Computer Vision and Image Understanding
journal homepage: www.elsevier.com

A Dual-Source Approach for 3D Human Pose Estimation from Single Images

Umar Iqbal[a,**], Andreas Doering[a], Hashim Yasin[b], Björn Krüger[c], Andreas Weber[d], Juergen Gall[a]

[a]*Computer Vision Group, University of Bonn, Germany*
[b]*National University of Computer and Emerging Sciences, Pakistan*
[c]*Gokhale Method Institute, Stanford, USA*
[d]*Multimedia, Simulation, Virtual Reality Group, University of Bonn, Germany*

ABSTRACT

In this work we address the challenging problem of 3D human pose estimation from single images. Recent approaches learn deep neural networks to regress 3D pose directly from images. One major challenge for such methods, however, is the collection of large amounts of training data. Particularly, collecting a large number of unconstrained images that are annotated with accurate 3D poses is impractical. We therefore propose to use two independent training sources. The first source consists of accurate 3D motion capture data, and the second source consists of unconstrained images with annotated 2D poses. To incorporate both sources, we propose a dual-source approach that combines 2D pose estimation with efficient 3D pose retrieval. To this end, we first convert the motion capture data into a normalized 2D pose space, and separately learn a 2D pose estimation model from the image data. During inference, we estimate the 2D pose and efficiently retrieve the nearest 3D poses. We then jointly estimate a mapping from the 3D pose space to the image and reconstruct the 3D pose. We provide a comprehensive evaluation of the proposed method and experimentally demonstrate the effectiveness of our approach, even when the skeleton structures of the two sources differ substantially.

1. Introduction

3D human pose estimation has a vast range of applications such as virtual reality, human-computer interaction, activity recognition, sports video analytics, and autonomous vehicles. The problem has traditionally been tackled by utilizing multiple images captured by synchronized cameras capturing the person from multiple views (Belagiannis et al., 2014; Sigal et al., 2012; Yao et al., 2012). In many scenarios, however, capturing multiple views is infeasible which limits the applicability of such approaches. Since 3D human pose estimation from a single image is very difficult due to missing depth information, depth cameras have been utilized for human pose estimation (Baak et al., 2011; Shotton et al., 2011; Grest et al., 2005). However, current depth sensors are also limited to indoor environments and cannot be used in unconstrained scenarios. Therefore, estimating 3D pose from single, in particular unconstrained, images is a highly relevant task.

One approach to address this problem is to follow a fully-supervised learning paradigm, where a regression model (Bo and Sminchisescu, 2010; Ionescu et al., 2014b; Kostrikov and Gall, 2014; Ionescu et al., 2014a; Agarwal and Triggs, 2006; Bo et al., 2008; Li and Chan, 2014; Tekin et al., 2015) or a deep neural network (Li et al., 2015; Tekin et al., 2016, 2017; Zhou et al., 2016a; Moreno-Noguer, 2017; Popa et al., 2017) can be learned to directly regress the 3D pose from single images. This approach, however, requires a large amount of training data where each 2D image is annotated with a 3D pose. In contrast to 2D pose estimation, manual annotation of such training data is not possible due to ambiguous geometry and body part occlusions. On the other hand, automatic acquisition of accurate 3D pose for an image requires a very sophisticated setup. The popular datasets like HumanEva (Sigal et al., 2010) or Human3.6M (Ionescu et al., 2014b) use synchronized multiple cameras with a commercial marker-based system to acquire accurate 3D poses for images. This, however, requires a very expensive hardware setup and also limits the applicability of such systems primarily to indoor laboratory environments due to the requirements of marker-based system like studio envi-

**Corresponding author: Tel.: +49-228-73-4690;
e-mail: uiqbal@iai.uni-bonn.de (Umar Iqbal)

ronment and attached markers. Some recent approaches such as EgoCap (Rhodin et al., 2016) allows to capture 3D poses in outdoor environments, but image data in such cases is restricted only to ego-centric views of the person.

In this work, we propose a dual-source method that does not require training data consisting of pairs of an image and a 3D pose, but rather utilize 2D and 3D information from two independent training sources as illustrated in Fig. 1. The first source is accurate 3D motion capture data containing a large number of 3D poses, and is captured in a laboratory setup, e.g., as in the CMU motion capture dataset (CMU, 2014) or the Human3.6M dataset (Ionescu et al., 2014b). Whereas, the second source consists of images with annotated 2D poses as they are provided by 2D human pose datasets, e.g., MPII Human Pose (Andriluka et al., 2014), Leeds Sports Pose (Johnson and Everingham, 2010), and MSCOCO (Lin et al., 2014). Since 2D poses can be manually annotated for images, they do not impose any restriction regarding the environment from where the images are taken. In fact any image from the Internet can be annotated and used. Since both sources are captured independently, we do not know the 3D pose for any training image. In order to bring the two sources together, we map the motion capture data into a normalized 2D pose space to allow for an efficient retrieval based on 2D body joints. Concurrently, we learn a 2D pose estimation model from the 2D images based on convolutional neural networks. During inference, we first estimate the 2D pose and retrieve the nearest 3D poses using an effective approach that is robust to 2D pose estimation errors. We then jointly estimate the projection from the 3D pose space to the image and reconstruct the 3D pose.

A preliminary version of this work was presented in (Yasin et al., 2016). In this work we leverage the recent progress in 2D pose estimation (Toshev and Szegedy, 2014; Iqbal et al., 2017; Carreira et al., 2016; Pishchulin et al., 2016; Wei et al., 2016; Hu and Ramanan, 2016; Insafutdinov et al., 2016; Newell et al., 2016; Bulat and Tzimiropoulos, 2016; Gkioxari et al., 2016; Rafi et al., 2016; Chu et al., 2017), and improve the performance of (Yasin et al., 2016) by a large margin. We further show that with the availability of better 2D pose estimates, the approach (Yasin et al., 2016) can be largely simplified. We extensively evaluate our approach on two popular datasets for 3D pose estimation namely Human3.6M (Ionescu et al., 2014b) and HumanEva (Sigal et al., 2010). We provide an in-depth analysis of the proposed approach. In particular, we analyze the impact of different MoCap datasets, the impact of the similarity of the training and test poses, the impact of the accuracy of the used 2D pose estimator, and also the differences of the skeleton structure between the two training sources. Finally, we also provide qualitative results for images taken from the MPII Human Pose dataset (Andriluka et al., 2014).

2. Related Work

Earlier approaches for 3D human pose estimation from single images (Bo et al., 2008; Mori and Malik, 2006; Bo and Sminchisescu, 2010; Agarwal and Triggs, 2004; Sminchisescu et al., 2005; Agarwal and Triggs, 2006) utilize discriminative methods to learn a mapping from hand-crafted local image features (e.g., HOG, SIFT, etc.) to 3D human pose. Since local features are sensitive to noise, Kostrikov and Gall (2014) proposed an approach based on a 3D pictorial structure model that combines generative and discriminative methods to obtain robustness to noise. For this, regression forests are trained to estimate the probabilities of 3D joint locations and the final 3D pose is inferred by the pictorial structure model. Since inference is performed in 3D, the bounding volume of the 3D pose space needs to be known and the inference requires a few minutes per frame. In addition to the local image features, the approach (Ionescu et al., 2014a) also utilizes body part segmentation with a second order hierarchical pooling process to obtain robust image descriptors. Instead of computing low level image features, the approach (Pons-Moll et al., 2014) uses boolean geometric relationships between body joints to encode body pose appearance. These features are then used to retrieve semantically similar poses from a large corpus of 3D poses.

With the advances in deep learning, more recent approaches learn end-to-end CNNs to regress the 3D joint locations directly from the images (Li and Chan, 2014; Li et al., 2015; Tekin et al., 2016; Rogez and Schmid, 2016; Chen et al., 2016; Zhou et al., 2016a; Moreno-Noguer, 2017; Park et al., 2016; Tekin et al., 2017; Lin et al., 2017; Sun et al., 2017; Pavlakos et al., 2017). In this direction, the work (Li and Chan, 2014) is one of the earliest methods that presents an end-to-end CNN architecture, where a multi-task loss is proposed to simultaneously detect body parts in 2D images and regress their locations in 3D space. In (Li et al., 2015), a max-margin loss is incorporated with a CNN architecture to efficiently model joint dependencies. Similarly, Zhou et al. (2016a) enforce kinematic constraints by introducing a differentiable kinematic function that can be combined with a CNN. The approach (Tekin et al., 2016) uses auto-encoders to incorporate dependencies between body joints and combines them with a CNN architecture to regress 3D poses. Sun et al. (2017) propose a bone-based pose representation and a compositional loss that encodes long range dependencies between body parts and allows efficient 3D pose regression. Approaches for data augmentation have also been proposed in (Rogez and Schmid, 2016) and (Chen et al., 2016) where synthetic training images are generated to enlarge the training data. The approaches (Park et al., 2016; Tekin et al., 2017; Popa et al., 2017) leverage the information about the locations of 2D body joints to aid 3D human pose estimation. While Park et al. (2016) directly use the 2D joint coordinates to regularize the training of a CNN, Tekin et al. (2017) and Popa et al. (2017) use confidence scoremaps of 2D body joints obtained using a CNN as additional features for 3D pose regression. All these approaches demonstrate very good performances for 3D pose estimation, but require a large amount of training data containing pairs of images and ground-truth 3D poses to train deep network architectures. This limits their applicability to the environments of the training data.

Estimating 3D human pose from a given 2D pose by exploiting motion capture data has also been addressed in the literature (Simo-Serra et al., 2012; Ramakrishna et al., 2012; Yasin et al., 2013; Simo-Serra et al., 2013; Wang et al., 2014; Zhou et al.,

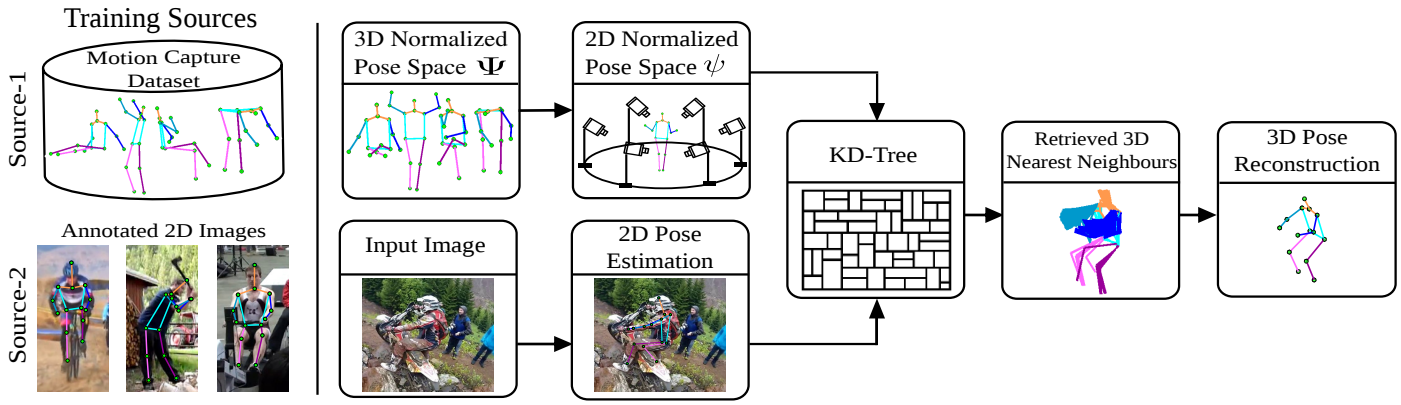

Fig. 1: **Overview.** Our approach utilizes two training sources. The first source is a motion capture database that consists of only 3D poses. The second source is an image database with manually annotated 2D poses. The 3D poses in the motion capture data are normalized and projected to 2D using several virtual cameras. This gives many pairs of 3D-2D poses where the 2D poses are used as features for 3D pose retrieval. The image data is used to learn a 2D pose estimation model based on a CNN. Given a test image, the pose estimation model predicts the 2D pose which is then used to retrieve nearest 3D poses from the normalized 3D pose space. The final 3D pose is then estimated by minimizing the projection error under the constraint that the solution is close to the retrieved poses.

2015; Bogo et al., 2016; Sanzari et al., 2016; Chen and Ramanan, 2017; Lassner et al., 2017; Tome et al., 2017). While early approaches (Ramakrishna et al., 2012; Simo-Serra et al., 2012; Yasin et al., 2013) used manually annotated 2D joint locations, Simo-Serra et al. (2013) and Wang et al. (2014) proposed one of the first approaches that estimate the 3D pose from estimated 2D poses. With the progress in 2D pose estimation methods (Toshev and Szegedy, 2014; Pishchulin et al., 2016; Carreira et al., 2016; Iqbal et al., 2017; Wei et al., 2016; Hu and Ramanan, 2016; Insafutdinov et al., 2016; Newell et al., 2016; Bulat and Tzimiropoulos, 2016; Gkioxari et al., 2016; Rafi et al., 2016; Chu et al., 2017), the number of approaches in this category also rose (Zhou et al., 2015; Bogo et al., 2016; Chen and Ramanan, 2017; Lassner et al., 2017; Tome et al., 2017). All these approaches have the benefit that they do not require training data containing images with annotated 3D poses, but rather only utilize pose data to build their models.

In (Yasin et al., 2013), ground-truth 2D pose is used in the first frame and tracked in a video. A nearest neighbor search is then performed to obtain the nearest 3D poses. The approach (Ramakrishna et al., 2012) constructs a sparse representation of 3D body pose using a MoCap dataset and fits it to manually annotated 2D joint positions. While Wang et al. (2014) extend the approach to handle estimated poses from an off-the-shelf 2D pose estimator (Yang and Ramanan, 2011), Du et al. (2016) extend it to leverage temporal information in video data. The approaches (Simo-Serra et al., 2012, 2013) use the information about the 2D body joints to constrain the search space of 3D poses. In (Simo-Serra et al., 2012) an evolutionary algorithm is proposed to sample poses from the pose space that correspond to the estimated 2D joint positions. This set is then exhaustively evaluated according to some anthropometric constraints. The approach is extended in (Simo-Serra et al., 2013) such that the 2D pose estimation and 3D pose estimation are iterated. In contrast to (Ramakrishna et al., 2012; Wang et al., 2014; Simo-Serra et al., 2012), the approach (Simo-Serra et al., 2013) deals with 2D pose estimation errors.

An expectation maximization algorithm is presented in (Zhou et al., 2015) to estimate 3D poses from monocular videos. Additional smoothness constraints are used to exploit the temporal information in videos. In addition to the 3D pose, Bogo et al. (2016) also estimate the 3D shape of the person. The approach exploits a high-quality 3D human body model and fits it to estimated 2D joints using an energy minimization objective. The approach is improved further in (Lassner et al., 2017) by introducing an extra fitting objective and generating additional training data. In (Chen and Ramanan, 2017) a non-parametric nearest neighbor model is used to retrieve 3D exemplars that minimize the reprojection error from the estimated 2D joint locations. Tome et al. (2017) propose a probabilistic 3D pose model and combine it with a multi-staged CNN, where the CNN incorporates evidences from the 2D body part locations and projected 3D poses to sequentially improve 2D joint predictions which in turn also results in better 3D pose estimates. Other approaches also learn deep neural networks to directly regress 3D pose from 2D joint information (Moreno-Noguer, 2017; Martinez et al., 2017). Martinez et al. (2017) propose a deep neural network with residual connections to directly regress 3D pose from 2D pose as input. Moreno-Noguer (2017), on the other hand, proposes to first encode 3D pose using an Euclidean distance matrix formulation that implicitly incorporates body joint relations and allows to regress 3D poses in form of a distance matrix.

Action specific priors learned from motion capture data have also been proposed for 3D pose tracking (Urtasun et al., 2006; Andriluka et al., 2010). These approaches, however, are more constrained by assuming that the type of motion is known in advance and therefore cannot deal with a large and diverse pose dataset.

3. Overview

In this work, we propose an approach to estimate the 3D pose from an RGB image. Since annotating 2D images with accu-

rate 3D pose data is infeasible and obtaining 3D body pose data in unconstrained scenarios using sophisticated MoCap systems is impractical, our approach does not require that the training data consists of images annotated with 3D pose. In contrast, we use two independent sources of training data. The first source contains only 3D poses captured by a motion capture system. Such data is publicly available in large numbers and can also be captured in controlled indoor environments. The second source contains unconstrained images with annotated 2D poses, which are also abundantly available (Andriluka et al., 2014; Lin et al., 2014) and can be easily annotated by humans. Apart from the requirement that the MoCap data contains poses that are related to the activities we are interested in, we do not assume any correspondence between the two sources. We therefore preprocess both sources separately as shown in Fig. 1. From the image data, we learn a CNN based 2D pose estimation model to predict 2D poses from images. This will be described in Section 4. The MoCap data is processed to efficiently retrieve 3D poses that could correspond to a 2D pose. This part is discussed in Section 5.1. We then estimate the 3D pose by minimizing the projection error under the constraint that the solution is close to the retrieved poses (Section 5.2). The source code of the approach is publicly available.[1]

4. 2D Pose Estimation

In this work, we use the convolutional pose machines (CPM) (Wei et al., 2016) for 2D pose estimation, but other CNN architectures, e.g. stacked hourglass (Newell et al., 2016) or multi-context attention models (Chu et al., 2017), could be used as well. Given an image I, we define the 2D pose of the person as $\mathbf{x} = \{x_j\}_{j\in\mathcal{J}}$, where $x_j \in \mathbb{R}^2$ denotes the 2D pixel coordinate of body joint j, and $\mathcal{J}$ is the set of all body joints. CPM consists of a multi-staged CNN architecture, where each stage $t \in \{1 \dots T\}$ produces a set of confidence scoremaps $\mathbf{s_t} = \{s_t^j\}_{j\in\mathcal{J}}$, where $s_t^j \in \mathbb{R}^{w\times h}$ is the confidence score map of body joint j at stage t, and w and h are the width and the height of the image, respectively. Each stage of the network sequentially refines the 2D pose estimates by utilizing the output of the preceding stage and also the features extracted from the raw input image. The final 2D pose $\mathbf{x}$ is obtained as

$$\mathbf{x} = \underset{\mathbf{x}'=\{x'_j\}_{j\in\mathcal{J}}}{\arg\max} \sum_{j\in\mathcal{J}} s_T^j(x'_j). \tag{1}$$

.

In our experiments we will show that training the network on publicly available dataset for 2D pose estimation in-the-wild, such as the MPII Human Pose dataset (Andriluka et al., 2014), is sufficient to obtain competitive results with our proposed method.

5. 3D Pose Estimation

While the 2D pose estimation model is trained using the images annotated with 2D poses as shown in Fig. 1, we now explain a method that utilizes the 3D poses from the second source to estimate the 3D pose from an image. Since both sources do not have any correspondence, we first have to establish correspondences between the 2D and 3D poses. For this, an estimated 2D pose is used as a query for 3D pose retrieval (Section 5.1). The retrieved 3D poses, however, contain many incorrect poses due to 2D-3D ambiguities, differences of the skeletons between the two training sources, and errors in the estimated 2D pose. It is therefore required to fit the 3D poses to the 2D observations. This is discussed in Section 5.2.

5.1. 3D Pose Retrieval

In order to efficiently retrieve 3D poses for a 2D pose query, we first preprocess the MoCap data by discarding the body location and orientation for each pose. This is achieved by applying the inverse transformation of the rigid transformation of the root joint, which is provided by the MoCap dataset, to all joints. After the transformation, the root joint is located at the origin of the coordinate system and the orientation of the pose is aligned with the x-axis. We denote the normalized 3D pose space with $\mathbf{\Psi}$, where $\mathbf{X} \in \mathbf{\Psi}$ denotes a normalized 3D pose. Similar to (Yasin et al., 2013), we project the normalized 3D poses $\mathbf{X} \in \mathbf{\Psi}$ to 2D using 120 virtual camera views with orthographic projection. We use elevation angles ranging between 0 and 60 degree and azimuth angles spanning 360 degrees, both sampled uniformly with a step size of 15 degrees. The projected 2D poses are further normalized by scaling such that the y-coordinates of the joints are within the range of $[-1, 1]$. The normalized 2D space does not depend on a specific coordinate system or a camera model and is denoted as ψ. This step is illustrated in Fig. 1. During inference, given a 2D pose estimated by the approach explained in Section 4, we first normalize it according to ψ, i.e., we translate and scale the pose such that the y-coordinates of the joints are within the range of $[-1, 1]$. The normalized 2D pose is then used to retrieve 3D poses. We use the average Euclidean distance between the joint positions to measure the distance between two normalized 2D poses. Finally, we use a kd-tree (Krüger et al., 2010) to efficiently retrieve K-nearest neighbors in ψ where the retrieved normalized 3D poses are the corresponding poses in $\mathbf{\Psi}$.

5.2. 3D Pose Estimation

In order to obtain the 3D pose $\mathbf{X}$, we have to estimate the unknown projection $\mathcal{M}$ from the normalized pose space $\mathbf{\Psi}$ to the image. To this end, we minimize the energy

$$E(\mathbf{X}, \mathcal{M}) = E_p(\mathbf{X}, \mathcal{M}) + \alpha E_r(\mathbf{X}) \tag{2}$$

over $\mathbf{X}$ and $\mathcal{M}$. The parameter α defines the weighting between the two terms E_p and E_r.

The first term $E_p(\mathbf{X}, \mathcal{M})$ measures the projection error of the 3D pose $\mathbf{X}$ and the projection $\mathcal{M}$:

$$E_p(\mathbf{X}, \mathcal{M}) = \left(\sum_{j\in\mathcal{J}} \| \mathcal{M}\left(X_j\right) - x_j \|^2 \right)^{\frac{1}{2}}, \tag{3}$$

where X_j is the 3D joint position of the unknown 3D pose and x_j is the joint position of the predicted 2D pose.

[1] http://pages.iai.uni-bonn.de/iqbal_umar/ds3dpose/

The second term ensures that the pose $\mathbf{X}$ is close to the retrieved 3D poses $\mathbf{X}^{\mathsf{k}}$:

$$E_r(\mathbf{X}) = \sum_{\mathsf{k}} \left(\sum_{j \in \mathcal{J}} \|X_j^{\mathsf{k}} - X_j\|^2 \right)^{\frac{1}{2}}. \quad (4)$$

The energy function (2) differs from the function that was proposed in (Yasin et al., 2016) in several ways. The energy function used in (Yasin et al., 2016) contains an additional term that enforces anthropometric constraints, it weights the retrieved 3D poses, and optimizes the energy in addition over five different joint sets. While these extensions improve the 3D pose estimation in case of noisy 2D pose estimates obtained by a pictorial structure model, we found that these extensions have a negligible impact on the accuracy if the 2D pose estimates are more accurate due to the used CNN for 2D pose estimation.

Minimizing the energy $E(\mathbf{X}, \mathcal{M})$ (2) over the continuous parameters $\mathcal{M}$ and $\mathbf{X}$ would be expensive. We therefore propose an approximate solution where we first estimate the projection $\mathcal{M}$ only. For the projection, we consider that the intrinsic parameters are provided and only estimate the global translation and orientation. The projection $\hat{\mathcal{M}}$ is estimated by minimizing

$$\hat{\mathcal{M}} = \arg\min_{\mathcal{M}} \left\{ \sum_{\mathsf{k}=1}^{\mathsf{K}} E_p(\mathbf{X}^{\mathsf{k}}, \mathcal{M}) \right\} \quad (5)$$

using non-linear gradient optimization with trust-region-reflective algorithm. We initialize the camera translation by $[0, 0, -Hf/h]$, where H is the mean height of the retrieved nearest neighbours and h corresponds to the height of the estimated 2D pose. In our experiments, we will also evaluate the case when the camera orientation and translation are also known. In this case, the projection $\mathcal{M}$ reduces to a rigid transformation of the 3D poses $\mathbf{X}$ from the normalized pose space $\mathbf{\Psi}$ to the camera coordinate system.

Given the estimated projection $\hat{\mathcal{M}}$, we minimize

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \left\{ E(\mathbf{X}, \hat{\mathcal{M}}) \right\} \quad (6)$$

to obtain the 3D pose $\mathbf{X}$.

The dimensionality of $\mathbf{X}$ can be reduced by applying PCA to the retrieved 3D poses $\mathbf{X}^{\mathsf{k}}$. Reducing the dimensions of $\mathbf{X}$ helps to decrease the optimization time without loss in accuracy, as we will show in the experiments.

6. Experiments

We evaluate the proposed approach on two publicly available datasets, namely Human3.6M (Ionescu et al., 2014b) and HumanEva-I (Sigal et al., 2010). Both datasets provide accurate 3D poses for each image and camera parameters. For all cases, 2D pose estimation is performed by convolutional pose machines (Wei et al., 2016) trained on the MPII Human Pose dataset (Andriluka et al., 2014) without any fine-tuning, unless it is stated otherwise.

6.1. Evaluation on Human3.6M Dataset

For evaluation on the Human3.6M dataset, a number of protocols have been proposed in the literature. The protocol originally proposed for the Human3.6M dataset (Ionescu et al., 2014b), which we denote by *Protocol-III*, uses the annotated bounding boxes and the training data only from the action class of the test data. This simplifies the task due to the small pose variations for a single action class and the known person bounding box. Other protocols have been therefore proposed in (Kostrikov and Gall, 2014) and (Bogo et al., 2016). In order to compare with other existing approaches, we report results for all three protocols (Kostrikov and Gall, 2014; Bogo et al., 2016) and (Ionescu et al., 2014b).

6.1.1. Human3.6M Protocol-I

Protocol-I, which was proposed by (Kostrikov and Gall, 2014), is the most unconstrained protocol. It does not make any assumption about the location and activity labels during testing, and the training data comprises all action classes. The training set consists of six subjects (S1, S5, S6, S7, S8 and S9), whereas the testing is performed on every 64^{th} frame taken from the sequences of S11. For evaluation, we use the *3D pose error* as defined in (Simo-Serra et al., 2012). The error measures the accuracy of the relative pose up to a rigid transformation. To this end, the estimated skeleton is aligned to the ground-truth skeleton by a rigid transformation and the average 3D Euclidean joint error is measured after alignment. The body skeleton consists of 14 body joints namely head, neck, ankles, knees, hips, wrists, elbows, and shoulders. In order to comply with the protocol, we do not use any ground truth bounding boxes, but estimate them using an off-the-shelf person detector (Ren et al., 2015). The detected bounding boxes are used by the convolutional pose machines for 2D pose estimation. We consider two sources for the motion capture data, namely the Human3.6M and the CMU motion capture dataset.

We first evaluate the impact of the parameters of our approach and the impact of different MoCap datasets. We then compare our approach with the state-of-the-art and evaluate the impact of the 2D pose estimation accuracy.

Nearest Neighbors. The impact of the number of nearest neighbors K used during 3D pose reconstruction is evaluated in Fig. 2. Increasing the number of nearest neighbors improves 3D pose estimation. This, however, also increases the reconstruction time. In the rest of this paper, we use a default value of $K = 256$ that provides a good trade-off between accuracy and run-time. The reconstruction of the 3D pose with $K = 256$ for a single image takes roughly 0.6 seconds[2]. We can see that using the CMU MoCap dataset results in a higher error as compared to the Human3.6M dataset. We will evaluate the impact of different MoCap datasets in more details later in this section.

PCA. PCA can be used to reduce the dimension of $\mathbf{X}$. While in (Yasin et al., 2016) a fixed number of principal components

[2]Measured on a 3.4GHz Intel processor using only one core.

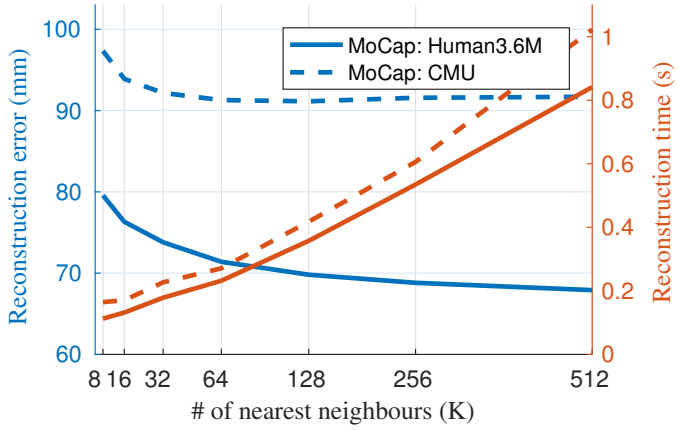

Fig. 2: Impact of the number of nearest neighbors K.

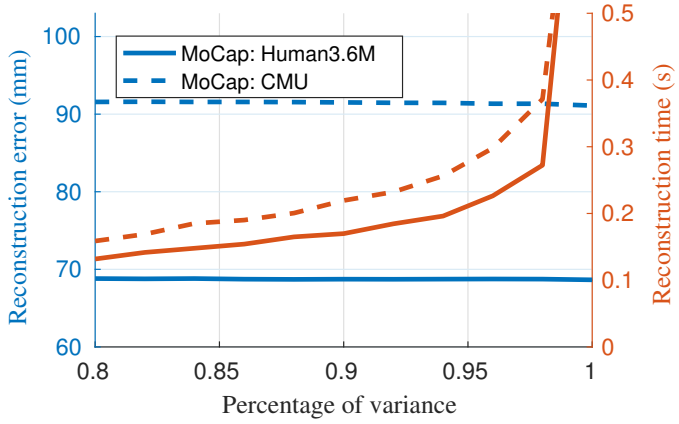

Fig. 3: Impact of PCA. The number of principle components are selected based on the minimum number of components that explain a given percentage of variation. The x-axis corresponds to the threshold for the cumulative amount of variation.

is used, we use a more adaptive approach and set the number of principal components based on the captured variance. The number of principal components therefore varies for each image. The impact of the threshold on the minimum amount of variation can be seen in Fig. 3. If the threshold is within a reasonable range, i.e. between 0.8 and 1, the accuracy is barely reduced while the runtime decreases significantly compared to 1, i.e. without PCA. In this work, we use the minimum number of principle components that explain at least 80% of the variance of the retrieved 3D poses $\mathbf{X}^k$.

Energy Terms. The impact of the weight α in (2) is reported in Fig. 4. If $\alpha = 0$, the term E_r is ignored and the error is very high. This is expected since E_r constrains the possible solution while E_p ensures that the estimated 3D pose projects onto the estimated 2D pose. In our experiments, we use $\alpha = 1$.

Impact of MoCap dataset size. We evaluate the impact of the size of the MoCap dataset in Fig. 5. In order to sub-sample the dataset, which consists of 469K 3D poses, we use a greedy approach that starts with an empty set and gradually adds a new pose if the distance to any previously selected pose is larger or equal to a threshold. Otherwise, the pose is discarded. Depending on the threshold (320mm, 160mm, 80mm, 40mm, 20mm), the dataset is reduced to 11K, 48K, 111K, 208K, and 329K poses, respectively. Using the entire 469K 3D poses of the Human3.6M training set as motion capture data results in a 3D pose error of 68.8mm. Reducing the size of the MoCap data to 329K reduces the error to 66.85mm. The reduction of the error is expected since the sub-sampling removes duplicates and very similar poses that do not provide any additional information when they are retrieved. However, decreasing the size of the MoCap dataset even further degenerates the performance. In the rest of our experiments, we use the MoCap dataset from Human3.6M with 329K 3D poses, where a threshold of 20mm is used to remove similar poses. While the runtime of the approach is linear with respect to the number of nearest neighbors (K) as it can be observed in Fig. 2, the sub-sampling of the MoCap dataset has a minimal impact on the runtime since the computational complexity of 3D pose retrieval is logarithmic with respect to the dataset size and the dataset size does not affect the energy function (2), in contrast to K.

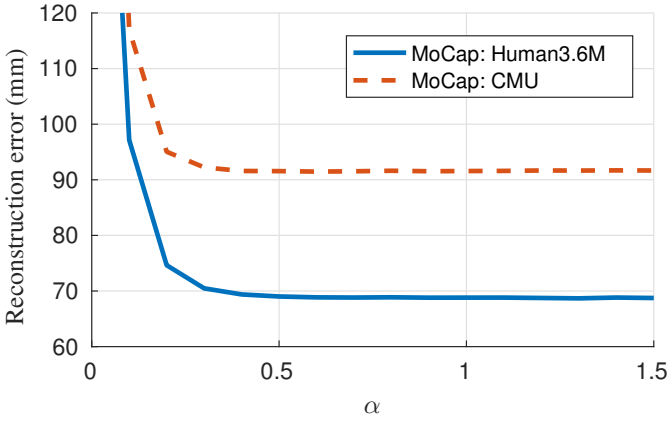

Fig. 4: Impact of α.

CMU Motion Capture Dataset. Our approach does not require images that are annotated by 3D poses but uses MoCap data as a second training source. We therefore also evaluate the proposed method using the CMU MoCap dataset (CMU, 2014) to construct the 3D pose space. We downsample the CMU dataset from 120Hz to 30Hz and use only one third of the 3D poses, resulting in 360K poses. We remove similar poses using the same threshold (20mm) as used for Human3.6M, which results in a final MoCap dataset with 303K 3D poses. Fig. 6 compares the pose estimation accuracy using both datasets, while the results for each activity can be seen in Tab. 1. As expected the error is higher due to the differences of the datasets.

To analyze the impact of the MoCap data in more detail, we

MoCap data	Direction	Discuss	Eating	Greeting	Phoning	Posing	Purchases	Sit	SitDown
Human3.6M	59.5	52.4	75.5	67.0	58.8	64.9	58.2	68.4	89.7
Human3.6M \ Activity	61.2	52.3	92.6	70.2	61.1	66.5	59.3	85.6	122.2
Human3.6M ∈ Activity	68.8	57.6	70.8	73.7	62.9	66.7	63.4	73.4	99.4
Human3.6M + GT 3D Poses	52.9	45.7	59.9	60.1	50.4	54.1	51.6	56.3	71.7
CMU	73.3	64.7	95.9	80.2	85.7	81.8	77.1	110.5	138.8

MoCap data	Smoking	Photo	Waiting	Walk	WalkDog	WalkTogether	Mean	Median
Human3.6M	73.0	88.5	67.7	52.1	73.0	54.1	66.9	61.5
Human3.6M \ Activity	74.8	92.6	72.4	64.5	74.6	69.0	74.5	67.3
Human3.6M ∈ Activity	74.8	89.5	77.4	49.3	70.8	55.9	70.4	65.3
Human3.6M + GT 3D Poses	64.2	69.2	60.4	47.8	60.6	44.9	56.7	51.3
CMU	100.9	95.3	90.6	82.9	87.6	91.3	91.0	83.3

Table 1: Impact of the MoCap dataset. While for Human3.6M \ Activity we removed all poses from the dataset that correspond to the activity of the test sequence, Human3.6M ∈ Activity only contains the poses of the activity of the test sequence. For Human3.6M + GT 3D Poses, we include the ground-truth 3D poses of the test sequences to the MoCap dataset.

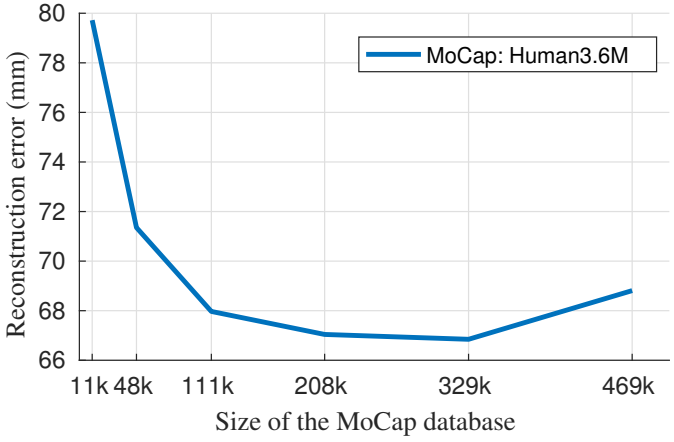

Fig. 5: Impact of the size of the MoCap dataset.

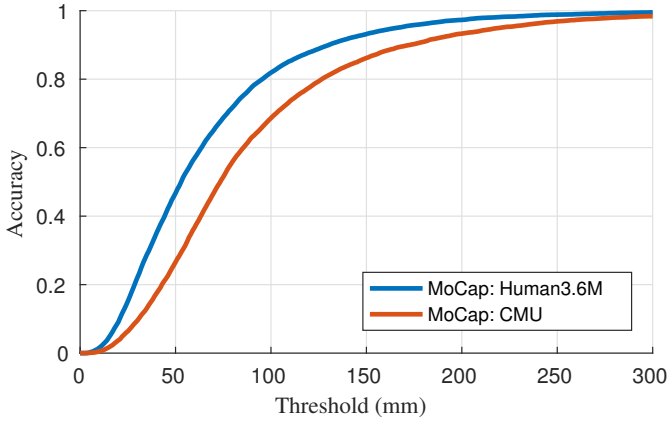

Fig. 6: Comparison of 3D pose error using different MoCap datasets. The plot shows the percentage of estimated 3D poses with an error below a specific threshold.

have evaluated the pose error for various modifications of the MoCap data in Tab. 1. First, we remove all poses of an activity from the MoCap data and evaluate the 3D pose error for the test images corresponding to the removed activity. The error increases since the dataset does not contain poses related to the removed activity anymore. While the error still stays comparable for many activities, e.g. Direction, Discussion, etc., a significant increase in error can be seen for activities that do not share similar poses with other activities, e.g. SitDown. However, even if all poses related to the activity of the test images are removed, the results are still good and better compared to the CMU dataset. This indicates that the error increase for the CMU dataset cannot only be explained by the difference of poses, but also other factors like different motion capture setups seem to influence the result. We will investigate the impact of the difference of the skeleton structure between two datasets in Section 6.2.

We also evaluate the case when the MoCap dataset contains only the poses of a specific activity. This also results in an increased mean pose estimation error and shows that having a diverse MoCap dataset is helpful to obtain good performance. Finally, we also report the error when the 3D poses of the test sequences are added to the MoCap dataset. In this case, the mean error is reduced from 66.9mm to 56.7mm.

Comparison with State-of-the-art. Tab. 2 compares the performance of the proposed method with the state-of-the-art approaches (Kostrikov and Gall, 2014; Yasin et al., 2016; Rogez and Schmid, 2016; Chen and Ramanan, 2017; Moreno-Noguer, 2017; Tome et al., 2017; Zhou et al., 2017; Sun et al., 2017) using both MoCap datasets. The proposed approach reduces the 3D pose error reported in (Yasin et al., 2016) from 108.3mm to 66.9mm when using the Human3.6M MoCap dataset. A similar decrease in error can also be seen for the CMU dataset (124.8mm vs. 91.0mm). The main improvement compared to (Yasin et al., 2016) stems from the better 2D pose estimation model. Our approach also outperforms the recent methods (Chen and Ramanan, 2017; Moreno-Noguer, 2017; Tome et al., 2017). While Moreno-Noguer (2017) utilizes 3D

Method	Direction	Discuss	Eating	Greeting	Phoning	Posing	Purchases	Sit	Sit Down
Kostrikov and Gall (2014)	-	-	-	-	-	-	-	-	-
Yasin et al. (2016)	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	170.8
Rogez and Schmid (2016)	-	-	-	-	-	-	-	-	-
Chen and Ramanan (2017)	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7	195.6
Moreno-Noguer (2017)	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5
Tome et al. (2017)	-	-	-	-	-	-	-	-	-
Zhou et al. (2017)	47.9	48.8	52.7	55.0	56.8	49.0	45.5	60.8	81.1
Sun et al. (2017)*	42.1	44.3	45.0	45.4	51.5	43.2	41.3	59.3	73.3
Ours	59.5	52.4	75.5	67.0	58.8	64.9	58.2	68.4	89.7
(MoCap from CMU dataset)									
Yasin et al. (2016)	102.8	80.4	133.8	120.5	120.7	98.9	117.3	150.0	182.6
Ours	73.3	64.7	95.9	80.2	85.7	81.8	77.1	110.5	138.8

Method	Smoking	Photo	Waiting	Walk	WalkDog	WalkTogehter	Mean	Median
Kostrikov and Gall (2014)	-	-	-	-	-	-	115.7	-
Yasin et al. (2016)	108.2	142.5	86.9	92.1	165.7	102.0	108.3	-
Rogez and Schmid (2016)	-	-	-	-	-	-	88.1	-
Chen and Ramanan (2017)	83.5	93.3	71.2	55.7	85.9	62.5	82.7	69.1
Moreno-Noguer (2017)	75.8	92.6	69.6	71.5	78.0	73.2	74.0	-
Tome et al. (2017)	-	-	-	-	-	-	70.7	-
Zhou et al. (2017)	53.7	65.5	51.6	50.4	54.8	55.9	55.3	-
Sun et al. (2017)*	51.0	53.0	44.0	38.3	48.0	44.8	48.3	-
Ours	73.0	88.5	67.7	52.1	73.0	54.1	66.9	61.5
(MoCap from CMU dataset)								
Yasin et al. (2016)	135.6	140.1	104.7	111.3	167.0	116.8	124.8	-
Ours	100.9	95.3	90.6	82.9	87.6	91.3	91.0	83.3

Table 2: Comparison with the state-of-the-art on the Human3.6M dataset using *Protocol-I*. *additional ground-truth information is used.

poses from Human3.6M as training data, Tome et al. (2017) use the 2D pose data from Human3.6M to learn a multistage deep CNN architecture for 2D pose estimation. We on the other hand do not use any 2D or 3D pose information for training and only utilize a pre-trained model trained on the MPII Human Pose Dataset (Andriluka et al., 2014) for 2D pose estimation. We also compare our performance with the most recent approaches (Zhou et al., 2017; Sun et al., 2017). These approaches perform better than our method. However, they use pairs of images and 3D poses to learn deep CNN models while our approach does not require 3D pose annotations for images. Moreover, in contrast to our method, none of the aforementioned approaches have shown that they can handle MoCap data that is from a different source than the test data.

Impact of 2D Pose. We also investigate the impact of the accuracy of the estimated 2D poses. If we initialize the approach with the 2D ground-truth poses, the 3D pose error is significantly reduced as shown in Tab. 3. This indicates that the 3D pose error can be further reduced by improving the used 2D pose estimation method. We also report the 3D pose error when both 3D and 2D ground-truth poses are available. In this case the error reduces even further which shows the potential of further improvements for the proposed method. We also compare our approach to (Yasin et al., 2016) and (Chen and Ramanan, 2017), which also report the accuracy for ground-truth 2D poses.

6.1.2. Human3.6M Protocol-II

The second protocol, *Protocol-II*, has been proposed in (Bogo et al., 2016). The dataset is split using five subjects (S1, S5, S6, S7, S8) for training and two subjects (S9 and S11) for testing. We follow (Lassner et al., 2017) and perform testing on every 5^{th} frame of the sequences from the frontal camera (cam-3) and trial-1 of each activity. The evaluation is performed in the same way as in *Protocol-I* with a body skeleton consisting of 14 joints. In contrast to *Protocol-I*, the *ground-truth bounding boxes* are, however, used during testing. Tab. 4 reports the comparison of the proposed method with the state-of-the-art approaches (Akhter and Black, 2015; Ramakrishna et al., 2012; Zhou et al., 2015; Bogo et al., 2016; Lassner et al., 2017; Tome et al., 2017; Moreno-Noguer, 2017; Martinez et al., 2017; Pavlakos et al., 2017; Tekin et al., 2017). While our approach achieves comparable results to (Akhter and Black, 2015; Ramakrishna et al., 2012; Zhou et al., 2015; Bogo et al., 2016; Lassner et al., 2017; Tome et al., 2017; Moreno-Noguer, 2017), more recent approaches (Martinez et al., 2017; Pavlakos et al., 2017; Tekin et al., 2017) perform better. The approaches (Pavlakos et al., 2017; Tekin et al., 2017), however, use pairs of images and 3D poses as training data, and the approach (Martinez et al., 2017) uses more recent improvements in the deep neural network architectures with exhaustive parameter selection to directly regress 3D pose from 2D joint information. Whereas, our approach does not require dataset specific training and therefore requires less supervision and can generalize better to different scenarios.

6.1.3. Human3.6M Protocol-III

The third protocol, *Protocol-III*, is the most commonly used protocol for Human3.6M. Similar to *Protocol-II*, the dataset is split by using subjects S1, S5, S6, S7 and S8 for training and subjects S9 and S11 for testing. The sequences are downsampled from the original frame-rate of 50fps to 10fps, and testing is performed on the sequences from all cameras and trials. The evaluation is performed without a rigid transformation, but both the ground-truth and estimated 3D poses are centered with respect to the root joint. We therefore have to use the provided camera parameters such that the estimated 3D pose is in the coordinate system of the camera. The training and testing is often performed on the same activity. However, some recent approaches also report results by training only once for all activities. In this work, we report results under both settings. In this protocol, a body skeleton with 17 joints is used and the ground-truth bounding boxes are used during testing. Note that even though the 3D poses contain 17 joints, we still use the 2D poses with 14 joints for nearest neighbor retrieval and only use the corresponding joints for optimizing objective (2). Tab. 5 provides a detailed comparison of the proposed approach with the state-of-the-art methods.

Finally, we present some qualitative results in Fig. 7. As it can be seen, our approach shows very good performance even for highly articulated poses and under severe occlusions.

6.2. Evaluation on HumanEva-I Dataset

We follow the same protocol as described in (Simo-Serra et al., 2013; Kostrikov and Gall, 2014) and use the provided training data to train our approach while using the validation data as test set. As in (Simo-Serra et al., 2013; Kostrikov and Gall, 2014), we report our results on every 5^{th} frame of the sequences *walking* (A1) and *jogging* (A2) for all three subjects (S1, S2, S3) and camera C1. The 3D pose error is computed as in *Protocol-I* for the Human3.6M dataset.

We perform experiments with the 3D pose data from the HumanEva and CMU MoCap datasets. For HumanEva, we use the entire 49K 3D poses of the training data as MoCap dataset. Since the joint positions of the skeleton used for HumanEva differs from the joint annotations that are provided by the MPII Human Pose dataset, we fine-tune the 2D pose estimation model on the HumanEva dataset using the provided 2D pose data. For fine-tuning, we run 500 iterations with a learning rate of 0.00008.

We also have to adapt the skeleton structure of the CMU dataset to the skeleton structure of the HumanEva dataset. As in (Yasin et al., 2016), we re-target the 3D poses in the CMU dataset to the skeleton of the HumaEva dataset using linear regression. For this, we first scale normalize the 3D poses in both datasets such that the height of each pose is equal to 1000mm. For each pose in the CMU dataset, we then search the nearest neighbor in the HumanEva dataset. For computing the distance between poses, we only use the joints that are common in both datasets. The pairs of poses that have a distance greater than 5mm are discarded and the remaining pairs are used to learn a linear mapping between the skeletons of the two datasets.

We analyze the impact of the difference between the skeletons of both datasets in Tab. 6. Using HumanEva as MoCap

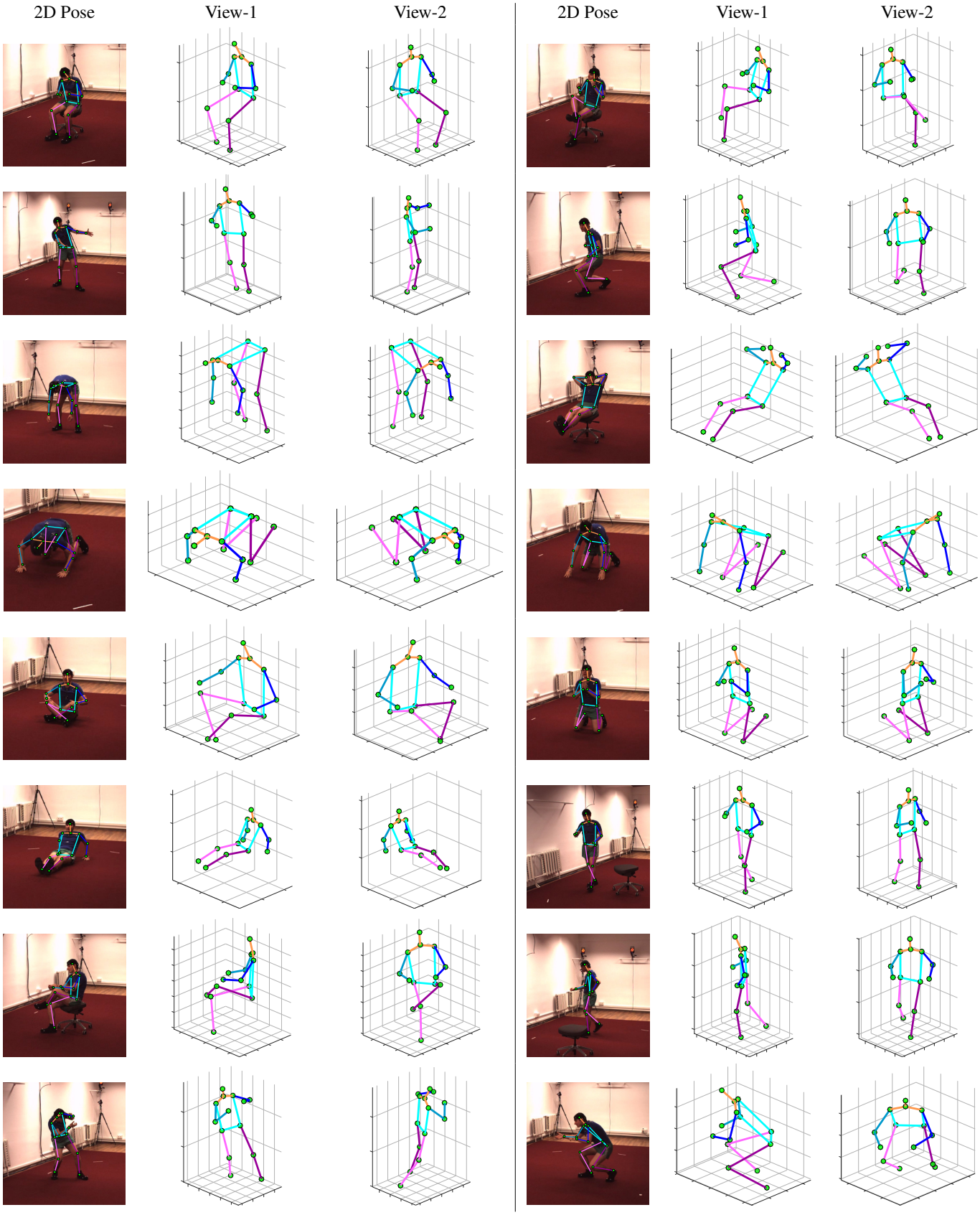

Fig. 7: Some qualitative results from the Human3.6M (Ionescu et al., 2014b) dataset.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Ours	59.5	52.4	75.5	67.0	58.8	64.9	58.2	68.4	89.7
Ours + GT 2D	51.9	45.3	62.4	55.7	49.2	56.0	46.4	56.3	76.6
Ours + GT 2D + GT 3D	40.9	35.3	41.6	44.3	36.6	43.7	38.0	40.3	53.4
Yasin et al. (2016) + GT 2D	60.0	54.7	71.6	67.5	63.8	61.9	55.7	73.9	110.8
Chen and Ramanan (2017) + GT 2D	53.3	46.8	58.6	61.2	56.0	58.1	48.9	55.6	73.4
(MoCap from CMU dataset)									
Ours + GT 2D	67.8	58.7	90.3	72.1	78.2	75.7	71.9	103.2	132.8

Method	Smoke	Photo	Wait	Walk	WalkDog	WalkTogether	Mean	Median
Ours	73.0	88.5	67.7	52.1	73.0	54.1	66.9	61.5
Ours + GT 2D	58.8	79.1	58.9	35.6	63.4	46.3	56.1	51.9
Ours + GT 2D + GT 3D	44.2	56.6	45.9	26.9	45.8	31.4	41.6	39.1
Yasin et al. (2016) + GT 2D	78.9	96.9	67.9	47.5	89.3	53.4	70.5	-
Chen and Ramanan (2017) + GT 2D	60.3	76.1	62.2	35.8	61.9	51.1	57.5	51.9
(MoCap from CMU dataset)								
Ours + GT 2D	91.3	91.6	84.7	70.9	81.2	76.7	83.7	75.6

Table 3: Impact of the 2D pose estimation accuracy. GT 2D denotes that the ground-truth 2D pose is used. GT 3D denotes that the 3D poses of the test images are added to the MoCap dataset as in Tab. 1.

dataset results in a 3D pose error of 31.5mm, whereas using CMU as MoCap dataset increases the error significantly to 80.0mm. Re-targeting the skeletons of the CMU dataset to the skeleton of HumanEva reduces the error from 80.0mm to 50.5mm, and re-targeting the skeleton of HumanEva to CMU increases the error from 31.5mm to 58.4mm. This shows that the difference of the skeleton structure between the two sources can have a major impact on the evaluation. This is, however, not an issue for an application where the MoCap dataset defines the skeleton structure.

We also compare our approach with the state-of-the-art approaches (Kostrikov and Gall, 2014; Wang et al., 2014; Radwan et al., 2013; Simo-Serra et al., 2013, 2012; Bo and Sminchisescu, 2010; Yasin et al., 2016; Popa et al., 2017; Martinez et al., 2017; Pavlakos et al., 2017; Moreno-Noguer, 2017) in Tab. 7. Our method is competitive to all methods except of the very recent approaches (Moreno-Noguer, 2017; Martinez et al., 2017; Pavlakos et al., 2017) that use more supervision or more recent CNN architectures. In particular, the ability to use MoCap data from a different source than the test data has so far not addressed by other works. This experimental protocol, however, is essential to assess the generalization capabilities of different methods.

Finally, we present qualitative results for a few realistic images taken from the MPII Human Pose dataset (Andriluka et al., 2014) in Fig. 8. The results show that the proposed approach generalizes very well to complex unconstrained images.

7. Conclusion

In this work, we have proposed a novel dual-source method for 3D human pose estimation from monocular images. The first source is a MoCap dataset with 3D poses and the other source are images with annotated 2D poses. Due to the separation of the two sources, our approach needs less supervision compared to approaches that are trained from images annotated with 3D poses, which is difficult to acquire under real conditions. The proposed approach therefore presents an important step towards accurate 3D pose estimation in unconstrained images. Compared to the preliminary work, the proposed approach does not require to train dataset specific models and can generalize across different scenarios. This is achieved by utilizing the strengths of recent 2D pose estimation methods and combining them with an efficient and robust method for 3D pose retrieval. We have performed a thorough experimental evaluation and demonstrated that our approach achieves competitive results in comparison to the state-of-the-art, even when the training data are from very different sources.

8. Acknowledgments

The work has been financially supported by the DFG projects GA 1927/5-1 and We 1945/11-1 (DFG Research Unit FOR 2535 Anticipating Human Behavior) and the ERC Starting Grant ARCA (677650).

References

Agarwal, A., Triggs, B., 2004. 3d human pose from silhouettes by relevance vector regression, in: IEEE Conference on Computer Vision and Pattern Recognition.

Agarwal, A., Triggs, B., 2006. Recovering 3d human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 44–58.

Akhter, I., Black, M.J., 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition.

Fig. 8: Some qualitative results from the MPII Human Pose Dataset.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sit
Akhter and Black (2015)	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7
Ramakrishna et al. (2012)	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6
Zhou et al. (2015)	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1
Bogo et al. (2016)	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3
Moreno-Noguer (2017)	64.1	76.6	70.6	80.8	93.0	96.3	74.0	65.5	87.9
Lassner et al. (2017)	-	-	-	-		-	-	-	-
Tome et al. (2017)	-	-	-	-		-	-	-	-
Martinez et al. (2017)	44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9	66.3
Pavlakos et al. (2017)	-	-	-	-		-	-	-	-
Tekin et al. (2017)	-	-	-	-		-	-	-	-
Ours	75.3	75.8	70.9	92.8	89.0	101.5	78.1	61.4	97.9
(MoCap from CMU dataset)									
Ours	89.7	88.6	94.1	101.1	106.3	104.1	85.9	81.0	121.7

	SitDown	Smoking	Waiting	WalkDog	Walk	WalkTogether	Mean	Median
Akhter and Black (2015)	173.7	177.8	181.9	176.2	198.6	192.7	181.1	158.1
Ramakrishna et al. (2012)	175.6	160.4	161.7	150.0	174.8	150.2	157.3	136.8
Zhou et al. (2015)	137.5	106.0	102.2	106.5	110.4	115.2	106.7	90.0
Bogo et al. (2016)	137.3	83.4	77.3	79.7	86.8	81.7	82.3	69.3
Moreno-Noguer (2017)	109.5	83.8	93.1	81.6	73.5	72.6	81.5	-
Lassner et al. (2017)	-	-	-	-	-	-	80.7	-
Tome et al. (2017)	-	-	-	-	-	-	79.6	-
Martinez et al. (2017)	77.6	54.0	58.8	49.0	35.9	40.7	52.1	-
Pavlakos et al. (2017)	-	-	-	-	-	-	51.9	-
Tekin et al. (2017)	-	-	-	-	-	-	50.1	-
Ours	121.6	84.2	85.8	75.8	67.8	65.0	83.8	75.3
(MoCap from CMU dataset)								
Ours	146.1	98.9	101.7	92.7	84.4	99.0	100.5	92.3

Table 4: Comparison with the state-of-the-art on the Human3.6M dataset using *Protocol-II*.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis, in: IEEE Conference on Computer Vision and Pattern Recognition.

Andriluka, M., Roth, S., Schiele, B., 2010. Monocular 3d pose estimation and tracking by detection, in: IEEE Conference on Computer Vision and Pattern Recognition.

Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C., 2011. A data-driven approach for real-time full body pose reconstruction from a depth camera, in: IEEE International Conference on Computer Vision.

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2014. 3d pictorial structures for multiple human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

Bo, L., Sminchisescu, C., 2010. Twin gaussian processes for structured prediction. Internation Journal of Computer Vision 87, 28–52.

Bo, L., Sminchisescu, C., Kanaujia, A., Metaxas, D., 2008. Fast algorithms for large scale conditional 3d prediction, in: IEEE Conference on Computer Vision and Pattern Recognition.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J., 2016. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, in: European Conference on Computer Vision.

Bulat, A., Tzimiropoulos, G., 2016. Human pose estimation via convolutional part heatmap regression, in: European Conference on Computer Vision.

Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., 2016. Human pose estimation with iterative error feedback, in: IEEE Conference on Computer Vision and Pattern Recognition.

Chen, C., Ramanan, D., 2017. 3d human pose estimation = 2d pose estimation + matching, in: IEEE Conference on Computer Vision and Pattern Recognition.

Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B., 2016. Synthesizing training images for boosting human 3d pose estimation, in: International Conference on 3D Vision.

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X., 2017. Multi-context attention for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

CMU, 2014. Carnegie mellon university graphics lab: Motion capture database. URL: http://mocap.cs.cmu.edu. mocap.cs.cmu.edu.

Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., Kankanhalli, M., Geng, W., 2016. Marker-less 3d human motion capture with monocular image sequence and height-maps, in: European Conference on Computer Vision.

Gkioxari, G., Toshev, A., Jaitly, N., 2016. Chained predictions using convolutional neural networks, in: European Conference on Computer Vision.

Grest, D., Woetzel, J., Koch, R., 2005. Nonlinear body pose estimation from depth images, in: Joint Pattern Recognition Symposium.

Hu, P., Ramanan, D., 2016. Bottom-up and top-down reasoning with hierarchical rectified gaussians, in: IEEE Conference on Computer Vision and Pattern Recognition.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in: European Conference on Computer Vision.

Ionescu, C., Carreira, J., Sminchisescu, C., 2014a. Iterated second-order label sensitive pooling for 3d human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014b. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1325–1339.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sit
Ionescu et al. (2014b)	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3	151.6
Li and Chan (2014)	-	136.9	96.9	124.7	-	168.7	-	-	-
Tekin et al. (2015)	102.4	158.5	88.0	126.8	118.4	185.0	114.7	107.6	136.2
Tekin et al. (2016)	-	129.1	91.4	121.7	-	162.2	-	-	-
Du et al. (2016)	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5
Chen and Ramanan (2017)	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1
Zhou et al. (2016b)	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5
Zhou et al. (2016a)	91.8	102.4	97.0	98.8	113.4	125.2	90.0	93.9	132.2
Sanzari et al. (2016)	48.8	56.3	96.0	84.8	96.5	105.6	66.3	107.4	116.9
Tome et al. (2017)	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2
Rogez et al. (2017)	76.2	80.2	75.8	83.3	92.2	105.7	79.0	71.7	105.9
Moreno-Noguer (2017)	67.5	79.0	76.5	83.1	97.4	100.4	74.6	72.0	102.4
Mehta et al. (2017)	62.6	78.1	63.4	72.5	88.3	93.8	63.1	74.8	106.6
Zhou et al. (2017)	68.7	74.8	67.8	76.4	76.3	98.4	84.0	70.2	88.0
Mehta et al. (2016)	59.7	69.5	60.9	68.7	76.6	85.7	58.9	78.7	90.9
Lin et al. (2017)	58.0	68.2	63.3	65.8	75.3	93.1	61.2	65.7	98.7
Pavlakos et al. (2017)	67.4	72.0	66.7	69.1	72	77.0	65.0	68.3	83.66
Tekin et al. (2017)	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1
Martinez et al. (2017)	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0
Sun et al. (2017)*	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7
Ours	90.9	98.4	98.2	118.3	118.0	130.5	95.9	112.1	146.1
(MoCap from CMU dataset)									
Ours	139.4	148.0	148.3	165.2	161.7	170.1	138.6	168.2	168.5

	SitDown	Smoking	Waiting	WalkDog	Walk	WalkTogether	Mean	Median
Ionescu et al. (2014b)	243.0	162.1	170.7	177.1	96.6	127.9	162.1	-
Li and Chan (2014)	-	-	-	132.2	70.0	-	-	-
Tekin et al. (2015)	205.7	118.2	146.7	128.1	65.9	77.2	125.3	-
Tekin et al. (2016)	-	-	-	130.5	65.8	-	-	-
Du et al. (2016)	226.9	120.0	117.7	137.4	99.3	106.5	126.5	-
Chen and Ramanan (2017)	240.1	106.7	106.2	114.1	87.0	90.6	114.2	93.1
Zhou et al. (2016b)	199.2	107.4	118.1	114.2	79.4	97.7	113.0	-
Zhou et al. (2016a)	159.0	106.9	94.4	126.1	79.0	99.0	107.3	-
Sanzari et al. (2016)	129.6	97.8	65.9	130.5	92.6	102.2	93.2	-
Tome et al. (2017)	173.9	85.0	85.8	86.3	71.4	73.1	88.4	-
Rogez et al. (2017)	127.1	88.0	83.7	86.6	64.9	84.0	87.7	-
Moreno-Noguer (2017)	116.7	87.7	94.6	82.7	75.2	74.9	85.6	-
Mehta et al. (2017)	138.7	78.8	73.9	82.0	55.8	59.6	80.5	-
Zhou et al. (2017)	113.8	78.0	90.1	75.1	62.6	73.6	79.9	-
Mehta et al. (2016)	125.2	71.2	68.9	82.6	54.0	60.0	74.1	-
Lin et al. (2017)	127.7	70.4	68.2	72.9	50.6	57.7	73.1	-
Pavlakos et al. (2017)	96.5	71.7	65.8	74.9	59.1	63.2	71.9	-
Tekin et al. (2017)	107.3	69.3	70.3	74.3‘	51.8	63.2	69.7	-
Martinez et al. (2017)	94.6	62.3	59.1	65.1	49.5	52.4	62.9	-
Sun et al. (2017)*	86.7	61.5	53.4	61.6	47.1	53.4	59.1	-
Ours	150.1	112.4	113.5	109.2	89.1	88.4	111.8	95.3
(MoCap from CMU dataset)								
Ours	186.7	154.8	154.4	163.7	140.9	160.3	157.3	141.7

Table 5: Comparison with the state-of-the-art on the Human3.6M dataset using *Protocol-III*. *additional ground-truth information is used.

MoCap Data	Walking (A1, C1)			Jogging (A2, C1)			Average
	S1	S2	S3	S1	S2	S3	
HumanEva	27.4	28.6	32.5	39.9	29.4	31.4	31.5
CMU	68.4	81.6	88.3	70.1	81.6	89.9	80.0
CMU → HumanEva	39.5	47.3	61.4	53.5	48.3	53.1	50.5
HumanEva → CMU	45.1	54.9	59.1	58.6	63.1	69.7	58.4

Table 6: Impact of different skeleton structures. The symbol → indicates retargeting of the skeleton structure of one dataset to the skeleton of another dataset.

Iqbal, U., Garbade, M., Gall, J., 2017. Pose for action - action for pose, in: IEEE Conference on Automatic Face and Gesture Recognition.

Johnson, S., Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation, in: British Machine Vision Conference.

Kostrikov, I., Gall, J., 2014. Depth sweep regression forests for estimating 3d human pose from images, in: British Machine Vision Conference.

Krüger, B., Tautges, J., Weber, A., Zinke, A., 2010. Fast local and global similarity searches in large motion capture databases, in: ACM SIGGRAPH Symposium on Computer Animation.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V., 2017. Unite the people: Closing the loop between 3d and 2d human representations, in: IEEE Conference on Computer Vision and Pattern Recognition.

Li, S., Chan, A.B., 2014. 3d human pose estimation from monocular images with deep convolutional neural network, in: Asian Conference on Computer Vision.

Li, S., Zhang, W., Chan, A., 2015. Maximum-margin structured learning with deep networks for 3d human pose estimation, in: IEEE International Conference on Computer Vision.

Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H., 2017. Recurrent 3d pose sequence machines, in: IEEE Conference on Computer Vision and Pattern Recognition.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation, in: IEEE International Conference on Computer Vision.

Mehta, D., Rhodin, H., Casas, D., Sotnychenko, O., Xu, W., Theobalt, C., 2016. Monocular 3d human pose estimation using transfer learning and improved CNN supervision, in: http://arxiv.org/abs/1611.09813.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H., Xu, W., Casas, D., Theobalt, C., 2017. Vnect: Real-time 3d human pose estimation with a single RGB camera, in: SIGGRAPH.

Moreno-Noguer, F., 2017. 3d human pose estimation from a single image via distance matrix regression, in: IEEE Conference on Computer Vision and Pattern Recognition.

Mori, G., Malik, J., 2006. Recovering 3d human body configurations using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1052–1062.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision.

Park, S., Hwang, J., Kwak, N., 2016. 3d human pose estimation using convolutional neural networks with 2d pose information, in: European Conference on Computer Vision Workshops.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose, in: IEEE Conference on Computer Vision and Pattern Recognition.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B., 2016. DeepCut: Joint subset partition and labeling for multi person pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

Pons-Moll, G., Fleet, D.J., Rosenhahn, B., 2014. Posebits for monocular human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

Popa, A., Zanfir, M., Sminchisescu, C., 2017. Deep multitask architecture for integrated 2d and 3d human sensing, in: IEEE Conference on Computer Vision and Pattern Recognition.

Radwan, I., Dhall, A., Goecke, R., 2013. Monocular image 3d human pose estimation under self-occlusion, in: IEEE International Conference on Computer Vision.

Rafi, U., I.Kostrikov, Gall, J., Leibe, B., 2016. An efficient convolutional network for human pose estimation, in: British Machine Vision Conference.

Ramakrishna, V., Kanade, T., Sheikh, Y.A., 2012. Reconstructing 3d human pose from 2d image landmarks, in: European Conference on Computer Vision.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks, in: Conference on Neural Information Processing Systems.

Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C., 2016. Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics 35, 162.

Rogez, G., Schmid, C., 2016. Mocap-guided data augmentation for 3d pose estimation in the wild, in: Conference on Neural Information Processing Systems.

Rogez, G., Weinzaepfel, P., Schmid, C., 2017. Lcr-net: Localization-classification-regression for human pose, in: IEEE Conference on Computer Vision and Pattern Recognition.

Sanzari, M., Ntouskos, V., Pirri, F., 2016. Bayesian image based 3d pose estimation, in: European Conference on Computer Vision.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition.

Sigal, L., Balan, A.O., Black, M.J., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Internation Journal of Computer Vision 87, 4–27.

Sigal, L., Isard, M., Haussecker, H., Black, M.J., 2012. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. Internation Journal of Computer Vision 98, 15–48.

Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F., 2013. A joint model for 2d and 3d pose estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition.

Simo-Serra, E., Ramisa, A., Alenyà, G., Torras, C., Moreno-Noguer, F., 2012. Single image 3d human pose estimation from noisy observations, in: IEEE Conference on Computer Vision and Pattern Recognition.

Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.N., 2005. Discriminative density propagation for 3d human motion estimation, in: IEEE Conference on Computer Vision and Pattern Recognition.

Sun, X., Shang, J., Liang, S., Wei, Y., 2017. Compositional human pose regression, in: International Conference on Computer Vision.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P., 2016. Structured prediction of 3d human pose with deep neural networks, in: British Machine Vision Conference.

Tekin, B., Marquez-Neila, P., Salzmann, M., Fua, P., 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation, in: IEEE International Conference on Computer Vision.

Tekin, B., Suna, X., Wanga, X., Lepetita, V., Fua, P., 2015. Predicting people's 3d poses from short sequences, in: IEEE Conference on Computer Vision and Pattern Recognition.

Tome, D., Russell, C., Agapito, L., 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition.

Toshev, A., Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition.

Urtasun, R., Fleet, D.J., Fua, P., 2006. 3d people tracking with gaussian process dynamical models, in: IEEE Conference on Computer Vision and Pattern Recognition.

Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W., 2014. Robust estimation of 3d human poses from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition.

Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines, in: IEEE Conference on Computer Vision and Pattern Recognition.

Yang, Y., Ramanan, D., 2011. Articulated pose estimation with flexible mixtures-of-parts, in: IEEE Conference on Computer Vision and Pattern Recognition.

Yao, A., Gall, J., Van Gool, L., 2012. Coupled action recognition and pose estimation from multiple views. International Journal of Computer Vision 100, 16–37.

Yasin, H., Iqbal, U., Krüger, B., Weber, A., Gall, J., 2016. A dual-source

Methods	Walking (A1, C1)			Jogging (A2, C1)			Average
	S1	S2	S3	S1	S2	S3	
Simo-Serra et al. (2012)	99.6	108.3	127.4	109.2	93.1	115.8	108.9
Radwan et al. (2013)	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang et al. (2014)	71.9	75.7	85.3	62.6	77.7	54.4	71.3
Simo-Serra et al. (2013)	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Kostrikov and Gall (2014)	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Bo and Sminchisescu (2010)*	38.2	32.8	40.2	42.0	34.7	46.4	39.1
Yasin et al. (2016)	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Lin et al. (2017)	26.5	20.7	38.0	41.0	29.7	29.1	30.8
Popa et al. (2017)	27.1	18.4	39.5	37.6	28.9	27.6	29.9
Martinez et al. (2017)	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Pavlakos et al. (2017)	22.3	19.5	29.7	28.9	21.9	23.8	24.3
Moreno-Noguer (2017)	19.8	12.6	26.2	43.8	21.8	22.1	24.4
Ours	27.4	28.6	32.5	39.9	29.4	31.4	31.5
MoCap from CMU dataset							
Yasin et al. (2016)	52.2	51.0	62.8	74.5	72.4	56.8	61.6
Ours	39.5	47.3	61.4	53.5	48.3	53.1	50.5

Table 7: Comparison with other state-of-the-art approaches on the HumanEva-I dataset. The average 3D pose error (mm) is reported for all three subjects (S1, S2, S3) and camera C1. * denotes a different evaluation protocol.

approach for 3d pose estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition.

Yasin, H., Krüger, B., Weber, A., 2013. Model based full body human motion reconstruction from video data, in: International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications.

Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y., 2016a. Deep kinematic pose regression, in: European Conference on Computer Vision Workshops.

Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K., 2015. Sparse representation for 3d shape estimation: A convex relaxation approach, in: IEEE Conference on Computer Vision and Pattern Recognition.

Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K., 2016b. Sparseness meets deepness: 3d human pose estimation from monocular video, in: IEEE Conference on Computer Vision and Pattern Recognition.

Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K., 2017. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. arXiv preprint arXiv:1701.02354 .