MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G. Derpanis and Kostas Daniilidis, *Fellow, IEEE*

Abstract—Recovering 3D full-body human pose is a challenging problem with many applications. It has been successfully addressed by motion capture systems with body worn markers and multiple cameras. In this paper, we address the more challenging case of not only using a single camera but also not leveraging markers: going directly from 2D appearance to 3D geometry. Deep learning approaches have shown remarkable abilities to discriminatively learn 2D appearance features. The missing piece is how to integrate 2D, 3D and temporal information to recover 3D geometry and account for the uncertainties arising from the discriminative model. We introduce a novel approach that treats 2D joint locations as latent variables whose uncertainty distributions are given by a deep fully convolutional neural network. The unknown 3D poses are modeled by a sparse representation and the 3D parameter estimates are realized via an Expectation-Maximization algorithm, where it is shown that the 2D joint location uncertainties can be conveniently marginalized out during inference. Extensive evaluation on benchmark datasets shows that the proposed approach achieves greater accuracy over state-of-the-art baselines. Notably, the proposed approach does not require synchronized 2D-3D data for training and is applicable to "in-the-wild" images, which is demonstrated with the MPII dataset.

Index Terms—Motion capture, human pose, deep learning, sparse representation.

1 Introduction

This paper is concerned with the challenge of recovering the 3D full-body human pose from a markerless monocular RGB image sequence. Potential applications of our work include human-computer interaction, surveillance, rehabilitation, sports, video browsing and indexing, and virtual reality. Typical solutions for this task include motion capture (MoCap) systems with multiple cameras and reflective markers and depth sensors, e.g., Microsoft Kinect [1]. These techniques require customized devices, are limited to applications in constrained environments, and cannot be applied to archival RGB images or videos. This paper addresses the pose recovery challenge by using a single camera and avoiding the use of markers: going directly from 2D appearance to 3D geometry.

From a geometric perspective, 3D articulated pose recovery is inherently ambiguous from monocular imagery [2]. A considerable amount of work has tackled the geometric problem to reconstruct 3D human pose from 2D correspondences via articulated constraints [3], low-rank priors [4], sparse representations [5], or tracking with a body model [6]. These approaches typically assume 2D correspondences are provided or require careful initialization for frame-to-frame tracking based on low-level image features. Finding 2D correspondences is rendered difficult due to the large variation in human appearance (e.g., clothing, body shape, and illumination), arbitrary camera viewpoint, and obstructed visibility due to self-occlusions and external entities. Notable successes in 2D pose estimation have used discriminatively trained 2D part models coupled with 2D deformation priors, e.g., [7], [8], [9], and more recently using deep learning, e.g., [10]. Here, the 3D pose geometry is not leveraged. Combining robust image-driven 2D part detectors, expressive 3D geometric pose priors and temporal models to aggregate information over time is a promising area of research that has been given limited attention, e.g., [11], [12]. The challenge posed is how to seamlessly integrate 2D, 3D and temporal information to fully account for the model and measurement uncertainties.

This paper presents a 3D human pose estimation framework called MonoCap that consists of a novel synthesis between discriminative image-based and 3D reconstruction approaches. In particular, the approach reasons jointly about image-based 2D part location estimates and model-based 3D pose reconstruction, so that they can benefit from each other. Further, to improve the approach's robustness against detector error, occlusion, and reconstruction ambiguity, temporal smoothness is imposed on the 3D pose and viewpoint parameters. Figure 1 provides an overview of our approach. Given the input video (Fig. 1, top-left), 2D joint heat maps capturing positional uncertainty are generated with a deep fully convolutional neural network (CNN) (Fig. 1, top-right). These heat maps are combined with a sparse model of 3D human pose (Fig. 1, bottom-left) within an Expectation-Maximization (EM) framework to recover the 3D pose sequence (Fig. 1, bottom-right).

1.1 Related work

Considerable research has addressed the challenge of human motion capture from imagery [13], [14], [15], [16], [17].

- *X.Z. is with the College of Computer Science, Zhejiang University, China. M.Z., G.P., S.L. and K.D. are with Computer and Information Science Department and GRASP Laboratory, University of Pennsylvania, USA. K.G.D. is with the Department of Computer Science, Ryerson University, Canada.*
  E-mail: xzhou@cad.zju.edu.cn

Manuscript received XXX; revised XXX.

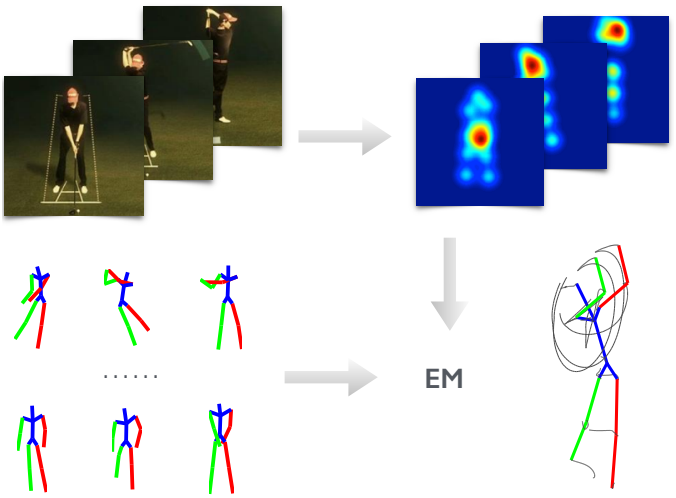


Fig. 1. Overview of proposed approach. (top-left) Input image sequence, (top-right) CNN-based heat map outputs representing the soft localization of 2D joints, (bottom-left) 3D pose dictionary, and (bottom-right) the recovered 3D pose sequence reconstruction. To fully account for uncertainty, the problem is addressed in a probabilistic framework where the 2D joint locations are modeled as latent variables and marginalized in an EM algorithm. Temporal smoothness in 3D is also imposed.

This work includes 2D human pose recovery in both single images (e.g., [7], [10], [18], [19], [20]), and video, e.g., [21], [22], [9], [23], [24], [25], [26]. In the current work, focus is placed on 3D pose recovery, where the pose model and prior are expressed in their natural 3D domain.

Early research on 3D monocular pose estimation in videos largely centered on generative models for frame-to-frame pose tracking, e.g., [6], [27]. These approaches rely on a given pose and dynamic model to constrain the pose search space. Notable drawbacks of this approach include: the requirement that the initialization be provided, and their inability to recover from tracking failures. To address these limitations, bottom-up models were proposed in more recent works, e.g., the "loose-limbed people" [28] and "tracking-by-detection" [11].

Another strand of research has focused on discriminative methods that predict 3D poses by searching a database of exemplars [29], [30], [31], [32] or via a discriminatively learned mapping from the image directly to human joint locations [33], [34], [35], [36], [37], [38]. Recently, deep convolutional networks (CNNs) have emerged as a common element behind many state-of-the-art approaches, including 3D human pose estimation, e.g., [39], [40], [41], [42], [43], [44]. To deal with the scarcity of training data, some recent works synthesize training images via graphics rendering [45] or image mosaicing [46].

Most closely related to our work are generic factorization approaches for recovering 3D non-rigid shapes from image sequences captured with a single camera [4], [47], [48], [49], [50], i.e., non-rigid structure from motion (NRSFM), and human pose recovery models based on known skeletons [2], [3], [51], [52], [53], [54] or sparse representations [5], [55], [56], [57], [58]. Much of this work has been realized by assuming manually labeled 2D joint locations; however, there is some recent work that has used a 2D pose detector to automatically provide the input joints [59], [60] or solved 2D and 3D pose estimation jointly [61], [12].

1.2 Contributions

In the light of previous work, the proposed approach advances the state-of-the-art in the following ways. First, in contrast to prediction approaches (e.g., [40], [41]), our approach does not require synchronized 2D-3D data, as captured by MoCap systems. The proposed approach only requires readily available annotated 2D imagery (e.g., the "in-the-wild" MPII dataset [62]) to train a CNN part detector and a separate 3D MoCap dataset (e.g., the CMU MoCap database) for the pose dictionary. The flexibility of using separate sources of training data makes the proposed approach more widely applicable. In comparison to examplar-based methods (e.g., [31], [32]), the proposed approach does not need to store and enumerate all possible 2D views and can generalize to unseen poses. Compared to other 3D reconstruction methods (e.g., [5], [58]), the proposed approach does not rely on a hard decision of 2D correspondences before reconstruction and considers an arbitrary pose uncertainty. In contrast to prior work that consider model-image alignment (e.g., [63], [28], [64]), the current approach leverages CNNs to learn better 2D representations and sparsity-driven 3D pose optimization to allow efficient and global inference. Finally, empirical evaluation demonstrates that the proposed approach is more accurate compared to extant approaches. In particular, in the case where 2D joint locations are provided, the proposed approach exceeds the accuracy of the state-of-the-art NRSFM baseline [48] on the Human3.6M dataset [37]. In the case where the 2D joints are unknown, empirical results on the HumanEva I [65], Human3.6M [37], and KTH Football II [66] datasets demonstrate overall improvement over published results. Further, the qualitative results on the MPII dataset [62] demonstrate that the proposed approach is able to reconstruct 3D poses from single "in-the-wild" images with a 3D pose prior learned from a separate MoCap dataset.

A preliminary version of this work appeared in CVPR 2016 [67]. Here, the work is extended in the following ways: the proposed approach integrates a perspective camera model (as opposed to an orthographic model), a corresponding optimization algorithm is introduced, a state-of-the-art 2D pose detector is used, and the empirical evaluations are significantly expanded to make them more comprehensive. The code is available at https://github.com/daniilidis-group/monocap.

2 Models

In this section, the models that describe the relationships between 3D poses, 2D poses, and images are introduced.

2.1 Sparse representation of 3D poses

The 3D human pose is represented by the 3D locations of a set of p joints, denoted by $\boldsymbol{S}_t \in \mathbb{R}^{3\times p}$ for frame t. In general, single-view reconstruction is an ill-posed problem. To address this problem, it is assumed that a 3D pose can be represented as a linear combination of predefined basis poses:

$$\boldsymbol{S}_t = \sum_{i=1}^{k} c_{it}\boldsymbol{B}_i, \tag{1}$$

where $\boldsymbol{B}_i \in \mathbb{R}^{3\times p}$ denotes a basis pose and c_{it} the corresponding weight. The basis poses are learned from training poses provided by a MoCap dataset. Instead of using the conventional active shape model [68], where the basis set is relatively small, a sparse representation is adopted which has been shown in recent work to be capable of modelling the large variability of human pose, e.g., [5], [56], [58]. That is, an overcomplete dictionary, $\{\boldsymbol{B}_1, \cdots, \boldsymbol{B}_k\}$, is learned with a relatively large number of basis poses, k, where the coefficients, c_{it}, are assumed to be sparse. In the remainder of this paper, $\boldsymbol{c}_t = [c_{1t}, \cdots, c_{kt}]^\top$ denotes the coefficient vector for frame t and $\boldsymbol{C}$ the matrix composed of all $\boldsymbol{c}_t$.

2.2 Dependence between 2D and 3D poses

2.2.1 Orthographic projection model

When the camera intrinsic parameters are unknown, an orthographic camera model is used to describe the dependence between a 3D pose and its imaged 2D pose:

$$\boldsymbol{W}_t = \boldsymbol{R}_t \boldsymbol{S}_t + \boldsymbol{T}_t \boldsymbol{1}^\top, \tag{2}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{2\times p}$ denotes the 2D pose in frame t, and $\boldsymbol{R}_t \in \mathbb{R}^{2\times 3}$ and $\boldsymbol{T}_t \in \mathbb{R}^2$ the camera rotation and translation, respectively. In the following, $\boldsymbol{W}$, $\boldsymbol{R}$ and $\boldsymbol{T}$ denote the collections of $\boldsymbol{W}_t$, $\boldsymbol{R}_t$ and $\boldsymbol{T}_t$ for all t, respectively.

Considering the observation noise and model error, the conditional distribution of the 2D poses given the 3D pose parameters is modeled as

$$\Pr(\boldsymbol{W}|\theta) \propto e^{-\mathcal{L}(\theta;\boldsymbol{W})}, \tag{3}$$

where $\theta = \{\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{T}\}$ is the union of all the 3D pose parameters and the loss function, $\mathcal{L}(\theta; \boldsymbol{W})$, is defined as

$$\mathcal{L}(\theta; \boldsymbol{W}) = \frac{\nu}{2} \sum_{t=1}^{n} \left\| \boldsymbol{W}_t - \boldsymbol{R}_t \sum_{i=1}^{k} c_{it} \boldsymbol{B}_i - \boldsymbol{T}_t \boldsymbol{1}^\top \right\|_F^2, \tag{4}$$

with $\|\cdot\|_F$ denoting the Frobenius norm. The model in (3) states that given the 3D poses and camera parameters, the 2D location of each joint belongs to a Gaussian distribution with a mean equal to the projection of its 3D counterpart and a precision (i.e., the inverse variance) equal to ν.

The loss function in (4) is equivalent to the one proposed in previous work [58] summed up over frames. In contrast to previous work [58], we extend the model to the case of a perspective camera, treat 2D poses as latent variables instead of fixed input, and impose temporal smoothness constraints, as introduced in the following subsections.

2.2.2 Perspective camera model

When the camera intrinsic parameters are given, denoted by the calibration matrix $\mathbf{K}$, the perspective camera model is used to describe the dependence between 2D and 3D poses:

$$\boldsymbol{K}^{-1} \boldsymbol{W}_t \boldsymbol{Z}_t = \boldsymbol{R}_t \boldsymbol{S}_t + \boldsymbol{T}_t \boldsymbol{1}^\top, \tag{5}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{3\times p}$ denotes the homogeneous coordinates of 2D joints, $\boldsymbol{R}_t \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{T}_t \in \mathbb{R}^3$ are the rotation and translation in 3D, and $\boldsymbol{Z}_t \in \mathbb{R}^{p\times p}$ is a diagonal matrix with diagonal elements denoting the depth values of joints. Correspondingly, the loss function in the perspective case is defined as:

$$\mathcal{L}(\theta; \boldsymbol{W}) = \frac{\nu}{2} \sum_{t=1}^{n} \left\| \boldsymbol{K}^{-1} \boldsymbol{W}_t \boldsymbol{Z}_t - \boldsymbol{R}_t \sum_{i=1}^{k} c_{it} \boldsymbol{B}_i - \boldsymbol{T}_t \boldsymbol{1}^\top \right\|_F^2, \tag{6}$$

where $\theta = \{\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{T}, \boldsymbol{Z}\}$.

Note that minimizing the loss in (6) yields a trivial solution where all variables converge to zero due to the inherent scale ambiguity in the perspective model. To avoid such a trivial solution, the depth of the root joint is enforced to be one by adding the following constraint during optimization:

$$z_{1t} = 1, \tag{7}$$

where z_{1t} denotes the first diagonal element of $\boldsymbol{Z}_t$ corresponding to the root joint.

2.3 Dependence between pose and image

When 2D poses are given, we assume that the distribution of 3D pose parameters is conditionally independent of the image data. Therefore, the likelihood function of θ can be factorized as

$$\Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) = \Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta), \tag{8}$$

where $\boldsymbol{I} = \{\boldsymbol{I}_1, \cdots, \boldsymbol{I}_n\}$ denotes the image data. The conditional distribution $\Pr(\boldsymbol{W}|\theta)$ is given in (3). We assume that the likelihood function $\Pr(\boldsymbol{I}|\boldsymbol{W})$ given the image data can be learned discriminatively using a CNN and written as

$$\Pr(\boldsymbol{I}|\boldsymbol{W}) \propto \Pi_t \Pi_j h_j(\boldsymbol{w}_{jt}; \boldsymbol{I}_t), \tag{9}$$

where $\boldsymbol{w}_{jt}$ denotes the image location of joint j in frame t, and $h_j(\cdot; \boldsymbol{I}_t)$ represents a mapping from an image $\boldsymbol{I}_t$ to the likelihood of the joint location (termed heat map). For each joint j, the mapping h_j is approximated by a CNN (described in Section 4).

2.4 Prior on model parameters

The following penalty function on the model parameters is introduced:

$$\mathcal{R}(\theta) = \alpha \|\boldsymbol{C}\|_1 + \frac{\beta}{2} \|\nabla_t \boldsymbol{C}\|_F^2 + \frac{\gamma}{2} \|\nabla_t \boldsymbol{R}\|_F^2, \tag{10}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm (i.e., the sum of absolute values), ∇_t the discrete temporal derivative operator, and α, β, γ are scalar weights. The first term penalizes the cardinality of the pose coefficients to induce a sparse pose representation. The second and third terms impose first-order smoothness constraints on both the pose coefficients and rotations. A similar smoothness constraint could be imposed on the translation component; however, empirically we did not observe an obvious performance difference with its inclusion.

3 3D POSE INFERENCE

In this section, our approach to 3D pose inference is described. Here, two cases are distinguished: (i) the image locations of the joints are provided (Section 3.1), and (ii) the joint locations are unknown (Section 3.2).

3.1 Given 2D poses

When the 2D poses, $\boldsymbol{W}$, are given, the model parameters, θ, are recovered via penalized maximum likelihood estimation (MLE):

$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmax}} \ \ln \Pr(\boldsymbol{W}|\theta) - \mathcal{R}(\theta) \\ &= \underset{\theta}{\operatorname{argmin}} \ \mathcal{L}(\theta; \boldsymbol{W}) + \mathcal{R}(\theta). \end{aligned} \tag{11}$$

The problem in (11) is solved via block coordinate descent, i.e., alternately updating one block of variables while fixing the others.

3.1.1 Orthographic projection model

Under the orthographic model, the loss function is given in (4) and the update rules in each iteration are given below.

The update of $\boldsymbol{C}$ needs to solve:

$$\boldsymbol{C} \leftarrow \underset{\boldsymbol{C}}{\operatorname{argmin}} \ \ \mathcal{L}(\boldsymbol{C}; \boldsymbol{W}) + \alpha\|\boldsymbol{C}\|_1 + \frac{\beta}{2}\|\nabla_t \boldsymbol{C}\|_F^2, \tag{12}$$

where the objective is the composite of two differentiable functions plus an ℓ_1 penalty. The problem in (12) is solved by accelerated proximal gradient (APG) [69]. Since the problem in (12) is convex, global optimality is guaranteed.

The update of $\boldsymbol{R}$ needs to solve:

$$\boldsymbol{R} \leftarrow \underset{\boldsymbol{R}}{\operatorname{argmin}} \ \ \mathcal{L}(\boldsymbol{R}; \boldsymbol{W}) + \frac{\gamma}{2}\|\nabla_t \boldsymbol{R}\|_F^2, \tag{13}$$

where the objective is differentiable and the variables are rotations restricted to $SO(3)$. Here, manifold optimization is adopted to update the rotations using the trust-region solver in the Manopt toolbox [70].

The update of $\boldsymbol{T}$ has the following closed-form solution:

$$\boldsymbol{T}_t \leftarrow \text{row mean}\left\{\boldsymbol{W}_t - \boldsymbol{R}_t \boldsymbol{S}_t\right\}. \tag{14}$$

3.1.2 Perspective camera model

Under the perspective model, the loss function is given in (6) and the update rules in each iteration are given below.

The update rules for $\boldsymbol{C}$ and $\boldsymbol{R}$ have the same forms as the orthographic case given in (12) and (13), respectively.

The update of $\boldsymbol{T}$ has the following closed-form solution:

$$\boldsymbol{T}_t \leftarrow \text{row mean}\left\{\boldsymbol{K}^{-1}\boldsymbol{W}_t \boldsymbol{Z}_t - \boldsymbol{R}_t \boldsymbol{S}_t\right\}. \tag{15}$$

The update of $\boldsymbol{Z}$ is also closed-form. Suppose z_{it} denotes the i-th diagonal element of $\boldsymbol{Z}_t$, $\boldsymbol{u}_{it}$ the i-th column of $\boldsymbol{K}^{-1}\boldsymbol{W}_t$, and $\boldsymbol{v}_{it}$ the i-th column of $\boldsymbol{R}_t\boldsymbol{S}_t + \boldsymbol{T}_t\boldsymbol{1}^\top$. Then the solution for each z_{it} is

$$z_{it} \leftarrow \begin{cases} 1 & \text{if } i = 1, \\ \frac{\boldsymbol{u}_{it}^\top \boldsymbol{v}_{it}}{\boldsymbol{u}_{it}^\top \boldsymbol{u}_{it}} & \text{otherwise} \end{cases}. \tag{16}$$

The entire algorithm for 3D pose inference given the 2D poses is summarized in Algorithm 1. The iterations are terminated once the objective value has converged. Since in each step the objective function is non-increasing, the algorithm is guaranteed to terminate; however, since the problem in (11) is nonconvex, the algorithm requires a suitably chosen initialization (described in Section 3.3).

Input: $\boldsymbol{W}$; // 2D joint locations
Output: θ ; // 3D pose parameters

initialize the parameters ; // Section 3.3
while *not converged* **do**
  update $\boldsymbol{C}$ by (12) with APG ;
  update $\boldsymbol{R}$ by (13) with Manopt ;
  update $\boldsymbol{T}$ by (14) if using orthographic model
    or by (15) if using perspective model ;
  update $\boldsymbol{Z}$ by (16) if using perspective model ;
end

Algorithm 1: Block coordinate descent to solve (11) under the orthographic or perspective projection camera models.

3.2 Unknown 2D poses

If the 2D poses are unknown, $\boldsymbol{W}$ is treated as a latent variable and is marginalized out during the estimation process. The marginalized likelihood function is

$$\Pr(\boldsymbol{I}|\theta) = \int \Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) d\boldsymbol{W}, \tag{17}$$

where $\Pr(\boldsymbol{I}, \boldsymbol{W}|\theta)$ is given in (8).

Direct marginalization of (17) is intractable. Instead, an EM algorithm is developed to compute the penalized MLE. In the expectation step, the expectation of the penalized log-likelihood is calculated with respect to the conditional distribution of $\boldsymbol{W}$ given the image data and the previous estimate of all the 3D pose parameters, θ':

$$\begin{aligned} Q(\theta|\theta') &= \int \left\{\ln \Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) - \mathcal{R}(\theta)\right\} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \left\{\ln \Pr(\boldsymbol{I}|\boldsymbol{W}) + \ln \Pr(\boldsymbol{W}|\theta) - \mathcal{R}(\theta)\right\} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \text{const} - \int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} - \mathcal{R}(\theta). \end{aligned} \tag{18}$$

It can be shown that (see Appendix for the derivation)

$$\int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} = \mathcal{L}(\theta; \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right]) + \text{const}, \tag{19}$$

where $\mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right]$ is the expectation of $\boldsymbol{W}$ given $\boldsymbol{I}$ and θ':

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right] &= \int \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') \ \boldsymbol{W} \ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta')}{M} \ \boldsymbol{W} \ d\boldsymbol{W}, \end{aligned} \tag{20}$$

and M is a scalar that normalizes the probability. Both $\Pr(\boldsymbol{I}|\boldsymbol{W})$ and $\Pr(\boldsymbol{W}|\theta')$ given in (9) and (3), respectively, are products of marginal probabilities of $\boldsymbol{w}_{jt}$. Therefore, the expectation of each $\boldsymbol{w}_{jt}$ can be computed separately. In particular, the expectation of each $\boldsymbol{w}_{jt}$ is efficiently approximated by sampling over the pixel grid.

In the maximization step, the following is computed:

$$\begin{aligned} \theta &\leftarrow \underset{\theta}{\operatorname{argmax}} \, Q(\theta|\theta') \\ &= \underset{\theta}{\operatorname{argmin}} \ \ \mathcal{L}(\theta; \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right]) + \mathcal{R}(\theta), \end{aligned} \tag{21}$$

which can be solved by Algorithm 1.

The entire EM algorithm is summarized in Algorithm 2 with the initialization scheme described next in Section 3.3.

Input: $h_j(\cdot; \boldsymbol{I}_t),\ \forall j, t$; // heat maps
Output: $\theta = \{\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{T}, \boldsymbol{Z}\}$; // pose parameters

initialize the parameters ; // Section 3.3
while *not converged* **do**
  $\theta' = \theta$;
  // Compute the expectation of $\boldsymbol{W}$
  $\mathbb{E}[\boldsymbol{W}|\boldsymbol{I}, \theta'] = \int \frac{1}{M} \Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta')\ \boldsymbol{W}\ d\boldsymbol{W}$;
  // Update θ by Algorithm 1
  $\theta = \operatorname{argmin}_\theta\ \mathcal{L}(\theta; \mathbb{E}[\boldsymbol{W}|\boldsymbol{I}, \theta']) + \mathcal{R}(\theta)$;
end

Algorithm 2: The EM algorithm for pose from video.

3.3 Initialization and dictionary learning

A convex relaxation approach [57], [58] is used to initialize the parameters. This convex formulation was initially proposed for pose recovery in a single image frame given 2D correspondences [57], a special case of (11). The approach was later extended to handle 2D correspondence outliers [58]. If the 2D poses are given, the model parameters are initialized for each frame separately with the convex relaxation [58]. Alternatively, if the 2D poses are unknown, for each joint, the image location with the maximum heat map value is used. Next, the robust estimation algorithm from [58] is applied to initialize the parameters.

A dictionary learning algorithm [58] is used to learn the pose dictionaries given training pose data. The dictionary size is empirically set to $K = 64$ for action specific dictionaries and $K = 128$ for the non-action specific case, based on the trade-off between reconstruction error and computational efficiency.

4 CNN-based Joint Uncertainty Regression

A CNN is used to learn the mapping $\boldsymbol{I}_t \mapsto h_j(\cdot; \boldsymbol{I}_t)$, where $\boldsymbol{I}_t$ denotes an input image and $h_j(\cdot; \boldsymbol{I}_t)$ a (spatial) heat map for joint j. Rather than training p separate networks for p joints, a fully convolutional neural network [71] is trained to regress p joint distributions simultaneously by taking into account the full-body information. The training labels to be regressed are multi-channel heat maps with each channel corresponding to the image location uncertainty distribution for a joint. The uncertainty is modeled by a Gaussian centered at the annotated joint location. Figure 2 illustrates the CNN-based 2D joint regressor.

The Stacked Hourglass model proposed by Newell et al. [26] is adopted as the network architecture, which represents the state-of-the-art for 2D human pose detection. The network is fully convolutional and the shape of the network is an hourglass structure consisting of a series of downsampling layers with decreasing resolutions followed by a series of upsampling layers. This implements bottom-up and top-down processing to integrate contextual information over the entire image. A second hourglass component is stacked at the end of the first one to refine the initial heat maps. The final outputs are 64×64 heat maps. The ℓ_2 loss is minimized during training and intermediate supervision is applied at the end of the first module. The convolutional layers are implemented with residual modules. Please refer to the original paper [26] for details.

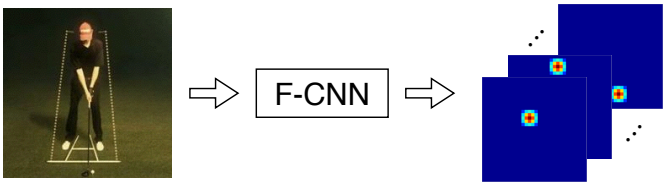


Fig. 2. CNN-based 2D joint regressor. The network is a fully convolutional neural network (F-CNN). The input is an image and the output is a multi-channel heat map with each channel capturing the spatial uncertainty distribution of a joint.

TABLE 2
Mean reconstruction errors (mm) on Human3.6M [37] given 2D joints.

	[5]	[48]	[58]	[32]	Orth.	Persp.
Protocol I	90.9	82.9	51.3	-	51.1	50.5
Protocol II	-	-	-	70.5	55.1	54.6

During testing, consistent with previous 3D pose methods (e.g., [40], [41]), a bounding box around the subject is assumed. The image patch in the bounding box, $\boldsymbol{I}_t$, is cropped in frame t and is provided to the network as input to predict the heat maps, $h_j(\cdot; \boldsymbol{I}_t),\ \forall j = 1, \ldots, n$.

5 Empirical Evaluation

5.1 Datasets

Empirical evaluation was performed on four standard datasets – Human3.6M [37], Human Eva I [65], KTH Football II [66], and MPII [62]. These datasets cover both controlled lab and more realistic scenarios. The first three datasets were used for quantitative evaluation and the last one for qualitative evaluation.

5.2 Evaluation metric

Given a set of estimated 3D joint locations $\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_n$ and the corresponding ground-truth locations $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_n^*$ in the same coordinates, the **per joint error** is defined as the average Euclidean distance over all the joints:

$$e = \frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i^*\|_2. \tag{22}$$

Note that the above metric depends on the absolute pose of the estimated structure, including scale, translation, and orientation. Scale and depth ambiguities are inherent to monocular reconstruction and in general cannot be resolved. The scale is directly learned from training subjects in prediction-based approaches [37], [40], [41]. For a fair comparison, our reconstruction output is scaled such that the mean limb length is identical to the average value of all training subjects. As the standard protocol in the Human3.6M and HumanEva datasets, the root locations of compared skeletons are aligned to make the evaluation translation invariant. Note that Procrustes alignment to the ground truth is not allowed.

TABLE 1
Mean per joint errors (mm) on Human3.6M [37] given 2D joints.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
PMP [5]	127.6	134.8	127.6	144.7	129.7	138.4	137.0	146.2
NRSFM [48]	137.2	140.7	129.8	118.0	134.3	120.0	156.5	161.7
Convex [58]	94.8	89.7	83.3	99.1	85.2	109.2	95.3	84.9
Orth.	92.8	88.2	82.3	97.4	83.3	107.1	93.2	83.7
Persp.	73.5	68.0	81.5	77.5	71.0	90.8	72.9	77.3
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
PMP [5]	149.8	161.5	132.0	156.7	120.2	159.3	159.8	141.4
NRSFM [48]	180.6	177.2	127.2	137.2	136.8	104.7	113.6	136.9
Convex [58]	80.3	97.2	82.2	94.4	83.4	81.4	92.3	90.2
Orth.	79.0	96.8	80.9	92.4	81.6	81.5	91.3	88.7
Persp.	68.3	91.4	66.8	75.0	67.0	60.1	71.5	74.2

The **reconstruction error** is defined as the 3D per joint error up to a similarity transformation, $\mathcal{T}$:

$$r = \min_{\mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{x}}_i - \mathcal{T}(\boldsymbol{x}_i^*)\|_2,$$

where the optimal parameters are obtained by Procrustes alignment. The 3D reconstruction error is widely used in structure-from-motion to evaluate the accuracy of recovered structure regardless of scale and rigid pose.

The **percentage of correct parts (PCP)** is defined as

$$\text{PCP} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left(\frac{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\| + \|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|}{2\|\boldsymbol{x}_i - \boldsymbol{y}_i\|} \leq \tau \right), \qquad (23)$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the coordinates of two ends of the i-th part and $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{y}}_i$ the corresponding estimates. $\mathbb{I}$ and τ denote the indicator function and the threshold, respectively. The PCP metric measures the fraction of correctly located parts with respect to a given threshold ($\tau = 0.5$ in this work).

5.3 Human3.6M

Human3.6M [37] is a recent large-scale dataset for 3D human sensing. It includes millions of 3D human poses acquired from a MoCap system with corresponding images from calibrated cameras. This setup provides synchronized videos and 2D-3D pose data for evaluation. It includes 11 subjects performing 15 actions, such as eating, sitting, and walking. The same data partition protocol as in previous work was used [40], [41]: the data from five subjects (S1, S5, S6, S7, S8) was used for training, and the data from two subjects (S9, S11) was used for testing. The original frame rate is 50 fps and is downsampled to 10 fps.

5.3.1 3D pose reconstruction with known 2D pose

First, the evaluation of the 3D reconstructability of the proposed approach with known 2D poses is presented. The generic approach to 3D reconstruction from 2D correspondences across a sequence is NRSFM. The proposed approach is compared to the state-of-the-art in NRSFM [48] on Human3.6M. A recent approach for single-view pose reconstruction, Projected Matching Pursuit (PMP) [5], and the initialization approach [58] used in our pipeline are also included in the comparison.

All sequences of S9 and S11 in Human3.6M were used for evaluation. A single pose dictionary from all the training pose data, irrespective of the action type, was used, i.e., a non-action specific dictionary. Dai et al.'s approach [48] requires a predefined rank K. Various values of K were considered with the best result for each sequence reported.

The per joint errors for each action are presented in Table 1, while the reconstruction errors are summarized in Table 2. In Table 2, Protocol I means the protocol introduced above in this paper, and Protocol II is the one proposed in [32] where only S11 and 14 joints are used in evaluation. The proposed approach with a perspective model ("Persp.") outperforms its orthographic counterpart ("Orth.") by a large margin in terms of per joint error. This performance difference is mostly due to the inaccuracy of rigid pose estimation with the orthographic model, which greatly increases the per joint error as Procrustes alignment is not allowed in the evaluation. Moreover, the orthographic model may suffer from a reflection ambiguity. In terms of reconstruction error which ignores rigid pose, the difference between the perspective and orthographic models is much smaller.

"Orth." performs slightly better than the initialization approach [58]. Note, given the 2D joints, "Orth." reduces to the initialization approach combined with temporal smoothness constraints.

The proposed approach also outperforms the NRSFM baseline [48]. The main reason is that the videos are captured by stationary cameras. Although the subject is occasionally rotating, the "baseline" between frames is generally small, and neighboring views provide insufficient geometric constraints for 3D reconstruction. In other words, NRSFM is very difficult to solve in the case of slow camera motion. This observation is consistent with prior findings in the NRSFM literature, e.g., [47]. In this case, the structure prior (even a non-action specific one) learned from existing Mo-

TABLE 3
Mean per joint errors (mm) on Human3.6M [37]. Marker † indicates using sequence information.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
LinKDE [37]	132.7	183.5	132.3	164.3	162.1	205.9	150.6	171.3
Li et al. [40]	-	136.8	96.9	124.7	-	168.6	-	-
Tekin et al. [41]†	102.4	147.7	88.8	125.2	118.0	182.7	112.3	129.1
Du et al. [42]†	85.0	112.6	104.9	122.0	139.0	135.9	105.9	166.1
Park et al. [43]	100.3	116.1	89.9	116.4	115.3	149.5	117.5	106.9
Zhou et al. [44]	91.8	102.4	96.6	98.7	113.3	125.2	90.0	93.8
Generic+nonspecific†	82.8	88.2	93.3	93.0	111.7	115.9	85.4	131.4
Generic+specific†	71.8	84.6	85.4	84.4	106.6	120.4	81.7	128.6
Fine-tuned+nonspecific†	77.8	74.0	85.0	83.7	80.1	96.4	79.2	85.3
Fine-tuned+specific†	69.2	75.2	75.8	73.6	75.4	99.6	76.1	73.6
Fine-tuned+specific w/o smooth	71.4	77.0	75.7	77.2	76.6	102.3	79.3	75.0

	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
LinKDE [37]	151.5	243.0	162.1	170.6	177.1	96.6	127.8	162.1
Li et al. [40]	-	-	-	-	132.1	69.9	-	-
Tekin et al. [41]†	138.8	224.9	118.4	138.7	126.2	55.0	65.7	124.9
Du et al. [42]†	117.4	226.9	120.0	117.6	137.3	99.2	106.5	126.4
Park et al. [43]	137.2	190.8	105.7	125.1	131.9	62.6	96.1	117.3
Zhou et al. [44]	132.1	158.9	106.9	94.4	126.0	79.0	98.9	107.2
Generic+nonspecific†	126.8	226.8	97.6	91.7	99.7	83.5	88.4	107.8
Generic+specific†	114.9	225.1	93.3	99.0	95.6	65.0	74.9	102.2
Fine-tuned+nonspecific†	79.1	118.4	75.3	80.3	75.3	67.8	81.8	82.7
Fine-tuned+specific†	75.0	109.6	73.7	88.9	71.8	55.6	73.5	77.8
Fine-tuned+specific w/o smooth	76.0	112.2	74.2	91.3	73.1	57.8	74.1	79.6

TABLE 4
Mean reconstruction errors (mm) on Human3.6M [37].

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
SMPLify [64]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3
Generic+nonspecific	52.0	54.0	59.1	61.7	74.2	70.7	51.5	60.3
Fine-tuned+nonspecific	46.7	47.7	54.9	54.1	56.3	65.4	46.9	49.1

	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
SMPLify [64]	100.3	137.3	83.4	77.3	79.7	86.8	81.7	82.3
Generic+nonspecific	83.9	119.9	66.9	54.8	64.5	55.6	59.1	65.9
Fine-tuned+nonspecific	60.1	81.5	53.2	49.7	54.2	47.1	53.7	54.7

TABLE 5
Mean per joint errors (mm) on the Human3.6M online test set ("H36M_NOS10") [37].

	Directions	Discussion	Eating	Greeting	Phoning	Posing	Purchases	Sitting
Ionescu et al. [37]	117	108	91	129	104	130	134	135
Li et al. [40]	-	92	76	98	92	107	141	136
Grinciunaite et al. [72]	91	89	94	102	105	99	112	151
Proposed	78	74	86	80	89	93	76	95

	SittingDown	Smoking	Photo	Waiting	Walking	WalkDog	WalkTogether	Mean
Ionescu et al. [37]	200	117	195	132	115	162	156	133
Li et al. [40]	265	97	171	105	99	139	110	122
Grinciunaite et al. [72]	239	109	151	106	101	141	106	119
Proposed	114	83	101	91	64	87	71	86

Cap data is critical for reconstruction.

5.3.2 3D pose reconstruction with unknown 2D pose

Next, results on the Human3.6M dataset are reported when 2D poses are predicted from images using the CNN described in Section 4. The proposed approach is compared to several recent baselines including both discriminative approaches (e.g., [37], [40], [41]) and two-stage reconstruction approaches (e.g., [64]). For the proposed approach, four combinations of training data sources were considered – the generic hourglass model trained on MPII ("generic") or the fine-tuned model trained on Human3.6M ("fine-tuned") combined with the nonspecific dictionary learned with all 3D pose data ("nonspecific") or the dictionary learned with action specific pose data ("specific").

The mean per joint errors are summarized in Table 3. The table shows that the proposed approach outperforms the baselines on most of the actions, yielding a much lower average error compared to the baselines. The "walk" and "walk together" sequences involve very predictable and repetitive motions. This might favor the direct regression approach [41]. The reconstruction errors are provided in Table 4. The proposed approach with a generic 2D pose detector outperforms SMPLify [64], which is the state-of-the-art two-stage approach that reconstructs 3D poses by fitting a parametric body shape model to 2D joints detected by a CNN-based 2D pose detector.

Comparing the results using different sources of training data, a remarkable improvement was achieved by using the fine-tuned 2D pose detector and the action specific 3D pose dictionaries. Nevertheless, the proposed approach with the generic detector and non-action specific dictionary still attained very competitive performance compared to the state-of-the-art. Note that the proposed approach can also be applied to single frames without temporal smoothness, with corresponding results presented in the last row of Table 3.

5.3.3 Human3.6M online test set

Table 5 presents results on the Human3.6M online test set "H36M_NOS10". The test set includes 360 sequences from three subjects with hidden ground truth. As the standard protocol, all compared approaches used action specific training. We used the fine-tuned hourglass model and action specific dictionaries. The results show that the proposed approach achieves significant improvements for all actions compared to the previous approaches.

5.3.4 Ablative analysis

Table 6 shows the impact of each component in our approach. For our approach, the 2D pose estimates were obtained from the expectations of their posterior distributions given in (20). Note that the 2D errors are with respect to the normalized bounding box size 256×256. Two cases of 2D input were considered: heat maps from the generic detector (less accurate), and the fine-tuned detector (more accurate). In this evaluation, action specific pose dictionaries and the Stacked Hourglass for joint heat map prediction were used.

Table 6 suggests that, for the same model, both the 3D and 2D errors increase significantly if EM is not applied. This indicates the importance of marginalizing the 2D uncertainty. Removing the smoothness constraint also increases the error, while the difference is more apparent when the 2D input is more noisy (from the "generic" detector). The perspective model always outperforms the orthographic model with the gap in terms of per joint error being much larger than that of reconstruction error. This indicates that the performance gain using a perspective camera model is mainly due to the more accurate rigid pose estimation. Finally, with the same orthographic model, the proposed approach clearly improves the initial solution [58] by taking advantage of EM and smoothness.

An alternative to the proposed EM algorithm is the maximum a posteriori (MAP) estimation which minimizes $-\ln \Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) + \mathcal{R}(\theta)$ over θ and $\boldsymbol{W}$. We implemented the block coordinate descent to solve the MAP estimation. With the fine-tuned detector, the mean per joint error of MAP is 79.3 mm, which is worse than that of EM (77.8 mm).

TABLE 6
Estimation errors for the variants of the proposed approach. "PJ", "Rec" and "2D" correspond to per joint error (mm), reconstruction error (mm), and 2D error (pixel), respectively.

	Generic			Fine-tuned		
	PJ	Rec	2D	PJ	Rec	2D
Perspective	102.2	63.3	10.1	77.8	53.2	6.0
w/o smooth	106.0	65.0	10.2	79.6	54.1	6.0
w/o smooth/EM	108.1	67.2	11.3	81.4	55.3	6.5
Orthographic	118.4	64.4	10.1	92.5	55.6	6.0
w/o smooth	122.2	65.4	10.2	93.8	55.3	6.0
w/o smooth/EM [58]	127.3	67.2	11.3	100.5	58.2	6.5

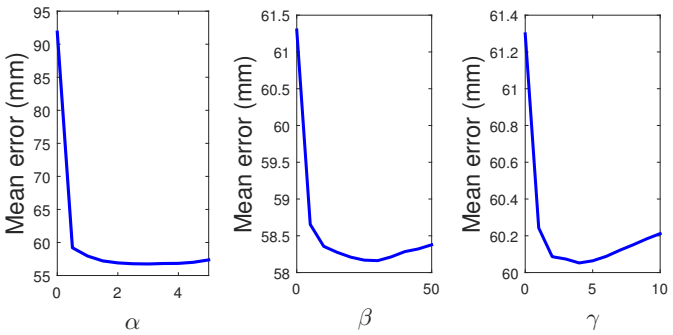


Fig. 3. Sensitivity to model parameters. The mean reconstruction error versus model parameter α, β or γ is shown, respectively.

5.3.5 Sensitivity to parameters

Figure 3 shows the mean reconstruction error as a function of each parameter in (10) while fixing the others, evaluated on a subset of test sequences (first five actions from S9). The first curve shows that the error initially decreases rapidly when α becomes nonzero. This indicates the importance of the sparsity constraint. After the initial rapid decrease, the error changes very smoothly, which indicates that the solution is not very sensitive to α when its value is in a proper range. Similar observations are made for the weights of the smoothness terms, β and γ. In practice, the model parameters were fixed without specific tuning in all other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2816031, IEEE Transactions on Pattern Analysis and Machine Intelligence

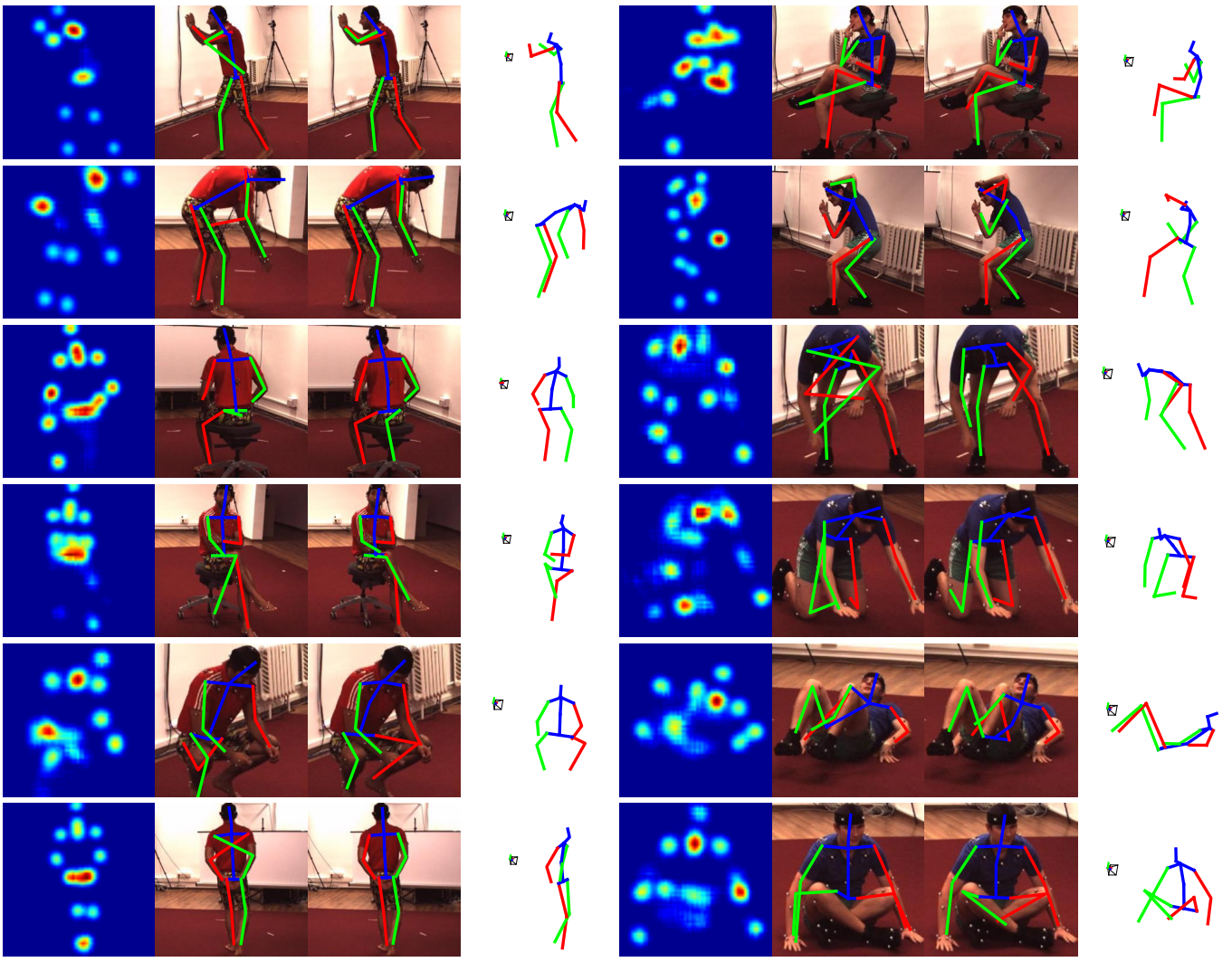


Fig. 4. Example comparative frame results on Human3.6M [37]. Each row includes two examples. The images from left-to-right correspond to the heat map (all joints shown simultaneously), the 2D pose found by greedily locating each joint separately according to the heat map, the estimated 2D pose by the proposed EM algorithm, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown. Notice that the errors in the 2D heat maps are corrected after considering the pose and temporal smoothness priors.

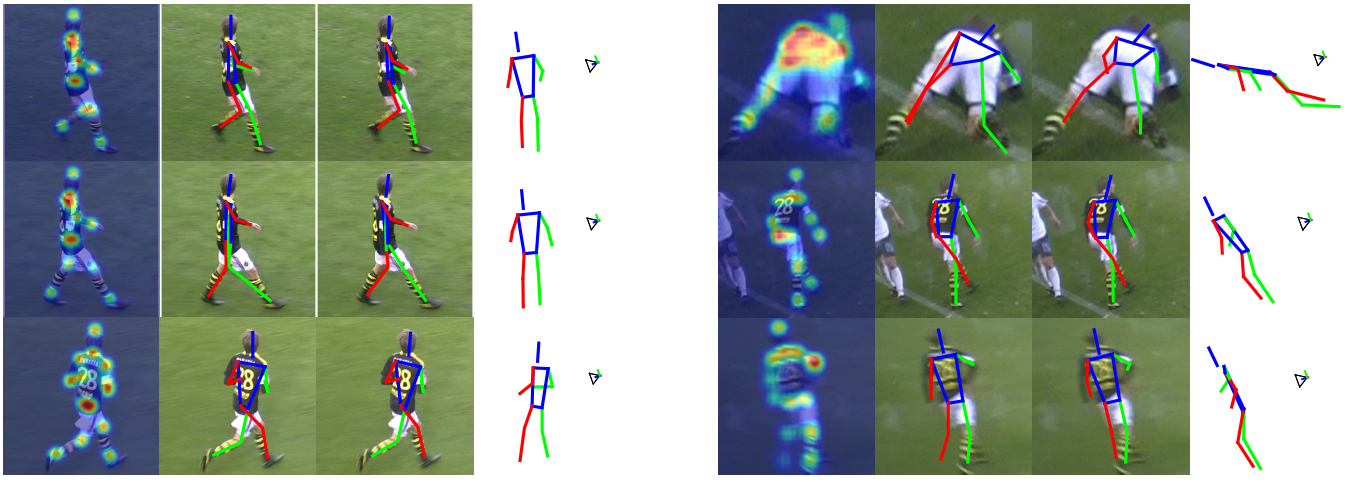


Fig. 5. Example comparative frame results on KTH Football II [73]. The images from left-to-right in each example correspond to the heat map (all joints shown simultaneously), the 2D pose found by greedily locating each joint separately according to the heat map response, the estimated 2D pose by the proposed EM algorithm, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

TABLE 7
Mean reconstruction errors (mm) on HumanEva I [65].

	Walking			Jogging			Average
	S1	S2	S3	S1	S2	S3	
Radwan et al. [53]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang et al. [60]	71.9	75.7	85.3	62.6	77.7	54.4	71.3
Simo-Serra et al. [61]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Bo et al. [34]	46.4	30.3	64.9	64.5	48.0	38.2	48.7
Kostrikov et al. [38]	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Yasin et al. [32]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Proposed	34.3	31.6	49.3	48.6	34.0	30.0	37.9

experiments presented in this paper ($\alpha = 0.5$, $\beta = 20$ and $\gamma = 2$ in a normalized 2D coordinate system).

5.3.6 Qualitative illustration

Figure 4 visualizes the results on several example frames. While the heat maps may be erroneous due to occlusion, left-right ambiguity, and other sources of uncertainty from the detectors, the proposed EM algorithm can effectively correct the errors by leveraging the pose prior, integrating temporal smoothness, and modeling the uncertainty.

5.4 HumanEva I

In this section, the evaluation results on HumanEva I [65] are presented. The evaluation protocol described elsewhere [61] was adopted. The walking and jogging sequences from camera C1 of all subjects were used for evaluation. The 2D joint detector trained on Human3.6M was fine-tuned with the training sequences for each action separately. Action specific pose dictionaries were learned for each subject separately. Each 3D pose reconstructed by the proposed approach was scaled to have the same average limb length as the training data.

The mean reconstruction errors for the evaluation sequences are reported in Table 7. The results of the compared baselines are taken from prior work [32]. Due to the large overlap between training and test data and less variability of poses, higher accuracies are generally obtained on this dataset compared to Human3.6M for all approaches. While none of the approaches dominate across all sequences, ours achieves the best overall accuracy.

5.5 KTH Football II

KTH Multiview Football II [66] contains images of professional footballers playing a match. It includes image sequences with 3D ground truth for 14 annotated joints captured from three calibrated views. The 3D ground truth was generated using multiview reconstruction with manual 2D annotations. Our evaluation was performed using the standard protocol [41], where the sequences of "Player 2" from "Camera 1" were used for testing. The generic hourglass model trained on MPII was used as the 2D detector without fine-tuning, while the pose dictionary was learned using the 3D poses associated with the training images provided in this dataset. Each 3D pose reconstructed by the proposed approach was scaled to have the same average limb length as the training poses and then aligned to the ground truth by a translation according to the root location.

TABLE 8
Mean PCP scores on KTH Football II [73].

	Sequence 1			Sequence 2
	[73]	[41]	Proposed	Proposed
Upper Arms	14	74	89	61
Lower Arms	06	49	78	49
Upper Legs	63	98	99	77
Lower Legs	41	77	85	56

To compare with the baselines, reported results are based on the percentage of correct pose (PCP) to measure part localization in 3D. Table 8 presents a summary of PCP results. It shows that the proposed approach achieves improved accuracy over the state-of-the-art. Results on selected frames are visualized in Figure 5.

5.6 MPII

Finally, the applicability of our proposed approach to in-the-wild imagery is qualitatively illustrated with the MPII Human Pose dataset [62]. MPII is a large-scale 2D human pose dataset that includes 25K single images extracted from YouTube videos containing over 40K people and 410 activities. This dataset does not include 3D pose data. The original hourglass model [26] trained on this dataset was used as the 2D detector and combined with the nonspecific action pose dictionary learned on Human3.6M to reconstruct the 3D human poses. The test images are from the validation set defined in previous work [26].

Figure 6 shows successful examples on MPII. Note that the input data consists of single images rather than sequences. While the pose dictionary is learned from another dataset, the proposed approach is able to produce visually reasonable 3D reconstructions from single images for a large variety of activities and viewpoints. Figure 7 presents some examples with larger uncertainties in the 2D heat maps, which result in incorrect 2D poses if the joints are simply determined by the heat map responses. After integrating

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2816031, IEEE Transactions on Pattern Analysis and Machine Intelligence

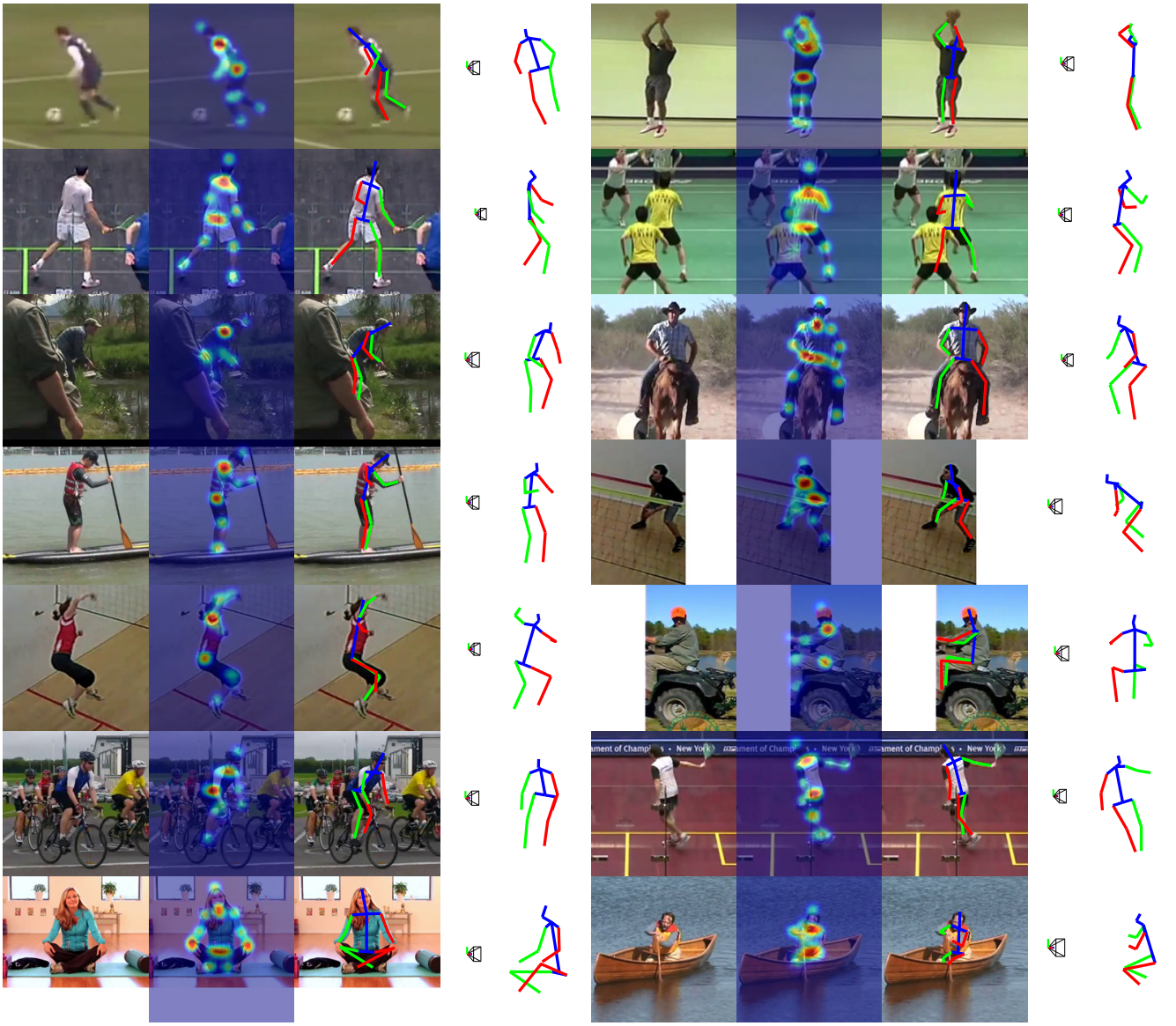


Fig. 6. Example successes on MPII [62]. In each example, the images from left-to-right correspond to the input image, the heat map (all joints shown simultaneously), the estimated 2D pose, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

the 3D pose prior by the proposed approach, better pose estimates are obtained. Figure 8 provides several failure examples. Visual inspection of these results suggests that the failures are mostly due to heavy occlusion, ambiguities from left-right symmetry, overlapping people, and extremely rare 3D poses that are beyond the representational capacity of the learned pose dictionary.

5.7 Running time

The experiments were performed on a desktop machine with an Intel i7 3.4G CPU, 8GB RAM and a GeForce GTX Titan X 6GB GPU. The running times for CNN-based heat map generation (with the hourglass model) and convex initialization were roughly 0.3s and 0.6s per frame, respectively; both steps can be easily parallelized. The EM algorithm usually converged in 20 iterations with a CPU time less than 100s for a sequence of 300 frames.

The running time of our approach depends on the dictionary size. There is a trade-off between accuracy and efficiency. For instance, for dictionary sizes of 32, 64, 96, and 128 tested on the first "Directions" sequence of S9, the mean reconstruction error (in mm) was 48.1, 46.0, 45.6 and 44.4, respectively, and the computation time (in seconds) was 91, 197, 317 and 488, respectively.

6 Summary

In summary, a 3D human pose estimation framework from a monocular image or video has been presented that consists of a novel synthesis between a deep learning-based 2D part regressor, a sparsity-driven 3D reconstruction approach, and a 3D temporal smoothness prior. This joint consideration combines the discriminative power of state-of-the-art 2D part detectors, the expressiveness of 3D pose models, and regularization by way of aggregating information over time. In practice, alternative part detectors, pose representations,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2816031, IEEE Transactions on Pattern Analysis and Machine Intelligence

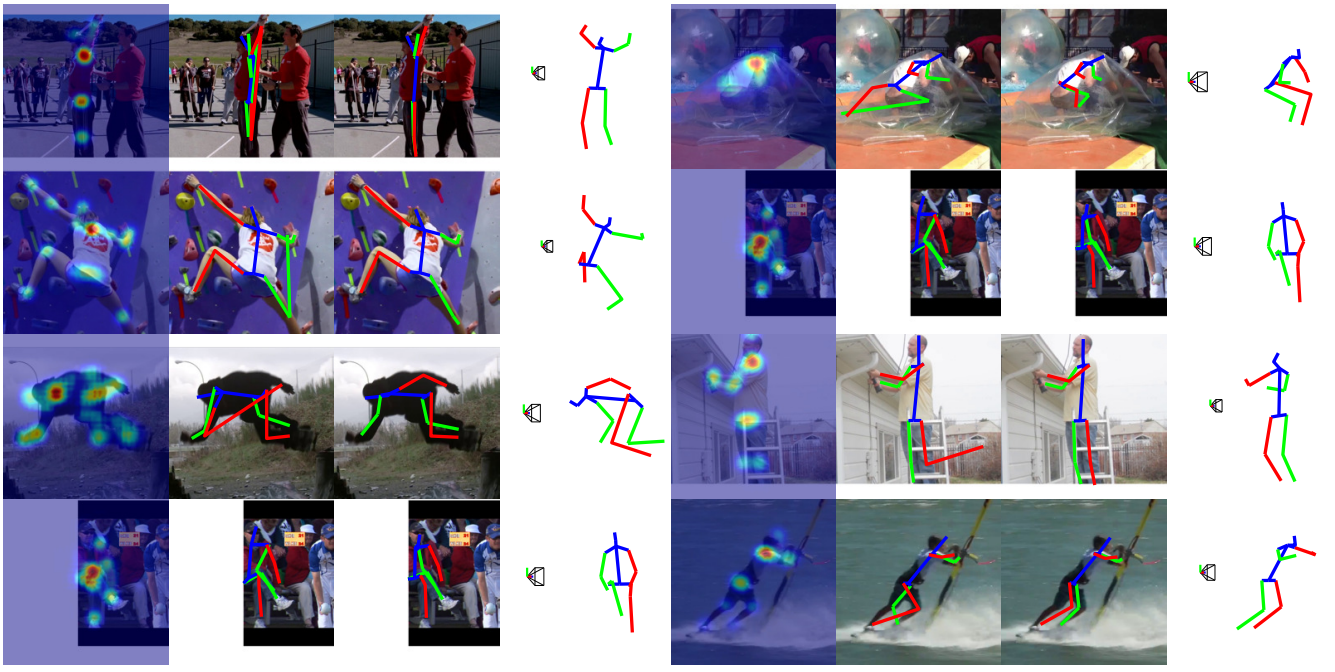


Fig. 7. Example comparative frame results on MPII [62]. In each example, the images from left-to-right correspond to the heat map (all joints shown simultaneously), the 2D pose found by greedily locating each joint separately according to the heat map, the estimated 2D pose by the proposed EM algorithm, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown. Notice that the errors in the 2D heat maps are corrected after considering the 3D pose prior.

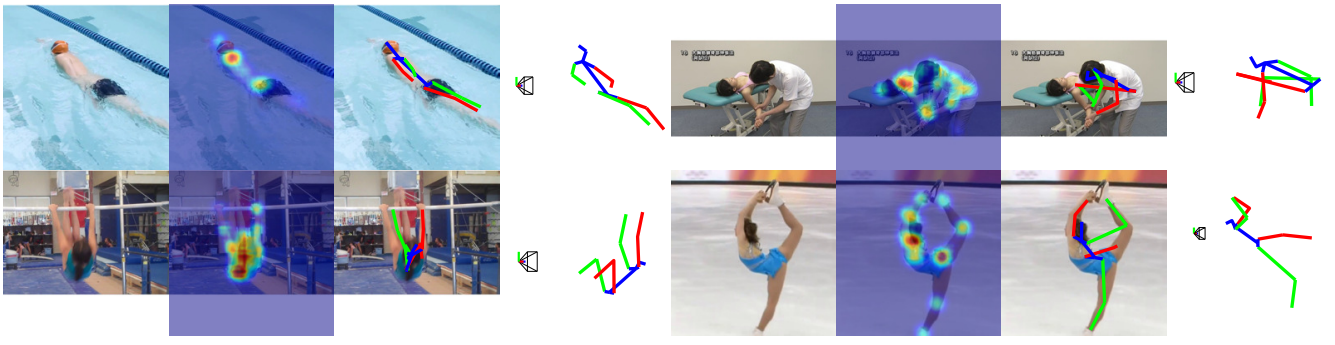


Fig. 8. Example failures on MPII [62]. In each example, the images from left-to-right correspond to the input image, the heat map (all joints shown simultaneously), the estimated 2D pose, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

and temporal models can be conveniently integrated in the proposed framework by replacing the original components. Experiments demonstrated that 3D geometric priors and temporal coherence can not only help 3D reconstruction but also improve 2D part localization.

APPENDIX

PROOF OF EQUATION (19)

For simplicity, $\mathcal{L}(\theta; \boldsymbol{W})$ is rewritten as

$$\begin{aligned}\mathcal{L}(\theta; \boldsymbol{W}) &= \sum_{t=1}^{n} \left\| \boldsymbol{W}_t - \boldsymbol{R}_t \sum_{i=1}^{k} c_{it} \boldsymbol{B}_i - \boldsymbol{T}_t \mathbf{1}^T \right\|_F^2 \\ &= \|\boldsymbol{W} - \boldsymbol{M}(\theta)\|_F^2 , \end{aligned} \tag{24}$$

where $\boldsymbol{W}$ is the stack of all $\boldsymbol{W}_t$ and $\boldsymbol{M}(\theta)$ is the stack of all $\boldsymbol{R}_t \sum_{i=1}^{k} c_{it} \boldsymbol{B}_i - \boldsymbol{T}_t \mathbf{1}^T$. Note that the constant $\frac{\nu}{2}$ is ignored for brevity.

Then, we have:

$$\begin{aligned}&\int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \|\boldsymbol{W} - \boldsymbol{M}(\theta)\|_F^2 \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \left\{ \|\boldsymbol{W}\|_F^2 - 2 \langle \boldsymbol{W}, \boldsymbol{M}(\theta) \rangle + \|\boldsymbol{M}(\theta)\|_F^2 \right\} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \left\{ \text{const} - \int 2 \langle \boldsymbol{W}, \boldsymbol{M}(\theta) \rangle \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} + \|\boldsymbol{M}(\theta)\|_F^2 \right\} \\ &= \left\{ \text{const} - 2 \left\langle \int \boldsymbol{W} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \ , \ \boldsymbol{M}(\theta) \right\rangle + \|\boldsymbol{M}(\theta)\|_F^2 \right\} \\ &= \left\| \int \boldsymbol{W} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} - \boldsymbol{M}(\theta) \right\|_F^2 + \text{const} \\ &= \|\mathrm{E}[\boldsymbol{W}|\boldsymbol{I}, \theta'] - \boldsymbol{M}(\theta)\|_F^2 + \text{const} \end{aligned} \tag{25}$$

DERIVATION OF EQUATION (20)

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{W}|\boldsymbol{I},\theta'\right] &= \int \Pr(\boldsymbol{W}|\boldsymbol{I},\theta')\ \boldsymbol{W}\ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{W},\boldsymbol{I}|\theta')}{\Pr(\boldsymbol{I}|\theta')}\ \boldsymbol{W}\ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{I}|\boldsymbol{W})\Pr(\boldsymbol{W}|\theta')}{M}\ \boldsymbol{W}\ d\boldsymbol{W}, \end{aligned} \tag{26}$$

where M is a constant.

REFERENCES

[1] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011. 1

[2] H. Lee and Z. Chen, "Determination of 3D human body postures from a single view," *CVGIP*, vol. 30, no. 2, pp. 148–168, 1985. 1, 2

[3] C. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," *CVIU*, vol. 80, no. 3, pp. 349–363, 2000. 1, 2

[4] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *CVPR*, 2000. 1, 2

[5] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *ECCV*, 2012. 1, 2, 3, 5, 6

[6] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *CVPR*, 1998. 1, 2

[7] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011. 1, 2

[8] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014. 1

[9] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *CVPR*, 2015. 1, 2

[10] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *CVPR*, 2014. 1, 2

[11] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *CVPR*, 2010. 1, 2

[12] F. Zhou and F. D. la Torre, "Spatio-temporal matching for human detection in video," in *ECCV*, 2014. 1, 2

[13] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, no. 2, pp. 90–126, 2006. 1

[14] C. Sminchisescu, "3D human motion analysis in monocular video techniques and challenges," in *AVSS*, 2007. 1

[15] M. A. Brubaker, L. Sigal, and D. J. Fleet, "Video-based people tracking," in *Handbook of Ambient Intelligence and Smart Environments*. Springer, 2010, pp. 57–87. 1

[16] D. Ramanan, "Part-based models for finding people and estimating their pose," in *Visual Analysis of Humans - Looking at People*. Springer, 2011, pp. 199–223. 1

[17] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation: A review of the literature and analysis of covariates," *CVIU*, vol. 152, pp. 1–20, 2016. 1

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

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

[20] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014. 2

[21] B. Sapp, D. J. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *CVPR*, 2011, pp. 1281–1288. 2

[22] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in *CVPR*, 2014, pp. 2361–2368. 2

[23] D. Park and D. Ramanan, "Articulated pose estimation with tiny synthetic videos," in *ChaLearn Workshop on Looking at People, CVPR*, 2015. 2

[24] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *ICCV*, 2015. 2

[25] D. Zhang and M. Shah, "Human pose estimation in videos," in *ICCV*, 2015. 2

[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016. 2, 5, 10

[27] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *CVPR*, 2003. 2

[28] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation," *IJCV*, vol. 98, no. 1, pp. 15–48, 2012. 2

[29] G. Shakhnarovich, P. A. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *ICCV*, 2003. 2

[30] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *PAMI*, vol. 28, no. 7, pp. 1052–1062, 2006. 2

[31] H. Jiang, "3D human pose reconstruction using millions of exemplars," in *ICPR*, 2010. 2

[32] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *CVPR*, 2016. 2, 5, 6, 10

[33] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *PAMI*, vol. 28, no. 1, pp. 44–58, 2006. 2

[34] L. Bo and C. Sminchisescu, "Twin Gaussian processes for structured prediction," *IJCV*, vol. 87, no. 1-2, pp. 28–52, 2010. 2, 10

[35] M. Salzmann and R. Urtasun, "Implicitly constrained Gaussian process regression for monocular non-rigid pose estimation," in *NIPS*, 2010. 2

[36] T. Yu, T. Kim, and R. Cipolla, "Unconstrained monocular 3D human pose estimation by action detection and cross-modality regression forest," in *CVPR*, 2013. 2

[37] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *PAMI*, vol. 36, no. 7, pp. 1325–1339, 2014. 2, 5, 6, 7, 8, 9

[38] I. Kostrikov and J. Gall, "Depth sweep regression forests for estimating 3d human pose from images." in *BMVC*, 2014. 2, 10

[39] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *ACCV*, 2014. 2

[40] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3D human pose estimation," in *ICCV*, 2015. 2, 5, 6, 7, 8

[41] B. Tekin, A. Rozantsev, , V. Lepetit, and P. Fua, "Direct prediction of 3D body poses from motion compensated sequences," in *CVPR*, 2016. 2, 5, 6, 7, 8, 10

[42] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng, "Marker-less 3D human motion capture with monocular image sequence and height-maps," in *ECCV*, 2016. 2, 7

[43] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," in *ECCVW*, 2016. 2, 7

[44] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *ECCVW*, 2016. 2, 7

[45] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen, "Synthesizing training images for boosting human 3D pose estimation," in *3DV*, 2016. 2

[46] G. Rogez and C. Schmid, "MoCap-guided data augmentation for 3D pose estimation in the wild," in *NIPS*, 2016. 2

[47] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *PAMI*, vol. 33, no. 7, pp. 1442–1456, 2011. 2, 6

[48] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *IJCV*, vol. 107, no. 2, pp. 101–122, 2014. 2, 5, 6

[49] Y. Zhu, D. Huang, F. De la Torre, and S. Lucey, "Complex non-rigid motion 3D reconstruction by union of subspaces," in *CVPR*, 2014. 2

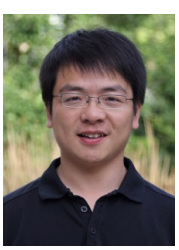
[50] J. Cho, M. Lee, and S. Oh, "Complex non-rigid 3D shape recovery using a Procrustean normal distribution mixture model," *IJCV*, pp. 1–21, 2015. 2

[51] J. Valmadre and S. Lucey, "Deterministic 3D human pose estimation using rigid structure," in *ECCV*, 2010. 2

[52] H. S. Park and Y. Sheikh, "3D reconstruction of a smooth articulated trajectory from a monocular image sequence," in *ICCV*, 2011, pp. 201–208. 2

[53] I. Radwan, A. Dhall, and R. Goecke, "Monocular image 3D human pose estimation under self-occlusion," in *ICCV*, 2013. 2, 10

[54] S. Leonardos, X. Zhou, and K. Daniilidis, "Articulated motion estimation from a monocular image sequence using spherical tangent bundles," in *ICRA*, 2016. 2
[55] X. Fan, K. Zheng, Y. Zhou, and S. Wang, "Pose locality constrained representation for 3D human pose reconstruction," in *ECCV*, 2014. 2
[56] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *CVPR*, 2015. 2, 3
[57] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3D shape estimation from 2D landmarks: A convex relaxation approach," in *CVPR*, 2015. 2, 5
[58] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *PAMI*, vol. 39, no. 8, pp. 1648–1661, 2017. 2, 3, 5, 6, 8
[59] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single Image 3D Human Pose Estimation from Noisy Observations," in *CVPR*, 2012. 2
[60] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," in *CVPR*, 2014. 2, 10
[61] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A Joint Model for 2D and 3D Pose Estimation from a Single Image," in *CVPR*, 2013. 2, 10
[62] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014. 2, 5, 10, 11, 12
[63] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *ICCV*, 2009. 2
[64] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *ECCV*, 2016. 2, 7, 8
[65] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *IJCV*, vol. 87, no. 1-2, pp. 4–27, 2010. 2, 5, 10
[66] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *BMVC*, 2013. 2, 5, 10
[67] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *CVPR*, 2016. 2
[68] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models–Their training and application," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995. 3
[69] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Tech. Rep., 2007. 4
[70] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *JMLR*, vol. 15, pp. 1455–1459, 2014. 4
[71] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 5
[72] A. Grinciunaite, A. Gudi, E. Tasli, and M. den Uyl, "Human pose estimation in space and time using 3D CNN," in *ECCVW*, 2016. 7
[73] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *CVPR*, 2013. 9, 10



Xiaowei Zhou is a Research Professor in the College of Computer Science, Zhejiang University. He was a Postdoctoral Researcher in Computer and Information Science, University of Pennsylvania. He obtained his Bachelor's degree in Optical Engineering from Zhejinag University, 2008, and his PhD degree in Electronic and Computer Engineering from The Hong Kong University of Science and Technology, 2013. His research interests are on 3D object recognition, human pose estimation, image matching, motion analysis and sparse/low-rank modeling.

Menglong Zhu is a Computer Vision Software Engineer at Google. He obtained a Bachelor's degree in Computer Science from Fudan University, in 2010, and a Master's degree in Robotics and a PhD degree in Computer and Information Science from University of Pennsylvania, in 2012 and 2016, respectively. His research interests are on object recognition, 3D object/human pose estimation, human action recognition, visual SLAM and text recognition.

Georgios Pavlakos is currently a doctoral student in Computer and Information Science, University of Pennsylvania. He received the BS degree in Electrical and Computer Engineering from the National Technical University of Athens, in 2014. His research interests lie at the intersection of computer vision and machine learning and include reconstruction and pose estimation of objects and humans from single images.

Spyridon Leonardos is currently a doctoral student in Computer and Information Science, University of Pennsylvania. He received the BS degree in Electrical and Computer Engineering (highest honors) from the National Technical University of Athens, in 2012 and the MS degree in Computer Science from University of Pennsylvania, in 2015. His research interests include multiple view geometry, reconstruction of articulated objects from video, Riemannian geometry for computer vision and sensor networks.

Konstantinos G. Derpanis is an Associate Professor of Computer Science, Ryerson University, Toronto. He received the Honours Bachelor of Science (BSc) degree in Computer Science from the University of Toronto, in 2000, and the MSc and PhD degrees in Computer Science from York University, Canada, in 2003 and 2010, respectively. Subsequently, he was a postdoctoral researcher in the GRASP Laboratory at the University of Pennsylvania. His main research field of interest is computer vision with emphasis on motion analysis and human motion understanding, and related aspects in image processing and machine learning.

Kostas Daniilidis is a Professor of Computer and Information Science, University of Pennsylvania. He obtained his MSE (Diploma) in Electrical Engineering from the National Technical University of Athens, 1986, and his PhD (Dr.rer.nat.) in Computer Science from the University of Karlsruhe, 1992. He is an IEEE Fellow and served as Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence from 2003 to 2007. His research interests are on visual motion and navigation, active perception, 3D object detection and localization, and panoramic vision.