

HUMANEVA: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion

Leonid Sigal · Alexandru O. Balan · Michael J. Black

Received: 5 May 2008 / Accepted: 10 July 2009 / Published online: 5 August 2009
© Springer Science+Business Media, LLC 2009

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. We present data obtained using a hardware system that is able to capture synchronized video and ground-truth 3D motion. The resulting HUMANEVA datasets contain multiple subjects performing a set of predefined actions with a number of repetitions. On the order of 40,000 frames of synchronized motion capture and multi-view video (resulting in over one quarter million image frames in total) were collected at 60 Hz with an additional 37,000 time instants of pure motion capture data. A standard set of error measures is defined

for evaluating both 2D and 3D pose estimation and tracking algorithms. We also describe a baseline algorithm for 3D articulated tracking that uses a relatively standard Bayesian framework with optimization in the form of Sequential Importance Resampling and Annealed Particle Filtering. In the context of this baseline algorithm we explore a variety of likelihood functions, prior models of human motion and the effects of algorithm parameters. Our experiments suggest that image observation models and motion priors play important roles in performance, and that in a multi-view laboratory environment, where initialization is available, Bayesian filtering tends to perform well. The datasets and the software are made available to the research community. This infrastructure will support the development of new articulated motion and pose estimation algorithms, will provide a baseline for the evaluation and comparison of new methods, and will help establish the current state of the art in human pose estimation and tracking.

Keywords Articulated pose estimation · Articulated tracking · Motion capture · Human tracking · Datasets and evaluation

This project was supported in part by gifts from Honda Research Institute and Intel Corporation. Funding for portions of this work was also provided by NSF grants IIS-0534858 and IIS-0535075. We would like to thank Ming-Hsuan Yang, Rui Li, Payman Yadollahpour and Stefan Roth 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.

The first two authors contributed equally to this work.

The work of L. Sigal was conducted at Brown University.

L. Sigal (✉)
University of Toronto, Dept. of Computer Science, 6 King's
College Rd, Toronto, ON M5S 3H5, Canada
e-mail: ls@cs.toronto.edu

A.O. Balan · M.J. Black
Brown University, Dept. of Computer Science, 115 Waterman St,
Box 1910, Providence, RI 02912, USA

A.O. Balan
e-mail: alb@cs.brown.edu

M.J. Black
e-mail: black@cs.brown.edu

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 (Hogg 1983; O'Rourke and Badler 1980). A variety of statistical (Agarwal and Triggs 2004a, 2004b; Balan et al. 2005; Deutscher and Reid 2005; Hua et al. 2005; Sigal et al. 2004; Sigal and Black 2006; Sminchisescu et al. 2005) as well as deterministic methods (Mori et al. 2004; Taylor 2000;

Shakhnarovich et al. 2003) have been developed for tracking people from single (Agarwal and Triggs 2004a, 2004b; Felzenszwalb and Huttenlocher 2005; Hua et al. 2005; Lan and Huttenlocher 2005; Mori 2005; Mori et al. 2004; Ramanan et al. 2005; Ramanan and Forsyth 2003; Ren et al. 2005; Ronfard et al. 2002; Sigal and Black 2006) or multiple (Balan et al. 2005; Deutscher and Reid 2005; Grauman et al. 2003; Sigal et al. 2004) 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.

Quantitative evaluation of human pose estimation and tracking is currently limited due to the lack of common datasets containing “ground truth” with which to test and compare algorithms. Instead qualitative tests are still 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 (Deutscher and Reid 2005; Felzenszwalb and Huttenlocher 2005; Ren et al. 2005). Another form of inspection involves applying the estimated motion to a virtual character to see if the movements appear natural (Sminchisescu et al. 2005). The lack of the quantitative experimentation at least in part can be attributed to the difficulty of obtaining 3D ground-truth data that specify the true pose of the body observed in video sequences.

To obtain some form of ground truth, previous approaches have resorted to custom action-specific schemes; *e.g.* motion of the arm along a circular path of known diameter (Kakadiaris and Metaxas 1996). Alternatively, synthetic data have been extensively used (Agarwal and Triggs 2004a, 2004b; Grauman et al. 2003; Shakhnarovich et al. 2003; Sminchisescu et al. 2005) for quantitative evaluation. With packages such as POSER (*e frontier*, Scotts Valley, CA) or MAYA (Autodesk, San Rafael, 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 allow quantitative evaluation, current datasets are still too simplistic to capture the complexities of natural images of people and scenes.

In the last few years, there have been a few successful attempts (Gall et al. 2006; Knossow et al. 2008; Muendermann et al. 2007; Rosenhahn et al. 2006) to simultaneously capture video and ground truth 3D motion data (in the form of marker-based tracking); some groups were also able to capture 2D motion ground truth data in a similar fashion

(Wang and Rehg 2006). Typically hardware systems similar to the one proposed here have been employed (Knossow et al. 2008) where the video and motion capture data were captured either independently (and synchronized in software off-line) or with hardware synchronization. While this allowed some quantitative analysis of results (Gall et al. 2006; Knossow et al. 2008; Muendermann et al. 2007; Rosenhahn et al. 2006; Wang and Rehg 2006), to our knowledge none of the synchronized data captured by these groups (with the exception of (Wang and Rehg 2006), discussed in Sect. 2) has been made available to the community at large, making it hard for competing approaches to compare performance directly. For 2D human pose/motion estimation, quantitative evaluation is more common and typically uses hand-labeled data (Hua et al. 2005; Ramanan et al. 2005; Ramanan and Forsyth 2003). Furthermore, for both 2D and 3D methods, no standard error measures exist and results are reported in a variety of ways which prevent direct comparison; *e.g.* average root-mean-squared (RMS) angular error (Agarwal and Triggs 2004a, 2004b, Sminchisescu et al. 2005), normalized error in joint angles (Shakhnarovich et al. 2003), silhouette overlap (Ramanan et al. 2005; Ramanan and Forsyth 2003), joint center distance (Balan et al. 2005; Grauman et al. 2003; Lan and Huttenlocher 2005; Lee and Nevatia 2006; Li et al. 2006; Sigal et al. 2004; Sigal and Black 2006), *etc.*

Here we describe two datasets containing 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 these datasets, which we call HUMANEVA, 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 measures. The motivation behind the design of the HUMANEVA datasets 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 affect 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 datasets contain a number of subjects performing repetitions (trials) of a varied set of predefined actions. The datasets are broken into training, validation, and test sub-sets. For the testing subset, the ground truth data are withheld and a web-based

evaluation system is provided. A set of error measures is defined and made available as part of the dataset. These error measures 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 datasets. This support code shows how the data and error measures can be used and provides an easy-to-use Matlab (*The Mathworks*, Natick, MA) interface to the data. This allows different methods to be fairly compared using the same data and the same error measures.

In addition we provide a baseline algorithm for 3D articulated tracking in the form of simple Bayesian filtering. We analyze the performance of the baseline algorithm under a variety of parameter choices and show how these parameters affect the performance. The reported results on the HUMANEVA-II dataset are intended to be the baseline against which future algorithms that use the dataset can be compared. In addition, this Bayesian filtering software is freely available, and can serve as a foundation for new algorithm development and experimentation with image likelihood models and new prior models of human motion.

In systematically addressing the problems of articulated human pose estimation and tracking using the HUMANEVA datasets, other related research areas may benefit as well, such as 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 (Phillips et al. 2000, 2002), human gait identification (Gross and Shi 2001; Sarkar et al. 2005), dense stereo vision (Scharstein and Szeliski 2002) and optical flow (Baker et al. 2007). These efforts have helped advance the state-of-the-art in their respective fields. Our hope is that the HUMANEVA datasets will lead to similar advances in articulated human pose and motion estimation. In the short time that the dataset has been made available to the research community, it has already helped with the development and evaluation of new approaches for articulated motion estimation (Bissacco et al. 2007; Bo et al. 2008; Lee and Elgammal 2007; Li et al. 2006, 2007; Ning et al. 2008; Rogez et al. 2008; Urtasun and Darrell 2008; Vondrak et al. 2008; Xu and Li 2007). The dataset has also served as a basis for a series of workshops on Evaluation of Human Motion and Pose Estimation (EHuM)¹ set forth by the authors.

¹While the workshops did not have any printed proceedings, submissions can be viewed on-line: <http://www.cs.brown.edu/people/ls/ehum/>, <http://www.cs.brown.edu/people/ls/ehum2/>.

2 Related Work

2.1 Articulated Pose and Motion Estimation

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 an image sequence. Tracking approaches often assume knowledge of the initial pose of the body in the first frame and focus on the evolution of this pose over time. These approaches can be combined (Sigal et al. 2004; Sminchisescu et al. 2005), such that tracking benefits from automatic initialization and failure recovery in the form of static pose estimation and pose estimation benefits 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 while 2.5D approaches also allow the model to have relative depth information. Finally 3D approaches typically model the human body using simplified 3-dimensional parts such as cylinders or superquadrics. A short summary of different approaches with evaluation and error measures employed (when appropriate) can be seen in Table 1; for a more complete taxonomy, particularly of older work, we refer readers to (Gavrila 1999) and (Moeslund and Granum 2001).

2.2 Common Datasets

While HUMANEVA is the most extensive dataset for evaluation of human pose and motion estimation, there have been several related efforts. A similar approach was employed by Wang and Rehg (2006) where synchronized motion capture and monocular video was collected. The dataset, used by the authors to analyze performance of 2D articulated tracking algorithms, is available to the public.² The dataset, however, only contains 4 sequences (2 of which come from old movie footage and required manual labeling); only 2D ground truth marker positions are provided. The INRIA Perception Group also employed a similar approach for collection of ground truth data (Knossow et al. 2008), however,

²<http://www.cc.gatech.edu/grads/w/Ping.Wang/Project/FigureTracking.html>.

Table 1 Short survey of the human motion and tracking algorithms. Methods 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 (★) and (★★) evaluation measures are consistent with the evaluation measures proposed in this paper

Year	Reference	Model type	Parts	Dim	Type	Evaluation	Measure
1983	Hogg (1983)	Cylinders	14	2.5	T	Qualitative	
1996	Gavrila and Davis (1996)	Superquadric Ellip.	12	3	T	Quantitative	
1996	Ju et al. (1996)	Patches	2	2	T	Qualitative	
1996	Kakadiaris and Metaxas (1996)	D Silhouettes	2	3	T	Quantitative	
1998	Bregler and Malik (1998)	Ellipsoids	10	3	T	Qualitative*	
2000	Rosales and Sclaroff (2000)	Stick-Figure	10	3	P	Synthetic	★ ¹
2000	Sidenbladh et al. (2000)	Cylinders	2/10	3	T	Qualitative	
2002	Ronfard et al. (2002)	Patches	15	2	P	Hand Labeled	
2002	Sidenbladh et al. (2002)	Cylinders	2/10	3	T	Qualitative	
2003	Grauman et al. (2003)	Mesh	N/A	3	P	Synthetic/POSER	★
2003	Ramanan and Forsyth (2003)	Rectangles	10	2	T,P	Hand Labeled	◇◇
2003	Shakhnarovich et al. (2003)	Mesh	N/A	3	P	Synthetic/POSER	‡
2003	Sminchisescu and Triggs (2003a, 2003b)	Superquadric Ellip.	15	3	T	Qualitative ²	
2004	Agarwal and Triggs (2004a, 2004b)	Mesh	N/A	3	P	Synthetic/POSER	†
2004	Deutscher and Reid (2005)	R-Elliptical Cones	15	3	T	Qualitative	
2004	Lan and Huttenlocher (2004)	Rectangles	10	2	T,P	Qualitative	
2004	Mori et al. (2004)	Stick-Figure	9	3	P	Qualitative	
2004	Roberts et al. (2004)	Prob. Template	10	2	P	Qualitative	
2004	Sigal et al. (2004)	R-Elliptical Cones	10	3	T,P	Motion Capture	★★
2005	Balan et al. (2005)	R-Elliptical Cones	10	3	T	Motion Capture	★★
2005	Felzenszwalb and Huttenlocher (2005)	Rectangles	10	2	P	Qualitative	
2005	Hua et al. (2005)	Quadrangular	10	2	P	Hand Labeled	‡
2005	Lan and Huttenlocher (2005)	Rectangles	10	2	P	Motion Capture	★
2005	Ramanan et al. (2005)	Rectangles	10	2	T,P	Hand Labeled	◇◇
2005	Ren et al. (2005)	Stick-Figure	9	2	P	Qualitative	
2005	Sminchisescu et al. (2005)	Mesh	N/A	3	T,P	Synthetic/POSER	†
2006	Gall et al. (2006)	Mesh	N/A	3	T	Motion Capture	†
2006	Lee and Nevatia (2006)	R-Elliptical Cones	5/10	3	T,P	Hand Labeled	★★ ³
2006	Li et al. (2006)	R-Elliptical Cones	10	3	T	HUMANEVA	★★
2006	Rosenhahn et al. (2006)	Free-form surface patches	N/A	3	T	Motion Capture	†
2006	Sigal and Black (2006)	Quadrangular	10	2	P	Motion Capture	★
2006	Urtasun et al. (2006)	Stick-figure	15	3	T	Qualitative	
2006	Wang and Rehg (2006)	SPM + templates	10	2	T	Motion Capture	★ and ◇

only the multi-view video data is currently made available to the public.

The CMU Graphics Lab Motion Capture Database (CMU) is by far the most extensive dataset of publicly available motion capture data. It has been used by many researchers within the community to build prior models of human motion. The dataset, however, is not well suited for evaluating video-based tracking performance. While, for many of the motion capture sequences, low-resolution monocular videos are available, the calibration information required to project the 3D models into the images is not. Nevertheless, the video data has proved useful for the analysis of dis-

criminative methods that do not estimate 3D body location *e.g.* (Navaratnam et al. 2007). In addition, the subjects are dressed in tight fitting motion capture suits and hence lack the realistic clothing variations exhibited in less controlled environments.

The CMU Motion of Body (MoBo) Database (Gross and Shi 2001), initially developed for gait analysis, has also proved useful in analyzing the performance of articulated tracking algorithms (Fathi et al. 2007; Zhang et al. 2006). While the initial dataset, which contains an extensive collection of walking motions, did not contain joint-level ground

Table 1 (Continued)

Year	Reference	Model type	Parts	Dim	Type	Evaluation	Measure
2007	Balan et al. (2007)	SCAPE	15	3	P	Qualitative	
2007	Lee and Elgammal (2007)	Joint centers	N/A	3	T	HUMANEVA	**
2007	Muendermann et al. (2007)	SCAPE	15	3	T	Motion Capture	** and \diamond
2007	Navaratnam et al. (2007)	Mesh	N/A	3	P	Motion Capture	\dagger
2007	Srinivasan and Shi (2007)	Exemplars	6	2	P	Hand Labeled	* and \diamond
2007	Xu and Li (2007)	Cylinders	10	3	T	HUMANEVA	**
2008	Bo et al. (2008)	Joint centers	N/A	3	P	HUMANEVA	**
2008	Ning et al. (2008)	Stick-figure	10	3	P	HUMANEVA	\dagger
2008	Rogez et al. (2008)	Joint centers	10	2/3	P	HUMANEVA	*
2008	Urtasun and Darrell (2008)	Joint centers	N/A	3	P	HUMANEVA	**
2008	Vondrak et al. (2008)	Ellipsoids + prisms	13	3	T	HUMANEVA	**

*Mean squared distance in 2D between the set of $\mathcal{M} = 15$ (or fewer) virtual markers corresponding to the joint centers and limb ends. Measured in pixels (pix). $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|m_i(\mathbf{x}) - m_i(\hat{\mathbf{x}})\|$, where $m_i(\mathbf{x}) \in \mathbb{R}^2$ is the location of 2D marker i with respect to pose \mathbf{x}

**Mean squared distance in 3D between the set of $\mathcal{M} = 15$ virtual markers corresponding to the joint centers and limb ends. Measured in millimeters (mm). $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|m_i(\mathbf{x}) - m_i(\hat{\mathbf{x}})\|$, where $m_i(\mathbf{x}) \in \mathbb{R}^3$ is the location of 3D marker i with respect to pose \mathbf{x}

\dagger Root mean square (RMS) error in joint angle. Measured in degrees (deg). $D(\theta, \hat{\theta}) = \frac{1}{N} \sum_{i=1}^N |(\theta_i - \hat{\theta}_i) \bmod \pm 180^\circ|$, where $\theta \in \mathbb{R}^N$ is the pose in terms of joint angles

\ddagger Normalized error in joint angle. Measured as a fraction from 0 to 1. $D(\theta, \hat{\theta}) = \sum_{i=1}^N 1 - \cos(\theta_i - \hat{\theta}_i)$, where $\theta \in \mathbb{R}^N$ is the pose in terms of joint angles

\diamond Pixel overlap

$\diamond\diamond$ Pixel overlap based threshold resulting in binary 0/1 detection measure

\P Mean distance from 4 endpoints of quadrangular shape representing the limb

\dagger Error units were in fractions of the subject's height

\ddagger While only qualitative analysis of the overall tracking performance was presented, a quantitative analysis of the number of local minima in the posterior was performed

\P Additional per-limb weighting was applied to downweight the error proportionally with the size of the limb

truth information, manually labeled data has been made available³ by Zhang et al.

A more direct comparison of HUMANEVA to other datasets that are available to the community is given in Table 2.

3 HUMANEVA Datasets

To simultaneously capture video and motion information, our subjects wore natural clothing (as opposed to tight-fitting motion capture suits typically used for pure motion capture sessions) on which reflective markers were attached using invisible adhesive tape.⁴ Our motivation was to obtain

“natural” looking image data that contained all the complexity posed by moving clothing. One negative outcome of this is that the markers tend to move more than they would with a tight-fitting motion capture suit. As a 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, motion capture data for some frames in some sequences are missing markers or are inaccurate. We made an effort to detect such cases and exclude them from the 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

³<http://www.cs.cmu.edu/~zhangj/>.

⁴Participation 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. A copy of the consent form for the “Video and Motion Capture Project” is available by writing to 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.

Table 2 Comparison of HUMANEVA to other datasets available and employed by the community

	HUMANEVA datasets	Wang and Rehg (2006)	INRIA perception Knossow et al. (2008) multi-cam dataset	CMU MoCap dataset (CMU)	CMU MoBo dataset Gross and Shi (2001)
# of subjects	4	3	Unknown	> 100	25
# of frames	≈ 80,000	≈ 450	Unknown	Unknown	≈ 200,000
# of sequences	56	4	13	2605	100
<i>Video data</i>					
# of cameras	4/7	1	8/34	1	6
Calib. available	Yes	No	Yes	No	Yes
<i>Dataset content</i>					
Motion	Walk	Walk	Dance	Many	Walk
	Jog	Dance	Exercise		
	Throw/catch	Jumping jacks			
	Gesture				
	Box				
Appearance	Combo				
	Natural	Natural/ MoCap suit	Natural/ MoCap suit	MoCap suit	Natural
<i>Ground truth</i>					
Content	3D	2D	None	3D	2D
Source	MoCap	MoCap/ Manual label	None	MoCap	Manual label Zhang et al. (2006)

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 at video resolution and hence will likely have an insignificant impact on the performance of image-based tracking algorithms.

We have developed two datasets that we call HUMANEVA-I and HUMANEVA-II. HUMANEVA-I was captured earlier and is the larger of the two sets. HUMANEVA-II was captured using a more sophisticated hardware system that allowed better quality motion capture data and hardware synchronization. The differences between these two datasets are outlined in Fig. 1.

Since all the data was captured in a laboratory setting, the sequences do not contain any external occlusions or significant clutter, but do exhibit the challenges imposed by strong illumination (*e.g.* strong shadows that tend to confuse background subtraction); grayscale cameras used in the HUMANEVA-I dataset present additional challenges when it comes to background subtraction and image features. Even at 60 Hz the images still exhibit a fair amount of motion blur.

The split of the training and test data was specifically designed to emphasize the ability of the pose and motion estimation approaches to generalize to novel subjects and unobserved motions. To this end, one subject and one motion

for all subjects were withheld from the training and validation dataset for which ground truth is given out. We believe the proposed datasets exhibit a moderately complex and varied set of motions under realistic indoor imaging conditions that are applicable to most pose and motion estimation techniques proposed to date.

3.1 HUMANEVA-I

HUMANEVA-I contains data from 4 subjects performing a set of 6 predefined actions in three repetitions (twice with video and motion capture, and once with motion capture alone). A short description of the actions is provided in Fig. 1. Example images of a subject walking are shown in Fig. 2 where data from 7 synchronized video cameras is illustrated with an overlay of ground truth body pose.

3.1.1 Hardware

Ground truth motion of the body was captured using a commercial motion capture (MoCap) system from ViconPeak.⁵ 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.

⁵<http://www.vicon.com/>.

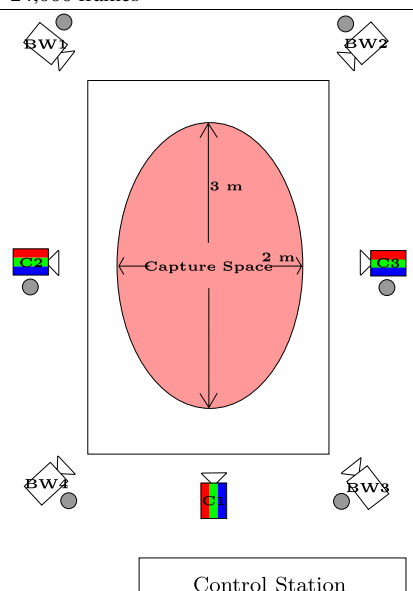
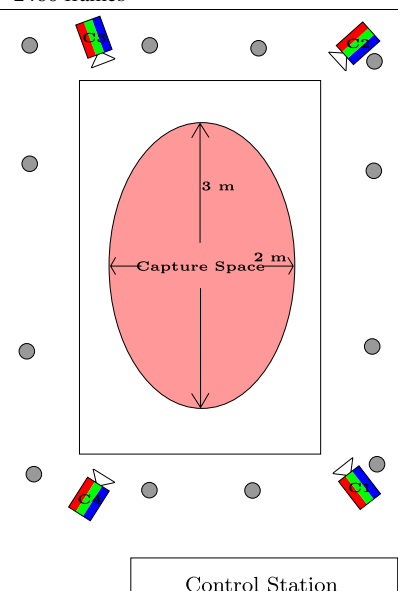
		HUMANEVA-I	HUMANEVA-II
MoCap	<i>Hardware system</i>		
	Manufacturer	ViconPeak	ViconPeak
	Number of cameras	6	12
	Camera resolution	1M-pixel	MX13 1.3M-pixel
	Frame rate	120 Hz	120 Hz
Video Capture System	<i>Color cameras</i>		
	Number of cameras	3	4
	Frame grabber	IO Industries	ViconPeak
	Camera model	UniQ UC685CL	Basler A602fc
	Sensor	Progressive Scan	Progressive Scan
	Camera resolution	659 × 494 pixels	656 × 490 pixels
	Frame rate	60 Hz	60 Hz
	<i>Grayscale cameras</i>		
	Number of cameras	4	
	Frame grabber	Spica Tech	
	Camera model	Pulnix TM6710	
	Sensor	Progressive Scan	
	Camera resolution	644 × 448 pixels	
	Frame rate	60 Hz	
Data	Synchronization	Software	Hardware
	Actions	(1) Walking, (2) Jogging, (3) Gesturing (4) Throwing and Catching a ball, (5) Boxing, (6) Combo	Combo
	Number of subjects	4	2
	<i>Number of frames</i>		
	Training (synchronized)	6800 frames	
	Training (MoCap only)	37,000 frames	
	Validation	6800 frames	
	Testing	24,000 frames	2460 frames
Capture Space Layout			

Fig. 1 HUMANEVA Datasets. The table illustrates the hardware system and configuration used to capture the two datasets, HUMANEVA-I and HUMANEVA-II. The main difference between the hardware systems lies in hardware synchronization employed in HUMANEVA-II. The contents of the two datasets in terms subjects, motion and amount of data are also noted. The bird's eye view sketch of the capture configuration is also shown with rough dimensions of the capture space and placement of video and motion capture cameras. The color video cameras (C) are designated by RGB striped pattern, grayscale video cameras (BW) by the empty camera icon and motion capture cameras are denoted by gray circles

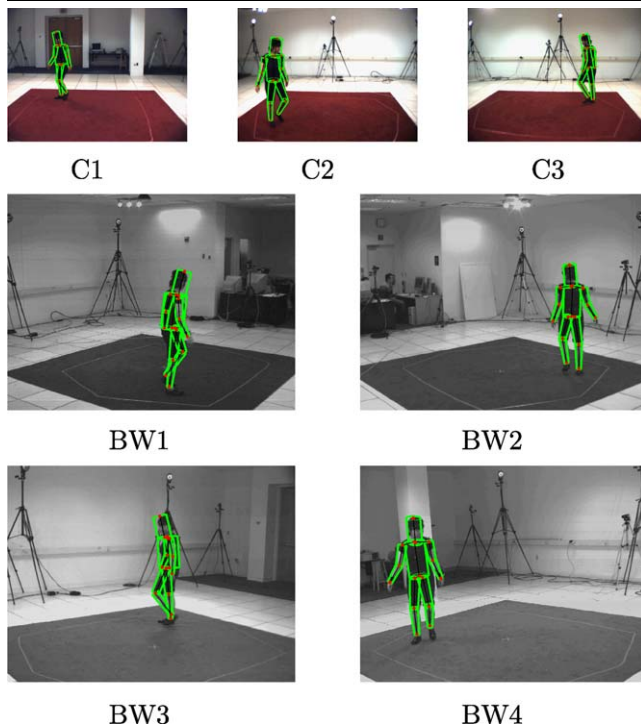


Fig. 2 Example data from the HUMANEVA-I database. Example images of walking subject (S1) from 7 synchronized video cameras (three colored and four grayscale) are shown with overlaid synchronized motion capture data

Video data was captured using two commercial video capture systems, one from Spica Technology Corporation⁶ and one from IO Industries.⁷ The Spica system captured video using four Pulnix⁸ TM6710 grayscale cameras (grayscale, progressive scan, 644×488 resolution, frame rate of up to 120 Hz). The IO Industries system used three UniQ⁹ UC685CL 10-bit color cameras with 659×494 resolution and a frame rate of up to 110 Hz. The raw frames were re-scaled from 659×494 to 640×480 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 Fig. 1 (left).

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 Sects. 3.3 and 3.4 respectively.

⁶<http://www.spicatek.com/>.

⁷<http://www.ioindustries.com/>.

⁸<http://www.pulnix.com/>.

⁹<http://www.uniqvision.com/>.

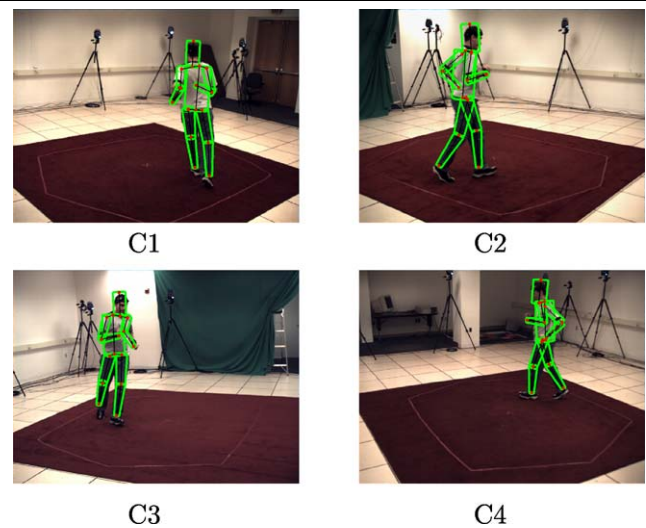


Fig. 3 Example data from the HUMANEVA-II database. Example images of subject (S4) from 4 synchronized color video cameras performing a combo motion (that includes jogging as shown)

3.2 HUMANEVA-II

HUMANEVA-II contains only 2 subjects (both also appear in the HUMANEVA-I dataset) performing an extended sequence of actions that we call *Combo*. In this sequence (see Fig. 3) a subject starts by walking along an elliptical path, then continues on to jog in the same direction and concludes with the subject alternatively balancing on each of the two feet roughly in the center of the viewing volume. Unlike HUMANEVA-I, this later dataset contains a relatively small test set of synchronized frames (≈ 2500). The HUMANEVA-I training and validation data is intended to be shared across the two datasets with test results primarily being reported on HUMANEVA-II.

3.2.1 Hardware

As with HUMANEVA-I, the ground truth motion capture data was acquired using a system from ViconPeak. However, here we used a more recent Vicon MX system with twelve 1.3M-pixel cameras. This newer system produced more accurate motion capture data.

Video data was captured using a 4-camera reference system provided by ViconPeak which allowed for frame-accurate synchronization (using the Vicon MX Control module) of the video and motion capture data. Video was captured using four Basler¹⁰ A602fc progressive scan cameras with 656×490 resolution operated at 60 Hz. The rough relative placement of cameras is illustrated in Fig. 1 (right). A calibration procedure to align the Vicon and Basler coordinate systems is discussed in the next section.

¹⁰<http://www.baslerweb.com/>.

3.3 Calibration

The motion capture system was calibrated using Vicon's proprietary software and protocol. Calibration of the intrinsic parameters for the video capture systems was done using a standard checker-board calibration grid and the Camera Calibration Toolbox for Matlab (Bouguet). 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 $c \in \mathcal{C}$. We assume square pixels and let the skew $\alpha_c = 0$ for all cameras $c \in \mathcal{C}$.

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 solved for using a semi-automated procedure to align the global coordinate axis of each video camera with the global coordinate axis of the Vicon motion capture system. A single moving marker was captured by the video cameras and the motion capture system for a number of synchronized frames (> 1000). The resulting 3D tracked position of the marker $\Gamma_t^{(3D)}$, $t \in \{1 \dots T^{(3D)}\}$ was recovered using the Vicon software. The 2D position of the marker in the video, $\Gamma_t^{(2D)}$, $t \in \{1 \dots T^{(2D)}\}$, was recovered using a Hough circle transform (Hough 1962) that was manually initialized in the first frame and subsequently tracked. 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

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

for the rotation, R_c , and translation, T_c . Note that the video cameras were calibrated with respect to the calibration parameters of the Vicon system, as opposed to from the images directly.

In the HUMANEVA-I dataset, the video and motion capture systems were not temporally synchronized in hardware, hence we also solved for 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 we assumed that the temporal scaling was constant over the length of a capture sequence¹¹ (i.e. no temporal drift). The 3D position $f(\Gamma_{tA_c+B_c}^{(3D)}; R_c, T_c)$ was linearly interpolated to cope with non-integer indices $tA_c + B_c$. Finally, in (1), $\delta(t; A_c, B_c)$ is defined as:

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

¹¹In practice $A_c \approx 2$ since the frame rate of motion capture system was roughly 120 Hz and video system is 60 Hz.

The calibration accuracy of the video cameras appears most accurate in the center of the viewing volume (close to the world origin).

For the HUMANEVA-II data, frame-accurate synchronization was achieved in hardware and we used fixed values $A_c = 2$ and $B_c = 0$ for the temporal scaling and offset.

3.4 Synchronization

While the extrinsic calibration parameters and temporal scaling, A_c , can be estimated once per camera (the Vicon system was only re-calibrated when cameras moved¹²), without hardware synchronization, the temporal offset B_c was different for every sequence captured. To temporally synchronize the motion capture and the video in software, for HUMANEVA-I we manually labeled visible markers on the body for a small sub-set of images (6 images were used with several marker positions labeled per frame). These labeled frames were subsequently 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.

4 Evaluation Measures

Various evaluation measures have been proposed for human motion tracking and pose estimation. For example, a number of papers have suggested using joint-angle difference 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 introduce a more widely applicable error measure based on a sparse set of virtual markers that correspond to the locations of joints and limb endpoints. This error measure was first introduced for 3D pose estimation and tracking in (Sigal et al. 2004) and later extended in (Balan et al. 2005). It has since been also used for 3D tracking in (Li et al. 2006) and for 2D pose estimation evaluation in (Lan and Huttenlocher 2005; Sigal and Black 2006).

Let \mathbf{x} represent the pose of the body. We define $\mathcal{M} = 15$ virtual markers as $\{m_i(\mathbf{x})\}$, $i = 1 \dots \mathcal{M}$, where $m_i(\mathbf{x}) \in \mathbb{R}^3$ (or $m_i(\mathbf{x}) \in \mathbb{R}^2$ if a 2D body model is used) is a function of the body pose that returns the position of the i th marker in the world (or image respectively). Notice that defining functions $m_i(\mathbf{x})$ for any standard representation of the body pose \mathbf{x} is trivial. The error between the estimated pose $\hat{\mathbf{x}}$ and the

¹²Calibration 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.

ground truth pose \mathbf{x} is expressed as the average Euclidean distance between individual virtual markers:

$$D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|m_i(\mathbf{x}) - m_i(\hat{\mathbf{x}})\|. \quad (3)$$

To ensure that we can compare algorithms that use different numbers of parts, we add a binary selection variable per-marker $\hat{\Delta} = \{\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_{\mathcal{M}}\}$ and obtain the final error function

$$D(\mathbf{x}, \hat{\mathbf{x}}, \hat{\Delta}) = \frac{1}{\sum_{j=1}^{\mathcal{M}} \hat{\delta}_j} \sum_{i=1}^{\mathcal{M}} \hat{\delta}_i \|m_i(\mathbf{x}) - m_i(\hat{\mathbf{x}})\|, \quad (4)$$

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

For the sequence of T frames we compute the average performance using the following:

$$\mu_{seq} = \frac{1}{T} \sum_{t=1}^T D(\mathbf{x}_t, \hat{\mathbf{x}}_t, \hat{\Delta}_t). \quad (5)$$

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 (Balan et al. 2005; Lan and Huttenlocher 2005; Sigal et al. 2004; Sigal and Black 2006) that have already used this error measure, we compute the 3D error in millimeters (mm) and the 2D error directly in the image in pixels (pix).

The error measures formulated above are appropriate for measuring the performance of approaches that are able to recover the full 3D articulated pose of the person in space or the 2D articulated pose of the person in an image. Some approaches, however, are inherently developed to recover the pose but not the global position of the body (most discriminative approaches fall into this category, e.g. Agarwal and Triggs 2004b; Navaratnam et al. 2007; Sminchisescu et al. 2005). To make the above error measures appropriate for this class of approaches we employ a *relative* variant

$$\tilde{D}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|\tilde{m}_i(\mathbf{x}) - \tilde{m}_i(\hat{\mathbf{x}})\|, \quad (6)$$

with $\tilde{m}_i(\mathbf{x}) = m_i(\mathbf{x}) - m_0(\mathbf{x})$, where $m_i(\mathbf{x})$ is defined as before and $m_0(\mathbf{x})$ is the position of the marker corresponding to the origin of the root segment. The rest of the equations can also be modified accordingly. It is worth noting that this measure assumes that the orientation of the body relative to the camera is recovered; this is typical of most discriminative methods.

Note that the error measures assume that an algorithm returns a unique body pose estimate rather than a distribution

over poses. For algorithms that model the posterior distribution over poses as uni-modal, the mean pose is likely to give a good estimate of \mathbf{x} . Most recent methods, however, model multi-modal posterior distributions implicitly or explicitly. Here the maximum-a posteriori estimate may be a more appropriate choice for \mathbf{x} . This is discussed in greater detail in (Balan et al. 2005). Alternative error measures that compute lower-bounds for sample- or kernel-based representations of the posterior are discussed in (Balan et al. 2005).

5 Baseline Algorithm

In addition to the datasets and quantitative evaluation measures, we provide a baseline algorithm¹³ against which future advances can be measured. While no “standard” algorithm exists in the community, we implemented a fairly common Bayesian filtering method based on the methods of Deutscher and Reid (2005) and Sidenbladh et al. (2002). Several variations on the base algorithm are explored with the goal of giving some insight into the important design choices for human trackers. Quantitative results are presented in the following section.

5.1 Bayesian Filtering Formulation

We pose the tracking problem in a standard way as one of estimating the *posterior* probability distribution $p(\mathbf{x}_t | \mathbf{y}_{1:t})$ for the state \mathbf{x}_t of the human body at time t given a sequence of image observations $\mathbf{y}_{1:t} \equiv (\mathbf{y}_1, \dots, \mathbf{y}_t)$. Assuming a first-order Markov process

$$p(\mathbf{x}_t | \mathbf{x}_{1:t-1}) = p(\mathbf{x}_t | \mathbf{x}_{t-1}),$$

with a sensor Markov assumption

$$p(\mathbf{y}_t | \mathbf{x}_{1:t}, \mathbf{y}_{1:t-1}) = p(\mathbf{y}_t | \mathbf{x}_t),$$

a recursive formula for the posterior can be derived (Arulampalam et al. 2002; Doucet et al. 2000):

$$p(\mathbf{x}_t | \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t | \mathbf{x}_t) \int p(\mathbf{x}_t | \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}, \quad (7)$$

where the integral in (7) computes the *prediction* using the previous posterior and the *temporal diffusion* model $p(\mathbf{x}_t | \mathbf{x}_{t-1})$. The prediction is weighted by the *likelihood* $p(\mathbf{y}_t | \mathbf{x}_t)$ of the new image observation conditioned on the pose estimate.

¹³The implementation is available for download from <http://vision.cs.brown.edu/humaneva/baseline.html>.

5.1.1 Optimization

Non-parametric approximate methods represent posterior distributions by a set of N random samples or particles with associated normalized weights that are propagated over time using the temporal model and assigned new weights according to the likelihood function. This is the basis of the Sequential Importance Resampling (SIR) algorithm, or Condensation (Arulampalam et al. 2002; Isard and Blake 1998). A variation of SIR is the Annealed Particle Filter (APF) introduced for human tracking by Deutscher and Reid (2005). An APF iterates these steps multiple times at each time instant in order to better localize the modes of the posterior distribution, and relies on simulated annealing to avoid local optima.

We briefly summarize our implementation of the Annealed Particle Filter algorithm used here since this forms the core of our baseline algorithm in the experiments that follow. The Sequential Importance Resampling algorithm is also tested in the following section but is not described in detail as it is similar to APF.

At each time instant the APF algorithm proceeds in a set of “layers”, from layer M down to layer 1, that update the probability density over the state parameters. The state density at layer $m + 1$ is represented using a set of N particles with associated normalized weights $\mathcal{S}_{t,m+1} \equiv \{\mathbf{x}_{t,m+1}^{(i)}, \pi_{t,m+1}^{(i)}\}_{i=1}^N$. For the prediction step at layer m , a Gaussian diffusion model is implemented (Sect. 5.1.4). Specifically, hypotheses are drawn with replacement using Monte Carlo sampling from the state probability density at the previous layer $m + 1$ using

$$\{\mathbf{x}_{t,m}^{(i)}\}_{i=1}^N \sim \sum_{j=1}^N \pi_{t,m+1}^{(j)} \mathcal{N}(\mathbf{x}_{t,m+1}^{(j)}, \alpha^{M-m} \Sigma). \quad (8)$$

The sampling covariance matrix Σ controls the breadth of the search at each layer with a large Σ spreading sampled particles more widely. From layer to layer we scale Σ by a parameter α . This parameter is used to gradually reduce the diffusion covariance matrix Σ at lower layers in order to drive the particles towards the modes of the posterior distribution. Typically α is set to 0.5.

Sampled poses that exceed the joint angle limits of the trained action model or result in inter-penetration of limbs are rejected and not re-sampled within a layer. The remaining particles are assigned new normalized weights based on an “annealed” version of the likelihood function (Sect. 5.1.3)

$$\pi_{t,m}^{(i)} = \frac{p(\mathbf{y}_t | \mathbf{x}_{t,m}^{(i)})^{\beta^m}}{\sum_{j=1}^N p(\mathbf{y}_t | \mathbf{x}_{t,m}^{(j)})^{\beta^m}}, \quad i \in \{1, \dots, N\}, \quad (9)$$

where β^m is a temperature parameter optimized so that approximately half the particles get selected for propagation/diffusion to the next layer by the Monte-Carlo sampler (8). The resulting particle set $\mathcal{S}_{t,m} \equiv \{\mathbf{x}_{t,m}^{(i)}, \pi_{t,m}^{(i)}\}_{i=1}^N$ is then used to compute layer $m - 1$ by re-applying (8), (9). In tracking, the top layer is initialized with the particle set of the bottom layer at the previous time instant: $\mathcal{S}_{t,M+1} = \mathcal{S}_{t-1,1}$.

The *expected* as well as the *maximum a posteriori* poses at frame t can be computed from the particle set $\mathcal{S}_{t,1}$ at the bottom layer using:

$$\hat{\mathbf{x}}_t = \sum_{i=1}^N \pi_{t,1}^{(i)} \mathbf{x}_{t,1}^{(i)}, \quad (10)$$

$$\hat{\mathbf{x}}_t^{\text{MAP}} = \mathbf{x}_{t,1}^{(j)}, \quad \pi_{t,1}^{(j)} = \max_i (\pi_{t,1}^{(i)}). \quad (11)$$

SIR is a special case of APF which has only one annealing layer ($M = 1$) and for which the effect of the annealing temperature parameter is removed ($\beta^m = 1$).

5.1.2 Parametrization of the Skeleton

As is common in the literature, the skeleton of the body is modeled as a 3D kinematic tree with the limbs represented by truncated cones (Fig. 4(b)). We consider 15 body parts: pelvis area, torso, head, upper and lower arms and legs, hands and feet. There are two types of parameters that describe the pose and shape of the body. The shape is given by the length and width of the limbs, which in our case are assumed known and fixed. Our objective is to recover the pose of the body, which is parametrized by a reduced set of 34 parameters comprising the global position and orientation of the pelvis and the relative joint angles between neighboring limbs. The hips, shoulders and thorax are modeled as ball and socket joints (3 DoF), the clavicles are allowed 2 DoFs, while the knees, ankles, elbows, wrists and head are assumed to be hinge joints with 1 DoF.

The subjects in the dataset were all manually measured using a standard Vicon protocol to obtain their height, weight, limb width and shoulder joint offsets. Motion capture training data was then used to estimate limb lengths for each subject as well as to learn static and dynamic priors for different motion styles. The raw data provided by the Vicon motion capture system consists of the location and orientation of local coordinate systems at each joint, with consecutive joints along the skeleton not constrained to be a fixed distance from each other. Limb lengths are computed as the median distance between pairs of corresponding joint locations over a large set of training motions and are kept fixed during testing. We also derive joint angle limits and inter-frame joint angle variations from the statistics of the relative joint angles between neighboring body parts.

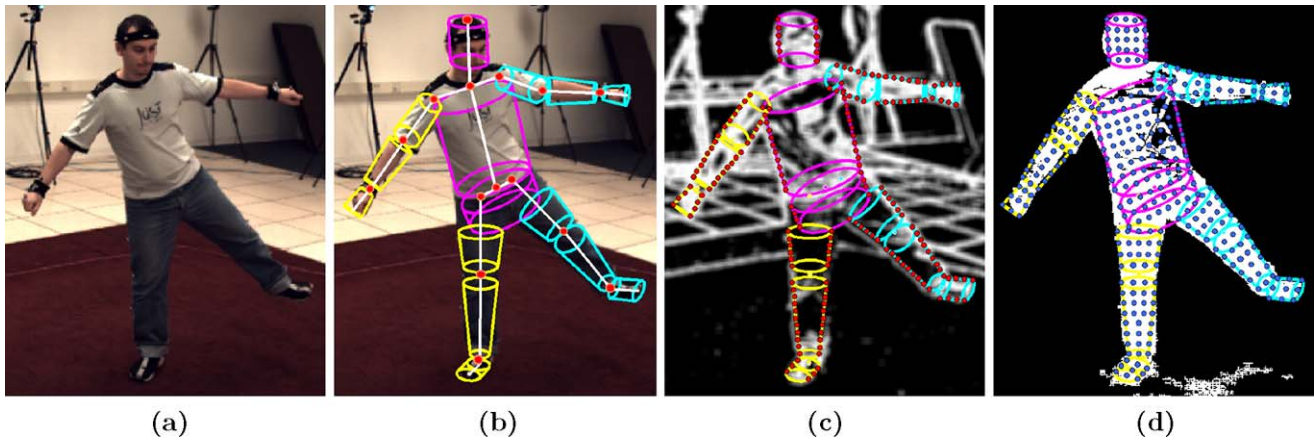


Fig. 4 (a) Input image. (b) Body model. The body is represented as a kinematic tree with 15 body parts. The red spheres represent the joint locations where virtual markers are placed for computing 3D error: pelvis joint, hips, knees and ankles, shoulders, elbows and wrists, neck and the top of the head. (c) Smoothed gradient edge map M_t^e , with values ranging from 0 (pure black) to 1 (pure white). Sparse points shown in red $\{\xi_{x_t}^e(j)\}$ along the edges of the body model are matched against the edges in the image. (d) Foreground silhouette map M_t^f , with the background being 0 and foreground 1. Sparse points shown in blue $\{\xi_{x_t}^f(j)\}$ selected in a grid inside the body model are matched against the foreground silhouette

5.1.3 Likelihoods

For each particle in the posterior representation, its likelihood represents how well the projection of a given body pose fits the observed image(s). Many image features could be used, including appearance models and optical flow constraints, however, most common approaches rely on silhouettes and edges (Deutscher and Reid 2005).

5.1.3.1 Edge-based Likelihood Functions We detect edges using image gradients that have been thresholded to obtain binary maps (Deutscher and Reid 2005). An edge distance map M^e is then constructed for each image to determine the proximity of a pixel to an edge. This can be achieved by convolving the binary edge map with a Gaussian kernel, and then re-mapping it between 0 and 1. This can be thought of as representing the edge probability (Deutscher and Reid 2005) at a given pixel.

The negative log-likelihood is then estimated by projecting into the edge map sparse points (for computational efficiency) along the apparent boundaries of all model parts and computing the mean square error (MSE) of the edge map responses:

$$-\log p^e(\mathbf{y}_t | \mathbf{x}_t) \propto \frac{1}{|\{\xi_{x_t}^e\}|} \sum_j (1 - M_t^e(\xi_{x_t}^e(j)))^2, \quad (12)$$

where $\{\xi_{x_t}^e\}$ is the set of pixel locations corresponding to all projected points (indexed by j) along all body part edges induced by pose \mathbf{x}_t , and M_t^e is the edge distance map at time t (Fig. 4(c)). The reader is referred to (Deutscher and Reid 2005) for a more detailed discussion.

5.1.3.2 Silhouette-based Likelihood Function Binary foreground silhouette maps M_t^f are generated using a learned Gaussian model for each pixel; the model is learned from 10 static background images and silhouettes subsequently obtained by comparing the background pixel probability to that of a uniform foreground model. We model the constraint that the silhouette of the body model should project inside the image silhouette. As before, for computational efficiency, we only check for a sparse number of points within the limbs (Fig. 4(d)). The negative log-likelihood of the observations given pose \mathbf{x}_t is then estimated by taking a number of visible points inside all limbs and projecting them into the image $\{\xi_{x_t}^f\}$. The MSE between the predicted and observed silhouette values for these points is computed (Deutscher and Reid 2005):

$$-\log p^f(\mathbf{y}_t | \mathbf{x}_t) \propto \frac{1}{|\{\xi_{x_t}^f\}|} \sum_j (1 - M_t^f(\xi_{x_t}^f(j)))^2. \quad (13)$$

5.1.3.3 Bi-directional Silhouette-based Likelihood Function The advantage of the previous silhouette likelihood formulation is computational efficiency and similarity to the edge-based likelihood formulation. However, this comes at the expense of being asymmetric: the body is constrained to lie inside the image silhouette, but not vice versa. This becomes a problem when the model predicts occluded parts and consequently does not fully cover the image silhouette. In Fig. 5(b) both legs track the same physical leg, but the penalty is minimal using $p^f(\mathbf{y}_t | \mathbf{x}_t)$.

We can correct this by defining a symmetric silhouette likelihood (Sminchisescu and Telea 2002; Sminchisescu

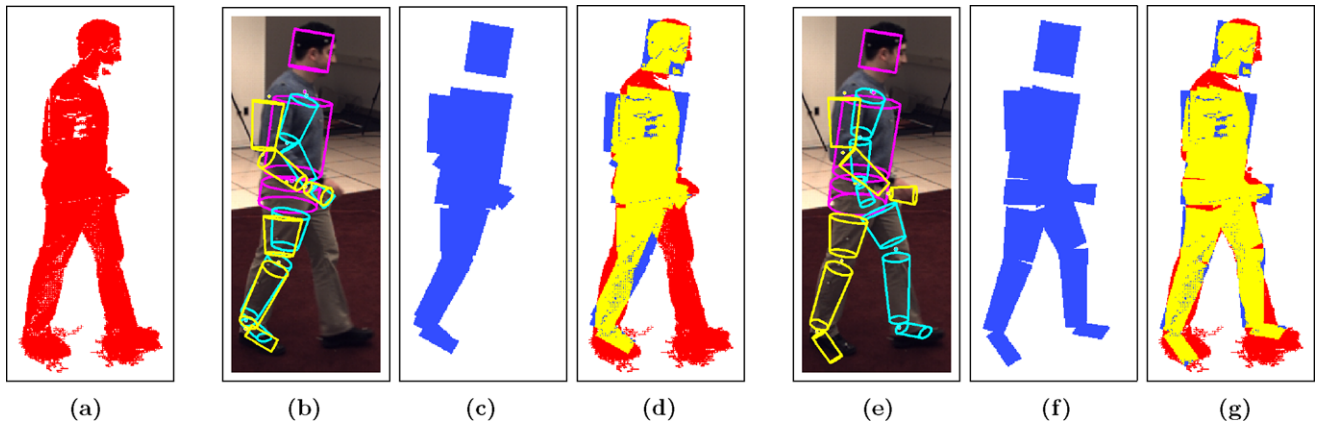


Fig. 5 Silhouette-based likelihood. **(b)–(d)**: Traditional silhouette likelihood. **(e)–(g)**: Bi-directional likelihood. **(a)** Foreground silhouette map M_t^f . The red pixels have value 1, and the background 0. **(b)** Tracking failure example using standard likelihood; two legs explain the same silhouette pixels. **(c)** Body model silhouette map M_t^b , obtained by rendering the cylinders to the image plane. The Blue pixels have value 1, and the background 0. **(d)** Silhouette overlap (Yellow). The standard silhouette likelihood does not penalize for the fact that the Red regions are not explained by the model. **(e)** Tracking result with bi-direction silhouette term; both legs now correct. **(f)** Body model silhouette projected into the image. **(g)** Silhouette overlap for bi-direction term; more image pixels are explained (Yellow pixels)

2002) that penalizes non-overlapping regions for both silhouettes. For this it is convenient to use a pixel-dense silhouette representation. Let M_t^b represent the binary silhouette map for the cylindrical body model and M_t^f the image foreground. Figure 5(d) shows the overlap between the two silhouettes. We seek to minimize the non-overlapping regions, Red and Blue, therefore maximizing the Yellow region. The size of each region can be computed by summing over all image pixels p using

$$R_t = \sum_p (M_t^f(p)(1 - M_t^b(p))), \quad (14)$$

$$B_t = \sum_p (M_t^b(p)(1 - M_t^f(p))), \quad (15)$$

$$Y_t = \sum_p (M_t^f(p)M_t^b(p)). \quad (16)$$

The negative log-likelihood of a pose is then defined as a linear combination of the fractions of each silhouette not explained by the other:

$$-\log p^d(\mathbf{y}_t|\mathbf{x}_t) \propto (1-a) \frac{B_t}{B_t + Y_t} + a \frac{R_t}{R_t + Y_t}. \quad (17)$$

We make the likelihood symmetric by setting $a = 0.5$. When a is 0, we effectively get the behavior of the previous 1-sided silhouette likelihood $p^f(\mathbf{y}_t|\mathbf{x}_t)$.

5.1.3.4 Combining Likelihood Functions We combine image measurements from multiple cameras or multiple likelihood formulations as follows:

$$-\log p(\mathbf{y}_t|\mathbf{x}_t) = \frac{1}{K} \frac{1}{|L|} \sum_{k=1}^K \sum_{l \in L} -\log p^l(\mathbf{y}_t^{(k)}|\mathbf{x}_t), \quad (18)$$

where K is the number of cameras, $\mathbf{y}_t^{(k)}$ is the image observation in the k -th camera and $L \subset \{e, f, d\}$ is a set of likelihood functions such as the ones in (12), (13), (17).

5.1.4 Action Models: Temporal Diffusion and Pose Priors

Predictions from the posterior are made using temporal models. The simplest model applicable to generic motions assumes no change in state from one time to the next: $\bar{\mathbf{x}}_t = \mathbf{x}_{t-1}$ (Deutscher and Reid 2005). The predictions are diffused using normally distributed random noise to account for errors in the assumption. The noise is drawn from a multi-variate Gaussian with diagonal covariance Σ where the sampling standard deviation of each body angle is set to equal the maximum absolute inter-frame angular difference for a particular motion style (Deutscher and Reid 2005).

We also implement a hard prior on individual poses to reduce the search space. Specifically, we reject (without re-sampling) any particle corresponding to an implausible body pose. We check for angles exceeding joint angle bounds and producing inter-penetrating limbs (Sminchisescu and Triggs 2003b). In our implementation we explicitly test for intersections between the torso and the lower arms and between the left and the right calves.

We use the term *action model* (AM) to denote the sampling covariance Σ used for particle filtering and the valid range of the joint angles. Action models can be learned specifically for a certain actor or for a particular motion style, or they can be generic. We only learned subject-generic action models by combining the data from all three available subjects in the training dataset.

Different motion styles influence the sampling covariance and joint angle limits used. The training data in the

HUMANEVA-I dataset contains walking, jogging, hand gestures, throwing and catching a ball, and boxing action styles from three different subjects. Subsets of these were used to learn style-specific action models. For example, the sampling covariance and the valid range of joint angles are typically smaller for walking than for jogging models, making the problem simpler for walking test sequences. For sequences containing both walking and jogging, it is typical for the flexion-extension movement of the elbow to cover disjoint angle intervals for the two styles. A combined action model for walking and jogging can be learned instead.

To represent a generic (style-independent) action model, we use the entire HUMANEVA-I training data to learn the sampling covariance Σ^G . For the joint limits, our training data is not diverse enough to be suitable for discovering the full anatomical range of every joint angle, particularly for the leg joints. Instead we rely on standard anatomical joint limits (AJL).

6 Experiments

We performed a series of experiments with the two different algorithms (APF and SIR), several likelihoods and various action models (“priors”); details of each variant are described along with the corresponding experiment. Most of these are variations of a *base configuration* (BC) that uses annealed filtering with 200 particles per layer, 5 layers of annealing, a likelihood based on bi-directional silhouette matching (BiS), and an action model appropriate for *generic* motions which enforces anatomical joint limits (G-AJL). We also reject samples where the limbs penetrate each other as described above. The experiments were conducted on the two sequences in the HUMANEVA-II dataset. In each case, ground truth was available in the first frame to initialize the tracker.

The error of an individual pose was computed using (4) which averages the Euclidean distance between virtual markers placed as shown in Fig. 4(b). Given the samples (particles) at each frame, we computed the error of the expected pose using (10). This is appropriate for the APF since we expect the posterior representation to be uni-modal at the bottom layer of annealing. Alternatively we could have estimated the error of the most likely pose in the posterior distribution. In our experiments we found this measure to be consistently worse than the error of the expected pose by an average of 2 mm with noisier reconstructed joint angle trajectories. We attribute this to the fact that particle filtering methods represent the posterior probability as a function of both the likelihood weights and the density of particles. The MAP estimate may miss a region that has a high posterior probability due to high particle density but small individual weights.

Our optimization strategy is stochastic in nature and produces different results when running experiments with the same configuration parameters. To get a measure of performance consistency and repeatability, we ran each experiment five times for each of the sequences, unless explicitly noted otherwise. We plot the mean of our error measure (3D or 2D depending on the experiment) for each time instant over all the runs, while for the BC we also highlight the standard deviation as a gray overlay in Figs. 6, 9, 10, 11, 12, 13, and in the corresponding rows in the error tables.

The errors at each frame are combined to compute the average error μ_{seq} (5) for each of the three activity phases (walking, jogging and leg balancing), as well as the overall error over the two sequences. We report the mean and standard deviation of the average error μ_{seq} over multiple runs.¹⁴

6.1 Computation Time

The computation time is directly proportional to the number of particles, number of camera views and number of layers used, and vastly depends on the choice of likelihood function. Performing full inference using the one-sided silhouette likelihood $p^f(\mathbf{y}_t|\mathbf{x}_t)$ jointly with the edge likelihood $p^e(\mathbf{y}_t|\mathbf{x}_t)$ with 1000 particles per frame for 4 camera views takes about 40 seconds on a standard PC with software written in Matlab, as opposed to 250 seconds when using the bi-directional silhouette likelihood $p^d(\mathbf{y}_t|\mathbf{x}_t)$. Likelihood evaluations dominate the overall computation time; particle diffusion and checking for limb inter-penetration are relatively insignificant by comparison.

6.2 Performance of the Base Configuration BC

Sample tracking results overlaid on the images using the BC are shown in Figs. 14 and 15, and illustrate visually what different amounts of error correspond to. The 5 runs of BC suggest that the tracking performance is fairly stable across runs. This is illustrated in Fig. 6 for 3D errors. Performance results using other error measures are included in Fig. 7 to allow easy comparison with other methods.

The occasional spikes in the error correspond to the tracker losing track of the arms or the legs swapping places (e.g. frame 656 in Fig. 14). Since walking and jogging are periodic motions, the arms and legs are usually found again during the next cycle. This is also illustrated when investigating the error for individual joints. Figure 8 shows that limb extremities are the hardest to track. Large errors for the wrists are obtained when the elbow becomes fully flexed and

¹⁴When computing the results, we ignore 38 frames (298–335) for sequence 2 where accurate ground truth was not available. The apparent gap in the error plots during the walking phase is a result of this.

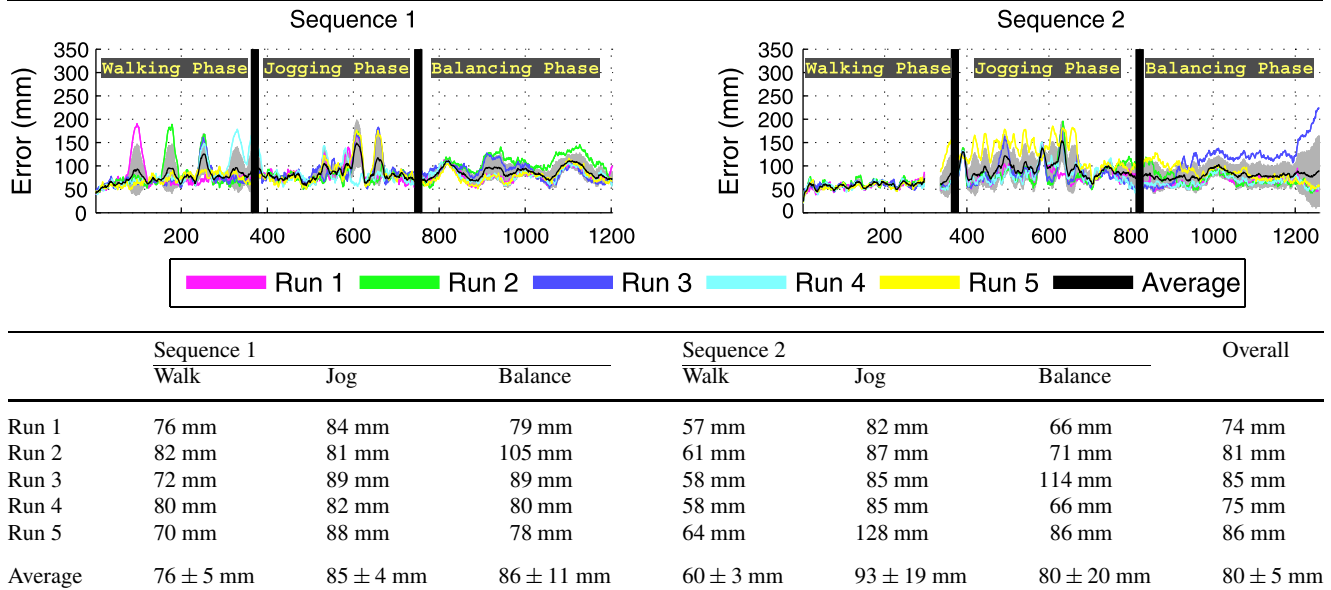


Fig. 6 Stability of the baseline algorithm. The *BC* was run 5 times to establish the stability of the method. Errors for each run in the two sequences are plotted along with the standard deviation of the error in gray as a gray band in the plots. The table shows the error for each run along with the average and standard deviation

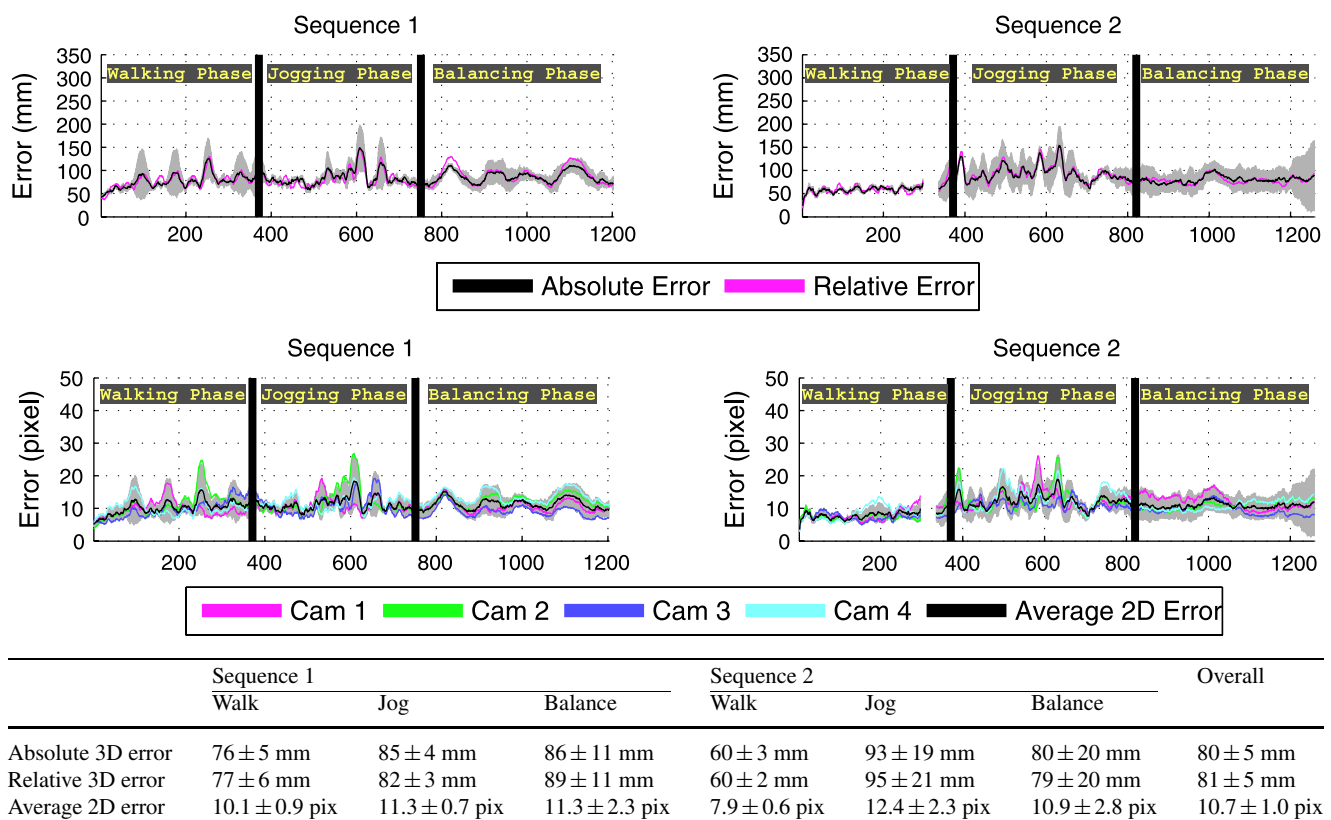


Fig. 7 Performance for various error measures. The performance for *BC* is shown for the various error measures. Top plots: absolute and relative 3D error. Middle plots: 2D pixel error for each camera and averaged over all cameras. The absolute error is given by the average distance between predefined locations on the body (4), while the relative error is computed after first globally aligning the ground truth model and the estimated model at the pelvic joint. We found the two measures to be comparable as the pelvis location is fairly well estimated by the tracker

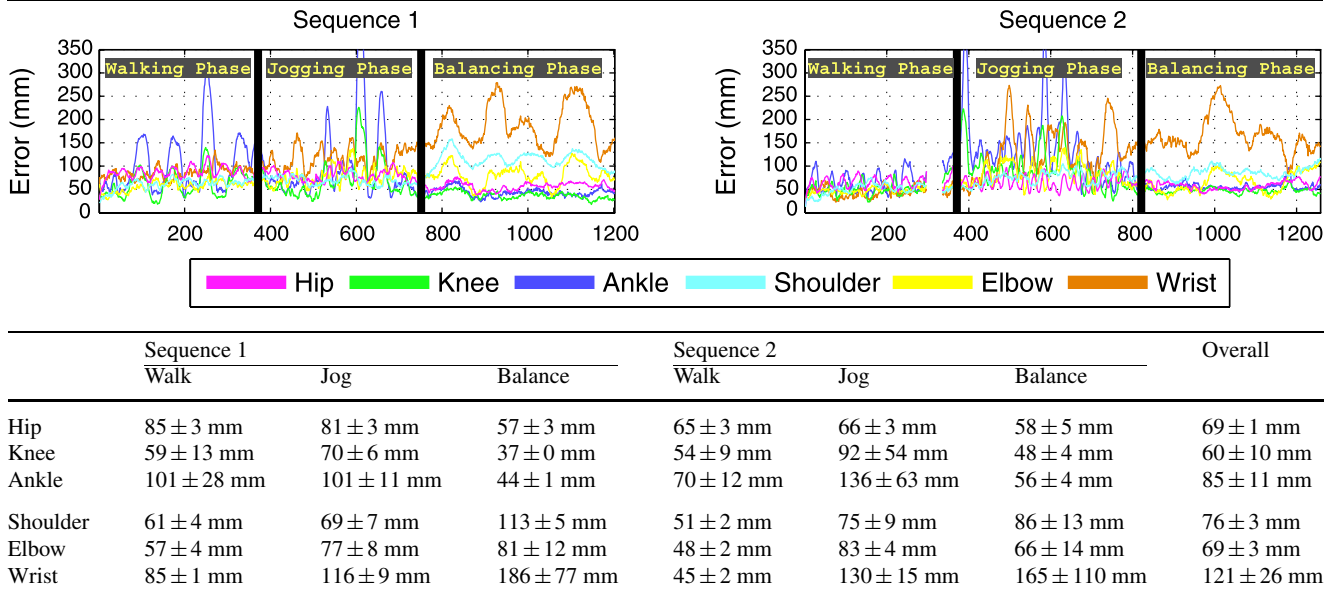


Fig. 8 Error for selected individual joint locations. Errors for individual joint locations averaged over the left and right side illustrate that the arms are harder to track than the legs due to occlusions by the torso. Limb extremities such as wrists and ankles are less constrained and tend to get lost more often than the shoulders and hips. The spikes in error for the ankles correspond to the two legs swapping places. The results come from tracking using the *BC* configuration

gets stuck in this position (e.g. frame 1030 in Fig. 15). From trial to trial, these events may or may not happen due to the stochastic nature of the optimization, making the error variance in these cases higher (identified in the plots in Fig. 6 as spikes in the gray overlay).

The results also highlight the relative degree of difficulty of the two sequences. They are relatively similar, except for the jogging phase where the second sequence is significantly more difficult to track than the first and presents a larger variance in performance. This is consistent with the fact that the second subject is jogging faster.

6.3 Comparing Temporal Diffusion Models

Recall that our APF implementation uses a Gaussian diffusion model to sample new poses. This is a very weak form of temporal prior which does not try to predict the pose at the next time instant from the current one; rather it adds Gaussian noise to the current pose to expand the search range at the next time instant. This diffusion model depends on the choice of the “sampling covariance” Σ .

The two test sequences contain walking and jogging styles, followed by balancing on both legs in turn. Training data, however, only covered walking and jogging. We have therefore considered the following subject-independent action models:

- Walking-style Action Model (**W**)—all walking training data were used to learn the sampling covariance and the joint angle limits.

- Walking and Jogging-style Action Model (**WJ**)—all walking and jogging training data were used to learn the sampling covariance and the joint angle limits.
- Generic Action Model with Anatomic Joint Limits (**G-AJL**)—all training data were used to learn the sampling covariance; joint limits were not derived from training data, but instead were set to bio-mechanical anatomical limits. Note that this is the model used in the *BC*.
- Generic Action Model without Joint Limits (**G-OJL**)—all training data were used to learn the sampling covariance; joint angle limits were not enforced.

All other tracking parameters were the same as the *BC*. Tracking results using the different models are shown in Fig. 9.

All of these remain very weak models in that they do not explicitly describe how the body moves. Rather they are heuristics that control how widely to search the state space around a given pose. We found that the most accurate results were obtained when the activity matched the style-specific action model used. The walking model **W** worked well for walking but not for other activities. Walking performance was good in part because the constrained joint limits prevented the legs from swapping places. Adding jogging to the training (**WJ**) increased the sampling covariance and extended the joint limits, and consequently improved performance slightly on balancing without significantly affecting performance on walking. As one might expect, however, the **WJ** model performed significantly better on jogging. Both

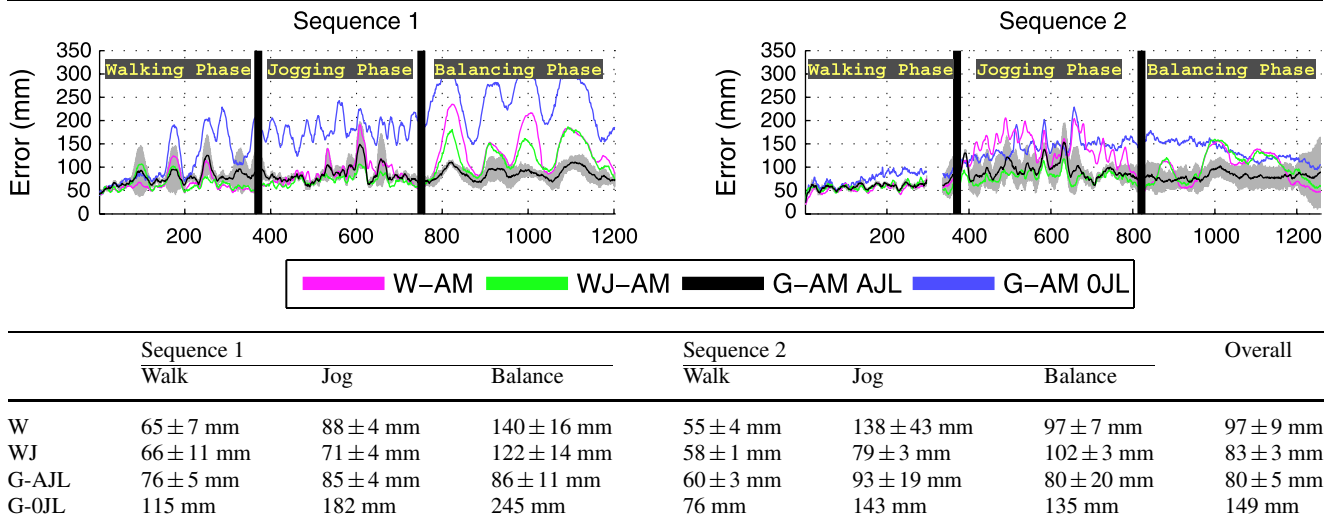


Fig. 9 Comparison of priors. Experimental results for *BC* using different *action models* are shown: walking (**W**), walking and jogging (**WJ**) and generic (**G**). For the generic action model, the joint angle limits were not derived from training data as in the case of walking or jogging, but rather were set to standard anatomical joint limits (**AJL**). We use **OJL** to denote when joint limits were not enforced. In this case performance degraded considerably and we only show results for one run. For **W**, **WJ** and **G-AJL** results were averaged over 5 different runs to more effectively compare different action models

W and **WJ** failed to generalize to the balancing style because the joint angle limits were too narrow at the hips and shoulders.

The **G-AJL** extends the joint limits to anatomical values and performed very well on the balancing portion. This is expected since balancing is a very simple motion and is intuitively easier to track than jogging for example. Clearly the learned anatomical joint limits for **W** and **WJ** prevented these models from generalizing to new poses. In addition, **G-AJL** was able to generalize well to each of the 3 different styles, remaining relatively competitive with the style-specific actions models. Finally, the performance of **G-OJL** illustrates the importance of enforcing joint angle limits. In this model the lack of such limits led to tracking failure even during the walking motions.

The experiments suggest that a generic action model with anatomic joint limits is the optimal choice for sequences with free-style motion.

6.4 Comparing Likelihood Functions

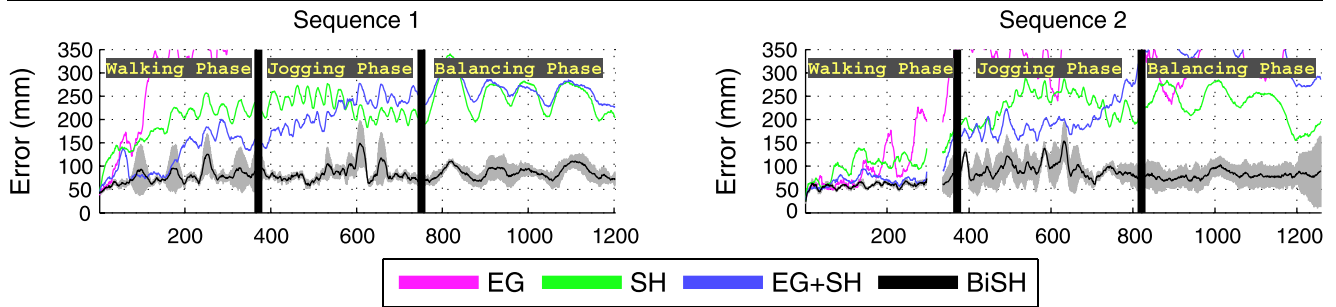
The bi-directional silhouette likelihood **BiS** provides symmetric constraints between the image and model silhouettes, but it is computationally expensive. The standard asymmetric silhouette likelihood **S** is computationally more efficient, but provides weaker constraints. Previous work (Balan et al. 2005; Deutscher and Reid 2005) has shown **S** performs well when combined with the edge likelihood **E** using (18), which we denote by **E + S**. We compare **BiS** with **E + S**, as well as with **E** and **S** separately, all in the context of the *BC* which uses a weak prior on motion (**G-AJL**).

The results in Fig. 10 illustrate that the **BiS** likelihood was the only one capable of tracking the subject over the full length of both sequences, with no other likelihood being able to cope with the fast jogging motion. For the first sequence even the walking motion turned out to be too hard to track. We therefore concentrate our analysis of the likelihoods on sequence 2 during the walking phase only.

We found that relying solely on edges caused the model to drift off the subject and onto the background, with little chance of recovering from tracking failures. Edges do help improve the performance of the standard silhouette likelihood during walking, which otherwise performs poorly as well.

We attribute the fact that the **E + S** likelihood eventually loses track to the combination of a simple likelihood formulation with a weak generic prior **G-AJL** that together allow for improbable poses that explain only part of the image observations. To test this, we combined the same likelihood with a more specific prior (**WJ**), and found it performed much better on walking and jogging data even with half the number of particles (*cf.* Fig. 10). This is consistent with the results reported in (Balan et al. 2005). At the same time, the stronger **BiS** can cope with the weaker prior.

Therefore, for methods that rely on strong priors, simple image observations may be enough, but in the absence of appropriate priors, richer image observation measurements are necessary. Clearly better edge detection methods could be employed and integration of edge information along the entire boundary (instead of sampling) might improve results.

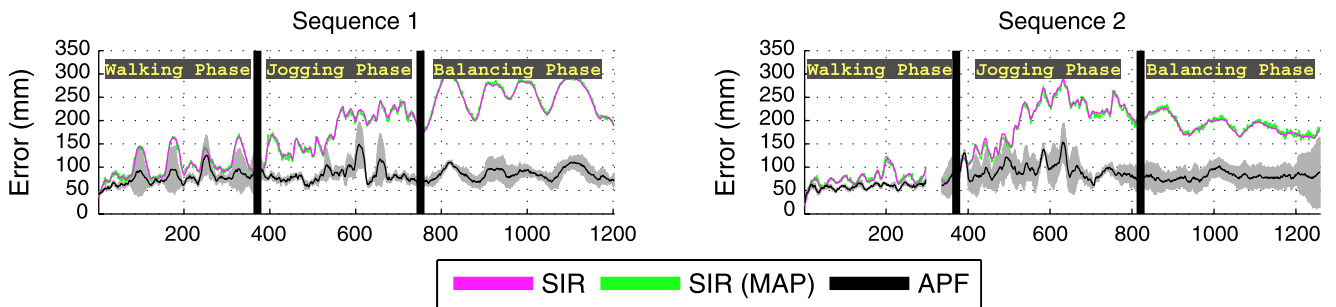


	Sequence 1			Sequence 2			Overall
	Walk	Jog	Balance	Walk	Jog	Balance	
E ^a	315 mm	1367 mm	1298 mm	116 mm	530 mm	344 mm	662 mm
S	184 ± 20 mm	229 ± 32 mm	253 ± 90 mm	105 ± 22 mm	229 ± 10 mm	233 ± 77 mm	205 ± 8 mm
E+S	118 ± 32 mm	216 ± 43 mm	265 ± 133 mm	75 ± 7 mm	199 ± 24 mm	356 ± 116 mm	205 ± 22 mm
BiS	76 ± 5 mm	85 ± 4 mm	86 ± 11 mm	60 ± 3 mm	93 ± 19 mm	80 ± 20 mm	80 ± 5 mm
E+S ^b	90 mm	95 mm	161 mm	68 mm	100 mm	132 mm	111 mm

^aPerformance with the edge likelihood alone was so poor that we only show results for a single run

^bThis experiment was run once and differs from *BC* in that it uses the WJ action model with only 100 particles instead of the G-AJL action model with 200 particles. Its plot is not shown in the graph above

Fig. 10 Comparison of likelihoods. Edge, standard silhouette and bi-directional silhouette likelihoods are compared. The bi-directional model is more computationally expensive, but it is the only one able to completely track the subject using a generic prior. The E+S model is shown to be competitive when combined with a stronger prior that matches the test motion



	Sequence 1			Sequence 2			Overall
	Walk	Jog	Balance	Walk	Jog	Balance	
SIR	101 ± 6 mm	178 ± 36 mm	251 ± 204 mm	75 ± 4 mm	201 ± 35 mm	188 ± 148 mm	166 ± 50 mm
SIR (MAP)	101 ± 6 mm	176 ± 36 mm	254 ± 203 mm	76 ± 5 mm	198 ± 35 mm	190 ± 147 mm	166 ± 49 mm
APF	76 ± 5 mm	85 ± 4 mm	86 ± 11 mm	60 ± 3 mm	93 ± 19 mm	80 ± 20 mm	80 ± 5 mm

Fig. 11 Algorithm comparison. Performance of the Annealed Particle Filter (APF) and Sequential Importance Resampling (SIR) methods is shown. SIR performed significantly worse than APF and started diverging during jogging, which affected the performance during the balancing phase

6.5 Algorithmic Choices

6.5.1 Comparing Regular and Annealed Particle Filtering

The main computational cost in both the APF and SIR is the evaluation of the likelihood for each particle. To fairly compare the methods we keep the number of likelihood evaluations constant across methods. Hence, the number of parti-

cles used for SIR (*i.e.* 1000) is the product of the number of layers (5) and the number of particles per layer (200) in the annealing method. A comparison of the methods is shown in Fig. 11.

In contrast to the APF, the SIR could maintain multi-modal posterior distributions in which case computing the error of the expected pose might not be appropriate. There-

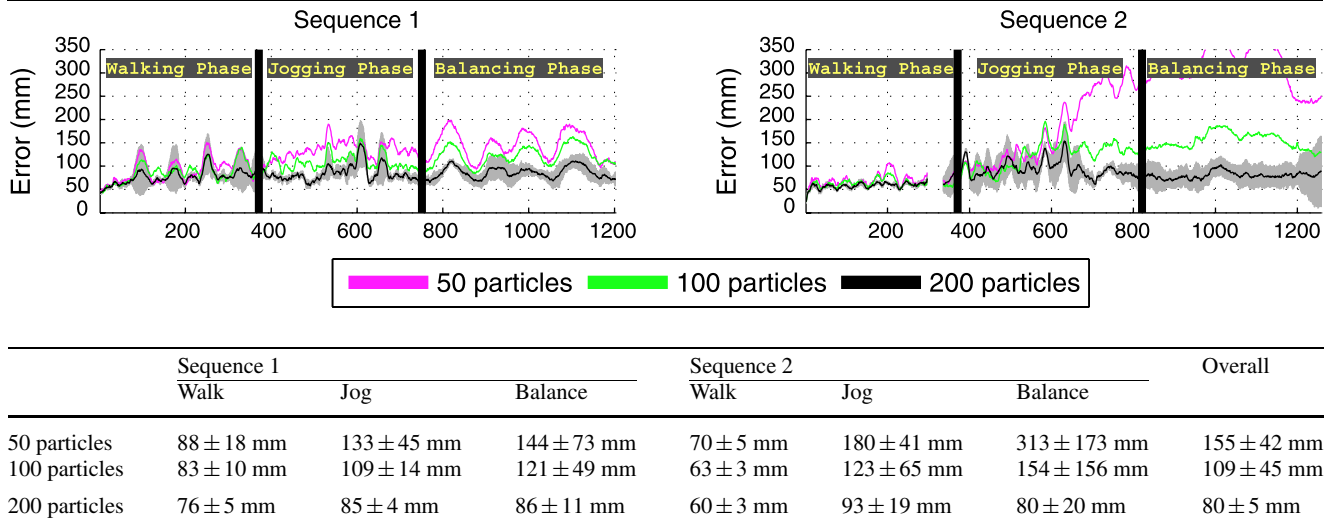


Fig. 12 Number of particles. The effect of the number of particles on accuracy is plotted for the baseline (200) as well as 50 and 100 particles. The number of particles needed depended on the type of motion being tracked, with more particles being needed for fast motions than for slow motions

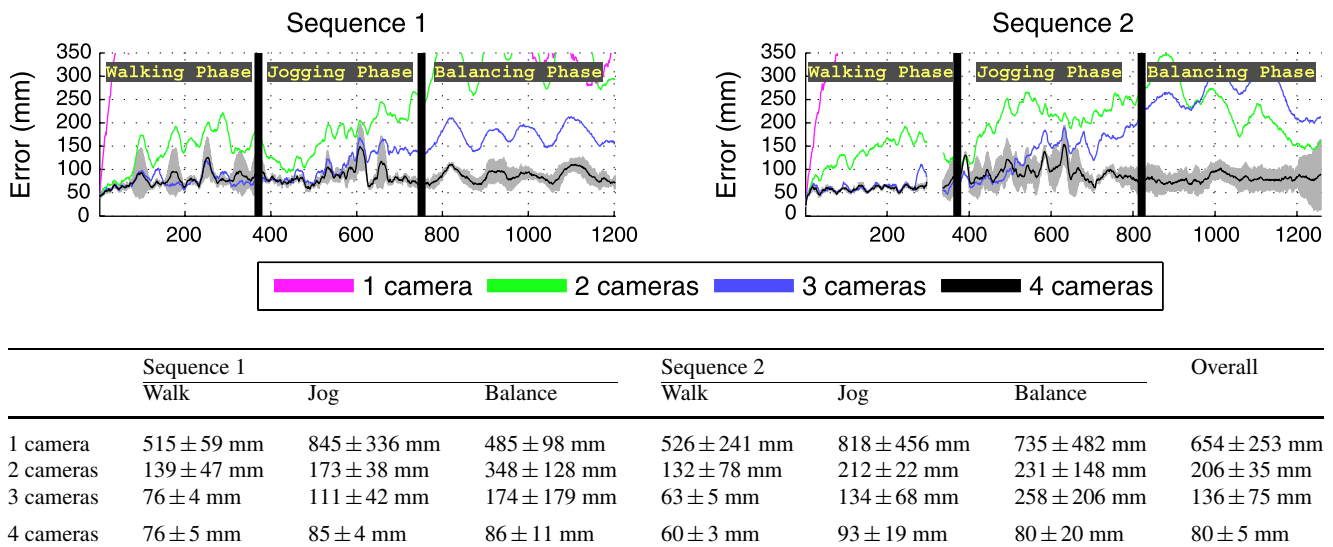


Fig. 13 Number of cameras. At least 3 camera views are needed to track walking motions and at least 4 are needed for more complex motions such as jogging

fore we also report the error of the most likely particle (MAP). We found, however, that the difference in error between the expected pose and the most likely pose was insignificant, and the error curves overlapped. Either way, relative to APF, SIR was significantly worse and more prone to losing track of the subject during fast motions.

6.5.2 Varying the Number of Particles

We also varied the number of particles used in the baseline configuration. Using more particles helps prevent the tracker from losing track and improves performance. The tracker is

much more stable when run using 200 particles. Using 100 particles or fewer makes the tracker unstable as illustrated by the significant increase in error variance in Fig. 12. Based on these results we conclude that the number of particles needed depends on the type of motion being tracked, with more particles being needed for fast motions than for slow motions.

6.5.3 Varying the Number of Camera Views

The HUMANEVA-II dataset used 4 cameras placed on the corners of a rectangular area as shown in Fig. 1. Assessing performance for different number of cameras depends

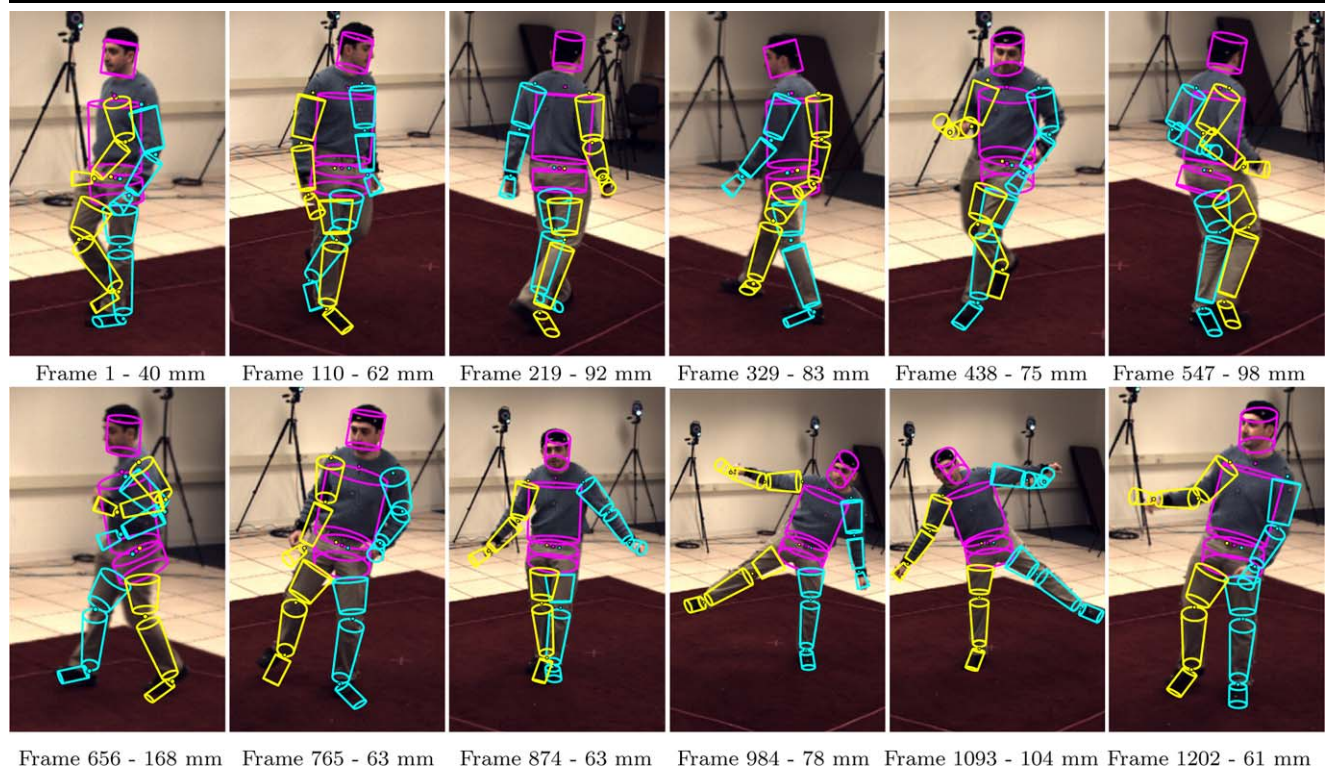


Fig. 14 Tracking results. Typical results obtained on sequence 1 using the baseline tracking configuration (*BC*) are shown for a few frames. The estimated body model is shown projected into the images, with the corresponding 3D error shown underneath. These provide some intuition for different levels of error (in mm). To help discriminate between the left and right sides, we draw the left side in blue and the right side in yellow. For example, frame 656 receives a high error for having mistaken one leg for the other

on the choice of cameras. We ran experiments with *BC* for all subsets of cameras, once for each camera configuration, combining the errors for configurations with the same number of cameras. Mean errors and standard deviations are reported in Fig. 13 over 4 configurations for the one camera case, 6 pairs and 4 triples, respectively.

The results clearly show that monocular tracking is beyond the abilities of the present algorithm. Adding a second camera view significantly improved the results but still the tracker could not cope with simple walking motions. At least 3 camera views were needed to track walking motions and 4 were needed for more complex motions such as jogging. For walking motions there was no statistical difference between using 3 or 4 camera views.

7 Analysis of Performance and Failures

7.1 Model

Our model of the body is an approximation to the true human body shape (though it is fairly typical of the state of the art). We make two key assumptions that (1) the body is made of rigid cylindrical or conical segments and (2) joints only

model the most significant degrees of freedom. We make no attempts to fit the shape of the limbs to the image measurements (Balan et al. 2007). More accurate body models may lead to more accurate tracking results but this hypothesis needs to be verified experimentally. Also, a more anatomically correct modeling of the DoF of the joints may be required for applications in bio-mechanics (Muendermann et al. 2007).

7.2 Image Likelihoods

One of the main observations of our experiments with the baseline algorithm is that results of the approach heavily rely on the quality of the likelihood model. It is our belief that one of the key problems in human motion tracking is the formulation of reliable image likelihood models that are general, do not require background subtraction, and can be applied over a variety of imaging and lighting conditions. We have implemented relatively standard likelihood measures, however, other likelihoods have been proposed and should be evaluated.

For example, more principled edge likelihoods have been formulated using measurable model edge segments (Wachter and Nagel 1999), phase information (Poon and Fleet 2002) and the learned statistics of filter responses

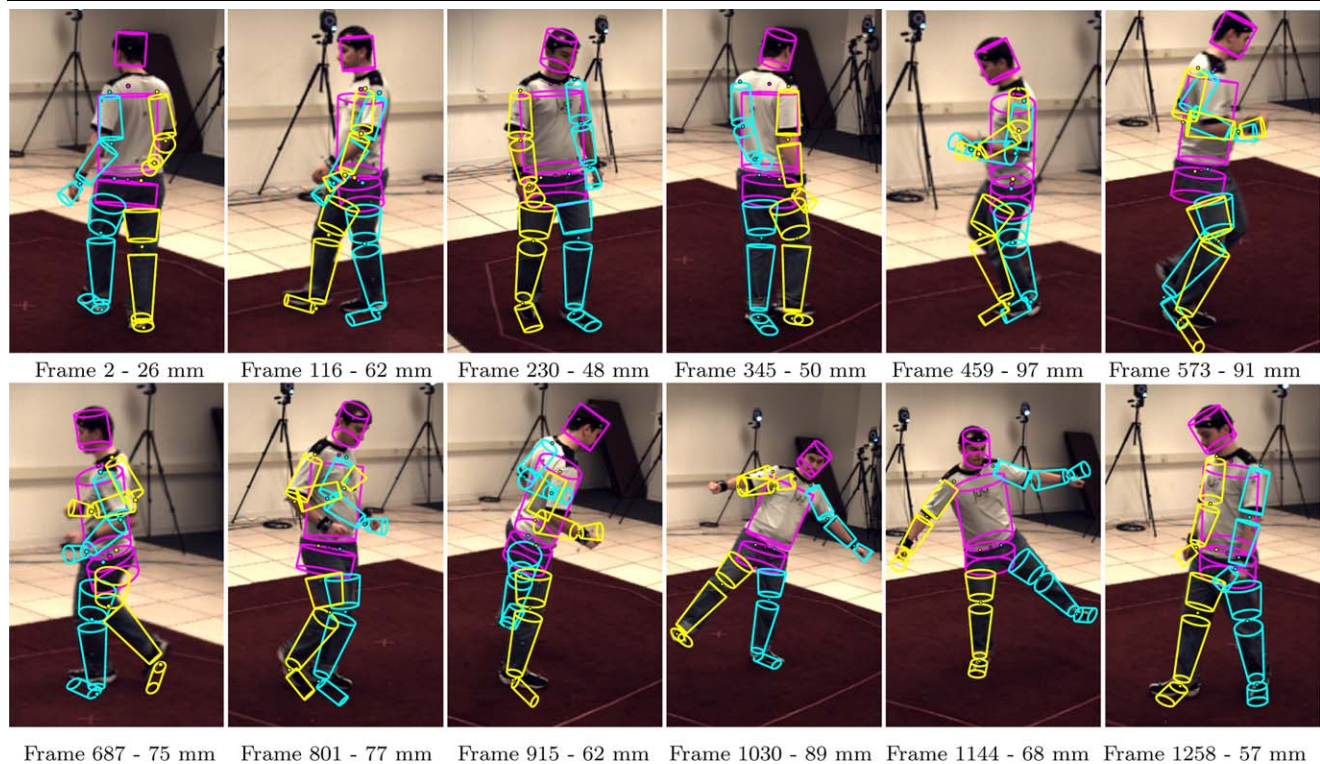


Fig. 15 Tracking results. Typical results obtained on sequence 2 using the baseline tracking configuration are shown

(Roth et al. 2004; Sidenbladh and Black 2003). Non-edge-based likelihood measures include optical flow (Bregler and Malik 1998; Sidenbladh et al. 2000), flow occlusion/disocclusion boundaries (Sminchisescu and Triggs 2003b), segmented silhouettes based on level sets (Rosenhahn et al. 2006), image templates (Wang and Rehg 2006), spatio-temporal templates (Dimitrijevic et al. 2006), principal component-based models of appearance (Sidenbladh et al. 2000), and robust on-line local (Balan and Black 2006; Jepson et al. 2003; Urtasun et al. 2006) and global appearance models (Balan and Black 2006).

7.3 Motion Priors

While the action models used for diffusion within our framework work relatively well in a multi-view setting, it is likely that monocular tracking can benefit from stronger prior models of human motion. The use of strong¹⁵ prior motion models are common with early work concentrating on switching dynamical models (Pavolovic et al. 1999) and eigen-models of cyclic motions (Ormoneit et al. 2000, 2001; Sidenbladh et al. 2000). More recently, motion priors that utilize latent spaces as a means of modeling classes of motions that are inherently low-dimensional in nature have be-

come popular. Low-dimensional non-linear latent variable priors were first (to our knowledge) introduced in (Sminchisescu and Jepson 2004) and later extended in (Lu et al. 2007); Gaussian Processes Latent Variable Models (Urtasun et al. 2005), Gaussian Processes Dynamical Models (Urtasun et al. 2006) and Factor Analyzers (Li et al. 2006) are popular and effective choices particularly for instances where little training data is available. Weaker implicit priors that utilize motion capture data directly (Sidenbladh et al. 2002) have also been effective. Lastly, priors based on abstracted (Brubaker et al. 2007) or full-body (Vondrak et al. 2008) physical simulations recently have been proposed for specific classes of motions (*e.g.* walking).

7.4 Inference

While we explored two inference algorithms, SIR and APF, other promising methods do exist and may lead to more robust or faster performance. For example, hybrid Monte Carlo sampling (Poon and Fleet 2002), partitioned sampling (MacCormick and Isard 2000), or covariance-scaled sampling (Sminchisescu and Triggs 2003b) are all promising alternatives. Kalman filtering (Wachter and Nagel 1999) is another alternative that may be appropriate for the applications where one can ensure that the likelihood and the dynamics are uni-modal.

¹⁵By strong prior motion models here we mean models that bias inference towards a particular pre-defined class of motions.

7.5 Failures

We have observed that it is generally harder to track the upper body, due to frequent occlusions between the arms and the torso. We attribute these difficulties to the poor likelihood functions that are not able to effectively model internal structure within the silhouette region. The upper body also tends to exhibit more stylistic variation across people; the lower body must provide support and hence is more constrained by the dynamics of the motion itself.

The infrequent failures of the baseline algorithm can be classified into two categories: (1) minor tracking failures for individual body parts and (2) 180-degree rotation in the overall body pose; the latter is much harder to recover from in practice. We suspect these failures at least to some extent can be attributed to the nature of annealing which may not represent multi-modal distributions in the posterior effectively.

8 Conclusions and Discussions

We have introduced a dataset for evaluation of human pose estimation and tracking algorithms. This is a comprehensive dataset that contains synchronized video from multiple camera views, associated 3D ground truth, quantitative evaluation measures, and a baseline human tracking algorithm. All the data and associated software is made freely available to the research community.¹⁶ We hope that this dataset will lead to further advances in articulated human motion estimation as well as provide the means of establishing the state of the art performance of current algorithms.

While not new, the baseline algorithm, in addition to providing performance against which future advances on this data can be measured, is designed to serve as a test-bed for future experiments with likelihood functions, prior models and inference methods within the context of Bayesian filtering. We found that the annealed particle filter with 5 layers and 200 particles per layer worked reliably in practice (better than SIR) and that four camera views were necessary for stable tracking. Furthermore we found that the bi-directional silhouette likelihood performed significantly better than the edges and/or standard silhouettes. A fairly weak (generic) “prior” (embodied here as the sampling covariance) that enforced anatomic joint limits and non-interpenetration of parts worked well across activities; stronger models should be explored.

While we treat the marker-based motion capture data as the “ground truth”, it is worth noting that the *true* human motion is somewhat elusive. Even with perfect marker-based motion capture data, deriving the location of joints in

the human body is not a trivial task. For example, hip joints are not well defined and can only be measured to about 2–10 (mm) accuracy given the marker protocol employed by the Vicon system (Camomilla et al. 2006). The true gold standard in localizing the position of hip joints is still debated in the bio-mechanics literature (Corazza et al. 2007). The placement of markers over regular clothes and limits on the calibration accuracy of the video cameras with respect to the Vicon calibration may lead to additional errors that are hard to quantify. While currently unavailable, it is clear that other methods of simultaneously capturing video and motion capture data are necessary if not to allow better ground truth, then to at least lift the need of performing the motion in a laboratory environment. Current research in non-marker-based methods for capturing human motion (Vlasic et al. 2008) may prove to be viable alternatives in a few years.

References

- Agarwal, A., & Triggs, B. (2004a). Learning to track 3D human motion from silhouettes. In *International conference on machine learning (ICML)* (pp. 9–16).
- Agarwal, A., & Triggs, B. (2004b). 3D human pose from silhouettes by relevance vector regression. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 882–888).
- Arulampalam, S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174–188.
- Baker, S., Scharstien, D., Lewis, J. P., Roth, S., Black, M. J., & Szeliski, R. (2007). A database and evaluation methodology for optical flow. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Balan, A., Sigal, L., Black, M. J., Davis, J., & Haussecker, H. (2007). Detailed human shape and pose from images. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Balan, A., & Black, M. J. (2006). An adaptive appearance model approach for model-based articulated object tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 758–765).
- Balan, A., Sigal, L., & Black, M. (2005). A quantitative evaluation of video-based 3D person tracking. In *IEEE workshop on visual surveillance and performance evaluation of tracking and surveillance (VS-PETS)* (pp. 349–356).
- Bissacco, A., Yang, M.-H., & Soatto, S. (2007). Fast human pose estimation using appearance and motion via multi-dimensional boosting, regression. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Bo, L., Sminchisescu, C., Kanaujia, A., & Metaxas, D. (2008). Fast algorithms for large scale conditional 3D prediction. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Bouguet, J.-Y. Camera calibration toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/, accessed on 7/24/2009.
- Bregler, C., & Malik, J. (1998). Tracking people with twists and exponential maps. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 8–15).
- Brubaker, M., Fleet, D. J., & Hertzmann, A. (2007). Physics-based person tracking using simplified lower-body dynamics. In *IEEE*

¹⁶Data and code available at <http://vision.cs.brown.edu/humaneva/>.

- conference on computer vision and pattern recognition (CVPR) (pp. 1–8).
- Camomilla, V., Cereatti, A., Vannozzi, G., & Cappozzo, A. (2006). An optimized protocol for hip joint centre determination using the functional method. *Journal of Biomechanics*, 39(6), 1096–1106.
- CMU Motion Capture Database, <http://mocap.cs.cmu.edu/>, accessed on 7/24/2009.
- Corazza, S., Mündermann, L., & Andriacchi, T. (2007). A framework for the functional identification of joint centers using markerless motion capture, validation for the hip joint. *Journal of Biomechanics*, 40(15), 3510–3515.
- Deutscher, J., & Reid, I. (2005). Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2), 185–205.
- Doucet, A., Godsil, S. J., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3), 197–208.
- Dimitrijevic, M., Lepetit, V., & Fua, P. (2006). Human body pose detection using bayesian spatio-temporal, templates. *Computer Vision and Image Understanding*, 104(2), 127–139.
- Fathi, A., & Mori, G. (2007). Human pose estimation using motion, exemplars. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Felzenszwalb, P., & Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 55–79.
- Gall, J., Rosenhahn, B., Brox, T., Kersting, U., & Seidel, H.-P. (2006). Learning for multi-view 3D tracking in the context of particle filters. In *LNCS: Vol. 4292. International symposium on visual computing (ISVC)* (pp. 59–69). Berlin: Springer.
- Gavrila, D. (1999). The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1), 82–98.
- Gavrila, D., & Davis, L. (1996). 3-D model-based tracking of humans in action: a multi-view approach. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 73–80).
- Grauman, K., Shakhnarovich, G., & Darrell, T. (2003). Inferring 3D structure with a statistical image-based shape model. In *IEEE international conference on computer vision (ICCV)* (pp. 641–648).
- Gross, R., & Shi, J. (2001). The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18. Robotics Institute, Carnegie Mellon University.
- Hogg, D. C. (1983). Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1, 5–20.
- Hough, P. V. C. (1962). *Method and means for recognizing complex patterns*. U.S. Patent 3,069,654.
- Hua, G., Yang, M.-H., & Wu, Y. (2005). Learning to estimate human pose with data driven belief propagation. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 747–754).
- Isard, M., & Blake, A. (1998). Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1), 5–28.
- Jepson, A., Fleet, D., & El-Maraghi, T. (2003). Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 1296–1311.
- Ju, S., Black, M., & Yacoob, Y. (1996). Cardboard people: a parameterized model of articulated motion. In *International conference on automatic face and gesture recognition* (pp. 38–44).
- Kakadiaris, I. A., & Metaxas, D. (1996). Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 81–87).
- Knossow, D., Ronfard, R., & Horaud, R. (2008). Human motion tracking with a kinematic parameterization of extremal contours. *International Journal of Computer Vision*, 79(3), 247–269.
- Lan, X., & Huttenlocher, D. (2005). Beyond trees: common factor models for 2D human pose recovery. In *IEEE international conference on computer vision (ICCV)*, vol. 1 (pp. 470–477).
- Lan, X., & Huttenlocher, D. (2004). A unified spatio-temporal articulated model for tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 722–729).
- Lee, C.-S., & Elgammal, A. (2007). Modeling view and posture manifold for tracking. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Lee, M., & Nevatia, R. (2006). Human pose tracking using multi-level structured models. In *European conference on computer vision (ECCV)*, vol. 3 (pp. 368–381).
- Li, R., Tian, T.-P., & Sclaroff, S. (2007). Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Li, R., Yang, M.-H., Sclaroff, S., & Tian, T.-P. (2006). Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers. In *European conference on computer vision (ECCV)*.
- Lu, Z., Perpinan, M. C., & Sminchisescu, C. (2007). People tracking with the laplacian eigenmaps latent variable model. In *Advances in neural information processing systems (NIPS)*, vol. 2 (pp. 137–150).
- MacCormick, J., & Isard, M. (2000). Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European conference on computer vision (ECCV)*, vol. 2 (pp. 3–19).
- Moeslund, T., & Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 18, 231–268.
- Mori, G. (2005). Guiding model search using segmentation. In *IEEE international conference on computer vision (ICCV)* (pp. 1417–1423).
- Mori, G., Ren, X., Efros, A., & Malik, J. (2004). Recovering human body configurations: combining segmentation and recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 326–333).
- Muenderrmann, L., Corazza, S., & Andriacchi, T. (2007). Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Navaratnam, R., Fitzgibbon, A., & Cipolla, R. (2007). The joint manifold model for semi-supervised multi-valued regression. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Ning, H., Xu, W., Gong, Y., & Huang, T. (2008). Discriminative learning of visual words for 3D human pose estimation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Ormoneit, D., Sidenbladh, H., Black, M. J., & Hastie, T. (2001). Learning and tracking cyclic human motion. In *Advances in neural information processing systems (NIPS)*, vol. 13 (pp. 894–900).
- Ormoneit, D., Sidenbladh, H., Black, M. J., & Hastie, T. (2000). Stochastic modeling and tracking of human motion, *Learning 2000*, Snowbird, UT.
- O'Rourke, J., & Badler, N. I. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6), 522–192.
- Pavlovic, V., Rehg, J., Cham, T.-J., & Murphy, K. (1999). A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *IEEE international conference on computer vision (ICCV)* (pp. 94–101).
- Phillips, P. J., Blackburn, D., Bone, M., Grother, P., Micheals, R., & Tabassi, E. (2002). Face recognition vendor test. <http://www.frvr.org/>.
- Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms.

- IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), 1090–1104.
- Poon, E., & Fleet, D. (2002). Hybrid Monte Carlo filtering: edge-based people tracking. In *IEEE workshop on motion and video computing* (pp. 151–158).
- Ramanan, D., Forsyth, D., & Zisserman, A. (2005). Strike a pose: tracking people by finding stylized poses (CVPR). In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 271–278).
- Ramanan, D., & Forsyth, D. (2003). Finding and tracking people from the bottom up. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 467–474).
- Ren, X., Berg, A., & Malik, J. (2005). Recovering human body configurations using pairwise constraints between parts. In *IEEE international conference on computer vision (ICCV)*, vol. 1 (pp. 824–831).
- Roberts, T., McKenna, S., & Ricketts, I. (2004). Human pose estimation using learnt probabilistic region similarities and partial configurations. In *European conference on computer vision (ECCV)*, vol. 4 (pp. 291–303).
- Rogez, G., Rihan, J., Ramalingam, S., Oritte, C., & Torr, P. H. S. (2008). Randomized trees for human pose estimation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Ronfard, R., Schmid, C., & Triggs, B. (2002). Learning to parse pictures of people. In *European conference on computer vision (ECCV)*, vol. 4 (pp. 700–714).
- Rosales, R., & Sclaroff, S. (2000). Inferring body pose without tracking body parts. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 721–727).
- Rosenhahn, B., Brox, T., Kersting, U., Smith, D., Gurney, J., & Klette, R. (2006). A system for marker-less human motion estimation. *Kuenstliche Intelligenz*, 1, 45–51.
- Roth, S., Sigal, L., & Black, M. J. (2004). Gibbs likelihoods for Bayesian tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 886–893).
- Sarkar, S., Phillips, P. J., Liu, Z., Robledo, I., Grother, P., & Bowyer, K. W. (2005). The human ID gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2), 162–177.
- Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3), 7–42.
- Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In *IEEE international conference on computer vision (ICCV)*, vol. 2 (pp. 750–759).
- Sidenbladh, H., & Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54(1–3), 183–209.
- Sidenbladh, H., Black, M. J., & Sigal, L. (2002). Implicit probabilistic models of human motion for synthesis and tracking. In *European conference on computer vision (ECCV)*, vol. 1 (pp. 784–800).
- Sidenbladh, H., De la Torre, F., & Black, M. J. (2000). A framework for modeling the appearance of 3D articulated figures. In *International conference on automatic face and gesture recognition (FG)* (pp. 368–375).
- Sidenbladh, H., Black, M., & Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *European conference on computer vision (ECCV)*, vol. 2 (pp. 702–718).
- Sigal, L., Bhatia, S., Roth, S., Black, M., & Isard, M. (2004). Tracking loose-limbed people. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 421–428).
- Sigal, L., & Black, M. (2006). Measure locally, reason globally: occlusion-sensitive articulated pose estimation. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 2041–2048).
- Sminchisescu, C., Kanaujia, A., Li, Z., & Metaxas, D. (2005). Discriminative density propagation for 3D human motion estimation. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 390–397).
- Sminchisescu, C., & Jepson, A. (2004). Generative modeling for continuous non-linearly embedded visual inference. In *International conference on machine learning (ICML)* (pp. 759–766).
- Sminchisescu, C., & Triggs, B. (2003a). Kinematic jump processes for monocular 3D human tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 69–76).
- Sminchisescu, C., & Triggs, B. (2003b). Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6), 371–391.
- Sminchisescu, C., & Telea, A. (2002). Human pose estimation from silhouettes a consistent approach using distance level sets. In *International conference on computer graphics, visualization and computer vision (WSCG)*.
- Sminchisescu, C. (2002). Consistency and coupling in human model likelihoods. In *International conference on automatic face and gesture recognition (FG)* (pp. 27–32).
- Srinivasan, P., & Shi, J. (2007). Bottom-up recognition and parsing of the human body. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Taylor, C. J. (2000). Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding*, 80(3), 349–363.
- Urtasun, R., & Darrell, T. (2008). Local probabilistic regression for activity-independent human pose inference. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3D people tracking with gaussian process dynamical models. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 238–245).
- Urtasun, R., Fleet, D. J., Hertzmann, A., & Fua, P. (2005). Priors for people tracking from small training sets. In *IEEE international conference on computer vision (ICCV)*, vol. 1 (pp. 403–410).
- Vlasic, D., Baran, I., Matusik, W., & Popović, J. (2008). Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3), 1–9.
- Vondrak, M., Sigal, L., & Jenkins, O. C. (2008). Physical simulation for probabilistic motion tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).
- Wang, P., & Rehg, J. M. (2006). A modular approach to the analysis and evaluation of particle filters for figure tracking. In *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1 (pp. 790–797).
- Wachter, S., & Nagel, H. H. (1999). Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3), 174–192.
- Xu, X., & Li, B. (2007). Learning motion correlation for tracking articulated human body with a Rao-Blackwellised particle filter. In *IEEE international conference on computer vision (ICCV)* (pp. 1–8).
- Zhang, J., Luo, J., Collins, R., & Liu, Y. (2006). Body localization in still images using hierarchical models and hybrid search. In *IEEE international conference on computer vision and pattern recognition (CVPR)*, vol. 2 (pp. 1536–1543).