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 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 the presented research 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 can hardly be applied to archival RGB images or videos. This paper addresses the challenging problem of not only using a single camera but also getting rid of the 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 prior [4], sparse representation [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 external entities and self-occlusions. Notable successes in pose estimation considered the challenge of 2D pose recovery using 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 the proposed approach. Given the input video (Fig. 1, top-left), 2D joint heat maps are generated with a deep 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., M.Z., G.P., S.L., 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: xiaowz@seas.upenn.edu

Manuscript received XXX; revised XXX.

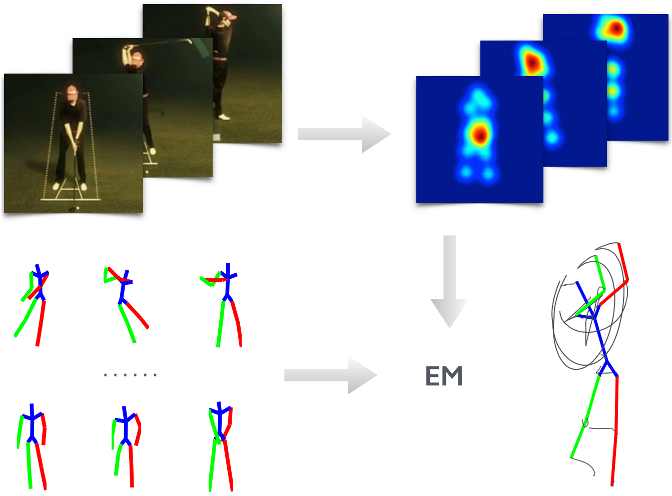

Fig. 1. Overview of the proposed approach. (top-left) Input image sequence, (top-right) CNN-based heat map outputs representing the soft localization of 2D joints, (bottom-left) 3D pose dictionary, and (bottom-right) the recovered 3D pose sequence reconstruction. To fully account for uncertainty, the 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 proposed to synthesize training images via graphics rendering [45] or image mosaicing [46].

Most closely related to the present paper 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 methods (e.g., [40], [41]), the proposed 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], [56]), 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 approaches are 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 landmarks are unknown, empirical results on the Human3.6M [37], HumanEva I [65] 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 3D pose prior learned from a separate MoCap dataset. A preliminary version of this work appeared in CVPR 2016 [67]. The code is available at http://cis.upenn.edu/~xiaowz/monocap.html.

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, which is denoted by $\boldsymbol{S}_t \in \mathbb{R}^{3\times p}$ for frame t. To reduce the ambiguity for 3D reconstruction, 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 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], [57]. 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, $\boldsymbol{c}_t$ denotes the coefficient vector $[c_{1t}, \cdots, c_{kt}]^\top$ for frame t and $\boldsymbol{C}$ denotes the matrix composed of all $\boldsymbol{c}_t$.

2.2 Dependence between 2D and 3D poses

The dependence between a 3D pose and its imaged 2D pose is modeled with a weak perspective camera model:

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

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. Note, the scale parameter in the weak perspective model is removed because the 3D structure, $\boldsymbol{S}_t$, can itself be scaled. 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 modelled 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.

2.3 Dependence between pose and image

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

$$\Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) = \Pr(\boldsymbol{I}|\boldsymbol{W})\Pr(\boldsymbol{W}|\theta), \tag{5}$$

where $\boldsymbol{I} = \{\boldsymbol{I}_1, \cdots, \boldsymbol{I}_n\}$ denotes the input images and $\Pr(\boldsymbol{W}|\theta)$ is given in (3). $\Pr(\boldsymbol{I}|\boldsymbol{W})$ is difficult to directly model, but it is proportional to $\Pr(\boldsymbol{W}|\boldsymbol{I})$ by assuming uniform priors on $\boldsymbol{W}$ and $\boldsymbol{I}$, and $\Pr(\boldsymbol{W}|\boldsymbol{I})$ can be learned from data.

Given the image data, the 2D distribution of each joint is assumed to be only dependent on the current image. Thus,

$$\Pr(\boldsymbol{I}|\boldsymbol{W}) \propto \Pr(\boldsymbol{W}|\boldsymbol{I}) = \Pi_t \Pi_j h_j(\boldsymbol{w}_{jt}; \boldsymbol{I}_t), \tag{6}$$

where $\boldsymbol{w}_{jt}$ denotes the image location of joint j in frame t, and $h_j(\cdot; \boldsymbol{Y})$ represents a mapping from an image $\boldsymbol{Y}$ to a probability distribution of the joint location (termed heat map). For each joint j, the mapping h_j is approximated by a CNN learned from training data. The details of the CNN learning step is described in Section 4.

2.4 Prior on model parameters

The following penalty function on the model parameters is introduced:

$$\mathcal{R}(\theta) = \alpha\|\boldsymbol{C}\|_1 + \frac{\beta}{2}\|\nabla_t \boldsymbol{C}\|_F^2 + \frac{\gamma}{2}\|\nabla_t \boldsymbol{R}\|_F^2, \tag{7}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm (i.e., the sum of absolute values), and ∇_t the discrete temporal derivative operator. 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 on both the pose coefficients and rotations.

3 3D Pose Inference

In this section, the proposed 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):

$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmax}}\ \ \ln \Pr(\boldsymbol{W}|\theta) - \mathcal{R}(\theta)\\ &= \underset{\theta}{\operatorname{argmin}}\ \ \mathcal{L}(\theta;\boldsymbol{W}) + \mathcal{R}(\theta).\end{aligned} \tag{8}$$

The problem in (8) is solved via block coordinate descent, i.e., alternately updating $\boldsymbol{C}$, $\boldsymbol{R}$ or $\boldsymbol{T}$ while fixing the others. The update of $\boldsymbol{C}$ needs to solve:

$$\boldsymbol{C} \leftarrow \underset{\boldsymbol{C}}{\operatorname{argmin}}\ \ \mathcal{L}(\boldsymbol{C};\boldsymbol{W}) + \alpha\|\boldsymbol{C}\|_1 + \frac{\beta}{2}\|\nabla_t\boldsymbol{C}\|_F^2, \tag{9}$$

where the objective is the composite of two differentiable functions plus an ℓ_1 penalty. The problem in (9) is solved by accelerated proximal gradient (APG) [69]. Since the problem in (9) is convex, global optimality is guaranteed. The update of $\boldsymbol{R}$ needs to solve:

$$\boldsymbol{R} \leftarrow \underset{\boldsymbol{R}}{\operatorname{argmin}}\ \ \mathcal{L}(\boldsymbol{R};\boldsymbol{W}) + \frac{\gamma}{2}\|\nabla_t\boldsymbol{R}\|_F^2, \tag{10}$$

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:

$$\boldsymbol{T}_t \leftarrow \text{row mean}\left\{\boldsymbol{W}_t - \boldsymbol{R}_t\sum_{i=1}^{k} c_{it}\boldsymbol{B}_i\right\}. \tag{11}$$

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 converge; however, since the problem in (8) is nonconvex, the algorithm requires a suitably chosen initialization (described in Section 3.3).

```
Input: $\boldsymbol{W}$ ;                       // 2D joint locations
Output: $\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{T}$ ;                  // pose parameters
initialize the parameters ;                // Section 3.3
while not converged do
    update $\boldsymbol{C}$ by (9) with APG;
    update $\boldsymbol{R}$ by (10) with Manopt;
    update $\boldsymbol{T}$ by (11);
end
```

Algorithm 1: Block coordinate descent to solve (8).

3.2 Unknown 2D poses

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

$$\Pr(\boldsymbol{I}|\theta) = \int \Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) d\boldsymbol{W}, \tag{12}$$

where $\Pr(\boldsymbol{I}, \boldsymbol{W}|\theta)$ is given in (5).

Direct marginalization of (12) is extremely difficult. 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':

$$\begin{aligned} Q(\theta|\theta') &= \int \{\ln \Pr(\boldsymbol{I}, \boldsymbol{W}|\theta) - \mathcal{R}(\theta)\} \; \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \{\ln \Pr(\boldsymbol{I}|\boldsymbol{W}) + \ln \Pr(\boldsymbol{W}|\theta) - \mathcal{R}(\theta)\} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \text{const} - \int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} - \mathcal{R}(\theta). \end{aligned} \tag{13}$$

It can be easily shown that

$$\int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} = \mathcal{L}(\theta; \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right]) + \text{const}, \tag{14}$$

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

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right] &= \int \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') \; \boldsymbol{W} \; d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta')}{Z} \; \boldsymbol{W} \; d\boldsymbol{W}, \end{aligned} \tag{15}$$

and Z is a scalar that normalizes the probability. The derivation of (14) and (15) is given in the appendix. Both $\Pr(\boldsymbol{I}|\boldsymbol{W})$ and $\Pr(\boldsymbol{W}|\theta')$ given in (6) 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:

$$\begin{aligned} \theta &\leftarrow \underset{\theta}{\operatorname{argmax}} \, Q(\theta|\theta') \\ &= \underset{\theta}{\operatorname{argmin}} \;\; \mathcal{L}(\theta; \mathbb{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right]) + \mathcal{R}(\theta), \end{aligned} \tag{16}$$

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

Algorithm 2: The EM algorithm for pose from video.

3.3 Initialization

A convex relaxation approach [57], [58] is used to initialize the parameters. A convex formulation was proposed to solve the single frame pose estimation problem given 2D correspondences [57], which is a special case of (8). 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 a convex method [57]. 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.

4 CNN-BASED JOINT UNCERTAINTY REGRESSION

A CNN is used to learn the mapping $\boldsymbol{Y} \mapsto h_j(\cdot; \boldsymbol{Y})$, where $\boldsymbol{Y}$ denotes an input image and $h_j(\cdot; \boldsymbol{Y})$ represents a heat map for joint j. Instead of learning p 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. Figure 2 gives an illustration. In this work, two CNN architectures are considered.

The first architecture is similar to the SpatialNet model proposed elsewhere [24] but without any spatial fusion or temporal pooling. The network consists of seven convolutional layers with 5×5 filters followed by ReLU layers and a last convolutional layer with $1 \times 1 \times p$ filters to provide dense prediction for all joints. A 2×2 max pooling layer is inserted after each of the first three convolutional layers. The network is trained by minimizing the l_2 loss between the prediction and the label with the open source Caffe framework [72]. Stochastic gradient descent (SGD) with momentum of 0.9 and a mini-batch size of 128 is used. During training, a rectangular patch is extracted around the subject from each image and is resized to 256×256 pixels. Random shifts are applied during cropping and RGB channel-wise random noise is added for data augmentation. Channel-wise RGB mean values are computed from the dataset and subtracted from the images for data normalization. The training labels to be regressed are multi-channel heat maps with each channel corresponding to the image location uncertainty distribution for each joint. The uncertainty is modeled by a Gaussian centered at the annotated joint location. The heat map resolution is reduced to 32×32 to decrease the CNN model size which allows a large batch size in training and prevents overfitting.

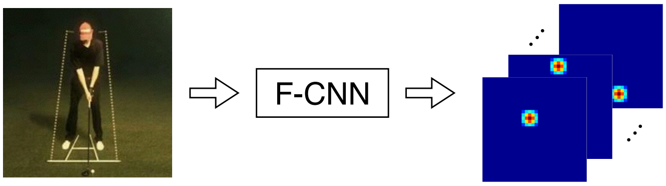

Fig. 2. Illustration of the CNN based 2D joint regressor. The network is a fully convolutional neural network (F-CNN). The input is an image and the output is a multi-channel heat map with each channel showing the spatial uncertainty distribution of a joint.

TABLE 2
Reconstruction accuracy given 2D poses on Human3.6M [37] with two input cases considered: the original 2D views from Human3.6M and synthesized views with artificial camera motion. The numbers are the mean reconstruction errors (mm).

	Original	Synthesized
PMP [5]	89.50	84.16
NRSFM [48]	72.98	48.88
Initial [58]	50.04	48.08
Optimized	**49.64**	**47.57**

The second architecture is the Stacked Hourglass model proposed by Newell et al. [26], which has achieved state-of-the-art performance for 2D human pose detection. Similar to the basic model described above, the hourglass model also consists of convolutional layers, but the main difference is that 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, which implements the bottom-up and top-down processing to integrate contextual information over the whole image. A second hourglass component is stacked at the end of the first one to refine the heat maps as a postprocessing step. 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.

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

5 Empirical Evaluation

5.1 Datasets and implementation details

Empirical evaluation was performed on four datasets – Human3.6M [37], Human Eva I [65], KTH Football II [66] and MPII [62], which cover both controlled lab and more realistic scenarios. The first three 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 joints:

$$e = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i^*\|_2. \tag{17}$$

Note that the above metric depends on the absolute pose of the estimated structure including scale, translation and orientation. The scale and depth ambiguities are inherent to monocular reconstruction and cannot be resolved in general. The scale is directly learned from training subjects in prediction-based methods [37], [40], [41]. For a fair comparison, the reconstruction by the proposed method 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.

The **reconstruction error** is defined as the 3D per joint error up to a similarity transformation:

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

where $\mathcal{T}$ denotes the transformation and the optimal parameters can be obtained by the Procrustes method. 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

$$\mathrm{PCP} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\frac{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\| + \|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|}{2\|\boldsymbol{x}_i - \boldsymbol{y}_i\|} \le \tau\right), \tag{18}$$

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.

5.3 Human3.6M

The Human3.6M dataset [37] is a recently published 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.

The algorithm in [58] was used to learn the pose dictionaries. The dictionary size was set to $K = 64$ for action-specific dictionaries and $K = 128$ for the nonspecific action case. For all experiments, the same set of parameters was used ($\alpha = 0.1$, $\beta = 5$, $\gamma = 0.5$, $\nu = 4$ in a normalized 2D coordinate system).

TABLE 1
Reconstruction accuracy given 2D poses on Human3.6M [37]. The numbers are the mean reconstruction errors (mm).

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
PMP [5]	68.56	77.53	95.79	86.49	73.94	95.21	78.31	97.83
NRSFM [48]	79.26	60.75	125.79	74.97	**37.14**	**58.74**	61.38	88.47
Initial [58]	38.25	45.00	47.23	48.51	45.95	63.77	47.02	43.88
Optimized	**37.88**	**43.02**	**46.51**	**48.42**	43.74	59.41	**44.76**	**42.64**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
PMP [5]	103.57	123.42	72.89	95.19	83.49	95.17	95.05	89.50
NRSFM [48]	112.74	114.68	45.09	78.00	62.79	63.20	**31.67**	72.98
Initial [58]	**53.52**	**70.12**	41.29	**48.40**	53.73	**47.37**	56.53	50.04
Optimized	55.30	73.34	**40.49**	48.82	**53.62**	50.42	56.16	**49.64**

TABLE 3
Mean **per joint errors** (mm) on Human3.6M [37]. Two variants of the proposed method were evaluated – (i) using the SpatialNet model or (ii) using the Hourglass model.

	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
Proposed+SpatialNet	87.3	109.3	87.0	103.1	116.1	143.3	106.8	99.7
Proposed+Hourglass	**68.7**	**74.8**	**67.8**	**76.4**	**76.3**	**98.4**	**84.0**	**70.2**
	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
Proposed+SpatialNet	124.5	199.2	107.4	118.0	114.2	79.3	97.7	113.0
Proposed+Hourglass	**88.0**	**113.8**	**78.0**	**90.1**	**75.1**	62.6	73.6	**79.9**

5.3.1 3D pose reconstruction with known 2D pose

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

The sequences of S9 and S11 from the first camera in the Human3.6M dataset were used for evaluation and frames beyond 30 seconds were truncated for each sequence. The 2D orthographic projections of the 3D poses provided in the dataset were used as the input. Performance was evaluated by the reconstruction errors, the standard protocol for evaluating NRSFM. To demonstrate the generality of the proposed approach, a single pose dictionary from all the training pose data, irrespective of the action type, was used, i.e., a non-action specific model. The method from Dai et al. [48] requires a predefined rank K. Here, various values of K were considered with the best result for each sequence reported.

The mean reconstruction errors for different actions are summarized in Table 1. The proposed method clearly outperforms the NRSFM baseline. The 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,

TABLE 4
Mean **reconstruction errors** (mm) on Human3.6M [37]. The hourglass model is used as the 2D joint detector. 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 specific dictionary learned with action-specific pose data ("specific").

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
SMPLify [64]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	79.7	86.8	81.7	82.3
Generic+nonspecific	52.5	53.6	58.5	62.0	73.3	70.3	52.1	58.2	82.9	118.6	65.7	54.9	64.5	55.6	59.1	65.5
Generic+specific	49.2	55.6	53.3	59.6	70.1	74.2	55.0	58.2	75.9	118.2	65.7	61.7	63.8	50.5	54.2	64.4
fine-tuned+nonspecific	47.9	**48.8**	52.7	55.0	56.8	**65.5**	**49.0**	**45.5**	**60.8**	81.1	**53.7**	**51.6**	**54.8**	50.4	55.9	**55.3**
fine-tuned+specific	**45.9**	52.0	**51.5**	**53.5**	**56.7**	69.2	51.7	45.8	61.7	**76.5**	56.9	57.9	55.4	**46.2**	**52.1**	55.6

TABLE 5
The mean reconstruction error (mm) for each joint on Human3.6M [37].

Head	Jaw	Thorax	Spine	Pelvis	Hip$_{left}$	Hip$_{right}$	Knee$_{left}$	Knee$_{right}$	Ankle$_{left}$	Ankle$_{right}$	Shoulder$_{left}$	Shoulder$_{right}$	Elbow$_{left}$	Elbow$_{right}$	Wrist$_{left}$	Wrist$_{right}$	Average
46.9	42.9	31.2	31.6	34.4	40.9	44.4	61.4	60.8	90.2	83.4	37.5	34.1	65.5	64.3	87.7	87.8	55.6

NRSFM is very difficult to compute with slow camera motion. This observation is consistent with prior findings in the NRSFM literature, e.g., [47]. To validate this issue, an artificial rotation was applied to the 3D poses by 15 degrees per second and the 2D joint locations were synthesized by projecting the rotated 3D poses into 2D. The corresponding results are presented in Table 2. In this case, the performance of NRSFM improved dramatically. Overall, the experiments demonstrate that the structure prior (even a non-action specific one) from existing pose data is critical for reconstruction. This is especially true for videos with small camera motion, which is common in real world applications. Comparing the initial results and optimized results, the temporal smoothness helps but the change is not significant since the single frame initialization is very stable with known 2D poses. Nevertheless, in the next section it is shown that the temporal smoothness is important when 2D poses are not given.

5.3.2 3D pose reconstruction with unknown 2D pose

Next, results on the Human3.6M dataset are reported when 2D poses are not given. The proposed method is compared to several recent baseline methods. The first baseline method is LinKDE which is provided with the Human3.6M dataset [37]. This baseline is based on single frame regression. The second one is a CNN-based method from Li et al. [40]. The third one is a recently proposed approach from Tekin et al. [41] which uses CNNs and explores motion information in a short sequence. The last one is from Bogo et al. [64] which fits a human body model to the 2D pose given by a state-of-the-art CNN-based detector. In this experiment, the sequences of S9 and S11 from all cameras were used for evaluation.

The standard evaluation protocol of the Human3.6M dataset was adopted, i.e., the mean per joint error (mm) in 3D is calculated between the reconstructed pose and the ground truth in the camera frame with their root locations aligned. In general, it is impossible to determine the scale of the object in monocular images. The baseline methods learn the scale from training subjects. For a fair comparison, the reconstructed pose by the proposed method was scaled such that the mean limb length of the reconstructed pose was identical to the average value of all training subjects. Note that the Procrustes alignment is not allowed here. It was empirically found that the 3D joint error was significantly affected by the bias of camera rotation estimates due to the adopted weak perspective camera model. To reduce the bias, the camera pose was further refined for each frame using a perspective-n-point (PnP) algorithm [73], given the camera intrinsic parameters provided by the dataset and the 3D/2D structures estimated by the EM algorithm.

The results are summarized in Table 3. The table shows that the proposed method outperforms the baselines on most of the actions except for "walk" and "walk together", which involve very predictable and repetitive motions and might favor the direct regression approach [41]. A remarkable improvement was achieved by using the hourglass

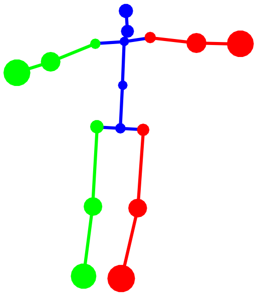

Fig. 3. Visualization of the mean estimation error for each joint on Human3.6M [37]. The radius of the circle around each joint represents the mean estimation error for that joint; see Table 5 for the actual values.

TABLE 6
The estimation errors before and after EM. The 3D errors are the mean reconstruction errors for all test data.

	Proposed+SpatialNet		Proposed+Hourglass	
	3D (mm)	2D (pixel)	3D (mm)	2D (pixel)
Before	79.8	14.9	58.2	6.5
After	**72.3**	**10.8**	**55.6**	**6.0**

model as the 2D pose detector which was pretrained on the MPII dataset and fine-tuned on the Human3.6M. The performance boost is attributed to both the better network design of the hourglass model and additional training data in the MPII dataset. This also demonstrates the flexibility of the proposed framework to leverage the state-of-the-art network design and the large-scale datasets with only 2D annotations, in contrast to the direct regression approaches that require synchronized image-MoCap pairs.

The 3D reconstruction errors are also provided in Table 4. Here the hourglass model was used as the 2D detector and the results with different sources of training data are reported. The reconstructions errors are clearly reduced by fine-tuning the hourglass model on Human3.6M with the 2D annotations provided in the dataset, while the difference between using an action-specific 3D pose dictionary and a nonspecific dictionary learned with all data is negligible.

5.3.3 Error of each joint

The mean reconstruction error for each separate joint averaged over all testing frames of Human3.6M is provided in Table 5 and visualized in Figure 3. As expected, the results show that the extremities of the four limbs are much more difficult to localize due to the high degree of freedom and possibility of being occluded.

5.3.4 The effect of EM

Table 6 shows the estimation errors before and after the EM optimization. Note that the 2D errors are with respect to the normalized bounding box size 256 × 256. The table shows that the convex initialization provides suitable initial estimates, which are further improved by the EM algorithm that integrates joint detection uncertainty and temporal smoothness. As for 2D errors, while the hourglass model is extremely powerful achieving a very small error, the EM algorithm is still able to further improve it by integrating the 3D and temporal information.

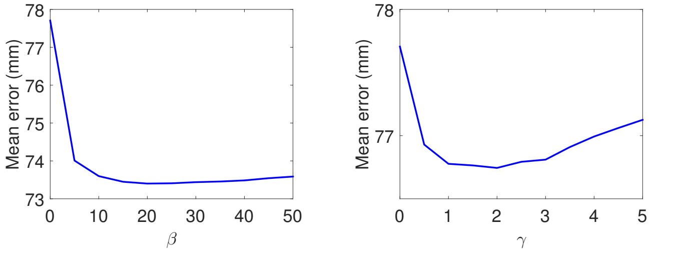

Fig. 4. The effect of smoothness on pose estimates. The mean 3D reconstruction error versus the model parameters β and γ is shown, which control the temporal smoothness of coefficients and camera rotation, respectively.

5.3.5 The effect of smoothness

To illustrate the effect of the smoothness terms, Figure 4 shows the mean reconstruction error of as a function of the weights of the smoothness terms in (7). In the left plot β is varying and $\gamma = 0$, and vise versa in the right plot. The experiment was performed on all sequences of S9. The curves show that the error decreases quickly when the parameters become nonzero indicating the importance of smoothness. After that, the error changes very smoothly in a certain range, which means that the solutions are insensitive to the parameters in a proper range. In all our experiments, the model parameters were fixed without specific tuning.

5.3.6 Qualitative illustration

Figure 5 visualizes the results of some example frames, where the heat maps were produced by the SpatialNet model. While the heat maps may be erroneous due to occlusion, left-right ambiguity, and other uncertainty from the detectors, the proposed EM algorithm can largely 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 the HumanEva I dataset [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 SpatialNet model trained on the Human3.6M dataset 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 method was scaled to have the same average limb length as the training data and then aligned to the ground truth with the Procrustes method allowing a rigid transformation.

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

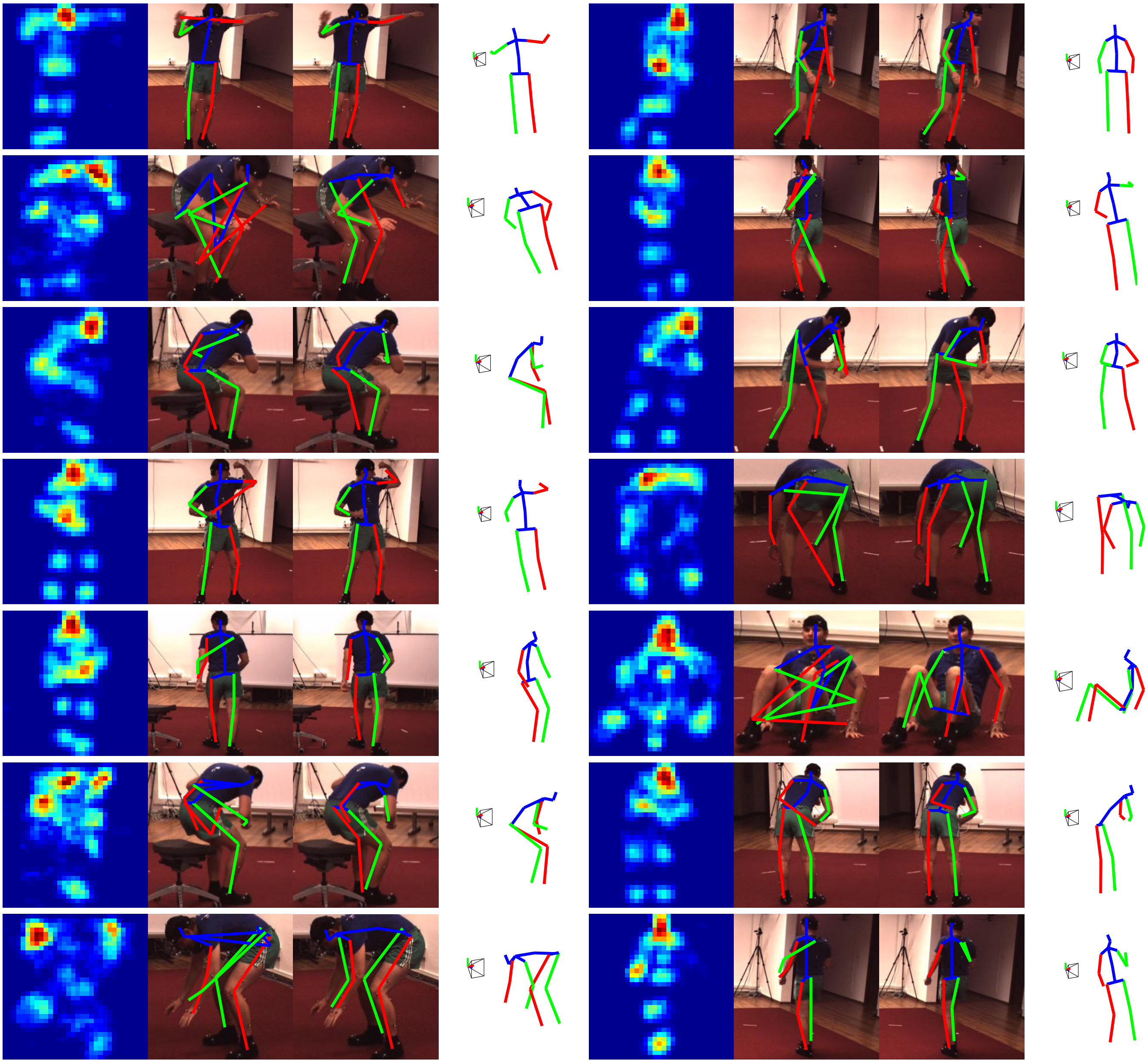

Fig. 5. Example frame results on Human3.6M [37], where the errors in the 2D heat maps are corrected after considering the pose and temporal smoothness priors. Each row includes two examples from two actions. The images from left-to-right correspond to the heat map (all joints combined), the 2D pose found 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.

TABLE 7
Quantitative results on HumanEva I [65]. The table presents the mean reconstruction errors in millimeters.

	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**

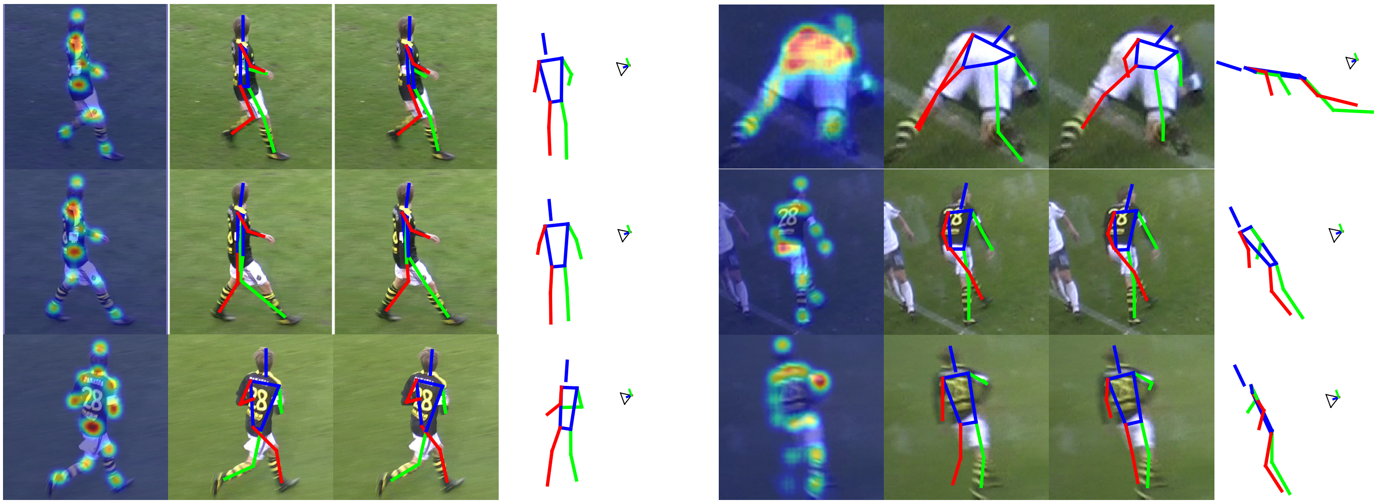

Fig. 6. Example frame results on KTH Football II [74]. The images from left-to-right in each example correspond to the heat map (all joints combined), the 2D pose found by greedily locating each joint separately according to the heat map response, the estimated 2D pose by the proposed EM algorithm, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

TABLE 8
Quantitative results on KTH Football II [74]. The table presents the mean PCP scores (the higher the better). The sequences of Player 2 from Camera 1 are used.

	Sequence 1			Sequence 2
	[74]	[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

5.5 KTH Football II

The KTH Multiview Football II [66] dataset 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 from the multiview reconstruction with manual 2D annotations. Evaluation was performed using the standard protocol [41], where "Player 2" was 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 method 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.

To compare with the baseline methods, 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 method achieves improved accuracy over the state-of-the-art. The results of selected frames are visualized in Figure 6.

5.6 MPII

Finally, the applicability of the proposed method to Internet images is qualitatively illustrated with the MPII dataset [62]. The MPII human pose dataset is a large-scale 2D human pose dataset that includes 25K single images extracted from YouTube videos containing over 40K people and 410 activities. No 3D pose data is available in the dataset. The original hourglass model [26] trained on this dataset was used as the 2D detector and combined with the nonspecific 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 7 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 method is able to produce visually reasonable 3D reconstructions from single images for a large variety of activities and viewpoints. Figure 8 presents some examples with larger uncertainties in the 2D heat maps, which result in incorrect 2D poses if the joints are located simply by heat map responses. After integrating the 3D pose prior by the proposed method, better pose estimates are obtained. Figure 9 provides several failed examples. Empirical observations suggested that the failures were mostly due to heavy occlusion, ambiguities from left-right symmetry, multiple overlapping persons, and extremely rare 3D poses that could hardly be represented by the pose dictionary.

5.7 Running time

The experiments were performed on a desktop with an Intel i7 3.4G CPU, 8G 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.

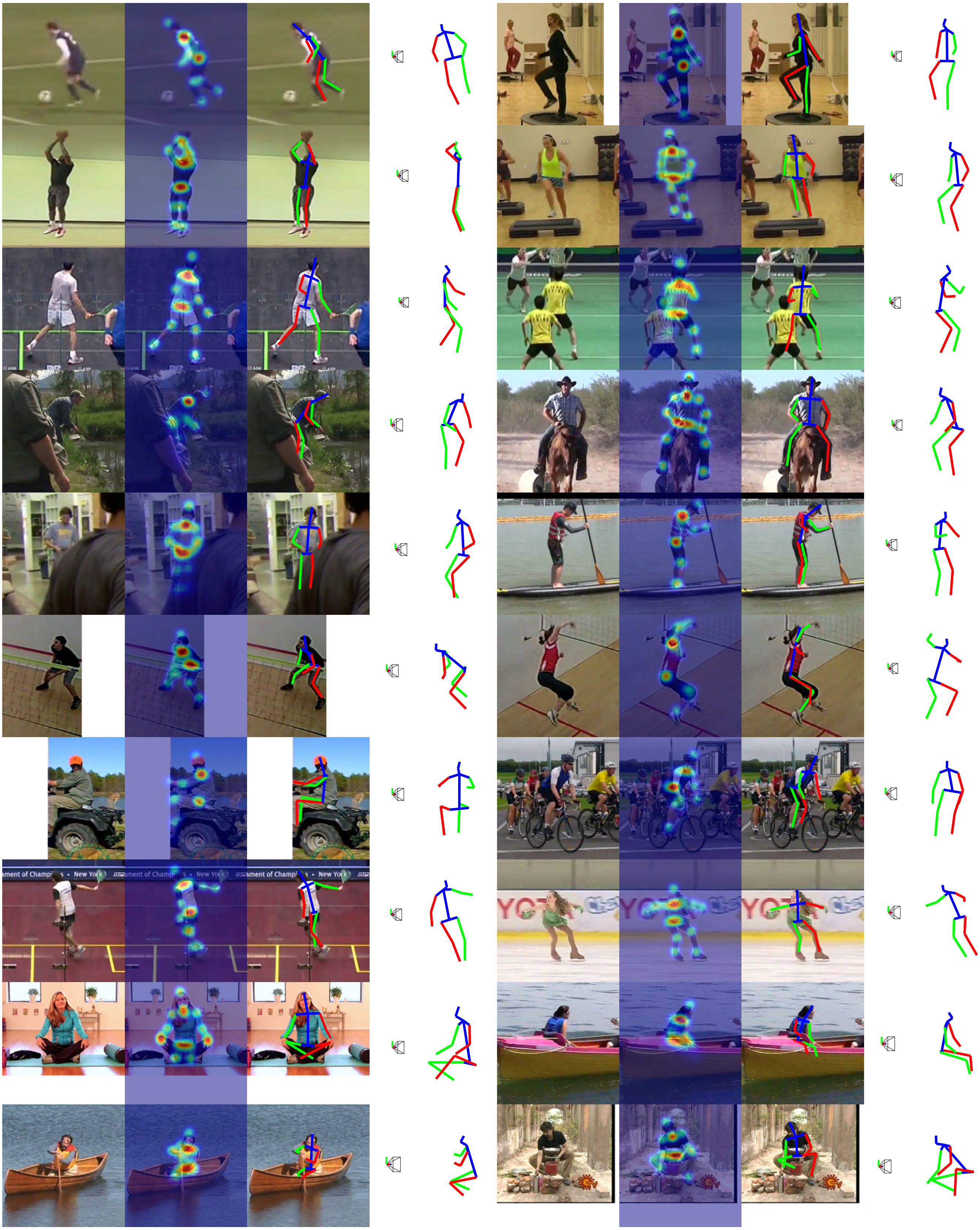

Fig. 7. Successful examples on MPII [62]. In each example, the images from left-to-right correspond to the input image, the heat map (all joints combined), the estimated 2D pose, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

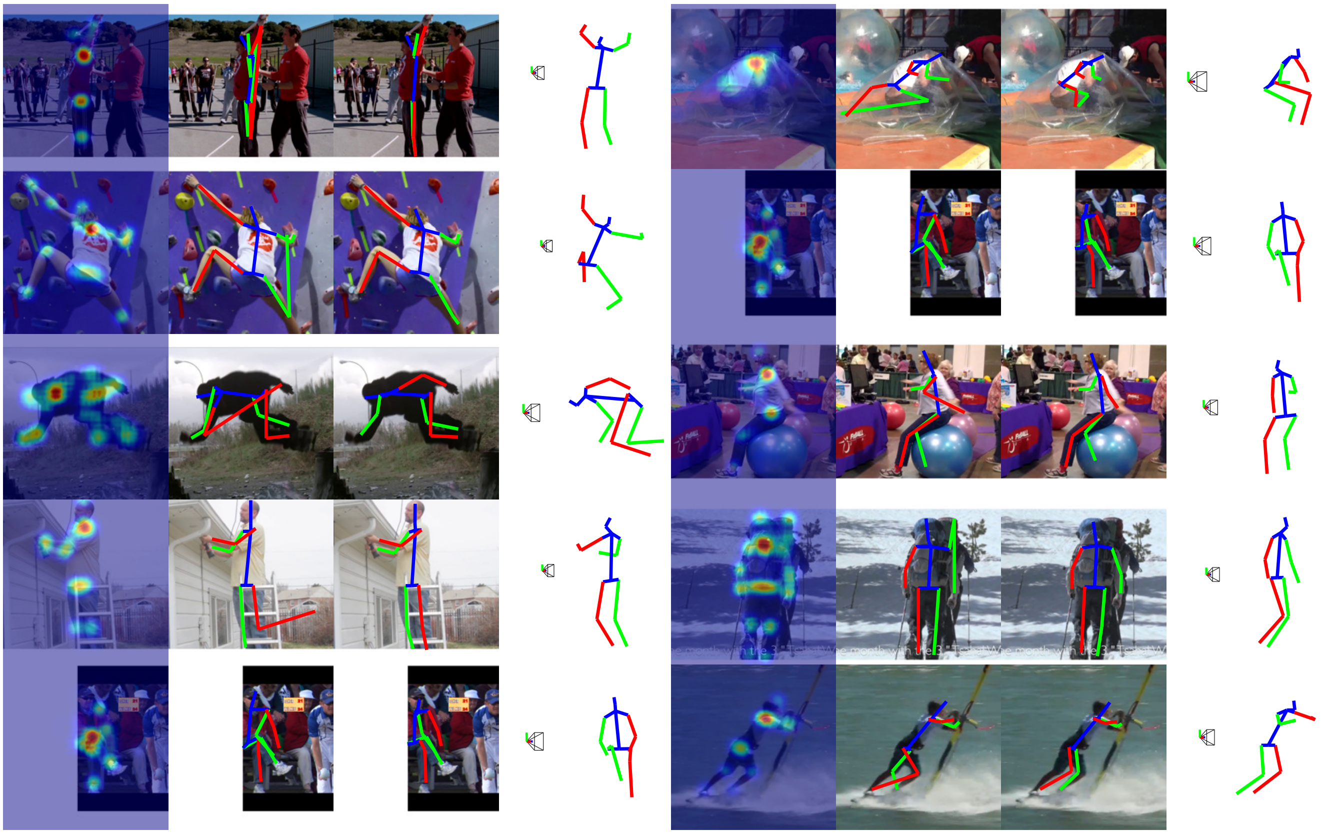

Fig. 8. Example frame results on MPII [62], where the errors in the 2D heat maps are corrected after considering the 3D pose prior. In each example, the images from left-to-right correspond to the heat map (all joints combined), the 2D pose found by greedily locating each joint separately according to 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.

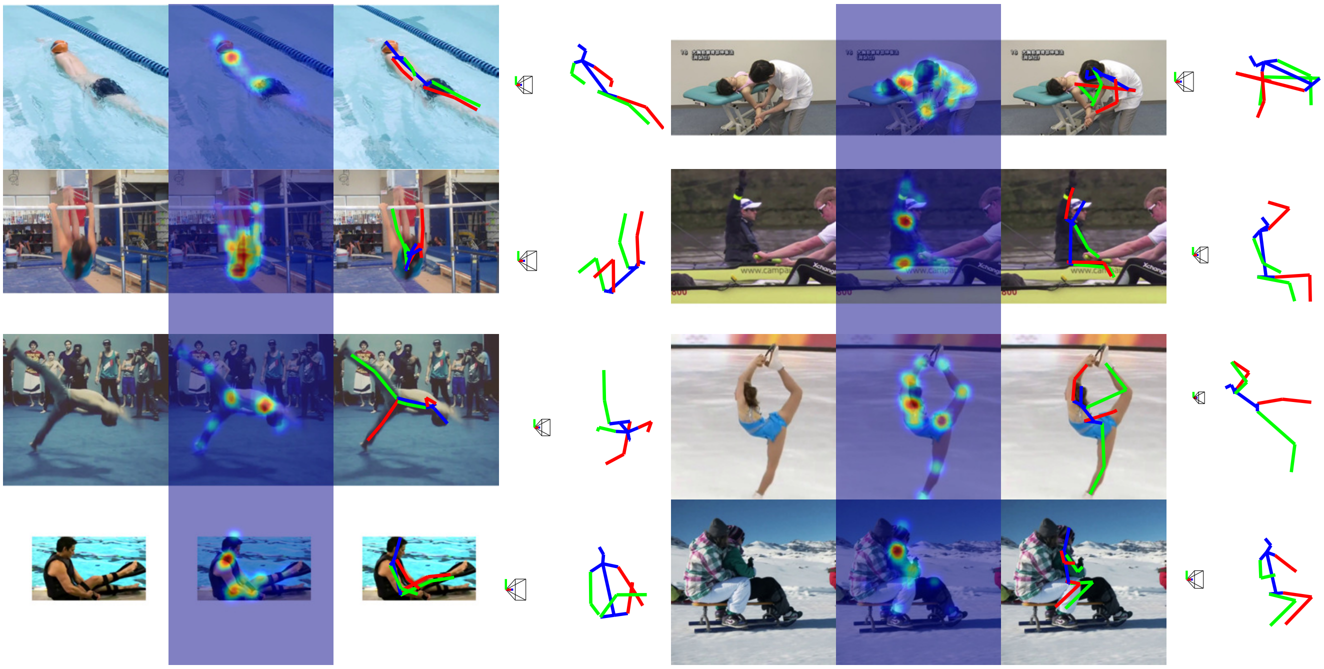

Fig. 9. Failed examples on MPII [62]. In each example, the images from left-to-right correspond to the input image, the heat map (all joints combined), the estimated 2D pose, and the estimated 3D pose visualized in a novel view. The original viewpoint is also shown.

6 Summary

In summary, a 3D pose estimation framework from 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 joint detectors, pose representations 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 joint localization. Future extensions may include incremental algorithms for online tracking-by-detection and handling multiple subjects.

Appendix

Proof of Equation (14)

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

$$\begin{aligned} \mathcal{L}(\theta; \boldsymbol{W}) &= \sum_{t=1}^{n} \left\| \boldsymbol{W}_t - \boldsymbol{R}_t \sum_{i=1}^{k} c_{it} \boldsymbol{B}_i - \boldsymbol{T}_t \boldsymbol{1}^T \right\|_F^2 \\ &= \|\boldsymbol{W} - \boldsymbol{Z}(\theta)\|_F^2 , \end{aligned} \tag{19}$$

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

$$\begin{aligned} &\int \mathcal{L}(\theta; \boldsymbol{W}) \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \|\boldsymbol{W} - \boldsymbol{Z}(\theta)\|_F^2 \ \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \int \left\{ \|\boldsymbol{W}\|_F^2 - \langle \boldsymbol{W}, \boldsymbol{Z}(\theta) \rangle + \|\boldsymbol{Z}(\theta)\|_F^2 \right\} \ \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \\ &= \ \left\{ \text{const} - \int \langle \boldsymbol{W}, \boldsymbol{Z}(\theta) \rangle \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} + \|\boldsymbol{Z}(\theta)\|_F^2 \right\} \\ &= \ \left\{ \text{const} - \left\langle \int \boldsymbol{W} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} \ , \ \boldsymbol{Z}(\theta) \right\rangle + \|\boldsymbol{Z}(\theta)\|_F^2 \right\} \\ &= \left\| \int \boldsymbol{W} \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') d\boldsymbol{W} - \boldsymbol{Z}(\theta) \right\|_F^2 + \text{const} \\ &= \|\mathrm{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right] - \boldsymbol{Z}(\theta)\|_F^2 + \text{const} \end{aligned} \tag{20}$$

Derivation of Equation (15)

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{W}|\boldsymbol{I}, \theta'\right] &= \int \Pr(\boldsymbol{W}|\boldsymbol{I}, \theta') \ \boldsymbol{W} \ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{W}, \boldsymbol{I}, \theta')}{\Pr(\boldsymbol{I}, \theta')} \ \boldsymbol{W} \ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta') \Pr(\theta')}{\Pr(\boldsymbol{I}|\theta') \Pr(\theta')} \ \boldsymbol{W} \ d\boldsymbol{W} \\ &= \int \frac{\Pr(\boldsymbol{I}|\boldsymbol{W}) \Pr(\boldsymbol{W}|\theta')}{Z} \ \boldsymbol{W} \ d\boldsymbol{W} \end{aligned} \tag{21}$$

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, 4

[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, 8, 9

[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, 9

[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, 9

[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

[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, 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, 6

[43] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," in *ECCVW*, 2016. 2, 6

[44] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *ECCVW*, 2016. 2, 6

[45] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen *et al.*, "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, 7

[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, 9

[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, 3, 4

[58] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *arXiv preprint arXiv:1509.04309*, 2015. 2, 4, 5, 6

[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, 9

[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, 8, 9

[62] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, June 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*. Springer, 2016. 2, 7

[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, 8, 9

[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. 2

[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. 3

[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. 3

[71] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 4

[72] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. 4

[73] C.-P. Lu, G. D. Hager, and E. Mjolsness, "Fast and globally convergent pose estimation from video images," *PAMI*, vol. 22, no. 6, pp. 610–622, 2000. 7

[74] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *CVPR*, 2013. 10