A Review Of 3D Human Pose Estimation From 2D Images

Kristijan BARTOL*[1], David BOJANIĆ[1], Tomislav PETKOVIĆ[1],
Nicola D'APUZZO[2], Tomislav PRIBANIĆ[1]
[1] University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia;
[2] Hometrica Consulting, Ascona, Switzerland

https://doi.org/10.15221/20.29

Abstract

Human pose estimation task takes images as input and extracts a set of locations representing the predefined body joints and the sparse connections between the joints, called the body parts. A pose can be estimated from single or multiple frames, in a single (monocular) or multi-view (stereo) setup and for a single person or multiple people in the scene. In this work, we provide an overview of the classic and deep learning-based 3D pose estimation approaches. We also point out relevant evaluation metrics, pose parametrizations, body models, and 3D human pose datasets. Finally, we review state-of-the-art pose estimation results, briefly discuss open problems, and propose possible future research directions.

Keywords: 3d computer vision, human pose estimation, review

1. Introduction

Human pose estimation is one of the fundamental computer vision tasks, whose applications span from action detection [37] and recognition [36] to human tracking [27], augmented reality [5], video surveillance [63], sports [8] and other. A human pose can be described as an articulated body [68]. An articulated body is an object composed of more than one rigid part connected by joints, allowing rotational and translational motion under three degrees of freedom [65]. The pose estimation can be defined as the search for a specific pose P in a space of all articulated poses Π. Another definition of the pose estimation is the problem of extracting a set of image features that correspond to human body joints (also known as keypoints). The set of joints J is a set of pixel coordinates $J_{2d} \in R_{2n}$, in case of 2D pose estimation, and a set of 3D location coordinates $J_{3d} \in R_{3n}$, expressed in millimeters, in case of 3D pose estimation (Fig. 1). The joints are sparsely connected via rigid body parts. In case of multi-person pose estimation, the goal is to extract multiple joint sets, $J_i \in M$, i.e., multiple poses, $P \in \Pi$. Note that the expected number of joints $x \in J$ in the pose model may vary between the datasets (Tab. 2) and between the algorithms (Tab. 1).

The pose estimation is generally difficult due to unknown location of a human body in an image, unknown number of people in the scene, (self-) occlusions, a variety of environments [71, 43, 28] and actions [37, 36] and diverse body shapes [7, 41] and clothes [58]. Reconstruction of a 3D human pose from a single 2D image is particularly challenging due to 2D-to-3D elevation ambiguities [69, 29, 42, 44, 67, 66, 54,49]. To compensate for the lack of information in a single-view, multi-view [2, 30, 12, 59, 47] and multi-frame (video stream) [74, 76, 33, 26, 56] approaches to 3D human pose estimation have also been proposed. In most of the cases, 3D pose estimation approaches are top-down, meaning that they first locate a bounding box of the person and then apply the pose reconstruction procedure [30, 67, 48, 77], or simply assume a single-person prediction in the given image. On the other hand, bottom-up approaches do not know the number of people upfront; they first detect individual body parts and then compose them into complete human poses [9].

Recent deep learning-based approaches are often classified into regression- and detection-based [54, 66, 67]. The detection-based methods generate a likelihood heat map foreach joint and locate the joint as the point with the maxi-mum likelihood in the map [54]. The regression-based methods produce a continuous output directly, without using the likelihoods [42, 66]. Both detection- and regression-based approaches use multi-stage refinements [67, 44, 54, 59] to improve the performance. Body priors (for example, constant body part lengths through the frames or human body symmetries [35]) are also proven to be beneficial for the optimization [66, 69, 29]. Some deep learning models are able to learn the priors without the explicit clues [30, 79].

In general, the existing 3D human pose estimation approaches can therefore be classified based on several distinctive properties, summarized in Tab. 1: the number of people in the scene, the number of cameras, whether they exploit the temporal context, whether the method is regression- or detection-based approach, is it a top-down or bottom-up approach. The Tab. 1 is discussed in detail in the Sec. 5.

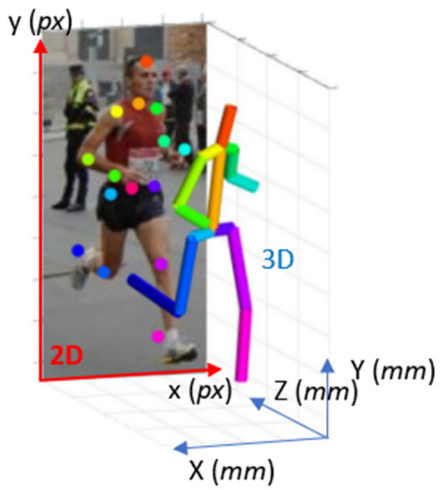

Fig 1. An example of 2D and 3D human pose estimation on the MPI-INF-3DHP dataset [43]. The 2D pose locations are expressed in pixels and 3D pose in millimeters. The estimations are made using the 3DPoseNet model [43].

A recent review paper by Chen et al. [15] covers 2D and 3D deep learning-based human pose estimation approaches. In contrast, Perez-Sala et al. [57] published a paper in 2014.that covers classical approaches. In 2016., a review focused on monocular pose estimation was published by Gong et al. [22]. A work most similar to ours was published by Sarafianos et al. [62], covering 3D human pose estimation approaches. They provide a comprehensive review over state-of-the-art, propose a taxonomy of the approaches and discuss the open problems at that point in time. In this work, we use a similar approach to 3D human pose estimation overview as [62], but considering the most recent line of work, that are mostly based on deep learning. We also reflect on the pioneering works in 2D-to-3D pose elevation [74, 35, 68].

Tab 1. An overview of the common properties of the recent state-of-the-art 3D pose estimation approaches. Note that TEMP stands for temporary, ABS for absolute pose, REG for regression, DET for detection, TD for top-down and BU for bottom-up.

	#PEOPLE	#CAMERA	TEMP?	ABS?	REG/ DET	TD/BU	YEAR
[30]	single	multi	no	no	det.	TD	2019.
[67]	single	mono	no	no	both	TD	2018.
[8]	multi	multi	yes	no	reg.	TD	2019.
[13]	single	mono	yes	no	reg.	TD	2020.
[73]	single	mono	no	no	reg.	TD	2019.
[51]	single	mono	no	no	det.	TD	2018.
[44]	multi	mono	no	no	det.	BU	2018.
[32]	single	multi	no	no	det.	TD	2019.
[26]	single	mono	yes	no	reg.	TD	2018.
[59]	single	multi	no	no	det.	BU	2019.
[47]	single	multi	yes	no	reg.	TD	2019.
[60]	single	mono	no	no	reg.	TD	2017.
[49]	single	mono	no	no	reg.	TD	2017.
[78]	single	mono	no	no	reg.	TD	2019.
[46]	multi	mono	yes	no	det.	TD	2019.
[21]	multi	mono	no	no	det.	BU	2020.
[38]	single	mono	no	no	reg.	TD	2019.
[10]	single	mono	no	no	det.	TD	2019.
[48]	multi	mono	no	yes	both	TD	2019.
[40]	single	mono	yes	no	reg.	TD	2019.
[25]	single	multi	no	no	det.	TD	2020.
[14]	single	multi	no	no	det.	TD	2019.
[16]	single	mono	yes	no	det.	TD	2019.
[20]	single	mono	no	no	-	TD	2019.
[17]	single	mono	yes	no	det.	TD	2020.

In the remainder of the paper, we review human pose estimation datasets and then provide a comprehensive review of 3D pose estimation approaches based on the taxonomy in the Tab. 2. In the evaluation section (Sec. 4), we list and explain the metrics used for the quantitative evaluation. Due to the amount of work, diversity and the success of deep learning-based methods, we separately summarize their properties in the Tab. 3, 4, 5. In the Sec. 4, we analyze the results of the state-of-the-art methods, based on the most commonly used evaluation metrics, namely mean per-joint position error (MPJPE). In the discussion, we gather the insights obtained from reviewing the literature and point out the current open problems of 3D pose estimation. Finally, we conclude by identifying possible future research directions.

2. Datasets

Large high-quality 3D human pose estimation datasets are crucial for the success of deep learning models. The precise 3D annotations of human body joints serve as a direct supervision for models to learn how to detect the joints and resolve 2D-to-3D elevation ambiguities [30, 59, 17, 42, 25, 16]. However, acquiring 3D data in the real world is challenging and is done in specially designed studios [31] and indoor environments, using wearable IMU sensors [70]. Therefore, labeled 3D pose estimation datasets often lack a diversity of environments and backgrounds, especially in-the-wild examples (outdoors and sports activities, unusual and spontaneous human poses, etc.). One way to cope with the problem of environment and background diversity is to generate the data synthetically [71, 28]. In the remainder of the section, we briefly review the existing 3D human pose datasets - an overview is given in Tab. 2.

Table 2. An overview of 3D human pose estimation datasets. The last column is discussed in the Sec. 5.

	#Camera	#People	#Joints	#Frames		
				Train	Val.	Test
Camp. & Shelf [4]	3-5	2-3	14	-	-	3k
Dense [23]	1	Many	-	All (50k)		
Panoptic [31]	521	2-10	15	All (65 videos)		
Unite [34]	1	1	-	-	-	8k
3DPeople [58]	1	1	16	All (2M)		
HumanEva [64]	7	1	15	6.8k	6.8k	24k
Human3.6M [28]	4	1	17	1.5M	0.6M	1.5M
3DHP [43]	14	1-8	15	All (1.3M)		
Surreal [71]	1	1	-	All (6M)		
Faust [7]	1	1	-	100	100	200
Total Capture [70]	8	1	19	All (2M)		
PosePrior [1]	1	1	18	-	-	-

One of the first publicly available (multi-person) 3D pose estimation datasets were Campus and Shelf [4], consisting of a series of frames sampled from two few-minute videos. The Campus and Shelf datasets are too small to be used for training deep learning models but are used for an evaluation of multi-person pose estimation methods [4, 12, 19]. In terms of modern pose estimation datasets, Campus and Shelf are relatively small, lacking variety of people, poses, activities and in-the-wild examples. Instead of Campus and Shelf, CMU Panoptic [31] and MuCo-3DHP [43] datasets are used for the evaluation of multi-person pose estimation methods.

CMU Panoptic dataset is a multi-person 3D human pose dataset, capturing the 3D motion of people engaged in different social interactions (dancing, playing instruments, social games, etc.). The dataset is recorded in a specially designed, Panoptic Studio, with over 500 VGA and HD cameras and10 Kinect v2 RGB+D sensors, distributed over the surface of geodesic sphere with a 5.49m diameter. MuCo-3DHPdataset is an extension of MPI-INF-3DHP dataset [43] for multiple people. MuCo-3DHP is interesting as it copes with aforementioned diversity of the environments (indoors and outdoors), backgrounds (extended with synthetically generated ones) and in-the-wild examples.

Regarding single-person datasets, an MPI PosePrior [1] captures a variety of stretching poses from an actor using wearable sensors. The data is used to learn pose-dependent joint angle limits. Total Capture dataset [70] uses 8 HD cameras, 4 male and 1 female actors, each performing four diverse performances. There is a total of almost 2M captured frames of synchronized video. For capturing the data, they used IMU sensors. Human3.6M [28] is a large-scale single-person dataset, acquired by recording the performance of 5 female and 6 male subjects, under 4 different viewpoints (indoors).To extend to outdoor scenes and enrich the dataset, they also provide controlled mixed reality evaluation scenarios where3D human models are animated using motion capture and inserted in complex real

environments, viewed with moving cameras, under occlusion. A similar, but an order of magnitude smaller dataset is HumanEva [64]. They used 7 cameras (4 grayscale and 3 color) and did not extend above the indoor configuration.

There are two popular synthetic datasets, SURREAL [71] and 3DPeople [58]. SURREAL is the first large-scale person dataset, it contains 2D and 3D pose skeletons, depth and body part segmentation maps, optical flow and surface normal ground truth for video input. In total, it contains 6 million frames. 3DPeople dataset is very similar to SURREAL; it contains around 2 million images of 40 male and 40 female actors performing 70 actions. Every subject-action sequence is captured from 4 camera views.

Finally, there are a few datasets [7, 23, 34] that do not provide 3D pose annotations directly; instead, the annotations can be extracted from the volumetric or surface representations.

3. Methods

One of the early works, by Webb [74], analyzes the problem of interpreting images of moving jointed objects (i.e. the articulated bodies like the human pose [68]), in particular, the problem of estimating the body parts' lengths. Similar to the modern definition of a body pose model [28], Webb defines the jointed object model as a set of: joints, rigid parts (body parts) and feature points (keypoints), where the keypoints coincide with the feature points. The proposed algorithm is based on a simple observation that the rigid part length is correctly observed when the angle between the rigid part and the camera's z-axis is right. The observation is based on the assumption that the rigid part is observed under the right angle during the movement and that the object's distance from the observer is known.

Lee and Chen in 1985. [35] already knew that the 3D pose estimation from 2D feature points reduces to determining the rigid parts' lengths and directions (whether they point towards or against the camera). To resolve some of the binary direction ambiguities for the rigid parts, they introduced the angle constraints on the flexion of shoulders, hips, elbows, knees, pelvis and neck. A work by Taylor from 2000. [68] uses a weak perspective projection (an ortographic projection with an unknown scale) to find a family of 3D pose solutions, given a single uncalibrated image. Taylor assumes that the correspondence between the joints in the model and point features in the image, as well as the the relative lengths of the segments in the models, are known. Additionally, given the rigid part lengths, the scale of the weak perspective can be determined. To recover the pose unambiguously, the user specifies which end of the body part is closer to the observer. A work by Wei and Chai from 2010. [76] propose 3D pose recovery from 2D pose estimations in a two-step optimization procedure - the first to estimate a skeletal size and the camera parameters and then use the estimations to reconstruct the poses in a so-called joint-angle space.

A 3D pictorial structures model, first published in 2013. [2], and then revisited in 2016. [4], cope with the multi-person pose estimation problem from multiple views. Due to a large number of possible poses in a multi-view setup, they first generate a reduced state space by triangulation of the corresponding pairs of body parts, obtained by the part detectors in each camera view. The key idea of the 3D pictorial structures model is the use of multi-view unary potential functions, that take into account the 2D detection confidences, multi-view part visibility, temporal consistency and the pixel reprojection error between the views. Finally, they balance the potentials' influence by learning the model parameters using a variation of an SVM algorithm. A dual-source approach [29] from 2016. trains a complex machine learning model given two inputs - a generated 3D pose space and an annotated 2D image. The image is used to learn a pictorial structures model for 2D pose estimations. The 3D pose is selected from the pose space based on minimizing the reprojection error between the 2D and the 3D pose.

In the recent years, almost all of 3D pose estimation methods are based on deep learning. Still, there are a few successful classical attempts. A movement of a human body can be described using a kinematic chain model. A kinematic chain space (KCS) algorithm [72] first translates a human body into a kinematic chain space and then optimizes a nuclear norm. The algorithm is not applicable only to human skeletons, but also to other kinematic chains like animals or industrial robots. An improvement over the pictorial structures' idea [4] is to first cluster the detected bodies throughout the frames and then apply the 3D pictorial structures model [19]. Even though the clustering step reduces the overall algorithm's complexity, it does not scale well for the larger number of views. Instead of processing the inputs from multiple views simultaneously, a cross-view tracking algorithm [12] uses an iterative strategy. The algorithm takes noisy 2D pose estimations as inputs and associate them among all pairs of views by exploiting temporal consistency. A convincing real-time performance was presented with 12 to 28 camera views. Structure-from-articulated-motion [33] is an optimization algorithm that ap-plies an additional articulated structure term as a soft constraint on top of the classic non-rigid structure-from-motion problem [18].

3.1. Deep learning

An approach by Chen and Ramanan [11] is similar to a dual-source approach [29] in a way that they also generate a library of 3D poses before learning. The 3D poses are projected to virtual camera views to obtain correspondent (2D, 3D) pose pairs. The prediction is made by selecting the most similar 3Dpose to an estimated 2D pose. The 2D pose estimation is based on the predicted heatmaps using a deep learning model called convolutional pose machine (CPM) [75], that was proven to be successful for 2D pose estimation [9]. An early deep learning-based approach, called a kinematic pose regression [80], is the first model to exploit the structural constraints of a human body in a fully-differentiable manner. The model's architecture is simple, consisting of the convolutional layers that end with a fully connected layer producing the motion parameters. The motion parameters are mapped to the joints in a so-called kinematic layer and the ground truth 3D joints are used as a supervision. Sparseness-meets-deepness [79] is the first method that uses a variation of standard sparse pose representation (Fig. 1) to learn a deep model. They use the idea from [1] to learn a pose prior and also employ the temporal smoothness. To integrate pose prior and smoothness learning into the model in a differentiable manner, they use 2D pose estimations as latent variables in a form of heatmaps. A work by Park et al. [53] concatenates 2D pose estimations with the extracted image features. Instead of using relative position between of the 3D joints and the root joint (pelvis) as a ground truth, they have shown that using the relative positions with respect to multiple joints improves their learning. Li and Lee [38] learn a mixture density network [6] to generate multiple possible 3D poses' hypotheses from a single monocular image. Another approach that generates the hypotheses is a deep pose consensus approach [10]. In contrast to [38], it generates partial hypotheses, for each group of the joints. The estimated joints are aggregated into poses in the final part of the model.

A simple yet effective deep learning-based approach to 3D pose reconstruction from 2D observations [42] is the first remarkably successful regression-based model. The model was able to learn 2D-to-3D pose correspondences and ambiguities by using only 2D pose annotations as input and 3D pose annotations as an expected output, with no image data. The work sparkled the interest in the regression-based methods for 3D pose estimations and is still used as an example for the evaluation protocols (see Sec. IV). Similar to the simple yet effective approach [42], a distance matrix regression [49] also learns 2D-to-3D pose correspondence. Instead of directly using 2N- and 3N- dimensional pose representations, they first represent the poses using N×N matrices of Euclidean distances between every pair of joints, and then formulate a problem as 2D-to-3D distance matrix regression. An adversarial approach to 3D-from-2D correspondences [20] learns a discriminator to distinguish between the real and the generated 3D poses. The generator first randomly generates relative depths between the joints but, with time, learns the feasible depth offsets and therefore - the pose prior. Another adversial approach is RepNet [73]. In contrast to the methods described above [42, 49, 20], it completely ignores 2D-to-3D joint correspondences. The discriminator learns a distribution of 3D poses and the the generator learns a distribution of detected 2D poses (obtained using the Stacked-Hourglass 2Dpose detector [50]) to a distribution of 3D poses, supervised by the discriminator.

A maximum-margin approach [39] learns two separate sub-networks to embed the estimated pose and the 3D pose into a common space. The score of the prediction is the dot-product between the two embeddings. LCR-Net [61] predicts a multi-person 3D poses from a single image in three steps: by first generating multiple 2D pose proposals, then scoring the proposals using the classifier and finally refining the poses both in 2D and 3D using the trained regressor. The pose proposals are obtained by first finding the bounding boxes and then generating multiple sets of joint locations for every bounding box. A compositional human pose regression [66] uses bones instead of joints as pose representation and show that the bones more stable and easier to learn than joints. The key to the method's success is the separation of the 2D part (pixel locations) of the joint predictions from the depth part. They also exploit the joint connection structure to define a loss function that encodes long-range interactions between the bones. A semi-supervised approach [47] exploits temporal relations between the multi-camera views to handle un annotated and uncalibrated videos. They formulate a multiview-consistent and rigid rotation-invariant 3D pose representation and refer to it as a canonical pose. The canonical pose is obtained by constraining the bone connecting the pelvis to the left hip joint to be always parallel to XZ plane. The advantage of a canonical pose compared to a view-specific pose is that it does not need to change the orientation with variations in the camera view.

Regarding recent detection-based approaches, the com-pressed volumetric heatmaps [21] propose to use high resolution input images, transform them into a compressed volumetric representation using an autoencoder network and then use a second model to decode the compression. A coarse-to-fine approach [54] uses a volumetric (voxel) representation of the scene. Multi-stage architecture first outputs coarse voxel predictions and refines them throughout many encoder-decoder subnetworks. A

cross-view fusion [59] extracts heatmaps from multiple views and then use a newly proposed recursive pictorial structure model [4] to generate 3D pose predictions. A very similar approach was seen in a 2-year earlier work by Pavlakos et al. [55], but they used a regular 3D pictorial structures model. A marginal heatmaps approach [51] uses heatmaps to predict object's margins instead of predicting joints directly. The approach is very successful, being top #1 on the MPI-3DHP dataset's benchmark (see Sec. 4). The marginal heatmaps approach uses integral regression [67].

An integral human pose regression [67] shows that the detection- and regression-based approaches can be combined. Problem with the detection-based approaches is the max-operator applied on the heatmaps that is not differentiable, effectively acting as a postprocessing step. Instead of max-operator, the authors propose to use the soft-argmax and call the approach an integral regression. As the operation is now differentiable, the joint predictions can be regressed using 3Dground truth. Learnable triangulation approach [30] exploits the combination of detection- and regression-based approaches from [67] in a multi-view camera configuration. Assuming that the camera locations are known, they pose a problem as a differentiable triangulation. The learning triangulation approach was proven to be the most successful multi-view and 3D pose estimation approach of all, at the moment of writing this paper (see Sec. 4). Inspired by classical stereo matching problem [24], the epipolar transformer model [25] employs differentiable epipolar constraints on the pairs of views, assuming known camera parameters. Instead of combining 2D features via triangulation, the idea is to search for the correspondences on the epipolar lines, hopefully leading to 3D-aware features. Another approach that exploits the epipolar geometry [32] shows that the epipolar constraint can be exploited even without the unknown camera parameters, in a self-supervised fashion. Camera parameters are found implicitly, using the known correspondences between the joints. The model predicts 2D poses from the two views and then uses these poses to predict the 3D pose. In a separate branch, 3D poses are directly predicted from each image, separately. Supervision comes from the similarity between the 3D pose estimation from the two branches. An approach by Xuet al. [78] proposes to learn a pose grammar to explicitly incorporate knowledge about human body configuration, for example, kinematics, symmetry and motor coordination. The grammar is learned using a recurrent-type network on top of a base convolutional network that captures pose-aligned features from a single image.

Trajectory space factorization [40] utilizes matrix factorization for sequential 3D human pose estimation. The 3D poses in all frames are represented as a motion matrix factorized into trajectory bases matrix and a trajectory coefficient matrix. The trajectory bases matrix is precomputed from matrix factorization approaches such as Singular Value Decomposition (SVD), and the problem of sequential 3D pose estimation is reduced to training a deep network to regress the trajectory coefficient matrix. A method by Hossain et al. [26] uses an LSTM network to utilize the temporal information across a sequence of 2D pose locations to estimate a sequence of 3D poses. VideoPose3D [56] is a semi-supervised model that jointly learns trajectory and pose submodels based on 3D-to-2D projection error. They also learn a supervised model based on labeled 2D poses and add a soft constraint to match the mean bone lengths of the unlabeled predictions to the labeled ones. The proposed architecture used in Video-Pose3D is called temporal convolutional networks (TCNs). Anatomy3D [13] decomposes the task into bone direction and bone length prediction. A shown in [35], knowing bones' lengths and directions is sufficient to derive 3D joint locations. Additionally, they employ an implicit attention mechanism to feed the 2D keypoint visibility scores into the model as extra guidance, which significantly mitigates the depth ambiguity in many challenging poses. VNect [45] is a real-time, single-person monocular pose estimation system. The system tracks a bounding box throughout the frames. Based on a person-centered bounding-box crop, the trained convolutional network predicts 2D heatmaps and 3D location maps for all joints. The temporal filtering, smoothing and the skeleton fitting is applied on the 2D and 3D pose estimations to obtain temporally coherent 3D poses. XNect [46] is a multi-person improvement over VNect, consisting of three stages. The first stage uses convolutional network to predict heatmaps representing the 3Djoints, in a bottom-up fashion. The second stage uses a fully connected network for every person in the image, in parallel, to obtain the complete 3D poses. The third stage is similar to VNect as it applies a temporally stable kinematic skeleton fitting.

Occlusion problem is tackled in the occlussion-aware-network approach [16]. The model generates 2D confidence heatmaps to detect the unreliable, occluded joints. The occluded joints are fed into a temporal convolutional network [56] that predicts the joint locations based on optical flow and temporal consistency. Single-shot multi-person approach [44] proposes so-called occlussion-robust pose-maps (ORPM)that introduce redundancy into the location maps proposed by VNect [45]. Redundancy is added by allowing the read-out of the complete pose in some of the location maps. The number of ORPM has a fixed number of outputs, but the redundancy still enables the encoding the poses of multiple overlapping people. A very practical and successful monocular approach (see Sec. 4) is

proposed by Cheng et al. [17], coping with the problem of human pose estimation in video. As people in videos appear in different scales and have various motion speeds, they apply multi-scale spatial features for 2Djoints' predictions in every frame, and multi-stride temporal convolutional networks (TCNs) [56] to estimate 3D joints. They also explicitly mask random joints during training to specifically cope with the occlusions using data augmentation.

A model by Rhodin et al. [60] learns to predict 3D human pose from a single view, by learning in a multi-view con-figuration. The key elements of their approach are the multi-view constraints that enforce the correct rough estimates of the pose orientations. The weak supervision, however, was not sufficient, so they also used 3D ground truth, when available.

4. Evaluation

The most common evaluation metric for 3D human pose estimation is a mean per-joint precision error (MPJPE). For a single frame f and a single pose p in the frame f, MPJPE is computed as an L2-norm:

$$E_{MPJPE}(f,p) = \frac{1}{N_p}\sum_{1}^{N_p} \left\| m_{f,p}^{(f)}(i) - m_{gt,p}^{(f)}(i) \right\|_2, \qquad (1)$$

Where N_p is the number of joints in the pose p. In multi-person pose estimation tasks, for a collection of frames f_i in Φ, the error is the average over the MPJPEs of all frames and poses Π:

$$E_{MPJPE}(\Phi,\Pi) = \frac{1}{N_\Phi}\sum_{1}^{N_\Phi} \frac{1}{N_{\Pi_i}}\sum_{j=1}^{N_{\Pi_i}} E_{MPJPE}\,(f_{ij}, p_{ij}), \qquad (2)$$

Where N_Φ is the number of frames and N_Π number of poses in the frame f_i. This evaluation procedure was named protocol #1 in the simple-yet-effective paper [42]. A protocol #2 first aligns the predicted and ground truth poses using the Procrustes alignment. We believe that today's state-of-the-art has reached a level where the only relevant evaluation protocol should be #1, without the need for the prior alignment. Therefore, we report the protocol #1 results in the Tab. 3 on Humans3.6M dataset, in the Tab. 4 on MPI-3DHP dataset and in the Tab. 5 on HumanEva dataset. We also point out the important features of the approaches, some of which overlap with the ones in Tab. 1.

Tab 3. Quantitative comparison of the state-of-the-art methods for 3D pose estimation on Human3.6M dataset [28] (protocol #1).

	MPJPE	SUPERV.	#CAM	TEMP.	DET./REG.	EXTRA DATA
[46]	63.6	superv.	mono	yes	det.	no
[3]	63.3	Superv.	mono	yes	Reg.	no
[55]	56.9	Unsuperv.	multi	no	Det.	no
[73]	50.9	Unsuperv.	mono	no	det/	no
[56]	46.8	Semi-sup.	mono	yes	Reg.	no
[40]	46.6	Super.	mono	yes	Reg.	no
[13]	44.1	Superv.	mono	yes	Reg.	no
[16]	42.9	Superv.	mono	yes	reg	no
[38]	42.6	Superv.	mono	N	Reg.	no
[17]	40.1	Superv.	mono	yes	Det.	no
[59]	26.2	Superv.	multi	no	Det.	yes
[25]	19.0	Superv.	multi	no	Det.	yes
[30]	17.7	Superv.	mult	no	Det.	Yes

Tab 4. Quantitative comparison of the state-of-the-art methods for 3D pose estimation on MPI-3DHP dataset [43] (protocol #1).

	MPJPE	SUPERV.	TEMP.	DET./REG.	EXTRA DATA
[46]	98.4	superv.	Yes	det.	no
[73]	92.5	unsuperv.	No	det.	yes
[51]	60.1	superv.	No	det.	no

Tab 5. Quantitative comparison of the state-of-the-art methods for 3D pose estimation on HumanEva dataset [64] (protocol #1).

	MPJPE	SUPERV.	TEMP.	DET./REG.
[29]	38.9	superv.	No	reg.
[49]	26.9	superv.	No	reg.
[42]	24.6	superv.	No	reg.
[54]	24.3	Superv.	No	Det.
[78]	22.9	Superv.	no	Reg.
[26]	22.0	Superv.	Yes	Reg.
[16]	14.3	Semi-sup.	No	Det.
[17]	13.5	Semi-sup.	yes	Det.

5. Discussion

From the Tab. 1, we can observe that most of the approaches for 3D pose estimation are top-down and therefore limited to a single-person pose estimation. The reason why most of the methods are top-down is because the bottom-up approaches require a whole scene processing to achieve invariance on the number of people, which is resource-demanding. Resource consumption problem is directly related to detection-based approaches that are bound to represent the whole scene, either using multiple heatmaps that represent depth [67] or using voxels [54]. There are a few successful attempts to cope with the high resource consumption problem of detection-based methods, using compressed volumetric heatmaps [21] and using location maps instead of heatmaps [46, 44]. Regression-based methods, on the other hand, were never used in a bottom-up fashion. We may explain this by the fact that single-person pose estimation is still too difficult for the regression-based methods - the top performing methods on all the three presented datasets (Tab. 3, 4 and 5), both multi-view and monocular are detection-based. However, most of the top performing monocular methods on Human3.6M (Tab. 3) are regression-based, so learning bottom-up regression might be worthwhile.

The approaches that apply an iterative refinement approach [44, 59, 32, 11, 55] (Tab. 1) report an increase in performance. Most notably, a cross-view fusion approach [59] is the third best multi-view approach on Human3.6M dataset (Tab.3). Another important observation is that only a single method from the Tab. 1 [48] reconstructs 3D poses in absolute coordinates. Most of the methods reconstruct the poses up-to-scale, i.e., either the ground truth poses are normalized in the preprocessing step or the predicted poses are first scaled to match the ground truth size and then evaluated. This means that, even if the method reconstructs multiple people in the scene, it still does not know the depths (distances from the camera) of these poses.

Observing Tab. 3, it is obvious that multi-view approaches are significantly better performing, which is expected. Supervised approaches are more successful on Human3.6M (Tab. 3 and MPI-3DHP (Tab. 4) but, interestingly, semi-supervised approaches [17, 16] are the most successful on the HumanEva dataset (Tab. 5). Note that the best performing method in HumanEva is also exploiting the time component (the model is learning from a video). Based on Tab. 3, 4 and 5, we cannot say whether exploiting the time component brings a performance gain in terms of MPJPE. Also, it is obvious that the three datasets are not equally difficult. The best methods achieve 17.7mm, 60.1mm and 13.5mm for the Human3.6M, MPI-3DHP and HumanEva datasets, respectively. We can therefore conclude that MPI-3DHP dataset is currently the most challenging 3D human pose estimation dataset.

6. Conclusion

In this work, we have reviewed 3D human pose estimation approaches and datasets, extracted and commented on their common features, presented state-of-the-art methods and compared them based on a single evaluation protocol (MPJPE, protocol #1). Deep learning-based 3D pose estimation achieves remarkably low MPJPE, but still lacks the ability the recover the absolute scale and is unable to reconstruct multiple 3Dposes in a fully-differentiable manner. The future works might be motivated by the success of the detection-based 2D pose detection methods [9, 50], focusing on the low-resource implementation that avoids directly embedding the whole scene into a deep learning model via heatmaps. Regarding low-resource and non-GPU approaches, a lightweight OpenPose [52] is an example of a successful GPU-to-CPU transition, keeping real-time performance, while sacrificing a bit of accuracy. In the future, we might expect more real-time and low-resource proposals that are able to reconstruct the absolute pose dimensions in an end-to-end learning fashion.

7. Acknowledgements

This work has been supported by Croatian Science Foundation under the grant number HRZZ-IP-2018-01-8118 (STEAM).

References

[1] Ijaz Akhter and Michael J. Black. “Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction”. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015. June 2015.

[2] Sikandar Amin et al. “Multi-view Pictorial Structures for 3D Human Pose Estimation”. In: Jan. 2013, pp. 45.1–45.11. ISBN: 1-901725-49-9.DOI: 10.5244/C.27.45.

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. “Exploiting Temporal Context for 3D Human Pose Estimation in the Wild”. In:2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), pp. 3390–3399.

[4] Vasileios Belagiannis et al. “3D Pictorial Structures Revisited: Multiple Human Pose Estimation”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence38 (Oct. 2016), pp. 1929–1942.DOI: 10.1109 /TPAMI.2015.2509986.

[5] Hayet Belghit et al. “Vision-based Pose Estimation for Augmented Reality: A Comparison Study”. In: ArXiv: abs/1806.09316 (2018).

[6] Christopher M. Bishop. Mixture density networks. Tech. rep. 1994.

[7] Federica Bogo et al. “Dynamic FAUST: Registering Human Bodies in Motion”. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). July 2017.

[8] Lewis Bridgeman et al. “Multi-Person 3D Pose Estimation and Tracking in Sports”. In:2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019), pp. 2487–2496.

[9] Z. Cao et al. “OpenPose: Realtime Multi-Person 2DPose Estimation using Part Affinity Fields”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[10] Geonho Cha et al. “Deep Pose Consensus Networks”. In: *ArXiv* abs/1803.08190 (2019).

[11] Ching-Hang Chen and Deva Ramanan. “3D Human Pose Estimation = 2D Pose Estimation + Matching”. In:2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2017), pp. 5759–5767.

[12] Long Chen et al. “Cross-View Tracking for Multi-Human 3D Pose Estimation at over 100 FPS”. In: *ArXiv* abs/2003.03972 (2020).

[13] Tianlang Chen et al. “Anatomy-aware 3D HumanPose Estimation in Videos”. In: *ArXiv* abs/2002.10322(2020).

[14] Xipeng Chen et al. “Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation”. In:2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), pp. 10887–10896.

[15] Yucheng Chen, Yingli Tian, and Mingyi He. “Monocular human pose estimation: A survey of deep learning-based methods”. In: Computer Vision and Image Understanding192 (2020), p. 102897.ISSN: 1077-3142.DOI: https://doi.org/10.1016/j.cviu.2019.102897. URL: http://www.sciencedirect.com/science/article/pii/S1077314219301778.

[16] Y. Cheng et al. “Occlusion-Aware Networks for 3DHuman Pose Estimation in Video”. In:2019 IEEE/CVF International Conference on Computer Vision (ICCV).2019, pp. 723–732.

[17] Yu Cheng et al. “3D Human Pose Estimation using Spatio-Temporal Networks with Explicit Occlusion Training”. In: *ArXiv* abs/2004.11822 (2020).

[18] Y. Dai, H. Li, and M. He. “A simple prior-free method for non-rigid structure-from-motion factorization”. In:2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012, pp. 2018–2025.

[19] Junting Dong et al. “Fast and Robust Multi-Person3D Pose Estimation From Multiple Views”. In:2019IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)(2019), pp. 7784–7793.

[20] Dylan Drover et al. “Can 3D Pose be Learned from 2DProjections Alone?” In: *ArXiv* abs/1808.07182 (2018).

[21] Matteo Fabbri et al. “Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation”. In: *ArXiv* abs/2004.00329 (2020).

[22] Wenjuan Gong et al. “Human Pose Estimation from Monocular Images: A Comprehensive Survey”. In: Sensors (Basel, Switzerland)16 (2016).

[23] Rıza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. “Densepose: Dense human pose estimation in the wild”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 7297–7306.

[24] Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. 2nd ed. New York, NY, USA: Cambridge University Press, 2003.ISBN:0521540518.

[25] Yihui He et al. “Epipolar Transformers”. In: *ArXiv* abs/2005.04551 (2020).

[26] Mir Rayat Imtiaz Hossain and James J. Little. “Exploiting Temporal Information for 3D Human Pose Estimation”. In:ECCV. 2018.

[27] Eldar Insafutdinov et al. “Articulated Multi-person Tracking in the Wild”. In: *ArXiv* abs/1612.01465(2016).

[28] Catalin Ionescu et al. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensingin Natural Environments”. In:IEEE Transactions on Pattern Analysis and Machine Intelligence 36.7 (July 2014), pp. 1325–1339.

[29] Umar Iqbal et al. “A Dual-Source Approach for 3D Human Pose Estimation from a Single Image”. In: *ArXiv* abs/1705.02883 (2018).

[30] Karim Iskakov et al. “Learnable Triangulation of Human Pose”. In:2019 IEEE/CVF International Conference on Computer Vision (ICCV)(2019), pp. 7717–7726.

[31] Hanbyul Joo et al. “Panoptic Studio: A Massively Multiview System for Social Interaction Capture”. In: IEEE Transactions on Pattern Analysis and MachineIntelligence41 (2019), pp. 190–204.

[32] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. “Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)(2019), pp. 1077–1086.

[33] Onorina Kovalenko et al. “Structure from Articulated Motion: Accurate and Stable Monocular 3D Reconstruction without Training Data”. In: Sensors (Basel, Switzerland)19 (2019).

[34] Christoph Lassner et al. “Unite the People: Closing the Loop Between 3D and 2D Human Representations”. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). July 2017. URL: http://up.is.tuebingen.mpg.de.

[35] Hsi-Jian Lee and Zen Chen. “Determination of 3D human body postures from a single view”. In: Comput.Vis. Graph. Image Process. 30 (1985), pp. 148–168.

[36] Bo Li et al. “Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN”. In:2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (2017), pp. 601–604.

[37] Bo Li et al. “Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network”. In:2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (2017), pp. 613–616.

[38] Chen Li and Gim Hee Lee. “Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network”. In:2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), pp. 9879–9887.

[39] Sijin Li, Weichen Zhang, and Antoni B. Chan. “Maximum-Margin Structured Learning with Deep Net-works for 3D Human Pose Estimation”. In: International Journal of Computer Vision 122 (2015), pp. 149–168.

[40] Jiahao Lin and Gim Hee Lee. “Trajectory Space Factorization for Deep Video-Based 3D Human Pose Estimation”. In: *ArXiv* abs/1908.08289 (2019).

[41] Matthew Loper et al. “SMPL: A Skinned Multi-Person Linear Model”. In: ACM Trans. Graphics (Proc. SIG-GRAPH Asia)34.6 (Oct. 2015), 248:1–248:16.

[42] Julieta Martinez et al. “A Simple Yet Effective Base-line for 3d Human Pose Estimation”. In:2017 IEEE International Conference on Computer Vision (ICCV) (2017), pp. 2659–2668.

[43] Dushyant Mehta et al. “Monocular 3D Human PoseEstimation In The Wild Using Improved CNN Super-vision”. In:3D Vision (3DV), 2017 Fifth InternationalConference on. IEEE. 2017. DOI: 10.1109/3dv.2017.00064. URL: http://gvv.mpi-inf.mpg.de/3dhpdataset.

[44] Dushyant Mehta et al. “Single-Shot Multi-Person 3DPose Estimation from Monocular RGB”. In: 3D Vision(3DV), 2018 Sixth International Conference on. IEEE.Sept. 2018. URL: http://gvv.mpi-inf.mpg.de/projects/SingleShotMultiPerson.

[45] Dushyant Mehta et al. “VNect: real-time 3D human pose estimation with a single RGB camera”. In: *ArXiv* abs/1705.01583 (2017).

[46] Dushyant Mehta et al. “XNect: Real-time Multi-person3D Human Pose Estimation with a Single RGB Cam-era”. In: *ArXiv* abs/1907.00837 (2019).

[47] Rudrajit Mitra et al. “Multiview-Consistent Semi-Supervised Learning for 3D Human Pose Estimation”. In: *arXiv* Computer Vision and Pattern Recognition(2019).

[48] Gyeongsik Moon, Ju Yong Chang, and Kyoung MuLee. “Camera Distance-Aware Top-Down Approach for3D Multi-Person Pose Estimation from a Single RGB Image”. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV)(2019), pp. 10132–10141.

[49] Francesc Moreno-Noguer. “3D Human Pose Estimationfrom a Single Image via Distance Matrix Regression”.In:2017 IEEE Conference on Computer Vision andPattern Recognition (CVPR)(2017), pp. 1561–1570.

[50] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked Hourglass Networks for Human Pose Estimation”. In: ECCV. 2016.

[51] Aiden Nibali et al. “3D Human Pose Estimation With2D Marginal Heatmaps”. In:2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (2019), pp. 1477–1485.

[52] Daniil Osokin. “Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose”. In: ICPRAM.2018.

[53] Sungheon Park, Jihye Hwang, and Nojun Kwak. “3DHuman Pose Estimation Using Convolutional Neural Networks with 2D Pose Information”. In: *ArXiv* abs/1608.03075 (2016).

[54] Georgios Pavlakos et al. “Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose”. In:2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), pp. 1263–1272.

[55] Georgios Pavlakos et al. “Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations”. In:2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2017), pp. 1253–1262.

[56] Dario Pavllo et al. “3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training”. In:2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)(2019), pp. 7745–7754.

[57] Xavier Perez-Sala et al. “A Survey on Model Based Approaches for 2D and 3D Visual Human Pose Recovery”. In: Sensors (Basel, Switzerland)14 (2014), pp. 4189–4210.

[58] Albert Pumarola et al. “3DPeople: Modeling the Geometry of Dressed Humans”. In: International Conference in Computer Vision (ICCV). 2019.

[59] Haibo Qiu et al. “Cross View Fusion for 3D Hu-man Pose Estimation”. In:2019 IEEE/CVF Interna-tional Conference on Computer Vision (ICCV)(2019), pp. 4341–4350.

[60] Helge Rhodin et al. “Learning Monocular 3D Human Pose Estimation from Multi-view Images”. In:2018IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), pp. 8437–8446.

[61] G. Rogez, P. Weinzaepfel, and C. Schmid. “LCR-Net: Localization-Classification-Regression for HumanPose”. In:2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 1216–1224.

[62] Nikolaos Sarafianos et al. “3D Human pose estimation: A review of the literature and analysis of covariates”.In: Computer Vision and Image Understanding152(2016), pp. 1–20.ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2016.09.002. URL: http://www.sciencedirect.com/science/article/pii/S1077314216301369.

[63] A. M. Sharma, K. Venkatesh, and A. Mukerjee. “Human pose estimation in surveillance videos using temporal continuity on static pose”. In:2011 International Conference on Image Information Processing. 2011, pp. 1–6.

[64] L. Sigal, A. Balan, and M. J. Black. “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion”. In: International Journal of Computer Vision87.1 (Mar. 2010), pp. 4–27.

[65] Georgios Stamou et al. “4.11 - 2D and 3D Motion Tracking in Digital Video”. In: Handbook of Image and Video Processing (Second Edition). Ed. by ALBOVIK. Second Edition. Communications, Networking and Multimedia. Burlington: Academic Press, 2005, pp. 491–XVIII.ISBN: 978-0-12-119792-6.DOI: https://doi.org/10.1016/B978-012119792-6/50093-0. URL: http://www.sciencedirect.com/science/article/pii/B9780121197926500930.

[66] Xiao Sun et al. “Compositional Human Pose Regression”. In:2017 IEEE International Conference on Computer Vision (ICCV) (2017), pp. 2621–2630.

[67] Xiao Sun et al. “Integral Human Pose Regression”. In: ECCV. 2018.

[68] Camillo Taylor. “Reconstruction of articulated objects from point correspondences in a single uncalibrated image”. In: vol. 1. Feb. 2000, 677–684 vol.1. ISBN: 0-7695-0662-3. DOI: 10.1109/CVPR.2000.855885.

[69] Denis Tome, Chris Russell, and Lourdes Agapito. “Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image”. In:2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), pp. 5689–5698.

[70] Matt Trumble et al. “Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors”. In:2017British Machine Vision Conference (BMVC). 2017.

[71] Gul Varol et al. “Learning from Synthetic Humans”. In: CVPR. 2017.

[72] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. “A Kinematic Chain Space for Monocular Motion Capture”. In: *ArXiv* abs/1702.00186 (2018).

[73] Bastian Wandt and Bodo Rosenhahn. “RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation”. In: 2019IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)(2019), pp. 7774–7783.

[74] Jon A. Webb. “Static Analysis of Moving Jointed Objects”. In: Proceedings of the First AAAI Conference on Artificial Intelligence. AAAI’80. Stanford, California: AAAI Press, 1980, pp. 35–37.

[75] Shih-En Wei et al. “Convolutional Pose Machines”. In:2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp. 4724–4732.

[76] X. K. Wei and Jinxiang Chai. “Modeling 3D human poses from uncalibrated monocular images”. In:2009 IEEE 12th International Conference on Computer Vision. 2009, pp. 1873–1880.

[77] Bin Xiao, Haiping Wu, and Yichen Wei. “Simple Baselines for Human Pose Estimation and Tracking”. In: ECCV. 2018.

[78] Yuanlu Xu et al. “Learning Pose Grammar for Monocular 3D Pose Estimation”. In: 2019.

[79] Xiaowei Zhou et al. “Sparseness Meets Deepness: 3DHuman Pose Estimation from Monocular Video”. In:2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2016), pp. 4966–4975.

[80] Xingyi Zhou et al. “Deep Kinematic Pose Regression”. In: *ArXiv* abs/1609.05317 (2016).