

TOWARDS A LOW COST MULTI-CAMERA MARKER BASED HUMAN MOTION CAPTURE SYSTEM

C. Canton-Ferrer, J.R. Casas, M. Pardàs

Image Processing Group, Technical University of Catalonia (Spain)

ABSTRACT

This paper presents a low cost real-time alternative to available commercial human motion capture systems. First, a set of distinguishable markers are placed on several human body landmarks and the scene is captured by a number of calibrated and synchronized cameras. In order to establish a physical relation among markers, a human body model (HBM) is defined. Markers are detected on all camera views and delivered as the input of an annealed particle filter scheme where every particle encodes an instance of the pose of the HBM to be estimated. Likelihood between particles and input data is performed through the generalized symmetric epipolar distance and kinematic constraints are enforced in the propagation step towards avoiding impossible poses. Tests over the HumanEva annotated dataset yield quantitative results showing the effectiveness of the proposed algorithm. Results over sequences involving fast and complex motions are also presented.

Index Terms— Human motion capture, multi-camera analysis, particle filtering

1. INTRODUCTION

Accurate retrieval of the configuration of an articulated structure from information provided by multiple cameras is a field that found numerous applications in the recent years. The grown of computer graphics technology together with motion capture systems have been extensively used by the cinema and video games industry to generate virtual avatars or fantastic characters. Medicine also benefited from these advances in the field of orthopedics, locomotive pathologies assessment or sports performance improvement. However, all these applications require accurate input data to work and, nowadays, only human motion capture (HMC) systems aided by markers placed on some body landmarks may produce the desired degree of accuracy.

Optical systems based on photogrammetric methods are more used than others requiring special suits embedding skeletal-like structures [1] or magnetic devices [2]. However, optical systems are usually expensive and require a dedicated hardware involving a high number of cameras (typically, more than 7) and/or a high frame rate (typically, 60-120 Hz) to produce an accurate output in form of a set of the 3D positions corresponding to the markers attached to the performer's body. The more usual are retroreflective markers that reflect back light that is generated near the cameras lens [3].

These systems require to reconstruct the 3D position of markers from its 2D projections taking into account occlusions and detection noise. Since errors occur when crucial markers become occluded or their trajectories are confused, temporal tracking is also employed. Finally, most applications require the transformation of the markers localization to the parameters of a HBM. Commercial tools that perform this transformation are generally semi-automatic, thus involving a labor-intensive and prone to errors task.

In many systems, the estimation of the markers' 3D position and the fitting of the HBM are decoupled. One of the first attempts to use an anatomical human model to enhance marker detection and trajectory tracking was presented in [4]. Another approach based on 3D marker clustering and topology connectivity analysis to fit a HBM was presented in [1] and, along the same line, [5] performed an optimization process over 3D markers to estimate the pose of the performer. Detection of 2D markers in separate images and its analysis using calibration information has been used in [6] enforcing a HBM afterwards. A similar technique using a Kalman filter involving the HBM in the data association step was presented in [7].

In this paper, a low cost real-time multi-camera algorithm for marker based human motion capture is presented. Marker detection and HBM pose estimation is performed in the same analysis loop by means of an annealed particle filter. Epipolar geometry is exploited in the particle likelihood evaluation by means of the symmetric epipolar distance being robust to noisy marker detections and false measurements. Kinematic restrictions are applied in the particle propagation step towards avoiding impossible poses. Finally, effectiveness of the proposed algorithm is assessed by means of objective metrics defined in the framework of HumanEva dataset [8] and metrics introduced in [9].

2. HBM BASED ANNEALED PARTICLE FILTERING

Let us define a state space $\mathcal{Y} \subset \mathbb{R}^D$ formed by the D defining parameters of an articulated HBM, in our case, the angles at the joints and the global translation and rotation of the model w.r.t. the real world, adding up to $D = 27$ (see an example in Fig.1). Estimating the optimal pose of this HBM at time t , $\hat{\mathcal{Y}}_t$, given a set of noisy observations $\mathbf{z}_{1:t}$ up to time t , involves computing a representation of the posterior likelihood $p(\mathcal{Y}|\mathbf{z}_{1:t})$, that usually exhibits a multimodal shape. Particle filtering (PF) [10] has been found suitable to tackle such problems but, due to the high dimensionality of the state space, the number of particles required to efficiently explore \mathcal{Y} turns out to be computationally unfeasible.

Annealed particle filtering (APF) [11] has been presented in the context of HMC as a technique to efficiently estimate $p(\mathcal{Y}|\mathbf{z}_{1:t})$ requiring far less particles than PF, hence allowing real-time implementations. APF introduces a layered posterior estimation where a set of N_p weighted particles $\{\mathcal{Y}_t^j \in \mathcal{Y}, w_t^j \in \mathbb{R}\}_{j=1}^{N_p}$ are evaluated and propagated through a set of N_L progressively smoothed versions of the likelihood function (also called annealing layers) thus avoiding getting trapped in local minima. Finally, once reaching the last annealing layer, pose $\hat{\mathcal{Y}}_t$ is computed as the weighted average of all particles. An example of the APF operation is depicted in Fig.1.

Following the standard APF algorithm, some factors are to be taken into account when implementing it: the measurement generation, the likelihood evaluation and the propagation model.

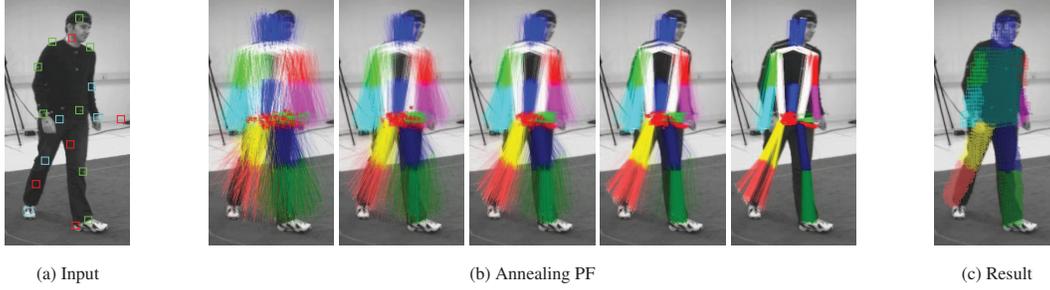


Fig. 1. APF operation example. In (a), the output of the employed marker detector where color boxes stand for correct (green), false (red) and missed (blue) detections. In (b), the progressive fitting of particles driven by the annealing process and, in (c), the final pose estimation $\hat{\mathcal{Y}}_t$.

2.1. Measurement generation

For a given frame in the video sequence, a set of N_C images are obtained from the N_C cameras. Each camera is modeled using a pinhole camera model based on perspective projection [12]. Accurate calibration information is available. The input data \mathbf{z}_t to our tracking system will be the 2D projection of the set of distinguishable markers attached to the body of the performer onto these N_C images. Let $\mathcal{D}_n = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{Q_n}\}$ be the set of Q_n locations detected in the image captured in the n -th view, \mathbf{I}_n , $1 < n \leq N_C$. Ideally, this set would contain the 2D projections of the markers that are not affected by the occlusions produced by the body itself onto the n -th camera view. In order to generate \mathcal{D}_n , a marker detection algorithm $\Gamma : \mathbf{I}_n \rightarrow \mathcal{D}_n$ is employed whose performance is assessed by the triplet: detection rate (\overline{DR}), the false positive rate (\overline{FP}) and the variance estimation error (σ_{Γ}^2). This generic formulation of Γ will allow performance comparisons of the tracking algorithm when using different marker detection algorithms.

Markers are usually placed at the joints, the end of the limbs, the top of the head and the chest of the subject. In this paper, some experiments were conducted using little yellow balls as body markers thus a color-based marker detection algorithm was employed to retrieve their 2D positions. However, the proposed method is general enough to be applied to any type of markers detectable onto a set of 2D planes under perspective projection. An example of the markers measurement delivered to the tracking algorithm is shown in Fig.1a.

2.2. Likelihood evaluation

In order to evaluate the likelihood between the body pose represented by a given particle state $\mathbf{y}_t^j \in \mathcal{Y}$ with reference to the input data $\mathbf{z}_t = \{\mathcal{D}_n\}_{n=1}^{N_C}$, a fitness function $w(\mathbf{z}_t, \mathbf{y}_t^j)$ should be defined.

The M 3D positions of the HBM landmarks (the joints and the end of the limbs) corresponding to the pose described by the state variable \mathbf{y} are computed using forward kinematics [6]. Let us denote these coordinates as the set $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$, $\mathbf{x}_m \in \mathbb{R}^3$. The fitness function relating the 3D locations set X with the 2D observations $\{\mathcal{D}_n\}_{n=1}^{N_C}$ should measure how well these 2D points fit as projections of the set X . We have tackled a similar problem in [13] in a Bayesian framework and the underlying idea is applied in this context. For every element \mathbf{x}_m from the set X , we compute its projection onto every camera as

$$\mathbf{p}_{m,n} = P_n(\mathbf{x}_m), \quad 1 \leq m \leq M, \quad 1 \leq n \leq N_C, \quad (1)$$

where $P_n(\cdot)$ is the perspective projection operator from 3D to 2D on the n -th view [12]. Then, the set $T_m = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{N_C}\}$ containing

the closest measurement in every camera view associated to every HBM landmark \mathbf{x}_m is constructed as follows:

$$\mathbf{t}_n = \min_{\mathbf{d}_q} \|\mathbf{p}_{m,n} - \mathbf{d}_q\|, \quad \mathbf{d}_q \in \mathcal{D}_n, \quad \forall n. \quad (2)$$

However, not all the 3D points \mathbf{x}_m may have a projection onto every view due to occlusions or a miss detection of the marker detection algorithm. In order to detect such cases, a thresholding is applied to the elements \mathbf{t}_n dismissing those measurements above a threshold ρ . In this case, $\mathbf{t}_n = \emptyset$. At this point, it is needed to measure how likely are the set of 2D measurements T_m to be projections of the 3D HBM landmark \mathbf{x}_m . This can be done by means of the generalized symmetric epipolar distance $d_{SE}(\cdot)$ presented in [13].

Let $l(\mathbf{x}^i, j)$ be the epipolar line generated by the point \mathbf{x} in a given view i onto another view j . Symmetric epipolar distance between two points $d_{SE}(\mathbf{x}^i, \mathbf{x}^j)$, in the two views i, j , is defined as:

$$d_{SE}(\mathbf{p}^i, \mathbf{p}^j) \triangleq \sqrt{d^2(l(\mathbf{x}^i, j), \mathbf{x}^j) + d^2(l(\mathbf{x}^j, i), \mathbf{x}^i)}, \quad (3)$$

where $d(l(\mathbf{x}^i, j), \mathbf{x}^j)$ is defined as the Euclidean distance between the epipolar line $l(\mathbf{x}^i, j)$ and the point \mathbf{x}^j as depicted in Fig.2. It has been shown in [13] that the extension of the symmetric epipolar distance for $k \geq 2$ points (in k different views) $d_{SE}(\mathbf{x}^1, \dots, \mathbf{x}^k)$ can be written in terms of the distance defined in Eq.3 as:

$$d_{SE}(\mathbf{x}^1, \dots, \mathbf{x}^k) = \sqrt{\sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} d_{SE}^2(\mathbf{x}^i, \mathbf{x}^j)}. \quad (4)$$

This distance produces low values when the 2D points are coherent, that is when they are projections from the same 3D location. The score s_m associated to T_m , and therefore to \mathbf{x}_m , is defined as:

$$s_m(\mathbf{z}_t, \mathbf{x}_m) \equiv s_m(\mathbf{z}_t, T_m) \propto d_{SE}(T_m), \quad (5)$$

and normalized such that $s_m(\mathbf{z}_t, T_m) \leq 1$. In the case where the non-empty elements of T_m is below 2, the distance $d_{SE}(T_m)$ can not be computed. Under these circumstances, we set $s_m(\mathbf{z}_t, T_m) = 1$.

Finally, the cost function $C(\mathbf{z}_t, \mathbf{y}_t^j)$ is constructed as the average of the distances over the M HBM 3D landmark points:

$$C(\mathbf{z}_t, \mathbf{y}_t^j) = \frac{1}{M} \sum_{m=1}^M s_m(\mathbf{z}_t, \mathbf{x}_m). \quad (6)$$

The associated weighting function is defined as:

$$w(\mathbf{z}_t, \mathbf{y}) = \exp\left(-\frac{C(\mathbf{z}_t, \mathbf{y}_t^j)^2}{2\sigma^2}\right). \quad (7)$$

In our experiments, $\rho = 10$ pixels and $\sigma = 1$ provided satisfactory results being ρ the most discriminative parameter driving the accuracy of the algorithm.

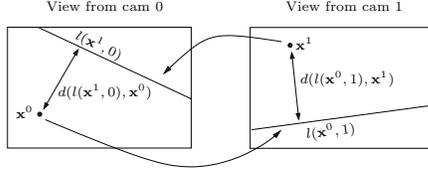


Fig. 2. Symmetric epipolar distance between two points $d_{SE}(x^0, x^1)$.

2.3. Propagation model

Kinematic restrictions imposed by the angular limits at each joint may produce a more robust tracking output. Employing a previously learnt motion model in the particle propagation step can improve tracking results if annotated data is available [14]. However, these methods are constrained to deal with motions present in the training corpus thus being not suitable for unconstrained motion tracking. In this paper, angular constraints are enforced in the propagation step of the APF scheme. Usually, the propagation step consists in adding a random component to the state vector of a particle as:

$$y_t^k = y_{t-1}^k + \mathcal{N}(\mathbf{0}, \Sigma) = \mathcal{N}(y_{t-1}^k, \Sigma), \quad (8)$$

where $\mathcal{N}(\mu, \Sigma)$ stands for a random multivariate Gaussian distribution of mean μ and covariance matrix Σ . Diagonal elements of covariance matrix Σ are set to half of the maximum expected variation of each variable of the state space over one time step. However, this propagation may lead to poses out of the joint legal angular ranges. In this work, we present the following technique: when propagating particles, angular constraints are taken into account and samples of a truncated Gaussian distribution, denoted as \mathcal{N}^* , are generated instead of a complete Gaussian distribution, as shown in Fig.3. In this way, particles are always generated within the allowed ranges.

3. EXPERIMENTS AND RESULTS

In order to test the proposed algorithm, two tracking scenarios have been chosen. The first scenario is the standard HumanEva-I dataset [8] containing a set of 5 actions performed by 3 different subjects at a resolution of 640x480 pixels and a frame-rate of 25 fps. Ground truth data is available and two metrics are defined (mean, μ , and standard deviation, σ , of the estimation error) towards providing quantitative and comparable results. Moreover, metrics proposed in [9] for 3D human pose tracking evaluation are also employed. Assuming that landmark positions \hat{x}_m associated to particle y_t^i have been computed through forward kinematics, we can define a *matched* marker estimation \hat{x}_m with respect to the ground truth position x_m as the one fulfilling $\epsilon = \|x_m - \hat{x}_m\| < \delta$. This stands for those estimations that fall δ -close to the ground truth position. Then, the **Multiple Marker Tracking Accuracy (MMTA)**, is defined as the percentage of markers $x_m \in X$ fulfilling the $\epsilon < \delta$ condition, and the **Multiple Marker Tracking Precision (MMTP)**, as the average of the metric error between \hat{x}_m and x_m , of all pairs fulfilling $\epsilon < \delta$. Finally, these scores are averaged for all frames in the analysis sequence.

3.1. HumanEva results

A first experiment has been defined towards deciding the optimal number of layers N_L and particles per layer N_p . Available ground truth data allowed generating synthetic input data to our algorithm with a controlled degree of corruption driven by the triplet \overline{DR} , \overline{FP} and σ_F^2 . Hence, the robustness of the algorithm was evaluated in all

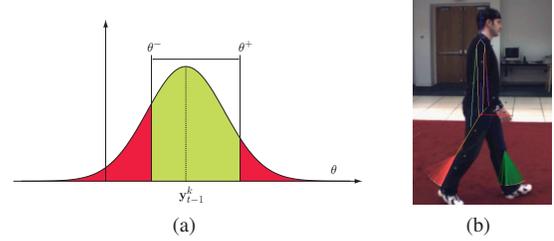


Fig. 3. Angular constraints enforcement. In (a), particles are propagated using a truncated Gaussian distribution \mathcal{N}^* centered at y_{t-1}^k with covariance matrix Σ bounded between θ^- and θ^+ (green zone). In (b), an example of particle propagation in the knee angle displaying how propagated particles never fall out the legal ranges ($\theta < 0$).

possible cases. The presented metrics have been computed for all possible combinations of N_L , N_p , \overline{DR} , \overline{FP} and σ_F^2 in a large simulation. As a summary, we contemplated the worst case scenario where we fixed $\overline{DR} \geq 0.9$ and explored the influence of \overline{FP} and σ_F^2 . *MMTA* score has been chosen as the most significant figure to assess the APF performance as shown in Fig.5. When fixing the variance error estimation σ_F^2 and increasing the false positive rate \overline{FP} , it is observed that it does not affect the overall performance of the algorithm since these false measurements in separate images do not hold a 3D consistency and, therefore, the generalized symmetric epipolar distance can reject them. On the other hand, when fixing \overline{FP} and increasing σ_F^2 , there is a progressive degradation of the performance of the tracker. The optimal operation point was set to be $N_L = 3$ and $N_p = 700$.

HumanEva-I dataset was analyzed using the proposed tracking system producing the results shown in Table 1. Averaged *MMTA* and *MMTP* scores indicates that in 95% of analyzed frames, difference between the estimation and the ground truth is below $\delta = 10$ cm and the committed error in these frames has an average of 45 mm. When comparing the performance for individual actions, it can be seen that those involving fast motion (boxing and jogging) exhibit a lower tracking performance than the others (walking or gesturing).

	Marker based APF			
	μ	σ	<i>MMTP</i>	<i>MMTA</i>
Walking	56.01	14.46	45.81	96.15
Jog	62.51	18.71	47.77	90.12
Throw/Catch	58.31	18.64	47.13	91.72
Gesture	44.70	4.31	42.42	97.46
Box	77.89	30.64	46.12	87.03
Average	59.88	17.35	45.85	95.32

Table 1. Quantitative results for the HumanEva-I dataset when using a marker detection algorithm with $\overline{DR} = 0.9$, $\overline{FP} = 20$ and $\sigma_F^2 = 4$ cm. PF parameters were set to $N_L = 3$ and $N_p = 700$. Distances are measured in millimeters and $\delta = 100$ mm.

3.2. Real scenario results

The presented body tracking algorithm has been applied to capture motion figures from 4 different types of dances: *salsa*, *belly dancing* and two Turkish folk dances. The analysis sequences were recorded with 6 calibrated cameras with a resolution of 1132x980 pixels at 30 fps. Markers attached to the body of the dance performer were little yellow balls and a color-based detection algorithm Γ has been used to generate the sets \mathcal{D}_n for every incoming multi-view frame. The original images are processed in the YCrCb color space which gives



Fig. 4. APF tracking examples in a real scenario involving fast motion.

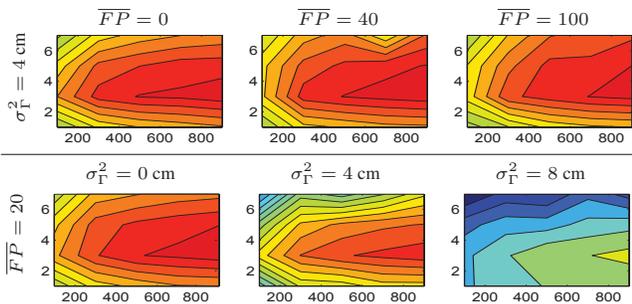


Fig. 5. MMTA results for several operation conditions fixing $\overline{DR} = 0.9$. Vertical and horizontal axes stand for N_L and N_p , respectively.

flexibility over intensity variations in the frames of a video as well as among the videos captured from different views. In order to learn the chrominance information of the marker color, markers on the dancer are manually labeled in one frame for all camera views. It was assumed that the distributions of Cr and Cb channel intensity values belonging to marker regions are Gaussian. Thus, the mean can be computed over each marker region (a pixel neighborhood around the labeled point). Then, a threshold in the Mahalanobis sense is applied to all images in order to detect marker locations. An empirical analysis showed that the detector Γ had the following performance triplet: $\overline{DR} = 0.98$, $\overline{FP} = 4$ and $\sigma_\Gamma^2 = 2$ cm. In this particular scenario, the algorithm had to cope with very fast motion associated to some figures. Even though this harsh conditions, the results were satisfactory and visually accurate as shown in Fig.4.

In this realistic scenario, a distributed computing system was employed to process the input images and generate the \mathcal{D}_n sets using the Γ marker detector. These data was fed to the APF algorithm that, due to the low complexity of the involved operations, was able to attain real-time performance (25 fps) in a 3 GHz computer.

4. CONCLUSIONS AND FUTURE WORK

This paper presents a robust real-time low cost approach to marker based human motion capture using multiple cameras. Progressive fitting of a HBM through the APF algorithm using a multi-view consistency likelihood function and a kinematically constrained particle propagation model allowed an accurate estimation of the body pose. Quantitative evaluation based on HumanEva dataset assessed the robustness of the algorithm when dealing faulty input data. Fast dance motion was also analyzed proving the adequateness of our technique to deal with a real scenario data.

Future work aims at exploiting the scalability of the HBM towards designing fitting algorithms able to cope with occluded body parts observed when there are occluding elements in the scene. Other research lines aim at gait and motion disorders analysis.

5. REFERENCES

- [1] A. Kirk, J. O'Brien, and D. Forsyth, "Skeletal parameter estimation from optical motion capture data," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 782–788.
- [2] "Moven-Inertial Motion Capture," <http://www.moven.com>.
- [3] "VICON," <http://www.vicon.com>.
- [4] L. Herda, P. Fua, R. Plankers, R. Boulic, and D. Thalmann, "Using skeleton-based tracking to increase the reliability of optical motion capture," *Human Movement Science*, vol. 20:3, pp. 313–341, 2001.
- [5] E. Aguiar, C. Theobalt, and H. Seidel, "Automatic learning of articulated skeletons from 3D marker trajectories," in *Proc. Int. Symposium on Advances in Visual Computing*, 2006, pp. 4291 of *Lecture Notes on Computer Science*, pp. 485–494.
- [6] G. Guerra-Filho, "Optical motion capture: theory and implementation," *Journal of Theoretical and Applied Informatics*, vol. 12:2, pp. 61–89, 2005.
- [7] P. Cerveri, A. Pedotti, and G. Ferrigno, "Robust recovery of human motion from video using Kalman filters and virtual humans," *Human Movement Science*, vol. 22, pp. 377–404, 2003.
- [8] L. Sigal and M.J. Black, "HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Tech. Rep. CS-06-08, Department of Computer Science, Brown University, 2006.
- [9] C. Canton-Ferrer, J.R. Casas, and M. Pardàs, "Towards a fair evaluation of video-based 3D human pose algorithms," in *Proc. IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (submitted)*, 2009.
- [10] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50:2, pp. 174–188, 2002.
- [11] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *Int. Journal of Computer Vision*, vol. 61:2, pp. 185–205, 2005.
- [12] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.
- [13] C. Canton-Ferrer, J.R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," 2005, vol. 3515 of *Lecture Notes on Computer Science*, pp. 281–289.
- [14] F. Caillette, A. Galata, and T. Howard, "Real-time 3D human body tracking using variable length Markov models," in *Proc. British Machine Vision Conference*, 2005, vol. 1, pp. 469–478.