An Approach for Model-based 3D Human Pose Tracking, Animation and Evaluation

Tomasz Krzeszowski and Bogdan Kwolek

Rzeszów University of Technology
W. Pola 2, 35-959 Rzeszów, Poland
{tkrzeszo, bkwolek}@prz.edu.pl

Summary. This work presents an approach for 3D human pose tracking, animation and evaluation. The tracking of the full body is done using a modified particle swarm optimization and two synchronized cameras. On the basis of the 3D pose estimates we generate animated human motion. The animated images are processed in the same way as videos taken from the CCD cameras. This way we obtained ground-truth and utilized it in evaluations of the motion tracker.

1 Introduction

3D articulated human body tracking consists in determining the location of the person and orientation of each body part. The ability to track 3D pose is an important one, not least in the areas of visual surveillance, clinical analysis and sport (biomechanics). Tracking human body articulation is a difficult task because of high dimensionality of the state space. Another reason that makes this task difficult is the problem of self-occlusion, where body parts occlude each other depending on the body configuration. In vision based-pose tracking, inherent ambiguity arises as 3D pose parameters are estimated on the basis of 2D image features. Typically, multiple cameras are used to reduce ambiguities in a single view caused by occlusion.

Typically, tracking involves searching for the current pose using the estimate from the previous frame guided by a locomotion model. Particle filters [1] are commonly employed in 3D pose tracking as they can maintain multiple hypotheses. However, in ordinary particle filtering the number of samples needed for accurate tracking increases exponentially with the number of parameters. Moreover, even for low dimensional spaces it can be observed a tendency of particles to become concentrated in a single mode of the probability distribution and thus the tracking performance depends considerably on the quality of the importance sampler. As human body contains no less than 10 body parts, equating to more than 20 degrees-of-freedom (DOF), the number of particles in an ordinary particle filter might be huge. In such

spaces, sample impoverishment may prevent the particle filter from maintaining multimodal probability distributions over long periods of time. Therefore, considerable research was done in order to develop methods with improved concentration of particles near true body poses. Deutscher *et al.* [2] proposed an annealed particle filter, which employs the annealing process to gradually move the particles towards the global maximum.

Gavrila and Davis [3] utilize an explicit hierarchical search to subsequently locate parts of the three-based kinematic model, reducing the search complexity. In the discussed approach, the torso was localized using color cues. However, in practice, it is not easy to localize the torso and hence to provide a good starting guess for the search. Furthermore, imprecision in the localization of the torso, among others due to occlusion, can easily lead to unrecoverable failure.

Recently, particle swarm optimization (PSO) [4], a population based stochastic optimization technique has gained considerable interest in the field of full-body articulated tracking [5][6]. Unlike the independent samples in the particle filter, the simple agents in the PSO interact with one another and with their environment in the course of searching for the best solution. Although there is no centralized control, such interactions between agents lead to the intelligent global behavior, unknown to the individual agents, which in turn results in more effective exploration of the high-dimensional search space.

In this paper we discuss a cascaded algorithm for 3D pose tracking, which is based on particle swarm optimization. In the first step, it determines the pose of the whole body using reduced number of particles. Afterwards, given the location of the torso that was determined in such a way we perform the rediversification of the particles in the part of the vector state that describes the pose of the legs. The rediversification is done on the basis of the pose of legs determined in the global stage. That means that for the best pose of the torso, which was determined in advance, we generate several hypothesized configurations of the legs. Finally, we carry out optimization using only the part of the state vector that describes pose of the legs. At this stage, in the objective function we consider only legs. In a similar manner we determine the pose of the hands. We present the experimental results that were obtained using two synchronized and calibrated cameras, overlooking the same scene. Our attention was restricted to walking motions. Using the estimated 3D poses, a computer animation of human walking has been done. For each 3D pose, such a virtual human has been overlaid on the background image. The images were then processed in the same way as videos taken from the CCD cameras. This way, for a given pose, which can be perceived as ground-truth, we got animated human. On the basis of such images we estimated 3D poses and performed qualitative evaluations of the tracking algorithm.

2 The algorithm

2.1 Tracking algorithm

Particle swarm optimization [4] 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) [7], at the beginning of each frame the estimation of the whole body pose takes place. In the first step, it determines the pose of the whole body using reduced number of particles. Afterwards, given the location of the torso that was determined in such a way we perform the rediversification of the particles in the part of the vector state that describes the pose of the legs. The rediversification is done on the basis of the best pose of legs determined in the global stage. That means that for the pose of the torso, which was determined in advance, we generate several hypothesized configurations of the legs. Finally, we carry out optimization using only the part of the state vector that describes the pose of the legs. At this stage, in the objective function we consider only legs. In a similar manner we determine the pose of the hands.

2.2 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 joint angles. 3D projection is used in mapping the model onto 2D image plane. 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.

2.3 Fitting cost

In the PSO each particle represents the hypothesized 3D pose. The fitness score reflects how well the projection of a given body pose fits the observed images. The person's silhouette is typically delineated by background subtraction. It is then used to calculate silhouette-overlap term. In addition, image cues such as edges, ridges, color, optical flow are often utilized. However, most common algorithms rely on silhouettes and edges. The most common type of edge detection process uses a gradient operator.

Figure 1 depicts input images and corresponding foreground images in lateral and frontal view. The background images were extracted using algorithm that has been proposed in [8].

Fig. 1. Input image (view 1), foreground image, input image (view 2), foreground image.

Figure 2 depicts the subsequent stages of distance map extraction, which serves as edge-proximity term. The distance transform assigns each pixel a value that is the distance between that pixel and the nearest nonzero edge pixel. The dilated binary image, see Fig. 2b, was employed to extract background-subtracted edge image, shown at Fig. 2d, which was utilized in extraction of the distance map. The projected line segments of the 3D model are aligned with such a distance map.

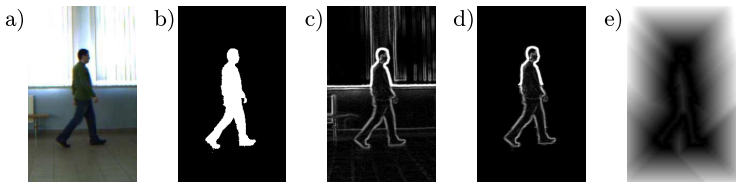

Fig. 2. Input image a), foreground image b), gradient magnitude c), masked gradient magnitude d), edge distance map e).

3 Experimental results

The tracking performance of the algorithm has been evaluated experimentally in a scenario with a walking person. While we restricted our focus of attention to tracking person's torso and legs, the 3D pose of both arms as well as of the head has also been estimated. The images were captured by two synchronized cameras that are oriented perpendicular to each other.

Figure 3 depicts some experimental results that were achieved using the discussed above camera setup. Image a) depicts initial 3D pose seen from lateral view, whereas the image c) illustrates the 3D pose seen from the frontal view. Figure 3b) (images left-to-right, top-to-bottom) presents the 3D tracking poses, that are overlaid on the images seen from the camera's lateral view, whereas Fig. 3d) illustrates the model overlaid on the images from the frontal view. The results were achieved by GLPSO in 20 iterations and using 200 particles. As we can observe the algorithm is capable of estimating the 3D pose of a walking person. Thanks to the use of two cameras the occlusions are handled quite well. Overall, similar tracking results were observed for the other loosely dressed individuals.

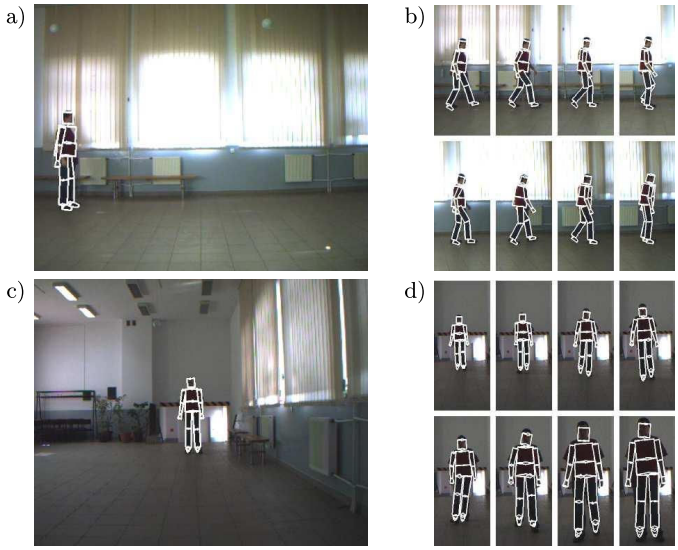

Fig. 3. Input image with the overlaid model, lateral-view a), frontal-view c). Tracking in frames #15, 30, 45, 60, 75, 90, 105, 120 (left-to-right, top-to-bottom), lateral-view b), frontal-view d).

The 3D pose estimates were recorded in the BVH files to perform skeletal animation. The BVH format is utilized as a standard representation of movements in the animation of the humanoid structures. Skeletal animation

is a technique used in 3D rendering, which employs an exterior shell (called skin or mesh) composed of vertices representing the object surface and an internal skeleton for the animation. The skeleton consists of hierarchical set of interconnected bones. Each of them has a three dimensional transformation, which determines its position, scale and orientation, and an optional parent bone. Every bone in such a skeleton is coupled with some portion of the character's visual representation. In the skeletal animation the skeleton is used as a controlling mechanism to deform attached mesh data via so called skinning.

Figure 4 depicts some images from a sequence, which have been generated using 3ds Max. The animation is done using 3D pose estimates that were obtained in the 3D pose tracking and stored in BVH files. The images were then processed in the same way as videos taken from the CCD cameras. This way, for a given pose, which can be perceived as ground-truth, we got animated human. On the basis of such images we estimated 3D poses and performed qualitative evaluations of the tracking algorithm.

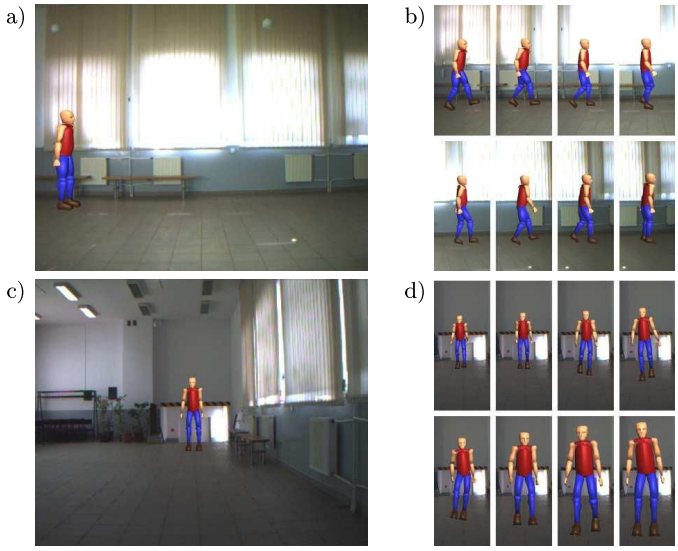

Fig. 4. Input image with the overlaid virtual human, lateral-view a), frontal-view c). Tracking in frames #15, 30, 45, 60, 75, 90, 105, 120 (left-to-right, top-to-bottom), lateral-view b), frontal-view d).

In Tab. 1 are shown the average errors, which were obtained for $M = 32$ markers. 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 as:

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

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 ground-truth, whereas T stands for the number of frames. 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, 4 markers were placed on the head, 7 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 used in calculating the average Euclidean distance (2). The errors that are shown in Tab. 1 are averages of 10 runs of the algorithm and were obtained using frame sequences, which are shown in Fig. 4. As we can observe, the GLPSO algorithm outperforms the PSO based tracker.

For fairness, in all experiments we use the equivalent particle number. 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 200 particles in the PSO corresponds to the exploitation of 150, 25 and 25 particles, respectively, whereas the use of 100 particles equals to use of 80 particles for tracking the global configuration of the body, along with 10 and 10 particles for tracking hands and legs, respectively.

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

	#particles	#it.	error [mm]	$\overline{\sigma}$ [mm]
PSO	100	10	87.14	61.21
	100	20	82.37	60.64
	200	10	84.48	58.06
	200	20	79.50	59.56
GLPSO	100	10	74.18	48.72
	100	20	69.37	48.56
	200	10	70.16	46.56
	200	20	65.23	45.95

The 3D pose tracking algorithm was written in C/C++. The experiments were done using images that were recorded at 15 Hz. They were acquired by two synchronized cameras, overlooking the same scene. The evaluation of the algorithm was done on a desktop PC with 4 GB RAM, Intel Core i5, 2.8 GHz. The system operates on color images with spatial resolution of 640×512 pixels. The algorithm operates at $\sim$1 frame per second (100 particles, 10 it.). There

has also been implemented a parallel version of the algorithm using OpenMP threads, which was then executed on mentioned above multi-core processor. Through parallelization of the fitness function the 3D pose tracking was done at ∼1.5 fps. The initial body pose has been determined manually.

4 Conclusions

We presented an approach for 3D human pose tracking and evaluation. In experiments we tracked the 3D pose of a walking person. On the basis of 3D pose estimates we generated animated human motion. The images with animation were processed in the same way as videos taken from cameras. This way we obtained ground-truth and utilized it in evaluations of the algorithm.

Acknowledgement

This work has been partially supported by the National Science Centre (NCN) within the project N N516 483240.

References

1. Isard, M. and Blake, A. (2006) CONDENSATION - conditional density propagation for visual tracking. *Int. J. of Computer Vision*, **29**, 5–28.
2. Deutscher, J., Blake, A., and Reid, I. (2000) Articulated body motion capture by annealed particle filtering. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 126–133.
3. Gavrila, D. M. and Davis, L. S. (1996) 3-D model-based tracking of humans in action: a multi-view approach. *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 73–80, CVPR '96, IEEE Computer Society.
4. Kennedy, J. and Eberhart, R. (1995) Particle swarm optimization. *Proc. of IEEE Int. Conf. on Neural Networks*, pp. 1942–1948, IEEE Press, Piscataway, NJ.
5. Zhang, X., Hu, W., Wang, X., Kong, Y., Xie, N., Wang, H., Ling, H., and Maybank, S. (2010) A swarm intelligence based searching strategy for articulated 3D human body tracking. *IEEE Workshop on 3D Information Extraction for Video Analysis and Mining in conjuction with CVPR*, pp. 45–50.
6. Krzeszowski, T., Kwolek, B., and Wojciechowski, K. (2010) GPU-accelerated tracking of the motion of 3D articulated figure. *Int. Conf. on Comp. Vision and Graphics*, pp. 155–162, Lecture Notes in Computer Science, Springer, vol. 6374.
7. Krzeszowski, T., Kwolek, B., and Wojciechowski, K. (2011) Model-based 3D human motion capture using global-local particle swarm optimizations. *Computer Recognition Systems 4*, vol. 95 of *Advances in Intelligent and Soft Computing*, pp. 297–306, Springer Berlin / Heidelberg.
8. 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.