UMPM benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction

N.P. van der Aa[1,2], X. Luo[1], G.J. Giezeman[1], R.T. Tan[1], R.C. Veltkamp[1]
[1]Utrecht University [2]Noldus Information Technology
n.vanderaa@noldus.nl, {x.luo, g.j.giezeman, r.t.tan, r.c.veltkamp}@uu.nl

Abstract

Analyzing human motion, including tracking and pose estimation, is a major topic in computer vision. Many methods have been developped in the past and will be developped in the future. To have a systematic and quantitative evaluation of such methods, ground truth data of the 3D human motion is scientifically required. Some publicly available data sets exist, like HumanEva, that provide synchronized video sequences with detailed ground truth 3D data for scenes limited to only a single person. However, for multiple persons, such a data set currently does not exist. In this paper, we present the Utrecht Multi-Person Motion (UMPM) benchmark, which includes synchronized motion capture data and video sequences from multiple viewpoints for multi-person motion including multi-person interaction. The data set is available to the research community to promote research in multi-person articulated human motion analysis. This paper describes the design of the benchmark, the technical problem solutions, and the resulting data sets.

1. Introduction

The development of articulated human motion analysis shows a wide spread of approaches. There are many possible models to describe the human body and its motion. Available solutions can be classified according to 2D or 3D human models [7], model-based or model-free [16], view-invariant or view-dependent [9], etc. Although most methods are restricted to one person only, recent advances also include the extension to multiple persons, including the challenges of occlusions by other people and objects. Examples can be found in multi-person tracking [14], and estimating poses from still images [6] and videos [11, 13].

The large amount of available methods makes systematic quantitative evaluation a requisite to determine how well a method performs compared to the state-of-the-art. For this purpose, data sets have been made publicly available, providing video sequences with ground truth 3D information. The most used example is the HumanEva data set [17, 18], where actions of a single person have been captured on video together with marker-based motion capture (MoCap) data. A similar data set for multiple persons should be provided to stimulate research for the multi-person case.

In this paper we introduce such a benchmark for multiple persons. This data set is called the UMPM benchmark and its general purposes are (1) to provide synchronized videos and MoCap data of multi-person scenarios, including multi-person interactions, and (2) to be used as a benchmark to evaluate multi-person motion capturing techniques. The extension to a multi-person benchmark is not trivial. If one person is present in the scene, only self-occlusions can deteriorate the results. With more persons in the scene, inter-person occlusions are unavoidable. In contrast to the HumanEva data set, our data set also includes static occluders in the scene like a table and a chair. Although the data set is primarily meant for multi-person articulated human motion capturing, the supplementary data such as background images and the assignment of the 3D MoCap data to a specific subject, ensures that this data set can also be used for background subtraction and tracking research in general.

The remainder of this paper is organized as follows. An overview of publicly available data sets is presented in Section 2. The design of our benchmark is provided in Section 3 and the data acquisition in Section 4. Section 5 discusses the main challenges of creating the benchmark. Section 6 provides guidelines for using this data set. Finally, the main limitations are discussed in Section 7.

2. Related work

Human motion analysis includes detection and tracking, which is a requisite step for pose estimation. Many data sets are available for evaluating people tracking methods, e.g. the PETS benchmarks [4]. However, these data sets only provide ground truth data about a central point or a bounding box. For pose estimation, this is too restrictive.

Several benchmarks do exist for pose estimation. Such

Name	Year	No. cameras	Frame rate	Resolution	No. subjects	No. frames	No. sequences	Ground truth
CMU-MoBo [8]	2001	6	30	640 × 480	25	200,000	100	-
IXMAS [21]	2006	5	23	390 × 291	11	-	39	reconstructed volumes
HumanEva [18]	2007	7[1]	60	640 × 480	4	80,000	56	Vicon MoCap data *12 cams, 195 mpp*[3]
CMU-MMAC [2]	2009	6	30-60	640 × 480 / 1024 × 768	43	-	-	Vicon MoCap data *16 cams, 40 mpp*
MuHAVi-MAS [3]	2009	8	25	720 × 576	14	-	119	Manual annotations
TUM Kitchen [19]	2009	4	25	384 × 288	4	-	20	Markerless MoCap data
MPI08 [15]	2010	8	40	1004 × 1004	-	24,000	54	3D laser scans
UMPM benchmark	2011	4[2]	50	644 × 484	30	400,000	36	Vicon MoCap data *14 cams, 37 mpp*

[1] The first HumanEva dataset is recorded with 4 color and 3 greyscale cameras. The second HumanEva dataset only uses the four color cameras.
[2] The 4 color cameras do not face each other directly to avoid similar sillouettes. [3] *cams*: cameras and *mpp*: markers per person.

Table 1. Properties of multicamera benchmarks for pose estimation and gesture recognition.

benchmarks should have (1) sufficiently high resolution images to capture details, (2) a high frame rate to detect movements, and (3) multiple cameras to see a subject from varying view points. Table 1 provides an overview of the multicamera benchmarks. All data sets assume a controlled environment to facilitate detection and tracking: the lighting conditions remain static and the only object changing the scenery is the person appearing in the scene. In that case, simple background subtraction methods [23] can be applied to detect the person. Since these benchmarks contain rough gestures like waving, jumping, etc., the requirements on synchronisation between video recordings and the 3D ground truth information are not that urgent. The synchronization in the HumanEva data set [17, 18] was done by software in the first part and by hardware in the second, while the MuHAVi-MAS data set [3] has no explicit synchronization at all. However, if the movements become faster like in a fight, or subtle like in sign language, the synchronization should be done better.

An important difference in the data sets is the way they provide ground truth 3D information. Since positions of body parts should be measured, a logical choice is to use a MoCap system to provide the ground truth. The HumanEva data set and CMU Multi-Modal Activity Database (CMU-MMAC) [2] obtain MoCap data captured by a Vicon system[1]. This is an industry standard for optical marker-based motion capturing. The system uses infrared cameras to recover the 3D positions of reflective markers attached to the subject. HumanEva used twelve 1.3 megapixel cameras and CMU-MMAC used twelve 4 megapixel cameras. The placements of the markers (typically 20-60) are positioned to measure the 3D position of the entire human body. Some benchmarks provide alternative ground truth 3D information. For example, the Multimodal Motion Capture Indoor Dataset (MPI08) [15] uses 3D laser scans of the human body. The TUM Kitchen Data Set [19] provides MoCap data extracted from videos using their own markerless full-body MeMoMan tracker [5].

The main drawback of the state-of-the-art benchmarks for pose and gesture recognition including MoCap data is that they are restricted to one person only. There are data sets for multiple persons with MoCap data such as the CMU Graphics Lab Motion Capture Database [1], the data set used by Liu [12] and the Stereo Pedestrian Detection Evaluation Dataset [10]. However, the first one provides only MoCap data and no video, the second one provides MoCap data for one person only in a two person interaction scenario and the last one is aimed at pedestrians only.

Our UMPM benchmark uses a triangular setup of four color video cameras with a resolution of 644 × 484 pixels at 50 fps that are synchronized with a Vicon system consisting of 14 four megapixel cameras at 100 fps, to optimally capture information of a multi-person environment.

3. Benchmark design

To simultaneously capture video and 3D MoCap data, while keeping the ”natural” appearance of the subjects, the subjects are equipped with reflective markers attached to the clothes using transparent ribbon. This implies that the markers are not as tight as when MoCap suits are used. The subjects are represented as natural as possible with respect to motion and visual appearance. Therefore, no restrictions have been made on clothing, haircut, make-up, etc., as long as the markers are detectable (e.g. subjects do not wear any shiny objects like glasses). Since there is no unlimited access in variety of persons, we depend on the availability of persons with a specific race, length, body size, etc. The participants are master students, PhD students and staff members of the Multimedia & Geometry group of the

[1] http://www.vicon.com/

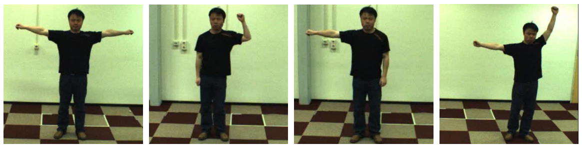


Figure 1. Example poses used in the synthetic motions.

Informatics and Computer Science Department of Universiteit Utrecht. Participation in the collection process was voluntary and each subject was required to read, understand and sign a consent form for collection and distributation of data. A copy of the consent form is available by writing the authors. Subjects were informed that the data, including video images, would be made available to the research community and could appear in scientific publications.

The recordings should capture as much variation in poses and gestures as possible and include challenging environment settings to ensure that this data set is representative for the real world. The recordings should also show the main challenges of multi-person motion, which are visibility (a (part of a) person is not visible because of occlusions by other persons or static objects, or by self-occlusions) and ambiguity (body parts are identified ambiguously when persons are close to each other). The body poses and gestures are classified as natural (commonly used in daily life) and synthetic (special human movements for some particular purpose such as human-computer interaction, sports or gaming). Each of these two classes is subdivided into a few scenarios. In total, our data set consists of 9 different scenarios. Each scenario is recorded with 1, 2, 3 and 4 persons in the scene and is recorded multiple times to provide variations, i.e. different subject combination, order of poses and motion patterns. For natural motion we defined 5 different scenarios where the subjects (1) walk, jog and run in an arbitrary way among each other, (2) walk along a circle or triangle of a predetermined size, (3) walk around while one of them sits or hangs on a chair, (4) sit, lie, hang or stand on a table or walk around it, and (5) grab objects from a table. These scenarios include individual actions, but the number of subjects moving around in the restricted area cause inter-person occlusions. We also include two scenarios with interaction between the subjects: (6) a conversation with natural gestures, and (7) the subjects throw or pass a ball to each other while walking around. The scenarios with synthetic motions include poses as shown in Figure 1, performed when the subjects (8) stand still and (9) move around. These scenarios are recorded without any static occluders to focus only on inter-person occlusions.

Before starting any actual action, all subjects perform a T-pose at their starting position. In this pose all markers and body parts are maximally visible, which ensures a proper way to initialize the motion capturing. To check the synchronization of the video cameras, one person claps his hands before the scenario starts. At the end of each scenario, each subject returns to its starting position, one person claps again to obtain a second check of the synchronization, and each subject adopts the T-pose again.

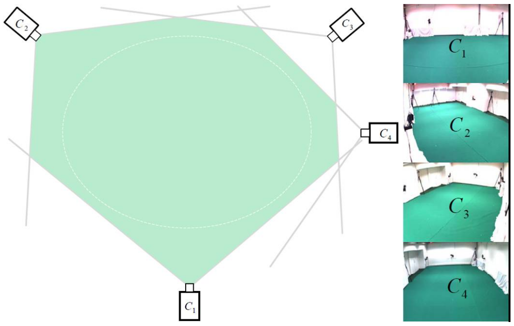


Figure 2. Top-view of the camera placement of the four monocular color cameras (*left*) and the camera views (*right*).

4. Acquisition data

To capture the video sequences, the room is equipped with 4 Basler PiA A640-210-gc color cameras with a resolution of 644 × 484 and a frame rate of maximal 210 fps[2]. The cameras are placed such that 3 cameras form an equilateral triangle together with the fourth camera as shown in Figure 2. This choice ensures that (1) the cameras surround the acquisition area and the subjects, (2) there is sufficient overlap between the cameras' field of views, and (3) the cameras do not face each other directly to avoid similar silhouettes. The cameras have a wide angle lens (3.5 mm) to capture wider views such that subjects can perform closer to the camera and the size requirement of the scene is reduced. Although a wide angle lens shows more radial distortion than a normal lens, this can be corrected. Example images from the videos are shown in Figure 3.

The ground truth 3D data is captured with a Vicon MoCap system, which consists of 8 Vicon MX-40+ cameras (4 megapixel resolution, maximum speed of 160 fps, infrared), and 6 Vicon MX-F40 cameras (4 megapixel resolution, maximum speed of 370 fps, near-infrared). The actual recording speed is set to 100 fps, which is four times the frame rate of the color cameras. Both the color and Vicon cameras have been mutually synchronized by a hardware module of the Vicon system (Ultranet HD). The Vicon system identifies the 3D position of the reflective markers attached to the subjects. In our UMPM benchmark the markers are positioned as illustrated in Figure 4. For each person we used 37 markers: 3 around the head; 2 around the neck; 4 around the waist; 1 on the shoulder; 3 around each elbow; 3 around each wrist; 1 on the outside of the hip; 3 around

[2] http://www.baslerweb.com/

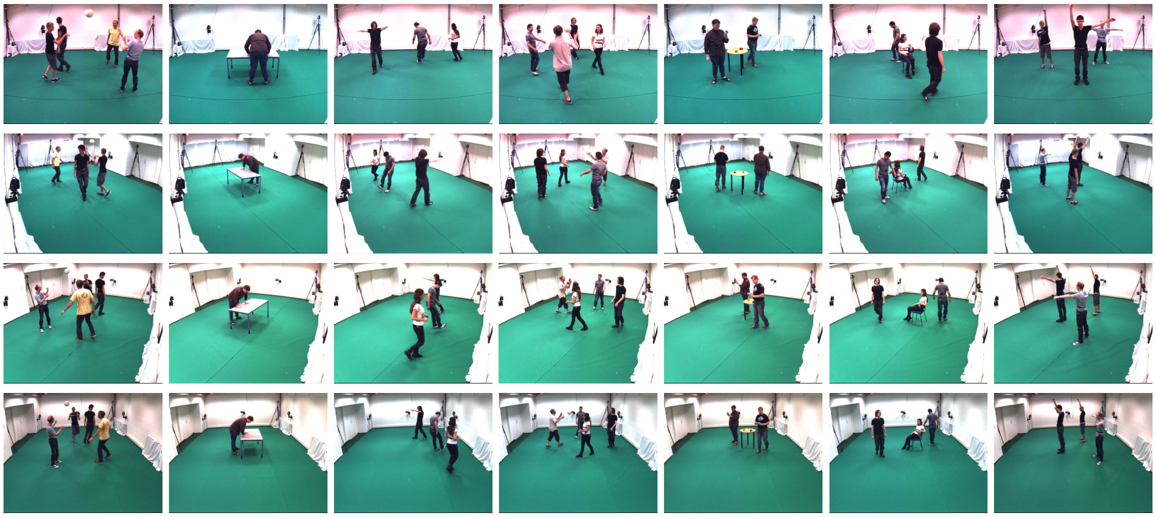


Figure 3. Example data from the UMPM benchmark where each column is a screen shot of a scenario taken by all four color cameras.

each knee; and 3 around each ankle. The positioning of the markers is customized to handle inter-person occlusions. Each joint of the human body (wrist, ankle, neck, etc.), except the shoulder and tigh, is measured by more than one marker. For example, the wrist has three markers to indicate the center. If a marker is occluded, other markers for this joint might still be detected.

Each scene setup and camera placement need proper calibration to relate the camera views to the 3D world, including accurate estimation of the cameras intrinsic/extrinsic parameters and precise alignment of the global origin. The Vicon cameras are calibrated by the Vicon calibration wand (a tool with 5 markers) and the calibration function embedded in the Vicon Nexus software [3]. To calibrate the video cameras, the conventional checkerboard-based calibration method [24] is used. To align the global origin of both camera systems, the wand is placed on top of the checker-board such that the zero-coordinates of both tools are located at the same spot, and the coordinate axes overlay each other.

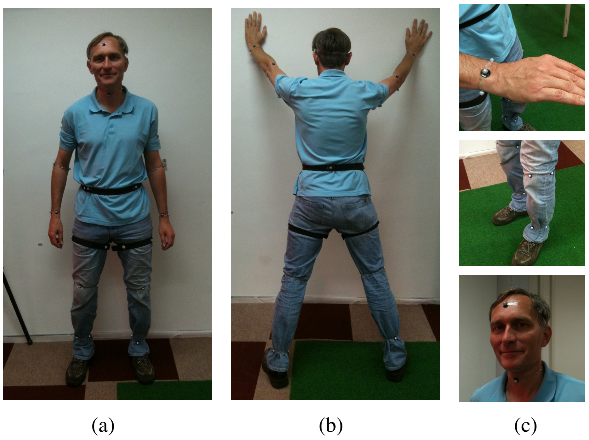
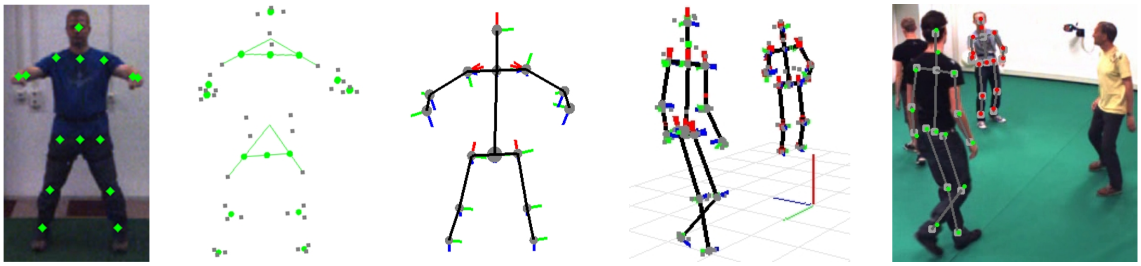


Figure 4. Placement of the reflective markers: frontal view (a), back view (b) and detailed views of the wrist, the knees and ankles, and the head and nek (c).

The Nexus software combines the data from the Vicon cameras to obtain 3D marker positions. After manually assigning labels to the markers in one frame, the Nexus software reconstructs the trajectories and assigns labels to marker positions in the other frames. This reconstruction is not always correct and relabeling is necessary. This implies that (1) missing markers should be identified, (2) erroneous measurements should be removed, and (3) each marker should get a label of the body part and the person it belongs to. Manually checking the marker positions - typically half a million per recording- is a non-trivial task. We developed an approach to subsequently (1) check the continuity of trajectories, meaning that a sudden change in the speed of a label points to a place that needs special attention (this reduces the number of positions to inspect to about 100), (2) assign labels to marker positions that were not yet assigned a label (anonymous markers), where we iteratively check for discontinuities, relabel manually and label anonymous markers until no serious discontinuities are left, and finally, (3) interpolate trajectories and throw away anonymous markers. This procedure is repeated until no significant improvements are found. The 3D marker positions, labels and the way these labels are found, are made available in the appropriate fields in a C3D file structure[4].

Next to the 37 marker positions per subject, we also pro-

[3] http://www.cacs.louisiana.edu/labs/ecrg/vicon/
[4] http://www.c3d.org/index.html



(a) (b) (c) (d) (e)

Figure 5. Estimating the subjects' joints ground truth. (a) Example frame superimposed with virtual markers (green circles). (b) Estimating the virtual markers (green circles) by the 3D coordinates of the reflective markers (gray squares). (c) The kinematically constrained human skeleton model. (d) Virtual markers (colored squares) drive the skeletons (black bones and gray joints). (e) The virtual markers (colored circles) and the skeletons (gray lines and circles) superimposed on one example view.

vide two sets with 15 virtual 3D positions to describe the bone joints, namely head, neck, shoulders, elbows, wrists, pelvis, tighs, knees and ankles (Figure 5(a)). In the first set, the joint positions (except for the shoulders and tighs) are computed by averaging the coordinates of the corresponding markers, since these are placed around the joints. The corresponding mean is therefore ”inside” the body part, which properly represents the bone joints. The shoulders and pelvis joints are estimated by the positions of the neck, waist and corresponding limbs. The left pelvis joint is approximated as a point on the line between the center of the waist and the marker on the left upper leg, and depends on the subjects pelvis width (Figure 5(b)). Similarly, the center of the neck and the marker on the upper arm define the shoulder coordinates. Finally, the means of the pelvis joints and shoulder joints define the pelvis and chest centers, respectively. The second set adds a kinematically constrained human skeleton model (23 degrees of freedom) to the 15 virtual joints of the first set to overcome the problem of moving markers caused by the deformation of muscles or moving clothes. It uses the Cyclic-Coordinate Descent (CCD)-based inverse kinematics approaches [22] (Figure 5(c)). The 3D coordinates of the virtual joints drive the skeleton (Figure 5(d)). The bone lengths of the skeleton are scaled according to the corresponding positions of the virtual joints in the first frame. The skeleton imposes strict poses and joint position constraints, which makes the estimation of the joint ground truth more robust against moving markers and measurement errors (Figure 5(e)).

5. Challenges

The main challenge in creating the UMPM benchmark is to handle intra-/inter-person and other occlusions. A consequence of such occlusions is that some markers cannot be detected. However, to provide complete ground truth 3D information of the subjects, we must identify and locate these missing markers. To reduce the possibility of missing markers, we adopted a different setting of the markers (see Figure 4). Although the probability increases to find a marker for each joint, these marker positions have to be translated to the actual joint position, which is defined as the average of the marker positions of this joint. Hence, if a marker is missed, this average position still has to be estimated. Another consequence concerns the labelling and tracking of each marker. The closer markers are to each other, the higher the probability that a marker is labelled wrongly. The problem deteriorates if some markers are missed.

The shoulder joint must be approximated from the upper arm and neck markers. However, simply averaging the marker positions is not adequate, since a moving upper arm does not imply a moving shoulder joint. Thus, a more advanced model is needed. Similarly, the pelvis joint is approximated using the upper leg and waist markers.

6. Usage of data set

This benchmark is meant to evaluate human motion capturing for multiple subjects in a similar way as HumanEva does for a single subject. In [17] an evaluation measure has been introduced to compare algorithms with the provided ground truth. The benchmark provides four synchronized color videos as input of the articulated human motion capture method. The ground truth 3D information is available in 3 formats: (1) 37 marker positions per subject, (2) 15 joint positions obtained directly by averaging the marker positions, and (3) 15 corrected joint postions by enforcing kinematic constraints. The user may choose which one is used, but any pose recovery/tracking method has to translate its results to either one of these formats.

To facilitate the use of this data set, we provide some additional material. First, to match the video sequences to the ground truth, the calibration parameters for each camera are given, namely (1) internal and (2) external camera parameters, and (3) distortion parameters. Second, we provide backround images for methods that rely on back-

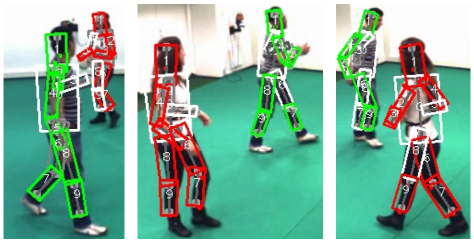


Figure 6. Per view ccclusion detection. The white and colored trapezoid represent occluded and non-occluded parts, respectively.

ground subtraction. Third, extensive documentation concerning this benchmark is available in a technical report [20]. The video recordings, C3D files, calibration parameters, background images, documentation and software of the UMPM benchmark can be downloaded from the following URL: http://www.projects.science.uu.nl/umpm/.

The benchmark can also be used to focus on specific parts of an algorithm. An example is the visibility measure for each body part and camera, which was first presented in [13] for tracking purposes. The kinematic skeletons are used to model the motion and poses, where the visibility is computed for each body part for a selected camera view. The result is shown in Figure 6. The best visibility view is expected to provide robust cues for locating a body part.

7. Discussion

Although this data set is created with great care, it has its limitations. The quality of the video images is determined by the hardware used. The wide angle lenses widened the captured scene, but also increased radial distortion.

The marker positions were chosen to detect inter-person occlusions. Although the current setting of the markers is an improvement of the single person setting, capturing the shoulder and pelvis joint positions might be improved further. Therefore, for each measured marker position, we indicate in the C3D file if it is measured and labeled by Nexus software, or corrected by our software. In this way, the users may decide which marker positions they want to use.

8. Acknowledgement

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO).

References

[1] CMU GLMC. http://mocap.cs.cmu.edu/.

[2] CMU MMAC. http://kitchen.cs.cmu.edu/.

[3] MuHAVi-MAS. http://dipersec.king.ac.uk/MuHAVi-MAS/.

[4] PETS 2010. http://pets2010.net/.

[5] J. Bandouch, F. Engstler, and M. Beetz. Accurate human motion capture using an ergonomics-based anthropometric human model. In *AMDO*, 2008.

[6] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010.

[7] D. Gavrilla. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, January 1999.

[8] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, 2001.

[9] X. Ji and H. Liu. Advances in view-invariant human motion analysis: A review. *SMC-C*, 40(1):13–24, January 2010.

[10] P. Kelly, N. O'Connor, and A. Smeaton. A framework for evaluating stereo-based pedestrian detection techniques. *CirSysVideo*, 18(8):1163–1167, August 2008.

[11] M. Lee and R. Nevatia. Human pose tracking using multi-level structured models. In *ECCV*, 2006.

[12] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011.

[13] X. Luo, R. Tan, and R. Veltkamp. Multi-person tracking based on vertical reference lines and dynamic visibility analysis. In *ICIP*, 2011.

[14] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[15] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3D full-body human motion capture. In *CVPR*, 2010.

[16] R. Poppe. Vision-based human motion analysis: An overview. *CVIU*, 108(1-2):4–18, 2007.

[17] L. Sigal, A. Balan, and M. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, March 2010.

[18] L. Sigal and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[19] M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *THEMIS*, 2009.

[20] N. van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp. Utrecht multi-person motion (umpm) benchmark. Technical Report UU-CS-2011-027, Utrecht University, 2011.

[21] D. Weinland, R. Ronfarda, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3):249–257, 2006.

[22] C. Welman. Inverse kinematics and geometric constraints for articulated figure manipulation. Master's thesis, Simon Fraser University, April 1993.

[23] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.

[24] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, November 2000.