Synchronized Video and Motion Capture Dataset and Quantitative Evaluation of Vision Based Skeleton Tracking Methods for Robotic Action Imitation

Selamawet Atnafu[1(✉)] and Conci Nicola[2]

[1] Bahir Dar Institute of Technology, Bahir Dar, Ethiopia
wselame7@gmail.com
[2] University of Trento, Trento, Italy
conci@disi.unitn.it

Abstract. Marker-less skeleton tracking methods are being widely used for applications such as computer animation, human action recognition, human robot collaboration and humanoid robot motion control. Regarding robot motion control, using the humanoid's 3D camera and a robust and accurate tracking algorithm, vision based tracking could be a wise solution. In this paper we quantitatively evaluate two vision based marker-less skeleton tracking algorithms (the first, Igalia's Skeltrack skeleton tracking and the second, an adaptable and customizable method which combines color and depth information from the Kinect.) and perform comparative analysis on upper body tracking results. We have generated a common dataset of human motions by synchronizing an XSENS 3D Motion Capture System, which is used as a ground truth data and a video recording from a 3D sensor device. The dataset, could also be used to evaluate other full body skeleton tracking algorithms. In addition, sets of evaluation metrics are presented.

Keywords: Joint angle · Accuracy · Tracking ability · Human motion dataset 3D camera · Ground truth

1 Introduction

The intention of making humanoid robots perform human like motions, which needs the development of easy and simple motion control approaches, has attracted the attention of many researchers [5]. Rather than using very complicated motion planning techniques, learning by demonstration, which combines motion capture and control systems, has being considered as an efficient and intuitive way to control the motion of humanoid robots and to teach them how to perform human like motions [5]. The first task in the process of learning by imitation is to have a robust and accurate motion capture system. Even if a number of marker-less skeleton tracking algorithms, which are convenient for this purpose, have been being proposed for years, there needs a way to quantitatively evaluate the performance of each method. Most of the evaluation schemas provide qualitative results due to lack of a common human motion dataset with a ground truth data [11]. The earliest efforts to collect synchronized motion

F. Mekuria et al. (Eds.): ICT4DA 2017, LNICST 244, pp. 150–158, 2018.
https://doi.org/10.1007/978-3-319-95153-9_14

capture and ground truth data have been used to analyze only 2D tracking methods [14]. Others also have tried to provide their synchronized dataset available to the public but some lack joint level ground truth information [15] and others do not provide calibration information [16]. HumanEva is the most recent and complete dataset which is made available to the public [11]. Since the subjects wore natural closing, on to which markers are attached, the accuracy of the ground truth data is reduced due to the movement of the markers [11]. In this work, two marker-less skeleton tracking methods are evaluated quantitatively. The two methods are chosen because they showed good tracking performances and are open sources which make them accessible for investigation. Both algorithms take input data from a 3D camera. RGB-D is a vision based method that takes depth and color information from the camera and outputs 3D location of human skeleton without requiring unnatural initialization poses [2]. It has two iterations, each one performing pixel-wise labeling, body part proposal and kinematic tree search steps. It doesn't involve any pre-processing of the input data [2]. While Skeltrack takes a buffer containing the depth image. From the depth image in the buffer a search for extreme points is performed [4]. Starting from those points the position for other joints is computed using heuristics and mathematics [4].

Skeltrack is implemented to track only upper body joints while, RGB-D has full body skeleton extraction ability [2, 4]. In this paper we are focused only on the upper body skeleton tracking. A common dataset of human motions is generated which is used as a ground truth data. Evaluation metrics and comparative results are presented here.

2 Related Works

Motion capture has been considered as a way for imitation learning of a robot from humans doing some motion in human-robot interaction/collaboration applications [5].

Marker based solutions, which are the most available real-time motion tracking methods in the market, are being used for most motion capture applications. In the work of Ott [6] and Shon [7] markers are attached to the body to get the measure of body position which is used by the tracking algorithm. Even though marker based systems provide a good tracking results, they are expensive and need the user to wear such a suit or a marker every time a person interacts with a robot which makes it inconvenient to use and difficult to maintain.

Also, the accuracy of joint positions depends on the precise placement of the markers [13]. Marker-less, Vision based methods on the other hand, are usually simple and do not need additional arrangements [13]. The research about this approach, would lead to one big step towards autonomous on line learning movements [8]. Before the introduction of low cost depth cameras, image based techniques have been used widely. But generally they have a drawback of segmenting foreground from background. Due to this they are effective for stationary camera and static backgrounds [2].

The advent of low cost depth sensors has initiated the study of depth-based methods to be considered as the latest tracking solutions. Real time and reliable pose estimation results, which are also robust to occlusions, have been found from tracking systems using multiple depth sensors [9]. But since our task is to use the data collected by the depth sensor on the robot's head, they are not very much convenient. Monocular kinect

based approaches by Shotton [1], OPenNI/Nite and Microsoft Solutions Microsoft Kinect SDK and Microsoft Kinect for windows are the recent and mostly used skeleton tracking solutions. Even though they show good performances, the source code is closed and do not let us do further researching and modifications.

In this study the most recent motion tracking solutions [2, 4] are explored to discover opportunities that push the performance of these methods towards accurate, stable and robust enough, with regard to joint position and joint angle calculations, to be used for robotic motion imitation applications.

3 Experimental Method

After exploring the two algorithms in detail, comparative experiment is done to see the performances quantitatively. Since a common dataset of human motions is required, we have recorded data from both the camera and from the sensor suit, which is considered as a ground truth data, by synchronizing the recordings in time. A small Asus Pro Live Xtion depth sensor is used to record a video data. The ground truth data is captured using the MVN motion capture suit. It consists of inertial sensors which are used to measure translation and orientation of body segments [12]. Although in our analysis we have only consider upper body motions, a full body tracking configuration is chosen to make the dataset useful for other researches. Then we have tested both on a common set of actions by extracting joint positions and joint angles.

The 3D camera was placed at a height of 1.8 m at a distance of one up to two and half meters from the subjects. Two subjects are participated to collect the motion dataset. It was necessary to take body measurements for the calibration of the MVN Moven Motion capture system. Table 1 shows the body segment dimension measures taken for each subject.

Table 1. Body dimensions for the two subjects

Body segment	Body segment measure (cm)	
	Subject 1	Subject 2
Body height	150	172
Shoe size	23	27
Arm span	142	161
Knee height	40	51
Hip height	70	82

We have implemented a Motion Sensor Suit Receiver which uses a UDP network streaming protocol. We have also implemented a software synchronization to start and stop both recordings simultaneously. Every recording has a length of 30–40 s. Both recordings from the camera and the suit are compressed as binary files and dumped on to the Hard Disk which helps to reduce the size of the files and save memory space. A synchronized play back code is also implemented here.

3.1 Synchronized Video and Motion Capture Dataset

In our dataset we have tried to include a wide variety of poses to show the performances of the algorithms in different cases. HumanEvaII provides a complex sequence of action, walking along an elliptical path, jogging and body balancing, which involve a full body motion [11]. Instead in our dataset even though complex sequence of actions is included, we have mainly focused on simple upper body motions which involve hand movements. We started from simple set of upper body motions that involve hand and head movements. Hands stretched to the sides, hands up, hands close to the body, waving one hand and both hands are some of the cases to mention. We took captures of poses with the subjects facing the camera from the side and turning to the back to evaluate the performances of the methods in unusual body positions. We have also included complex motions, such as walking, rotating the whole upper body part to the left and to the right with legs fixed, one hand pointing up and the other down and others. Each pose is repeated two times to insure that a backup is taken. In total we have collected 120 different sets of motion captures.

3.2 Data Processing

The two algorithms are made to run on the recorded binary data, which contains 3D information and camera intrinsic parameters for the calibration purpose, and the outputs are saved as text files. The files contain the 3D locations, given as x, y, z position, of each upper body joint. The data is then imported into Matlab for further processing and analysis. The joint positions are given relative to the global coordinate system of the Xtion sensor. While for the suit, during the calibration step, the global coordinate frame is fixed at the right heel point of the subject (Fig. 1).

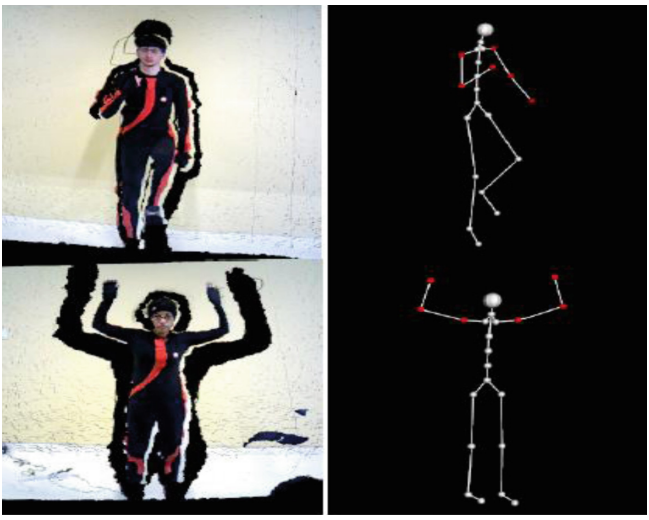

Fig. 1. A snapshot from point cloud play back and reconstructed skeleton from the recorded motion capture data

All the segment positions are computed with respect to this reference frame. Due to the different coordinate systems we did the comparison of the skeleton outputs based on the joint angle measures. From the 3D position of each joint, we have calculated joint angles using Matlab functions. Tracking of elbow and shoulder angles for both right and left side are computed and are plotted with respect to time.

$$\theta = atan2(norm(cross(vec_1, vec_2)),\ dot(vec_1, vec_2)) \tag{1}$$

3.3 Evaluation Metrics

Robustness of each method is calculated as the rate of detection of a joint or a body part over all frames of a recorded stream. We have used an L − 1 norm distance metric to compute joint angle errors and we presented joint angle error plots with respect to time. A list of average joint errors is also included to determine the accuracy of the methods with respect to joint angles we got from the suit. We did a statistical analysis on the joint angle error distribution and precision values are used to explain the tracking ability. Standard error of the mean is used as a way to measure the precision of each set of motions. This parameter measures how close the angle errors are distributed to the average joint angle error. The tracking ability is represented as precision values. It is calculated as:

$$SEM = \frac{SD}{\sqrt{N}} \tag{2}$$

Where, SD is the standard deviation which is computed as

$$SD^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - X_{mean})^2 \tag{3}$$

Where, N is the total number of frames.

4 Results and Discussion

Elbow and Shoulder joint angles for both hands are plotted with respect to time. In each of the plots below, elbow and shoulder joint angles together with angle deviations from the ground truth are listed. For both methods, Shoulder joint angles are better estimated than elbow angles. In addition, the plots illustrate the results in detail to give quantitative explanations. The tables list average joint angle errors for each joint for a video stream with 30–40 s length (Figs. 3, 4, 6 and 7).

In Fig. 2 joint angles are plotted for the subject performing two hands waving motion. This is the pose where the two methods show better performances and give slightly smaller joint angle errors. The reason for this is that for the skeltrack the pose allows detecting all the extreme points (the head and the two hands) and if so, all the rest joints can be determined correctly. For RGB-D also small body parts, hands and elbow points which have lesser chance of being detected, are further from the larger body part (the torso) and hence are detected correctly.

Fig. 2. Figure showing waving the two hands motion

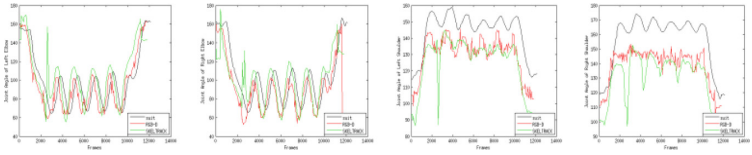

Fig. 3. Joint angle plot for waving the two hands

Fig. 4. A sequence of motions, Y pose, hands down, T pose, hands up, hands down

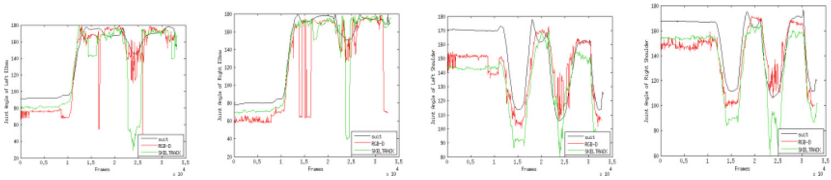

Fig. 5. Joint angle plot for a sequence of motions; Y pose, hands down, T pose, hands up, hands down

Fig. 6. Sequence of motions while turning to the back

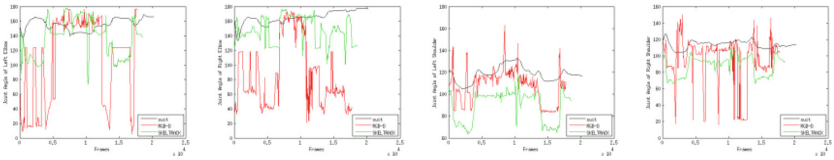

Fig. 7. Joint angle plot while turning to the back

In the second sequence of actions, Fig. 5, due to hands down and hands up poses, joint angle errors are increased for both methods.

The Tables 2 and 3, show that RGB-D gives lesser joint angle error and hence it is a more accurate tracking algorithm than Igalia's Skeltrack. But this works for the subject performing actions facing to the camera. For un-natural poses like, turning to the back and to the sides, RGB-D fails to estimate the pose. Here Skeltrack performance is better. This can be shown from Table 4.

Table 2. Joint angle estimation for two hands waving motion

Joint	Average angle error		Precision	
	RGB-D	Skeltrack	RGB-D	Skeltrack
Left elbow	13.1	18.08	1.99	1.18
Right elbow	18.85	20.63	1.84	1.61
Left shoulder	14.65	18.31	0.45	0.41
Right shoulder	19.29	26.42	0.8	0.54

Table 3. Joint angle estimation for a subject performing a sequence of actions; Y pose, hands down, T pose, hands up, hands down

Joint	Average angle error		Precision	
	RGB-D	Skeltrack	RGB-D	Skeltrack
Left elbow	14.52	18.48	1.17	0.79
Right elbow	10.42	21.26	1.2	0.81
Left shoulder	15.25	20.15	0.81	0.35
Right shoulder	15.75	15.42	2.04	0.52

In all the cases precision values for Skeltrack are lower than RGB-D. This clearly indicates that Skeltrack has better tracking ability. In the source codes RGB-D has no tracking implemented in to it. Each new frame is detected and the pose is estimated without the prior information from the previous frame.

Table 4. Joint angle estimations for the sequence of actions while turning to the back

Joint	Average angle error		Precision	
	RGB-D	Skeltrack	RGB-D	Skeltrack
Left elbow	54.45	31.61	3.44	2.03
Right elbow	73.29	18.96	2.69	1.18
Left shoulder	29.69	15.16	0.88	0.54
Right shoulder	27.64	26.22	2.04	0.52

5 Conclusion

In this paper we have generated a human motion dataset which can be used to quantitatively evaluate skeleton tracking methods. A 3D Asus Xtion camera to record the video and MVN sensor suit to collect the ground truth data are involved in the experiment. A quantitative evaluation scheme is also introduced. Finally two open source tracking algorithms are tested for the application of upper body robot action imitation task. The results demonstrate, RGB-D gives better angle estimations for a variety of poses but for those where the body is facing the camera. We observed a total failure of detection and pose estimation for unusual poses like turning to the side and to the back.

Regarding the tracking ability, skeltrack gives better results than RGB-D producing smooth and stable joint angles.

On average both have shown an angle error of 10–20° which make them hard to use directly for the robot motion control application. By incorporating a tracking algorithm in to the RGB-D method and by compensating the angle error, good results would be found. In the case of Skeltrack, incorporating markers (Inertial Sensors) at the extreme points would improve the accuracy and hence would produce good tracking results.

References

1. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real time human pose recognition in parts from single depth images. In: CVPR. Microsoft Research Cambridge and Xbox Incubation (2011)
2. Buys, K., Cagniart, C., Baksheev, A., De Laet, T., De Schutter, J., Pantofaru, C.: An adaptable system for RGB-D based human body detection and pose estimation. J. Vis. Commun. Image Represent. **25**, 39–52 (2014)
3. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision. Advances in Computer Vision and Pattern Recognition (ACVPR), pp. 70–98. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4640-7_5
4. Joaquim, R.: IgaliaSkeltrack

5. Luo, R.C., Shih, B.H., Lin, T.W: Real time human motion imitation of anthropomorphic dual arm robot based on cartesian impedance control
6. Ott, C., Lee, D., Nakamura, Y.: Motion capture based human motion recognition and imitation by direct marker control
7. Shon, P., Keith, G., Rao, P.N.: Robotic imitation from human motion capture using Gaussian processes
8. Azad, P., Ude, A., Asfour, T., Dillmann, R.: Stereo-based markerless human motion capture for humanoid robot systems. In: IEEE (2007)
9. Zhang, L., Sturm, J., Cremers, D., Lee, D.: Real-time human motion tracking using multiple depth cameras
10. The point cloud documentation. http://pointclouds.org/
11. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion
12. MVN Sensor suit Manual
13. Ong, A., Harris, I.S., Hamill, J.: The efficacy of a video-based marker-less tracking system for gait analysis. Comput. Methods Biomech. Biomed. Eng. **20**, 1089–1095 (2017)
14. Wang, P., Rehg, J.M.: A modular approach to the analysis and evaluation of particle filters for tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 790–797 (2006)
15. Gross, R., Shi, J.: The CMU motion of body (MoBo) database. Robotics Institute, Carnegie Mellon University, Technical report CMU-RI-TR-01-18 (2001)
16. CMU motion capture database. http://mocap.cs.cmu.edu/