HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion

Leonid Sigal and Michael J. Black

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-06-08
September 2006

HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion

Leonid Sigal **Michael J. Black**
Department of Computer Science
Brown University
Providence, RI 02912
{ls,black}@cs.brown.edu

Technical Report CS-06-08

Abstract

While research on articulated human motion and pose estimation has progressed rapidly in the last few years, there has been no systematic quantitative evaluation of competing methods to establish the current state of the art. Current algorithms make many different choices about how to model the human body, how to exploit image evidence and how to approach the inference problem. We argue that there is a need for common datasets that allow fair comparison between different methods and their design choices. Until recently gathering ground-truth data for evaluation of results (especially in 3D) was challenging. In this report we present a novel dataset obtained using a unique setup for capturing synchronized video and ground-truth 3D motion. Data was captured simultaneously using a calibrated marker-based motion capture system and multiple high-speed video capture systems. The video and motion capture streams were synchronized in software using a direct optimization method. The resulting HumanEva-I dataset contains multiple subjects performing a set of predefined actions with a number of repetitions. On the order of 50,000 frames of synchronized motion capture and video was collected at 60 Hz with an additional 37,000 frames of pure motion capture data. The data is partitioned into training, validation, and testing sub-sets. A standard set of error metrics is defined that can be used for evaluation of both 2D and 3D pose estimation and tracking algorithms. Support software and an on-line evaluation system for quantifying results using the test data is being made available to the community. This report provides an overview of the dataset and evaluation metrics and provides pointers into the dataset for additional details. It is our hope that HumanEva-I will become a standard dataset for the evaluation of articulated human motion and pose estimation.

1 Introduction

The recovery of articulated human motion and pose from video has been studied extensively in the past 20 years with the earliest work dating to the early 1980's [10, 22]. A variety of statistical [1, 2, 3, 6, 12, 36, 37, 38] as well as deterministic methods [21, 41, 35] have been developed for tracking people from single [1, 2, 7, 12, 15, 20, 21, 25, 26, 27, 29, 37] as well as multiple [3, 6, 9, 36] views. All these methods make different choices regarding the state space representation of the human body and the image observations required to infer this state from the image data. Despite clear advances in the field, evaluation of these methods remains mostly

heuristic and qualitative. As a result, it is difficult to evaluate the current state of the art with any certainty or even to compare different methods with any rigor. This technical report describes a database of human motion and evaluation metrics that can be used to address these problems.

Quantitative evaluation of human pose estimation and tracking is currently limited due to the lack of common datasets with "ground truth" with which to test and compare algorithms. Instead qualitative tests are widely used and evaluation often relies on visual inspection of results. This is usually achieved by projecting the estimated 3D body pose into the image (or set of images) and visually assessing how the estimates explain the image [6, 7, 27]. Another form of inspection involves applying the estimated motion to a virtual character to see if the movements appear natural [38]. The lack of the quantitative experimentation at least in part can be attributed to the difficulty of obtaining 3D ground-truth data that specifies the true pose of the body observed in video data.

To obtain some form of ground truth, previous approaches have resorted to custom action-specific schemes (or tricks); e.g. motion of the arm along the circular plate of known diameter [14]. Alternatively synthetic data has been extensively used [1, 2, 9, 35, 38] for quantitative evaluation. With packages such as POSER (*e frontier*, Scotts Valley, CA), semi-realistic images of humans can be rendered and used for evaluation. Such images, however, typically lack realistic camera noise, often contain very simple backgrounds and provide simplified types of clothing. While synthetic data allows quantitative evaluation (given known 3D pose), current datasets are still too simplistic to capture the complexities of natural images of people and scenes.

For 2D human pose/motion estimation, quantitative evaluation is more common and typically uses hand labeled data [12, 25, 26]. While quantitative evaluation does occur, the datasets are typically unique to each research group, preventing direct comparison of methods. Furthermore, for both 2D and 3D methods, no standard error metrics exist and results are reported in a variety of ways which prevent direct comparison; e.g. average root-mean-squared (RMS) angular error, silhouette overlap, joint center distance, etc.

Here we describe a database of human activity with associated ground truth that can be used for quantitative evaluation and comparison of both 2D and 3D methods. We hope that the creation of this database, which we call HUMANEVA-I[1], will advance the state of the art in human motion and pose estimation by providing a structured, comprehensive, development dataset with support code and quantitative evaluation metrics. The motivation behind the design of the HUMANEVA-I dataset is that, as a research community, we need to answer the following questions:

- What is the state-of-the art in human pose estimation?
- What is the state-of-the art in human motion tracking?
- What algorithm design decisions effect human pose estimation and tracking performance and to what extent?
- What are the strengths and weaknesses of different pose estimation and tracking algorithms?
- What are the main unsolved problems in human pose estimation and tracking?

In answering these questions, comparisons must be made across a variety of different methods and models to find which choices are most important for a practical and robust solution. To support this analysis, the HUMANEVA-I database contains a number of subjects performing 3 repetitions (trials) of a varied set of predefined actions. The database is broken into training, validation, and test datasets (for which the ground truth data is withheld and an on-line evaluation system is made available). A set of error metrics is defined and made available as part

[1]The "I" in HUMANEVA-I is an acknowledgment that the current database has limitations and what we learn from this first database will most likely lead to improved database in the future. Consequently we leave open the possibility of future, improved, HUMANEVA datasets.

of the dataset. These error metrics are general enough to be applicable to most current pose estimation and tracking algorithms and body models. Support software for manipulating the data and evaluating results is also made available as part of the HumanEva-I dataset. This support code shows how the data and error metrics can be used and provides an easy-to-use Matlab interface to the data. This allows different methods to be fairly compared using the same data and the same error metrics.

In systematically addressing the problems of articulated human pose estimation and tracking using the HumanEva-I database, other related research areas may benefit as well, such as, for example, foreground/background segmentation, appearance modeling and voxel carving.

It is worth noting that similar efforts have been made in related areas including the development of datasets for face detection [23, 24], human gait identification [31] and dense stereo vision [32]. These efforts helped advance the state-of-the-art in the respective fields. Our hope is that the HumanEva-I dataset will lead to similar advances in articulated human pose and motion estimation.

2 Related work

Classically the solutions to articulated human motion estimation fall into two categories: pose estimation and tracking. **Pose estimation** is usually formulated as the inference of the articulated human pose from a single image (or in a multi-view setting, from multiple images captured at the same time). **Tracking**, on the other hand, is formulated as inference of the human pose over a set of consecutive image frames throughout the image sequence. Tracking approaches tend to make strong assumptions about existence of the initial pose of the body at the first frame and only concern themselves with evolution of this pose over time. In recent years progress has been made towards combining these two sets of approaches [36, 38], such that tracking can benefit from automatic initialization and failure recovery in the form of static pose estimation and pose estimation in turn can benefit from temporal coherence constraints.

It is important to note that both tracking and pose estimation can be performed in 2D, 2.5D, or 3D corresponding to different ways of modeling the human body. In each case, the body is typically represented by an articulated set of parts corresponding naturally to body parts (limbs, head, hands, feet, etc.). Here 2D refers to models of the body that are defined directly in the image plane or in the world plane parallel to the image plane. 2.5D approaches tend to model the body in the image plane but also allow the model to have relative depth information. Finally 3D refers to approaches that model the human body as a 3-dimensional structure which is often composed of simplified parts represented as cylinders or superquadrics. A short summary of different approaches with evaluation and error metrics employed (when appropriate) can be seen in Table 1; for a more complete taxonomy, particularly of older work, we refer readers to [8] and [19].

3 Summary of the data

The HumanEva-I database consists of 4 subjects performing a set of 6 predefined actions three times (twice with video and motion capture, and once with motion capture alone). A description of the actions is provided in Table 2. The dataset includes separate training, validation and test sets. To test the generalization ability of algorithms, we provide some test sets for which the actions are not in the training or validation set. We also withhold all the activities of one subject for testing. Details of the data and evaluation methods are provided in Section 6.

Participation in the collection process was voluntary and each subject was required to read, understand, and sign an Institutional Review Board (IRB) approved consent form for collection and distribution of data.[2]

[2] A copy of the consent form for the "Video and Motion Capture Project" is available by writing to the authors.

Table 1: Short summary of the human motion and tracking algorithms. Works are listed in the chronological order by the first author. *Type* refers to the type of the approach, where (P) corresponds to the pose-estimation and (T) to tracking. Approaches that employ ($\star$) and ($\star\star$) evaluation metrics are consistent with the evaluation metrics proposed in this paper.

Year	First Author	Model Type	Parts	Dim	Type	Evaluation	Metric
1983	Hogg [10]	Cylinders	14	2.5	T	Qualitative	
1996	Ju [13]	Patches	2	2	T	Qualitative	
1996	Kakadiaris [14]	D Silhouettes	2	3	T	Quantitative	
1998	Bregler [5]	Ellipsoids	10	3	T	Qualitative*	
2000	Rosales [30]	Stick-Figure	10	3	P	Synthetic	$\star$[a]
2000	Sidenbladh [34]	Cylinders	2/10	3	T	Qualitative	
2002	Ronfard [29]	Patches	15	2	P	Hand Labeled	
2002	Sidenbladh [33]	Cylinders	2/10	3	T	Qualitative	
2003	Grauman [9]	Mesh	N/A	3	P	Synthetic/POSER	$\star$
2003	Ramanan [26]	Rectangles	10	2	T,P	Hand Labeled	$\diamond$
2003	Shakhnarovich [35]	Mesh	N/A	3	P	Synthetic/POSER	$\ddagger$
2003	Sminchisescu [39, 40]	Superquadric Ellip.	15	3	T	Qualitative[b]	
2004	Agarwal [1, 2]	Mesh	N/A	3	P	Synthetic/POSER	†
2004	Deutscher [6]	R-Elliptical Cones	15	3	T	Qualitative	
2004	Lan [16]	Rectangles	10	2	T,P	Qualitative	
2004	Mori [21]	Stick-Figure	9	3	P	Qualitative	
2004	Roberts [28]	Prob. Template	10	2	P	Qualitative	
2004	Sigal [36]	R-Elliptical Cones	10	3	T,P	Motion Capture	$\star\star$
2005	Balan [3]	R-Elliptical Cones	10	3	T	Motion Capture	$\star\star$
2005	Felzenszwalb [7]	Rectangles	10	2	P	Qualitative	
2005	Hua [12]	Quadrangular	10	2	P	Hand Labeled	$\natural$
2005	Lan [15]	Rectangles	10	2	P	Motion Capture	$\star$
2005	Ramanan [25]	Rectangles	10	2	T,P	Hand Labeled	$\diamond$
2005	Ren [27]	Stick-Figure	9	2	P	Qualitative	
2005	Sminchisescu [38]	Mesh	N/A	3	T,P	Synthetic/POSER	†
2006	Lee [17]	R-Elliptical Cones	5/10	3	T,P	Hand Labeled	$\star\star$[c]
2006	Li [18]	R-Elliptical Cones	10	3	T	Motion Capture	$\star\star$
2006	Sigal [37]	Quadrangular	10	2	P	Motion Capture	$\star$

$\star$ - Mean squared distance in 2D between the set of $M = 15$ virtual markers corresponding to the joint centers and limb ends. Measured in pixels (*pix*).

$D(X, \hat{X}) = \frac{1}{M} \sum_{i=1}^{M} \parallel x_i - \hat{x}_i \parallel$, where $x_i \in R^2$ is location of 2D marker i, and $X = \{x_1, x_2, ..., x_M\}$.

$\star\star$ - Mean squared distance in 3D between the set of $M = 15$ virtual markers corresponding to the joint centers and limb ends. Measured in millimeters (*mm*).

$D(X, \hat{X}) = \frac{1}{M} \sum_{i=1}^{M} \parallel x_i - \hat{x}_i \parallel$, where $x_i \in R^3$ is location of 3D marker i, and $X = \{x_1, x_2, ..., x_M\}$.

† - Root mean square (RMS) error in joint angle. Measured in degrees (*deg*).

$D(\theta, \hat{\theta}) = \frac{1}{M} \sum_{i=1}^{M} |(\theta_i - \hat{\theta}_i) mod \pm 180^\circ|$, where $\theta \in R^M$ is the pose in terms of joint angles.

$\ddagger$ - Normalized error in joint angle. Measured as a fraction from 0 to 1.

$D(\theta, \hat{\theta}) = \sum_{i=1}^{M} 1 - cos(\theta_i - \hat{\theta}_i)$, where $\theta \in R^M$ is the pose in terms of joint angles.

$\diamond$ - Pixel overlap threshold results in binary 0/1 detection measure.

$\natural$ - Mean distance from 4 endpoints of quadrangular shape representing the limb.

[a] Error units were in fraction of the subject's height.

[b] While only qualitative analysis of the overall tracking performance was presented. A quantitative analysis of the number of minima in the posterior was given.

[c] Additional per limb weighting is applied to downweight the error proportionally to the size of the limb.

Action	Description
Walking	Subjects walked in an elliptical path at the edge of the capture space.
Jog	Subjects jogged (slow running) in an elliptical path at the edge of the capture space.
Gesture	Subjects were instructed to perform "hello" and "good-bye" gestures in repetition.
Throw/Catch	Subjects tossed and caught a baseball with the help of the lab assistant (who stood outside the capture volume). Subjects were instructed to explore a variety of styles (e.g. overhead throw, under arm).
Box	Subjects imitated boxing. No instruction was given on how this action should be performed.
Combo	Subjects were instructed to perform a series of actions that consisted of walking followed by jogging and then balancing on each one of two feet. The series of actions were performed in sequence without interruption.

Table 2: Summary of the 6 actions exhibited in the HumanEva-I dataset. For actions that did not require subjects to move in the capture space (Gesturing, Throw/Catch, Box), with subjects performed these standing roughly in the center of the viewing volume and facing toward camera $C1$ (see Figure 1).

To simultaneously capture video and motion information, our subjects wore natural clothing (as opposed to motion capture suits, as is often done for pure motion capture sessions) on which reflective markers were attached using invisible adhesive tape. Our motivation was to obtain "natural" looking image data that contains all the complexity posed by moving clothing. One negative outcome of this is that the markers tend to move more than they might with a tight-fitting motion capture suit. As result, our ground truth motion capture data may not always be as accurate as that obtained by more traditional methods; we felt that the trade-off of accuracy for realism here was acceptable. We have applied minimal post-processing to the motion capture data, steering away from the use of complex software packages (e.g. Motion Builder) that may introduce biases or alter the motion data in the process. As a result, our motion capture data for particular frames in some sequence may be missing markers or may be mislabeled. This results in invalid poses for these frames. We made every effort to detect such cases and exclude them from quantitative comparison. Note that the presence of markers on the body may also alter the *natural* appearance of the body. Given that the marker locations are known, it would be possible to provide a pixel mask in each image covering the marker locations; these pixels could then be excluded from further analysis. We felt this was unnecessary since the markers are often barely noticeable in the video data and hence will likely have an insignificant impact on the performance of image-based tracking algorithms.

Example images of two different subjects jogging and boxing are shown in Figure 3. Data from 7 synchronized video cameras is shown with an overlay of ground truth motion. More detailed descriptions of the collection process and the hardware employed are given in the next section.

Subjects were informed that the data, including video images, would be made available to the research community and could appear in scientific publications.

4 Data collection

4.1 Hardware

Ground truth motion of the body was captured using a commercial motion capture (MoCap) system from ViconPeak (`http://www.vicon.com/`). The ViconPeak MoCap system is an industry standard for optical marker-based motion capture and has been successfully employed in a variety of entertainment applications for over 10 years. The system uses reflective markers and six 1M-pixel cameras to recover the 3D position of the markers and thereby estimate the 3D articulated pose of the body.

Video data was captured using two commercial video capture systems. One from Spica Technology Corporation (`http://www.spicatek.com/`) and one from IO Industries (`http://www.ioindustries.com/`). The Spica system captured video using four Pulnix (`http://www.pulnix.com/`) TM6710 grayscale cameras. These were grayscale progressive scan cameras with 644x488 resolution and a frame rate of up to 120 Hz. The IO Industries system used three UniQ (`http://www.uniqvision.com/`) UC685CL 10-bit color cameras with 659x494 resolution and a frame rate of up to 110 Hz. The raw frames were re-scaled from 659x94 to 640x480 by IO Industries software. To achieve better image quality under natural indoor lighting conditions both video systems were set up to capture at 60 Hz. The rough relative placement of cameras is illustrated in Figure 1.

The motion capture system and video capture systems were not synchronized in hardware, and hence a software synchronization was employed. The synchronization and calibration procedures are described in the next sections. Example images from the HumanEva-I database are shown in Figure 3. The appearance of the 4 subjects is illustrated in Figure 2.

4.2 Calibration

The motion capture system was calibrated using Vicon's proprietary software and protocol. Calibration of the intrinsic parameters for the two video capture systems was done using a standard checker-board calibration grid and the Camera Calibration Toolbox for Matlab [4]. Focal length ($F_c \in \mathbb{R}^2$), principle point ($C_c \in \mathbb{R}^2$) and radial distortion coefficients ($K_c \in \mathbb{R}^5$) were estimated for each camera $c \in \{BW1, BW2, BW3, BW4, C1, C2, C3\}$. We assumed that pixels were square and let the skew $\alpha_c = 0$ for all c.

The extrinsic parameters corresponding to the rotation, $R_c \in SO(3)$, and translation, $T_c \in \mathbb{R}^3$, of the camera with respect to the global (shared) coordinate frame were optimized over using a semi-automated procedure to align the global coordinate axis of each video camera with the global coordinate axis of Vicon motion capture system. A single moving marker was captured by the video cameras and the motion capture system for a number of frames (> 1000) at the same time. The resulting 3D tracked position of the marker $\Gamma_t^{(3D)}$, $t \in \{1..T^{(3D)}\}$ was recovered using the Vicon software. The 2D position of the marker in video, $\Gamma_t^{(2D)}$, $t \in \{1..T^{(2D)}\}$, was recover using a Hough circle transform [11] that was manually initialized at the first frame and subsequently tracked. Notice that $T^{(3D)} \neq T^{(2D)}$ because the video and motion capture systems are not synchronized in hardware and have different frame rates.

The projection of the 3D marker position $f(\Gamma_t^{(3D)}; R_c, T_c)$ onto the image was then optimized directly for each camera by minimizing

$$\min_{R_c,T_c,A_c,B_c} \sum_{t=1}^{T^{(2D)}} \delta(t; A_c, B_c) \|\Gamma_t^{(2D)} - f(\Gamma_{t*A_c+B_c}^{(3D)}; R_c, T_c)\|^2. \quad (1)$$

In addition to optimizing over R_c and T_c we also optimized over the relative temporal scaling, $A_c \in \mathbb{R}$, between the video and Vicon cameras, and the temporal offset $B_c \in \mathbb{R}$. In doing so

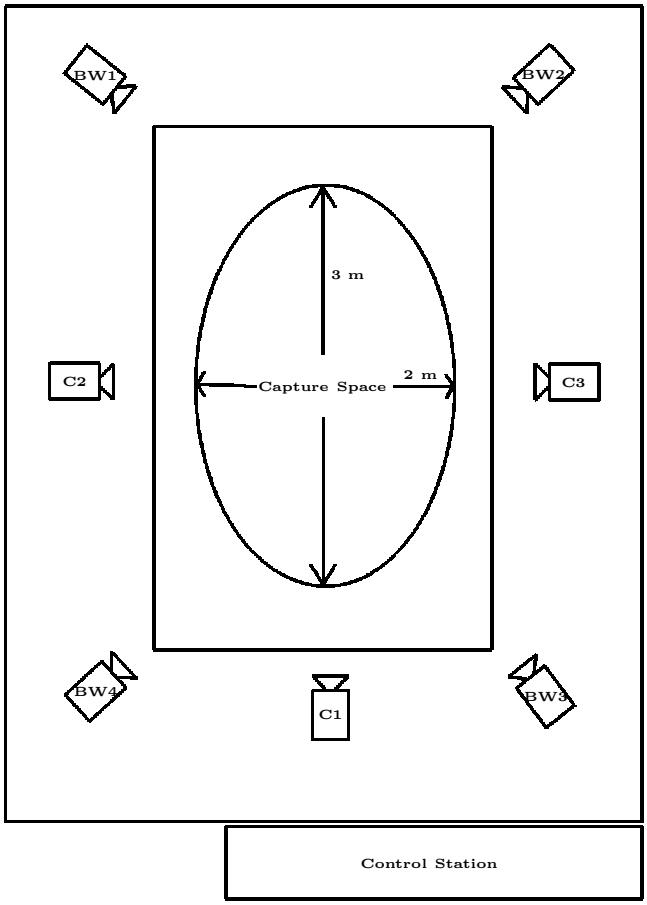


Figure 1: Camera setup for the HumanEva-I data acquisition. The bird's eye view sketch is shown with rough dimensions of the capture space and the placement of 7 video cameras. The color cameras are labeled $C1$, $C2$, $C3$ and the grayscale cameras by $BW1$ to $BW4$.

we assume that the temporal scaling is constant[3] and hence there is no temporal drift. The 3D position $f(\Gamma^{(3D)}_{t*A_c+B_c}; R_c, T_c)$ is linearly interpolated to cope with non-integer indices $t*A_c+B_c$. Finally, in Eqn. (1), $\delta(t; A_c, B_c)$ is defined as:

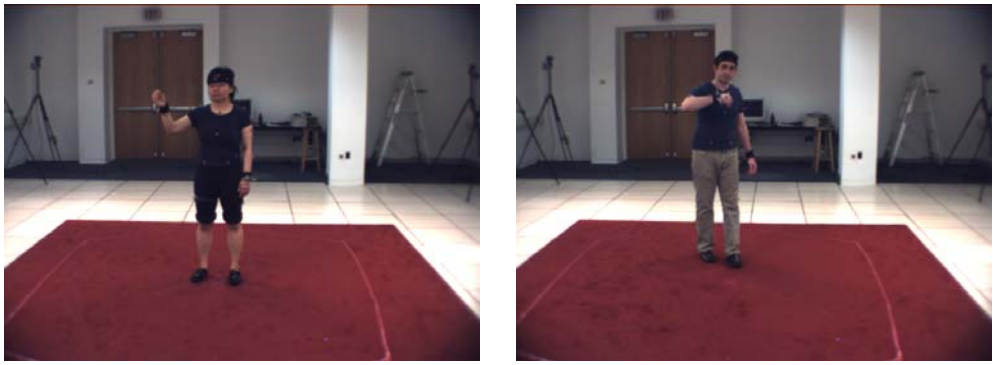
$$\delta(t; A_c, B_c) = \begin{cases} 0 & if \quad t*A_c+B_c > T^{(3D)} \\ 0 & if \quad t*A_c+B_c < 1 \\ 1 & otherwise. \end{cases} \tag{2}$$

4.3 Synchronization

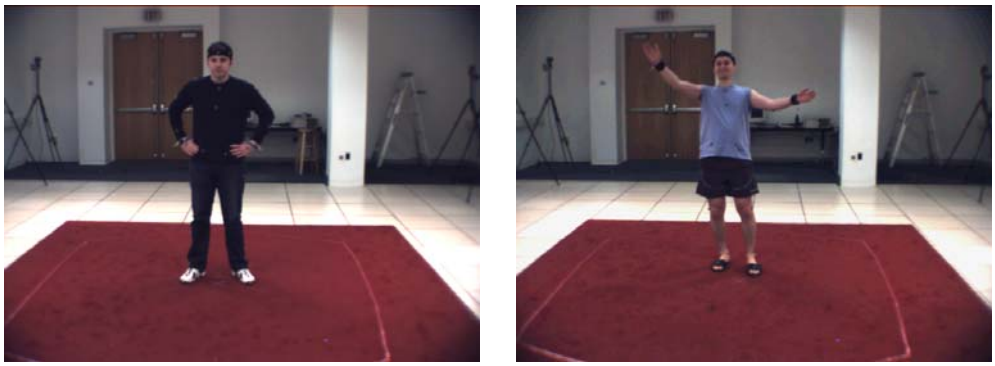
While extrinsic calibration parameters and temporal scaling, A_c, can be estimated once per camera (so long as the setup is not moved and Vicon system is not re-calibrated[4]), the temporal

[3] In practice $A_c \approx 2$ since the frame rate of motion capture system is roughly 120 Hz and video system is 60 Hz.

[4] Calibration of the Vicon motion capture system changes the global coordinate frame and hence requires re-calibration of extrinsic parameters of the video cameras as well.



(S1) (S2)



(S3) (S4)

Figure 2: Sample images of the 4 subjects used in the HumanEva-I dataset. Notice that appearance of the subjects varies significantly in the type, style, and color of clothing. We had 3 male and 1 female subject.

offset B_c will be different for every sequence captured. To synchronize the motion capture and the video in software we manually labeled visible markers on the body for a small sub-set of images (6 images were used with generally a couple of marker positions labeled per frame). These labeled frames can be used in the optimization procedure above but with fixed values for R_c, T_c, and A_c to recover a least squares estimate of the temporal offset B_c for every sequence captured.

5 Evaluation Metrics

Various evaluation metrics have been proposed for human motion tracking and pose estimation. For example, a number of papers have suggested using joint-angle distance as the error measure (see Table 1). This measure however assumes a particular parameterization of the human body and cannot be used to compare methods where the body models have different degrees of freedom or have different parameterizations of the joint angles.

For this dataset we aim to define an error measure that will be (1) widely applicable and (2) relatively fast to compute. Hence, we propose an error measure based on the sparse set of virtual markers that correspond to the locations of joints and limb endpoints and that can uniquely encode the pose of the body. This error metric was first introduced for 3D pose estimation and

tracking in [36] and later extended in [3]. It has since been also used for 3D tracking in [18] and for 2D pose estimation evaluation in [15, 37].

Assuming that we can represent the pose of the body using M virtual markers, we can write the state of the body as $X = \{x_1, x_2, ...x_M\}$, where $x_m \in \mathbb{R}^3$ (or $x_m \in \mathbb{R}^2$ if a 2D body model is used) is the position of the marker m in the world (or image respectively). Notice, that converting from any standard representation of the body pose to X is trivial. The error in estimated pose $\hat{X}$ to the ground truth pose X can then be expressed as the average absolute distance between individual markers,

$$D(X, \hat{X}) = \sum_{m=1}^{M} \frac{\parallel x_m - \hat{x}_m \parallel}{M}. \tag{3}$$

To ensure that we can compare algorithms that use different numbers of parts we add a binary selection variable per-marker $\hat{\Delta} = \{\hat{\delta}_1, \hat{\delta}_2, ...\hat{\delta}_M\}$ obtaining the final proposed error metric,

$$D(X, \hat{X}, \hat{\Delta}) = \sum_{m=1}^{M} \frac{\hat{\delta}_m \parallel x_m - \hat{x}_m \parallel}{\sum_{i=1}^{M} \hat{\delta}_i}, \tag{4}$$

where $\hat{\delta}_m = 1$ if the proposed algorithm is able to recover marker m, and 0 otherwise.

For the sequence of T frames we can compute the average performance and the standard deviation of the performance using the following:

$$\mu_{seq} = \frac{1}{T} \sum_{t=1}^{T} D(X_t, \hat{X}_t, \hat{\Delta}_t), \tag{5}$$

$$\sigma_{seq} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} [D(X_t, \hat{X}_t, \hat{\Delta}_t) - \mu_{seq}]^2}. \tag{6}$$

Since many tracking algorithms are stochastic in nature, an average error and the standard deviation computed over a number of runs is most useful. As a convention from previous methods [3, 15, 37, 36] that have already used this error metric we compute the 3D error in millimeters (mm) and 2D error directly in the image in pixels (pix).

Notice that for algorithms that model the posterior distribution using a unimodal distribution over the pose of the body the mean over the posterior is likely to give a good estimate for the pose X. For others, however, that may model the posterior using distributions that can be multi-modal, better results can be obtained by choosing $\hat{X}$ to be, for example, the most likely sample from the posterior. This is discussed in greater detail in [3]. Alternative error metrics that compute lower-bounds for sample- or kernel-based representations of the posterior are discussed in [3].

6 Dataset structure

The HumanEva-I dataset contains 4 subjects performing a set of 6 actions each in 3 separate trials (two with synchronized motion and video and one with motion capture alone). All video data was captured in uncompressed format and later compressed using XviD codec[5] (version 1.1.0) to make web distribution of the video data practical. The directory structure for the dataset is as follows:

[5] An open source video codec that is based on MPEG-4. Both XviD and DivX evolved from the Mayo open source project; however, DivX became a commercial product, while XviD (DivX backwards) is distributed under the GPL license. For more information, visit `www.xvid.org`.

⟨Subject⟩/Image_Data/ - Contains image/video data. Each AVI file in the directory has the following name structure, ⟨Action⟩_⟨Trial⟩_(⟨Camera⟩).avi.

⟨Subject⟩/Mocap_Data/ - Contains Vicon motion capture data in C3D format (http://www.c3d.org). Each C3D file in directory has the following name structure, ⟨Action⟩_⟨Trial⟩.c3d.

⟨Subject⟩/Calib_Data/ - Contains image/video calibration data. These human-readable text .CAL files can be read using the packaged software and contain intrinsic and extrinsic calibration parameters for each video camera.

⟨Subject⟩/Sync_Data/ - Contain synchronization data between the image and the motion capture streams expressed in the human-readable text .OFS files (one for every camera, action and trial). Each OFS file in the directory has the following name structure, ⟨Action⟩_⟨Trial⟩_(⟨Camera⟩).ofs.

Background/ - Contains 3 sets of background template videos that were taken in the beginning, at the end and in the middle of a week-long capture session. These videos can be used to derive rough silhouettes of the body. In some actions foreign objects are used and interactions with a lab assistant are required, in these cases silhouettes will be poor. These challenging scenarios however will correspond to realistic real-world imaging conditions.

Release_Code/ - Matlab sample code for loading and using the data. Please carefully study this code. A GUI for viewing the dataset and the code for the synchronized motion capture and video viewer is provided, as well as example applications for loading the data, and computing the error using the proposed error metrics. See the README file in this directory for details.

Release_Docs/ - Documentation materials for the *HumanEva-I* dataset.

In the above ⟨___⟩ designate variable names. The valid values for all mentioned variable names are given bellow.

6.1 Training, validating, and testing

For convenience and fairness the *HumanEva-I* data is broken down into three disjoint sub-sets: training, validation, and testing. The details of how the dataset is broken into the three sub-sets is given in Table 4. For training and validation sets the motion capture is available and error

Variable	Valid (string) value
⟨Subject⟩	∈ {S1, S2, S3, S4}
⟨Action⟩	∈ {Box, Combo, Gesture, Jog, ThrowCatch, Walking}
⟨Trial⟩	∈ {1, 2, 3}
⟨Camera⟩	∈ {C1, C2, C3, BW1, BW2, BW3, BW4}

metrics are provided for the evaluation. Users of this dataset are free to use these as they like though we recommend developing or training algorithms on the training set and saving the validation set for initial testing. For quantitative evaluation of the test set ground-truth motion capture data is withheld and an on-line evaluation tool is provided instead. The purpose of this is to prevent parameter tuning on the test dataset[6].

Notice, that trial 3 contains only motion capture data and is intended to be used by groups interested in learning motion priors. Trial 1 contains synchronized motion capture and video and is broken (equally) into validation and training segments. In all cases trial 2 is reserved in its entirety for testing, and the motion capture for those trials is withheld[7].

The data and the on-line evaluation is available from the project web page `http://www.cs.brown.edu/research/vision/humaneva/`.

Sequence			Set Partition			Description		
Subject	Action	Trial	Validate	Train	Test	FPS	Video	MoCap
S1	Walking	1	$1-590$	$591-1180$	-	60 Hz	$\surd$	$\surd$
S1	Walking	2	-	-	$1-980$	60 Hz	$\surd$	W
S1	Walking	3	-	$(1-3238)$	-	120 Hz	$\times$	$\surd$
S1	Jog	1	$1-367$	$368-735$	-	60 Hz	$\surd$	$\surd$
S1	Jog	2	-	-	$1-856$	60 Hz	$\surd$	W
S1	Jog	3	-	$(1-3175)$	-	120 Hz	$\times$	$\surd$
S1	Throw/Catch	1	$1-473$	$474-946$	-	60 Hz	$\surd$	$\surd$
S1	Throw/Catch	2	-	-	$1-929$	60 Hz	$\surd$	W
S1	Throw/Catch	3	-	$(1-3453)$	-	120 Hz	$\times$	$\surd$
S1	Gesture	1	$1-395$	$396-790$	-	60 Hz	$\surd$	$\surd$
S1	Gesture	2	-	-	$1-1059$	60 Hz	$\surd$	W
S1	Gesture	3	-	$(1-2127)$	-	120 Hz	$\times$	$\surd$
S1	Box	1	$1-385$	$386-770$	-	60 Hz	$\surd$	$\surd$
S1	Box	2	-	-	$1-607$	60 Hz	$\surd$	W
S1	Box	3	-	$(1-1653)$	-	120 Hz	$\times$	$\surd$
S1	Combo	2	-	-	$1-2602$	60 Hz	$\surd$	W
Total (S1)			2268	2266/(13095)	7172			
S2	Walking	1	$1-438$	$439-877$	-	60 Hz	$\surd$	$\surd$
S2	Walking	2	-	-	$1-1097$	60 Hz	$\surd$	W
S2	Walking	3	-	$(1-1523)$	-	120 Hz	$\times$	$\surd$
S2	Jog	1	$1-398$	$399-796$	-	60 Hz	$\surd$	$\surd$
S2	Jog	2	-	-	$1-733$	60 Hz	$\surd$	W
S2	Jog	3	-	$(1-1573)$	-	120 Hz	$\times$	$\surd$
S2	Throw/Catch	1	$1-550$	$551-1101$	-	60 Hz	$\surd$	$\surd$
S2	Throw/Catch	2	-	-	$1-1346$	60 Hz	$\surd$	W

Table 4 Continued on next page –

[6]We will log evaluations and, if a group is abusing the test set in a manner consistent with parameter tuning, we reserve the right to restrict their access to the test set and future versions of the HumanEva database.

[7]For fairness, the authors of this document will follow the same procedure as all other users and will only access the test data through the on-line service. Logs of our access will be made available upon request (in fact, we may make logs of all access to the evaluation data visible on the website).

Table 4 – continued from previous page

S2	Throw/Catch	3	-	$(1-3340)$	-	120 Hz	$\times$	$\surd$
S2	Gesture	1	$1-500$	$501-1000$	-	60 Hz	$\surd$	$\surd$
S2	Gesture	2	-	-	$1-1025$	60 Hz	$\surd$	W
S2	Gesture	3	-	$(1-3551)$	-	120 Hz	$\times$	$\surd$
S2	Box	1	$1-382$	$383-765$	-	60 Hz	$\surd$	$\surd$
S2	Box	2	-	-	$1-975$	60 Hz	$\surd$	W
S2	Box	3	-	$(1-3108)$	-	120 Hz	$\times$	$\surd$
S2	Combo	2	-	-	$1-1996$	60 Hz	$\surd$	W
Total (S2)			2210	2206/(13646)	7033			
S3	Walking	1	$1-448$	$449-896$	-	60 Hz	$\surd$	$\surd$
S3	Walking	2	-	-	$1-806$	60 Hz	$\surd$	W
S3	Walking	3	-	$(1-2358)$	-	120 Hz	$\times$	$\surd$
S3	Jog	1	$1-401$	$402-803$	-	60 Hz	$\surd$	$\surd$
S3	Jog	2	-	-	$1-842$	60 Hz	$\surd$	W
S3	Jog	3	-	$(1-1973)$	-	120 Hz	$\times$	$\surd$
S3	Throw/Catch	1	$1-493$	$494-987$	-	60 Hz	$\surd$	$\surd$
S3	Throw/Catch	2	-	-	$1-967$	60 Hz	$\surd$	W
S3	Throw/Catch	3	-	$(1-2074)$	-	120 Hz	$\times$	$\surd$
S3	Gesture	1	$1-533$	$534-1067$	-	60 Hz	$\surd$	$\surd$
S3	Gesture	2	-	-	$1-554$	60 Hz	$\surd$	W
S3	Gesture	3	-	$(1-1789)$	-	120 Hz	$\times$	$\surd$
S3	Box	1	$1-512$	$513-1024$	-	60 Hz	$\surd$	$\surd$
S3	Box	2	-	-	$1-719$	60 Hz	$\surd$	W
S3	Box	3	-	$(1-1573)$	-	120 Hz	$\times$	$\surd$
S3	Combo	2	-	-	$1-1761$	60 Hz	$\surd$	W
Total (S3)			2387	2385/(9767)	5649			
S4	Walking	2	-	-	$1-670$	60 Hz	$\surd$	W
S4	Jog	2	-	-	$1-593$	60 Hz	$\surd$	W
S4	Throw/Catch	2	-	-	$1-776$	60 Hz	$\surd$	W
S4	Gesture	2	-	-	$1-462$	60 Hz	$\surd$	W
S4	Box	2	-	-	$1-577$	60 Hz	$\surd$	W
S4	Combo	2	-	-	$1-1105$	60 Hz	$\surd$	W
Total (S4)					4183			
Total			6865	6857/(36508)	24037			

Table 4: HUMANEVA-I composition of the Training, Validation and Testing sets. "Set Partition" refers to the frames in the specified range that disjointly corresponds to either one of the three sets. W indicates that the motion capture data is Withheld for evaluation.

7 Background subtraction

Since the majority of the current pose estimation and tracking algorithms make use of silhouette features, background images have been collected to allow foreground/background segmentation of the scene. For each camera 3 sets of the background images have been collected; before, after and in the middle of a week-long capture session. Since some parts of the scene are non-rigid and may have moved slightly over the session (e.g. the carpet may have moved due to the forces applied by the legs) all 3 sets should be used for a more robust background segmentation. Obtaining good silhouettes in the case of color cameras is easier because the background is less ambiguous; in the grayscale cameras good silhouette segmentation is challenging.

Software for learning a per-pixel background model using a mixture of Gaussians and then segmenting the image into foreground and background layers is provided as part of the software package. While the background distribution modeled by a mixture density at every pixel is learned from the collected background images using EM, the foreground object (person) is assumed to have uniform distribution over all colors. This results in the classification criterion for foreground/background.

8 Baseline algorithm

A baseline particle filtering algorithm [3] for tracking human pose with ground-truth initialization is under development and will be provided for quantitative performance comparison when available at a future date.

Acknowledgments. This project was supported in part by gifts from Honda Research Institute and Intel Corporation. We would like to thank Ming-Hsuan Yang, Rui Li, Alexandru Balan and Payman Yadollahpour for help in data collection and post-processing. We also would like to thank Stan Sclaroff for making the color video capture equipment available for this effort.

References

[1] A. Agarwal and B. Triggs. Learning to track 3D human motion from silhouettes. *International Conference on Machine Learning (ICML)*, pp. 9–16, 2004.

[2] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 882–888, 2004.

[3] A. Balan, L. Sigal and M. Black. A quantitative evaluation of video-based 3D person tracking. *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pp. 349–356, 2005.

[4] J.-Y. Bouguet. Camera calibration toolbox for Matlab. *http://www.vision.caltech.edu/bouguetj/calib_doc/*.

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8–15, 1998.

[6] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV)*, 61(2):185–205, 2004.

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55-79, Jan. 2005.

[8] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–98, 1999.

[9] K. Grauman, G. Shakhnarovich and T. Darrell. Inferring 3D structure with a statistical image-based shape model. *IEEE International Conference on Computer Vision (ICCV)*, pp. 641–648, 2003.

[10] D.C. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, vol. 1, pp. 5–20, 1983.

[11] P. V. C. Hough. Method and means for recognizing complex patterns. *U.S. Patent 3,069,654*, 1962.

[12] G. Hua, M.-H. Yang and Y. Wu. Learning to estimate human pose with data driven belief propagation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 747–754, 2005.

[13] S. Ju, M. Black and Y. Yacoob. Cardboard people: A parametrized model of articulated motion. *International Conference on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[14] I.A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 81–87, 1996.

[15] X. Lan and D. Huttenlocher. Beyond trees: Common factor models for 2D human pose recovery. *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp. 470–477, 2005.

[16] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 722–729, 2004.

[17] M. Lee and R. Nevatia. Human pose Tracking using multi-level structured models. *European Conference on Computer Vision (ECCV)*, vol. 3, pp. 368–381, 2006.

[18] R. Li, M.-H. Yang, S. Sclaroff and T.-P. Tian. Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers. *European Conference on Computer Vision (ECCV)*, 2006.

[19] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding (CVIU)*, 18:231–268, 2001.

[20] G. Mori. Guiding model search using segmentation. *IEEE International Conference on Computer Vision (ICCV)*, pp. 1417–1423, 2005.

[21] G. Mori, X. Ren, A. Efros and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 326–333, 2004.

[22] J. O'Rourke and N.I. Badler. Model-Based Image Analysis of Human Motion Using Constraint Propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2:6, pp. 522–192, 1980.

[23] P. J. Phillips, D. Blackburn, M. Bone, P. Grother, R. Micheals and E. Tabassi. Face recognition vendor test. *http://www.frvt.org/*, 2002.

[24] P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[25] D. Ramanan, D. Forsyth and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 271–278, 2005.

[26] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 467–474, 2003.

[27] X. Ren, A. Berg and J. Malik. Recovering human body configurations using pairwise constraints between parts. *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[28] T. Roberts, S. McKenna and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. *European Conference on Computer Vision (ECCV)*, vol. 4, pp. 291–303, 2004.

[29] R. Ronfard, C. Schmid and B. Triggs. Larning to parse pictures of people. *European Conference on Computer Vision (ECCV)*, vol. 4, pp. 700–714, 2002.

[30] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 721–727, 2000.

[31] S. Sarkar, P. J. Phillips, Z. Liu, I. Robledo, P. Grother and K. W. Bowyer. The Human ID Gait Challenge Problem: Data Sets, Performance, and Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[32] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47(1/2/3):7–42, 2002.

[33] H. Sidenbladh, M. J. Black and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *European Conference on Computer Vision (ECCV)*, vol. 1, pp. 784–800, 2002.

[34] H. Sidenbladh, M. Black and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *European Conference on Computer Vision (ECCV)*, vol. 2, pp. 702–718, 2000.

[35] G. Shakhnarovich, P. Viola and T. Darrell. Fast pose estimation with parameter-sensitive hashing. *IEEE International Conference on Computer Vision (ICCV)*, vol.2, pp. 750–759, 2003.

[36] L. Sigal, S. Bhatia, S. Roth, M. Black and M. Isard. Tracking loose-limbed people. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 421–428, 2004.

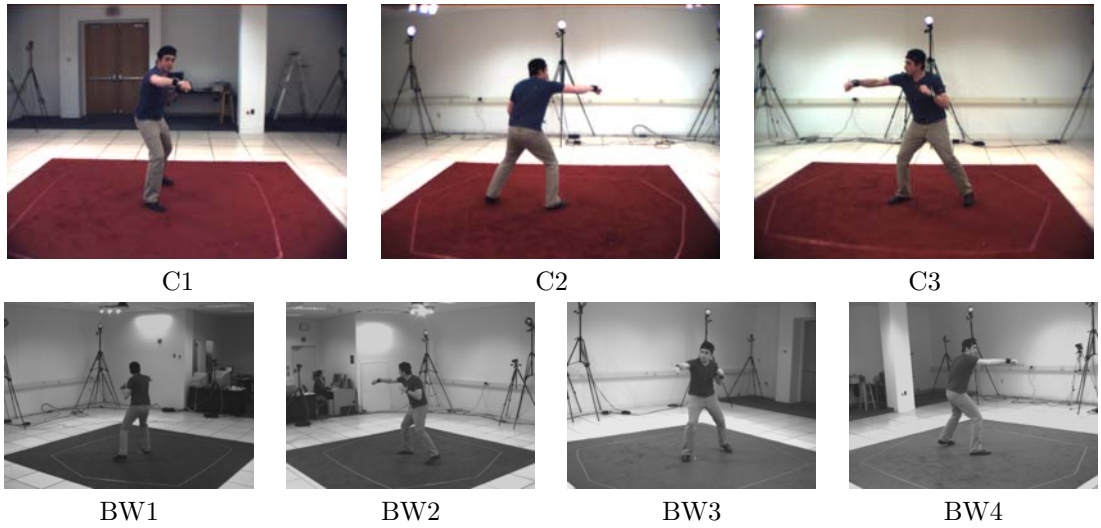
[37] L. Sigal and M. Black. Measure Locally, Reason Globally: Occlusion-sensitive articulated pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[38] C. Sminchisescu, A. Kanaujia, Z. Li and D. Metaxas. Discriminative density propagation for 3D human motion estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 390–397, 2005.

[39] C. Sminchisescu and B. Triggs Kinematic jump processes for monocular 3D human tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 69–76, 2003.

[40] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6), pp. 371–391, 2003.

[41] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding (CVIU)*, 80(3):349–363, 2000.



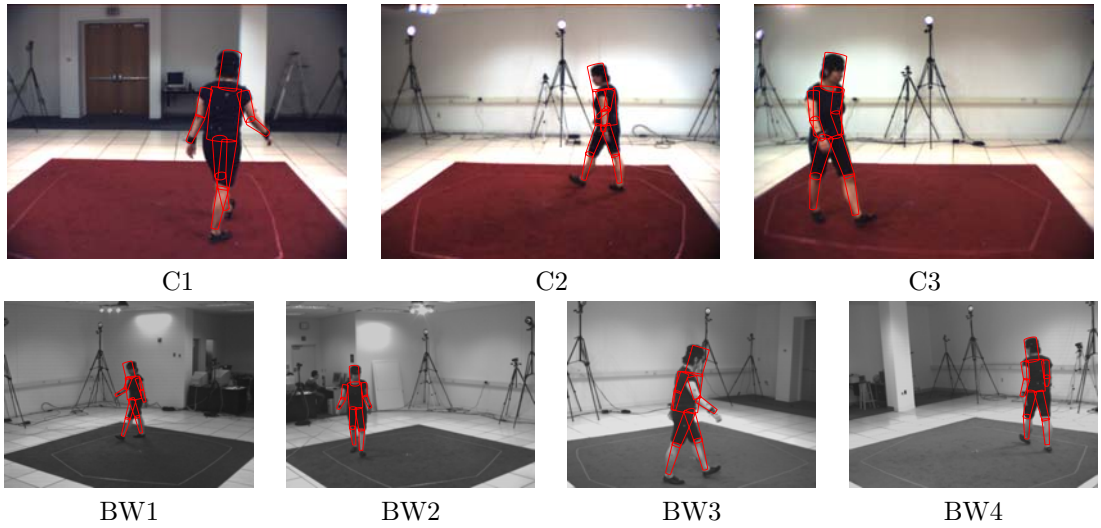


Figure 3: Example data from the HumanEva-I database. Example images of boxing from 7 synchronized video cameras (three colored and four grayscale) are shown on **top**. The synchronized motion capture data overlaid on the multi-view image data for walking of a different subject is shown on **bottom**. Notice that motion capture is available for the top sequence as well, but is not shown for clarity.