Harvesting Multiple Views for Marker-less 3D Human Pose Annotations

Georgios Pavlakos[1], Xiaowei Zhou[1], Konstantinos G. Derpanis[2], Kostas Daniilidis[1]

[1] University of Pennsylvania   [2] Ryerson University

Abstract

Recent advances with Convolutional Networks (ConvNets) have shifted the bottleneck for many computer vision tasks to annotated data collection. In this paper, we present a geometry-driven approach to automatically collect annotations for human pose prediction tasks. Starting from a generic ConvNet for 2D human pose, and assuming a multi-view setup, we describe an automatic way to collect accurate 3D human pose annotations. We capitalize on constraints offered by the 3D geometry of the camera setup and the 3D structure of the human body to probabilistically combine per view 2D ConvNet predictions into a globally optimal 3D pose. This 3D pose is used as the basis for harvesting annotations. The benefit of the annotations produced automatically with our approach is demonstrated in two challenging settings: (i) fine-tuning a generic ConvNet-based 2D pose predictor to capture the discriminative aspects of a subject's appearance (i.e., "personalization"), and (ii) training a ConvNet from scratch for single view 3D human pose prediction without leveraging 3D pose groundtruth. The proposed multi-view pose estimator achieves state-of-the-art results on standard benchmarks, demonstrating the effectiveness of our method in exploiting the available multi-view information.

1. Introduction

Key to much of the success with Convolutional Networks (ConvNets) is the availability of abundant labeled training data. For many tasks though this assumption is unrealistic. As a result, many recent works have explored alternative training schemes, such as unsupervised training [17, 26, 45], auxiliary tasks that improve learning representations [42], and tasks where groundtruth comes for free, or is very easy to acquire [31]. Inspired by these works, this paper proposes a geometry-driven approach to automatically gather a high-quality set of annotations for human pose estimation tasks, both in 2D and 3D.

ConvNets have had a tremendous impact on the task of 2D human pose estimation [40, 41, 27]. A promising research direction to improve performance is to automatically
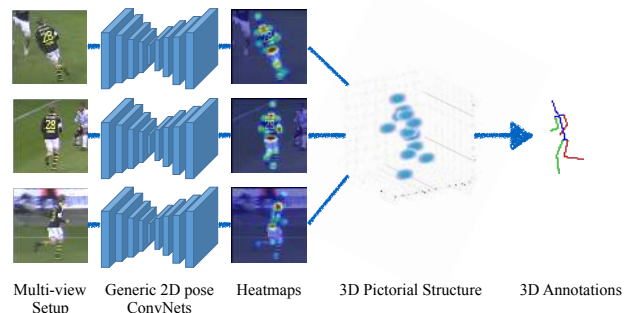

Figure 1: Overview of our approach for harvesting pose annotations. Given a multi-view camera setup, we use a generic ConvNet for 2D human pose estimation [27], and produce single-view pose predictions in the form of 2D heatmaps for each view. The single-view predictions are combined optimally using a 3D Pictorial Structures model to yield 3D pose estimates with associated per joint uncertainties. The pose estimate is further probed to determine reliable joints to be used as annotations.

adapt (i.e., "personalize") a pretrained ConvNet-based 2D pose predictor to the subject under observation [11]. In contrast to its 2D counterpart, 3D human pose estimation suffers from the difficulty of gathering 3D groundtruth. While gathering large-scale 2D pose annotations from images is feasible, collecting corresponding 3D groundtruth is not. Instead, most works have relied on limited 3D annotations captured with motion capture (MoCap) rigs in very restrictive indoor settings. Ideally, a simple, marker-less, multi-camera approach could provide reliable 3D human pose estimates in general settings. Leveraging these estimates as 3D annotations of images would capture the variability in users, clothing, and settings, which is crucial for ConvNets to properly generalize.

Towards this goal, this paper proposes a geometry-driven approach to automatically harvest reliable annotations from multi-view imagery. Figure 1 provides an overview of our approach to automatically harvest reliable joint annotations. Given a set of images captured with a calibrated multi-view setup, a generic ConvNet for 2D human pose [27] pro-

duces single-view confidence heatmaps for each joint. The heatmaps in each view are backprojected to a common discretized 3D space, functioning as unary potentials of a 3D pictorial structure [16, 15], while a tree graph models the pairwise relations between joints. The marginalized posterior distribution of the 3D pictorial structures model for each joint is used to identify which estimates are reliable. These reliable keypoints are used as annotations.

Besides achieving state-of-the-art performance as compared to previous multi-view human pose estimators, our approach provides abundant annotations for pose-related learning tasks. In this paper, we consider two tasks. In the first task, we project the 3D pose annotations to the 2D images to create "personalized" 2D groundtruth, which is used to adapt the generic 2D ConvNet to the particular test conditions (Figure 2a). In the second task, we use the 3D pose annotations to train from scratch a ConvNet for single view 3D human pose estimation that is on par with the current state-of-the-art. Notably, in training our pose predictor, we limit the training set to the harvested annotations and do not use the available 3D groundtruth (Figure 2b).

In summary, our four main contributions are as follows:

- We propose a geometry-driven approach to automatically acquire 3D annotations for human pose without 3D markers;

- the harvested annotations are used to fine-tune a pretrained ConvNet for 2D pose prediction to adapt to the discriminative aspects of the appearance of the subject under study, i.e., "personalization"; we empirically show significant performance benefits;

- the harvested annotations are used to train from scratch a ConvNet that maps an image to a 3D pose, which is on par with the state-of-the-art, even though none of the available 3D groundtruth is used;

- our approach for multi-view 3D human pose estimation achieves state-of-the-art results on standard benchmarks, which further underlines the effectiveness of our approach in exploiting the available multi-view information.

2. Related work

Data scarcity for human pose tasks: Chen *et al*. [12] and Ghezelghieh *et al*. [18] create additional synthetic examples for 3D human pose to improve ConvNet training. Rogez and Schmid [34] introduce a collage approach. They combine human parts from different images to generate examples with known 3D pose. Yasin *et al*. [44] address the data scarcity problem, by leveraging data from different sources, e.g., 2D pose annotations and MoCap data. Wu *et al*. [42] also integrate dual source learning within a single ConvNet.

(a) "Personalizing" a 2D pose ConvNet
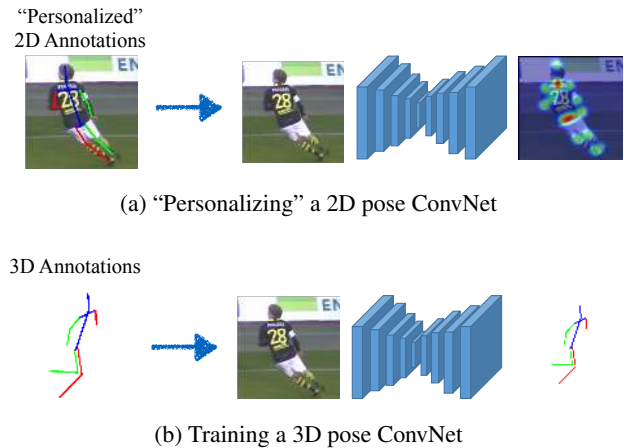

(b) Training a 3D pose ConvNet

Figure 2: The quality of the harvested annotations is demonstrated in two applications: (a) projecting the 3D estimates into the 2D imagery and using them to adapt ("personalize") a generic 2D pose ConvNet to the discriminative appearance aspects of the subject, (b) training a ConvNet that predicts 3D human pose from a single color image.

Instead of creating synthetic examples, or bypassing the missing data, the focus of our approach is different. In particular, our goal is to gather images with corresponding 2D and 3D automatically generated annotations and use them to train a ConvNet. This way we employ images with statistics similar to those found in-the-wild, which have been proven to be of great value for ConvNet-based approaches.

2D human pose: Until recently, the dominant paradigm for 2D human pose involved local appearance modeling of the body parts coupled with the enforcement of structural constraints with a pictorial structures model [3, 43, 32]. Lately though, end-to-end approaches using ConvNets have become the standard in this domain. The initial work of Toshev and Szegedy [40] regressed directly the x, y coordinates of the joints using a cascade of ConvNets. Tompson *et al*. [39] proposed the regression of heatmaps to improve training. Pfister *et al*. [30] proposed the use of intermediate supervision, with Wei *et al*. [41] and Carreira *et al*. [10] refining iteratively the network output. More recently, Newell *et al*. [27] built upon previous work to identify the best practices for human pose prediction and propose an hourglass module consisting of ResNet components [19], and iterative processing to achieve state-of-the-art performance on standard benchmarks [2, 36]. In this work, we employ the hourglass architecture as our starting point for generating automatic 3D human pose annotations.

Single view 3D human pose: 3D human pose estimation from a single image has been typically approached by applying more and more powerful discriminative methods on the image and combining them with expressive 3D priors to

recover the final pose [37, 47, 7]. As in the 2D pose case, ConvNets trained end-to-end have grown in prominence. Li and Chan [24] regress directly the x, y, z spatial coordinates for each joint. Tekin *et al*. [38] additionally use an autoencoder to learn and enforce structural constraints on the output. Pavlakos *et al*. [29] instead propose the regression of 3D heatmaps instead of 3D coordinates. Li *et al*. [25] follow a nearest neighbor approach between color images and pose candidates. Rogez and Schmid [34] use a classification approach, where the classes represent a sample of poses. To demonstrate the quality of our harvested 3D annotations, we also regress the x, y, z joint coordinates [24, 38], while employing a more recent architecture [27].

Multi-view 3D human pose: Several approaches [6, 1, 9, 22, 4, 5] have extended the pictorial structures model [16, 15] to reason about 3D human pose taken from multiple (calibrated) viewpoints. Earlier work proposed simultaneously reasoning about 2D pose across multiple views, and triangulating 2D estimates to realize actual 3D pose estimates [6, 1]. Recently, Elhayek *et al*. [13, 14] used ConvNet pose detections for multi-view inference, but with a focus on tracking rather than annotation harvesting, as pursued here. Similar to the current paper, 3D pose has previously been directly modelled in 3D space [9, 22, 4, 5]. A straightforward application of the basic pictorial structures model to 3D is computationally expensive due to the six degrees of freedom for the part parameterization. Our parameterization instead models only the 3D joint position, something that has also been proposed in the context of single view 3D pose estimation [23]. This instantiation of the pictorial structure makes inference tractable since we deal with three degrees of freedom rather than six.

Personalization: Consideration of pose in video presents an opportunity to tune the appearance model to the discriminative appearance aspects of the subject and thus improve performance. Previous work [33] leveraged this insight by using a generic pose detector to initially identify a set of high-precision canonical poses. These detections are then used to train a subject-specific detector. Recently, Charles *et al*. [11] extended this idea using a generic 2D pose ConvNet to identify a select number of high precision annotations. These annotations are propagated across the video sequence based on 2D image evidence, e.g., optical flow. Regarding identifying confident predictions, the work of Jammalamadaka *et al*. [21] is related, where they extract features from the image and the output and train an evaluator to estimate whether the predicted pose is correct. In our work, rather than using 2D image cues to identify reliable annotations, our proposed approach leverages the rich 3D geometry presented by the multi-view setting and the constraints of 3D human pose structure, to combine and consolidate single view information. Such cues are highly reliable and complementary to image-based ones.

3. Technical approach

The following subsections describe the main components of our proposed approach. Section 3.1 gives a brief description of the generic ConvNet used for 2D pose predictions. Section 3.2 describes the 3D pictorial structures model used to aggregate multi-view image-driven keypoint evidence (i.e., heatmaps) provided as output by a ConvNet-based 2D pose predictor with 3D geometric information from a human skeleton model. Section 3.3 describes our annotation selection scheme that identifies reliable keypoint estimates based on the marginalized posterior distribution of the 3D pictorial structures model for each keypoint. The proposed uncertainty measure inherently integrates image evidence across all viewpoints and geometry. Finally, Sections 3.4 and 3.5 present two applications of our annotation harvesting approach. Section 3.4 describes the use of the harvested annotations to fine-tune an existing 2D pose ConvNet predictor. The resulting adapted predictor is sensitive to the discriminative aspects of the appearance of the subject under consideration, i.e., "personalization". Section 3.5 describes how we use the harvested annotations to train from scratch a 3D pose ConvNet predictor that maps a single image to 3D pose.

3.1. Generic ConvNet

The initial component of our approach is a generic ConvNet for 2D human pose estimation that provides the initial set of noisy predictions for single view images. Since our approach is agnostic to the particular network architecture, any of the top-performing ConvNets is sufficient for this step, e.g., [41, 8, 27]. Here, we adopt the state-of-the-art stacked hourglass design [27]. The main architectural component of this network is the hourglass module which consists of successive convolutional and pooling layers, followed by convolutional and upsampling layers, leading to a symmetric hourglass design. Stacking multiple hourglasses together allows for iterative processing of the image features. Best performance is achieved by the use of intermediate supervision, forcing the network to produce one set of predictions at the end of each hourglass.

The prediction of the network is in the form of 2D heatmaps for each joint. The entire heatmap output includes useful information regarding the confidence of predictions, and can be considered as a 2D distribution of the joint locations. To take advantage of the entire heatmap prediction, we backproject the 2D distributions of the joints in a discretized 3D cube. This is used to accommodate the predictions for all the views and serves as the inference space for 3D pictorial structures model, described in Sec. 3.2.

3.2. Multi-view optimization

The pose model used to aggregate information across views is based on a 3D generalization of the classical pic-

torial structures model [16, 15]. A major departure of the current work from prior 3D instantiations of multi-view approaches (e.g., [9]) is the use of a joint representation, $\mathbf{S} = \{\mathbf{s}_i | i = 1, \ldots, N\}$, where $\mathbf{s}_i \in \mathbb{R}^3$ encodes the 3D position of each joint, rather than the 3D configuration of parts, i.e., limbs. The simplified parameterization and tree structure for the pairwise terms admit efficient 3D joint configuration inference via dynamic programming, i.e., the sum-product algorithm.

Articulation constraints: The pairwise relation between joints is modelled by a tree structure of the human skeleton. The edge set is denoted by $\mathcal{E}$ and the edge (i.e., limb) lengths by $\{L_{ij} | (i, j) \in \mathcal{E}\}$. The prior distribution is given by the articulation constraints and can be written as

$$p(\mathbf{S}) \propto \prod_{(i,j) \in \mathcal{E}} p(\mathbf{s}_i, \mathbf{s}_j). \tag{1}$$

The pairwise terms, $p(\mathbf{s}_i, \mathbf{s}_j)$, constrain the lengths of the human limbs L_{ij}:

$$p(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} 1, & \text{if } L_{ij} - \varepsilon \leq \|\mathbf{s}_i - \mathbf{s}_j\| \leq L_{ij} + \varepsilon \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $\varepsilon = 1$ is used as a tolerance for the variability from the expected limb length L_{ij} of the subject. More sophisticated pairwise terms can also be adopted if MoCap data are available, e.g., [23].

Data likelihood: Given a 3D pose, the likelihood of seeing M synchronized images from M calibrated cameras is modeled as

$$p(\mathbf{I}|\mathbf{S}) \propto \prod_{k=1}^{M} \prod_{i=1}^{N} p(\mathbf{I}_k | \pi_k(\mathbf{s}_i)), \tag{3}$$

where $\pi_k(\mathbf{s}_i)$ denotes the 2D projection of $\mathbf{s}_i$ in the k-th view given the camera parameters. The data likelihood, $p(\mathbf{I}_k | \pi_k(\mathbf{s}_i))$, is modelled by the multi-channel heatmap outputs of the ConvNet (Sec. 3.1).

Inference: Finally, the posterior distribution of a 3D pose given 2D images from different views is given by:

$$p(\mathbf{S}|\mathbf{I}) \propto \prod_{k=1}^{M} \prod_{i=1}^{N} p(\mathbf{I}_k | \pi_k(\mathbf{s}_i)) \prod_{(i,j) \in \mathcal{E}} p(\mathbf{s}_i, \mathbf{s}_j). \tag{4}$$

The solution space of the 3D joint position is restricted to a 3D bounding volume around the subject and quantized by a $64 \times 64 \times 64$ grid. Pose estimates are computed as the mean of the marginal distribution of each joint given the multi-view images. The marginal distribution of the discrete variables is efficiently computed by the sum-product algorithm [15].

3.3. Annotation selection

The 3D reconstructions provided by the multi-view optimization offer a very rich but noisy set of annotations. We are effectively equipped with automatic 3D annotations for all the images of the multi-view setup. Moreover, these annotations integrate appearance cues from each view (2D pose heatmaps), geometric constraints from the multiple views (backprojection in a common 3D space), as well as constraints from the articulated structure (3D pictorial structure). This allows us to capitalize on the available information from the images and the 3D geometry to provide a robust set of annotations.

For further benefits, we proceed to a selection step over the annotations provided from the 3D reconstruction. A useful property of our multi-view optimization using the pictorial structures model is that the marginalized distribution of each joint offers a measure of the prediction's uncertainty. This means that we are provided with a selection cue for free. For example, the determinant of the 3D covariance matrix for each joint's marginalized distribution can be used as a confidence measure to decide whether the joint will be used as an annotation. In our experiments, we identify as reliable annotations the 70% most confident predictions for each joint in terms of the determinant of the 3D covariance matrix, although other measures are also possible.

3.4. "Personalizing" 2D pose ConvNet

The goal of "personalization" is to adapt the original detector such that it captures the discriminative appearance aspects of the subject of interest, such as clothing. Both Ramanan *et al.* [33] and Charles *et al.* [11] proposed methods to "personalize" a detector using 2D evidence (e.g., optical flow) from monocular video. Instead, our proposed approach focuses on cues provided by image evidence, geometric properties of the multi-view setup, and structural constraints of the human body.

Given the set of selected annotations, we use them to fine-tune a generic 2D pose ConvNet with backpropagation, such that it adapts to the testing conditions of interest. The procedure is very similar to the one used to train the ConvNet in the first place, with the difference that we leverage our automatically generated annotations as targets for the available images. The target heatmaps consist of a 2D Gaussian with a standard deviation $\sigma = 1$ pixel, centered on the annotation location of the joint. A separate heatmap is synthesized for each joint. During training, we use a Mean Squared-Error loss between the predicted and the target heatmaps. If the joint is not within the selected annotation set (i.e., the localization is not confident), we simply ignore the loss incurred by it during optimization. We terminate refinement after four epochs through our auto-annotated data to avoid overfitting on the given examples.

3.5. 3D pose ConvNet training

For 3D human pose estimation, we train a ConvNet from scratch that takes a single image as input and predicts the 3D pose. Our formulation follows the coordinate regression paradigm [24, 38], but more sophisticated methods can also be employed, e.g., the volumetric representation for 3D pose [29]. The target of the network is the x, y, z coordinates of N human body joints. For x, y we use pixel coordinates, while z is expressed in metric depth with respect to a specified root joint (here the pelvis is defined as the root). We organize the output in a single $3N$-dimensional vector. The network is supervised with an $\mathcal{L}_2$ regression loss:

$$\mathcal{L} = \sum_{n=1}^{N} \| \boldsymbol{x}_{gt}^n - \boldsymbol{x}_{pr}^n \|_2^2, \tag{5}$$

where $\boldsymbol{x}_{gt}^n$ is the groundtruth and $\boldsymbol{x}_{pr}^n$ is the predicted location for joint n. The architecture we use is a single hourglass module [27] with the addition of a fully connected layer at the end to allow every output to have a connection with each activation of the previous feature volume.

4. Empirical evaluation

This section is dedicated to the empirical evaluation of our proposed approach. First, we give a description of the datasets used (Section 4.1). Next, we briefly discuss the implementation details of our approach (Section 4.2). Finally, we present the quantitative (Sections 4.3 to 4.5) and the qualitative evaluations (Section 4.6).

4.1. Datasets

For our quantitative evaluation we focused on two datasets that target human pose estimation and provide a multiple camera setup; (i) KTH Multiview Football II [9], a small-scale outdoor dataset with challenging visual conditions, and (ii) Human3.6M [20], a large-scale indoor dataset, with a variety of available scenarios.

KTH Multiview Football II [9] contains images of professional footballers playing a match. Evaluation for 3D pose was performed using the standard protocol introduced with the dataset [9] and used elsewhere [22, 5], where Sequence 1 of "Player 2" is used for testing. Reported results are based on the percentage of correct parts (PCP) to measure 3D part localization using the two and three camera setups. Additional evaluation for 2D pose was performed using Sequence 2 of "Player 2" for testing, where reported results are based on the percentage of correct parts in 2D.

Human3.6M [20] is a recent large-scale dataset for 3D human sensing captured in a lab setting. It includes 11 subjects performing 15 actions, such as walking, sitting, and phoning. Following previous work [25, 47], we use two subjects for testing (S9 and S11), and report results based on the average 3D joint error.

	Two cameras				Three cameras				
	[9]	[4]	[5]	Ours	[9]	[22]	[4]	[5]	Ours
Upper arms	53	64	96	**98**	60	89	68	98	**100**
Lower arms	28	50	68	**92**	35	68	56	72	**100**
Upper legs	88	75	98	**99**	**100**	**100**	78	99	**100**
Lower legs	82	66	88	**97**	90	99	70	92	**100**
Average	62.7	63.8	87.5	**96.5**	71.2	89.0	68.0	90.3	**100**

Table 1: Quantitative comparison of multi-view pose estimation methods on KTH Multiview Football II. The numbers are the percentage of correct parts (PCP) in 3D using two and three cameras. Baseline numbers are taken from the respective papers. In constrast to the compared methods, no training data from this dataset was used for our approach.

It is crucial to mention that in the experiments presented below, no groundtruth data was leveraged for training from the respective datasets. We relied solely on the generic 2D ConvNet (trained on MPII [2]) and the knowledge of the geometry from the calibrated camera setup.

4.2. Implementation details

For the generic 2D pose ConvNet, we use a publicly available model [27], which is trained on the MPII human pose dataset [2]. To "personalize" a given 2D pose ConvNet through fine-tuning, we maintain the same training details as the ones described in the original work. The learning rate is set to 2.5e-4, the batch size is 4, rmsprop is used for optimization and data augmentation is used, that includes rotation ($\pm 30^o$), scale (± 0.25), and left-right flipping.

To train the 3D pose ConvNet, we employ the same architecture, but we use only one hourglass component and add a fully connected layer at the end, to regress the N joints coordinates. The training details with respect to optimization and data augmentation are the same as for the initial network, but training is done from scratch (we do not use a pretrained model).

4.3. Multi-view pose estimation

First of all, we need to assess the accuracy of the annotations provided from our multi-view optimization scheme. Since our ConvNets are not trained using groundtruth data from the aforementioned datasets, we heavily rely on the quality of these automatic annotations. Therefore, we evaluate multi-view pose estimation using our approach, described in Section 3.2

First, we report results of our approach on the small-scale, yet challenging KTH dataset. Even though relevant methods train specialized 2D detectors for pose estimation, they are all outperformed by our approach using only a generic ConvNet for 2D joint prediction. The relative improvement is illustrated in Table 1.

For Human3.6M we apply the same method to multi-

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Li et al. [25]	-	134.13	97.37	122.33	-	166.15	-	-
Zhou et al. [47]	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Tekin et al. [38]	-	129.06	91.43	121.68	-	162.17	-	-
Zhou et al. [46]	91.83	102.41	96.95	98.75	113.35	125.22	90.04	93.84
Ours	**41.18**	**49.19**	**42.79**	**43.44**	**55.62**	**46.91**	**40.33**	**63.68**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
Li et al. [25]	-	-	-	-	134.13	68.51	-	-
Zhou et al. [47]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	113.01
Tekin et al. [38]	-	-	-	-	130.53	65.75	-	-
Zhou et al. [46]	132.16	158.97	106.91	94.41	126.04	79.02	98.96	107.26
Ours	**97.56**	**119.90**	**52.12**	**42.68**	**51.93**	**41.79**	**39.37**	**56.89**

Table 2: Quantitative evaluation of our approach on Human3.6M. The numbers are the average 3D joint errors (mm). Baseline numbers are taken from the respective papers. Note that Zhou *et al*. [47] use video, while our proposed method is multi-view.

view pose estimation. Since this dataset was published very recently, there are no reported results for multi-view pose estimation methods. It is interesting though to compare with the top-performing works for single view 3D pose such that we can quantify the current gap between single view and multi-view estimation. The full results are presented in Table 2. Our approach reduces the error of the state-of-the-art single view approach of Zhou *et al*. [46] by almost a half. We note that Zhou *et al*. [47] use video instead of prediction from a single frame. We do not include results from Bogo *et al*. [7] and Sanzari *et al*. [35] which report average errors of 82.3mm and 93.15mm, respectively, since they use a rigid alignment between the estimated pose and the groundtruth, making it not comparable with the other methods. Moreover, as a weak multi-view baseline, we averaged the per view 3D estimates from one of the state-of-the-art approaches [47]. This naive combination achieves an average error of 103.10mm which is a minimal improvement compared to the original error of 113.01mm for the corresponding single view approach. This demonstrates that handling the views independently and combining the single view 3D pose results in a late stage does not leverage the rich 3D geometric constraints available and significantly underperforms compared to our multi-view optimization.

4.4. "Personalizing" 2D pose ConvNet

Having validated the accuracy of our proposed multi-view optimization scheme, the next step is to actually leverage the automatic annotations for learning purposes. The most immediate benefit comes from using them to refine the generic ConvNet and adapt it to the particular test conditions. This can be considered as an application of "personalization", similar in nature to the goal of Charles *et al*. [11], where significant pose estimation gains in terms of accuracy were reported.

For KTH we use the two available sequences from "Player 2" to evaluate the online adaptation of our network.

	Seq 1		Seq2	
	Generic	Refined	Generic	Refined
Upper arms	98	**100**	80	**89**
Lower arms	89	**92**	64	**74**
Upper legs	98	**100**	85	**91**
Lower legs	96	**100**	79	**86**
Average	95.3	**98.0**	77.0	**84.5**

Table 3: Quantitative comparison of the generic ConvNet versus the refined version for the two sequences of "Player 2" from KTH Multiview Football II. The numbers are percentage of correct parts (PCP) in 2D. Performance improvement is observed across all parts.

Since our focus is to purely evaluate the quality of the 2D predictions before and after refinement, we report 2D PCP results in Table 3. We observe performance improvement across all parts of the subject. Moreover, for the second sequence which is considerably more challenging, the benefit from our refinement is even greater. This underlines the importance of refinement when the original detector fails.

For Human3.6M we evaluate the quality of 2D heatmaps through their impact on the multi-view optimization. Achieving better results for 2D pose estimation is definitely desirable, but ideally, the predicted heatmaps should benefit other post-processing steps as well, e.g., our multi-view optimization. In Table 4, we provide a more detailed ablative study comparing different sets of annotations for refinement. Starting with the "Generic" ConvNet, one naive approach we compare against is using the heatmap maximum predictions as annotations ("HM"), or a subset of the most confident of those predictions ("HM+sel"). For "HM+sel" we use the heatmap value to indicate detection confidence, and identify only the top 70% for each joint as reliable 2D annotations. These serve as baselines for refining the ConvNet. We also employ the complete annotation set that is provided from our multi-view optimization

	Purchases	Sitting	Sitting Down	Average (15 actions)
Generic	63.68	97.56	119.90	56.89
HM	57.57	86.37	100.39	55.13
HM+sel	52.50	91.49	110.30	55.62
PS	51.32	79.39	97.26	51.18
PS+sel	**45.98**	**68.09**	**73.91**	**47.83**

Table 4: Quantitative comparison of multi-view optimization after fine-tuning the ConvNet with different annotation sets and evaluating on Human3.6M. We present results for the three most challenging actions (based on Table 2), along with the average across all actions. The numbers are the average 3D joint error (mm). "Generic", "HM", "HM+sel", "PS" and "PS+sel" are defined in Section 4.4.

	Average (6 actions)	Average (15 actions)
Li *et al*. [25]	121.31	-
Tekin *et al*. [38]	116.77	-
Park *et al*. [28]	111.12	117.34
Zhou *et al*. [46]	104.73	107.26
Rogez *et al*. [34]	-	121.2
Ours	113.65	118.41

Table 5: Quantitative comparison of single image approaches on Human3.6M. The numbers are the average 3D joint errors (mm). Baseline numbers are taken from the respective papers. In contrast to the other works, we do not use 3D groundtruth for training, instead we rely solely on the harvested 3D annotations. Despite that, our performance is on par with the state-of-the-art.

("PS"), and a high quality version of this by selecting the most confident joint predictions only (denoted as "PS+sel" and described in Section 3.3). The reported results include both the average performance across all 15 actions, as well as the performance on the three actions with the highest error, according to Table 2, namely, Purchases, Sitting, and Sitting Down. Again, the performance benefits are greater for more challenging actions, which justifies the use of our method to overcome dataset bias and adapt to the scenario of interest. Also, the naive approach to recover more 2D annotations and bootstrapping on the output of the generic ConvNet ("HM" and "HM+sel") is only marginally helpful on average, which underlines the benefit of the rich geometrical information we employ to recover annotations. Finally, the proposed selection scheme ("PS+sel") outperforms the model that uses all annotations of the multi-view optimization ("PS") which exemplifies the importance of selecting only a high-quality subset of the annotations for refinement.

4.5. Training a 3D pose ConvNet

A great challenge, but also a very interesting application of our method is to use the automatically generated annotations to train a ConvNet for 3D pose estimation. Since KTH is a small-scale dataset, we focus on Human3.6M. We leverage the high-quality annotations from the multi-view optimization scheme, and train the network described in Section 3.5 from scratch. The results are presented in Table 5, along with other approaches. Even though we only use the noisy annotations recovered by our approach for training and ignored the groundtruth from the dataset, the final trained ConvNet is on par with the state-of-the-art.

4.6. Qualitative results

For "personalization", Figures 3 and 4 show qualitative sample results of the proposed approach with and without fine-tuning on annotations recovered from the input imagery on KTH Multiview Football II and Human3.6M, re-

spectively. Despite the generic ConvNet being quite reliable, it might fail for the most challenging poses which are underrepresented in the original generic training set. The benefit from the "personalized" ConvNet is greater in these cases since it adapts to the discriminative appearance of the user and recovers the pose successfully.

For the 3D pose ConvNet trained from scratch, we present example 3D reconstructions in Figure 5. Notice the challenging poses of the subject and the very accurate poses predicted by the ConvNet.

5. Summary

This paper presented an automatic way to gather 3D annotations for human pose estimation tasks, using a generic ConvNet for 2D pose estimation and recordings from a multi-view setup. The automatically generated annotations were used to adapt a generic ConvNet to the particular task, demonstrating important performance benefits from this "personalization". Additionally, we trained a ConvNet for 3D pose estimation which performs on par with the current state-of-the-art, even though we only used automatically harvested annotations, and ignored the provided groundtruth.

One promising direction for future work is using the automatic annotation setup in an outdoor environment, (where MoCap systems and depth sensors are not applicable) to collect 3D annotations for in-the-wild images. This would allow us to train a generic 3D human pose ConvNet, similar to the 2D counterparts, by overcoming the bottleneck of limited color images with 3D groundtruth.
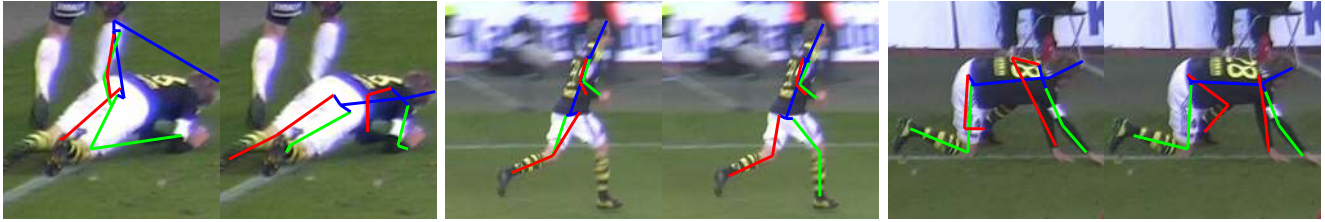
Figure 3: Examples on KTH Multiview Football II showing the performance gain from "personalization". For each pair of images, pose estimation results are presented from the generic (left) and the "personalized" ConvNet (right).
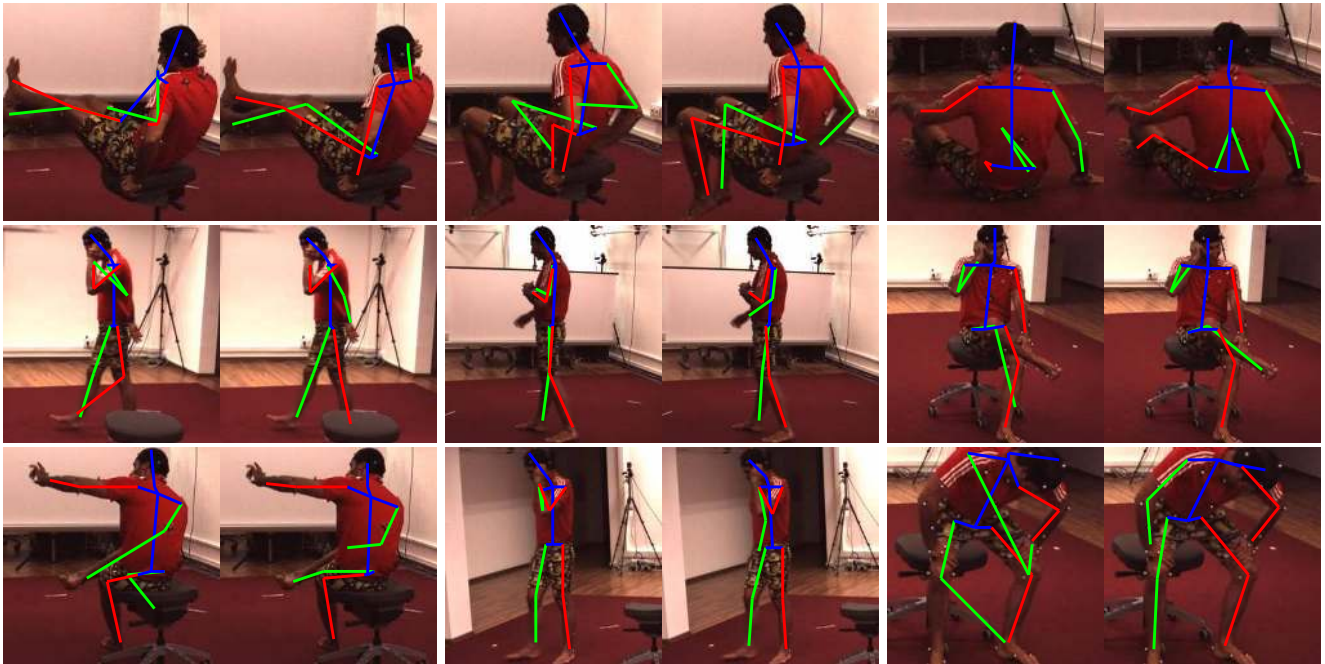

Figure 4: Examples on Human3.6M showing the performance gain from "personalization". For each pair of images, pose estimation results are presented from the generic (left) and the "personalized" ConvNet (right).
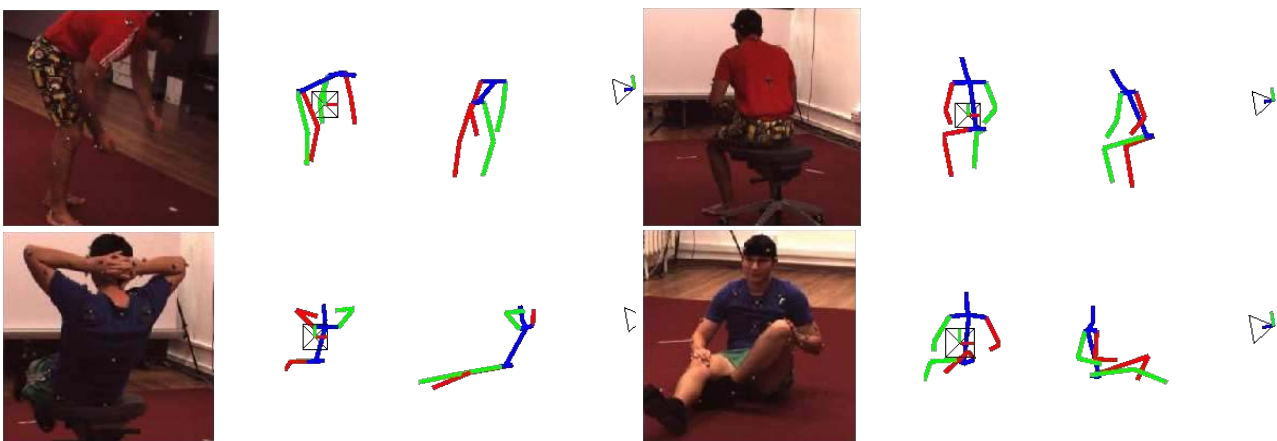

Figure 5: Example predictions on Human3.6M from the ConvNet trained to estimate 3D pose from a single image. For each example, we present (left-to-right) the input image, the predicted 3D pose from the original view, and a novel view. Red and green indicate left and right, respectively.

References

[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2013. 3

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2, 5

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[4] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 3, 5

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures revisited: Multiple human pose estimation. *PAMI*, 38(10):1929 – 1942, 2105. 3, 5

[6] M. Bergtholdt, J. H. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *IJCV*, 87(1-2):93–117, 2010. 3

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 3, 6

[8] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 3

[9] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 3, 4, 5

[10] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[11] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman. Personalizing human video pose estimation. In *CVPR*, 2016. 1, 3, 4, 6

[12] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen, et al. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 2

[13] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015. 3

[14] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. MARCOnI - ConvNet-based MARker-less motion Capture in Outdoor and Indoor scenes. *PAMI*, 39(3):501–514, 2017. 3

[15] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2, 3, 4

[16] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *TC*, C-22(1):67–92, 1973. 2, 3, 4

[17] R. Garg and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 1

[18] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar. Learning camera viewpoint using CNN to improve 3D body pose estimation. In *3DV*, 2016. 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[20] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 5

[21] N. Jammalamadaka, A. Zisserman, M. Eichner, V. Ferrari, and C. Jawahar. Has my algorithm succeeded? An evaluator for human pose estimators. In *ECCV*, 2012. 3

[22] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *BMVC*, 2013. 3, 5

[23] I. Kostrikov and J. Gall. Depth sweep regression forests for estimating 3D human pose from images. In *BMVC*, 2014. 3, 4

[24] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 3, 5

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

[26] G. Long, L. Kneip, J. M. Alvarez, and H. Li. Learning image matching by simply watching video. In *ECCV*, 2016. 1

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3, 5

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

[29] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. *CVPR*, 2017. 3, 5

[30] T. Pfister, J. Charles, and A. Zisserman. Flowing ConvNets for human pose estimation in videos. In *ICCV*, 2015. 2

[31] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta. The curious robot: Learning visual representations via physical interactions. In *ECCV*, 2016. 1

[32] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2

[33] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65–81, 2007. 3, 4

[34] G. Rogez and C. Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 2, 3, 7

[35] M. Sanzari, V. Ntouskos, and F. Pirri. Bayesian image based 3D pose estimation. In *ECCV*, 2016. 6

[36] B. Sapp and B. Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013. 2

[37] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3D human pose estimation from noisy observations. In *CVPR*, 2012. 3

[38] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 3, 5, 6, 7

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

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

[41] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1, 2, 3

[42] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 1, 2

[43] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2013. 2

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

[45] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCVW*, 2016. 1

[46] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCVW*, 2016. 6, 7

[47] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 3, 5, 6