A Motion Capture System for Sport Performance Analysis Based on Inexpensive RGB-D Sensors

Cyrille ANDRE*[1], Cédric LEMAITRE[1], Matthieu VOIRY[1], Antoine LAVAULT[1]
[1] Apeira Technologies, Le Creusot, France

https://doi.org/10.15221/19.176

Abstract

Motion capture for gesture and performance analysis in sports usually requires tracking of several anatomical landmarks with high spatial accuracy and sampling frequency. In addition, a model of the outer surface of the body can also be useful, e.g. for aerodynamic studies. In this context, we developed a motion capture system based on four inexpensive commercial RGB-D cameras. Our system produces both skeletal poses and 3D meshes of moving human bodies at high frequency with good accuracy.

Keywords: motion capture, 4D scanning, RGB-D sensors, ICP, gesture analysis in sports.

1. Introduction

Motion capture for sport gesture and performance analysis requires tracking of several anatomical landmarks with high spatial accuracy and sampling frequency. In this area, marker-based optical systems remain the most popular solution. It is indeed one of the most reliable and accurate techniques. However, the need to place markers, the quite long set-up time, and the price of such systems are major drawbacks in some situations.

In this context, we developed an inexpensive and markerless motion capture system. On the one hand, it allows capturing the dynamic pose of the athlete, and on the other hand, the external surface of his body, which is useful for aerodynamic studies. This system operates four inexpensive RGB-D cameras connected to a single-board computer. This multi-view approach increases the number of observations and reduces the hidden areas of the body, which ultimately improves accuracy and frequency of measurements, but also causes some difficulties.

First, we have to manage the large amount of data generated by the four sensors working independently. Furthermore, due to the fact that the RGB-D cameras cannot be triggered, we must deal with data captured asynchronously. Thus, we must ensure that the timestamps are consistent and take into account the movements in the scene in order to properly merge data of all the sensors. It is also essential to calibrate the cameras with sufficient accuracy in the volume of interest. Proper calibration avoids reconstruction error due to inaccurate estimates of the rigid transformations between sensors. Finally, we need to address specific issues related to the body model that we use to produce dynamic skeletons and watertight meshes of the moving athlete.

In this paper, we will first briefly discuss the related works. Then, we will present our motion capture system and the proposed solutions for the various problems outlined above. We will finally show some results obtained with the system and make a first evaluation of its performance.

2. Related Works

Optical systems using data captured by image sensors to triangulate the 3D positions of passive or active markers are the most widely used motion capture systems [1]. Markerless solutions have also been studied for decades [2,3], but, as stated in [4], they implicitly admit their lower accuracy by using the results of marker-based methods as reference. However, over the last few years the markerless approaches tend to fill the gap [4,5,6]. Finally, several methods to capture 3D skeletal pose using a single RGB camera have also been published [12, 13].

In addition to optical sensors, depth sensors are also widely used to capture dynamic scenes [7], including human pose [8, 9] or attitude [5]. Such approach can take advantage of direct depth measurements by light set-up with few sensors [10,11].

* c.andre@apeira-technologies.fr.; +33 *9 72 36 66 48*; www.apeira-technologies.fr

The markerless methods generally track changes in a 3D model between consecutive frames. The model can consist of elementary geometric shapes [14,15], be created during an initialization step [16, 17], or alternatively be a generative body model [18,19,20] fitted to the observations.

Often, markerless approaches also capture the surface of the bodies being tracked in addition to skeleton pose [21]. The 3D model and the skeletal pose determine the shape of this surface, but some local deformations can be updated dynamically [4,5,22].

3. Methods

3.1. Overview

We propose to infer both skeletal pose and surface of a moving human body by fitting a template model to depth maps measured asynchronously by four RGB-D sensors.

The appearance of the body model is driven by three sets of parameters. The shape parameters describe the morphology. Each of them affects the position of all points of the surface and should remain constant for an individual. The pose parameters describe the relative orientation of the skeletal joints. Each point of the surface is affected by the displacements of the joints to which it is attached, according to a blend skinning function. Finally, each point of the surface can be moved in the normal direction. This third set of parameters allows to represent geometric details of the clothes or the hair. Note that like the pose parameters, these parameters can also evolve over time.

After calibrating the system, the parameters are estimated by fitting the model to the observations. This process includes the following steps:

- The pose and shape parameters are first estimated simultaneously using the data recorded during a static initialization pose. Assuming the subject does not move, the set of parameters can be evaluated with data from different points of view. In this way, the measured point cloud better covers the model surface and the shape parameters can be better estimated. A first guess of local deformations can also be made.
- While studying the movements of interest, the pose is estimated frame by frame while the morphology remains constant. Starting with the previous pose, the model is fitted to the observations by minimizing a cost function. This function takes into account the distance between each point of the surface and the closest observation, found iteratively.
- Once the pose estimation has been completed for the current frame, the local deformations of the observed points of the surface are updated according to the measurements.

3.2. Data acquisition

Our system consists of 4 RGB-D cameras (Microsoft Kinect v2) plugged on single-board computers (Odroid XU4). All sensors are connected via an Ethernet network to a consumer-grade master computer dedicated to control, monitoring and data processing.

Data from all sensors are sent to the master at low resolution and low frame rate, for monitoring purpose only. In order to not saturate the network, data in full resolution are stored locally and transmitted to the master at once at the end of the recording.

The master also synchronizes the sensors' clocks to ensure the consistency of recordings' timestamps. This synchronization is achieved by sending customized messages in a similar way as the Precision Time Protocol (PTP).

3.2. Calibration

The rigid transformations from each camera coordinates system to a world reference frame have to be evaluated before each experiment. Moreover, even a minor error at this stage will significantly degrade multi-view processing. It is therefore essential to calibrate the cameras with sufficient precision and this treatment should be as simple as possible.

To do so, our calibration method uses reflective markers sticked on the ground inside the area observed by all sensors. The markers are visible in the IR image and form a regular grid (see Fig. 1.a). Using additional markers we added two orthogonal axes, which, together with the vertical direction, define the world reference frame.

To compute the rigid transformation from a camera coordinates system to this frame, we first compute the 3D positions of the markers detected in the IR image using the depth map. Then, the two orthogonal axes are identified and used to roughly define the reference frame in the camera frame, assuming it's a right-handed coordinate system and that the sensor is not upside down (see Fig.1.b). Finally, in order to reinforce the consistency of the transformations, they are refined by minimizing the sum of quadratic distances between markers coordinates computed using depth maps and transformations corresponding to different cameras.

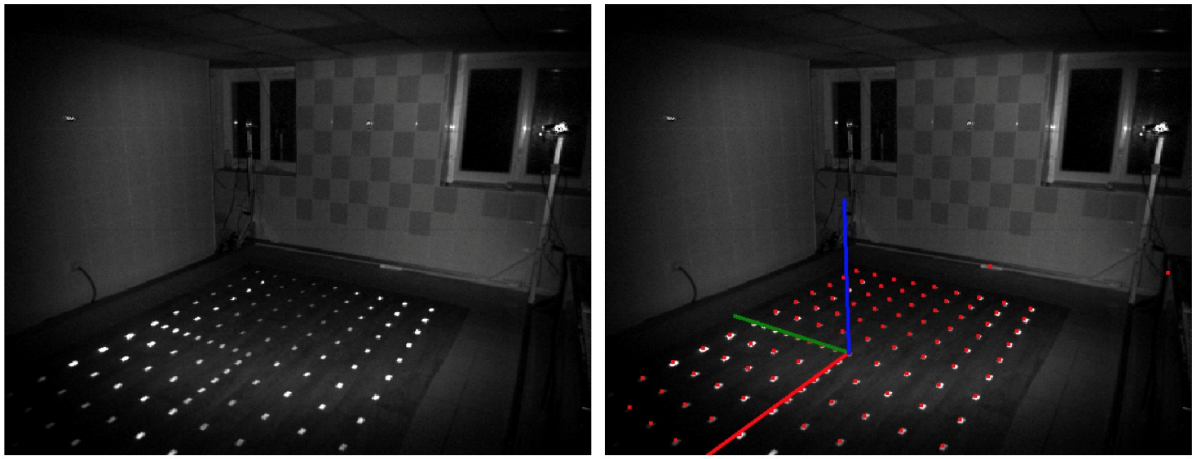

Fig. 1: a) IR image of marker grid ; b) Defined reference frame and detected markers.

The use of markers lying on the ground has several advantages:

- There is no need to install and remove markers for calibration: it greatly simplifies the operation.
- All sensors can detect the same markers and therefore the calibration error is not propagated through sensors as in a two-by-two calibration process.
- The calibration process is focused on the volume of interest. Some approaches based on targets placed outside this area may tend to reduce the residual error on the targets but not necessarily where it is needed.
- Two sensors facing each other can be calibrated, which is quite difficult using a vertical target.
- The reference frame is visible in the scene and consistent with the ground orientation. This is useful to avoid post-processing of 3D output.

3.3. Body model fitting

3.3.1. Body model

We use a parametric human shape model based on the Skinned Multi-Person Linear model (SMPL). SMPL [20] includes an outer surface with 6890 vertices and a skeleton with 24 joints. Each joint has 3 rotational degrees of freedom. Thus, by adding the translation of the root joint, there are 75 pose parameters.

Before applying the pose, the neutral body model T is deformed according to the shape parameters β and the pose parameters θ. The body shape $T(\beta, \theta)$ is formulated as the following sum:

$$T(\beta, \theta) = T + B_s(\beta) + B_p(\theta)$$

where B_s and B_p are vertex offsets, representing shape blendshapes and pose blendshapes respectively. The term B_p is designed to compensate for the distortions induced by skinning when the pose is applied.

The posed body model $M(\beta, \theta)$ is obtained by applying a skinning function to the body shape:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta, w)$$

where W is the blend skinning function, $J(\beta)$ the joint locations, and w the skinning weights. The parameters of the model are learned from data in order to produce realistic shapes in any poses.

3.3.2. Shape and pose fitting

At each time step *t*, the shape and pose parameters are estimated by minimizing an energy function E_{tot} that includes a data term and a prior term:

$$E_{tot} = \lambda_{data} E_{data} + \lambda_{prior} E_{prior}$$

The prior term E_{prior}, which prevents unobserved parameters from diverging, is formulated as:

$$E_{prior} = \lambda_\beta \sum_i \beta_i + \lambda_T f_T\big(\theta(t), \theta(t-1)\big) + \lambda_R f_R\big(\theta(t), \theta(t-1)\big)$$

The first term favors shape parameters with low values. Since these parameters represent deviations with respect to a morphological average, they should indeed not be too far from 0. The other terms penalize sudden changes of pose parameters; the function f_T is a simple difference between translation parameters, and the function f_R is the sum of distances between rotation parameters (expressed as quaternions).

The data term E_{data}, which estimates the fitting quality of the body model with a measured depth map, is expressed as:

$$E_{data} = \sum_{(v_m;u)\in P} \psi\big(n_{v_m}^T (v_m - u)\big)$$

where *P* is a correspondence set (a correspondence is a pair consisting of a point v_m of the model associated with the nearest point u in the depth map), n_{v_m} is the normal of the model surface at v_m, and the function ψ can be the Tukey penalty or a simple pruning function.

The energy E_{tot} is minimized using an Iterative Closest Point (ICP) approach. We alternate data association and Gauss-Newton optimization of the energy function based on the temporary correspondence set *P*.

Note that during the initialization step, both shape and pose parameters are optimized using depth maps from multiple sensors. Here, the energy E_{tot} therefore includes several data terms E_{data}. On the other hand, during the motion tracking process, the shape parameters remain constant; in that case, only pose parameters are estimated.

3.3.3. Local deformations

The SMPL model apprehends the overall appearance of bodies but does not capture irregularities due to clothes or hair. To take into account these local deformations, each vertex v_m of the model can be moved along the normal direction n_{v_m} as follows:

$$\widetilde{v_m} = \delta_m n_{v_m} + v_m$$

where the scalar δ_m is intended to compensate the difference between the model and the observed point cloud.

The δ_m parameters are first estimated for each visible vertex using closest observations in the following manner:

$$\delta_m = \frac{1}{W_m} \sum_{u_i \in \Omega_m} \Big(n_{v_m}^T (u_i - v_m)\Big) e^{\frac{-\|u_i - v_m\|_{rad}^2}{\sigma^2}}$$

where Ω_m is the subset of measured points close to v_m, $\|u - v\|_{rad}^2$ is the radial distance between *u* and v given by $\|u - v\|^2 - \big(n_v^T (u - v)\big)^2$, the normalization term W_m is the sum of the weights $e^{\frac{-\|u_i - v_m\|_{rad}^2}{\sigma^2}}$ and the parameter σ is set according to the point density on the surface of the model.

In practice, we keep the unnormalized sum S_m and the normalization W_m for each vertex. In order to take into account the evolution of deformations over time, new observations are included in the model using an exponential filtering:

$$\delta_m = \alpha S_m^{t-1} + (1-\alpha) \sum_{u_i \in \Omega_m} \Big(n_{v_m^t}^T (u_i^t - v_m^t)\Big) e^{\frac{-\|u_i^t - v_m^t\|_{rad}^2}{\sigma^2}}$$

$$W_m^t = \alpha W_m^{t-1} + (1-\alpha) \sum_{u_i \in \Omega_m} e^{\frac{-\|u_i^t - v_m^t\|_{rad}^2}{\sigma^2}}$$

$$\delta_m^t = \frac{S_m^t}{W_m^t}$$

Finally, before rendering, the deformations values of vertices are smoothed iteratively according to the values of their neighbors:

$$\hat{\delta}_m^{t,k+1} = \frac{1}{\widehat{W}_m^{k+1}} \sum_{j \in \omega_m} \hat{\delta}_j^{t,k} \widehat{W}_j^k \, e^{\frac{-\|v_j - v_m\|^2}{\sigma^2}}$$

$$\widehat{W}_m^{k+1} = \sum_{j \in \omega_m} \widehat{W}_j^k \, e^{\frac{-\|v_j - v_m\|^2}{\sigma^2}}$$

where ω_m is the set of the indices of neighboring vertices of v_m. The weights $\widehat{W}_j^0$ are initialized as the invert of the weighted variance of the samples belonging to Ω_j. As a result, the deformations propagate smoothly over the mesh, while the model remains close to the observations where the variance is small.

3.4. Motion tracking

3.4.1. Initialization

The generative nature of the method presented above assumes a relatively good initialization of the model. To ensure such proper initialization, we use a 2D skeleton detected in the RGB image. This skeleton is projected into the 3D scene using the depth map associated with the image and then the pose parameters are roughly estimated according to the alignment of body segments.

3.4.2. Tracking and resampling

During the motion tracking process, the model cannot be fitted to the depth maps from all sensors simultaneously due to their asynchronous aspect. Moreover, the visible part of the model varies with time, depending on the point of view of each sensor.

To tackle these problems, we integrate data from different sensors sequentially. For a given depth map, we first predict the pose parameters according to the body pose and angular speed at previous acquisition time. The pose is then corrected using the proposed body model fitting process. Finally, the temporal sequence is resampled at a regular frame rate by SLERP interpolation.

4. Results

4.1. Qualitative evaluation

Figure 2 and 3 illustrate the processing steps of our method for two subjects; it depicts (from left to right): color image, depth map, obtained skeleton, obtained mesh (without local deformations), and finally obtained mesh (with local deformations).

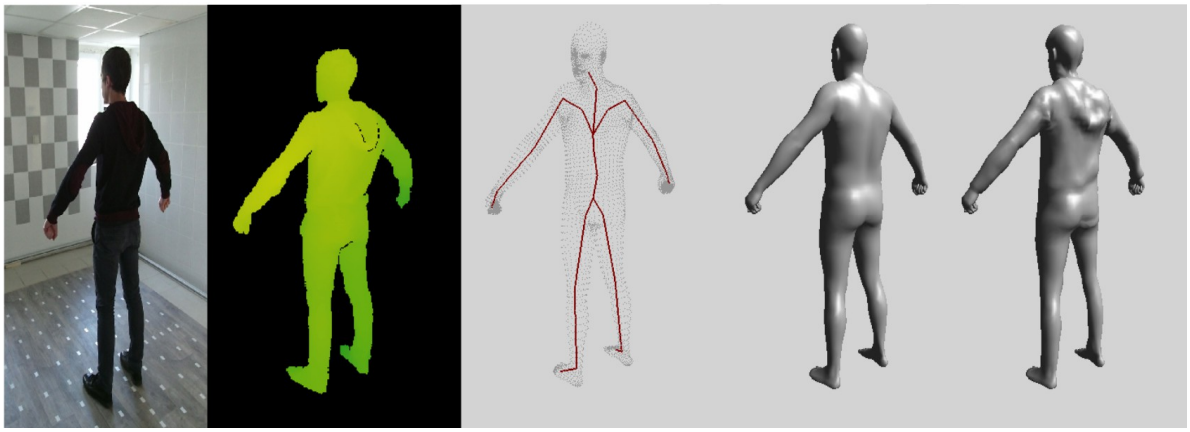

Fig. 2: Processing steps of our method (subject 1).

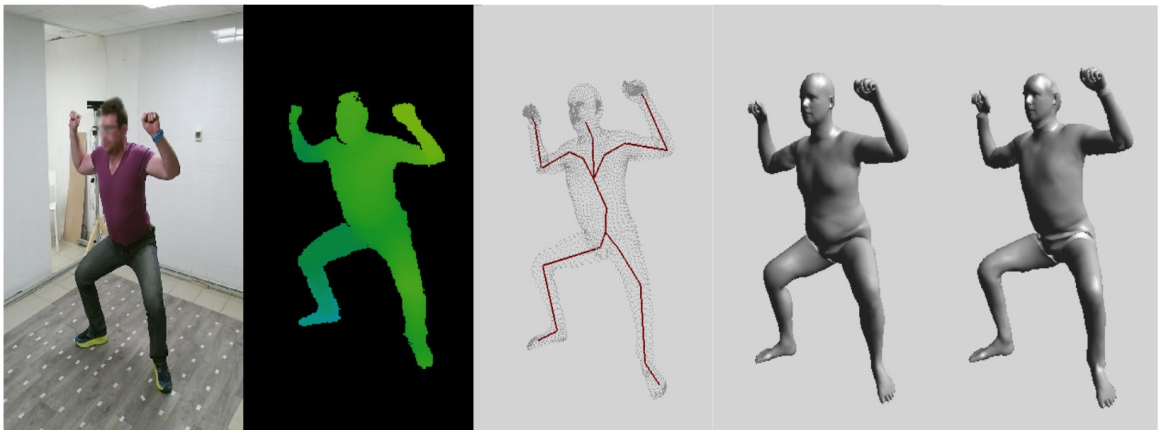

Fig. 3: Processing steps of our method (subject 2).

Figure 4 illustrates specifically the interest of using local deformations model; we can clearly see that the clothes are well captured.

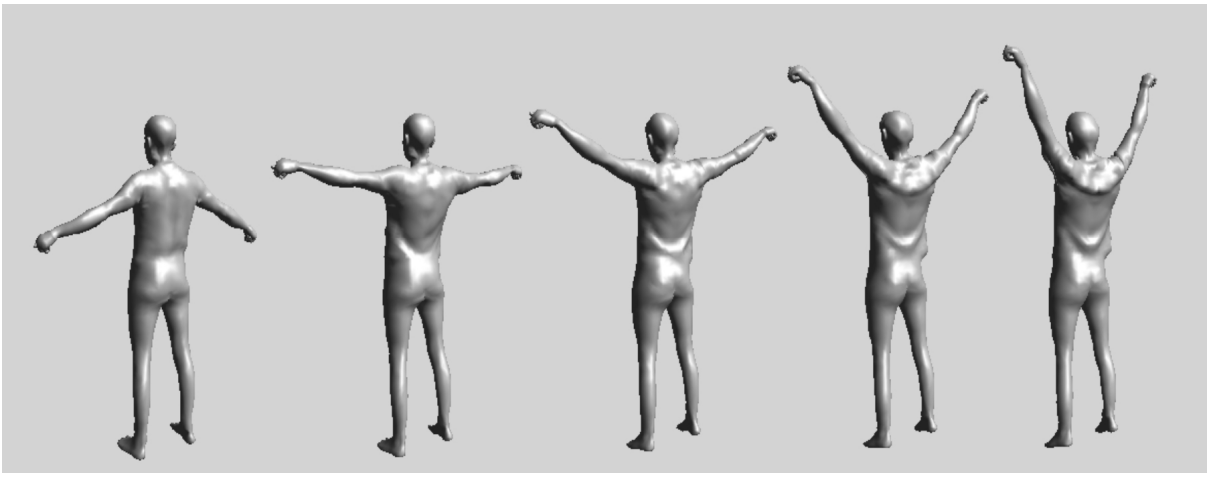

Fig. 4: Illustration of interest of using local deformations model.

Figure 5 shows other results for subjects with different morphologies. These results were obtained by processing data from MHAD database (see 4.2).

All these results demonstrate that our method can produce accurate skeletons and body meshes for subjects with a wide range of morphologies. Moreover, using local deformations model, we are able to capture fine surface details.

4.2. Quantitative evaluation

The quantitative evaluation of our system would require knowing human body poses exactly in dynamic scenes. Unfortunately it is almost impossible to obtain such data and therefore, the best we can do is to use another indirect measurement as reference. Marker-based optical motion capture systems are usually considered as reliable and are often used for this task.

Following this idea, we evaluated our method using the Berkeley Multimodal Human Action Database (MHAD). This database includes 11 actions performed multiple times by 12 different subjects, recorded simultaneously by multiple sensors including two Kinects v1 and a marker-based optical motion tracking system [23]. Although the used experimental system includes only two Kinects v1 (against four Kinect v2 for our system), the great number of actions performed by subjects with a wide range of morphologies and the simultaneous recording of different sensors make the database very useful.

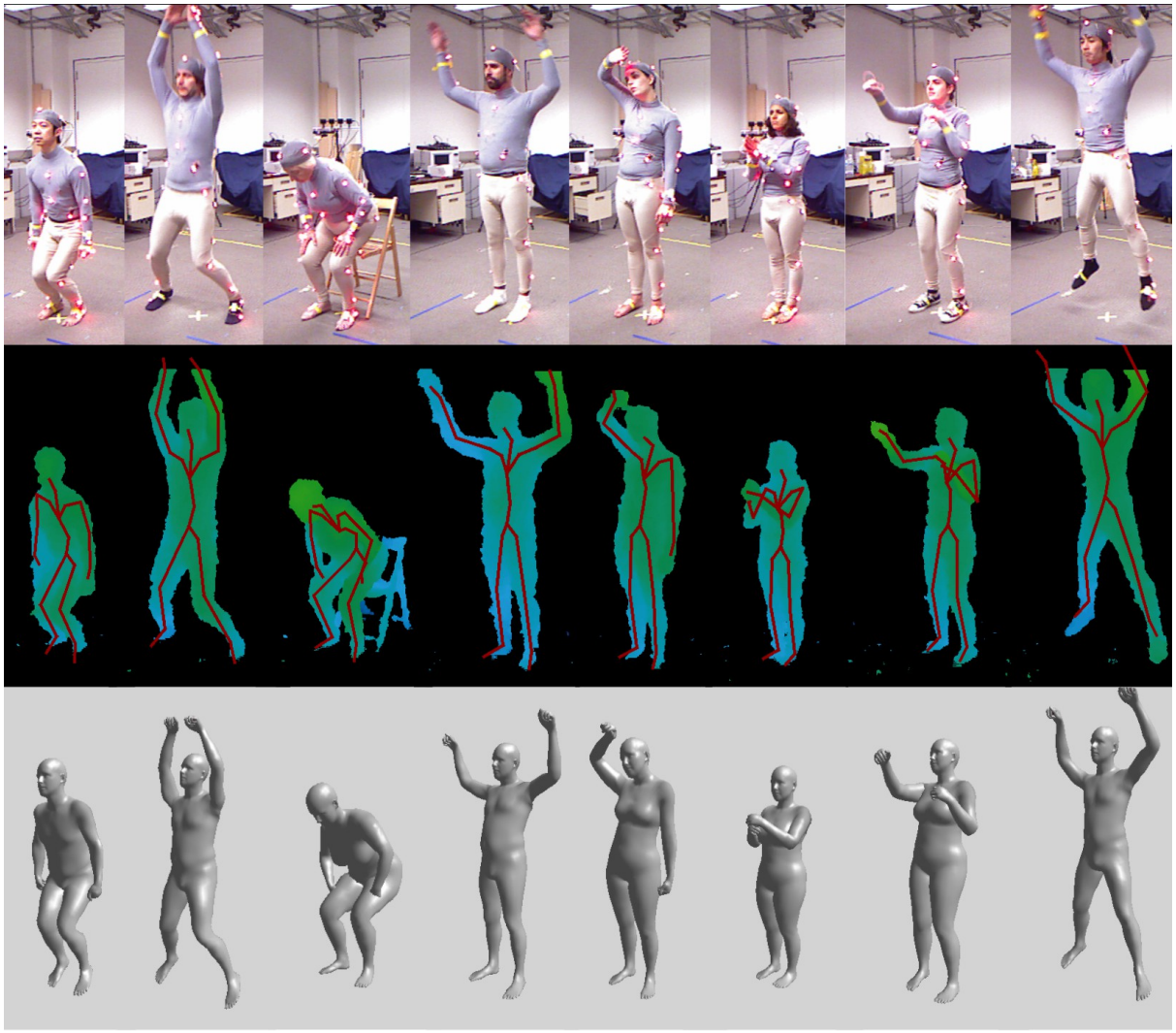

Fig. 5: Results obtained for subjects with different morphologies.

For each tested sequence, we computed the root mean square error between the expected position of each marker and the corresponding observation. The expected position is obtained using a normal vector on the body surface calculated in the first frame of the considered sequence. All RMSE are calculated independently for each marker (except those on the hands and feet) and finally, the results for all tested sequences are cumulated in a histogram (Fig.6).

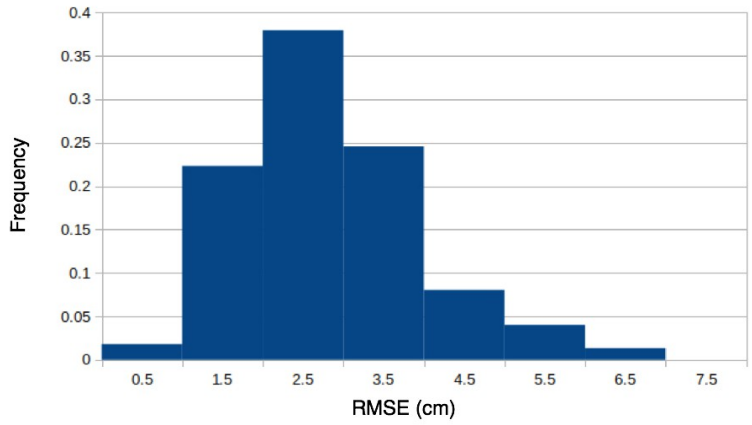

Fig. 6: Histogram of RMSE (cm) between the expected position of each marker and the observations.

These results are pretty good but it is important to note that they must be considered as a lower bound of the actual accuracy of our system. Indeed, processing data from Kinects v1 is disadvantageous because these sensors are less accurate than the Kinects v2 [24] that we use in our system. Moreover, we operate four sensors, which increases the number of observations and most likely improves the system accuracy. Finally, part of the measured error can be explained by the fact that, due to soft-tissue, markers can move about 2 cm from the body surface during motion [25].

5. Conclusion

In this paper, we presented a markerless and inexpensive motion capture system. It operates four commercial RGB-D cameras plugged on single-board computers and a desktop computer. After an automated and robust calibration step and a quick initialization stage, which allows to learn the individual morphological characteristics of the subject, the system is able to capture its movements. By fitting a template model to depth maps measured asynchronously by the sensors, it produces both skeletal poses and 3D meshes of the moving human body at high frequency (at least 60 Hz). We have showed experimentally that the system is able to track motion of people with a wide range of morphologies with good accuracy. However, further work is needed to better characterize the accuracy of the system.

References

[1] E. van der Kruk and M. M. Reijne, “Accuracy of human motion capture systems for sport applications ; state-of-the-art review,” European journal of sport science, vol. 18, no. 6, pp. 806–819, 2018. https://doi.org/10.1080/17461391.2018.1463397

[2] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, Performance capture from sparse multi-view video, vol. 27. ACM, 2008. https://doi.org/10.1145/1399504.1360697

[3] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, “Motion capture using joint skeleton tracking and surface estimation,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1746–1753, IEEE, 2009. https://doi.org/10.1109/CVPR.2009.5206755

[4] H. Joo, T. Simon, and Y. Sheikh, “Total capture: A 3d deformation model for tracking faces, hands, and bodies,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8320–8329, 2018. arXiv:1801.01615v1

[5] F. Bogo, M. J. Black, M. Loper, and J. Romero, “Detailed full-body reconstructions of moving people from monocular rgb-d sequences,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 2300–2308, 2015. https://doi.org/10.1109/ICCV.2015.265

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299, 2017. https://doi.org/10.1109/CVPR.2017.143

[7] R. A. Newcombe, D. Fox, and S. M. Seitz, “Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 343–352, 2015. https://doi.org/10.1109/CVPR.2015.7298631

[8] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt, “A data-driven approach for real-time full body pose reconstruction from a depth camera,” in Consumer Depth Cameras for Computer Vision, pp. 71–98, Springer, 2013. https://doi.org/10.1109/ICCV.2011.6126356

[9] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kip- man, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in CVPR 2011, pp. 1297–1304, Ieee, 2011. https://doi.org/10.1109/CVPR.2011.5995316

[10] B. Müller, W. Ilg, M. A. Giese, and N. Ludolph, “Validation of enhanced kinect sensor based motion capturing for gait assessment,” PloS one, vol. 12, no. 4, p. e0175813, 2017. https://doi.org/10.1371/journal.pone.0175813

[11] S. Meerits, D. Thomas, V. Nozick, and H. Saito, “Fusionmls : Highly dynamic 3d reconstruction with consumer-grade rgb-d cameras,” Computational Visual Media, vol. 4, no. 4, pp. 287–303, 2018. https://doi.org/10.1007/s41095-018-0121-0.

[12] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, “Vnect: Real-time 3d human pose estimation with a single rgb camera,” ACM Transactions on Graphics (TOG), vol. 36, no. 4, p. 44, 2017. https://doi.org/10.1145/3072959.3073596

[13] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis, “Monocap: Monocular human motion capture using a cnn coupled with a geometric prior,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 4, pp. 901–914, 2018. https://doi.org/10.1109/TPAMI.2018.2816031

[14] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, “Fast articulated motion tracking using a sums of gaussians body model,” in 2011 International Conference on Computer Vision, pp. 951–958, IEEE, 2011. https://doi.org/10.1109/ICCV.2011.6126338

[15] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, “Real-time human pose tracking from range data,” in European conference on computer vision, pp. 738– 751, Springer, 2012. https://doi.org/10.1007/978-3-642-33783-3_53

[16] Q. Zhang, B. Fu, M. Ye, and R. Yang, “Quality dynamic human body modeling using a single low-cost depth camera,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 676–683, 2014. https://doi.org/10.1145/2816795.2818013

[17] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, “Monoperfcap : Human performance capture from monocular video,” ACM Transactions on Graphics (ToG), vol. 37, no. 2, p. 27, 2018. https://doi.org/10.1145/3181973

[18] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, “Scape: shape completion and animation of people,” in ACM transactions on graphics (TOG), vol. 24, pp. 408–416, ACM, 2005. https://doi.org/10.1145/1186822.1073207

[19] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, “Dyna: A model of dynamic human shape in motion,” ACM Transactions on Graphics (TOG), vol. 34, no. 4, p. 120, 2015. https://doi.org/10.1145/2766993

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: A skinned multi-person linear model,” ACM transactions on graphics (TOG), vol. 34, no. 6, p. 248, 2015. https://doi.org/10.1109/CVPR.2014.92

[21] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, “Bodyfusion : Real-time capture of human motion and surface geometry using a single depth camera,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 910–919, 2017. https://doi.org/10.1109/ICCV.2017.104

[22] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, “Doublefusion : Real-time capture of human performances with inner body shapes from a single depth sensor,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7287–7296, 2018. https://doi.org/10.1109/TPAMI.2019.2928296

[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Berkeley mhad: A comprehensive multimodal human action database,” in 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60, IEEE, 2013. https://doi.org/10.1109/wacv.2013.6474999

[24] O. Wasenmüller and D. Stricker, “Comparison of kinect v1 and v2 depth images in terms of accuracy and precision,” in Asian Conference on Computer Vision, pp. 34–45, Springer, 2016. https://doi.org/10.1007/978-3-319-54427-4_3

[25] Leardini, L. Chiari, U. Della Croce, and A. Cappozzo, “Human movement analysis using stereophotogrammetry : Part 3. soft tissue artifact assessment and compensation,” Gait & posture, vol. 21, no. 2, pp. 212–225, 2005. https://doi.org/10.1016/j.gaitpost.2004.05.002