Multimodal Motion Capture Dataset TNT15

Timo v. Marcard, Gerard Pons-Moll, Bodo Rosenhahn

January 2016
v1.1

Contents

1 Introduction

Video-based human motion capture has been a very active research area for decades now. The articulated structure of the human body, occlusions, partial observations and image ambiguities makes it very hard to accurately track the high number of degrees of freedom of the human pose. Recent approaches have shown, that adding sparse orientation cues from Inertial Measurement Units (IMUs) helps to disambiguate and improves full-body human motion capture. As a complementary data source, inertial sensors allow for accurate estimation of limb orientations even under fast motions.
In the research landscape of marker-less motion capture, publicly available benchmarks for video-based trackers (e.g. *HumanEva* [1], *Human3.6M* [2], TUM kitchen dataset [3]) generally lack inertial data. One exception is the *MPI08* dataset [4] published along [5] and [6], which provides inertial data of 5 IMUs along with video data.
This new dataset, called *TNT15*, consists of synchronised data streams from 8 RGB-cameras and 10 IMUs. In contrast to *MPI08* it has been recorded in a normal office room environment and the high number of 10 IMUs can be used for new tracking approaches or improved evaluation purposes. Four subjects perform five activities, namely *walking*, *running on the spot*, *rotating arms*, *jumping* and *punching*. In total, the *TNT15* dataset contains more than 4:30 minutes of video and sensor data, which amounts to almost 13 thousand frames at a frame rate of 50 Hz.
The *TNT15* dataset is freely available for your own tests and experiments. However it is restricted to research purposes only. If you use this data, please acknowledge the effort that went into data collection by citing the corresponding publication *Human Pose Estimation from Video and IMUs* [7].

2 Technical Recording Setup

The *TNT15* database was recorded in a normal office room environment. The cameras were arranged along the walls of the room and covered a recording volume of approximately $2m \times 2m \times 2,5m$.

2.1 Video Data

The video data has been recorded by a IMAGO VisionBox AGE-X Cluster and 8 RGB-cameras. The synchronized image streams were recorded at a frame rate of 50Hz, each having a resolution of $800px \times 600px$. Figure 1 shows a sketch of the camera setup.

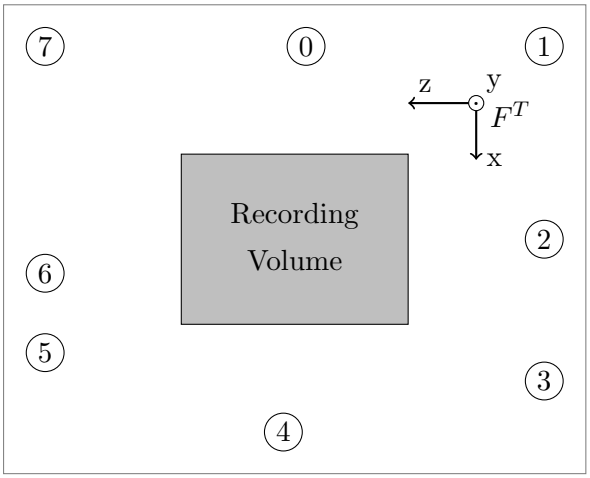

Figure 1: Sketch of the camera setup. Eight cameras (IDs 0 to 7) are positioned around the recording volume. Additionally, the common coordinate system F_T is illustrated. It is shifted outside of the recording volume for better visibility.

Figure 2: Sensor placement: 10 sensors are strapped to body extremities (shank, thigh, forearm, upper arm), chest and waist.

2.2 Inertial Data

Inertial data has been recorded by a XSens MTw Development Kit [8], consisting of a receiver and multiple synchronized wireless IMU sensors. We have used 10 sensors, which were strapped to the following body parts: shanks, thighs, lower arms upper arms, neck and hip. Each sensor provides acceleration and orientation data at a frame rate of 50Hz.

2.3 Calibration

The cameras have been calibrated to a common coordinate system F_T, which can be seen in Figure 1. We assume a standard pinhole camera model, consisting of internal and external camera parameters. The internal camera parameters were determined with a checkerboard pattern and the Mathworks® Camera Calibrator App. Radial distortion has also been removed during this step. In order to estimate the external camera parameters, we have used a calibration cube that was placed in the recording volume.
The MTw units provide orientation data relative to a static global inertial frame F^I, which is computed internally in each of the sensor units at the initial static position. It is defined as follows: the Z-axis is the negative direction of gravity measured by the internal accelerometer. The X-axis is the direction of the magnetic north pole measured by the magnetometer. Finally, the Y-axis is defined by the cross product $Z \times X$. For each sensor the absolute orientation data is provided by a stream of quaternions that define, at every frame, the map or coordinate transformation from the local sensor coordinate system to the global one $R^{IS}(t) : F^S \Rightarrow F^I$. In order to relate the sensor orientation measurements to the video coordinate system F_T, one has to determine the mapping between those coordinate systems. Since the Y-axis of the calibration cube for the tracking frame is perpendicular to the ground, the Y-axis of the tracking frame and the Z-axis of the inertial frame are aligned. Therefore, R^{TI} is a one parametric planar rotation that can be estimated beforehand using a calibration sequence [6]. We have avoided this calibration step by aligning the sensors with the tracking frame and performing a heading reset. This action basically rotates the inertial frame such that its X-axis is adjusted to the MTw units X-axis. We also considered a settling time of approx. 3 minutes in rest position before applying the heading-reset.

2.4 Synchronization

To synchronize the cameras with the IMU measurements, the actors were asked to perform a foot stamp at the beginning and end of every sequence. This motion is very prominent in the camera images and IMU acceleration data. Then, we simply synchronized the measurements by manually inspecting and aligning the data.

2.5 Subject Models

We provide surface meshes along this database. For each actor, a 3D laser scan has been captured before and after the measurements. They have a very high resolution and are stored in standard Wavefront '.obj' format. Additionally, we provide rigged, lower resolution meshes. For this purpose, we fitted a template mesh to the laser scans, placed a skeleton

and finally registered the mesh vertices with Pinocchio [9]. The rigged meshes are parametrised using twists and exponential maps as explained in [10], [11],citevonPon2016. Note: higher quality riggs will be made available soon.

2.6 Recording scheme

The overall recording procedure is summarised in the following list:

1. Camera & sensor calibration
2. Sensor instrumentation
3. Laser scan
4. Subject recordings
5. Record background for 10s
6. Repeat steps 2.-6. for every subject

3 Action Scripts

In order to capture multimodal motion data, the actors were asked to perform 5 activities. These activities range from rather simple motions such as *walking* and *running on the spot* to more complex motions such as *rotating arms*, *jumping* and *punching*.
The *walking* sequence consists of simple locomotion along a path with a 180° turn on the spot. In *running on the spot* the actors were asked to run on the spot at three different velocities. The *rotating arms* sequence contains forward, backward, synchronized and unsynchronized arm rotations, while *jumping* covers jumping jacks and skiing exercises. The *punching* sequence includes some dynamic boxing motions.
Each motion sequence is performed according to motion scripts, which are described in the following. All sequences start and end with the actor being in a so called scan-pose and a short foot stamp with the left foot. This foot stamp is used to synchronize video and IMU data.

0. Walking:
   (a) start at corner
   (b) scan-pose with left foot stamp
   (c) walk 2 steps
   (d) walk 2 steps while turning 180 to left
   (e) walk 2 steps

(f) walk 2 steps while turning 180 to left

(g) conclude with scan-pose and left foot stamp

1. Running on the spot:

   (a) start at center

   (b) scan-pose with left foot stamp

   (c) walk 6 steps on the spot (start with left foot)

   (d) jog 6 steps on the spot (start with left foot)

   (e) run 6 steps on the spot (start with left foot)

   (f) conclude with scan-pose and left foot stamp

2. Rotating arms:

   (a) start at center

   (b) scan-pose with left foot stamp

   (c) 4 forward rotations (both arms synchronized)

   (d) 4 backward rotations (both arms synchronized)

   (e) 4 forward rotations (both arms phase-shifted)

   (f) 4 backward rotations (both arms phase-shifted)

   (g) conclude with scan-pose and left foot stamp

3. Jumping jacks and skiing:

   (a) start at center

   (b) scan-pose with left foot stamp

   (c) 4 jumping jacks

   (d) 4 times skiing exercise / 8 jumps

   (e) conclude with scan-pose and left foot stamp

4. Dynamic punching:

   (a) start at center

   (b) scan-pose with left foot stamp

   (c) approx. 8 random punches

   (d) conclude with scan-pose and left foot stamp

<scene>	Action
00	Walking
01	Running on the spot
02	Rotating arms
03	Jumping jacks and skiing
04	Dynamic punching

Table 1: Action script numbering

4 Database Structure

The *TNT15* dataset is organized in the following directories:

- Images
- InertialData
- InputFiles
  - Models31Par
  - PoseInits31Par
  - CameraParameters
  - LaserScans
- Documentation

All recordings stored in the Images and InertialData directories are sorted by actor and scene. The four actors are denoted as <actor> $\in \{mr, pz, sg, sp\}$ and the action scripts (scenes) are numbered according to Table 1.

4.1 Image Data

The directory Images in the *TNT15* database contains RGB-videos, silhouette-videos and frame-wise silhouette image files. The file names are chosen according to the following naming conventions:

- RGB-video:
  TNT15_<actor>_<scene>_<camera>.mp4
- Silhouette-video:
  TNT15_<actor>_<scene>_<camera>.mp4
- Silhouette-images:
  TNT15_<actor>_<scene>_<camera>_<frame>_segmented.png.

The <camera>-variable denotes the two digits camera ID, see Fig. 1. The <frame>-variable is the five digits frame number, starting at 0. Note that in all images, radial distortion has been removed during the camera calibration procedure.
The silhouette files were generated by applying a background subtraction method based on a pixel-wise Gaussian model, similar to [12]. In order to keep the size of the *TNT15* database reasonably small we have excluded the background recordings. Please refer to the webpage of this database [13] to access this data.

4.2 Inertial Data

The directory InertialData in the *TNT15* database contains all calibrated inertial data, sorted by actor and scene. We provide sensor orientations in terms of quaternions, which map the sensor coordinate system to the global (video) coordinate system. Additionally, we have recorded acceleration measurements of the IMUs, denoted as acceleration and free acceleration. For the latter one, gravity has been removed from the acceleration data. The file names are chosen according to the following naming conventions:

- Quaternion data:
  `SensorQuat_<actor>_<scene>.txt`
- Acceleration data:
  `SensorAcc_<actor>_<scene>.txt`
- Free acceleration data:
  `SensorFreeAcc_<actor>_<scene>.txt`.

Within the inertial data files, each row lists the measurements of all sensors of a single time instance. The inertial data file header relates each column to a distinct sensor, which are named according to the closest parent joint of the human skeleton, see Fig. 2. The parent joint names are in line with the skeleton joint names in the provides mesh-models. Table 2 relates sensor placements to parent joints and the corresponding parent joint names.
Inertial data are recorded at the same frequency as the images (50Hz) and the first measurement in the inertial data files corresponds to the first frame of the respective image files.

4.3 Input Files

The directory InputFiles in the *TNT15* database contains camera projection matrices, actor models and respective initial poses for all sequences. Note that initial poses provide only a rough pose estimate of the first frame. The camera projection matrix file names follow the naming convention `proj<camera>.txt`. For each camera, the respective file contains the following matrices:

Body part	Parent joint	Sensor name
left shank	left knee	lknee
right shank	right knee	rknee
left thigh	left hip	lhip
right thigh	right hip	rhip
left forearm	left forearm	lelbow
right forearm	right forearm	relbow
left upper arm	left upper arm	lshoulder
right upper arm	right upper arm	rshoulder
chest	belly	belly
waist	root	root

Table 2: IMU sensor placement, their respective parent joints and sensor names in the corresponding inertial data files.

- P: 4×4 Total camera projection matrix ($P = [K \cdot M; 0_{1\times3}1]$)
- K: 3×3 Intrinsic camera parameters
- M: 3×4 Extrinsic camera parameters.

Revision History

Revision	Date	Author(s)	Description
1.0	22.01.16		initial version
1.1	11.01.16		added references to MPI08 papers and TNT15 paper; added usage guidelines to introduction
1.1	08.03.16		added references to TUM kitchen dataset and body model parametrization papers, fixed wrong references

References

[1] L. Sigal, A. Balan, and M. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal on Computer Vision (IJCV)*, vol. 87, no. 1, pp. 4–27, 2010.

[2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1325–1339, jul 2014.

[3] M. Tenorth, J. Bandouch, and M. Beetz, "The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 1089–1096, IEEE, 2009.

[4] "Multimodal human motion database MPI08." http://www.tnt.uni-hannover.de/project/MPI08_Database/.

[5] G. Pons-Moll, A. Baak, T. Helten, M. Muller, H.-P. Seidel, and B. Rosenhahn, "Multisensor-fusion for 3d full-body human motion capture," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010.

[6] A. Baak, T. Helten, M. Müller, G. Pons-Moll, B. Rosenhahn, and H. Seidel, "Analyzing and evaluating markerless motion tracking using inertial sensors," in *Proceedings of the 3rd International Workshop on Human Motion. In Conjunction with ECCV*, vol. 6553 of *Lecture Notes of Computer Science (LNCS)*, pp. 137–150, Springer, 2010.

[7] T. von Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and imus," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, 2016.

[8] "Xsens motion technologies." http://www.xsens.com/.

[9] I. Baran and J. Popović, "Automatic rigging and animation of 3d characters," in *ACM Transactions on Graphics (TOG)*, vol. 26, p. 72, ACM, 2007.

[10] G. Pons-Moll and B. Rosenhahn, "Ball joints for marker-less human motion capture," in *IEEE Workshop on Applications of Computer Vision (WACV),*, Dec. 2009.

[11] G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," in *Visual Analysis of Humans: Looking at People*, ch. 9, pp. 139–170, Springer, 2011.

[12] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 780–785, 1997.

[13] "Multimodal human motion database TNT15." http://www.tnt.uni-hannover.de/project/TNT15/.