3D Human Pose Estimation = 2D Pose Estimation + Matching

Ching-Hang Chen
Carnegie Mellon University
chinghac@andrew.cmu.edu

Deva Ramanan
Carnegie Mellon University
deva@cs.cmu.edu

Abstract

We explore 3D human pose estimation from a single RGB image. While many approaches try to directly predict 3D pose from image measurements, we explore a simple architecture that reasons through intermediate 2D pose predictions. Our approach is based on two key observations (1) Deep neural nets have revolutionized 2D pose estimation, producing accurate 2D predictions even for poses with self-occlusions (2) "Big-data"sets of 3D mocap data are now readily available, making it tempting to "lift" predicted 2D poses to 3D through simple memorization (e.g., nearest neighbors). The resulting architecture is straightforward to implement with off-the-shelf 2D pose estimation systems and 3D mocap libraries. Importantly, we demonstrate that such methods outperform almost all state-of-the-art 3D pose estimation systems, most of which directly try to regress 3D pose from 2D measurements.

1. Introduction

Inferring 3D human pose from image measurements is classic task in computer vision, dating back to the iconic work of Hogg [11] and O'Rourke and Badler [22]. Such a technology has immediate applications in various tasks such as action understanding, surveillance, human-robot interaction, and motion capture, to name a few. As such, it has a long and storied history. We refer the reader to various surveys for a broad overview of the popular topic [8, 19].

Previous approaches often make use of a highly sensored environment, including video streams [37, 30], multiview cameras [3, 10], depth images [23, 36, 27]. In this work, we focus on the "pure" and challenging setting of recovering 3D body pose with a single 2D RGB image [17, 35, 25, 32].

Our key insight to the problem is leveraging recent advances in 2D image understanding, made possible through the undeniable impact of deep learning. While originally explored for coarse recognition tasks such as image classification, recent methods have extended such network architectures to "fine-grained" human pose estimation, where the task is formulated as one of 2D heatmap
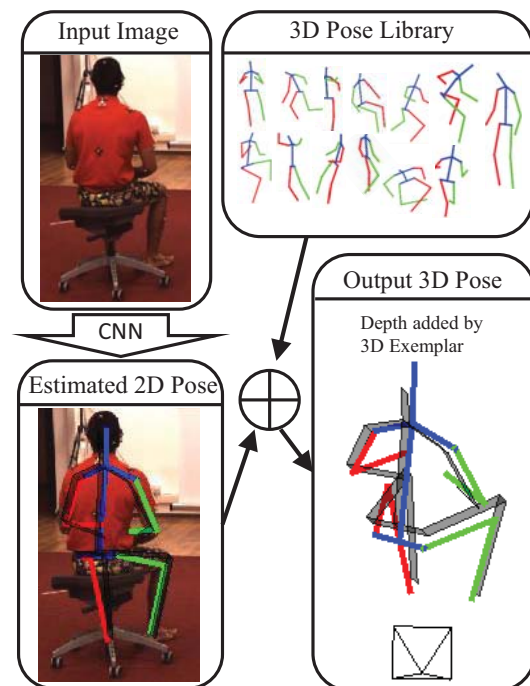

Figure 1. Overview of our approach for 3D pose estimation: given an input image, first estimate a 2D pose and then estimate its depth by matching to a library of 3D poses. The final prediction is given by the colored skeleton, while the ground-truth is shown in gray. Our approach works surprisingly well because 2D pose estimation is accurate even during occlusions (as illustrated by both wrists above), suggesting that 2D pose estimates need only be refined by adding depth values.

prediction [33, 21, 31, 12]. One of the long standing challenges in 2D human pose estimation has been estimating poses under self-occlusions. Indeed, reasoning about occlusions has been one of the underlying motivations for working in a 3D coordinate frame rather than 2D. But one of our salient conclusions is that state-of-the-art methods do a surprisingly good job of 2D pose estimation *even under occlusion*. Given this observation, the remaining challenge is predicting depth values for the estimated 2D joints.

Inferring 3D structure from 2D correspondences is also a

well-studied problem in computer vision, often addressed in multiview setting as structure from motion. In the context of monocular human pose estimation, the relevant cues seem to be semantic rather than geometric. One can estimate 3D postures from a 2D skeleton based on high-level knowledge derived from anthropometric, kinematic, and dynamic constraints. Inspired by the success of data-driven architectures, we explore a simple non-parametric encoding of such high-level constraints: given a 3D pose library, we generate a large number of 2D projections (from virtual camera views). Given this training set of paired (2D,3D) data and predictions from a 2D pose estimation algorithm, we report back the depths from the 3D pose associated with the closest matching 2D example from our library. Our entire pipeline is summarized in Fig. 1.

Generalization: One desirable property of our two-stage approach is generalization. Due to the difficulty of annotation in 3D, training datasets with 3D labels are typically collected in a lab environment, while 2D datasets tend to be more diverse. Our two-stage pipeline makes use of different training sets for different stages, resulting a system that can predict 3D poses from "in-the-wild" images.

Evaluation: Though we present qualitative results on in-the-wild-imagery, we also perform an extensive quantitative evaluation of our method on widely benchmarked 3D human-pose datasets, such as Human3.6M [13]. We follow standard train/test protocol splits, but our analysis reveals that there has been inconsistent reporting in the literature, both in terms of test sets and evaluation criteria. To make our results as transparent as possible, we report performance for all metrics and splits we could find. One of our surprising findings is the impressive performance of our simple pipeline: we outperform essentially all prior work on all metrics. Our entire pipeline, even given the non-parametric matching step, returns a 3D pose given a 2D image in **under 200ms** (160ms for 2D estimation by a CNN, 26ms for exemplar matching with a training library of 200,000 poses). Finally, to promote future progress, we perform an exhaustive analysis of additional baselines with upper bounds that reveal the continued benefit of working with intermediate 2D representations and data-driven encoding of 3D constraints.

2. Related work

Here we review related works on 3D human pose prediction most relevant to our approach.

(Deep) Regression: Most existing work that makes use of deep features tends to formulate the problem as a direct 2D image to 3D pose regression task. Li *et al.* [17] use deep learning to train a regression model to predict 3D pose directly from images. Tekin *et al.* [30] integrate spatio-temporal features via an image sequence to learn regression model for 3D pose mapping. We provide both a theoretical

and empirical analysis that suggests that 2D pose may be a useful intermediate representation.

Intermediate 2D pose: Other approaches have explored pipelines that use 2D poses as an intermediate result. Most focus on the second-stage that lifts 2D estimates to 3D. This is classically treated as a constrained optimization problem who's objective minimizes the 2D reprojection error of an unknown 3D pose and unknown camera [37, 32, 24, 2]. The optimization problem is often subject to kinematic constraints [34, 29], and sometimes 3D poses are assumed to live a in low-dimensional subspace to better condition the optimization [37]. Such optimization-based approaches could be sensitive to initialization and local minima, and often require expensive constrained solvers. We use data-driven matching, that when combined with a simple closed-form warping algorithm, yields a fast and accurate 3D solution.

Exemplar-based: Previous work has also explored example-based methods, dating back at least to [26]. A central challenge is generalization to novel poses outside the training set. [14] propose matching upper and lower bodies individually, to allow for novel compositions at test-time. [35] adapt exemplars to better match image measurements with an energy minimization approach. [25] synthesize new 2D images with image-based rendering. Other methods also warp 3D exemplars to 2D image descriptors, often based on shape-context [1, 20] or silhouette features [5]. In our work, we show that a modest number of exemplars (200,000), combined with a simple closed-form algorithm for warping a 3D exemplar to exactly project to 2D pose estimates, outperforms more complex methods.

3. Approach

In this section, we describe our method for estimating 3D human pose given a single RGB image. We make use of a probabilistic formulation over variables including the image I, the 3D pose $\mathbf{X} \in \mathbb{R}^{N \times 3}$, and the 2D pose $\mathbf{x} \in \mathbb{R}^{N \times 2}$, where N is the number of articulated joints. We write the joint probability as:

$$p(\mathbf{X}, \mathbf{x}, I) = p(\mathbf{X}|\mathbf{x}, I) \cdot p(\mathbf{x}|I) \cdot p(I) \qquad (1)$$

where the above makes no limiting assumptions by itself.

Conditional independence: Let us now assume that the 3D pose $\mathbf{X}$ is *conditionally independent* of image I given the 2D pose $\mathbf{x}$. This is equivalent to the implication that given a 2D skeleton, the prediction of its corresponding 3D skeleton would not be affected by 2D image measurements. While this is not quite true (we show a counter example in Fig. 2), it seems to be a reasonable first-order approximation. Moreover, this factorization still allows for $p(\mathbf{x}|I)$ to be *arbitrarily complex*, which is likely needed to accurately model complex interactions between 2D projections
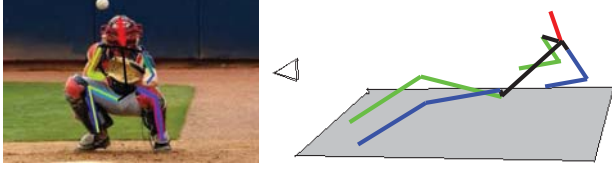
Figure 2. A failure case where the 3D pose is *not* conditionally independent of the image given the 2D pose: $p(\mathbf{X}|\mathbf{x}, I) \neq p(\mathbf{X}|\mathbf{x})$. We show the output of our system given the ground-truth 2D pose, with the (incorrect) best-matching 3D exemplar on the right (visualized from a novel viewpoint, where the estimated camera is drawn as a view frustum). Our experiments suggest that such cases are rare, and that much of the time 3D can be inferred from 2D projections.

and image features during occlusions. Given this conditional Independence, one can write:

$$p(\mathbf{X}, \mathbf{x}, I) = \underbrace{p(\mathbf{X}|\mathbf{x})}_{NN} \cdot \underbrace{p(\mathbf{x}|I)}_{CNN} \cdot p(I) \quad (2)$$

We tackle the second term with a image-based CNN that predicts 2D keypoint heatmaps. We tackle the first term with a non-parametric nearest-neighbor (NN) model. We describe each term in turn below.

3.1. Image-Based 2D Pose Estimation

Given the above Independence assumption, we would first like to predict 2D pose given image measurements. We model the conditional of 2D pose given an image as

$$P(\mathbf{x}|I) = CNN(I) \quad (3)$$

where we assume CNN is a nonlinear function that returns N 2D heatmaps (or marginal distributions over the location of individual joints). We make use of convolutional pose machines (CPMs) [33], which return precisely N heatmaps for individual body joints. We normalize the heatmaps so that they can be interpreted as marginal distributions for each joint. CPM is a near-state-of-the-art pose estimation system (88.5% PCKh on MPII dataset [4], quite close to the state-of-the-art value of 90.9% [21]). Note the off-the-shelf CPM model was trained on MPII dataset, which is a somewhat limited dataset in that annotations are provided through manual inspection. We fine-tune this model on the large scale Human3.6M [13] training set, which contains annotations acquired by a mocap system (allowing for larger-scale labeling).

3.2. Nonparametric 3D shape model

We model $P(\mathbf{X}|\mathbf{x})$ with a non-parametric nearest neighbor model. We will follow a notational convention where $\mathbf{X} = [X, Y, Z]$ and $\mathbf{x} = [x, y]$. Assume that we have library
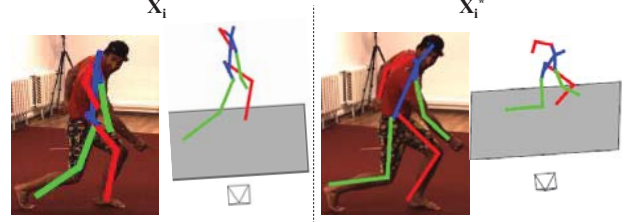

Figure 3. On the left, we show the 3D exemplar $\mathbf{X_i}$ that best matches the ground-truth 2D pose $\mathbf{x}$. While the overall pose is roughly correct, the arms and legs are bent incorrectly. By simply copying the depth values from the exemplar (and copying the (x, y) values from the 2D pose under a weak perspective model, as given in (6)), we can obtain a warped exemplar $\mathbf{X_i^*}$ that better matches the 2D pose.

of 3D poses $\{\mathbf{X}_i\}$ paired with a particular camera projection matrix $\{M_i\}$, such that the associated 2D poses are given by $\{M_i(\mathbf{X}_i)\}$. If we want to consider multiple cameras for a single 3D pose, we add another copy of the 3D pose with a different camera matrix to our library. We define a distribution over 3D poses based on reprojection error:

$$P(\mathbf{X} = \mathbf{X}_i|\mathbf{x}) \propto e^{-\frac{1}{\sigma^2}||M_i(\mathbf{X}_i) - \mathbf{x}||^2} \quad (4)$$

where the MAP estimate is given by the 1-nearest neighbor (1NN). We explore two extensions to the above basic framework.

Virtual cameras: We can further reduce the squared reprojection error by searching over small perturbations of each camera. This involves solving a camera resectioning problem [9], where an iterative solver can be initialized with M_i:

$$M_i^* = \underset{M}{\operatorname{argmin}} ||M(\mathbf{X}_i) - \mathbf{x}||^2 \quad (5)$$

In practice, we construct a shortlist of k candidates that score well according to (4), and resort them according to optimal camera matrix. We found that optimizing over cameras produced a small but noticeable improvement in our experiments. Unless otherwise specified, we choose $k = 10$ in our experiments.

Warped exemplars: Much previous work on exemplars introduce methods for warping exemplars to better match the 2D pose estimates, often formulated as an inverse kinematics optimization problem. We describe an extremely lightweight method for doing so here. We first align the 3D exemplar to the camera-coordinate system used to compute the projection $\mathbf{x}$. This is done with a 3D rigid transformation given by the camera extrinsics encoded in M_i (or M_i^*). In practice we use a training set $\{\mathbf{X}_i\}$ where 3D exemplars are already aligned to their projections $\{\mathbf{x}_i\}$, implying that extrinsics in M_i reduce to an identity matrix (which is the case for the Human3.6M dataset [13], since 3D poses are

specified in camera coordinates of their associated image projections). Given this alignment, we simply replace the (X_i, Y_i) exemplar coordinates with their scaled 2D counterparts (x, y) under a weak perspective camera model:

$$\mathbf{X}_i^* = \begin{bmatrix} sx & sy & Z_i \end{bmatrix}, \quad \text{where} \quad s = \frac{\text{average}(Z_i)}{f} \quad (6)$$

where f is the focal length of the camera (given by the intrinsics in M_i) and average(Z_i) is the average depth of the 3D joints. Such weak-perspective approximations are commonly used to initialize algorithms for perspective (PnP) camera calibration [18], and will be reasonable when the depth variation of the human skeleton is small relative to the overall distance to the camera. Our results suggest that such closed-form solutions for 3D warping rival the accuracy of complex energy minimization methods (see Fig. 3).

4. Experiments

In our experiments, we test a variety of variations of our proposed pipeline.

Qualitative results: We first present some qualitative results. Fig. 4 shows results on challenging examples from subject S11 of Human3.6M. We choose examples with self-occlusions and sitting poses. To demonstrate the accuracy of the 3D predictions, we visualize novel viewpoints. We then apply the proposed method on Leeds Sports Pose(LSP) dataset [15] to test cross-dataset generalization. We posit that our pipeline will generalize across image variation (due to the underlying robustness of our 2D pose estimation system) but maybe limited in the 3D estimates due to the library used (from Human3.6M). Importantly, our approach produces plausible 3D poses even when the activity class is not included in Human3.6M. This implies that our method can reliably estimate 3D poses in the wild!

4.1. Evaluation protocols

We use Human3.6M for quantitative evaluation and analysis. It appears that multiple train/test splits have been used in the literature, as well as different approaches to computing mean per joint position error (MPJPE), measured in millimeters. We summarize them here.

Protocol 1: In [35, 16, 25], the entire dataset was partitioned into six training subjects (S1, S5, S6, S7, S8, S9), and one testing subject (S11). Evaluation is performed on every 64th frame of S11's video clips. In this configuration, there are total 1.8 million 3D poses available in the training set. MPJPE between the ground truth 3D pose and the estimated 3D pose is computed by first aligning poses with a rigid transformation [16].

Protocol 2: Others [37, 30, 17] use five subjects (S1, S5, S6, S7, S8) for training, and two subjects (S9, S11) for testing. We follow [37]'s setup that downsamples the

videos from 50 fps to 10 fps. Here, MPJPE is evaluated *without* a rigid transformation, following the original *h36m* protocol: both the ground-truth and predicted 3D pose is centered with respect to a root joint (*ie.* pelvis). In contrast to Protocol 1, this evaluation can be sensitive to a single poorly-predicted joint, particularly if it is the root [13].

To compare to published performance numbers, we use the appropriate protocol as needed. From our own experience, we find Protocol 1 to be simpler and more intuitive, and so focus on it for our diagnostic evaluations.

4.2. Comparison to state-of-the-art (Protocol 1)

Final system: Table 1 compare MPJPE for each activity class. Our approach clearly outperforms [35] and [25]. ("Ours" in the tables of comparison throughout the experiment refers to the warped exemplar $\mathbf{X}^*$ described in Section 3.2)

Performance given ground-truth 2D: A common diagnostic is evaluating performance given ground-truth 2D poses, written as gt. Table 2 shows that our simple matching + warping outperforms [35], who use a complex iterative algorithm for matching and warping exemplars to image evidence. Our diagnostics will later show that even matching exemplars *without* warping outperforms prior art, indicating the remarkable power of a simple NN baseline.

Size of trainset: Table 3 shows the MPJPE versus the training data size. Since approaches deal with 2D and 3D sources differently, we list both sizes. Yasin *et al.* [35] project multiple 2D poses from each 3D exemplar (with virtual cameras) to create 2D poses for matching, and Rogez *et al.* [25] directly synthesize 2D images for training. Our approach makes use of the default training data in Human3.6M, where each 3D pose is paired with a single 2D projection. We max out performance with a modest pose library of 180k 3D-2D pairs, but produce competitive accuracy even for 18k. The slight increase in MPJPE for larger training sets seems to be related to noise from 2D pose estimation, since we observe a monotonically decrease when ground truth 2D poses are given (Fig. 7).

4.3. Comparison to state-of-the-art (Protocol 2)

Final system: Table 4 provides the comparison to [37] and [30] using Protocol 2. Note that in both these works, temporal smoothness was exploited by taking a short image sequences as input. Even though we do not use temporal information, our system is quite close to state-of-the-art. A qualitative comparison to [37] is also provided in Fig. 5.

Performance given ground-truth 2D: Our strong performance in Fig. 5 might be attributed to better 2D pose estimation. Therefore, we investigate the case given ground truth 2D pose, following Zhou's diagnostic protocol [37]: evaluate MPJPE up to a 3D rigid body transformation including scale, only on the first 30 seconds of the first cam-
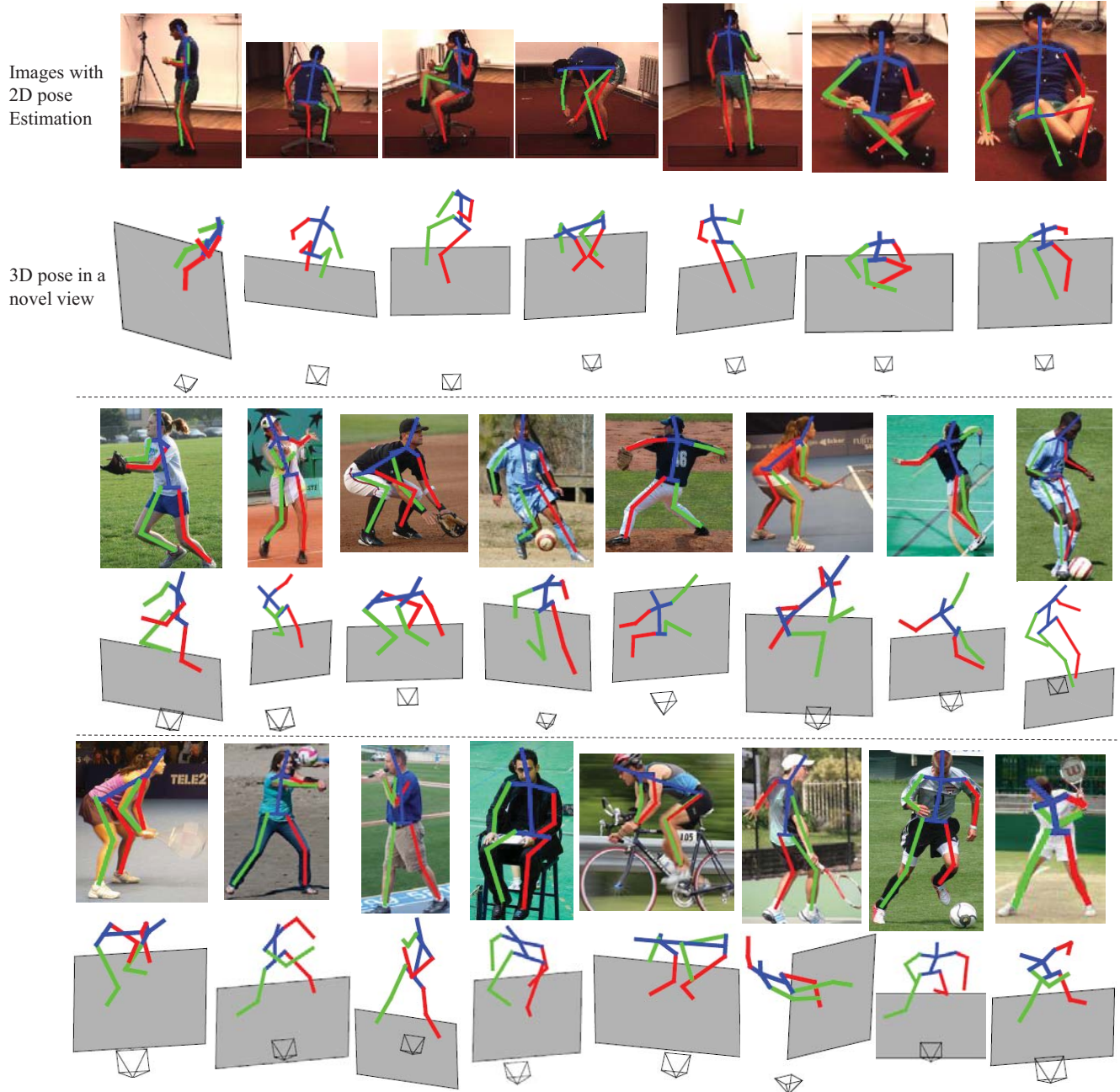
Figure 4. We show qualitative results on Human3.6M-test (**top**) and LSP-test (**bottom**). Our method produces plausible results for challenging images with self-occlusions and extreme poses, and can generalize to activities and poses not in the train set (Human3.6M-train).

era in Human3.6M. For a fair comparison, we make use the same training set of 3D-2D training data for both methods. The results are shown in Table 5. With a shortlist of $k = 10$ matches, camera resectioning (5) and exemplar warping (6) produces a slightly lower error than [37]'s approach without a 3D prior. Qualitative results are provided in Fig. 6. Our approach produces lower 2D reprojection error, while Zhou's method appears to suffers from the restriction of 3D

poses to a low-dimensional subspace.

4.4. Diagnostics

We now perform an extensive set of diagnostics to reveal the strength of our individual components, as well as upper-bound analysis that is useful for guiding future work. For simplicity, we restrict ourselves to **Protocol 1**.

Mean Per Joint Position Error (MPJPE), in mm

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Yasin [35]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	**170.8**
Rogez [25]	-	-	-	-	-	-	-	-	-
Ours	**71.63**	**66.60**	**74.74**	**79.09**	**70.05**	**67.56**	**89.30**	**90.74**	195.62
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Yasin [35]	108.2	142.5	86.9	92.1	165.7	102.0	108.3	-	-
Rogez [25]	-	-	-	-	-	-	88.1	-	-
Ours	**83.46**	**93.26**	**71.15**	**55.74**	**85.86**	**62.51**	**82.72**	**69.05**	-

Table 1. Comparison to [35] by **Protocol 1**. Our results are clearly state-of-the-art. Please see text for more details.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Yasin [35]	60.0	54.7	71.6	67.5	63.8	61.9	55.7	73.9	110.8
$\mathbf{X}^*|gt$ (Ours)	**53.27**	**46.75**	**58.63**	**61.21**	**55.98**	**58.13**	**48.85**	**55.60**	**73.41**
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Yasin [35]	78.9	96.9	67.9	47.5	89.3	53.4	70.5	-	-
Ours	**60.25**	**76.05**	**62.19**	**35.76**	**61.93**	**51.08**	**57.50**	**51.93**	-

Table 2. Comparison to [35] by **Protocol 1** given 2D ground truth. Our approach is clearly state-of-the-art, indicating the effectiveness of our simple approach to NN matching and warping. Table 7 shows that even simple NN matching produces an average accuracy of 70.93, rivaling prior art.

Method	2D source	3D source	Avg. MPJPE
Yasin [35]	64,000k	380k	108.3
Rogez [25]	207k	190k	88.1
Ours	18k	18k	85.94
Ours	180k	180k	**82.37**
Ours	1,800k	1,800k	82.72

Table 3. Comparison to [35] and [25] under different amounts of training data, under **Protocol 1**. Our approach yields the best performance at the source size of 180k.

Effect of warping: We evaluate the benefits of warping ($\mathbf{X}^*$ vs $\mathbf{X}$) in Table 6. It is clear that warping exemplars $\mathbf{X}^*$ is a simple and effective approach to reducing error. Quite surprisingly, *even without warping*, simply matching to a set of 3D exemplar projections outperforms the state-of-the-art (see Table 1 & Table 6)! To analyze an upper-bound for our warping approach, we combine 2D estimates (x, y) with depth values Z given by the ground-truth 3D pose Z_{GT}. In the last row, the performance of combining ground truth depth Z_{GT} with $\mathbf{x}$ is listed as a reference baseline. This suggests that one can still dramatically lower error by 2X even when continuing to use the output of current 2D pose estimation systems.

Warping given ground-truth 2D: Next, we compute the error for the case that ground truth 2D pose is given, as shown in Table 7. We write $|gt$ to emphasize that methods now have access to 2D ground-truth pose estimates. We first note that matching *unwarped* examples rivals the accuracy of state-of-the-art (see Table 2 & Table 7). This again suggests the remarkable power a simple NN baseline based on matching 2D projections. That said, warping still improves results by a considerable margin. A qualitative example is provided in Fig. 3.

Warping given optimal exemplar match: It is natural to ask what is the upper-bound on performance given our training set of (3D,2D) pairs. We first compute the optimal exemplar that minimizes 3D reprojection error (up to a rigid body transformation) to the true 3D test pose. We write the index of this best match from the training set as $i = GT$. We would like to see the effect of warping given this optimal match. We analyze this combination in Table 8. This suggests that, in principle, error can still be significantly reduced (by almost 2X) even given our fixed library of 3D poses. However, it is not clear that this is obtainable given our pipeline because it may require image evidence to select this optimal 3D exemplar (violating the conditional independence assumption from (2)).

Effect of trainset size: An important aspect to investigate is the influence of database size. Here we investigate the error versus the number of exemplars in the database. Fig. 7 evaluates performance versus a random fraction of our overall database. As expected, more data results in lower error, though diminishing results are observed (even in log scale). This is reasonable since training data is extracted from videos captured at 50 fps, implying that correlations over frames might limit the benefit of additional frames. We see that convergence is also effected by the quality of the 2D pose estimates: error given ground-truth 2D poses plateaus at 5×10^5, while 2D pose estimates plateau even sooner at 2×10^5. We posit that a more restricted 3D pose prior (implicitly enforced by a small

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Zhou [37]	**87.36**	109.31	**87.05**	**103.16**	116.18	106.88	**99.78**	**124.52**	**199.23**
Tekin [30]	102.41	147.72	88.83	125.38	118.02	112.38	129.17	138.89	224.9
Ours	89.87	**97.57**	89.98	107.87	**107.31**	**93.56**	136.09	133.14	240.12
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Zhou [37]	107.42	139.46	118.09	79.39	114.23	97.70	**113.01**	-	-
Tekin [30]	118.42	182.73	138.75	**55.07**	126.29	**65.76**	124.97	-	-
Ours	**106.65**	**139.17**	**106.21**	87.03	**114.05**	90.55	114.18	93.05	-

Table 4. Comparison to [37] and [30] by **Protocol 2**. Our results are close to state-of-the-art.
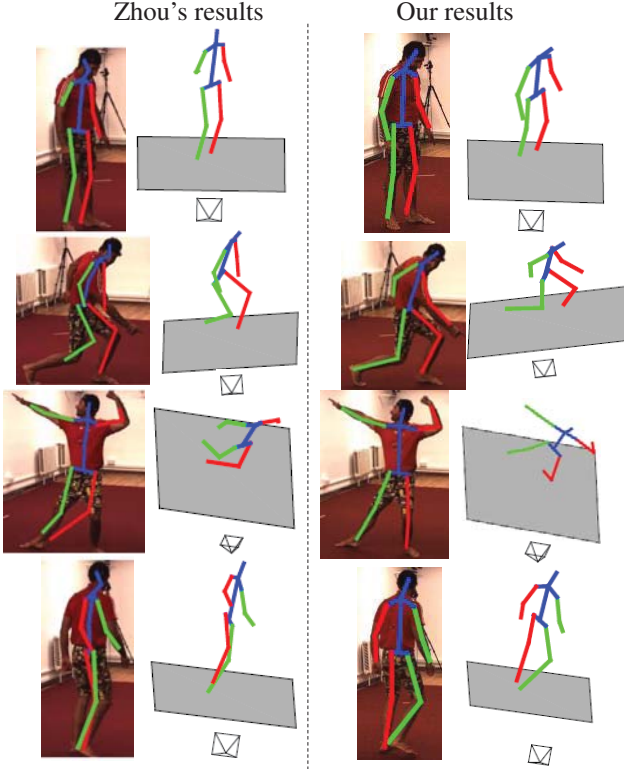

Figure 5. Qualitative comparison of Zhou [37] with our results. Our results are generally more accurate, but both methods struggle with left/right limb ambiguities (e.g., the second row). While much of our improved performance comes from better 2D pose estimation, we still compare favorably when using the same ground-truth 2D pose estimates (Fig. 6 and Table 5).

randomly-sampled 3D library) helps given inaccurate 2D pose estimates. But in either case, *exemplar-based 3D matching is effective even for modestly-size training sets (200,000)*. This analysis appears to suggest that better 2D pose estimates are needed to take advantage of "bigger" 3D datasets.

Since the joint prediction error is not a normal distribution, we also plot median error in Fig. 8. We see that the median is generally lower than mean error, and the difference between the two becomes smaller when ground truth 2D or 3D is given. This may suggest that errors are often
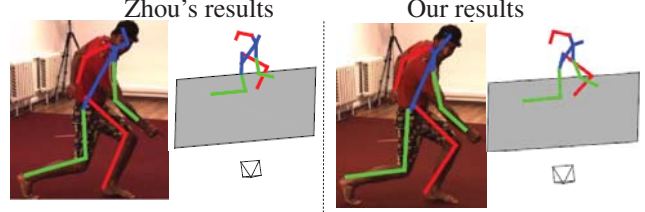

Figure 6. Qualitative comparison of Zhou [37] with our results, given access to the *same ground-truth 2D pose*. While both 3D estimates are plausible, Zhou's tends to produce higher 2D reprojection error since 3D poses are restricted to lie in a subspace (e.g., the incorrectly-articulated head).

Method	Avg. MPJPE
Ramakrishna [24]$\|gt$, multi-frame	89.50
Dai [6]$\|gt$, multi-frame	72.98
Zhou [37]$\|gt$, single-frame	50.04
Zhou [37]$\|gt$, multi-frame	49.64
$\mathbf{X}^*\|gt$, $k = 1$, single-frame	51.06
$\mathbf{X}^*\|gt$, $k = 10$, single-frame	**49.55**

Table 5. 3D pose estimation accuracy given ground-truth 2D poses, under **Protocol 2**. Here, k is the number of candidate exemplar extracted in the shortlist that are subsequently processed by searching over virtual cameras. Our single-frame results with $k = 10$ outperforms all prior art, including those which make use of multi-frame temporal cues.

Prediction	Avg.	Median
$\mathbf{X}\|\mathbf{x}$	85.52	75.04
$\mathbf{X}^*\|\mathbf{x}$	**82.72**	**69.05**
$[s\mathbf{x}\ Z_{GT}]$	43.86	30.19

Table 6. Given the predicted 2D pose $\mathbf{x}$, warped exemplars $\mathbf{X}^*$ outperform unwarped exemplars $\mathbf{X}$ by a reasonable margin. We also compute an upper-bound for warped exemplars that uses (x, y) estimates from the predicted pose and z estimates from the ground-truth 3D pose. The dramatic error reduction suggests that significant further improvement is possible by improving upon our 3D matching. Importantly, this improvement is realizable even given existing 2D pose estimation systems.

due to a single incorrect joint prediction, which would significantly impact average error but not the median.

Cross-dataset evaluation: To further examine general-

Prediction	Avg.	Median
$\mathbf{X}|gt$	70.93	65.35
$\mathbf{X}^*|gt$	**57.50**	**51.93**

Table 7. We compare matching to exemplars $\mathbf{X}$ and warped exemplars $\mathbf{X}^*$ given ground-truth 2D pose estimates. This suggests that our simple closed-form warping approach would be even more effective with better 2D pose estimates.

Prediction	Avg.	Median
$\mathbf{X}|GT$	60.11	55.36
$\mathbf{X}^*|GT$	**37.32**	**33.91**

Table 8. We analyze performance given the optimal matching 3D training exemplar "GT" (in terms of 3D error wrt the ground-truth test 3D pose). Simply reporting this optimal match produces an error of 60mm, around 10mm lower than the actual match found given an ideal 2D pose-estimation system (Table 7). Warping this exemplar $X^*|GT$ significantly improves accuracy. This suggests that our overall 3D matching stage could still be significantly improved even given the current size of the library of 3D poses.
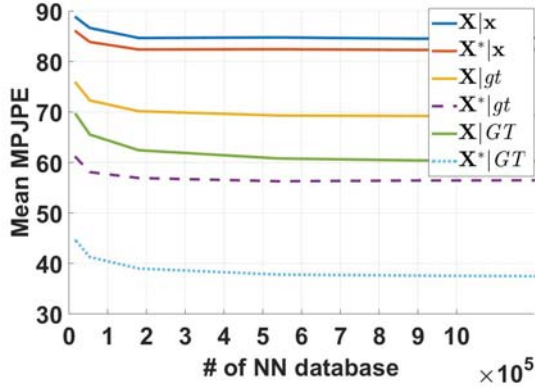

Figure 7. Mean MPJPE by **Protocol 1** versus size of the 3D pose library. We explore diagnostic variants using previously-introduced notation. In general, MPJPE decreases with a larger library. The error saturates at 2×10^5 when using CNN-predicted 2D poses "$|\mathbf{x}$", but saturates later at 5×10^5 when using ground-truth 2D poses "$|gt$". The results suggest that with better 2D pose estimates, our exemplar matching would benefit from larger training data.

ization of exemplar-matching, Table 9 quantitatively evaluates accuracy on HumanEva-I [28] given a model trained on Human3.6M. These results suggest that the 3D exemplars from HumanEva do generalize, and that generalization is *significantly* improved through our warping procedure.

5. Conclusion

We present an simple approach to 3D human pose estimation by performing 2D pose estimation, followed by 3D exemplar matching. The simplicity and efficiency of our method, combined with its state-of-the-art performance on both benchmark datasets and unconstrained "in-the-wild"
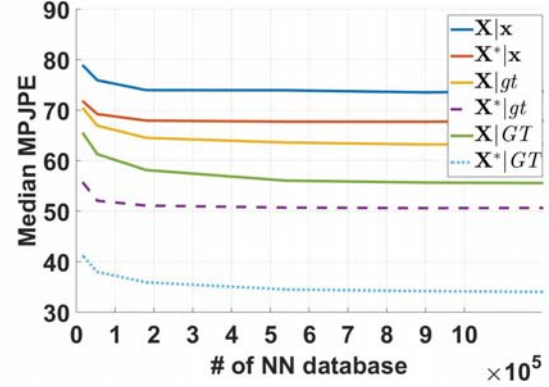

Figure 8. Median MPJPE by **Protocol 1** versus database size. Median error is lower than mean error in Fig. 7, suggesting that a few joints are responsible for a large mean error. Other trends follow the mean error curves from Fig. 7.

	Walk	Jog	Throw Catch	Gestures	Box	**Avg.**
Warped	64.46	69.88	59.99	67.89	79.22	68.29
Unwarped	90.17	95.27	82.74	88.82	103.85	92.17

Table 9. We evaluate a Human3.6M-trained model on HumanEva (under Protocol 1). To isolate the impact of 3D matching, we use ground-truth 2D keypoints. As a point of comparison, average error on Human3.6M test is 70.93 (unwarped) and 57.5 (warped) (Table 9). These results suggest that 3D exemplars do generalize across datasets, and importantly, warping *significantly* increases the amount of generalization. Note that the two datasets use different definitions of skeletons, implying that learning a mapping should reduce error even further.

imagery, suggests that such simple baselines should be used for future benchmarking in 3D pose estimation. A notable advantage of intermediate 2D representations is *modular training* – 2D datasets (which are typically larger and more diverse because of ease of annotation) can be used to train the initial-image processing module, while 3D motion capture data can be used to train the subsequent 3D-reasoning module. This allows our system to take immediate advantage of advances in 2D pose estimation, such as multi-body analysis [7]. Our results also suggest that 3D inference is, in some sense, "all about 2D", at least in the case of articulated objects. Indeed, one of our surprising findings was the high performance of 2D pose estimation systems even under occlusions, suggesting that 2D estimates can in fact be reliably estimated without directly reasoning about depth. Given such reliable 2D estimates, we show that one can efficiently impute depth through simple memorization and warping of a 3D pose library.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, 2004.

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015.

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

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

[5] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *ICCV*, 2011.

[6] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *IJCV*, 2014.

[7] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. *Computer Vision–ECCV 2010*, 2010.

[8] D. A. Forsyth, O. Arikan, and L. Ikemoto. *Computational Studies of Human Motion: Tracking and Motion Synthesis*. Now Publishers Inc, 2006.

[9] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[10] M. Hofmann and D. M. Gavrila. Multi-view 3d human pose estimation in complex environment. *IJCV*, 2012.

[11] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1983.

[12] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with convolutional latent-variable models. *arXiv preprint arXiv:1507.05699*, 2015.

[13] 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*, 2014.

[14] H. Jiang. 3d human pose reconstruction using millions of exemplars. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010.

[15] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the BMVC*, 2010.

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

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

[18] C.-P. Lu, G. D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on PAMI*, 2000.

[19] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 2001.

[20] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *IEEE Transactions on PAMI*, 2006.

[21] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016.

[22] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on PAMI*, 1980.

[23] U. Rafi, J. Gall, and B. Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *Proceedings of the IEEE Conference on CVPR Workshops*, 2015.

[24] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, 2012.

[25] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *NIPS*, 2016.

[26] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

[27] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.

[28] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010.

[29] 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.

[30] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *CVPR*, 2016.

[31] 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 *Advances in neural information processing systems*, 2014.

[32] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014.

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

[34] X. K. Wei and J. Chai. Modeling 3d human poses from uncalibrated monocular images. In *ICCV*, 2009.

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

[36] H. Yub Jung, S. Lee, Y. Seok Heo, and I. Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *CVPR*, 2015.

[37] 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.