Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low Number of Cameras

A. Elhayek, MPI Informatics
E. de Aguiar, MPI Informatics
A. Jain, New York University
J. Tompson, New York University
L. Pishchulin, MPI Informatics
M. Andriluka, Stanford University
C. Bregler, New York University
B. Schiele, MPI Informatics
C. Theobalt, MPI Informatics

Abstract

We present a novel method for accurate marker-less capture of articulated skeleton motion of several subjects in general scenes, indoors and outdoors, even from input filmed with as few as two cameras. Our approach unites a discriminative image-based joint detection method with a model-based generative motion tracking algorithm through a combined pose optimization energy. The discriminative part-based pose detection method, implemented using Convolutional Networks (ConvNet), estimates unary potentials for each joint of a kinematic skeleton model. These unary potentials are used to probabilistically extract pose constraints for tracking by using weighted sampling from a pose posterior guided by the model. In the final energy, these constraints are combined with an appearance-based model-to-image similarity term. Poses can be computed very efficiently using iterative local optimization, as ConvNet detection is fast, and our formulation yields a combined pose estimation energy with analytic derivatives. In combination, this enables to track full articulated joint angles at state-of-the-art accuracy and temporal stability with a very low number of cameras.

1. Introduction

Optical motion capture methods estimate the articulated joint angles of moving subjects from multi-view video recordings. Motion capture has many applications, for instance in sports, biomedical research, or computer animation. While most commercial systems require markers on the human body, marker-less approaches developed in research work directly on unmodified video streams [27, 32, 35]. Latest work shows that marker-less skeletal motion tracking is also feasible in a less controlled studio setting and outdoors, as well as in front of more general backgrounds where foreground segmentation is hard [15, 21]. Commonly these methods rely on a kinematic skeleton model with attached shape proxies, and they track the motion by optimizing an alignment metric between model and images in terms of the joint angles. Formulating and optimizing this usually highly non-convex energy is difficult. Global optimization of the pose is computationally expensive, and thus local methods are often used for efficiency, at the price of risking convergence to a wrong pose. With a sufficiently high number of cameras (≥ 8), however, efficient high accuracy marker-less tracking is feasible with local pose optimizers. Unfortunately, this strategy starts to fail entirely if only $2-3$ cameras are available, even when recording simple scenes inside a studio.

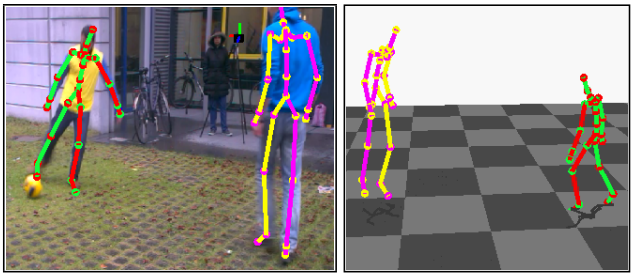

Figure 1. Our ConvNet-based marker-less motion capture algorithm reconstructs joint angles of multiple people performing complex motions in outdoor settings, such as in this scene recorded with only three mobile phones: (left) 3D pose overlaid with one camera view, (right) 3D visualization of captured skeletons.

In a separate strand of work, researchers developed learning-based discriminative methods for body part detection in a single image. Since part detection alone is often unreliable, it is often combined with higher-level graphical models, such as in pictorial structures [3], to improve robustness of 2D part or joint localization. Recently, these 2D pose estimation methods were extended to the multi-view case, yielding 3D joint positions from a set of images taken at the same time step [5]. Detection-based pose estimation can compute joint locations from a low number of images taken under very general conditions. However, accuracy of

estimated joint locations is comparably low, mainly due to the uncertainty in the part detections, and pose computation is far from real-time. Also, the approaches merely deliver joint positions, not articulated joint angles, and results on video exhibit notable jitter.

This paper describes a new method to fuse marker-less skeletal motion tracking with body part detections from a convolutional network (ConvNet) for efficient and accurate marker-less motion capture with few cameras. Through fusion, the individual strengths of either strategy are fruitfully enforced and individual weaknesses compensated. The core contribution is a new way to combine evidence from a ConvNet-based monocular joint detector [39] with a model-based articulated pose estimation framework [38]. This is done by a new weighted sampling from a pose posterior distribution guided by the articulated skeleton model using part detection likelihoods. This yields likely joint positions in the image with reduced positional uncertainty, which are used as additional constraints in a pose optimization energy. The result is one of the first algorithms to capture temporally stable full articulated joint angles from as little as 2-3 cameras, also of multiple actors in front of moving backgrounds. We tested our algorithm on challenging indoor and outdoor sequences filmed with different video and mobile phone cameras, on which model-based tracking alone fails. The high accuracy of our method is shown through quantitative comparison against marker-based motion capture, marker-less tracking with many cameras, and detection-based 3D pose estimation methods. Our approach can also be applied in settings where other approaches for pose estimation with a low number of sensors, that are based on depth cameras [4] or inertial sensors [31], are hard or impossible to be used, e.g. outdoors. The accuracy and stability of our method is achieved by carefully and cleverly combining all input information (i.e. 2D detections, the pose of the previous frame, several views, the 3D-model, and camera calibration). For instance, our method provides 1) strategies to select the correct scale of the ConvNet; 2) strategies to avoid tracking failure by weighting the final contribution of each estimate and by limiting the search space; 3) a new term which carefully integrates the body part detections from all cameras.

2. Related Work

Human motion capture algorithms from input video have seen great advances in recent years. We refer the reader to the surveys [27, 32, 35] for a detailed overview. In summary, the approaches can be divided into two categories: methods based on multi-view input and methods that rely on a single view.

The majority of the multi-view approaches combine a body model of the subject to be tracked, represented as a triangle mesh or simple primitives, with data extracted from the input images. Usually, they differ in the type of image features used and in the way optimization is performed. The multi-layer framework presented in [18] uses a particle-based optimization to estimate the pose from silhouette and color information. The approaches presented in [6, 25, 26] use training data to learn a motion model or a mapping from image features to the 3D pose. Tracking without silhouette information is also possible by combining segmentation with a shape prior and pose estimation. The approach described in [8] uses graph-cut segmentation, the techniques presented in [9, 19] use a level set segmentation with motion features or an analysis-by-synthesis approach. Alternatively, the approach in [38] presents an analytic formulation based on a Sums-of-Gaussians model. Usually, the performance of the approaches is measured on the HumanEVA benchmark [35].

More recent works make use of additional sensors, such as inertial sensors [31], or depth sensors [4, 43]. Other works try to overcome limitations of the multi-view motion capture approaches, allowing motion tracking with moving or unsynchronized cameras [21, 34, 14, 15]. However, most of them still rely on a sufficiently high number of cameras and they would fail if a small number of cameras are available, even when recording simple scenes.

In a second category of approaches, methods try to infer poses from single-view images, or motions from monocular video. Most of the methods for human pose estimation are based on the pictorial structures (PS) model [17, 16] that represents the body configuration as a collection of rigid parts and a set of pairwise part connections. A large number of algorithms have been proposed [13, 3, 42, 12, 33]. Yang&Ramanan [42] proposed a flexible mixture of templates based on linear SVMs. Approaches that model yet higher-order body-part dependencies have been proposed more recently. Pishchulin et al. [29, 30] model spatial relationships of body-parts using *Poselet* [7] priors and a DPM based part-detector. Sapp&Taskar [33] propose a multi-modal model which includes both holistic and local cues for mode selection and pose estimation. Similar to the *Poselets* method, using a semi-global classifier for part configuration, the *Armlets* approach by Gkioxari et al. [20] shows good performance on real-world data, however, it is demonstrated only on arms. Furthermore, all these approaches suffer from the fact that the features used (HoG features, edges, contours, and color histograms) are hand-crafted and not learnt.

Convolutional networks (ConvNets) are by far the best performing algorithms for many vision tasks. The state-of-the-art methods for human-pose estimation are also based on ConvNets ([40, 22, 39, 23, 11]). Toshev et al. [40] formulate the problem as a direct regression to joint location. Chen et al. [11] improve over [40] by adding an image dependent spatial prior. Jain et al. [22] train an image

patch classifier which is run in a sliding-window fashion at run time. Tompson et al. [39] use a multi-resolution ConvNet architecture to perform heat-map likelihood regression which they train jointly with a graphical model. However, apart from the new advances of these approaches, they still do not reach the same accuracy of multi-view methods, mainly due to the uncertainty in the part detections. In addition, they usually only work on very simplified models with few degrees of freedom, and the results usually exhibit jitter over time.

Only a few methods in the literature are able to combine the individual strengths of both strategies. Using a depth camera, Baak et al. [4] introduce a data-driven hybrid approach combining local optimization with global pose retrieval from a database for real-time full body pose reconstruction. Sridhar et al. [37] also uses a hybrid solution, combining a discriminative part-based pose retrieval technique with a generative pose estimation method, for articulated hand motion tracking using color and depth information. However, to the best of our knowledge, our paper presents one of the first algorithm to fuse marker-less skeletal motion tracking with body part detections from a convolutional network (ConvNet) for efficient and accurate marker-less motion capture with a few consumer cameras. This enables us to accurately capture full articulated motion of multiple people with as little as 2-3 cameras in front of moving backgrounds.

3. Overview

Input to our approach are multi-view video sequences of a scene, yielding n frames $I = I_1^c, ..., I_n^c$ for each static and calibrated camera $c \in C$. Cameras can be of varying types, and resolution, but run synchronized at the same frame rate.

We model each human in the scene with an articulated skeleton, comprising of 24 bones and 25 joints. Joint angles and global pose are parameterized through 48 pose parameters Θ represented as twists. Later, for 13 of the joints - mostly in the extremities - ConvNet detection constraints are computed as part of our fused tracker. In addition, 72 isotropic Gaussian functions are attached to the bones, with each Gaussian's position in space (mean) being controlled by the nearest bone. Each Gaussian is assigned a color, too. This yields an approximate 3D Sum of Gaussians (SoG) representation of an actor's shape, that was first introduced in [38]. In parallel, each input image is subdivided into regions of constant color using fast quad-tree clustering, and to each region a 2D Gaussian is fitted. Before tracking commences, the bone lengths and the Gaussians need to be initialized to match each tracked actor. Depending on the type of sequence (recorded by us or not), we employ an automatic initialization scheme by optimizing bone lengths, as described in [38]. If initialization poses were not captured, the model is manually initialized on the first frame of multi-view video; see [38] for more details.

The baseline generative model-based marker-less motion capture approach by Stoll et al. [38] and its extensions [14, 15] use the above scene representation and estimate pose by optimizing a color- and shape-based model-to-image similarity energy in Θ. This smooth and analytically differentiable energy can be optimized efficiently, which results in full articulated joint angles at state-of-the-art accuracy if enough cameras (typically $\geq$ 8) are available, and if the scene is reasonably controlled, well-lit, and with little background clutter. The method quickly fails, however, if the number of cameras is below five, and if - in addition - scenes are recorded outdoors, with stronger appearance changes, with multiple people in the scene, and with more dynamics and cluttered scene backgrounds.

To make this model-based tracking strategy scale to the latter more challenging conditions, we propose in this paper a new way to incorporate into the pose optimization additional evidence from a machine learning approach for joint localization in images based on ConvNets. ConvNet-based joint detection [39] shows state-of-the-art accuracy for locating joints in single images, even under challenging and cluttered outdoor scenes. However, computed joint likelihood heat-maps are rather coarse, with notable uncertainty, and many false positive detections. Extracting reliable joint position constraints for pose optimization directly from these detections is difficult.

To handle these uncertainties, we propose a model-guided probabilistic way to extract most likely joint locations in the multi-view images from the uncertain ConvNet detections. To this end, the pose posterior for the next frame is approximated by importance sampling with weights from the detection likelihood in the images. Here, the pose prior is modeled reliably based on articulated motion extrapolation from the previous time step's final pose estimate. From the sampled posterior, a most likely image location for each joint is computed, which is incorporated as constraint into the pose optimization energy, Sec. 5. In conjunction, this yields a new pose energy to be optimized for each time frame of multi-view video.

$$\begin{aligned} E(\Theta) =& w_{col} E_{col}(\Theta) + w_{BP} E_{BP}(\Theta) - \\ & w_l E_{lim}(\Theta) - w_a E_{acc}(\Theta) \end{aligned} \quad (1)$$

where $E_{col}(\Theta)$ is a color- and shape-based similarity term between projected body model and images (Sec. 4), $E_{BP}(\Theta)$ is the ConvNet detection term (Sec. 5), and w_{col} and w_{BP} control their weights. $E_{lim}(\Theta)$ enforces joint limits, and $E_{acc}(\Theta)$ is a smoothness term penalizing too strong accelerations [38]. The weights $w_{col} = 1$, $w_{BP} = 5$, $w_l = 0.1$ and $w_a = 0.05$ were found experimentally and are kept constant in all experiments.

This new energy remains to be smooth and analytically differentiable, and can thus be optimized efficiently using

standard gradient ascent initialized with the previous time step's extrapolated pose. ConvNet detections can be computed faster too. By optimizing this new energy we can track full articulated joint angles at state-of-the-art accuracy on challenging scenes with as few as two cameras.

4. Appearance-based Similarity Term

The appearance-based similarity term E_{col} [38] measures the overlap between a 3D model and the 2D SoG images for the images of each camera c. To this end, each 3D model Gaussian is projected using the operator Ψ into camera view c with current pose Θ, yielding a projected model SoG $\mathcal{K}_m(\Theta, c)$. The spatial and color overlap of each projected Gaussian basis function $\mathcal{B}_i(\mathbf{x})$ from $\mathcal{K}_m(\Theta, c)$ and $\mathcal{B}_j(\mathbf{x})$ from the color image SoG $\mathcal{K}_{I^c}$, is computed as:

$$\begin{aligned} &E(\mathcal{K}_{I^c}, \mathcal{K}_m(\Theta, c)) \\ &= \int_\Omega \sum_{i \in \Psi(\mathcal{K}_m(\Theta,c))} \sum_{j \in \mathcal{K}_{I^c}} d(\mathbf{c}_i, \mathbf{c}_j) \mathcal{B}_i(\mathbf{x}) \mathcal{B}_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \sum_{i \in \mathcal{K}_m(\Theta,c)} \sum_{j \in \mathcal{K}_{I^c}} E_{ij}, \end{aligned} \tag{2}$$

E_{ij} is the similarity between a pair of Gaussians $\mathcal{B}_i$ and $\mathcal{B}_j$ given their assigned colors $\mathbf{c}_i$ and $\mathbf{c}_j$:

$$\begin{aligned} E_{ij} &= d(\mathbf{c}_i, \mathbf{c}_j) \int_\Omega \mathcal{B}_i(\mathbf{x}) \mathcal{B}_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= d(\mathbf{c}_i, \mathbf{c}_j) 2\pi \frac{{\sigma_i}^2 {\sigma_j}^2}{{\sigma_i}^2 + {\sigma_j}^2} \exp\left(-\frac{\|\mu_i - \mu_j\|^2}{{\sigma_i}^2 + {\sigma_j}^2} \right). \end{aligned} \tag{3}$$

The smooth function $d(\mathbf{c}_i, \mathbf{c}_j)$ measures the Euclidean distance between $\mathbf{c}_i$ and $\mathbf{c}_j$ in the HSV color space and feeds the result into a Wendland function [41].

To approximate occlusion effects [38], projected 3D Gaussians are prevented from contributing multiple times in Eq. (2), and thus the final appearance similarity term computed over all cameras is

$$\begin{aligned} &E_{col}(\Theta) \\ &= \sum_{c \in C} \sum_{j \in \mathcal{K}_{I^c}} \min\left(\left(\sum_{i \in \Psi(\mathcal{K}_m(\Theta,c))} E_{ij} \right), E_{ii} \right). \end{aligned} \tag{4}$$

5. ConvNet Detection Term

We employ a ConvNet-based localization approach [39] to compute for each of the $n_{prt} = 13$ joints j in the arms, legs and head a Heat-map image $H_{j,c}$ for each camera view c at the current time step (Sect. 5.1). We employ a weighted sampling from a pose posterior guided by the kinematic model to extract most likely 2D joint locations $d_{j,c}$ in each image from the uncertain likelihood maps. These are used as additional constraints in the pose optimization energy (Sect. 5.2).

5.1. ConvNet Joint Detections

We briefly summarize the approach of [39, 23] which we use for 2D part detection. This method achieves state of the art results on several public benchmarks and is formulated as a Convolutional Network [24] to infer the location of 13 joints in monocular RGB images.

The model is a fully convolutional network and is therefore a translation invariant part detector, see [39] for details. It takes as input a single RGB image, creates a 3 level Gaussian pyramid and outputs 13 heat-maps $H_{j,c}$ describing the per-pixel likelihood for each of the 13 joints. Since the network consists of two 2×2 MaxPooling layers, the output heat-maps are at a decimated resolution. We do not explicitly train the ConvNet on frames used in this work, but use a net pre-trained on the MPII Human Pose Dataset [2], which consists of 28,821 training annotations of people in a wide variety of poses and static scenes. Note that training on our own sequences (or sequences similar to ours) may increase accuracy even further.

The first layer of the network is a local contrast normalization layer. This layer - in conjunction with the Gaussian pyramid input - creates 3 resolution images with non-overlapping spectral content. The advantage of this representation is that it promotes specialization amongst the 3 resolution banks, reducing network redundancy and thus improving generalization performance. Furthermore, the use of multiple resolutions increases the amount of spatial context seen by the network without a significant increase in the number of trainable parameters. Each of the 3 images is processed through a 4 stage Convolution-Non-Linearity[1]-MaxPooling network which creates a dense and high-level feature representation for each of the multi-resolution images.

The convolution features are then feed through a 4 layer Convolution-Non-Linearity network which simulates a fully connected neural network over a local receptive field of size 96×96 pixels in the highest resolution image. The first layer of this network (which is implemented as a 9×9 convolution layer) is split across the resolution banks, and then approximated by up-sampling the lower resolution features to bring them into canonical resolution before linearly combining them for processing into the three 1×1 convolution - Non-Linearity layers. To handle persons of different size, we precompute heat-maps $H^s_{j,c}$ at 4 different scales s. A major advantage of the ConvNet detections for 3D human pose estimation is that they do not suffer from the front/back ambiguity. We attribute this to their high discriminative capacity, efficient use of shared (high-level) convolutional features, learned invariance to input image transformations, and large input image context.

[1] For all non-linearity layers we use a Rectified Linear activation [28]

Figure 2. Refinement of the Body Part Detections using the pose posterior. **Left:** Overlay of the heat-map for the right ankle joint over the input image. **Middle:** sampling from pose posterior around the rough 2D position $p_{j,c}^{init}$ (black dots). **Right:** The final refined location of the body part $d_{j,c}$ (blue dot).

5.2. Refining Joint Detections

The joint detection likelihoods in $H_{j,c}^{s}$ exhibit notable positional uncertainty, false positives, and close-by multiple detections in multi-person scenes, Fig. 3 (Left). We therefore propose a new scheme to extract the most likely location $d_{j,c}$ of each joint in each camera view (and for each tracked person if multiple people are in the scene), given the history of tracked articulated poses. Our approach is motivated by weighted sampling from the 3D pose posterior distribution $P(D|\Theta)$.

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta)\ . \tag{5}$$

Here, D is short for the image evidence. The likelihood $P(D|\Theta)$ is represented by the ConvNet responses in the image plane. The pose prior $P(\Theta)$ is modeled by building a Gaussian pose distribution with a mean centred around the pose Θ_0^t predicted from the previous time steps as follows:

$$\Theta_0^t = \Theta^{t-1} + \alpha(\Theta^{t-1} - \Theta^{t-2})\ . \tag{6}$$

where $\alpha = 0.5$ for all sequences. In practice, we compute for each joint a most likely location $d_{j,c}$ by weighted sampling from the posterior. Instead of working on all joints and images simultaneously, we simplify the process, assuming statistical independence, and thus reduce the number of samples needed by performing the computation for each image and joint separately. First, an extrapolated mean 2D location of j in c is computed $p_{j,c}^{init}$ by projecting to joint location in pose Θ_0^t into the image. Then we sample $N = 250$ 2D pixel locations p from a 2D-Gaussian distribution with mean $\mu = p_{j,c}^{init}$ and $\sigma = 20$ pixel. This can be considered a per-joint approximation of the posterior $P(\Theta)$ from Eq (5), projected in the image. Fig. 2 illustrates this process.

For each sample p we compute a weight $w(p)$

$$w(p) = \begin{cases} H_{j,c}^{q}(p) & H_{j,c}^{q}(p) > H_{th} \\ 0 & H_{j,c}^{q}(p) \leq H_{th} \end{cases} \tag{7}$$

where we set H_{th} = 0.25. The final assumed position of the joint $d_{j,c}$ is calculated as the average location of the

Figure 3. **Left:** Joint detection likelihoods for the right ankle in the heat-maps $H_{j,c}$ exhibit notable positional uncertainty, and there are many false positives and close-by multiple detections. **Right:** Even though two body parts for the same class (i.e. lower wrist) are close to each other in the images, our approach is able to correctly estimate their individual locations.

weighted pose posterior samples

$$d_{j,c} = \sum_{i=1}^{N} p_i * w(p_i). \tag{8}$$

The latter step can be considered as finding the mode of the weighted samples drawn from the posterior $P(\Theta|D)$ using the ConvNet responses as likelihood. As a result, $d_{j,c}$ is an accurate estimate of the actual 2D position of the body part. Note that the size of the person in the image may vary significantly over time and across camera views. To cope with this, the scale q of the heat-map at which detections are computed best is automatically selected for each camera, joint, and time step as part of the computation of $d_{j,c}$. Specifically, q is the scale s at which in a 50×50 pixel neighborhood around $p_{j,c}^{init}$ the highest detection likelihood was found.

In case more than one body part of the same class (e.g. left wrist) are close to each other in one of the views, for instance if there are multiple actors in the scene (see Fig 3(Right)), the value $d_{j,c}$ can be wrongly found as the middle between the two detections. Since the heat-map value at $d_{j,c}$ is comparably low in the middle between two parts, such erroneous detections (e.g. with two nearby people in one view) can also be filtered out by the above weighting with a minimum threshold.

Our ConvNet joint detection term measures the similarity between a give pose Θ of our body model and the refined 2D body part locations. To this end, we first need to project the 3D joint positions defined by Θ into the respective camera image plane using the projection operator Ψ_c of camera c. We incorporate the detected joint locations $d_{j,c}$ into the SoG model-based pose optimization framework by adding

the following term to Eq. 1:

$$E_{BP}(\Theta) = \sum_{c \in C} \sum_{j=1}^{n_{prt}} w(d_{j,c}) \exp\left(- \frac{\|\mathbf{\Psi_c}(\mathbf{l_j}(\mathbf{\Theta})) - \mathbf{d_{j,c}}\|^2}{\sigma^2} \right). \quad (9)$$

Here, $w(d_{j,c})$ is a weight for a constraint computed as the detection likelihood of the most likely image location $d_{j,c}$; i.e. $w(d_{j,c})$ is the heat-map value at $d_{j,c}$. $\mathbf{l}_j(\Theta)$ is the 3D joint position of j if the model strikes pose Θ.

6. Experiments and Results

We evaluated our algorithm on six real world sequences, which we recorded in an uncontrolled outdoor scenario with varying complexity. The sequences vary in the numbers and identities of actors to track, the existence and number of moving objects in the background, and the lighting conditions (i.e. some body parts lit and some in shadow). Cameras differ in the types (from cell phones to vision cameras), the frame resolutions, and the frame rates. By quad-tree decomposition, all images are effectively downsampled to a small resolution used in the generative energy (i.e. blob frame resolution). For the joint detection computation, the full resolution images are used and four heatmaps, with different scales for the subject, are generated. Please note that all cameras are frame synchronized. In particular, the cell phone cameras and the GoPro cameras are synchronized using the recorded audio up to one frame's accuracy. Moreover, we recorded additional sequences in a studio for marker-based or marker-less quantitative evaluation of skeletal motion tracking. The ground truth of the $Soccer$ sequence was computed based on manual annotation of the 2D body part locations in each view. The ground truth of the $Marker$ sequence was acquired with a marker-based motion capture system and the ground truth of $Run1$ was estimated based on marker-less tracking with a dense setup (i.e. 11 cameras) using a variant of [38]. Table 1 summarizes the specifications of each sequence. Apart from body model initialization, which requires the user to apply a few strokes to background segment the images of four actor poses (see [38]), tracking is fully-automatic. Further, the run-time of our algorithm depends on the number of cameras and actors, and the complexity of the scene, e.g. the number of Guassians needed in 2D. For a single actor and three cameras (e.g. the Walk sequence from the HumanEva dataset [35]), our algorithm takes around $1.186s$ for processing a single frame.

Qualitative Results Figures 1 and 4 show example poses tracked from outdoor sequences with our approach. Please see also the accompanying video for additional results. Our algorithm successfully estimated the pose parameters of the actors in challenging outdoor sequences with two or three cameras. In particular, our algorithm successfully tracked the two actors in $Soccer$ and $Juggling$, who often occlude each other, it tracked the actors in highly cluttered scenes ($Walk2$, $Run2$) - each of which contains many moving people in the background, and it performed well in a sequence with strong lighting variations ($Walk1$). All of these sequences were challenging to previous methods.

Quantitative Results We evaluated the importance of each term of our combined energy function and compared our method against state-of-the-art multi-view and 3D body part detection methods. We evaluated the results of three variations of our approach: **gen** neglecting the ConvNet detection term (i.e. $w_{BP} = 0$ in Eq. 1), **disc** neglecting the Appearance-based Similarity term (i.e. $w_{col} = 0$ in Eq. 1), and **gen+disc**, our full combined energy (i.e. $w_{BP} = 5$ and $w_{col} = 1$). Please note that **gen** is similar to applying the generative marker-less motion capture method proposed by Stoll et al. [38].

In Table 2, we calculated the average overlap of the 3D SoG models against one additional input view not used for tracking for each sequence. This value is calculated using the E_{col} (Eq. 4) considering only the additional camera view. A higher number indicates that the reconstructed pose (and model) matches better the input image. As can be seen in Fig. 5, even small improves in the overlap value translates to great improves in the tracking, e.g. hands and feet. The results in the table show that our combined method achieves higher accuracy than applying [38] or only applying the ConvNet detection term. Please note that Max. Overlap is the average overlap of the 3D SoG models, defined by the ground truth model parameters. The method proposed in [38] is used as ground truth for some sequences. However, it fails even with many cameras for outdoor sequences (marked with * in the table). Fig. 5 shows the visual improvements of our solution. As shown in the images, by combining both energy terms, we are able to better reconstruct the positions of the hands and feet.

We also compared the individual components of our approach in terms of the average 3D joint position reconstruction error over time. Table 3 summarizes the comparison for the sequences that we have ground truth 3D joint positions (obtained with different methods depending on the sequence). Fig. 6(top) shows the plot of the 3D joint position reconstruction error over time for sequence $Marker$ for all three variants. Fig. 6(bottom) shows visual results for each variant. As seen in the images, our combined approach (**gen+disc**) is able to reconstruct the pose of the subject more accurately. Note that with a small camera setup (only 2-3 cameras), our approach is able to reach a similar level of accuracy achieved by a dense multi-view approach in controllable indoor scenes.

Table 1. Specification for each sequence evaluated by our approach.

Sequence	*Soccer*	*Kickbox*	*Marker*	*Run*1	*Run*2	*Walk*1	*Walk*2	*Juggling*
Num. of Cams.	3	3	2	3	2	3	3	4
Num. of Frames.	300	300	500	1000	600	600	210	300
Frame Rates	23.8	23.8	25	35	30	60	30	30
Camera Types	cell-phone (*HTC One X*)		PhaseSpace Vision Camera		GoPro			
Input Frame Resol.	1280x720		256x256	1296x972	1280x720			
Blob Frame Resol.	160x90		256x256	160x90	240x135			
Tracked Subjects	2	1	1	1	1	1	1	2
Moving backgroud	No	No	No	No	Yes	Yes	Yes	Yes

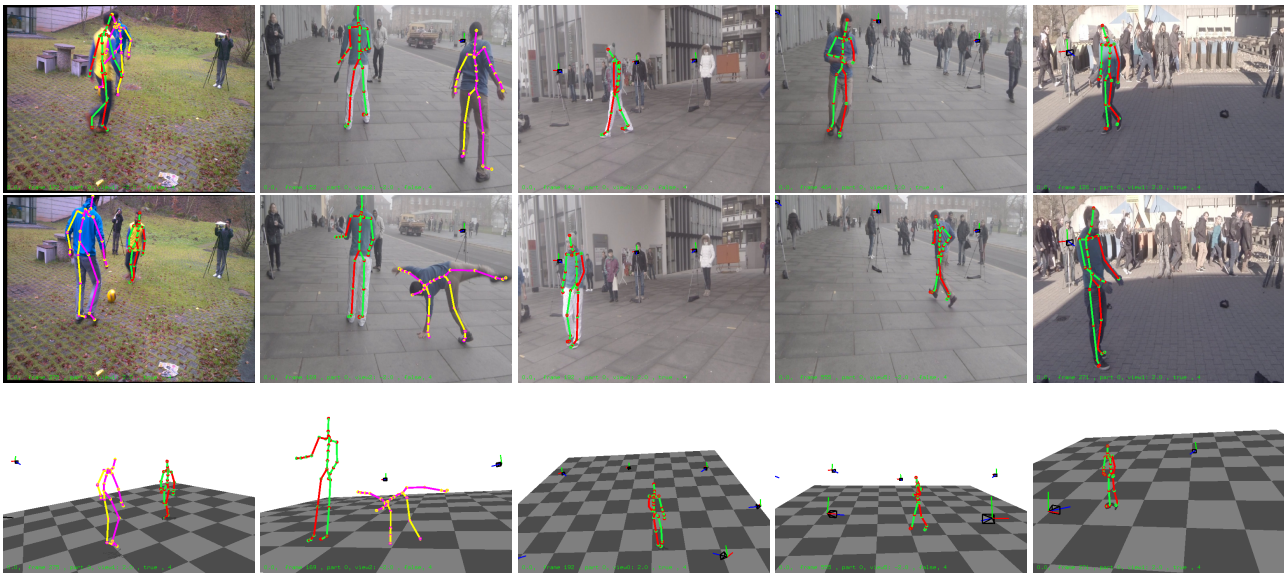

Figure 4. Qualitative results: From left to right particular frames for the *Soccer*, *Juggling*, *Walk*2, *Run*2 and *Walk*1 sequences recorded with only 2-3 cameras. For each sequence, from top to bottom, 3D pose overlaid with the input camera views for two frames and 3D visualizations of the captured skeletons.

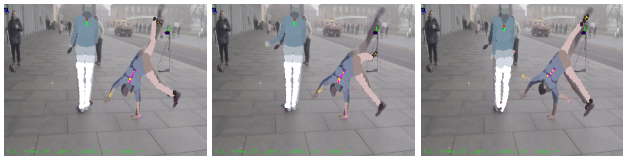

Figure 5. Particular frame for the *Juggling* sequence. From left to right, comparison between **gen+disc**, **disc** and **gen**, respectively. The individual strengths of both strategies are fruitfully enforced in our combined energy, which allows more accurate estimation of the positions of hands and feet.

Comparisons We evaluated our approach using the Boxing and Walking sequence from the HumanEva benchmark [35] and compared the results against Sigal et al.[36], Amin et al. [1] and Belagiannis et al.[10]. Table 4 summarizes the comparison results. As seen in the table, Amin et al. [1] shows very low average error but we also achieve similar results using our hybrid approach, outperforming the other methods; see supplementary document and video.

Table 3. Average 3D joint position error [cm].

Sequence	*Soccer*	*Marker*	*Run*1
gen (Gen. term only)	13.93	6.39	13.50
disc (Discr. term only)	**3.79**	5.69	6.11
gen+disc (Comb. energy)	3.95	**3.92**	**5.84**

Table 4. Average 3D joint position error for the HumanEva Walk and Box sequences.

Sequence	Walk [cm]	Box [cm]
Amin et al. [1]	5.45	4.77
Sigal et al. [36]	8.97	-
Belagiannis et al. [10]	6.83	6.27
Our approach	6.65	6.00

Discussion Our approach is subject to a few limitations. Currently, we can not track with moving cameras. With the current method motion tracking with a single camera view

Table 2. Average overlap of the 3D SoG models against an input view not used for tracking.

Sequence	*Soccer*	*Juggling*	*Marker*	*Run1*	*Run2*	*Walk1*	*Walk2*	*Kickbox*
gen (Gen. term only) [38]	43.58	58.48	49.33	47.04	17.93	31.78	34.12	57.07
disc (Discr. term only)	46.83	60.72	46.99	52.86	55.96	54.16	34.96	58.01
gen+disc (Combined energy)	**46.84**	**62.87**	**54.17**	**53.23**	**55.98**	**54.77**	**35.52**	**59.32**
Max. overlap	47.62	*	60.58	53.58	*	*	*	*

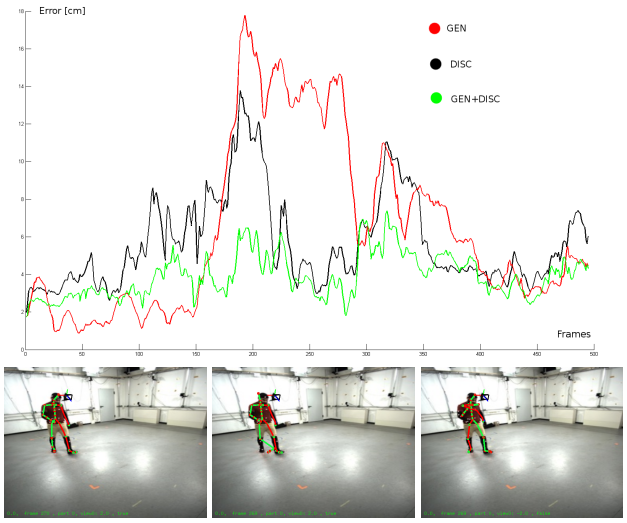

Figure 6. (top) Plot showing the average 3D joint position reconstruction error for sequence *Marker* using 2 input cameras only. (bottom) Visual results for variants **gen+disc**, **disc** and **gen**, respectively. Note that the correct reconstruction of the pose (e.g. hands and feet) is only possible with the combined terms in the energy function (**gen+disc**)

is not feasible. Also, the frame-rate of the camera needs to be adequate to handle the speed of the recorded motion. For example, if fast motions are captured with a lower frame-rate, we might not be able to track the sequence accurately, as shown in Fig. 7 for the Kickbox sequence, recorded at 23.8fps. However, this is also a common problem with approaches relying on a dense camera setup. Unlike purely generative methods, our approach is still able to recover from the tracking errors, even with such fast motion, and it can work correctly if a higher frame rate is used. Our approach works well even for challenging sequences like the juggling, which contains a cartwheel motion. However, for more complex motions, it might be necessary to re-train the ConvNet-based method for improved results.

7. Conclusion

We presented a novel and robust marker-less human motion capture algorithm that tracks articulated joint motion with only 2-3 cameras. By fusing the 2D body part detections, estimated from a ConvNet-based joint detection algorithm, into a generative model-based tracking algorithm,

Figure 7. Fast motions recorded with a lower frame-rate (23.8fps) generate blurred images, which makes it hard for our method to correctly track the foot with only 3 cameras.

based on the Sums of Gaussians framework, our system is able to deliver high tracking accuracy in challenging outdoor environments with only 2-3 cameras. Our method also works successfully when there is strong background motion (many people moving in the background), when very strong illumination changes are happening or when the human subject performs complex motions. By comparing against sequences recorded in controlled environments or recorded with many cameras, we also demostrated that our system is able to achieve state-of-the-art accuracy despite a reduced number of cameras. As future work, we would like to investigate the use of unsynchronized or moving cameras in our framework.

Acknowledgments: This research was funded by the ERC Starting Grant project CapReal (335545) and the Max Planck Center for Visual Computing and Communication.

References

[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *BMVC*, 2013. 7

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE CVPR*, June 2014. 4

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 1, 2

[4] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Proc. ICCV*, pages 1092–1099, 2011. 2, 3

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. *CVPR, IEEE*, 2014. 1

[6] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 87:28–52, 2010. 2

[7] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2

[8] M. Bray, E. Koller-Meier, and L. V. Gool. Smart particle filtering for high-dimensional tracking. *CVIU*, 106(1):116–129, 2007. 2

[9] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region and motion-based 3d tracking of rigid and articulated objects. *TPAMI*, 32:402–415, 2010. 2

[10] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 7

[11] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. *NIPS*, 2014. 2

[12] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013. 2

[13] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 2

[14] A. Elhayek, C. Stoll, N. Hasler, K. I. Kim, H.-P. Seidel, and C. Theobaltl. Spatio-temporal motion tracking with unsynchronized cameras. In *Proc. CVPR*, 2012. 2, 3

[15] A. Elhayek, C. Stoll, N. Hasler, K. I. Kim, and C. Theobaltl. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *Proc. CGF*, 2014. 1, 2, 3

[16] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 2005. 2

[17] M. A. Fischler and R. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, C-22(1), Jan 1973. 2

[18] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture – a multi-layer framework. *IJCV*, 87:75–92, 2010. 2

[19] J. Gall, B. Rosenhahn, and H.-P. Seidel. Drift-free tracking of rigid and articulated objects. In *CVPR*, 2008. 2

[20] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR*, 2013. 2

[21] N. Hasler, B. Rosenhahn, T. Thormählen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009. 1, 2

[22] A. Jain, J. Tompson, M. Andriluka, G. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. In *ICLR*, 2014. 2

[23] A. Jain, J. Tompson, Y. LeCun, and C. Bregler. Modeep: A deep learning framework using motion features for human pose estimation. *ACCV*, 2014. 2, 4

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[25] C.-S. Lee and A. Elgammal. Coupled visual and kinematic manifold models for tracking. *IJCV*, 87:118–139, 2010. 2

[26] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang. 3d human motion tracking with a coordinated mixture of factor analyzers. *IJCV*, 87:170–190, 2010. 2

[27] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2):90–126, 2006. 1, 2

[28] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814. Omnipress, 2010. 4

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2

[30] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 2

[31] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling. In *Proc. ICCV*, 2011. 2

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

[33] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013. 2

[34] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. Motion capture from body-mounted cameras. *ACM Trans. Graph.*, 30(4):31:1–31:10, July 2011. 2

[35] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2010. 1, 2, 6, 7

[36] L. Sigal, M. Isard, H. Haussecker, and M. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, 2012. 7

[37] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *IEEE ICCV*, 2013. 3

[38] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, 2011. 2, 3, 4, 6, 8

[39] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Join training of a convolutional network and a graphical model for human pose estimation. *NIPS*, 2014. 2, 3, 4

[40] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[41] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. In *Adv. in Comput. Math.*, 1995. 4

[42] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2

[43] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Performance capture of interacting characters with handheld kinects. In *ECCV*, 2012. 2