

# Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps

Yu Du<sup>1</sup>, Yongkang Wong<sup>2</sup>, Yonghao Liu<sup>1</sup>, Feilin Han<sup>1</sup>, Yilin Gui<sup>1</sup>,  
Zhen Wang<sup>1</sup>, Mohan Kankanhalli<sup>2,3</sup>, Weidong Geng<sup>1\*</sup>

<sup>1</sup>College of Computer Science, Zhejiang University

<sup>2</sup>Interactive & Digital Media Institute, National University of Singapore

<sup>3</sup>School of Computing, National University of Singapore

{answeror, yonghaoliu, hanfeilin, ylgui, wangzh\_cs, gengwd}@zju.edu.cn,  
yongkang.wong@nus.edu.sg, mohan@comp.nus.edu.sg

**Abstract.** The recovery of 3D human pose with monocular camera is an inherently ill-posed problem due to the large number of possible projections from the same 2D image to 3D space. Aimed at improving the accuracy of 3D motion reconstruction, we introduce the additional built-in knowledge, namely height-map, into the algorithmic scheme of reconstructing the 3D pose/motion under a single-view calibrated camera. Our novel proposed framework consists of two major contributions. Firstly, the RGB image and its calculated height-map are combined to detect the landmarks of 2D joints with a dual-stream deep convolution network. Secondly, we formulate a new objective function to estimate 3D motion from the detected 2D joints in the monocular image sequence, which reinforces the temporal coherence constraints on both the camera and 3D poses. Experiments with HumanEva, Human3.6M, and MCAD dataset validate that our method outperforms the state-of-the-art algorithms on both 2D joints localization and 3D motion recovery. Moreover, the evaluation results on HumanEva indicates that the performance of our proposed single-view approach is comparable to that of the multi-view deep learning counterpart.

**Keywords:** Human Pose Estimation, Height-map

## 1 Introduction

Marker-less motion capture is an active field of research in computer vision and graphics with applications in computer animation, video surveillance, biomedical research, and sports science. According to the recent study on world population aging [1], the life expectancy at age 60 and above is expected to grow in the next few decades. This anticipates an emerging need in video-based analysis systems to monitor the elderly in nursing home as an event alert system.

Existing motion capture approaches can be broadly divided into two categories: (1) methods based on monocular camera [2–5], and (2) methods that rely

---

\* Corresponding author.

on synchronous multi-view streams [6–8]. Nowadays, single view approaches are getting more attention in the industry. Although multi-view visual data presents richer information for marker-less motion capture, such data are not always available in reality, especially in the applications of video surveillance.

The recovery of 3D human poses with monocular image sequences is an inherently ill-posed problem, since the observed projection on a 2D image can be explained by multiple 3D poses and camera positions, when we try to infer poses from single-view images or motions from monocular video [9]. The problem becomes even more challenging if we consider realistic situations in which image features, such as the body silhouette, limbs or 2D joints, cannot be accurately detected due to environment factors or occlusions [10]. Nevertheless, human observers are able to accurately estimate the pose of a human body with a single eye. In most cases, they are also able to effortlessly organize the anatomical landmarks in three-dimensional space and predict the relative position of the camera, where the ambiguity is resolved by leveraging on vast memory of likely 3D configurations of humans [9]. A reasonable proxy for such capabilities can be available by learning from motion capture libraries, ensuring anthropometric plausibility while discarding impossible configurations. Motivated by this, we aim to achieve accurate 3D reconstruction of human motion from monocular image sequence recorded by a calibrated camera.

We propose a novel framework for marker-less 3D human motion capture with a single-view calibrated camera, where the 3D human pose is articulated as 3D pose or a skeleton model parameterized by joint locations. It consists of three key components, namely height-map generation, 2D joint localization, and 3D motion generation. Inspired by the recent success of deep learning approach on RGB-D camera [11,12], we propose a dual-stream Deep Convolution Network (ConvNet) to effectively detect 2D landmarks of human joints. The RGB images and the additional built-in knowledge (i.e., height-map) are independently modeled with one stream of ConvNet, which are then jointly fined-tuned for improved 2D joints detection. In addition, the nature of the dual-stream ConvNet architecture also allows the proposed method to be coupled with any improved RGB-based 2D joint detection algorithm in the future. Furthermore, in the 3D motion estimation stage, we propose to reinforce both the pose-conditioned joint velocity and the temporal coherence constraints of continuity of the camera and 3D poses in the optimization scheme. To the best of our knowledge, this is the first algorithm that utilizes the height-map to capture 3D articulated skeleton motion from a calibrated monocular camera.

The remaining of the paper is organized as follows. Section 2 reviews the related literature. Section 3 elaborates on the details of the proposed framework. Section 4 evaluates the proposed framework with both synthetic and real-world video dataset. Section 5 concludes the paper.

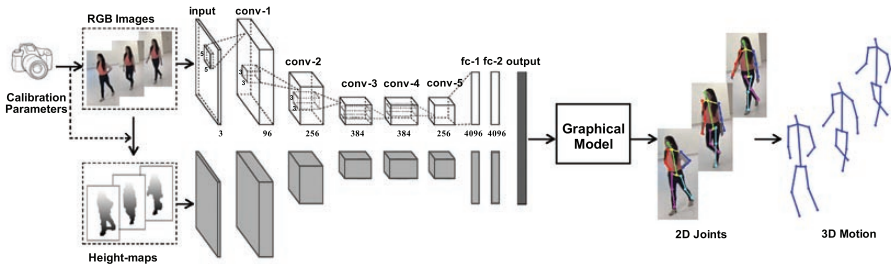
## 2 Literature Review

Human 3D pose estimation has received a lot of attention from the communities, and has been investigated using monocular camera [2–5], multi-view image sequences [6–8], and RGB-D sensor [11]. In the early research, Fischler and Elschlager [13] introduced the Pictorial Structures Model (PSM) to represent an object by a collection of parts in a deformable configuration. This is further adopted to represent human body as an articulated structure for tracking [4], recognition [14,15], and pose estimation [4,15–18] problems. Various approaches were proposed to learn PSM directly via RGB images [19,20] or depth images [17,21].

3D pose estimation from a single image is an inherently ill-posed problem due to the possibility of multiple plausible projections from the same 2D image to a variety of 3D poses. A common approach is to project the estimated 2D landmarks from single image to the 3D space by imposing certain constraints [5, 9,10,22,23]. Simo-Serra *et al.* [10] imposed kinematic constraints to guarantee that the resulting pose resembles a human shape. In [5] and [9], 3D pose is represented as a sparse linear combination of an overcomplete dictionary. In [9], the sparse model is computed while enforcing anthropometric regularity on pose structure, whereas [5] enforced eight limb length constraints to eliminate errors generated from inaccurate 2D landmarks estimation. Simultaneous estimation of both 3D pose and camera parameters further improved the performance [5, 9,23]. Recently, [24] considered the estimation of 3D pose as a Spatio-Temporal Matching problem that explore the correspondence between video and 3D motion capture data. The aforementioned approaches considered the 2D landmarks and 3D pose as separate problems, where [22] jointly models both problems using a Bayesian framework.

Recently, ConvNets has been applied to estimate human pose from a monocular camera. Specifically, it aims to label anatomical landmarks (or joints) on image. DeepPose [25] is the first work that holistically cast pose modeling as a joint regression problem. Chen and Yuille [2] further improved the estimation by modeling human pose as a graphical model where each local joint is considered as a node. They consider the part detection as a local image measurement and predict the spatial relationships between joints as an image dependent pairwise relations. Similarly, Tompson *et al.* [26] proposed a hybrid architecture that consists of ConvNet part detector and Markov Random Field inspired spatial-model to exploit the structural domain constraints. This approach is further improved by a cascaded architecture that combines fine and coarse scale ConvNet to accommodate the variance of human annotation errors [27]. Focusing on structured-output learning, Li *et al.* [28] embed image and pose into a high-dimensional space, whereas the image-pose embedding and score function are jointly trained using a maximum-margin cost function with a 2-stage optimization procedure. In contrast to the aforementioned approaches, [29–31] directly predicted the 3D pose from images.

While the single shot approach can be applied to model a human action from an image sequence, such approaches often result in unstable 3D motion recon-



**Fig. 1.** Conceptual illustration of the proposed 3D human motion capture framework with a calibrated monocular camera.

structions [3]. To address this, Wandt *et al.* [3] modeled 3D pose as a linear combination of base poses and proposed a periodic model for the mixing coefficients to improve the efficiency and accuracy for periodic motion (e.g., walking, running, *etc.*). The reconstruction on non-periodic motion is achieved with a regularization term on temporal bone length constancy. In [32], 3D pose estimation is considered as a sparsity-driven reconstruction problem with temporal smoothness prior. Furthermore, it regards 2D joints as a latent variable which the uncertainty maps can be jointly learned with deep learning based joint detector. Hasler *et al.* [7] proposed to perform automatic camera registration and audio synchronization for multiple cameras, followed by recovering 3D human pose by computing the correspondence between the extracted silhouettes. Hofmann and Gavrilova [6] proposed a multi-stage verification process for the shape hypotheses generated from each camera, and removed the temporal ambiguity by maximizing the best trajectories across cameras. Elhayek *et al.* [8] combined a ConvNet based part detection model [26] with a generative model-based tracking algorithm based on Sums of Gaussians framework, which captures temporally stable full articulated joints from multiple cameras.

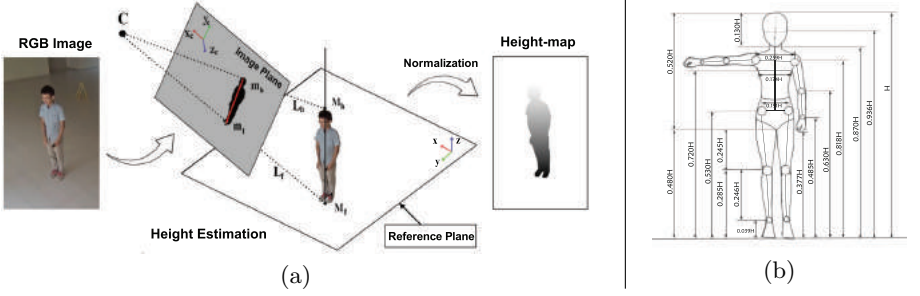
To the best of our knowledge, there exist no work that employs height-map as built-in knowledge together with color image for 2D joints detection.

### 3 Proposed Method

#### 3.1 Overview

In this work, the main objective is to accurately recover the 3D human poses with a calibrated monocular camera, where the 3D human motion is represented by a skeleton model parameterized by joint positions. Our proposed framework consists of three key components, namely, height-map generation, 2D joints localization, and 3D motion estimation. A conceptual diagram of the proposed framework is shown in Fig. 1.

The height-map is generated by existing height estimation algorithm [33] using calibrated camera parameters and the body silhouettes. Inspired by the



**Fig. 2.** (a) Illustration of height-map generation with pre-calibrated monocular camera, (b) Anatomical decomposition of Skeleton based on height [35].

recent success of skeleton pose recognition using RGB-D (color + depth) sensors [11,12], we propose a dual-stream deep ConvNet for 2D joints localization with RGB images and the computed height-maps (RGB-H). The dual-stream ConvNet is first trained with “Leeds Sports Poses” (LSP) dataset [34] (for the RGB stream), which is then used as an initial stream for the height-maps and trained with a synthetic dataset (for the H stream). The resulting model is then jointly fined-tuned on the target dataset with the computed RGB-H images. For the 3D human pose estimation, we consider both the reinforced temporal constraints of the camera and the pose-conditioned joint velocity.

### 3.2 Height-Map Generation

Height-map is a grayscale image designed to be an intermediate new representation of body parts, where pixels in a height-map indicate its height with respect to the reference plane rather than a measure of color, depth or intensity. For each pixel of the human body, we apply the height estimation method proposed by Park *et al.* [33] to calculate height from monocular RGB camera by back-projecting 2D features of an object into the 3D scene space (see Fig. 2). To accommodate variation in height across human subjects, we normalize the estimated height,  $H$ , on each pixel to relative height,  $\hat{H}$ , via:

$$\hat{H}(x, y) = k \cdot \frac{H(x, y)}{h_i} \quad (1)$$

where  $x$  and  $y$  is the pixel coordinate, and  $h_i$  indicates the body height of  $i$ th person.  $k$  is a scale constant to map the relative height-map to a desired range, which is empirically set to 255 to mimic an intensity channel (see Fig. 2a). Given a height-map, we implicitly encode the spatial relationships among joints of a skeleton structure [35] (see Fig. 2b).

### 3.3 2D Joints Localization

Given an image sequence with  $m$  frames  $\{I_1, \dots, I_m \mid I_t \in \mathbb{R}^{w \times h \times d}\}$ , where  $w$  and  $h$  are the width and height of an image, and  $d$  is the number of channels. The goal

is to localize the anatomical landmarks of human (i.e., 2D joints),  $\{p_1, \dots, p_m | p_t \in \mathbb{R}^{2n}\}$ , in each image using both the RGB images and the estimated height-maps, where  $n$  is the pre-defined number of 2D joints. In this work, we assume that one pose is observed at each frame to simplify the mathematical formulation.

We adapt a ConvNet-based 2D joints localization method [2], which achieved state-of-the-art results on several public benchmark datasets<sup>1</sup>. This method depicts human pose as a graphical model and predicts the spatial relationship between joints as an image dependent pairwise relation. Inspired by the hybrid approach that use RGB-D sensor data [11,12], we design a dual-stream deep learning architecture, which operates on both RGB image and height-map, and a fully connected layer is deployed to fuse these two streams (conceptual diagram is shown in Fig. 1). This architecture is similar to other recent multi-stream approaches for recognition and segmentation tasks [36–39].

The localization of 2D joints in each stream is formulated as the optimization of a score function over a part based graphical model [16]:

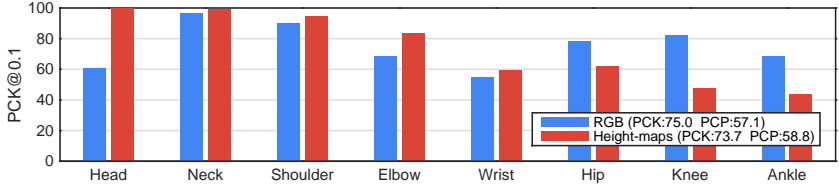
$$F(\mathbf{l}, \mathbf{t} | \mathbf{I}) = \sum_{i \in \mathcal{V}} U(\mathbf{l}_i | \mathbf{I}) + \sum_{(i,j) \in \mathcal{E}} R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji} | \mathbf{I}) + w_0 \quad (2)$$

where  $\mathbf{l} = \{\mathbf{l}_i | i \in \mathcal{V}\}$  is a set of joint positions,  $\mathbf{t} = \{t_{ij} | (i,j) \in \mathcal{E}\}$  is the pairwise relation type, and  $w_0$  is a bias term.  $\mathcal{V}$  and  $\mathcal{E}$  are the sets of vertices and edges of the graphical model, respectively.  $U$  and  $R$  contain mixtures of part types and pairwise relation types, which are specified as the marginalization of a joint distribution modeled by ConvNet. The input of the ConvNet is an image patch while the output is the evidence for a part to lie in this patch with a certain relationship to its neighbours. We refer the reader to [2] for more details. Given the learned models, we discard the output layers of both streams and employ a new output layer to fuse the output of the last fully connected layers.

The dual-stream ConvNet employs a stage-wise training strategy. The RGB stream is pre-trained on LSP dataset [34], and the resultant network is further applied on our synthetic height-maps dataset to obtain the initial weights of the height stream. Note that in order to reuse the pre-trained network on color images to initialize the height stream, we recreate a RGB image by replicating height-map three times as that in [40]. The entire network is then jointly fine-tuned on a target training set.

**Validation of Height-Map for 2D joints Localization** To evaluate the feasibility of using height-map for effective localization of 2D joints, we conducted a preliminary experiment on the 8-persons test set of a real-world surveillance dataset, namely Multi-Camera Action Dataset (MCAD) [41]. The height-map based single-stream ConvNet is trained on our synthetic dataset using the pre-trained ConvNet provided by [2]. The preliminary result (see in Fig. 3) shows that the pure height-map based approach is comparable and a complement to that based on the pre-trained model with RGB images in [2]. Therefore, we

<sup>1</sup> [http://human-pose.mpi-inf.mpg.de/#related\\_benchmarks](http://human-pose.mpi-inf.mpg.de/#related_benchmarks)



**Fig. 3.** Preliminary study of 2D joints localization with single-stream ConvNet on MCAD [41]. The values in the parenthesis are the mean value of PCK@0.1 and PCP [2].

argue that it is feasible to incorporate height-maps into the algorithmic pipeline of localizing landmark of joints from images. Please refer to Section 4 for details about databases and evaluation metrics.

### 3.4 3D Motion Estimation

Given a sequence of 2D joints  $\{p_1, \dots, p_m | p_t \in \mathbb{R}^{2n}\}$ , the corresponding 3D poses  $\{P_1, \dots, P_m | P_t \in \mathbb{R}^{3n}\}$  can be estimated by optimizing the following objective function

$$\min_{\theta} \mathcal{L}(\theta; \mathbf{p}) + \mathcal{R}_t(\theta) + \mathcal{R}_a(\theta) \quad (3)$$

where  $\theta = \{\mathbf{P}, \mathbf{V}, \mathbf{R}, \mathbf{T}\}$  is the union of all the 3D motion parameters, in which  $\mathbf{p} = [p_1^T \dots p_m^T]^T \in \mathbb{R}^{2mn}$ ,  $\mathbf{P} = [P_1^T \dots P_m^T]^T \in \mathbb{R}^{3mn}$ , and  $\mathbf{V} = [V_1^T \dots V_m^T]^T \in \mathbb{R}^{3mn}$  denote the 2D position, the 3D position, and the 3D velocity of each joint, respectively;  $p_t$  is the concatenation of  $\mathbf{l}$  at time  $t$ ;  $\mathbf{R} = \bigoplus_{t=1}^m (I_n \otimes R_t) \in \mathbb{R}^{3mn \times 3mn}$  and  $\mathbf{T} = [\mathbf{1}_{n \times 1} \otimes T_1^T \dots \mathbf{1}_{n \times 1} \otimes T_m^T]^T \in \mathbb{R}^{3mn}$  denote the orientation and position of the person in the camera frame;  $\otimes$  and  $\oplus$  are the Kronecker product and direct sum respectively;  $I$  is the identity matrix.

The first term is the reprojection error which is formulated as:

$$\mathcal{L}(\theta; \mathbf{p}) = \|\mathbf{p} - h(\mathbf{R}\mathbf{P} + \mathbf{T})\|^2 \quad (4)$$

where  $h: \mathbb{R}^{3mn} \rightarrow \mathbb{R}^{2mn}$  performs perspective projection of the 3D joints to the 2D image plane.

The second term enforces the temporal constraints on each joint's movement speed, the orientation of the person with respect to the camera, and the corresponding position

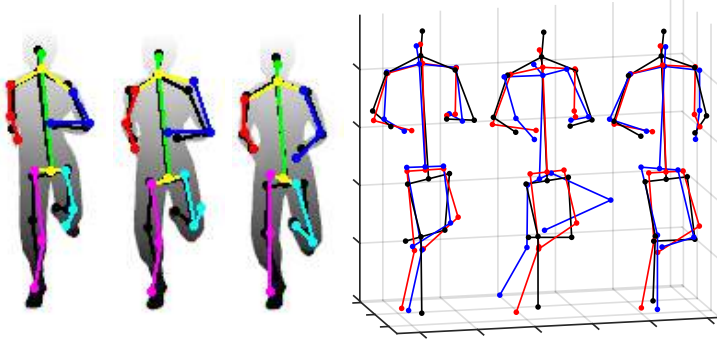
$$\mathcal{R}_t(\theta) = \alpha \|\nabla_t(\mathbf{P} - \mathbf{V})\|^2 + \beta_r \|\nabla_t \mathbf{R}\|^2 + \beta_t \|\nabla_t \mathbf{T}\|^2 \quad (5)$$

where  $\nabla_t$  is the discrete temporal derivative operator. The first sub-term penalizes the inconsistency between position and velocity. The second and third terms impose first-order smoothness on the orientation and position of the target person.

The last term imposes the anthropometric constraints on limb lengths

$$\mathcal{R}_a(\theta) = \gamma \|g(\mathbf{P})\|^2 \quad (6)$$

where  $g$  computes the length difference of arms and legs between the estimated poses and the training data.



**Fig. 4.** Qualitative illustration of the robustness of the temporal coherence constraints to inaccurate localization of 2D joints. The ground-truth 2D and 3D skeletons are colored in black. On the left are three consecutive synthetic height-maps of running motion, where the localization of the left ankle in the second frame is incorrect. On the right are the estimated 3D poses by [9] (in blue) and by our method (in red).

**Pose-Conditioned Joint Velocity** We represent a 3D human pose  $P_t$  and the joint velocity of this pose  $V_t$  at time  $t$  by a linear combination of a set of bases  $\mathbf{B} = \{b_1, \dots, b_k\}$  and a mean vector  $\mu$

$$X_t = [P_t^T, V_t^T]^T = \mu + \mathbf{B}_t^* \omega_t \quad (7)$$

$$\{b_i\}_{i \in \mathcal{I}_{\mathbf{B}_t^*}} \in \mathbf{B}_t^* \subset \mathbf{B} \quad (8)$$

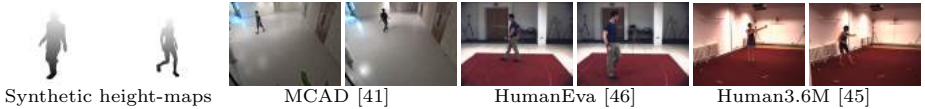
where  $\omega_t$  are the basis coefficients,  $\mathbf{B}_t^*$  is an optimal subset of an dictionary  $\mathbf{B}$  where each column of matrix  $\mathbf{B}_t^*$  is a basis  $b_i$  selected with index vector  $\mathcal{I}_{\mathbf{B}_t^*}$  from  $\mathbf{B}$ .  $\mathbf{B}$  is created by concatenating the bases computed from various types of motions using Principal Component Analysis (PCA).

When training the bases  $\mathbf{B}$ , each sample is formed by the concatenation of the 3D pose and the joint velocity of this pose. The joint velocity is approximated by the difference of joint positions in current and the  $k$ -th previous frames, where  $k = \lfloor s_3/s_2 + 0.5 \rfloor$ , in which  $s_2$  and  $s_3$  are the sampling rates of the input sequence and motion database respectively.

Based on this representation, the parameter  $P_t$  and  $V_t$  at time  $t$  are defined as  $[I_n \ 0_n](\mu + \mathbf{B}_t^* \omega_t)$  and  $[0_n \ I_n](\mu + \mathbf{B}_t^* \omega_t)$ , respectively. The parameter set can be re-written as  $\theta = \{\mathbf{I}, \mathbf{\Omega}, \mathbf{R}, \mathbf{T}\}$ , where  $\mathbf{I} = \{\mathcal{I}_{\mathbf{B}_1^*}, \dots, \mathcal{I}_{\mathbf{B}_m^*}\}$  is the index vectors, and  $\mathbf{\Omega} = [\omega_1^T \dots \omega_m^T]^T \in \mathbb{R}^{3mn}$  represents the coefficient vectors.

The sparse representation of human pose by an overcomplete dictionary has been adopted in recent work [9,23]. The key difference here is that our dictionary encodes not only the anthropomorphically plausible 3D poses, but also the pose-conditioned joint velocity. Fig. 4 shows that the implausible 3D poses estimated from the inaccurate localization of 2D joints can be corrected by our temporal coherence constraints.





**Fig. 5.** Samples from four datasets for evaluation.

**Optimization** The objective function in (3) is solved by Projected Matching Pursuit [9]. In each iteration, we first compute the loss function in (3) for each frame with the available basis, followed by a frame level optimal basis selection with basis that contribute to minimum loss. The selected optimal basis is excluded for the next iteration. Then we estimate  $\{\mathbf{\Omega}, \mathbf{R}, \mathbf{T}\}$  in (3) by Levenberg-Marquardt algorithm [42]. The optimization terminates if the reprojection error is less than a threshold  $\delta$  or the number of the basis selected for each frame reaches  $\phi$ .  $\mathbf{R}$  and  $\mathbf{T}$  are initialized by EPnP algorithm [43] using the known intrinsic parameters of the calibrated camera.

## 4 Experiments

In this section, we evaluate the performance of the proposed method from three perspectives. First, we evaluate the efficacy of the proposed dual-stream ConvNet for 2D joints localization, which include various single-stream and dual-stream configurations, as well as comparison against [44]. Second, the evaluation of 3D motion recovery is made with the ground-truth 2D joint locations, and compared against [9,23]. Third, we compare the entire pipeline of the proposed framework against [5,10,28,45]. To keep the consistency with the literature, we use a skeleton of 14 joints [44] where a virtual root joint is added merely for visualization. Before computing the 3D error in Section 4.3 and 4.4, the estimated 3D pose is rigidly aligned with the ground-truth as that in existing works [10,22,23]. For the 3D evaluation on Human3.6M, we do not perform the rigid alignment on the resulting motion.

Based on the preliminary experiment, we fix the parameters of the proposed 3D motion estimation method in all experiments, where  $\alpha = 0.1$ ,  $\beta_r = 10$ ,  $\beta_t = 1$ ,  $\gamma = 1$ ,  $\delta = 500$  and  $\phi = 15$ .

### 4.1 Datasets

We evaluate our approach on four datasets: (1) the synthetic height-maps dataset, (2) HumanEva dataset [46], (3) Human3.6M dataset [45], and (4) Multi-Camera Action Dataset (MCAD) [41]. The samples are shown in Fig. 5. We generate a large scale synthetic height-maps dataset, which consists of 184,872 synthetic height-maps along with the corresponding 2D and 3D joint locations, which are generated from 9 characters with 36 surrounding viewpoints. For each character there are around 570 poses extracted from five-hour motion capture data about dancing, walking, fighting, etc. HumanEva [46] is a benchmark dataset

for 3D pose estimation. It contains synchronized multi-view videos captured by calibrated cameras and 3D ground-truth motion of 4 subjects performing 6 pre-defined actions with 3 repetitions. We use the walking and jogging motions of three subjects in the HumanEva, as that in [5,10], to evaluate the localization of 2D joints and the overall performance of our method. The third dataset we used is Human3.6M [45], which is currently the largest video pose dataset. It contains over 3.6 million frames of different human poses, viewed from 4 different angles, using an accurate human motion capture system. The motions were performed by 11 human subjects under 15 activity scenarios. Following [28], we split the dataset to have 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9, S11) for testing. As far as the dataset is redundant, we select 1 out of 50 frames from all 4 cameras for training and every 5-th frame from camera 2 for testing, using the standard 17 joint skeleton from Human3.6M. The MCAD [41] consists of 20 persons and 18 actions recorded under 5 non-overlapping surveillance cameras, 14,298 action sequences in total. We manually labeled the 2D joints of all individuals in one of the cameras. 10 of the human subjects are used for training and the remaining ones are reserved for testing. All the data is converted into observer centric view during the pre-processing stage, as in [2].

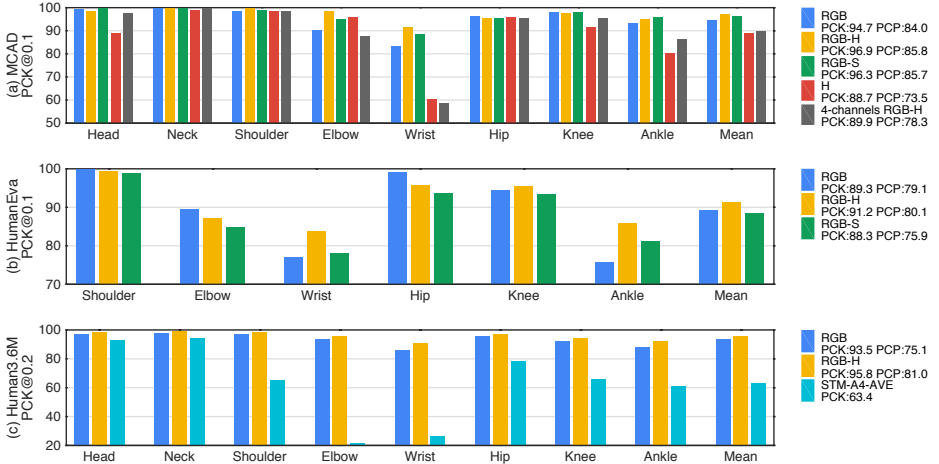
## 4.2 Evaluation of 2D Joints Localization

We consider two metrics as indicators to evaluate the performance of 2D joint localization. The performance analyzed in terms of the Probability of Correct Keypoints (PCK) metric proposed in [16], which measures the accuracy using a curve of the percentage of correctly localized joints by varying localization precision threshold. In this work, we also adopt the strict Probability of Correct Pose (PCP) proposed by Chen *et al.* [2], where a body part is considered as correct if both of its joints lie within 50% of the length of the ground-truth annotated endpoints. Based on the project site of [47]<sup>2</sup>, we select [2] as the baseline for 2D joint localization as it achieved the best performance for the time being.

**Evaluation on MCAD** We first compare the proposed 2D joints localization method (RGB-H) with the one solely relying on color images (RGB) [2] or height-maps on the test set of MCAD. The ConvNets of these three methods are fine-tuned on the training set of MCAD with 30,000 iterations and a learning rate of 0.001. Then the part based graphical models are also re-trained based on the fine-tuned ConvNets. As shown in Fig. 6(a), although the model solely based on height-maps achieves lower accuracy than [2], combining color images and height-maps indeed improves the precision.

Next, we compare our dual-stream ConvNet against another single-stream ConvNet on the test set of MCAD. The single-stream ConvNet has exactly the same structure as the one in [2] except that the input dimension of the first layer is 4 (denoted as “4-channels RGB-H”). This model is trained from scratch on the

<sup>2</sup> <https://cse.sc.edu/%7Efefan23/projects/cvpr15/cvpr15.html>



**Fig. 6.** Evaluation of 2D joints localization with RGB [2], RGB-H, RGB-S (RGB-Silhouette), H (height-maps), 4-channels RGB-H and STM-A4-AVE [44] respectively on the MCAD, HumanEva and Human3.6M.

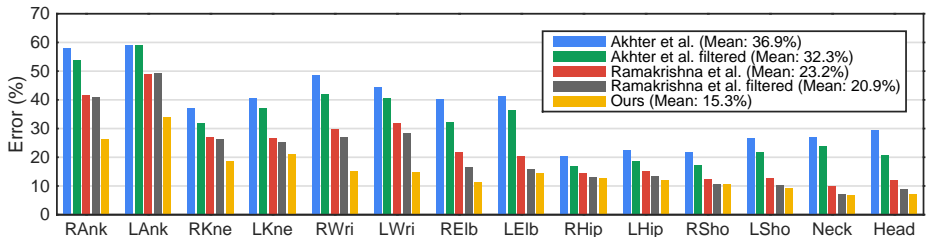
**Table 1.** Evaluation of 2D joints localization on Human3.6M. The numbers are PCK accuracy at threshold 0.2.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
STM-A4-AVE [44]	67.6	62.3	55.1	68.9	56.5	54.9	57.6	47.7
Chen <i>et al.</i> [2]	98.8	95.5	98.1	97.6	93.9	89.5	98.6	85.6
Ours	<b>99.0</b>	<b>96.9</b>	<b>98.9</b>	<b>98.4</b>	<b>96.5</b>	<b>94.0</b>	<b>99.2</b>	<b>93.8</b>
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Mean
STM-A4-AVE [44]	42.4	26.2	58.7	65.9	61.2	81.5	79.4	59.1
Chen <i>et al.</i> [2]	94.1	70.5	94.4	96.3	88.6	98.0	97.7	93.5
Ours	<b>95.0</b>	<b>77.2</b>	<b>96.1</b>	<b>98.2</b>	<b>94.3</b>	<b>98.7</b>	<b>98.9</b>	<b>95.8</b>

training set of MCAD. As shown in Fig. 6(a), the performance of dual-stream ConvNet is much better than that of single-stream ConvNet, especially for wrist joints which registers an improvement of 32.6 percentage points.

To investigate whether the body silhouette could achieve similar performance as height-map, we train and test a RGB-Silhouette (RGB-S) based model using the exactly same settings in the RGB-H case. Fig. 6(a) shows that RGB-H outperforms RGB-S.

**Evaluation on HumanEva** We compare three models (RGB, RGB-H and RGB-S) on the test set of HumanEva, where these models are trained on MCAD and not re-trained on this dataset. Because our definition of head and neck are different from HumanEva, we discard these two joints and evaluate with the remaining joints. As Fig. 6(b) shows, the precision of the estimated locations of the endsites are obviously improved by using RGB-H images, and the model based on the body silhouette does not generalize well on HumanEva.



**Fig. 7.** Evaluation of 3D motion recovery with known 2D joints. The respective average error is shown in the legend. The estimated poses of [23] and [9] are further filtered by zero-phase Butterworth filter (3rd order, 0.2 Hz for [23]; 2nd order, 1.7 Hz for [9]).

**Evaluation on Human3.6M** We compare the proposed method with [2] and STM-A4-AVE [44] on the test set (S9 and S11) of Human3.6M. Our model and [2] are fine-tuned on the training set of Human3.6M using the same settings in the experiment on MCAD. As shown in Fig. 6(c) and Table 1, our method significantly outperforms others, especially in terms of PCP metric.

### 4.3 Evaluation of 3D Motion Recovery with Ground-Truth 2D Joints

We compare the proposed 3D motion recovery method with others on a sequence of 154 consecutive frames of synthetic motion of running around a circle, where the 2D joints locations are known. The character is driven by the retargeted motion capture data of CMU motion capture database [48]. We use the source codes provided by [9] and [23]. We train the bases of our model and [9] on “running”, “walking”, “jumping” and “boxing” motions of CMU motion capture database by fixing the position and orientation of the root joint and concatenating PCA components which retrained 99% of the variance from each motion category. For [23], we directly test the provided model without re-training. We also report the result of [9,23] with simple smoothing filter. We use zero-phase Butterworth filter whose parameters are optimized with grid search. We report the relative reconstruction error proposed by [23], which is a distance measure relative to the length of the backbone of the ground-truth skeleton. Fig. 7 shows that our method achieves a lower reconstruction error.

### 4.4 Evaluation of 3D Motion Recovery with Predicted 2D Joints

In this section, we quantify the performance of 3D motion estimation as a distance measurement relative to the length of the backbone of the ground-truth skeleton [23]. Specifically, we report *Root Mean Square* (RMS) error on HumanEva and *mean per joint position error* on Human3.6M. Note that the difference in the evaluation scheme on HumanEva is to ensure consistency with [5]. Different from Section 4.3, we compare our entire pipeline which estimates 3D pose from raw RGB images and the corresponding height-maps.

**Table 2.** Evaluation of 3D motion estimation on 3 subjects of the HumanEva dataset. The value in each cell are the RMS error and standard deviation in millimeter.

	Walking				Jogging			
	S1	S2	S3	Mean	S1	S2	S3	Mean
[10]	99.6 (42.6)	108.3 (42.3)	127.4 (24.0)	111.8	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)	106.0
[5]	71.9 (19.0)	75.7 (15.9)	85.3 (10.3)	77.6	62.6 (10.2)	77.7 (12.1)	<b>54.4</b> (9.0)	64.9
Ours	<b>62.2</b> (18.6)	<b>61.9</b> (13.2)	<b>69.2</b> (22.4)	<b>64.4</b>	<b>56.3</b> (15.4)	<b>59.3</b> (14.4)	59.3 (15.5)	<b>58.3</b>

**Table 3.** Evaluation of 3D motion estimation on Human3.6M dataset. The error are reported in mean per joint position error (MPJPE) [45].

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sitting
LinKDE [45]	132.71	183.55	132.37	164.39	162.12	205.94	150.61	171.31	151.57
Li <i>et al.</i> [28]	-	136.88	<b>96.94</b>	124.74	-	168.68	-	-	-
Ours	<b>85.07</b>	<b>112.68</b>	104.90	<b>122.05</b>	<b>139.08</b>	<b>135.91</b>	<b>105.93</b>	<b>166.16</b>	<b>117.49</b>

	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Mean (6 actions)	Mean (15 actions)
LinKDE [45]	243.03	162.14	170.69	177.13	96.60	127.88	160.00	162.14
Li <i>et al.</i> [28]	-	-	-	<b>132.17</b>	<b>69.97</b>	-	121.56	-
Ours	<b>226.94</b>	<b>120.02</b>	<b>117.65</b>	137.36	99.26	<b>106.54</b>	<b>118.69</b>	<b>126.47</b>

We first evaluate our proposed framework against state-of-the-art [5,10] on the HumanEva. To ensure consistency with [5], the reconstruction error is computed on 12 joints<sup>3</sup>. As shown in Table 2, our method significantly outperforms others in 5 out of 6 tests and achieved the mean reconstruction error of 64.4 mm and 58.3 mm on walking and jogging motion respectively, which is around 17.0% and 10.2% reduction from [5]. In addition, our results is comparable to the state-of-the-art performance (66.5 mm) [8]. However, we would like to highlight that [8] is a multi-view deep-learning based approach, which has the advantage of richer information from multiple views. It should also be noted that we didn’t fine-tune our model on the HumanEva.

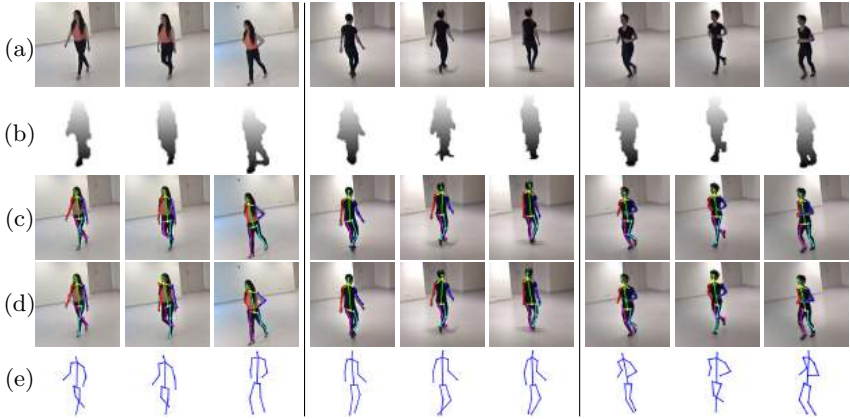
The second evaluation is conducted on the Human3.6 with results shown in Table 3. Our proposed approach outperforms [45] on almost all actions with an overall improvement of around 22%. Comparing with [28], we achieved better results on 3 out of 6 actions and the mean error favors our framework. Note that [28] is significantly better on the *Walking* action, while our approach stands out on the *Discussion* and *Photo* action.

And finally, we show the qualitative results of our proposed method on three persons from the MCAD [41]. As shown in Fig. 8, the localized 2D joints resemble that from the ground-truth label and the resultant 3D pose from the recovered 3D motion is good.

## 5 Conclusion

Monocular 3D human pose estimation is a highly ambiguous problem that requires introducing additional knowledge [11]. In this work, we studied the efficacy

<sup>3</sup> the left and right shoulders, elbows, wrists, hips, knees and ankles



**Fig. 8.** Qualitative result of the proposed framework of 3 persons (left, middle and right) from the MCAD [41]. (a) Image sequence, (b) Computed height-maps, (c) Ground-truth of 2D joints, (d) Localized 2D joints, and (e) Recovered 3D motion.

of height-map as a type of built-in prior knowledge to detect the anatomical landmarks of a human body, as well as enforce the temporal constraints on the camera and 3D poses for improved skeleton-based human pose estimation. Together with both components, we achieved state-of-the-art performance for both 2D joints localization and 3D motion estimation over two benchmark datasets (HumanEva & Human3.6M) and a real-world surveillance dataset (MCAD). The codes and the annotations of MCAD are available at <http://zju-capg.org/heightmap>.

Moreover, we evaluate our single view RGB-H approach with a state-of-the-art multi-view approach [8] on the walking motion from HumanEva dataset. On average, the spatial precision difference in detected joints is very close to each other on the mean reconstruction error. This suggests that our single view RGB-H method is very competitive for some real-world applications, such as human behavior analysis for event alert system, which usually require highly accurate 3D motion recovery from monocular video clips. This also enables us to utilize the millions of monocular cameras from the existing surveillance networks where camera can be calibrated with a reasonable amount of effort.

For future work, we aim to extend our framework to accommodate complex human motion (e.g., break dance, yoga exercise, *etc.*), where the height-map may fail to indicate the anatomical structure. We are also interested in scenarios to recover 3D human motion with sporadic partial human body occlusion.

**Acknowledgements.** This work was supported by a grant from the National High Technology Research and Development Program of China (Program 863, 2013AA013705), and the National Natural Science Foundation of China (No. 61379067). This research was partly supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

## References

1. United Nations, Department of Economic and Social Affairs, Population Division: World population ageing 2013 (2013) ST/SEA/SER.A/348.
2. Chen, X., Yuille, A.L.: Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. In: NIPS. (2014) 1736–1744
3. Wandt, B., Ackermann, H., Rosenhahn, B.: 3D human motion capture from monocular image sequences. In: CVPR Workshops. (2015) 1–8
4. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR. (2010) 623–630
5. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust Estimation of 3D Human Poses from a Single Image. In: CVPR. (2014) 2369–2376
6. Hofmann, M., Gavrilu, D.M.: Multi-view 3D Human Pose Estimation in Complex Environment. *International Journal of Computer Vision* **96**(1) (2012) 103–124
7. Hasler, N., Rosenhahn, B., Thormählen, T., Wand, M., Gall, J., Seidel, H.: Markerless Motion Capture with unsynchronized moving cameras. In: CVPR. (2009) 224–231
8. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: CVPR. (2015) 3810–3818
9. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D Human Pose from 2D Image Landmarks. *Lecture Notes in Computer Science* **7575** (2012) 573–586
10. Simo-Serra, E., Ramisa, A., Alenyà, G., Torras, C., Moreno-Noguer, F.: Single image 3D human pose estimation from noisy observations. In: CVPR. (2012) 2673–2680
11. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data. In: ICCV. (2013) 2456–2463
12. Gupta, S., Arbelaez, P., Girshick, R., Malik, J.: Aligning 3D Models to RGB-D Images of Cluttered Scenes. In: CVPR. (2015) 4731–4740
13. Fischler, M.A., Elschlager, R.A.: The Representation and Matching of Pictorial Structures. *IEEE Trans. Comput.* **22**(1) (1973) 67–92
14. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. *International Journal of Computer Vision* **61**(1) (2005) 55–79
15. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009) 1014–1021
16. Yang, Y., Ramanan, D.: Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12) (2013) 2878–2890
17. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM* **56**(1) (2013) 116–124
18. Zhang, D., Shah, M.: Human Pose Estimation in Videos. In: ICCV. (2015) 2012–2020
19. Yasin, H., Iqbal, U., Krüger, B., Weber, A., Gall, J.: A dual-source approach for 3d pose estimation from a single image. In: CVPR. (2016) 4948–4956
20. Ionescu, C., Carreira, J., Sminchisescu, C.: Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation. In: CVPR. (2014) 1661–1668
21. Li, S., Chan, A.B.: 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. *LNCS* **9004** (2014) 1–16

22. Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F.: A Joint Model for 2D and 3D Pose Estimation from a Single Image. In: CVPR. (2013) 3634–3641
23. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR. (2015) 1446–1455
24. Zhou, F., la Torre, F.D.: Spatio-temporal Matching for Human Pose Estimation in Video. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8) (2016) 1492–1504
25. Toshev, A., Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Networks. In: CVPR. (2014) 1653–1660
26. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In: NIPS. (2014) 1799–1807
27. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: CVPR. (2015) 648–656
28. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3D human pose estimation. In: ICCV. (2015) 2848–2856
29. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In: CVPR. (2016) 991–1000
30. Kostrikov, I.: Depth Sweep Regression Forests for Estimating 3D Human Pose from Images. BMVC (2014) 1–13
31. Hong, C., Yu, J., Wan, J., Tao, D., Wang, M.: Multimodal Deep Autoencoder for Human Pose Recovery. IEEE Trans. Image Process **24**(12) (2015) 5659–5670
32. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K., Daniilidis, K.: Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In: CVPR. (2016) 4966–4975
33. Park, S.W., Kim, T.E., Choi, J.S.: Robust Estimation of Heights of Moving People Using a Single Camera. In: LNEE. Volume 120. (2012) 389–405
34. Johnson, S., Everingham, M.: Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In: BMVC. (2010) 1–11
35. Benbakreti, S., Benyettou, M.: Gait recognition based on leg motion and contour of silhouette. In: ICITeS. (2012) 1–5
36. Srivastava, N., Salakhutdinov, R.R.: Multimodal Learning with Deep Boltzmann Machines. In: NIPS. (2012) 2222–2230
37. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous Detection and Segmentation. Lecture Notes in Computer Science **8695** (2014) 297–312
38. Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: NIPS. (2014) 568–576
39. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M.A., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (Sept 2015) 681–687
40. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning Rich Features from RGB-D Images for Object Detection and Segmentation. LNCS **8695** (2014) 345–360
41. Li, W., Wong, Y., Liu, A.A., Li, Y., Su, Y.T., Kankanhalli, M.: Multi-camera action dataset (MCAD): A dataset for studying non-overlapped cross-camera action recognition. CoRR **abs/1607.06408** (2016)
42. Moré, J.J.: The levenberg-marquardt algorithm: Implementation and theory. In: Numerical analysis. Springer (1978) 105–116
43. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An accurate  $O(n)$  solution to the PnP problem. International Journal of Computer Vision **81**(2) (2009) 155–166
44. Zhou, F., la Torre, F.D.: Spatio-temporal matching for human pose estimation in video. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8) (2016)



45. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(7) (2014) 1325–1339
46. Sigal, L., Balan, A., Black, M.: HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision* **87**(1-2) (2010) 4–27
47. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: *CVPR*. (2015) 1347–1355
48. CMU: Carnegie-Mellon Mocap Database