Multimodal 3D Human Pose Estimation from a Single Image

Scott Spurlock
Department of Computer Science
Elon University
sspurlock@elon.edu

Richard Souvenir
Department of Computer and Information Sciences
Temple University
souvenir@temple.edu

Abstract

In this paper, we propose a method for estimating 3D human pose from a single RGB image. Compared to methods that either provide point estimates for coordinate regression or unimodal predictions of joint locations, our approach predicts joint locations using multimodal distributions. In addition, we apply a data-driven approach to learn the conditional dependencies of the relative positions of joints. Our end-to-end approach takes as input images with either 2D or 3D labels and performs on par or better than the state-of-the-art on the Human3.6M and MPII datasets.

1. Introduction and Related Work

3D human pose estimation from a single image is a challenging inference problem with a long history in the fields of computer vision, and more recently, machine learning. A recent survey provides a general overview [17]. The challenges are due to the wide variation in possible human pose, and additional variability due to viewpoint, background, and occlusion, particularly the self occlusion of human body parts in particular poses from certain viewpoints.

In this paper, we propose an end-to-end method that focuses on three significant challenges for this task: (1) the ambiguity of 3D inference from a 2D image, (2) the relationship between the positions of different joints, and (3) the limited amount 3D labels for real-world images.

1.1. 3D Inference Ambiguity

Estimating 3D parameters from a 2D image leads to an inherent complication where many different 3D poses can be consistent with a given image. The majority of approaches sidestep this ambiguity and structure the problem as regressing the 3D locations of the joints by minimizing a least squares loss function and reporting a single (best) prediction [12, 5, 18]. Other approaches, rather than directly predicting 3D coordinates, predict related structures. One method learns to regress a 2D distance matrix encoding relationships between joints to a 3D distance matrix [14]. Pavlakos et al. proposes a volumetric representation to predict 3D heatmaps rather than coordinates [16]. Yang et al. adapts adversarial learning to distinguish between ground truth 3D annotations and 3D annotations created by a generator [21]. All of these approaches use least squares minimization and report the conditional average of the predicted 3D joint positions.

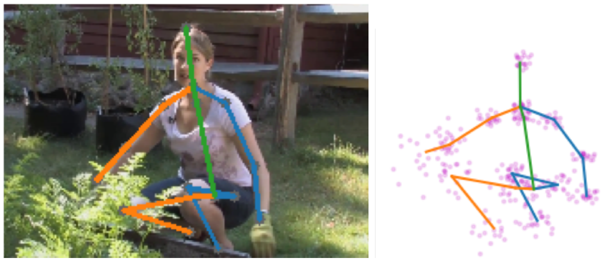

Figure 1. Our method predicts a distribution over the 3D location of each joint. On the right, 3D samples from each joint's predicted distribution are shown in magenta, with a skeleton drawn through the most likely estimates (and overlaid on the image). Joint positions that can be confidently estimated (head, shoulders) have a smaller variance, while joints that are occluded (right wrist and knee) have a larger variance, indicating more uncertainty.

There have been some alternatives to least squares minimization. One recent method lifts 2D poses to 3D by fitting a probabilistic 3D pose model [19]. Another method uses nearest-neighbor to find a suitable 3D pose based on matching similar 2D poses [3]. A similar approach matches a 2D pose estimate with a projected 3D pose by searching over possible projections from a set of virtual cameras [22]. These nonparametric approaches avoid the limitation inherent in least squares minimization, but require computationally expensive iterative fitting. Further, accuracy is limited by the number of exemplars, which imposes a trade-off between error and latency. Our approach is explicitly multimodal, estimating the conditional density over the space of possible 3D poses, given an input image, rather than a single average pose.

Concurrent to our work, a recent approach also provides multimodal predictions for human pose estimation [11].

This method differs from our work in several ways, including utilizing a two-stage training protocol rather than end-to-end training and estimating isotropic covariances.

We propose a model that predicts a multimodal probability distribution over the possible output values by incorporating Mixture Density Networks (MDNs) [2]. This approach allows for more accurate modeling of the ambiguity in human pose compared with traditional averaging methods. Predicted distributions can also be naturally aggregated over a sequence of frames for video prediction or a set of cameras for multi-view prediction.

1.2. Modeling Joint Dependencies

There have been a variety of approaches to model the interaction among human joints in the structure of the learned model. Most existing work treats the prediction of each joint separately (e.g., [12, 16]), while other approaches focus on the kinematic chain, defining parent-child relationships between, for example, knee and ankle, or elbow and wrist (e.g., [7, 9]). One recent approach proposed a pose grammar consisting of human joint dependencies based on kinematics as well as symmetry and motor coordination [5]. A potential drawback of these manually-defined parent-child relationships is that they may not fully capture the complex interaction among human joints.

We propose instead to use a data-driven approach to learn the joint relationships from the data. This idea is similar to a related method that incorporated mutual information to create a Bayesian network to model prior probabilities for 2D pose configurations [10]. Our approach learns which joints best predict the location of other (potentially occluded) joints through iterative refinement.

1.3. Overcoming Limited 3D Labels

There is a trade-off between 3D labels and image diversity; datasets with 3D annotations are typically collected in lab environments. By contrast, images labeled with 2D poses are abundant, and include real-world (or ”in the wild”) scenes. A common workaround is to follow a two-stage pipeline, where an image is first passed through a 2D estimation process, such as Convolutional Pose Machines (CPM) [20] or Stacked Hourglass [15], and then, as a separate process, the 2D estimate is lifted to 3D. Several methods take this lifting approach, which has the advantage that the relatively large amount of available 2D-annotated data can be used to train the first stage, while the more limited amount of 3D-annotated data is needed only to train the second stage [12, 5, 14]. However, potentially valuable image features are discarded due to the decoupling of the stages. In addition, error in the 2D phase is propagated to the 3D phase. Other methods propose an end-to-end training model, but require strictly 3D annotations, limiting their application to data collected in a lab environment [16, 18]. Recently, some methods combine both 2D- and 3D-annotated data in a single phase. One method uses transfer learning, allowing features learned with 2D-annotated examples to be re-used for predicting 3D [13]. Zhou et al. introduces a geometric constraint to allow loss computation for mixed batches of 2D and 3D examples [23].

Our approach follows recent work and employs an end-to-end training process that uses both 2D- and 3D-annotated data as input to learn a single integrated model. This integration helps by allowing the model to incorporate features from earlier stages in the final prediction. In addition, training with mixed 2D- and 3D-annotated data helps the learned model better generalize to real-world environments.

1.4. Contributions

In this paper, we propose an end-to-end method that focuses on three significant challenges for 3D human pose estimation from a single image. Our approach performs well compared to the state-of-the-art, both qualitatively and quantitatively. The main contributions include:

- an approach to model the inherently ambiguous one-to-many nature of 3D human pose estimation as a multimodal conditional distribution over possible poses;
- a technique for iteratively refining joint location estimates using mutual information to identify which joints most influence each other; and
- an end-to-end training approach that simultaneously integrates both 2D- and 3D-annotated examples.

2. Method

Given an image, I, the goal is to predict the 3D position of a set of human keypoints (joints). Following the most common formulation of the problem, the camera frame serves as the coordinate system, with the first two dimensions corresponding to image coordinates, and the third indicating depth in millimeters [23, 12, 16]. Poses are zero-centered about a root joint (pelvis). The objective is to minimize the mean prediction error:

$$\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{y^j} - \mathbf{\hat{y}^j}\| \tag{1}$$

where $\mathbf{\hat{y}^j}$ is the predicted position of the j^{th} joint, $\mathbf{y^j}$ is the ground truth position, and J is the number of joints in the model. In our approach, we learn the parameters of a multimodal distribution for each 3D joint location that minimizes the negative log likelihood of the target 3D joint location under the estimated probability density.

2.1. Model Architecture

Figure 2 shows the proposed architecture for our model. An input image, I, is first processed by a 2D pose estimation module, based on the stacked hourglass network [15],

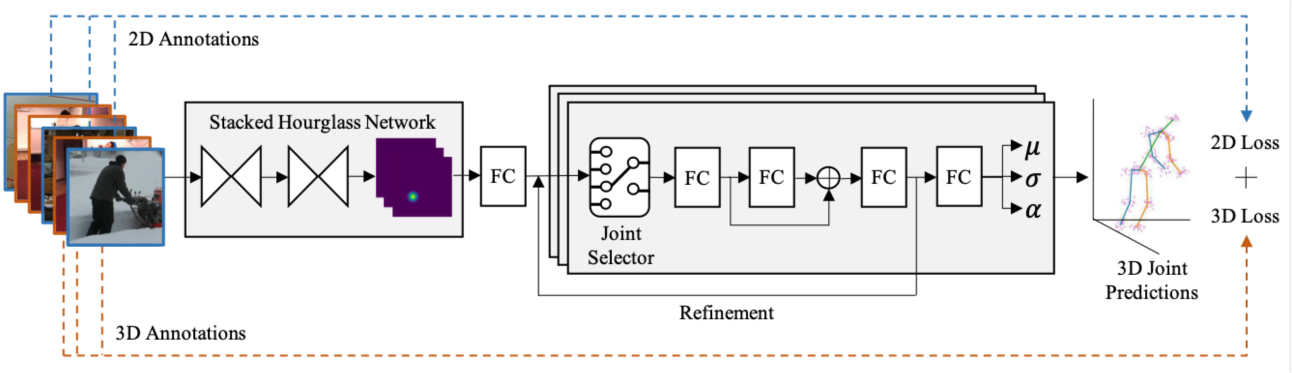

Figure 2. The proposed network is trained with mixed batches of both 2D- and 3D-annotated examples. The first stage of the network performs 2D inference and follows the stacked hourglass architecture of [15]. The second stage combines the output from the 2D stage to generate 3D predictions for each joint. The 3D stage incorporates a data-driven joint selection module to iteratively incorporate the prediction locations of related joints. The output is modeled by a mixture density network (MDN), which predicts the parameters (μ, σ, and α) for a distribution over the position for a given joint.

which produces 2D heatmaps for each joint. Next, a 3D module predicts a distribution over the position of each joint. This module iteratively refines the prediction for each joint based on the related joints (encoded by the selector) as well as the features from the 2D pose estimation module. During training, input examples are provided with either 2D or 3D labels. In Section 2.4, we describe how each case is handled.

2.2. Selecting Related Joints

The relative positions of body joints are not, in general, conditionally independent. In addition to the structural constraints imposed by the human body, the relative positions of particular joints can be both over- and under-represented in images from real-world scenes. We seek to learn, in a data-driven fashion, the mutual dependence of pairs of joints. In our model, we take a recurrent refinement approach where the predicted position of selected joints is fed back as input to the prediction module.

Mutual information is a measure of the dependence between two random variables and can help predict the utility of one joint position in determining another. To calculate the mutual information between each pair of human joints, we adopt a recent non-parametric method for estimating mutual information between dependent, continuous variables [6]. Using this method, we can identify the most informative related joints for each joint from the training data.

Figure 3 shows a graph representing the three highest mutual information relationships between joints based on the Human3.6M dataset [8]. Many of the relationships follow the kinematic chain (e.g., shoulder to elbow to wrist). Interestingly, some symmetry relationships are captured (e.g., between left and right knees), while others are not (e.g., between left and right elbows). And some less expected relationships are evident, as between elbows and hips. These mutual information relationships are represented in our network model by concatenating, as the input to a given refinement stage, the outputs from the prior stage of the joints with the highest mutual information with the given joint. This allows for the multi-stage refinement of the predicted position of each joint.

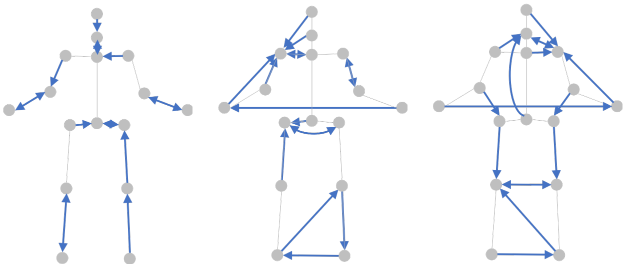

Figure 3. Data-driven joint dependencies. For the H3.6M dataset and 16-joint skeleton, the outgoing edges on the graphs show the first, second, and third most related joints, respectively, based on mutual information.

2.3. Multimodal Joint Predictions

The output of the network is a multimodal distribution for each joint location. Mixture density networks (MDN) [2] have been used to predict such distributions with neural architectures. Here, we model the conditional density over the 3D joint position, $\mathbf{y}$, for each joint using a mixture of K Gaussian distributions. This density is modeled as

$$p(\mathbf{y}|\mathbf{x},\mathbf{\Theta}) = \sum_{k=1}^{K} \alpha_k(\mathbf{x},\mathbf{\Theta})\phi_k(\mathbf{y}|\mathbf{x},\mathbf{\Theta}) \tag{2}$$

where $\mathbf{x}$ represents the input image, $\mathbf{\Theta}$ are the learned parameters of the network, α represents the weights of the components, and ϕ is the conditional density. We use the

Gaussian density,

$$\phi_k(\mathbf{y}|\mathbf{x},\mathbf{\Theta}) = \frac{1}{(2\pi)^{c/2}\sigma_k(\mathbf{x},\mathbf{\Theta})^c} \exp\left(-\frac{\|\mathbf{y}-\mu_k(\mathbf{x},\mathbf{\Theta})\|^2}{2\sigma_k(\mathbf{x},\mathbf{\Theta})^2}\right) \quad (3)$$

where μ_k is the mixture component mean, σ_k is the mixture component standard deviation, and c represents the dimensionality.

For an MDN, the outputs of the last layer are constrained so that the output can be interpreted as the parameters of a multimodal distribution. Let $\mathbf{z}$ represent the vector of outputs from the last layer of a neural network. The outputs corresponding to the standard deviations, z_k^σ, are constrained to be positive by applying the exponential function,

$$\sigma_k = \exp(z_k^\sigma) \quad (4)$$

Similarly, the outputs corresponding to the mixture weights, z_k^α, are constrained to sum to one using the softmax function,

$$\alpha_k = \frac{\exp(z_k^\alpha)}{\sum_{i=1}^K \exp(z_i^\alpha)} \quad (5)$$

We employ such a multimodal model for each joint. This mixture density formulation allows us to explicitly model the positional ambiguity in the 3D joint estimates. The output prediction is an approximation of the mode of the multimodal distribution; we select the mean, μ, of the component with the largest mixture weight, α. In Section 3, we show how this approach outperforms methods that rely on single point and unimodal distribution predictions.

2.4. 2D + 3D Hybrid Loss

The proposed network supports integrated training with batches of both 2D- and 3D-annotated examples, as shown in Figure 2. This capability is important to allow the model to learn to predict 3D positions, but at the same time to take advantage of the greater variation present in in-the-wild, 2D datasets.

In the 3D case, the loss for a given image with J joints is the sum of the negative log-likelihood of the probability estimates. Incorporating Equation 2 gives:

$$\sum_{j=1}^{J} -\log \sum_{k=1}^{K} \alpha_k(\mathbf{x},\mathbf{\Theta}^j)\phi_k(\mathbf{y}^j|\mathbf{x},\mathbf{\Theta}^j) \quad (6)$$

In the 2D case, we integrate the 3D prediction over the depth dimension and compute the loss using the 2D variant of Equation 3 with the 2D ground truth annotations. As both versions derive from proper posterior distributions, we do not perform any additional weighting or normalization to account for 2D versus 3D annotated examples.

3. Results

We implemented our model using PyTorch and trained using an Nvidia Titan Xp GPU.

3.1. Experiment Design

We follow the training protocol outlined in recent related work (e.g., [16, 23, 12]), where a hybrid (2D + 3D labels) dataset is created by combining the training sets of the Human3.6M [8] and MPII Human Pose [1] datasets.

3.1.1 Training Data

Human3.6M contains 3.6 million images captured in a controlled indoor environment. There are 7 actors performing 15 common actions, such as sitting, walking, and taking a photograph. Ground-truth 3D positions are provided for 16 joints, as well as camera calibration parameters. Following the most common protocol, we retain every 5th frame for our experiments, and group 5 of the actors (S1, S5, S6, S7, and S8) into the training set, with actors S9 and S11 reserved for testing. *MPII Human Pose* contains approximately 25 thousand in-the-wild images collected from YouTube videos of several hundred human activities and annotated with the 2D locations of 16 joints. The dataset is split into training and testing subsets.

3.1.2 Training Details

Training proceeds in two phases. Initially, the 2D pose estimation module, based on the stacked hourglass architecture, is trained from scratch using the 2D-annotated dataset, MPII, as described in Newell et al. [15]. Then the combined network is trained together using mixed batches of 2D- and 3D-annotated examples from both MPII and H3.6M. Training proceeds for 25 epochs using ADAM with a learning rate of 1e-4 and a batch size of 6. For 30 additional epochs, the learning rate is gradually reduced to 1e-6. Training images are extracted from full frames by cropping a square image based on annotations provided with the dataset, as is common in other work [8, 23].

3.1.3 Evaluation Metric

The most common evaluation metric for this task is the mean per joint position error (MPJPE) in millimeters between ground-truth and predicted 3D position. Poses are first aligned based on a root joint (e.g., center pelvis) and by scaling the predicted pose so that the total length of edges between joints matches the average total length of actors in the training set [23]. This alignment removes the necessity of using camera calibration parameters during testing. For our approach, we compute the MPJPE to the mean of the

Method	MPJPE
$K = 1$	68.7
$K = 3$	61.0
$K = 5$	**59.9**
$K = 7$	60.6

Method	MPJPE
$J_R = 0$	62.0
$J_R = 1$	61.3
$J_R = 2$	61.1
$J_R = 3$	**59.9**

Table 1. Component evaluation using MPJPE (lower is better) on the Human3.6M dataset. (Left) Comparison of the number of mixture components, K. (Right) Comparison of the number of conditionally related joints, J_R.

component with the highest mixture weight in the multimodal distribution.

3.2. Component Evaluation

We first evaluate the effect of the key contributions introduced in our model: (1) multimodal posterior predictions, (2) iterative refinement, and (3) data-driven mutual joint selection. These results were evaluated on the Human3.6M testing set, which includes 3D annotations.

Table 1 (left) shows the effect of varying the number of mixture components, K, on the 3D pose estimation accuracy on the Human3.6M dataset. In general, MPJPE decreases as K increases, saturating at $K = 5$. For the remaining experiments, we choose $K = 5$ mixture components.

We also evaluated the contributions of two intertwined contributions, iterative refinement and the number of related joints. For each variation, we train and test a model following the above protocol with $K = 5$ mixture components. In general, incorporating the iterative refinement stages improves prediction accuracy (e.g., for $J_R = 3$, the MPJPE increases from 59.9 mm with iterative refinement to 62.0 mm without). Table 1 (right) shows the results using an increasing number of related joints, J_R for each prediction. The performance tends to increase as additional joints are added at the cost of additional weights and memory usage for the network. In the following experiments, our model employs $K = 5$ mixture components, iterative refinement, and $J_R = 3$ conditional dependencies.

3.3. Comparison to Recent Methods

Table 2 details our results for 3D pose estimation on the Human3.6M dataset, as well as results reported by several related methods. We report results for both our full model ($K = 5$) and a baseline ($K = 1$), which is equivalent to providing a unimodal prediction for each joint. The average error for the baseline ($K = 1$) model, 68.7, is similar to several other recent methods. With the exception of a few outlier poses, our full model ($K = 5$) consistently performs at or near the top of the list compared to recent related approaches. The overall best performer [11] was developed concurrently to our method and also primarily relies on multimodal predictions using MDNs.

Figure 4 shows visual results from our model. Each input image is overlaid with the projected skeleton based on the most likely estimates from the model. The skeleton is shown in 3D from a different view angle to the right of each input image. Our model is able to accurately predict 3D joint locations for a wide variety of human poses, even in the presence of occlusion. We also evaluate our method's ability to generalize against the MPII validation set, which includes challenging in-the-wild images not previously seen by our model during training. Because this dataset includes only 2D annotations, we show qualitative results only (bottom three rows of Figure 4).

Figure 5 shows examples of predicted distributions over 3D joint positions for input images from the Human3.6M dataset. In the first example, samples drawn from each joint's predicted distribution (shown as magenta circles) are generally tightly clustered around the actual positions, although greater spread can be observed for the elbows and, particularly, the wrists, suggesting greater uncertainty for these predictions. The following examples are increasingly difficult due to greater self-occlusion, with several joints unobserved in the bottom input images. The samples drawn from the predicted distributions reflect greater uncertainty, with much more variation around the actual joint positions, particularly the occluded joints.

3.4. Cross-dataset Evaluation

We demonstrate the ability of our method to generalize beyond the datasets used for training by evaluating on images from an unseen dataset. For this experiment, we train a model as described above using the MPII training and H3.6M datasets. For testing, we use MPI-INF-3DHP [13], a 3D-annotated dataset collected using markerless motion capture, which contains a variety of actors, actions, and backgrounds. The test set consists of 2929 images from 7 different actions performed by 6 actors from three different scenarios: (1) studio with a greenscreen background, (2) studio without greenscreen, and (3) outdoors.

Following prior work [13, 23] using this data, the root joint (pelvis) is first aligned, and it is assumed that the pose scale is known. The evaluation metrics include the 3D Percentage of Correct Keypoints (3DPCK) with a threshold of 150 mm, as well as the area under the curve (AUC) for a range of PCK thresholds. For both, higher is better.

Table 3 compares our method to a recent approach [13] on this dataset. Our approach outperforms [13] on average and across all scenarios except for the examples from the greenscreen. We note our method greatly outperforms [13] on the outdoor sequences, which are the least similar to the images with 3D annotations we used for training, which were collected in an indoor studio environment. This further highlights the ability of model to incorporate the abundant 2D labels to generalize to new 3D predictions.

Method	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.
Ionescu et al. PAMI-16 [8]	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3
Du et al. ECCV-16 [4]	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2
Tekin et al. ICCV-16 [18]	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2
Chen & Ramanan CVPR-17 [3]	89.9	97.6	89.9	107.9	107.3	139.2	93.6	136.0
Pavlakos et al. CVPR-17 [16]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3
Mehta et al. 3DV-17 [13]	52.6	64.1	55.2	62.2	71.6	79.5	52.8	68.6
Zhou et al. ICCV-17 [23]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6
Martinez et al. ICCV-17 [12]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Fang et al. AAAI-18 [5]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7
Yang et al. CVPR-18 [21]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7
Li and Lee CVPR-19 [11]	**43.8**	**48.6**	**49.1**	**49.8**	**57.6**	61.5	**45.9**	**48.3**
Ours ($K = 1$)	54.8	63.9	59.5	67.8	68.4	55.5	60.0	83.9
Ours ($K = 5$)	48.1	57.0	50.8	61.4	58.3	**48.6**	53.3	71.2
Method	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Ionescu et al. PAMI-16 [8]	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Du et al. ECCV-16 [4]	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Tekin et al. ICCV-16 [18]	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Chen & Ramanan CVPR-17 [3]	133.1	240.1	106.6	106.2	87.0	114.0	90.5	114.1
Pavlakos et al. CVPR-17 [16]	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Mehta et al. 3DV-17 [13]	91.8	118.4	65.7	63.5	49.4	76.4	53.5	68.6
Zhou et al. ICCV-17 [23]	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Martinez et al. ICCV-17 [12]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang et al. AAAI-18 [5]	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Yang et al. CVPR-18 [21]	69.2	85.2	57.4	58.4	**43.6**	60.1	47.7	58.6
Li and Lee CVPR-19 [11]	**62.0**	73.4	**54.8**	**50.6**	56.0	**43.4**	**45.5**	**52.7**
Ours ($K = 1$)	128.5	67.4	79.0	64.1	49.9	69.1	56.6	68.7
Ours ($K = 5$)	108.2	**58.6**	68.9	57.7	44.2	59.3	47.9	59.9

Table 2. Comparison to several recent approaches using Mean Per Joint Position Error (MPJPE) in millimeters on the Human3.6M dataset. Some results are reported from [21].

Scenario	3DPCK		AUC	
	[13]	Ours	[13]	Ours
Greenscreen	**70.8**	68.6	-	38.6
Studio	62.3	**66.9**	-	36.7
Outdoor	58.5	**71.5**	-	43.2
All	64.7	**68.6**	31.7	**39.0**

Table 3. Results for cross-dataset evaluation on unseen data compared to [13] on the MPI-INF-3DHP dataset using 3DPCK and AUC (where reported). For both metrics, higher is better.

4. Conclusions

We presented an end-to-end method for 3D human pose estimation from a single image that (1) explicitly models the ambiguity inherent in 3D inference from a 2D image using multimodal distributions, (2) learns the conditional dependencies between the positions of different joints, and (3) incorporates both 2D and 3D labels to account for the limited amount 3D labels for real-world images. The experimental results demonstrate that our method outperforms approaches which rely on a single or ”average” prediction. Future directions include hyperparameter (e.g., number of mixture components, joint selection) learning as part of the end-to-end training process and combining multiple multimodal predictions for multiview or video settings.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014. 4

[2] C. M. Bishop. Mixture density networks. Technical report, Aston University, Birmingham, 1994. 2, 3

[3] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 1, 6

[4] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*. Springer, 2016. 6

[5] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI Conference on Artificial Intelligence*, 2018. 1, 2, 6

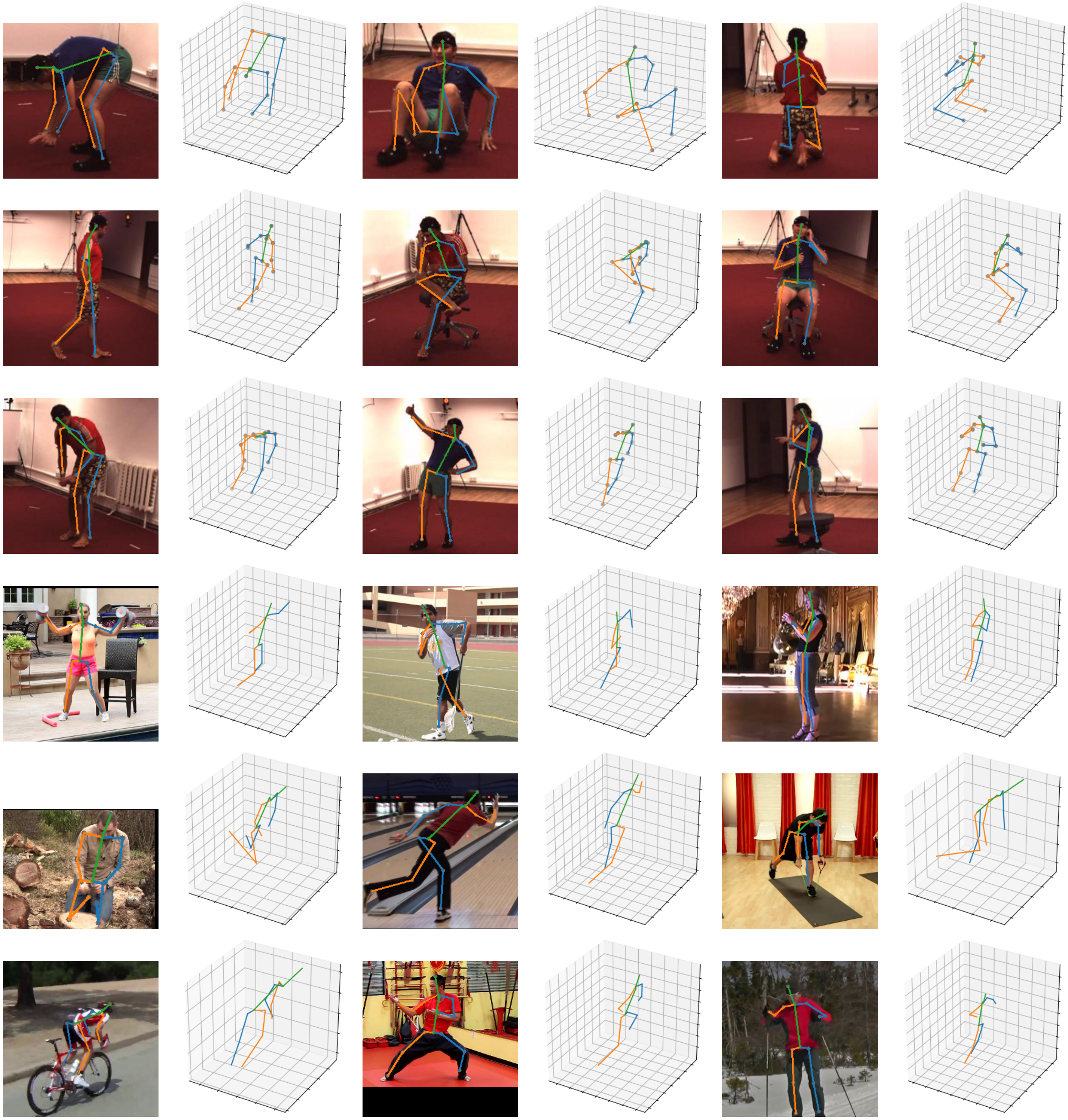

Figure 4. Predicted poses for example images from the Human3.6M dataset (top three rows) and MPII validation dataset (bottom three rows). For each input image (left), the predicted pose is overlaid, and the 3D prediction is shown (right).

[6] S. Gao, G. Ver Steeg, and A. Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, 2015. 3

[7] S. Hauberg, S. Sommer, and K. S. Pedersen. Gaussian-like spatial priors for articulated tracking. In *European Conference on Computer Vision*, pages 425–437. Springer, 2010. 2

[8] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on*

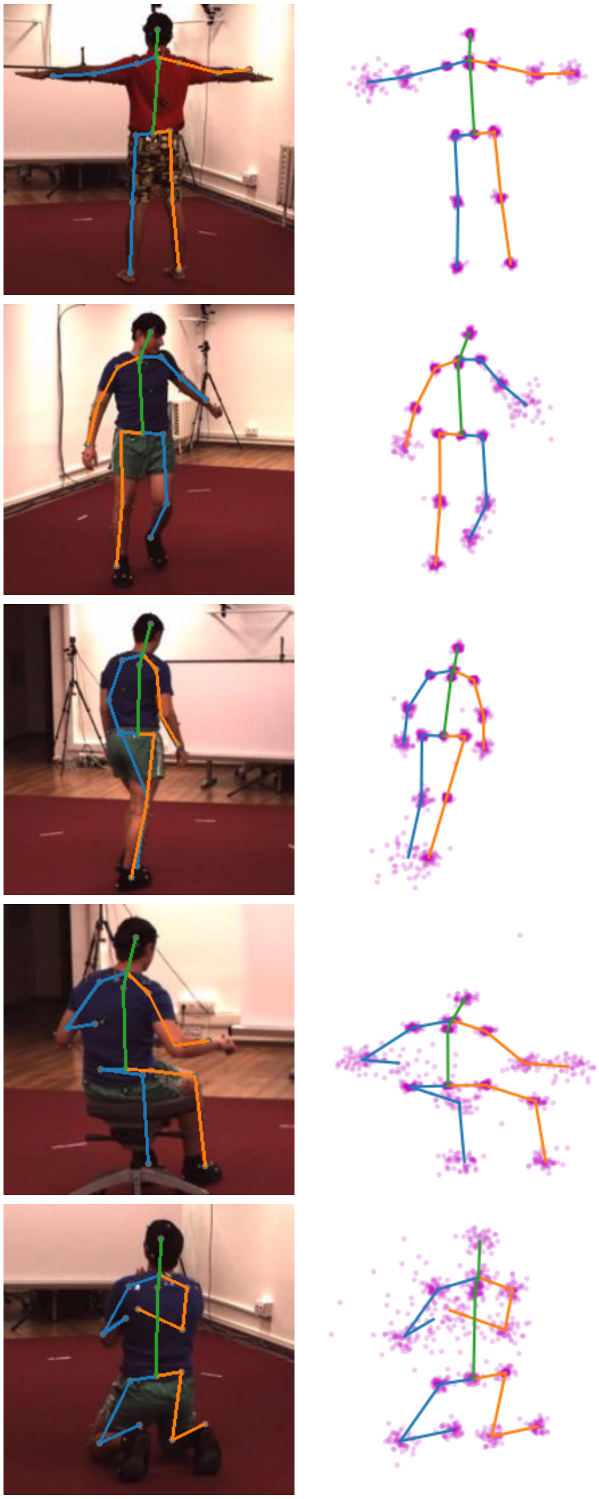

Figure 5. Predicted poses for example images from the Human3.6M dataset. For each input image (left), the predicted pose is overlaid, and the 3D prediction is shown, including a random sampling of 50 predictions from each joint's distribution (right).

Pattern Analysis and Machine Intelligence, 36(7), 2014. 3, 4, 6

[9] M. W. Lee and I. Cohen. Human upper body pose estimation in static images. In *European Conference on Computer Vision*, pages 126–138. Springer, 2004. 2

[10] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. A non-parametric bayesian network prior of human pose. In *International Conference on Computer Vision*. IEEE, 2013. 2

[11] C. Li and G. H. Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2019. 1, 5, 6

[12] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision*. IEEE, 2017. 1, 2, 4, 6

[13] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision*. IEEE, 2017. 2, 5, 6

[14] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 1, 2

[15] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*. Springer, 2016. 2, 3, 4

[16] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 1, 2, 4, 6

[17] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152, 2016. 1

[18] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *International Conference on Computer Vision*. IEEE, 2017. 1, 2, 6

[19] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *Conference on Computer Vision and Pattern Recognition*, 2017. 1

[20] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016. 2

[21] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 1, 6

[22] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016. 1

[23] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *International Conference on Computer Vision*. IEEE, 2017. 2, 4, 5, 6