlike initializations. The results were obtained on image sequence consisting of 240 frames, see Fig. 3, in 40 iterations using PSO consisting of 512 particles, and a configuration for hierarchical PSO with 40 iterations, 102 particles for torso, 205 for legs, and 205 particles for hands.

Table 1. Average errors for $M = 39$ markers.

		full body	torso	left hand	right hand	left ankle	right ankle
PSO	avg. err [mm]	222.3	112.6	258.5	660.8	225.8	227.1
	std. dev. [mm]	83.9	40.1	122.8	150.7	95.4	153.0
HPSO	avg. err [mm]	167.8	110.1	242.4	223.8	228.8	239.4
	std. dev. [mm]	66.7	22.9	128.1	123.0	86.5	105.8

The results for the full body, see also first column in Tab. 1, were obtained for $M = 39$ markers. From the above set of markers, 4 markers were placed on the head, 7 markers on each arm, 12 on the legs, 5 on the torso and 4 markers were attached to the pelvis. Given such a placement of the markers on the human body and the estimated human pose, which has been calculated by our algorithm, the corresponding positions of virtual markers were determined and then utilized in calculating the average Euclidean distance between corresponding markers. The average Euclidean distance $\overline{d}_i$ for each marker i was calculated using real world locations $m_i \in R^3$ on the basis of the following equation:

$$\overline{d}_i = \frac{1}{T} \sum_{t=1}^{T} ||m_i(\hat{x}_t) - m_i(x_t)|| \tag{3}$$

where $m_i(\hat{x})$ stands for marker's position that was calculated using the estimated pose, $m_i(x)$ denotes the position, which has been determined using ground-truth, whereas T stands for the number of frames. The errors reported in columns 2-6 of Tab. 1 indicate the distance errors for single markers on the considered limbs. For each marker i the standard deviation σ_i was calculated as follows:

$$\sigma_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(||m_i(\hat{x}_t) - m_i(x_t)|| - \overline{d}_i\right)^2} \tag{4}$$

The standard deviation $\overline{\sigma}$ shown in Tab. 1 is the average over all markers. The errors were obtained in scenarios with walking person, see Fig. 3. As we can observe, the hierarchical PSO algorithm outperforms the PSO based tracker. The results shown in Tab. 1 demonstrate that in our scenario with walking person that was shot by a monocular HD camera, the Particle Swarm Optimization algorithm is capable of estimating the full body motion with promising accuracy. The 3D reconstruction of human motion in monocular walking sequences is reliable in almost the whole sequence. The errors of the left hand are slightly larger for the reason that it has undergone complete occlusion in considerable number of frames. The mean distance error for *Lee walk* sequence recorded at 30 fps and 20 fps, that was obtained in [5] is equal to 283.6 ± 113.0 and 299.0 ± 121.9, respectively. The error obtained by our method on our walking sequence is far smaller owing to person segmentation into individual body parts as well as taking into account the direction of walking and the points of floor contact. Since the full body pose is estimated hierarchically, a large distance error of the torso can lead to considerable distance error of the whole body. A demo illustrating full body pose tracking using single monocular camera is available at: `http://prz.edu.pl/~bkwolek/res/icaisc12/sv_hmt.avi`.

The complete human motion capture system was written in C/C++. The system runs on Windows in both 32 bit and 64 bit modes. The entire tracking process takes approximately 7 sec. per frame on a PC with dual CPU Intel Xeon X5690 3.46 GHz using a configuration with 512 particles and 40 iterations for PSO and a configuration for hierarchical PSO with 40 iterations, 102 particles for torso, 205 for legs, 205 for hands. The image processing and analysis takes about 0.45 sec. Although the customization of the model can be completed automatically, the model is adjusted manually for each person to be tracked.

5 Conclusions

In this paper, we have shown that a successful full body motion tracking in monocular image sequences can be achieved using Particle Swarm Optimization and reliable segmentation of person into body parts. To show the advantages of the hierarchical PSO algorithm, we have conducted several experiments on sequences with a walking individual. The ordinary and hierarchical PSO algorithms were compared by analyses carried out both through qualitative visual evaluations as well as quantitatively through the use of the motion capture data as ground truth.

Acknowledgment

This paper has been supported by the research project OR00002111: "Application of video surveillance systems to person and behavior identification and threat detection, using biometrics and inference of 3D human model from video."

References

1. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. 28, 44–58 (2006)
2. Arsic, D., Lyutskanov, A., Rigoll, G., Kwolek, B.: Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. In: IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance. pp. 30–37. IEEE Press, Piscataway, NJ (2009)
3. Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for bayesian filtering. Statistics and Computing 10(1), 197–208 (2000)
4. Gavrila, D.M., Davis, L.S.: 3-D model-based tracking of humans in action: a multi-view approach. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR '96). pp. 73–80. IEEE Computer Society, Washington, DC, USA (1996)
5. John, V., Trucco, E., Ivekovic, S.: Markerless human articulated tracking using hierarchical Particle Swarm Optimisation. Image Vis. Comput. 28, 1530–1547 (2010)
6. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proc. of IEEE Int. Conf. on Neural Networks. pp. 1942–1948. IEEE Press, Piscataway, NJ (1995)
7. Kovac, J., Peer, P., Solina, F.: Human skin color clustering for face detection. In: Int. Conf. on Computer as a Tool. EUROCON 2003. vol. 2, pp. 144–148 (2003)
8. Krzeszowski, T., Kwolek, B., Wojciechowski, K.: GPU-accelerated tracking of the motion of 3D articulated figure. In: Computer Vision and Graphics, Lecture Notes in Computer Science, vol. 6374, pp. 155–162. Springer Berlin / Heidelberg (2010)
9. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2-3), 90–126 (2006)
10. Mori, G., Ren, X., Efros, A.A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: Proc. of the Conf. on Computer Vision and Pattern Recognition - vol. 2. pp. 326–333. IEEE Comp. Society (2004)
11. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A.: Challenges of human behavior understanding. In: Proc. of the First Int. Conf. on Human Behavior Understanding. pp. 1–12. HBU'10, Springer-Verlag, Berlin, Heidelberg (2010)
12. Schmidt, J., Fritsch, 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. pp. 567–572. IEEE Computer Society Press (2006)
13. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: Proc. of IEEE Int. Conf. on Computer Vision - Volume 2. pp. 750–757. ICCV '03, IEEE Computer Society, Washington, DC, USA (2003)
14. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3D human tracking. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. pp. 69–76. CVPR'03, IEEE Computer Society (2003)
15. Zhang, X., Hu, W., Wang, X., Kong, Y., Xie, N., Wang, H., Ling, H., Maybank, S.: A swarm intelligence based searching strategy for articulated 3D human body tracking. In: IEEE Workshop on 3D Information Extraction for Video Analysis and Mining in conjuction with CVPR. pp. 45–50. IEEE Press, Piscataway, NJ (2010)