IEEE Copyright Notice

Vasileiadis, M., Giakoumis, D., Malassiotis, S., Kostavelis, I., & Tzovaras, D. (2017, June). Body-part tracking from partial-view depth data. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2017* (pp. 1-4). IEEE. DOI = https://doi.org/10.1109/3DTV.2017.8280408

BODY-PART TRACKING FROM PARTIAL-VIEW DEPTH DATA

Manolis Vasileiadis, Dimitris Giakoumis, Sotiris Malassiotis, Ioannis Kostavelis and Dimitrios Tzovaras

Information Technologies Institute, Centre for Research and Technology Hellas
57001, Thermi - Thessaloniki, Greece

ABSTRACT

This paper presents a high-accuracy body-part tracking algorithm, capable of achieving efficient human motion analysis from partial view depth-data, suitable for deployment in real-life applications. The algorithm uses a consumer-grade depth camera for data input and combines a discriminative body part estimator along with a generative tracker, utilizing a realistic human body model, in order to track individual body limbs in short camera-distance, partial-view scenarios. Additionally, a shape adaptation feature is also introduced in order to further morph the human model based on the observations. The implementation is tested in a lower-body limbs tracking scenario, achieving promising accuracy and performance on consumer-grade hardware. Moreover, a lower-body motion dataset is also provided, consisting of 16 real-world sequences using automatic ground-truth annotations from a commercial motion capture system.

Index Terms — limb tracking, pose estimation, motion analysis, shape adaptation, partial view

1. INTRODUCTION

Human motion analysis and pose estimation techniques have found usage in a large variety of technology domains, aiming towards the development of applications with life-like, human-centric interfaces, which can increase the immersiveness and realism of the services provided to the user, while simultaneously enhancing the human-machine interaction experience, both on the virtual and physical level. Using natural body motions and gestures, such as head noding and arm waving, can enhance traditional computer input methods, while utilizing the human body as a controller (i.e. Microsoft Kinect, Nintendo Wii), can increase the immersiveness of modern video-games, especially in the fast-developing area of Augmented Reality (AR) games. Similarly, efficient human motion capture can provide realistic animations in virtual environments. Some of the areas where human pose estimation techniques are typically utilized include: human computer interaction (HCI), entertainment, security, healthcare, robotics [1][2].

Estimating the human pose can be an intricate and complex task, as the human body presents high variability in shape and size, along with a very large range of motion for each body part. While earlier attempts used complex marker-based motion capture systems towards this end, the demand for adaptable body tracking solutions suitable for user-friendly real-life applications, has turned the focus of the research community towards marker-less human pose estimation techniques, using low-cost consumer-grade RGB and more recently, depth cameras.

The utilization of such cameras, however, does present some drawbacks, which mainly derive from the cameras' single viewpoint and limited field of view [3]. While, these drawbacks rarely cause any major problems in controlled environments, such as laboratory trials, where the monitoring conditions tend to be ideal (i.e. full body view, no occlusions etc.), they can affect the motion analysis accuracy in real-life applications, where commonly encountered environmental factors, such as short camera-user distance, can significantly reduce the viewing area of the sensor unit. Moreover, the increased utilization of marker-less motion analysis techniques in a variety of commercial and scientific applications, has rendered high-accuracy limb-tracking necessary in order to achieve a satisfactory user experience, such as tracking fine gestures in free-hand HCI [4], realistically animating avatars in VR applications, ensuring safety and effectiveness in physical human-machine interactions (i.e. object handover in service robotics tasks) etc.

Towards this end, we propose a high-accuracy body-part tracking algorithm, capable of efficient tracking in scenarios where there is only a partial view of the tracked human available. The developed implementation uses a commercially available depth sensor for data input, and achieves real time performance on consumer-grade hardware, thus making it suitable for deployment on commercial and scientific applications.

2. RELATED WORK

A lot of research effort has been put towards human full-body pose estimation and tracking, leading to the development of both discriminative pose estimators, which estimate the human pose from a single frame using large training datasets and machine learning techniques [2][5][6][7], and generative body pose trackers, which track the detected body parts through consecutive frames by matching the input data to articulated body templates and minimizing an objective function through the utilization of various optimization techniques [8][9][10][11]. These methods, have been developed mainly targeting views where the whole human body is visible; however, they can be customised and re-purposed for body part estimation in partial-view scenarios as well.

Plagemann et al. [5] propose a novel interest point detector suitable for mesh and depth data. The interest points, called Accumulative Geodesic Extrema (AGEX), are computed by incrementally maximizing geodesic distances on the surface of the 3D mesh. Small depth image patches surrounding these points are then used as local descriptors in order to train a boosted classifier. In [6][7] randomized decision trees and forests are used for body part detection and treat the body part segmentation as a per-pixel classification task, with each pixel in the depth image being evaluated separately.

The problem of hand detection and tracking for short distance views is tackled in [12][13]. The hand area is segmented using RGB-based skin detection and filtering of the depth data, in order to provide an initial estimation of the hand position. Next, a model-based approach for 3D tracking of hand articulations is utilized with the hand pose being detected by minimizing the difference between the possible instances of a 3D hand model and the real visual observations of the human hand, while an objec-

978-1-5386-1635-2/17/$31.00

tive function measures that difference and defines the distance between the hand pose hypothesis and the observation. In similar fashion, Schmidt et al. [9] achieve robust hand pose tracking using the general-purpose DART tracker, in which the objects are represented by a symmetric version of the articulated Signed Distance Function (SDF) and gradient based optimization is used to estimate the pose.

Towards lower limb tracking, usually within the context of human gait analysis, Hu et al. [14] propose a method for tracking the legs of a walking human. One depth and two RGB cameras with direct frontal view of the walker's legs are used in order to track them, while a Hidden Markov Model (HMM) is utilized to estimate the pose of the legs from the observed data. In [15], the authors combine particle filtering [16] with the human locomotion model [17] for feet tracking from a robot-mounted RGB camera, while in [18] a complementary laserscan-based leg detection module is proposed within the context of a larger indoor human tracking framework.

Our method builds upon current pose estimation techniques, by combining a discriminative body part estimator along with a generative tracker, which utilizes a realistic human body model, in order to accurately track body limbs in close distance, partial view scenarios. Moreover, a shape adaptation step is also introduced, in order to further morph the human model based on the observed limb, resulting in high accuracy reconstruction and tracking of the body part. The implementation described in the following sections focuses on the scenario of lower-body limbs tracking, namely the shins and feet, which can be encountered in areas such as healthcare (lower body motion analysis i.e. testing maximal range and speed for leg extension), assistive service robotics etc. However, the same methodology can be easily utilized for upper-body limbs tracking. Finally, we have constructed a lower-body motion dataset comprised of real-world test sequences, along with ground truth data, which is published openly [19] for future benchmarks.

3. PROPOSED METHODOLOGY

The developed lower-limb tracker estimates the exact position of the human's legs, by trying to match a realistic articulated human body model, created offline using the MakeHuman [20] open-source tool, to the observed human 3D point cloud. The 3D data is produced by back-projecting the input depth image to the camera's 3D world coordinate system, while background / foreground segmentation and a rough estimation of the human's torso position are considered to be known beforehand.

3.1. Model Initialization

For the lower-limb tracker to successfully estimate the position of the human's legs, the model must be initialized to a pose approximating the actual pose of the observed legs. Towards this end an interest point detector is utilized, based on the work of Plagemann et al. [5], in order to find candidate interest points on the human's feet. Specifically, starting from the approximate position of the human's torso, the geodesic distances along the surface of the 3D human point cloud are calculated, utilizing Dijkstra's algorithm [21], in order find extrema points which correspond to the human leg (foot and shin) (Figure 1a). The estimated points are then used to segment the point cloud into foot and shin areas, by backtracking from these points and clustering all the 3D points that fall within an experimentally selected radius r_{limb}. Once the point cloud is segmented, the Articulated Iterative Closest Point (A-ICP) algorithm [22] is employed in order to register the human point cloud to the human body model (Figure 1b).

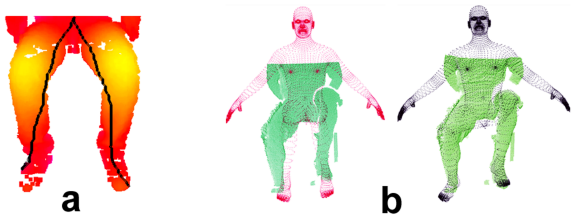

Figure 1. Model initialization: a) Geodesic distance calculation along the surface of the human point cloud, b) Left - rough initialization based on the detected extrema points, Right - pose refinement using A-ICP

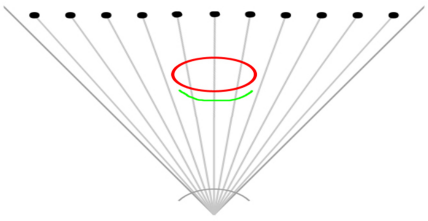

Figure 2. 2D slice of the ray casting model. Points of the 3D model (red) and the observation point cloud (green) that lie along the same ray are considered corresponding points. The model points are then translated along the ray in order to align with the observation points. In case of multiple model points along a single ray, only the closest point is taken into consideration.

3.2. Limb Tracking

After the completion of the initialization step, the human leg is tracked in subsequent frames: for each frame, starting from the last successful leg pose estimation, K-Dimensional (K-D) tree partitioning is performed in order to maintain points of interest between the human 3D point cloud and the initialized human model, followed by point-to-plane ICP which aligns the maintained points and the model. The overall ICP registration error is used as a tracking failure metric: if it exceeds an error threshold E_{foot}, the foot tracking accuracy is deemed inadequate and the lower-limb tracker is re-initialized following the initialization process described above.

3.3. Shape Adaptation

An additional tracking feature is also introduced in order to account for the changes in the leg shape due to the clothes that the tracked human may be wearing. Specifically, after each successful model / point cloud alignment, the human model is dynamically morphed based on the acquired point cloud; a ray casting model [23], originating from the camera point of view, is utilized in order to translate each point of the model to the position of the corresponding point of the 3D point cloud (Figure 2). Model points that do not have any correspondences (e.g. points on the rear side of the leg, which are not visible to the camera) are not taken into consideration during the tracking step on the next frame, resulting in faster convergence of the iterative algorithm and an accurate representation of the tracked limb.

4. EXPERIMENTAL EVALUATION

4.1. Dataset Generation

For the experimental evaluation of the developed lower-limb tracker, a lower-body motion dataset was captured, using a Kinect v1 RGB-

Figure 3. Left: dataset capture setup, Right: Kinect field of view, Bottom: Pattern poses during the Kinect/Vicon calibration process.

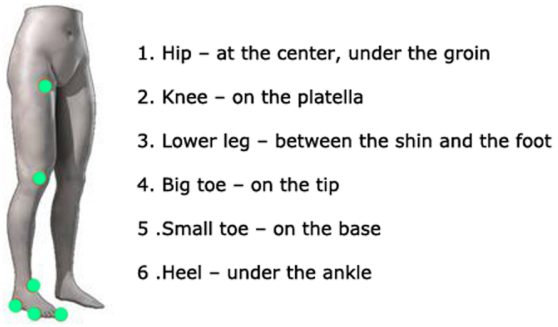

Figure 4. Positions of the ground truth markers tracked by Vicon.

D camera. The camera was positioned at height h=1.3m and tilted down at θ=45°, while the subjects were seated at a distance d=1.4m, facing towards the camera, thus providing a realistic camera viewpoint for lower-body motion analysis (Figure 3). Each subject was asked to lift and extend both legs while being monitored by the camera, which recorded RGB and Depth frames at 30fps. In total 16 subjects (13 male, 3 female) participated in the data recording procedure, resulting in 20000 captured Depth and RGB frames. Before commencing the data capture process, the RGB-D camera was externally calibrated to estimate the camera intrinsic parameters and distortion coefficients, in order to ensure correct pixel correspondence between the RGB and Depth frames and accurate 2D-to-3D projection of the depth data.

The leg position and pose ground truth data were captured using the Vicon motion capture system, which utilized 10 IR cameras in order to track 6 reflective spherical markers (r=0.5cm) positioned on each leg, at 100 Hz. The exact positioning of the markers is presented in Figure 4.

In order to transform the ground truth data provided by Vicon to the coordinate system of the Kinect camera, a Kinect / Vicon calibration step was executed before each recording session. Specifically, a large chessboard pattern was positioned in 5 different poses in order to cover the whole field of view of the camera (Figure 3). For each pose, 8 reflective markers were attached at predefined spots and tracked by Vicon, with the readings used to infer the positions of all the square edges of the pattern on the coordinate system of Vicon. Next, the pattern was captured by the Kinect sensor, with the square edges automatically detected on the RGB frame and their positions estimated on the coordinate system of the Kinect depth sensor. Finally, the transformation matrix between the two coordinate systems was calculated using a single iteration ICP. The average correspondence distance error across all the calibration sessions was E_{calib}=0.82cm.

A small deviation between the Vicon ground truth data and the 3D point cloud produced by the Kinect was also noticed in each frame. This systematic error, defined as the average distance between the Vicon readings and the corresponding points on the 3D point cloud, was manually estimated at E_{sys}=2.73cm, and can be attributed to three factors:

- accuracy of the Vicon readings
- accuracy of the Kinect readings and the intrinsic calibration
- temporal synchronization of the two streams

4.2. Experimental Results

Finally, the lower limb tracker was evaluated on all the pre-recorded sequences of the dataset, by computing the Euclidean distance among the detected Vicon markers on each point cloud and the fixed points on the human model, leading to an overall average positioning error of $E_{overall}$=4.6cm$\pm$0.3cm SD. Taking into consideration, and subsequently removing, the systematic error E_{sys} mentioned above, the algorithm's actual average positioning error was found to be E= $E_{overall}$-E_{sys}=1.9cm$\pm$0.3cm SD.

While there are not currently available any state-of-the-art methods targeting specifically partial-view lower-limb tracking, in order to provide a direct comparison to the proposed approach, the best accuracy results reported in similar works (3.4cm for full body tracking [10], 1.84cm for hand tracking [24]) serve as evidence that the presented method's performance is comparable to current state-of-the-art body-part tracking methods.

Moreover, the tracker was also tested under realistic conditions in an online fashion, achieving an operation framerate of 10-15fps, with the operation speed mainly affected by the complexity and speed of the tracked human's movement. An indicative successful foot tracking sequence is presented in Figure 5.

5. CONCLUSIONS

This paper introduced a high-accuracy body-part estimation and tracking algorithm, capable of robust performance in real-life applications, suitable for partial-view scenarios where due to limitations in the depth camera viewpoint and field of view only a part of the tracked human is visible. The algorithm builds upon modern motion analysis techniques by combining a body part estimator with an articulated tracker and further enhancing the limb-tracking accuracy by adapting the shape of the utilized human model to the observations. The algorithm is tested in a lower-body limb tracking scenario and achieves an average accuracy <2cm while performing at 10-15 fps, on non-optimized code. Additionally, a custom real-life lower-body motion dataset, with annotated ground truth data, is also provided for future benchmarking.

Future work may include extensive testing of the algorithm on upper-body limbs tracking scenarios, namely arms tracking, further enhancement of the shape-adaptation process, and GPU-optimization of the code in order to achieve a higher processing framerate.

6. ACKNOWLEDGEMENTS

This work has been supported by the EU Horizon 2020 funded project "Robotic Assistant for MCI Patients at home (RAMCIP)" under the grant agreement no. 643433

7. REFERENCES

[1] Thomas B Moeslund, Adrian Hilton, and Volker Krüger, "A survey of advances in vision-based human motion capture

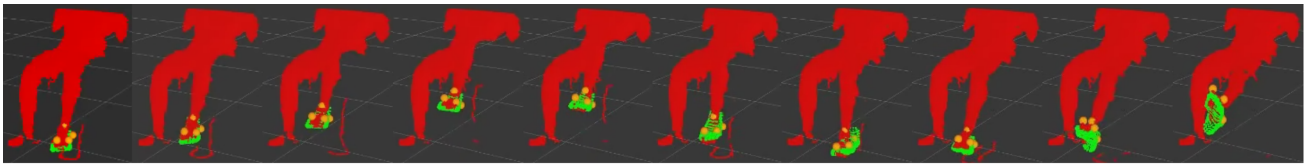

Figure 5. Sample foot tracking sequence from the lower-body motion dataset. Red: input point cloud, green: estimated foot pose, orange: ground truth markers.

and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[2] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen, "A survey of human pose estimation: the body parts parsing based methods," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 10–19, 2015.

[3] Tao Wei, Brian Lee, Yuansong Qiao, Alexandros Kitsikidis, Kosmas Dimitropoulos, and Nikos Grammalidis, "Experimental study of skeleton tracking abilities from microsoft kinect non-frontal views," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*. IEEE, 2015, pp. 1–4.

[4] Shun Zhang, Jinjun Wang, Yihong Gong, and Shizhou Zhang, "Free-hand gesture control with" touchable" virtual interface for human-3dtv interaction," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*. IEEE, 2015, pp. 1–4.

[5] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun, "Real-time identification and localization of body parts from depth images," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 3108–3113.

[6] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[7] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon, "Metric regression forests for correspondence estimation," *International Journal of Computer Vision*, vol. 113, no. 3, pp. 163–175, 2015.

[8] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun, "Real-time human pose tracking from range data," in *European conference on computer vision*. Springer, 2012, pp. 738–751.

[9] Tanner Schmidt, Richard A Newcombe, and Dieter Fox, "Dart: Dense articulated real-time tracking.," in *Robotics: Science and Systems*, 2014.

[10] Mao Ye and Ruigang Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2345–2352.

[11] Meng Ding and Guoliang Fan, "Articulated gaussian kernel correlation for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 57–64.

[12] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros, "Efficient model-based 3d tracking of hand articulations using kinect.," in *BmVC*, 2011, vol. 1, p. 3.

[13] Dan Song, Nikolaos Kyriazis, Iason Oikonomidis, Chavdar Papazov, Antonis Argyros, Darius Burschka, and Danica Kragic, "Predicting human intention in visual observations of hand/object interactions," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1608–1615.

[14] Richard Zhi-Ling Hu, Adam Hartfiel, James Tung, Adel Fakih, Jesse Hoey, and Pascal Poupart, "3d pose tracking of walker users' lower limb with a structured-light camera on a moving platform," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011, pp. 29–36.

[15] Ying Li, Sihao Ding, Qiang Zhai, Yuan F Zheng, and Dong Xuan, "Human feet tracking guided by locomotion model," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2424–2429.

[16] Shaohua Kevin Zhou, Rama Chellappa, and Baback Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491–1506, 2004.

[17] Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Yokoi, and Hirohisa Hirukawa, "The 3d linear inverted pendulum mode: A simple modeling for a biped walking pattern generation," in *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*. IEEE, 2001, vol. 1, pp. 239–246.

[18] Matthias Scheutz, John McRaven, and Gy Cserey, "Fast, reliable, adaptive, bimodal people tracking for indoor environments," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*. IEEE, 2004, vol. 2, pp. 1347–1352.

[19] "RAMCIP: Robotic Assistant for MCI Patients at home," http://www.ramcip-project.eu/ramcip-data-mng.

[20] "MakeHuman: Open source tool for making 3D characters," http://www.makehuman.org.

[21] Edsger W Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[22] Stefano Pellegrini, Konrad Schindler, and Daniele Nardi, "A generalisation of the icp algorithm for articulated bodies.," in *BMVC*. Citeseer, 2008, vol. 3, p. 4.

[23] Scott D Roth, "Ray casting for modeling solids," *Computer graphics and image processing*, vol. 18, no. 2, pp. 109–144, 1982.

[24] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun, "Realtime and robust hand tracking from depth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1106–1113.