

VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera

DUSHYANT MEHTA^{1,2}, SRINATH SRIDHAR¹, OLEKSANDR SOTNYCHENKO¹, HELGE RHODIN¹, MOHAMMAD SHAFIEI^{1,2}, HANS-PETER SEIDEL¹, WEIPENG XU¹, DAN CASAS³, CHRISTIAN THEOBALT¹

¹Max Planck Institute for Informatics, ²Saarland University, ³Universidad Rey Juan Carlos

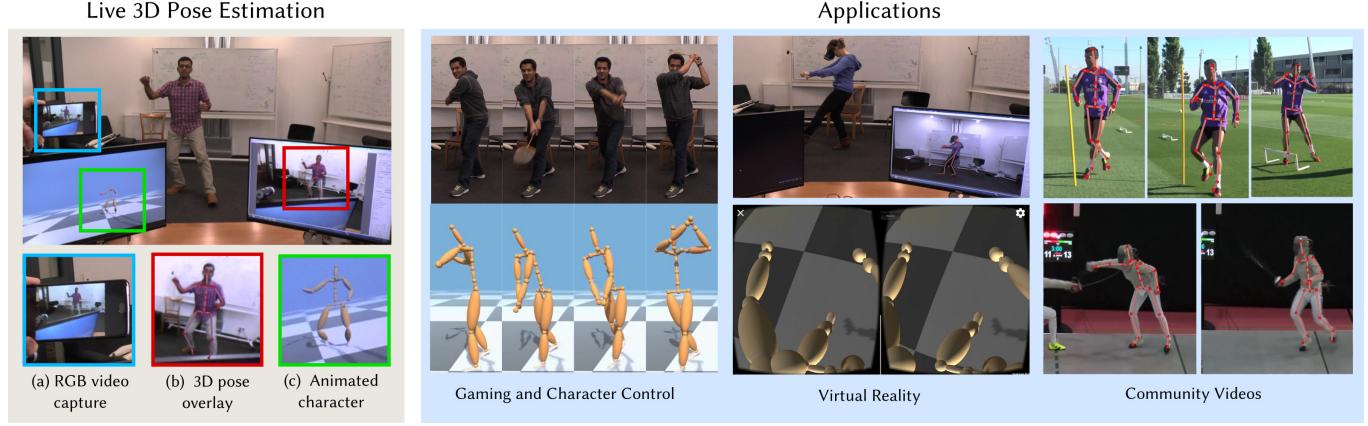


Fig. 1. We recover the full global 3D skeleton pose in real-time from a single RGB camera, even wireless capture is possible by streaming from a smartphone (left). It enables applications such as controlling a game character, embodied VR, sport motion analysis and reconstruction of community video (right). Community videos (CC BY) courtesy of Real Madrid C.F. [2016] and RUSFENCING-TV [2017].

We present the first real-time method to capture the full global 3D skeletal pose of a human in a stable, temporally consistent manner using a single RGB camera. Our method combines a new convolutional neural network (CNN) based pose regressor with kinematic skeleton fitting. Our novel fully-convolutional pose formulation regresses 2D and 3D joint positions jointly in real time and does not require tightly cropped input frames. A real-time kinematic skeleton fitting method uses the CNN output to yield temporally stable 3D global pose reconstructions on the basis of a coherent kinematic skeleton. This makes our approach the first monocular RGB method usable in real-time applications such as 3D character control—thus far, the only monocular methods for such applications employed specialized RGB-D cameras. Our method’s accuracy is quantitatively on par with the best offline 3D monocular RGB pose estimation methods. Our results are qualitatively comparable to, and sometimes better than, results from monocular RGB-D approaches, such as the Kinect. However, we show that our approach is more broadly applicable than RGB-D solutions, i.e., it works for outdoor scenes, community videos, and low quality commodity RGB cameras.

1 INTRODUCTION

Optical skeletal motion capture of humans is widely used in applications such as character animation for movies and games, sports and biomechanics, and medicine. To overcome the usability constraints imposed by commercial systems requiring marker suits [Menache 2000], researchers developed marker-less motion capture methods that estimate motion in more general scenes using multi-view

This work was funded by the ERC Starting Grant project CapReal (335545). Dan Casas was supported by a Marie Curie Individual Fellow, grant agreement 707326.

video [Moeslund et al. 2006], with recent solutions being real-time [Stoll et al. 2011]. The swell in popularity of applications such as real-time motion-driven 3D game character control, self-immersion in 3D virtual and augmented reality, and human–computer interaction, has led to new real-time full-body motion estimation techniques using only a single, easy to install, depth camera, such as the Microsoft Kinect [Microsoft Corporation 2010, 2013, 2015]. RGB-D cameras provide valuable depth data which greatly simplifies monocular pose reconstruction. However, RGB-D cameras often fail in general outdoor scenes (due to sunlight interference), are bulkier, have higher power consumption, have lower resolution and limited range, and are not as widely and cheaply available as color cameras.

Skeletal pose estimation from a single color camera is a much more challenging and severely underconstrained problem. Monocular RGB body pose estimation in 2D has been widely researched, but estimates only the 2D skeletal pose [Bourdev and Malik 2009; Felzenszwalb et al. 2010; Felzenszwalb and Huttenlocher 2005; Ferrari et al. 2009; Pishchulin et al. 2013; Wei et al. 2016]. Learning-based discriminative methods, in particular deep learning methods [Insafutdinov et al. 2016; Lifshitz et al. 2016; Newell et al. 2016; Tompson et al. 2014], represent the current state of the art in 2D pose estimation, with some of these methods demonstrating real-time performance [Cao et al. 2016; Wei et al. 2016]. Monocular RGB estimation of the 3D skeletal pose is a much harder challenge tackled by relatively fewer methods [Bogo et al. 2016; Tekin et al. 2016b,c; Zhou et al. 2015, 2016, 2015b]. Unfortunately, these methods are typically offline, and they often reconstruct 3D joint positions individually per image, which are temporally unstable, and do not enforce constant bone

lengths. Most approaches also capture local 3D pose relative to a bounding box, and not the full global 3D pose. This makes them unsuitable for applications such as real-time 3D character control.

In this paper, we present the first method that captures temporally consistent global 3D human pose—in terms of joint angles of a single, stable kinematic skeleton—in real-time (30 Hz) from a single RGB video in a general environment. Our approach builds upon the top performing single RGB 3D pose estimation methods using convolutional neural networks (CNNs) [Mehta et al. 2016; Pavlakos et al. 2016]. High accuracy requires training comparably deep networks which are harder to run in real-time, partly due to additional preprocessing steps such as bounding box extraction. Mehta et al. [2016] use a 100-layer architecture to predict 2D and 3D joint positions simultaneously, but is unsuitable for real-time execution. To improve runtime, we use a shallower 50-layer network. However, for best quality at real-time frame rates, we do not merely use a shallower variant, but extend it to a novel fully-convolutional formulation. This enables higher accuracy 2D and 3D pose regression, in particular of end effectors (hands, feet), in real-time. In contrast to existing solutions our approach allows operation on non-cropped images, and where run-time is a concern, it can be used to bootstrap a simple bounding box tracker. We also combine the CNN-based joint position regression with an efficient optimization step to fit a 3D skeleton to these reconstructions in a temporally stable way, yielding the global pose and joint angles of the skeleton. In summary, we contribute by proposing the first real-time method to capture global 3D kinematic skeleton pose from single RGB video. To strike a good compromise between computational complexity and accuracy, our method combines:

- A new real-time, fully-convolutional 3D body pose formulation using CNNs that yields 2D and 3D joint positions simultaneously and forgoes the need to perform expensive bounding box computations.
- Model-based kinematic skeleton fitting against the 2D/3D pose predictions to produce temporally stable joint angles of a metric global 3D skeleton, in real time.

Our real-time method achieves state-of-the-art accuracy comparable to the best offline RGB pose estimation methods on standard 3D human body pose benchmarks, particularly for end effector positions (Section 5.2). Our results are qualitatively comparable to, and sometimes better than, state-of-the-art single RGB-D methods [Girshick et al. 2011], even commercial ones [Microsoft Corporation 2015]. We experimentally show that this makes ours the first single-RGB method usable for similar real-time 3D applications—so far only feasible with RGB-D input—such as game character control or immersive first person virtual reality (VR). We further show that our method succeeds in settings where existing RGB-D methods would not, such as outdoor scenes, community videos, and even with low quality video streams from ubiquitous mobile phone cameras.

2 RELATED WORK

Our goal is stable 3D skeletal motion capture from (1) a single camera (2) in real-time. We focus the discussion of related work on approaches from the large body of marker-less motion capture research that contributed to attaining either of these properties.

Multi-view: With multi-view setups markerless motion-capture solutions attain high accuracy. Tracking of a manually initialized actor model from frame to frame with a generative image formation model is common. See [Moeslund et al. 2006] for a complete overview. Most methods target high quality with offline computation [Bregler and Malik 1998; Howe et al. 1999; Loper and Black 2014; Sidenbladh et al. 2000; Starck and Hilton 2003]. Real-time performance can be attained by representing the actor with Gaussians [Rhodin et al. 2015; Stoll et al. 2011; Wren et al. 1997] and other approximations [Ma and Wu 2014], in addition to formulations allowing model-to-image fitting. However, these tracking-based approaches often lose track in local minima of the non-convex fitting functions they optimize and require separate initialization, e.g. using [Bogo et al. 2016; Rhodin et al. 2016b; Sminchisescu and Triggs 2001]. Robustness could be increased with a combination of generative and discriminative estimation [Elhayek et al. 2016], even from a single input view [Rosales and Sclaroff 2006; Sminchisescu et al. 2006], and egocentric perspective [Rhodin et al. 2016a]. We utilize generative tracking components to ensure temporal stability, but avoid model projection through a full image formation model to speed up estimation. Instead, we combine discriminative pose estimation with kinematic fitting to succeed in our underconstrained setting.

Monocular Depth-based: The additional depth channel provided by RGB-D sensors has led to robust real-time pose estimation solutions [Baak et al. 2011; Ganapathi et al. 2012; Ma and Wu 2014; Shotton et al. 2013; Wei et al. 2012; Ye and Yang 2014] and the availability of low-cost devices has enabled a range of new applications. Even real-time tracking of general deforming objects [Zollhöfer et al. 2014] and template-free reconstruction [Dou et al. 2016; Innmann et al. 2016; Newcombe et al. 2015; Orts-Escalano et al. 2016] has been demonstrated. RGB-D information overcomes forward-backwards ambiguities in monocular pose estimation. Our goal is a video solution that overcomes depth ambiguities without relying on a specialized active sensor.

Monocular RGB: Monocular generative motion capture has only been shown for short clips and when paired with strong motion priors [Urtasun et al. 2006] or in combination with discriminative re-initialization [Rosales and Sclaroff 2006; Sminchisescu et al. 2006], since generative reconstruction is fundamentally underconstrained. Using photo-realistic template models for model fitting enables more robust monocular tracking of simple motions, but requires more expensive offline computation [de La Gorce et al. 2008]. Sampling-based methods avoid local minima [Balan et al. 2005; Bo and Sminchisescu 2010; Deutscher and Reid 2005; Gall et al. 2010]. However, real-time variants can not guarantee global convergence due to a limited number of samples, such as particle swarm optimization techniques [Oikonomidis et al. 2011]. Structure-from-motion techniques exploit motion cues in a batch of frames [Garg et al. 2013], and have also been applied to human motion estimation [Gotardo and Martinez 2011; Lee et al. 2013; Park and Sheikh 2011; Zhu et al. 2011]. However, batch optimization does not apply to our real-time setting, where frames are streamed sequentially. For some applications manual annotation and correction of frames is suitable, for instance to enable movie actor reshaping [Jain et al. 2010] and garment replacement in video [Rogge et al. 2014]. In combination

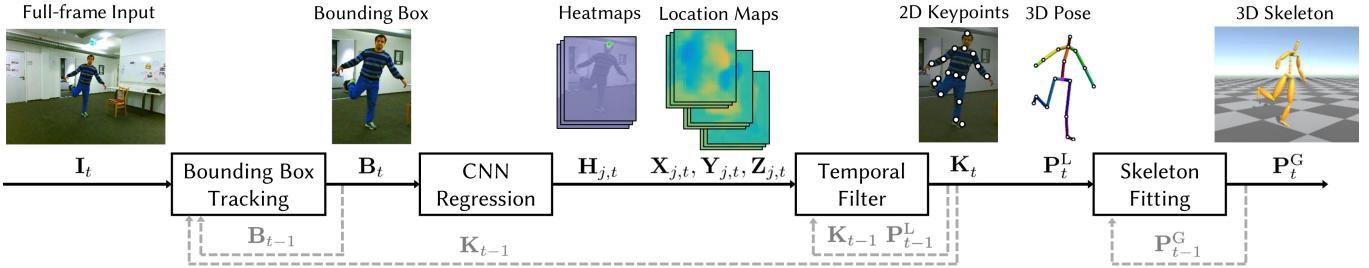


Fig. 2. Overview. Given a full-size image I_t at frame t , the person-centered crop B_t is efficiently extracted by bounding box tracking, using the previous frame's keypoints K_{t-1} . From the crop, the CNN jointly predicts 2D heatmaps $H_{j,t}$ and our novel 3D *location-maps* $X_{j,t}$, $Y_{j,t}$ and $Z_{j,t}$ for all joints j . The 2D keypoints K_t are retrieved from $H_{j,t}$ and, after filtering, are used to read off 3D pose P_t^L from $X_{j,t}$, $Y_{j,t}$ and $Z_{j,t}$. These per-frame estimates are combined to stable global pose P_t^G by skeleton fitting. Information from frame $t-1$ is marked in gray-dashed.

with physical constraints, highly accurate reconstructions are possible from monocular video [Wei and Chai 2010]. Vondrak et al. [2012] succeed without manual annotation by simulating biped-controllers, but require batch-optimization. While these methods can yield high-quality reconstructions, interaction and expensive optimization preclude live applications.

Discriminative 2D human pose estimation is often an intermediate step to monocular 3D pose estimation. Pictorial structure approaches infer body part locations by message passing over a huge set of pose-states [Agarwal and Triggs 2006; Andriluka et al. 2009; Bourdev and Malik 2009; Felzenszwalb and Huttenlocher 2005; Ferrari et al. 2009; Johnson and Everingham 2010] and have been extended to 3D pose estimation [Amin et al. 2013; Balan et al. 2007; Belagiannis et al. 2014; Sigal et al. 2012]. Recent approaches outperform these methods in computation time and accuracy by leveraging large image databases with 2D joint location annotation, which enables high accuracy prediction with deep CNNs [Belagiannis and Zisserman 2016; Hu et al. 2016; Insafutdinov et al. 2016; Pishchulin et al. 2016; Wei et al. 2016], on multiple GPUs, even at real-time rates [Cao et al. 2016]. Given 2D joint locations, lifting them to 3D pose is challenging. Existing approaches use bone length and depth ordering constraints [Mori and Malik 2006; Taylor 2000], sparsity assumptions [Wang et al. 2014; Zhou et al. 2015,a], joint limits [Akhter and Black 2015], inter-penetration constraints [Bogo et al. 2016], temporal dependencies [Rhodin et al. 2016b], and regression [Yasin et al. 2016]. Treating 3D pose as a hidden variable in 2D estimation is an alternative [Brau and Jiang 2016]. However, the sparse set of 2D locations loses image evidence, e.g. on forward-backwards orientation of limbs, which leads to erroneous estimates in ambiguous cases. To overcome these ambiguities, discriminative methods have been proposed that learn implicit depth features for 3D pose directly from more expressive image representations. Rosales and Sclaroff regress 3D pose from silhouette images with the *specialized mappings architecture* [2000], Agarwal and Triggs with linear regression [2006], and Elgammal and Lee through a joint embedding of images and 3D pose [2004]. Sminchisescu further utilized temporal consistency to propagate pose probabilities with a *Bayesian mixture of experts Markov model* [2007]. Relying on the recent advances in machine learning techniques and compute capabilities, approaches for direct 3D pose regression from the input image have

been proposed, using structured learning of latent pose [Li et al. 2015a; Tekin et al. 2016a], joint prediction of 2D and 3D pose [Li and Chan 2014; Tekin et al. 2016b; Yasin et al. 2016], transfer of features from 2D datasets [Mehta et al. 2016], novel pose space formulations [Pavlakos et al. 2016] and classification over example poses [Pons-Moll et al. 2014; Rogez and Schmid 2016]. Relative per-bone predictions [Li and Chan 2014], kinematic skeleton models [Zhou et al. 2016], or root centered joint positions [Ionescu et al. 2014a] are used as the eventual output space. Such direct 3D pose regression methods capture depth relations well, but 3D estimates usually do not accurately match the true 2D location when re-projected to the image, because estimations are done in cropped images that lose camera perspective effects, using a canonical height, and minimize 3D loss instead of projection to 2D. Furthermore, they only deliver joint positions, are temporally unstable, and none has shown real-time performance. We propose a method to combine 2D and 3D estimates in real-time along with temporal tracking. It is inspired by the method of Tekin et al. [2016c], where batches of frames are processed offline after motion compensation, and is related to the recently proposed per-frame combination of 2D and 3D pose [Tekin et al. 2016b].

Notably, only few methods target real-time monocular reconstruction. Exceptions are the regression of 3D pose from Haar features by Bissacco et al. [2007] and detection of a set of discrete poses from edge direction histograms in the vicinity of the previous frame pose [Taycher et al. 2006]. Both only obtain temporally unstable, coarse pose, not directly usable in our applications. Chai and Hodges obtain sufficient quality to drive virtual avatars in real-time, but require visual markers [Chai and Hodges 2005]. The use of CNNs in real time has been explored for variants of the object detection problem, for instance bounding box detection and pedestrian detection methods have leveraged application specific architectures [Angelova et al. 2015; Liu et al. 2016; Redmon et al. 2015] and preprocessing steps [Ren et al. 2015].

In a similar vein, we propose a 3D pose estimation approach that leverages a novel fully-convolutional CNN formulation to predict 2D and 3D pose jointly. In combination with inexpensive preprocessing and an optimization based skeletal fitting method, it enables high accuracy pose estimation, while running at more than 30 Hz.

3 OVERVIEW

Our system is capable of obtaining a temporally consistent, full 3D skeletal pose of a human from a monocular RGB camera. Estimating 3D pose from a single RGB camera is a challenging, under-constrained problem with inherent ambiguities. Figure 2 provides an overview of our method to tackle this challenging problem. It consists of two primary components. The first is a convolutional neural network (CNN) to regress 2D and 3D joint positions under the ill-posed monocular capture conditions. It is trained on annotated 3D human pose datasets [Ionescu et al. 2014b; Mehta et al. 2016], additionally leveraging annotated 2D human pose datasets [Andriluka et al. 2014; Johnson and Everingham 2010] for improved in-the-wild performance. The second component combines the regressed joint positions with a kinematic skeleton fitting method to produce a temporally stable, camera-relative, full 3D skeletal pose.

CNN Pose Regression: The core of our method is a CNN that predicts both 2D, and root (pelvis) relative 3D joint positions in real-time. The new proposed fully-convolutional pose formulation leads to results on par with the state-of-the-art offline methods in 3D joint position accuracy (see Section 5.2 for details). Being fully-convolutional, it can operate in the absence of tight crops around the subject. The CNN is capable of predicting joint positions for a diverse class of activities regardless of the scene settings, providing a strong basis for further pose refinement to produce temporally consistent full-3D pose parameters

Kinematic Skeleton Fitting: The 2D and the 3D predictions from the CNN, together with the temporal history of the sequence, can be leveraged to obtain temporally consistent full-3D skeletal pose, with the skeletal root (pelvis) localized in camera space. Our approach uses an optimization function that: (1) combines the predicted 2D and 3D joint positions to fit a kinematic skeleton in a least squares sense, (2) ensures temporally smooth tracking over time. We further improve the stability of the tracked pose by applying filtering steps at different stages.

Skeleton Initialization (Optional): The system is set up with a default skeleton which works well out of the box for most humans. For more accurate estimates, the relative body proportions of the underlying skeleton can be adapted to that of the subject, by averaging CNN predictions for a few frames at the beginning. Since monocular reconstruction is ambiguous without a scale reference, the CNN predicts height normalized 3D joint positions. Users only need to provide their height (distance from head to toe) once, so that we can track the 3D pose in true metric space.

4 REAL-TIME MONOCULAR 3D POSE ESTIMATION

In this section, we describe in detail the different components of our method to estimate a temporally consistent 3D skeletal motion from monocular RGB input sequences. As input, we assume a continuous stream of monocular RGB images $\{\dots, I_{t-1}, I_t\}$. For frame t in the input stream, the final output of our approach is P_t^G which is the full global 3D skeletal pose of the person being tracked. Because this output is already temporally consistent and in global 3D space, it can be readily used in applications such as character control.

We use the following notation for the output in the intermediate components of our method. The CNN pose regressor jointly estimates the 2D joint positions K_t and root-relative 3D joint positions P_t^L (Section 4.1). The 3D skeleton fitting component combines the 2D and 3D joint position predictions to estimate a smooth, temporally consistent pose $P_t^G(\theta, d)$, which is parameterized by the global position d in camera space, and joint angles θ of the kinematic skeleton S . J indicates the number of joints. We drop the frame-number subscript t in certain sections to aid readability.

4.1 CNN Pose Regression

The goal of CNN pose regression is to obtain joint positions, both, in 2D image space and 3D. For 2D pose estimation with neural nets, the change of formulation from direct regression of x, y body-joint coordinates [Toshev and Szegedy 2014] to a heatmap based body-joint detection formulation [Tompson et al. 2014] has been the key driver behind the recent developments in 2D pose estimation. The heatmap based formulation naturally ties image evidence to pose estimation by predicting a confidence heatmap $H_{j,t}$ over the image plane for each joint $j \in \{1..J\}$.

Existing approaches to 3D pose estimation lack such an image-to-prediction association, often directly regressing the root-relative joint locations [Ionescu et al. 2014a], leading to predicted poses whose extent of articulation doesn't reflect that of the person in the image. See Figure 9. Treating pose as a vector of joint locations also causes a natural gravitation towards networks with fully-connected formulations [Mehta et al. 2016; Rogez and Schmid 2016; Tekin et al. 2016a; Yu et al. 2016], restricting the inputs to tight crops at a fixed resolution, a limitation that needs to be overcome. These methods assume the availability of tight bounding boxes, which necessitates supplementation with separate bounding box estimators for actual usage, which further adds to the run-time of these methods. The fully-convolutional formulation of Pavlakos et al. [Pavlakos et al. 2016] seeks to alleviate some of these issues, but is limited by the expensive per-joint volumetric formulation, which still relies on cropped input and does not scale well to larger image sizes.

We overcome these limitations through our new formulation, by extending the 2D heatmap formulation to 3D using three additional *location-maps* X_j, Y_j, Z_j per joint j , capturing the root-relative locations x_j, y_j and z_j respectively. To have the 3D pose prediction linked more strongly to the 2D appearance in the image, the x_j, y_j and z_j values are read off from their respective location-maps at the position of the maximum of the corresponding joint's 2D heatmap H_j , and stored in $P^L = \{x, y, z\}$, where $x \in \mathbb{R}^{1 \times J}$ is a vector that stores the coordinate x location of each joint maximum. The pose formulation is visualized in Figure 3. Networks using this fully-convolutional formulation are not constrained in input image size, and can work without tight crops. Additionally, the network provides 2D and 3D joint location estimates without additional overhead, which we exploit in subsequent steps for real-time estimation. Section 5.2 shows the improvements afforded by this formulation.

Loss Term: To enforce the fact that we are only interested in x_j, y_j and z_j from their respective maps at joint j 's 2D location, the joint location-map loss is weighted stronger around the joint's 2D

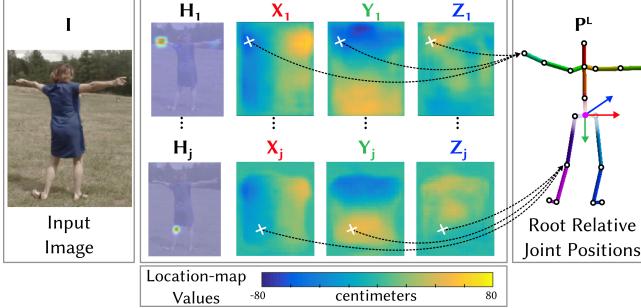


Fig. 3. Schema of the fully-convolutional formulation for predicting root relative joint locations. For each joint j , the 3D coordinates are predicted from their respective *location-maps* X_j , Y_j , Z_j at the position of the maximum in the corresponding 2D heatmap H_j . The structure observed here in the location-maps emerges due to the spatial loss formulation. See Section 4.1.

location. We use the L2 loss. For x_j is the loss formulation is

$$\text{Loss}(x_j) = \|H_j^{\text{GT}} \odot (X_j - X_j^{\text{GT}})\|_2, \quad (1)$$

where GT indicates ground truth and \odot is the Hadamard product. The location maps are weighted with the respective ground truth 2D heatmap H_j^{GT} , which in turn have confidence equal to a Gaussian with a small support localized at joint j 's 2D location. Note that no structure is imposed on the location-maps. The structure that emerges in the predicted location-maps is indicative of the correlation of x_j and y_j with root relative location of joint j in the image plane. See Figure 3.

Network Details: We use the proposed formulation to adapt the ResNet50 network architecture of He et al. [2016]. We replace the layers of ResNet50 from *res5a* onwards with the architecture depicted in Figure 5, producing the heatmaps and location-maps for all joints $j \in \{1..J\}$. After training, the Batch Normalization [Ioffe and Szegedy 2015] layers are merged with the weights of their preceding convolution layers to improve the speed of the forward pass.

Intermediate Supervision: We predict the 2D heatmaps and 3D location-maps from the features at *res4d* and *res5a*, tapering down the weights of intermediate losses with increasing iteration count. Additionally, similar to the root-relative location-maps X_j , Y_j and Z_j , we predict kinematic parent-relative location-maps ΔX_j , ΔY_j and ΔZ_j from the features at *res5b* and compute bone length-maps as:

$$BL_j = \sqrt{\Delta X_j \odot \Delta X_j + \Delta Y_j \odot \Delta Y_j + \Delta Z_j \odot \Delta Z_j}. \quad (2)$$

These intermediate predictions are subsequently concatenated with the intermediate features, to give the network an explicit notion of bone lengths to guide the predictions. See Figure 5.

Experiments showed that the deeper variants of ResNet offer only small gains for a substantial increase ($1.5\times$) in computation time, prompting us to choose ResNet50 to enable real-time, yet highly accurate joint location estimation with the proposed formulation.

Training: The network is pretrained for 2D pose estimation on MPII [Andriluka et al. 2014] and LSP [Johnson and Everingham 2010, 2011] to allow superior in-the-wild performance, as proposed



Fig. 4. Representative training frames from Human3.6m and MPI-INF-3DHP 3D pose datasets. Also shown are the background, clothing and occluder augmentations done on MPI-INF-3DHP training data.

by Mehta et al. [Mehta et al. 2016]. For 3D pose, we use MPI-INF-3DHP [Mehta et al. 2016] and Human3.6m [Ionescu et al. 2014b]. We take training sequences for all subjects except S9 and S11 from Human3.6m. We sample frames as per [Ionescu et al. 2014a]. For MPI-INF-3DHP, we consider all 8 training subjects. We choose sequences from all 5 chest-high cameras, 2 head-high cameras (angled down) and 1 knee-high camera (angled up) to learn some degree of invariance to the camera viewpoint. The sampled frames have at least one joint move by $> 200\text{mm}$ between them. We use various combinations of background, occluder (chair), upper-body clothing and lower-body clothing augmentation for 70% of the selected frames. We train with person centered crops, and use image scale augmentation at 2 scales ($0.7\times$, $1.0\times$), resulting in 75k training samples for Human3.6m and 100k training samples for MPI-INF-3DHP. Figure 4 shows a few representative frames of training data. In addition to the 17 joints typically considered, we use foot tip positions. The ground truth joint positions are with respect to a height normalized skeleton (knee-neck height 92 cm). We make use of the Caffe [2014] framework for training, and use the Adadelta [Zeiler 2012] solver with learning rate tapered down with increasing iterations.

Bounding Box Tracker: Existing offline solutions process each frame in a separate person-localization and bounding box (BB) cropping step [Mehta et al. 2016; Tekin et al. 2016c] or assume bounding boxes are available [Li and Chan 2014; Li et al. 2015b; Pavlakos et al. 2016; Tekin et al. 2016a; Zhou et al. 2016]. Although our fully-convolutional formulation allows the CNN to work without requiring cropping, the run-time of the CNN is highly dependent on the input image size. Additionally, the CNN is trained for subject sizes in the range of 250–340 px in the frame, requiring averaging of predictions at multiple image scales per frame (scale space search) if processing the full frame at each time step. Guaranteeing real-time rates necessitates restricting the size of the input to the network and tracking the scale of the person in the image to avoid searching the scale space in each frame. We do this in an integrated way. The 2D pose predictions from the CNN at each frame are used to determine the BB for the next frame through a slightly larger box around the predictions. The smallest rectangle containing the keypoints K is computed and augmented with a buffer area $0.2\times$ the height vertically and $0.4\times$ the width horizontally. To stabilize the estimates,

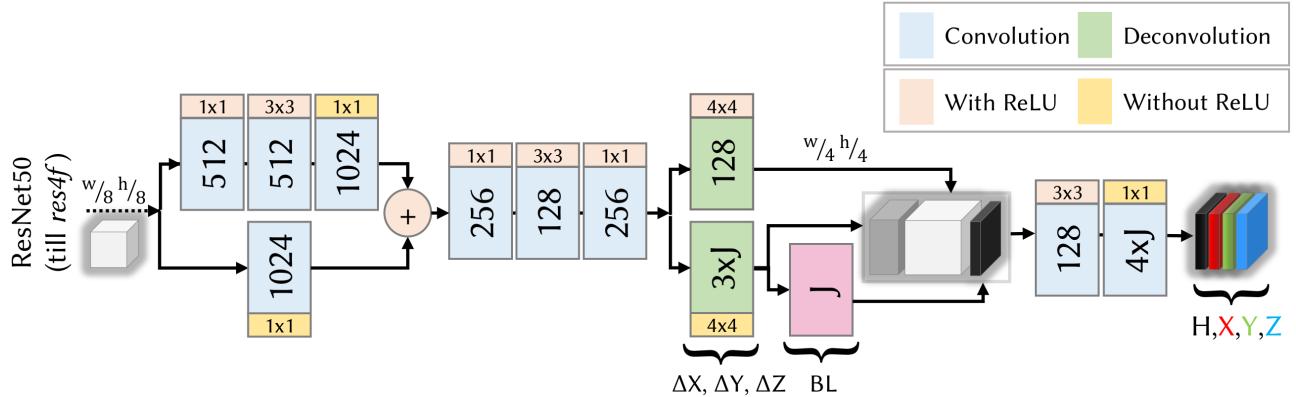


Fig. 5. Network Structure. The structure above is preceded by ResNet50/100 till level 4. We use kinematic parent relative 3D joint location predictions ΔX , ΔY , ΔZ as well as bone length maps BL constructed from these as auxiliary tasks. The network predicts 2D location heatmaps H and root relative 3D joint locations X, Y, Z. Refer to Section 4.1.

the BB is shifted horizontally to the centroid of the 2D predictions, and its corners are filtered with a weighted average with the previous frame’s BB using a momentum of 0.75. To normalize scale, the BB crop is resized to 368x368 px. The BB tracker starts with (slow) multi-scale predictions on the full image for the first few frames, and hones in on the person in the image making use of the BB-agnostic predictions from the fully convolutional network. The BB tracking is easy to implement and without runtime overhead, since the proposed fully-convolutional network outputs 2D and 3D pose jointly and operates on arbitrary input sizes.

4.2 Kinematic Skeleton Fitting

Applying per-frame pose estimation techniques on a video does not exploit and ensure temporal consistency of motion, and small pose inaccuracies lead to temporal jitter, an unacceptable artifact for most graphics applications. We combine the 2D and 3D joint positions in a joint optimization framework, along with temporal filtering and smoothing, to obtain an accurate, temporally stable and robust result. First, the 2D predictions K_t are temporally filtered [Casiez et al. 2012] and used to obtain the 3D coordinates of each joint from the location-map predictions, giving us P_t^L . To ensure skeletal stability, the bone lengths inherent to P_t^L are replaced by the bone lengths of the underlying skeleton in a simple retargeting step that preserves the bone directions of P_t^L . The resulting 2D and 3D predictions are combined by minimizing the objective energy

$$\begin{aligned} E_{\text{total}}(\theta, \mathbf{d}) = & E_{\text{IK}}(\theta, \mathbf{d}) + E_{\text{proj}}(\theta, \mathbf{d}) \\ & + E_{\text{smooth}}(\theta, \mathbf{d}) + E_{\text{depth}}(\theta, \mathbf{d}), \end{aligned} \quad (3)$$

for skeletal joint angles θ and the root joint’s location in camera space \mathbf{d} . The 3D inverse kinematics term E_{IK} determines the overall pose by similarity to the 3D CNN output P_t^L . The projection term E_{proj} determines global position \mathbf{d} and corrects the 3D pose by re-projection onto the detected 2D keypoints K_t . Both terms are implemented with the L2 loss,

$$E_{\text{proj}} = \|\Pi(P_t^G) - K_t\|_2 \text{ and } E_{\text{IK}} = \|(P_t^G - \mathbf{d}) - P_t^L\|_2, \quad (4)$$

where Π is the projection function from 3D to the image plane, and $P_t^G = P_t^G(\theta, \mathbf{d})$. We assume the pinhole projection model. If the camera calibration is unknown a vertical field of view of 54 degrees is assumed. Temporal stability is enforced with smoothness prior $E_{\text{smooth}} = \|P_t^G\|_2$, penalizing the acceleration \dot{P}_t^G . To counteract the strong depth uncertainty in monocular reconstruction, we penalize large variations in depth additionally with $E_{\text{depth}} = \|[\dot{P}_t^G]_z\|_2$ where $[\dot{P}_t^G]_z$ is the z component of 3D velocity \dot{P}_t^G . Finally, the 3D pose is also filtered with the 1 Euro filter [Casiez et al. 2012].

Parameters: The energy terms E_{IK} , E_{proj} , E_{smooth} and E_{depth} are weighted with $\omega_{\text{IK}} = 1$, $\omega_{\text{proj}} = 44$, $\omega_{\text{smooth}} = 0.07$ and $\omega_{\text{depth}} = 0.11$, respectively. The parameters of the 1 Euro Filter [Casiez et al. 2012] are empirically set to $f_{c_{\min}} = 1.7$, $\beta = 0.3$ for filtering K_t , to $f_{c_{\min}} = 0.8$, $\beta = 0.4$ for P_t^L , and to $f_{c_{\min}} = 20$, $\beta = 0.4$ for filtering P_t^G . Our implementation uses the Levenberg-Marquardt algorithm from the Ceres library [Agarwal et al. 2017].

5 RESULTS

We show live applications of our system at 30 Hz. The reconstruction quality is high and we demonstrate the usefulness of our method 3D character control, embodied virtual reality, and pose tracking from low quality smartphone camera streams. See Section 5.3 and Figure 1. Results are best observed in motion in the supplemental video. The importance of the steps towards enabling these applications with a video solution are thoroughly evaluated in more than 10 sequences. Results are comparable in quality to depth-camera based solutions like the Kinect [Microsoft Corporation 2013] and significantly outperform existing monocular video-based solutions. As the qualitative baseline we choose the state-of-the-art 2D to 3D lifting approach of Zhou et al. [2015] and the 3D regression approach of Mehta et al. [2016], which estimate joint-positions offline. The accuracy improvements are further quantitatively validated on the established H3.6M dataset [Ionescu et al. 2014b] and the MPI-INF-3DHP dataset [Mehta et al. 2016]. The robustness to diverse persons, clothing and scenes is demonstrated on several real-time examples

and community videos. Please see our project webpage for more results and details¹.

Computations are performed on a 6-core Xeon CPU, 3.8 GHz and a single Titan X (Pascal architecture) GPU. The CNN computation takes ≈ 18 ms, the skeleton fitting $\approx 7\text{--}10$ ms, and preprocessing and filtering 5 ms.

5.1 Comparison with Active Depth Sensors (Kinect)

We synchronously recorded video from an RGB camera and a co-located Kinect sensor in a living room scenario. Figure 6 shows representative frames. Although the depth sensor provides additional information, our reconstructions from just RGB are of a similar quality. The Kinect results are of comparable stability to ours, but yield erroneous reconstructions when limbs are close to scene objects, such as when sitting down. Our RGB solution, however, succeeds in this case, although is slightly less reliable in depth estimation. A challenging case for both methods is the tight crossing of legs. Please see the supplemental video for a visual comparison.

The video solution succeeds also in situations with direct sunlight (Figure 7), where IR-based depth cameras are inoperable. Moreover, RGB cameras can simply be equipped with large field-of-view (FOV) lenses and, despite strong distortions, successfully track humans [Rhodin et al. 2016a]. On the other hand, existing active sensors are limited to relatively small FOVs, which severely limits the tracking volume.

5.2 Comparison with Video Solutions

Qualitative Evaluation: We qualitatively compare against the 3D pose regression method of Mehta et al. [2016] and Zhou et al. [2015] on Sequence 6 (outdoor) of MPI-INF-3DHP test set. Our results are comparable to the quality of these offline methods (see Figure 10). However, the per frame estimates of these offline methods exhibits jitter over time, a drawback of most existing solutions. Our full pose results are temporally stable and are computed at real-time frame rate of 30 Hz.

The kinematic skeleton fitting estimates global translation \mathbf{d} . Figure 8 demonstrates that estimates are drift-free, the feet position matches with the same reference point after performing a circular walk. The smoothness constraint in depth direction limits sliding of the character away from the character, as pictured in the supplemental video sequences.

Quantitative Evaluation: We compare our method with the state-of-the-art approach of Mehta et al. [2016] on the MPI-INF-3DHP dataset, using the more robust Percentage of Correct Keypoints metric (3D PCK @150mm) on the 14 joints spanned by head, neck, shoulders, elbow, wrist, hips, knees and ankles. We train both, our model, as well as that of Mehta et al. on the same data (Human3.6m + MPI-INF-3DHP), as detailed in Section 4.1, to be compatible in terms of the camera viewpoints selected, and use ResNet100 as the base architecture for a fair comparison. Table 1 shows the results of the raw 3D predictions from our network on ground-truth bounding box cropped frames. We see that the results are comparable to that of Mehta et al. The slight increase in accuracy on going to a

¹<http://gvv.mpi-inf.mpg.de/projects/VNect/>

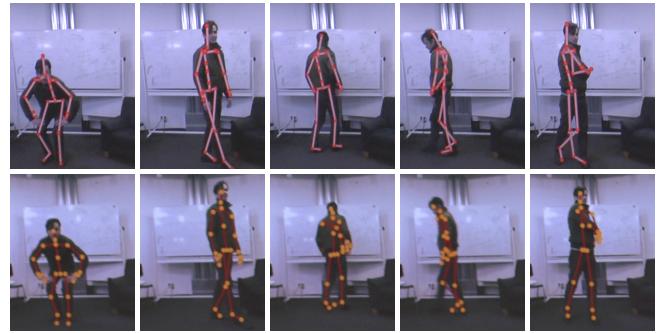


Fig. 6. Side-by-side pose comparison with our method (top) and Kinect (bottom). Overall estimated poses are of similar quality (first two frames). Both the Kinect (third and fourth frames) and our approach (fourth and fifth frames) occasionally predict erroneous poses.



Fig. 7. Our approach succeeds in strong illumination and sunlight (center right and right), while the IR-based depth estimates of the Microsoft Kinect are erroneous (left) and depth-based tracking fails (center left).



Fig. 8. The estimated 3D pose is drift-free. The motion of the person starts and ends at the marked point (orange), both in the real world and in our reconstruction.

50-layer network is possibly due to the better gradient estimates coming from larger mini-batches that can be fit into memory while training, on account of the smaller size of the network. Evidence that our method ties the estimated 3D positions strongly to image appearance than previous methods can also be gleaned from the fact that our approach performs significantly better for activity classes such as Standing/Walking, Sports and Miscellaneous without significant self-occlusions. We do lose some performance on activity classes with significant self-occlusion such as Sitting/Lying on the floor. We additionally report the Mean Per Joint Position Error (MPJPE) numbers in mm. Note that MPJPE is not a robust measure, and is heavily influenced by large outliers, and hence the worse performance on the MPJPE measure (124.7mm vs 117.6mm) despite the better 3D PCK results (76.6% vs 75.7%).

We further investigate the nature of errors of our method. We first look at the joint-wise breakup of accuracy of our fully-convolutional

Table 1. Comparison of our network against state of the art on MPI-INF-3DHP test set, using ground-truth bounding boxes. We report the Percentage of Correct Keypoints measure in 3D, and the Area Under the Curve for the same, as proposed by MPI-INF-3DHP. We additionally report the Mean Per Joint Position Error in mm. Higher PCK and AUC is better, and lower MPJPE is better.

Network	Scales	Stand/ Walk	Exercise	Sit On Chair	Crouch/ Reach	On the Floor	Sports	Misc.	Total		
		PCK	PCK	PCK	PCK	PCK	PCK	PCK	PCK	AUC	MPJPE(mm)
Ours (ResNet 100)	0.7, 1.0	87.6	76.4	71.4	71.6	47.8	82.5	78.9	75.0	39.5	127.8
	1.0	86.4	72.3	68.0	65.4	40.7	80.5	76.3	71.4	36.9	142.8
Ours (ResNet 50)	0.7, 1.0	87.7	77.4	74.7	72.9	51.3	83.3	80.1	76.6	40.4	124.7
	1.0	86.7	73.9	69.8	66.1	44.7	82.0	79.4	73.3	37.8	138.7
[Mehta et al. 2016] (ResNet 100)	0.7, 1.0	86.6	75.3	74.8	73.7	52.2	82.1	77.5	75.7	39.3	117.6
	1.0	86.3	72.4	71.5	67.6	49.2	81.0	76.2	73.2	37.8	126.6

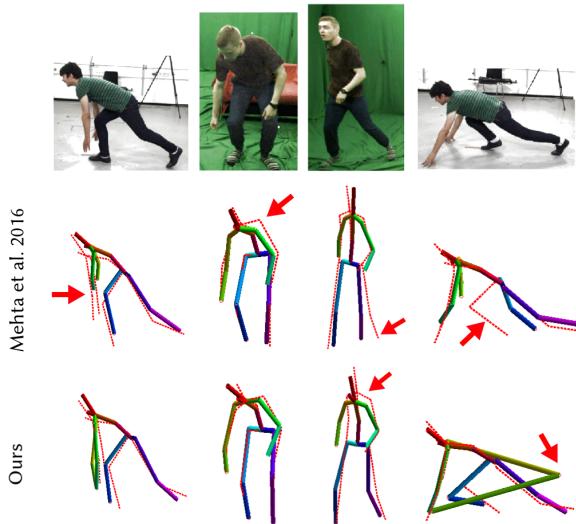


Fig. 9. A visual look at the direct 3D predictions resulting from our fully-convolutional formulation vs Mehta et al. Our formulation allows the predictions to be more strongly tied to image evidence, leading to overall better pose quality, particular for the end effectors. The red arrows point to mispredictions.

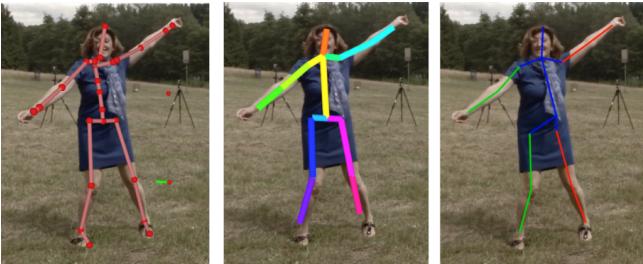


Fig. 10. Side-by-side comparison of our full method (left), against the offline joint-position estimation methods of Mehta et al. [2016] (middle) and Zhou et al. [2015] (right). Our real-time results are of a comparable quality to these offline methods. 2D joint positions for Zhou et al. are generated using Convolutional Pose Machines [2016].

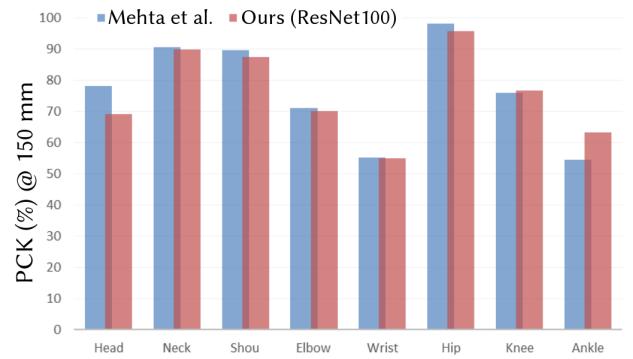


Fig. 11. Joint-wise breakdown of the accuracy of Mehta et al. and Our ResNet100 based CNN predictions on MPI-INF-3DHP test set.

ResNet100 CNN predictions vs Mehta et al.’s formulation with fully-connected layers. Figure 11 shows that the accuracy of ankles for our formulation is significantly better, while the accuracy of the head is markedly worse.

In Figure 9, we visually compare the two methods, further demonstrating the strong tie-in to image appearance that our formulation affords, and the downsides of the strong tie-in. We also show that our method is prone to occasional large mispredictions when the body joint 2D location detector misfires. It is these large outliers that obfuscate the reported MPJPE numbers. Figure 12, which plots the fraction of mispredicted joints vs. the error threshold on MPI-INF-3DHP test set shows that our method has a higher fraction of per-joint mispredictions beyond 300mm. It explains the higher MPJPE numbers compared to Mehta et al. despite equivalent PCK performance. The various filtering stages employed in the full pipeline ameliorate these large mispredictions.

For Human3.6m, we follow the protocol as in earlier work [Pavlakos et al. 2016; Tekin et al. 2016b,c], and evaluate on all actions and cameras for subject number 9 and 11, and report Mean Per Joint Position Error (mm) for root relative 3D joint positions from our network. See Table 3. Note that despite the occasional large outliers affecting the MPJPE measure, our predictions are still better than most of the existing methods.

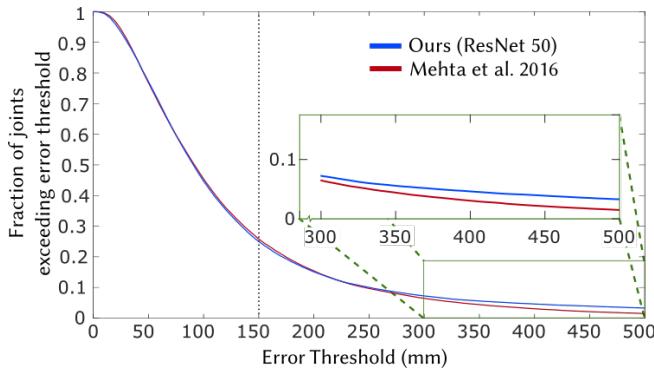


Fig. 12. Fraction of joints incorrectly predicted on MPI-INF-3DHP test set, as determined by the distance between the predicted joint location and the ground truth joint location being greater than the error threshold. The dotted line marks the threshold for which the 3D PCK numbers are reported. At bottom right we see that our method has larger occasional mispredictions, which result in higher MPJPE numbers despite otherwise similar performance.

Table 2. Results on MPI-INF-3DHP test set with the bounding box corners randomly jittered between ± 40 px to emulate noise from a BB estimator. Our fully-convolutional formulation is more robust than a comparative fully-connected formulation. The evaluation is at a single scale (1.0).

Network	Stand	Walk	Sit On	Crouch	On the	Floor	Sport	Misc.	Total
	PCK	PCK	PCK	PCK	PCK	PCK	PCK	PCK	
Ours (ResNet 100)	86.0	71.0	65.0	61.1	37.4	78.9	75.5	69.5	35.8
Ours (ResNet 50)	84.9	69.4	65.1	61.9	40.8	78.6	77.6	70.1	35.7
[Mehta et al. 2016]	81.2	64.2	67.1	62.1	43.5	76.0	71.1	67.8	34.0

The accuracy attained from single view methods is still below that of real-time multi-view methods, which can achieve a mean accuracy of the order of 10mm [Stoll et al. 2011].

Generalization to Different Persons and Scenes: We tested our method on a variety of actors, it succeeds for different body shapes, gender and skin tone. See supplemental video. To further validate the robustness we applied the methods to community videos from YouTube, see Figure 1. It generalizes well to the different backgrounds and camera types.

Model Components: To demonstrate that our fully-convolutional pose formulation is less sensitive to inexact cropping than networks using a fully-connected formulation, we emulate a noisy BB estimator by jittering the ground-truth bounding box corners of MPI-INF-3DHP test set uniformly at random in the range of ± 40 px. This also captures scenarios where one or more end effectors are not in the frame, so a loss in accuracy is expected for all methods. Table 2 shows that the fully-connected formulation of Mehta et al. suffers a worse hit in accuracy than our approach, going down by 7.9 PCK, while our comparable network goes down by only 5.5 PCK.

We show the effect of the various components of our full pipeline on the TS1 sequence of MPI-INF-3DHP test set in Figure 13. Without the E_{IK} component of E_{total} the tracking accuracy goes down to a

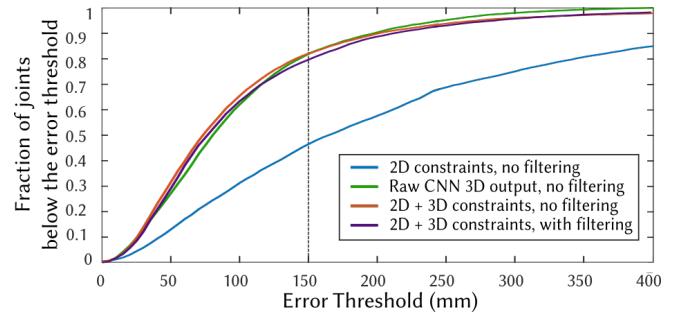


Fig. 13. Fraction of joints correctly predicted on the TS1 sequence of MPI-INF-3DHP test set, as determined by the distance between the predicted joint location and the ground truth joint location being below the error threshold. The dotted line marks the 150mm threshold for which the 3D PCK numbers are reported. We see that only using the 2D predictions as constraints for skeleton fitting (blue) performs significantly worse than using both 2D and 3D predictions as constraints (red). Though adding 1 Euro filtering (purple) visually improves the results, the slightly higher error here is due to the sluggish recovery from tracking failures. The 3D predictions from the CNN (green) are also shown.

PCK of 46.1% compared to a PCK of 81.7% when E_{IK} is used. The raw CNN 3D predictions in conjunction with the BB tracker result in a PCK of 80.3%. Using E_{IK} in E_{total} produces consistently better results for all thresholds lower than 150 mm. This shows the improvements brought about by our skeleton fitting term. Additionally, as shown in the supplementary video, using 1 Euro filtering produces qualitatively better results, but the overall PCK decreases slightly (79.7%) due to slower recovery from tracking failures. The influence of the smoothness and filtering steps on the temporal consistency are further analyzed in the supplemental video.

5.3 Applications

Our approach is suitable for various interactive applications since it is real-time, temporally stable, fully automatic, and exports data directly in a format amenable to 3D character control.

Character Control: Real-time motion capture solutions provide a natural interface for game characters and virtual avatars, which go beyond classical mouse and gamepad control. We applied our method on motions common in activities like tennis, dance, and juggling, see Figures 1 and 14. The swing of the arm and leg motion is nicely captured and could, for instance, be used in a casual sports and dancing game, but also for motion analysis of professional athletes to optimize their motion patterns. We also show successful results in non front-facing motions such as turning and writing on a wall, as well as squatting.

Virtual Reality: The recent availability of cheap head-mounted displays has sparked a range of new applications. Many products use handheld devices to track the user's hand position for interaction. Our solution enables them from a single consumer color camera. Beyond interaction, our marker-less full-body solution enables embodied virtual reality, see Figure 1. A rich immersive feeling is created by posing a virtual avatar of the user exactly to their own real pose. With our solution the real and virtual pose are aligned such that users perceive the virtual body as their own.

Table 3. Results of our raw CNN predictions on Human3.6m, evaluated on the ground truth bounding box crops for all frames of Subject 9 and 11. Our CNNs use only Human3.6m as the 3D training set, and are pretrained for 2D pose prediction. The error measure used is Mean Per Joint Position Error (MPJPE) in millimeters. Note again that the error measure used is not robust, and subject to obfuscation from occasional large mispredictions, such as those exhibited by our raw CNN predictions.

Method	Direct	Discuss	Eating	Greet	Phone	Posing	Purch.	Sitting	Sit Down	Smoke	Take Photo	Wait	Walk	Walk Dog	Walk Pair	All
[Zhou et al. 2015b]	87.4	109.3	87.1	103.2	116.2	106.9	99.8	124.5	199.2	107.4	143.3	118.1	79.4	114.2	97.7	113.0
[Tekin et al. 2016c]	102.4	147.7	88.8	125.3	118.0	112.3	129.2	138.9	224.9	118.4	182.7	138.8	55.1	126.3	65.8	125.0
[Yu et al. 2016]	85.1	112.7	104.9	122.1	139.1	105.9	166.2	117.5	226.9	120.0	135.9	117.7	137.4	99.3	106.5	126.5
[Ionescu et al. 2014b]	132.7	183.6	132.4	164.4	162.1	150.6	171.3	151.6	243.0	162.1	205.9	170.7	96.6	177.1	127.9	162.1
[Zhou et al. 2016]	91.8	102.4	97.0	98.8	113.4	90.0	93.8	132.2	159.0	106.9	125.2	94.4	79.0	126.0	99.0	107.3
[Pavlakos et al. 2016]	58.6	64.6	63.7	62.4	66.9	57.7	62.5	76.8	103.5	65.7	70.7	61.6	69.0	56.4	59.5	66.9
[Mehta et al. 2016]	52.6	63.8	55.4	62.3	71.8	52.6	72.2	86.2	120.6	66.0	79.8	64.0	48.9	76.8	53.7	68.6
[Tekin et al. 2016b]	85.0	108.8	84.4	98.9	119.4	98.5	93.8	73.8	170.4	85.1	95.7	116.9	62.1	113.7	94.8	100.1
Ours (ResNet 100)	61.7	77.8	64.6	70.3	90.5	61.9	79.8	113.2	153.1	80.9	94.4	75.1	54.9	83.5	61.0	82.5
Ours (ResNet 50)	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5

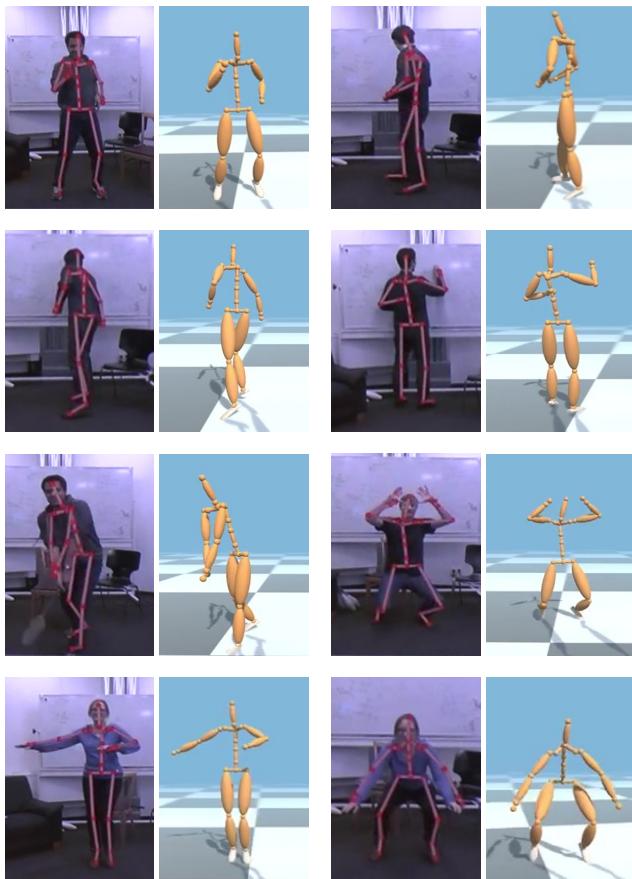


Fig. 14. Application to entertainment. The real-time 3D pose estimation method provides a natural motion interface, e.g. for sport games.

Ubiquitous Motion Capture with Smartphones: Real-time monocular 3D pose estimation lends itself to application on low quality smartphone video streams. By streaming the video to a machine with

sufficient capabilities for our algorithm, one can turn any smartphone into a lightweight, fully-automatic, handheld motion capture sensor, see Figure 15 and the accompanying video. Since smartphones are widespread, it enables the aforementioned applications for casual users without requiring additional sensing devices.

6 LIMITATIONS

Depth estimation from a monocular image is severely ill posed, slight inaccuracies in the estimation can lead to largely different depth estimates, which manifests also in our results in slight temporal jitter. We claim improved stability and temporal consistency compared to existing monocular RGB 3D pose estimation methods. This uncertainty could be further reduced with domain specific knowledge, e.g., foot-contact constraints when the floor location is known, and head-pose stabilization with the position of head-mounted-displays in VR applications, which is readily obtained with IMU-sensors.

A downside of our CNN prediction formulation is that mispredictions of 2D joint locations result in implausible 3D poses. This is ameliorated in the tracker through skeleton retargeting and pose filtering. This could be addressed directly in the CNN through imposition of stronger interdependencies between predictions. Additionally, the performance on poses with significant amounts of self occlusion remains a challenge.

Further, very fast motions can exceed the convergence radius of our IK optimization, but the integration of per frame 2D and 3D pose yields quick recovery from erroneous poses. Initial experiments with 256×256 px input to the CNN show that much higher frame rates are possible with no loss in accuracy.

7 DISCUSSION

The availability of sufficient annotated 3D pose training data remains an issue. Even the most recent annotated real 3D pose data sets, or combined real/synthetic data sets [Chen et al. 2016; Ionescu et al. 2014b; Mehta et al. 2016] are a subset of real world human pose, shape, appearance and background distributions. Recent top performing methods explicitly address this data sparsity by training

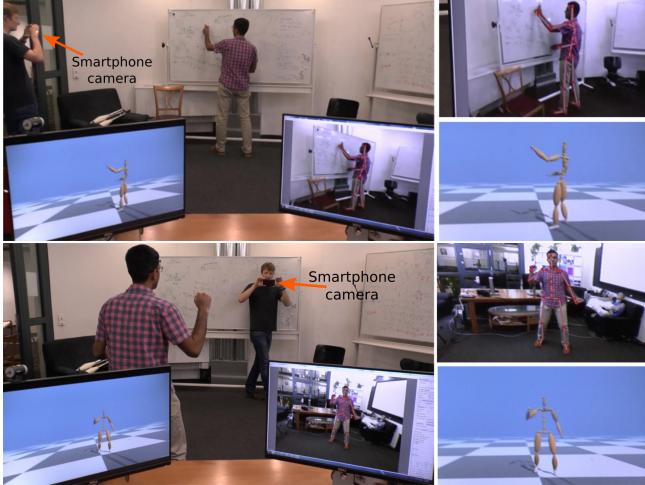


Fig. 15. Handheld recording with a readily available smartphone camera (left) and our estimated pose (right), streamed to and processed by a GPU enabled PC.

similarly deep networks, but with architectural changes enabling improved intermediate training supervision [Mehta et al. 2016].

Our implementation only supports a single person, although the proposed fully-convolutional formulation could be scaled to multiple persons. Such an extension is currently precluded due to the lack of multi-person datasets, required to train multi-person 3D pose regressors. One possible approach is to adapt the multi-person 2D pose methods of Insafutdinov et al. [2016] and Cao et al. [2016].

We also analyze the impact of 2D joint position mispredictions on the 3D joint position predictions from our fully-convolutional formulation. We decouple the 3D predictions from the 2D predictions by looking up the 3D joint positions from their location-maps using the ground truth 2D joint positions. See Table 4. We see a 3D PCK improvement of 2.8, which is congruent with the notion of a stronger tie-in of the predicted joint positions with the image plane, which causes the 3D joint predictions to be erroneous when 2D joint detection misfires. The upside of this is that the 3D predictions can be improved through improvements to 2D joint position prediction. Alternatively, optimization formulations that directly operate on the heatmaps and the location-maps could be constructed. Our fully-convolutional formulation can also benefit from iterative refinement, akin to heatmap-based 2D pose estimation approaches [Hu et al. 2016; Newell et al. 2016].

8 CONCLUSION

We have presented the first method that estimates the 3D kinematic pose of a human, including global position, in a stable, temporally consistent manner from a single RGB video stream at 30 Hz. Our approach combines a fully-convolutional CNN that regresses 2D and 3D joint positions and a kinematic skeleton fitting method, producing a real-time temporally stable 3D reconstruction of the motion. In contrast to most existing approaches, our network can operate without strict bounding boxes, and facilitates inexpensive bounding box tracking. We have shown results in a variety of challenging real-time scenarios, including live streaming from a smartphone

Table 4. Results on MPI-INF-3DHP test set with the 3D joint position lookup in the location-maps done using the ground truth 2D locations rather than the predicted 2D locations. We see that the location maps have captured better 3D pose information, which can perhaps be extracted through optimization methods operating directly on heatmaps and location-maps. The evaluation uses 2 scales (0.7, 1.0).

Network	Stand/	Walk	Sit On	Crouch/	On the	Sport	Misc.	Total	
	Exerc.	PCK	PCK	Chair	Reach	Floor	PCK	PCK	
Ours (ResNet 100)	88.1	80.9	74.0	76.1	56.3	82.9	80.2	77.8	41.0
Ours (ResNet 50)	88.0	81.8	78.6	77.4	59.3	82.8	81.2	79.4	41.6
[Mehta et al. 2016]	86.6	75.3	74.8	73.7	52.2	82.1	77.5	75.7	39.3

camera, as well as in community videos. A number of applications have been presented, such as embodied VR and interactive character control for computer games.

Qualitative and quantitative evaluations demonstrate that our approach compares to offline state-of-the-art monocular RGB methods and approaches the quality of real-time RGB-D methods. Hence, we believe our method is a significant step forward to democratizing 3D human pose estimation, lifting both the need for special cameras such as the IR-based depth cameras, as well as the long and heavy processing times.

REFERENCES

- Ankur Agarwal and Bill Triggs. 2006. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 28, 1 (2006), 44–58.
- Sameer Agarwal, Keir Mierle, and Others. 2017. Ceres Solver. <http://ceres-solver.org>. (2017).
- Ijaz Akhter and Michael J Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1446–1455.
- Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. 2013. Multi-view Pictorial Structures for 3D Human Pose Estimation. In *BMVC*.
- Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. 2009. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1014–1021.
- Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. 2015. Real-Time Pedestrian Detection With Deep Network Cascades. In *Proceedings of BMVC 2015*.
- Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. 2011. A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. In *IEEE International Conference on Computer Vision (ICCV)*.
- Alexandru O Balan, Leonid Sigal, and Michael J Black. 2005. A quantitative evaluation of video-based 3D person tracking. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 349–356.
- Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Hausseder. 2007. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8.
- Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 2014. 3D pictorial structures for multiple human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1669–1676.
- Vasileios Belagiannis and Andrew Zisserman. 2016. Recurrent Human Pose Estimation. *arXiv preprint arXiv:1605.02914* (2016).
- Alessandro Bissacco, Ming-Hsuan Yang, and Stefano Soatto. 2007. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- Liefeng Bo and Cristian Sminchisescu. 2010. Twin gaussian processes for structured prediction. *International Journal of Computer Vision* 87, 1-2 (2010), 28–52.
- Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and

- Shape from a Single Image. In *European Conference on Computer Vision (ECCV)*. Lubomir Bourdev and Jitendra Malik. 2009. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision (ICCV)*. 1365–1372.
- Ernesto Brau and Hao Jiang. 2016. 3D Human Pose Estimation via Deep Learning from 2D Annotations. In *International Conference on 3D Vision (3DV)*.
- Christoph Bregler and Jitendra Malik. 1998. Tracking people with twists and exponential maps. In *Conference on Computer Vision and Pattern Recognition*. 8–15.
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv preprint arXiv:1611.08050* (2016).
- Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1aCn filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2527–2530.
- Jinxiang Chai and Jessica K Hodgins. 2005. Performance animation from low-dimensional control signals. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 686–696.
- Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2016. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In *International Conference on 3D Vision (3DV)*.
- Martin de La Gorce, Nikos Paragios, and David J Fleet. 2008. Model-based hand tracking with texture, shading and self-occlusions. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference On*. IEEE, 1–8.
- Jonathan Deutscher and Ian Reid. 2005. Articulated body motion capture by stochastic search. *International Journal of Computer Vision* 61, 2 (2005), 185–205.
- Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts-Escalano, Christoph Rhemann, David Kim, Jonathan Taylor, and others. 2016. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 114.
- Ahmed Elgammal and Chan-Su Lee. 2004. Inferring 3D body pose from silhouettes using activity manifold learning. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 2. IEEE, II–681.
- Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt. 2016. MARConI - ConvNet-based MARker-less Motion Capture in Outdoor and Indoor Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2016).
- Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. In *IEEE transactions on pattern analysis and machine intelligence*. IEEE, 1627–1645.
- Pedro F Felzenszwalb and Daniel P Huttenlocher. 2005. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)* 61, 1 (2005), 55–79.
- Vittorio Ferrari, Manuel Marin-Jimenez, and Andrew Zisserman. 2009. Pose search: retrieving people using their pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8.
- Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. 2010. Optimization and Filtering for Human Motion Capture. *International Journal of Computer Vision (IJCV)* 87, 1–2 (2010), 75–92.
- Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. 2012. Real-time human pose tracking from range data. In *European conference on computer vision*. Springer, 738–751.
- Ravi Garg, Anastasios Roussos, and Lourdes Agapito. 2013. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1272–1279.
- Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. 2011. Efficient regression of general-activity human poses from depth images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 415–422.
- Paulo FU Gotardo and Aleix M Martinez. 2011. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 10 (2011), 2051–2065.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Nicholas R Howe, Michael E Leventon, and William T Freeman. 1999. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video.. In *NIPS*, Vol. 99. 820–6.
- Peiyun Hu, Deva Ramanan, Jia Jia, Sen Wu, Xiaohui Wang, Lianhong Cai, and Jie Tang. 2016. Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Matthias Inmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. (October 2016), 17.
- Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. DeepCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision (ECCV)*.
- Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of The 32nd International Conference on Machine Learning*. 448–456.
- Catalin Ionescu, Joao Carreira, and Cristian Sminchisescu. 2014a. Iterated second-order label sensitive pooling for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1661–1668.
- Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014b. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 36, 7 (2014), 1325–1339.
- Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. 2010. MovieReshape: Tracking and Reshaping of Humans in Videos. *ACM Transactions on Graphics* 29, 5 (2010). DOI: <https://doi.org/10.1145/1866158.1866174>
- Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 675–678.
- Sam Johnson and Mark Everingham. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference (BMVC)*. doi:10.5244/C.24.12.
- Sam Johnson and Mark Everingham. 2011. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Minsik Lee, Jungchan Cho, Chong-Ho Choi, and Songhwai Oh. 2013. Procrustean normal distribution for non-rigid structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1280–1287.
- Sijin Li and Antoni B Chan. 2014. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*. 332–347.
- Sijin Li, Weichen Zhang, and Antoni B Chan. 2015a. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. 2848–2856.
- Sijin Li, Weichen Zhang, and Antoni B Chan. 2015b. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. 2848–2856.
- Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. 2016. Human Pose Estimation using Deep Consensus Voting. In *European Conference on Computer Vision (ECCV)*.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott E. Reed. 2016. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision (ECCV)*.
- Matthew M Loper and Michael J Black. 2014. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision*. Springer, 154–169.
- Ziyang Ma and Enhua Wu. 2014. Real-time and robust hand tracking with a single depth camera. *The Visual Computer* 30, 10 (2014), 1133–1144.
- Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2016. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *arXiv preprint arXiv:1611.09813v2* (2016).
- Alberto Menache. 2000. *Understanding motion capture for computer animation and video games*. Morgan kaufmann.
- Microsoft Corporation. 2010. Kinect for Xbox 360. <http://www.xbox.com/en-US/xbox-360/accessories/kinect>. (2010).
- Microsoft Corporation. 2013. Kinect for Xbox One. <http://www.xbox.com/en-US/xbox-one/accessories/kinect>. (2013).
- Microsoft Corporation. 2015. Kinect SDK. <https://developer.microsoft.com/en-us/windows/kinect>. (2015).
- Thomas B. Moeslund, Adrian Hilton, and Volker Kräiger. 2006. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *CVIU* 104, 2–3 (2006), 90–126.
- Greg Mori and Jitendra Malik. 2006. Recovering 3d human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 28, 7 (2006), 1052–1062.
- Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alejandra Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *European Conference on Computer Vision (ECCV)*.
- Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect.. In *BmVC*, Vol. 1. 3.
- Sergio Orts-Escalano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, and others. 2016. Holoporation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 741–754.
- Hyun Soo Park and Yaser Sheikh. 2011. 3D reconstruction of a smooth articulated trajectory from a monocular image sequence. In *International Conference on Computer Vision (ICCV)*. 201–208.

- Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2016. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. *arXiv preprint arXiv:1611.07828* (2016).
- Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2013. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3487–3494.
- Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gerard Pons-Moll, David J Fleet, and Bodo Rosenhahn. 2014. Posebits for monocular human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2337–2344.
- Real Madrid C.F. 2016. Cristiano Ronaldo and Coentrao continue their recovery. https://www.youtube.com/watch?v=xqiPuX_buOo. (2016).
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640* (2015).
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
- Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016a. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* (2016).
- Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. 2016b. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision (ECCV)*. Springer, 509–526.
- Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. 2015. A Versatile Scene Model With Differentiable Visibility Applied to Generative Pose Estimation. In *ICCV*.
- Grégory Rogez and Cordelia Schmid. 2016. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. *arXiv preprint arXiv:1607.02046* (2016).
- Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. 2014. Garment replacement in monocular video sequences. *ACM Transactions on Graphics (TOG)* 34, 1 (2014), 6.
- Rómer Rosales and Stan Sclaroff. 2000. Specialized mappings and the estimation of human body pose from a single image. In *Human Motion, 2000. Proceedings. Workshop on*. IEEE, 19–24.
- Rómer Rosales and Stan Sclaroff. 2006. Combining generative and discriminative models in a framework for articulated pose estimation. *International Journal of Computer Vision* 67, 3 (2006), 251–276.
- RUSFENCING-TV. 2017. The Most Beautiful Strike / Saber Woman (Translated from Russian). <https://www.youtube.com/watch?v=0gOcMsWUkCU>. (2017).
- Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.
- Hedvig Sidenbladh, Michael J Black, and David J Fleet. 2000. Stochastic tracking of 3D human figures using 2D image motion. In *European conference on computer vision*. Springer, 702–718.
- Leonid Sigal, Michael Isard, Horst Haussecker, and Michael J Black. 2012. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision (IJCV)* 98, 1 (2012), 15–48.
- Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. 2006. Learning joint top-down and bottom-up processes for 3D visual inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1743–1752.
- Cristian Sminchisescu, Atul Kanaujia, and Dimitris N Metaxas. 2007. BM³E: Discriminative Density Propagation for Visual Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 11 (2007), 2030–2044.
- Cristian Sminchisescu and Bill Triggs. 2001. Covariance scaled sampling for monocular 3D body tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, 1–447.
- Jonathan Starck and Adrian Hilton. 2003. Model-based multiple view reconstruction of people. In *IEEE International Conference on Computer Vision (ICCV)*. 915–922.
- Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Fast articulated motion tracking using a sums of Gaussians body model. In *IEEE International Conference on Computer Vision (ICCV)*. 951–958.
- Leonid Taycher, David Demirdjian, Trevor Darrell, and Gregory Shakhnarovich. 2006. Conditional random people: Tracking humans with crfs and grid filters. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 222–229.
- Camillo J Taylor. 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 677–684.
- Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. 2016a. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference (BMVC)*.
- Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2016b. Fusing 2D Uncertainty and 3D Cues for Monocular Body Pose Estimation. *arXiv preprint arXiv:1611.05708* (2016).
- Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. 2016c. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*. 1799–1807.
- Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1653–1660.
- Raquel Urtasun, David J Fleet, and Pascal Fua. 2006. Temporal motion models for monocular and multiview 3d human body tracking. *Computer vision and image understanding* 104, 2 (2006), 157–177.
- Marek Vondrák, Leonid Sigal, Jessica Hodgins, and Odest Jenkins. 2012. Video-based 3D motion capture through biped control. *ACM Transactions On Graphics (TOG)* 31, 4 (2012), 27.
- Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. 2014. Robust estimation of 3d human poses from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2361–2368.
- Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiaolin Wei and Jinxiang Chai. 2010. Videomocap: modeling physically realistic human motion from monocular video sequences. In *ACM Transactions on Graphics (TOG)*, Vol. 29. ACM, 42.
- Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 188.
- Christopher Richard Wren, Ali Azarbajayani, Trevor Darrell, and Alex Paul Pentland. 1997. Pfnder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 19, 7 (1997), 780–785.
- Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall. 2016. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Mao Ye and Ruigang Yang. 2014. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2345–2352.
- Yongkang Yu, Feilinand Yonghao, Zhen Yilin, and Weidong Mohan. 2016. Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. In *European Conference on Computer Vision (ECCV)*.
- Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).
- Xiaowei Zhou, Spyridon Leonards, Xiaoyan Hu, and Kostas Daniilidis. 2015. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4447–4455.
- Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. 2016. Deep Kinematic Pose Regression. *ECCV Worktp on Geometry Meets Deep Learning*.
- Xiaowei Zhou, Menglong Zhu, Spyridon Leonards, and Kostas Daniilidis. 2015a. Sparse Representation for 3D Shape Estimation: A Convex Relaxation Approach. *arXiv preprint arXiv:1509.04309* (2015).
- Xiaowei Zhou, Menglong Zhu, Spyridon Leonards, Kosta Derpanis, and Kostas Daniilidis. 2015b. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yingying Zhu, Mark Cox, and Simon Lucey. 2011. 3D motion reconstruction for real-world camera motion. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1–8.
- Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. 2014. Real-time Non-rigid Reconstruction using an RGB-D Camera. *ACM Transactions on Graphics (TOG)* 33, 4 (2014).