Model-based 3D human motion capture using global-local particle swarm optimizations

Tomasz Krzeszowski, Bogdan Kwolek, and Konrad Wojciechowski

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

Summary. We present an approach for tracking the articulated motion of humans using image sequences obtained from multiple calibrated cameras. A 3D human body model composed of eleven segments that allows both rotation at joints and translation is utilized to estimate the pose. We assume that the initial pose estimate is available. A modified swarm intelligence based searching scheme is utilized to perform motion tracking. At the beginning of each optimization cycle, we estimate the pose of the whole body and then we refine locally the limb poses using smaller number of particles. The results that were achieved in our experiments are compared with those produced by other state-of-the-art methods, with analyses carried out both through qualitative visual evaluations as well as quantitatively by the use of the motion capture data as ground truth. They indicate that our method outperforms the algorithm based on the ordinary particle swarm optimization.

1 Introduction

Vision-based tracking of human bodies is an important problem due to various potential applications like recognition and understanding human activities, user friendly interfaces, surveillance, clinical analysis and sport (biomechanics). The goal of body tracking is to estimate the joint angles of the human body at any time. This is one of the most challenging problems in computer vision and pattern recognition because of self-occlusions, high dimensional search space and high variability in human appearance. An articulated human body can be thought of as a kinematic chain with at least 11 body parts. This may involve around 26 parameters to describe the full body articulation. By building a mapping from configuration space to observation space, 3D model-based approaches rely on searching the pose space to find the body configuration that best-matches the current observations [1]. Matching such complex and self-occluding model to human silhouette might be especially difficult in cluttered scenes. The major problems with 3D body tracking arise by reason of occlusions and depth ambiguities. Multiple cameras and simplified

backgrounds are often employed to ameliorate some of such practical difficulties. However, the use of multiple cameras is connected with difficulties such as camera calibration and synchronization, as well as correspondence.

The particle filtering is widely used in human motion tracking [2] owing to ability of dealing with high-dimensional probability distributions. Given the number of parameters needed to describe an articulated model, the number of particles of that is required to adequately approximate the underlying probability distribution in the pose space might be huge. Recent work demonstrates that particle swarm optimization (PSO) can produce similar or even superior results over particle filtering due to capability exploration of the high dimensional search space [3].

Markerless human motion capture has been studied since more than twenty years and is still a very active research area in computer vision and recognition. The advantage of markerless technique is that it eliminates the need for the specialized equipment as well as time needed to attach the markers. Complete survey of markerless human motion capture can be found in [4]. Despite huge research efforts in this area, there have only been a few successful attempts to simultaneously capture video and 3D motion data serving as ground truth for markerless motion tracking [2].

In this paper we present motion tracing results, which were obtained by a modified particle swarm optimization algorithm, together with analyses carried out both through qualitative visual evaluations as well as quantitatively by the use of the motion capture data as ground truth. The human body motion is modeled by a kinematic chain describing the movement of the torso/head, and both the arms as well as legs/feet. The tracking is done by particle swarm optimization with an objective function, which is built not only on cues like shape and edges but also on the segmented body parts. A global-local particle swarm optimization algorithm permits improved exploration of the search space and leads to better tracking, particularly regarding undesirable inconsistency in tracking legs and arms, resulting in swaps of such body parts, for instance, matching the right leg of the model to the left person's leg.

2 The algorithm

2.1 Tracking framework

The articulated model of the human body is built on kinematic chain with 11 segments. Such a 3D model consists of cuboids that model the pelvis, torso/head, upper and lower arm and legs. Its configuration 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. Each cuboid can be projected into 2D image plane via perspective projection. 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.

2.2 Person segmentation

In most of the approaches to articulated object tracking, background subtraction algorithms are employed to extract a person undergoing tracking [5]. Additionally, image cues such as edges, ridges, color are often employed to improve the extraction of the person [6]. In this work we additionally perform the segmentation of the person's shape into individual body parts. In our approach, the background subtraction algorithm [7] is used to extract the person and to store it in a binary image. We model the skin color using 16×16 histogram in rg color space. The patches of skin color are determined through histogram back-projection. The skin areas are then refined using the binary image as mask. Given the height of the extracted person we perform a rough segmentation of the legs and feet.

2.3 PSO-based motion tracking

Particle swarm optimization [8] is a global optimization, population-based evolutionary algorithm for dealing with problems in which a best solution can be represented as a point in n-dimensional space. The PSO is initialized with a group of random particles (hypothetical solutions) and then it searches hyperspace (i.e. R^n) of a problem for optima. Particles move through the solution space, and undergo evaluation according to some fitness function. Much of the success of PSO algorithms comes from the fact that individual particles have tendency to diverge from the best known position in any given iteration, enabling them to ignore local optima while the swarm as a whole gravitates towards the global extremum. If the optimization problem is dynamic, the aim is no more to seek the extrema, but to follow their progression through the space as closely as possible. Since the object tracking process is a dynamic optimization problem, the tracking can be achieved through incorporating the temporal continuity information into the traditional PSO algorithm. This means, that the tracking can be accomplished by a sequence of static PSO-based optimizations to calculate the best object location, followed by re-diversification of the particles to cover the possible object state in the next time step. In the simplest case, the re-diversification of the particle i can be realized as follows:

$$x_t^{(i)} \leftarrow \mathcal{N}(\hat{x}_{t-1}, \Sigma) \tag{1}$$

In the algorithm that we call global-local particle swarm optimization (GLPSO), at the beginning of each frame the estimation of the whole body pose takes place, see stage b) in Fig. 1. In the mentioned figure, the rectangular blocks represent state vectors with distinguished state variables of torso/head, arms, and legs/feet. To calculate such an estimate we initialize the particles using the estimated state $\hat{x}_{t-1}$ in time $t-1$, see stage a) in Fig. 1. In this stage the half of the particles is perturbed according to (1). In the remaining

part of the swarm, the state variables describing the location of the pelvis are initialized using the linear motion, perturbed by normally distributed random motion with zero mean, whereas the remaining state variables are initialized using only normally distributed random motion with the mean equal to the estimated state. Given the pose of the whole body, we construct state vectors consisting of the estimated state variables for torso/head and arms and state variables of the legs, which are constructed on the basis of the estimates of the corresponding state variables in time $t-1$, see Fig. 1 and the arrow connecting the state variables in time $t-1$ and time t. At this stage the discussed state variables are perturbed by normally distributed motion. Afterwards we execute particle swarm optimization in order to calculate the refined legs estimate, see stage d) as well as the right image in Fig. 1. Such refined state variables are then placed in the state vector, see stage e) in Fig. 1. The state variables describing the hands are refined analogously.

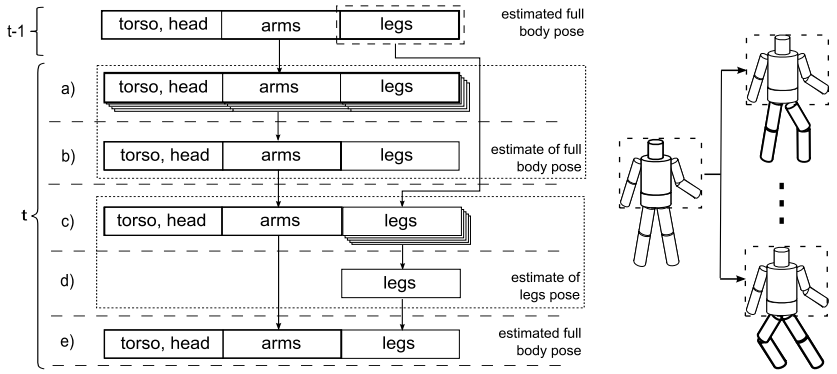

Fig. 1. Global-local particle swarm optimization for motion tracking (left), refinement of the legs configuration (right)

The fitness function of the PSO is determined on the basis of the following expression: $f(x) = w_1 f_1(x) + w_2 f_2(x) + w_3 f_3(x)$, where w_i stands for weighting coefficients that were determined experimentally. The function $f_1(x)$ reflects the degree of overlap between the segmented body parts and the projected segments of the model into 2D image. It is expressed as the sum of two components. The first component is the overlap between the binary image with distinguished body parts and the considered rasterized image of the model. The second component is the overlap between the rasterized image and the binary one. The larger the degree of overlap is, the larger is the fitness value. The function $f_2(x)$ reflects the degree of overlap of the edges. The last term, which reflects the overlap between arms and skin areas is not taken into account in the refinement of the pose of the legs.

3 Experiments

The algorithm has been tested in a multi-camera system consisting of four synchronized and calibrated cameras. The placement of the video cameras in our laboratory can be seen in Fig. 2. The cameras acquire images of size 1920×1080 with rate 24 fps. Ground truth motion of the body is provided by a commercial motion capture (MoCap) system from Vicon Nexus at rate 100 Hz. The system uses reflective markers and ten cameras to recover the 3D position of such markers. The synchronization between the MoCap and multi-camera system is done through hardware from Vicon Giganet Lab.

The tracking performance of our motion tracking algorithm was evaluated experimentally in a scenario with a walking person, see Fig. 2. Although we focused on tracking of torso and legs, we also estimated the pose of both arms as well as of the head. The body pose is described by position and orientation of the pelvis in the global coordinate system as well as relative angles between the connected limbs. Figure 2 depicts the projected model and overlaid on the input images from four cameras. It is worth noting that the analysis of the human way of walking (gait) can be employed in various applications ranging from medical applications to surveillance. Gait analysis is currently an active research topic.

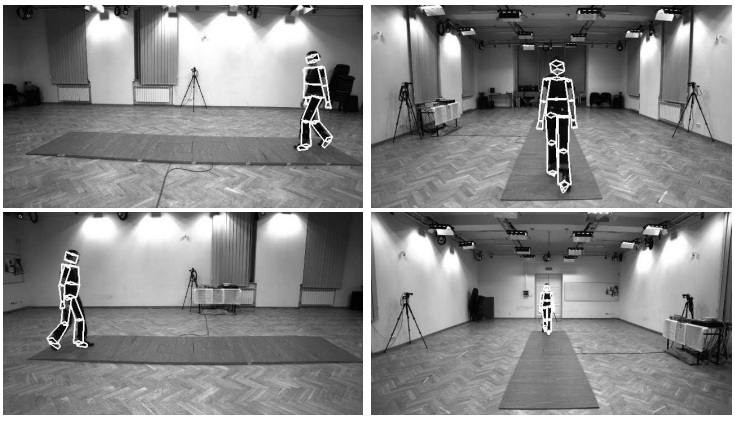

Fig. 2. Human motion tracking. Frame #20 in view 1 and 2 (upper row), in view 3 and 4 (bottom row)

Figure 3 illustrates some tracking results, which were obtained in one of the experiments. They were obtained using the global-local particle swarm optimization. The tracking of the whole body was done using 300 particles. In the pose refinement stage, the whole body is tracked using 200 particles, whereas 2×50 particles were used for tracking the legs as well as the arms.

The tracking was achieved using 20 iterations in each stage mentioned above. The number of frames in the discussed sequence is equal to 180.

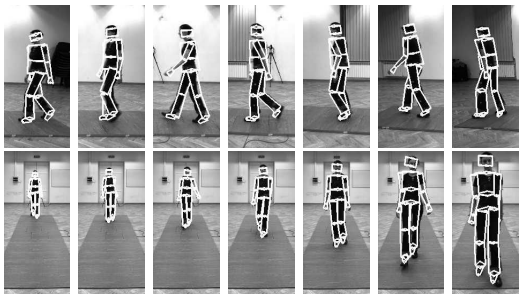

Fig. 3. Tracking results in a sequence 1. Frames #20, 40, 60, 80, 100, 120, 140 in view 1 (1st row) and in view 4 (2nd row)

Figure 4 depicts some experimental results, which were achieved in another test sequence. The results were obtained by global-local PSO algorithm using the same number of the particles and iterations. The tracking was done on 150 images.

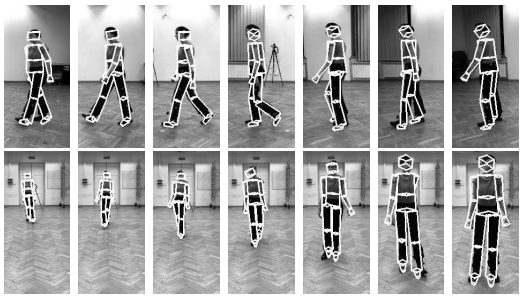

Fig. 4. Tracking results in a sequence 2. Frames #20, 40, 60, 80, 100, 120, 140 in view 1 (1st row) and in view 4 (2nd row)

The average Euclidean distance $\overline{d}_i$ for each marker i was calculated using real world locations $m_i \in R^3$. It was calculated on the basis of the following expression:

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

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

data from our motion capture system, whereas T stands for the number of frames. In Tab. 1 are shown the average distance errors $\overline{d}$ for $M = 39$ markers. For each marker i the standard deviation σ_i was calculated on the basis of the following equation:

$$\sigma_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(||m_i(\hat{x}_t) - m_i(x_t)|| - \overline{d}_i\right)^2} \tag{3}$$

The standard deviation $\overline{\sigma}$ shown in Tab. 1 is the average over all markers. From the above set of markers, four markers were placed on the head, seven markers on each arm, 6 on the legs, 5 on the torso and 4 markers were attached to the pelvis. Given the estimated human pose and such a placement of the markers on the human body, the corresponding positions of virtual markers were calculated and then utilized in calculating the average Euclidean distance given by (2). The errors that are shown in Tab. 1 were obtained using frame sequences, which are depicted in Fig. 3 and Fig. 4. As we can observe, our GLPSO algorithm outperforms significantly the tracker that is based on the ordinary PSO.

Table 1. Average errors and standard deviations of the whole body tracking

			Seq. 1		Seq. 2	
	#particles	#it.	error $\overline{d}$ [mm]	$\overline{\sigma}$ [mm]	error $\overline{d}$ [mm]	$\overline{\sigma}$ [mm]
PSO	100	10	79.41	50.19	82.25	53.48
	200	10	78.97	50.20	77.97	48.82
	300	10	74.54	46.24	77.84	50.31
	100	20	78.46	50.10	80.33	47.99
	200	20	72.32	44.28	77.29	49.17
	300	20	70.28	41.69	73.86	45.58
GLPSO	100	10	71.19	33.88	73.88	35.58
	200	10	67.17	27.83	66.36	25.58
	300	10	64.14	25.74	64.52	23.86
	100	20	68.03	28.31	67.01	27.72
	200	20	66.45	28.35	64.02	23.19
	300	20	62.46	23.77	62.59	22.45

For fairness, in all experiments we use the identical particle configurations. For the global-local PSO the sum of particles responsible for tracking the whole body, arms and legs corresponds to the number of the particles in the PSO. For instance, the use of 300 particles in PSO is equivalent to the use of 200 particles for tracking the full body, 50 particles for tracking the arms and 50 particles for tracking both legs in GLPSO. The use of 200 particles in the

PSO corresponds to the exploitation of 100, 50 and 50 particles, respectively, whereas the use of 100 particles equals to utilization 60 particles for tracking the global configuration of the body, along with 20 and 20 particles for tracking hands and legs, respectively.

Table 2 shows the average errors and standard deviations of some markers, which are located on lower body limbs, i.e. the right knee and the right tibia. The discussed results were obtained on images from the sequence 1. As we can see, the difference between errors obtained by GLPSO and PSO algorithms is more significant in comparison to the results shown in Tab. 1. The difference is larger because in experiments, whose results are illustrated in Tab. 1 we considered several markers on torso as well as pelvis, which typically produce small errors, see also Fig. 5, and in consequence contribute towards more smoothed results.

Table 2. Average errors and standard deviations of lower body tracking

			PSO		GLPSO	
	#particles	#it.	error $\overline{d}$ [mm]	$\overline{\sigma}$ [mm]	error $\overline{d}$ [mm]	$\overline{\sigma}$ [mm]
right knee	100	20	68.08	59.88	50.13	33.18
	300	20	63.01	56.87	44.84	27.17
right tibia	100	20	95.57	84.72	70.40	44.12
	300	20	93.25	77.91	72.39	36.44

In Fig. 5 we can observe a plot of the distance error over time for some body limbs. As we can observe, in some frames the PSO based algorithm produces considerably larger errors. The algorithm based on global-local PSO allows us to achieve superior results thanks to the decomposition of the search space.

In the last few years, there have only been a few successful attempts to simultaneously capture video and 3D motion data from the MoCap [2]. Hence, there are only few papers, which show the motion tracking accuracy by both qualitative visual analyses as well as quantitatively by the use of the motion capture data as ground truth. To our knowledge, the tracking accuracy presented above is as good as the accuracy presented in [2] and in other state-of-the-art systems. However, instead of the annealed particle filter we utilize global-local particle swarm optimization. Moreover, our algorithm performs the segmentation of the person's shape into individual body parts. The global-local particle swarm optimization algorithm allows better exploration of the search space as well as allows us obtaining better tracking, particularly regarding undesirable swaps of legs and arms. The segmentation of the body into individual body parts contributes towards improved tracking, mainly due to better matching the individual parts of the model to corresponding body parts, extracted on the image.

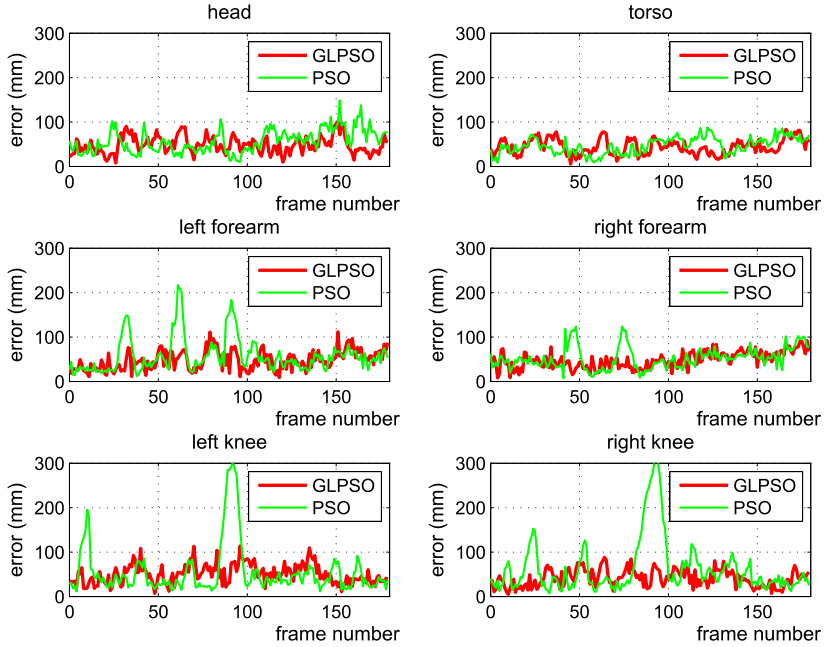

Fig. 5. The errors [mm] versus frame number

The complete human motion tracking system was written in C/C++. Currently, it requires an initial estimate of the body pose. The system operates on color images with spatial resolution of 960×540 pixels. The entire tracking process can be realized in approximately 1.1 sec. on a Intel Core i5 2.8 GHz using Open Multi-Processing (OpenMP), see Tab. 3. The image processing and analysis takes about 1 sec.

Table 3. Computation time on Intel Core i5 2.8 GHz

		time [sec.]	
#particles	#it.	1 core	4 cores
100	10	1.9	1.1
200	10	3.6	2.3
300	10	5.7	3.1
100	20	3.6	2.1
200	20	7.0	4.5
300	20	11.0	6.0

4 Conclusions

In this paper we presented an algorithm for tracking human pose using multiple calibrated cameras. The tracing is achieved by particle swarm optimization using both motion and shape cues. At the beginning the algorithm extracts person's silhouette and afterwards it segments the shape into individual body parts. A particle swarm optimization employs an objective function built on edges and such segmented body parts. We evaluated the algorithm through both qualitative visual analyses and quantitatively by the use of the motion capture data as ground truth. Experimental results show that algorithm achieves the tracking accuracy that is comparable to the accuracy produced by other state-of-art methods.

Acknowledgment

This work has been supported by the project "System with a library of modules for advanced analysis and an interactive synthesis of human motion" cofinanced by the European Regional Development Fund under the Innovative Economy Operational Programme - Priority Axis 1. Research and development of modern technologies, measure 1.3.1 Development projects.

References

1. Sidenbladh, H., Black, M., and Fleet, D. (2000) Stochastic tracking of 3D human figures using 2d image motion. *European Conf. on Computer Vision*, pp. 702–718.
2. Sigal, L., Balan, A., and Black, M. (2010) HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. Journal of Computer Vision*, **87**, 4–27.
3. Krzeszowski, T., Kwolek, B., and Wojciechowski, K. (2010) Articulated body motion tracking by combined particle swarm optimization and particle filtering. *Int. Conf. on Computer Vision and Graphics, Lecture Notes in Computer Science*, pp. 147–154, Springer, vol. 6374.
4. Poppe, R. (2007) Vision-based human motion analysis: An overview. *Comp. Vision and Image Understanding*, **108**, 4–18.
5. Sminchisescu, C., Kanaujia, A., Li, Z., and Metaxas, D. (2005) Discriminative density propagation for 3D human motion estimation. *In IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. I:390–397.
6. Schmidt, J., Fritsch, J., and Kwolek, B. (2006) Kernel particle filter for real-time 3D body tracking in monocular color images. *IEEE Int. Conf. on Face and Gesture Rec., Southampton, UK*, pp. 567–572, IEEE Computer Society Press.
7. Arsic, D., Lyutskanov, A., Rigoll, G., and Kwolek, B. (2009) Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 30–37, IEEE Press, Piscataway, NJ.
8. Kennedy, J. and Eberhart, R. (1995) Particle swarm optimization. *Proc. of IEEE Int. Conf. on Neural Networks*, pp. 1942–1948, IEEE Press, Piscataway, NJ.