•Review•

A survey on monocular 3D human pose estimation

Xiaopeng JI, Qi FANG, Junting DONG, Qing SHUAI, Wen JIANG, Xiaowei ZHOU*

State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058*, China*

*** Corresponding author,** xzhou@cad.zju.edu.cn

Received: 18 March 2020 **Accepted:** 25 April 2020

Supported by National Natural Science Foundation of China (61806176); the Fundamental Research Funds for the Central Universities (2019QNA5022).

Citation: Xiaopeng JI, Qi FANG, Junting DONG, Qing SHUAI, Wen JIANG, Xiaowei ZHOU. A survey on monocular 3D human pose estimation. Virtual Reality & Intelligent Hardware, 2020, 2(6): 471—500

DOI: 10.1016/j.vrih.2020.04.005

Abstract Recovering human pose from RGB images and videos has drawn increasing attention in recent years owing to minimum sensor requirements and applicability in diverse fields such as human-computer interaction, robotics, video analytics, and augmented reality. Although a large amount of work has been devoted to this field, 3D human pose estimation based on monocular images or videos remains a very challenging task due to a variety of difficulties such as depth ambiguities, occlusion, background clutters, and lack of training data. In this survey, we summarize recent advances in monocular 3D human pose estimation. We provide a general taxonomy to cover existing approaches and analyze their capabilities and limitations. We also present a summary of extensively used datasets and metrics, and provide a quantitative comparison of some representative methods. Finally, we conclude with a discussion on realistic challenges and open problems for future research directions.

Keywords Human pose estimation; Human motion capture

1 Introduction

Monocular human pose estimation is the process of estimating the configuration of human body parts from a single image or a sequence of images. Among many human-centered tasks (e.g., human detection, human tracking, and human action recognition) in computer vision, human pose estimation is particularly important as it provides rich geometric and motion information of the human body for a variety of applications, such as visual surveillance, behavior analysis, autonomous driving, service robots, healthcare, gaming, and animation. Human pose estimation is a challenging problem owing to the existence of large degrees of freedom, high variations in appearances, changes in viewpoints, and complex backgrounds. Some of these difficulties have been addressed in constrained settings by motion capture (MoCap) systems with reflective markers[1–4], depth sensors[5–10], or inertial measurement units (IMUs)[11–15]. However, such systems require particular hardware that is often expensive and exclusive, while the data acquisition process often restricts the range of human movement. Therefore, as a more accessible approach, estimating human pose from RGB images captured by regular cameras is garnering the attention of researchers.

Human pose is often represented in vision systems by the angles of various joints on a predefined skeleton. As shown in Figure 1a, a simplified representation of human pose is the locations of a set of

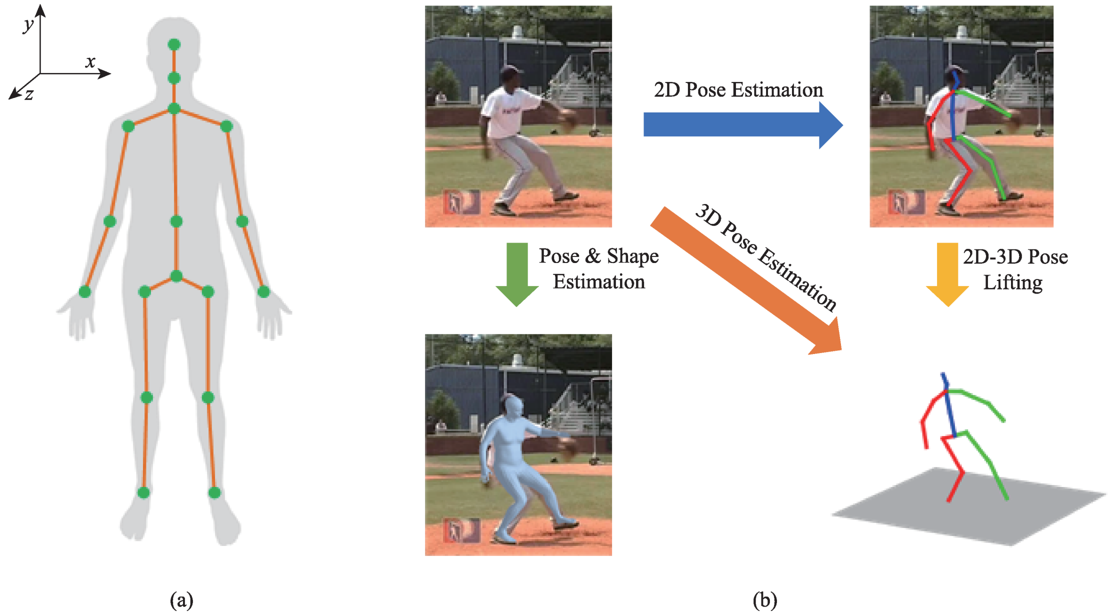

Figure 1 Illustration of human pose representation and pose estimation. (a) A common kinematic representation of the human body by 17 keypoints; (b) 3D human pose estimation, 2D-3D pose lifting and human pose and shape estimation.

keypoints on the human body, which are directly observed better in images than in the joint angles. With such a simplified representation, the objective of 2D human pose estimation is to locate the (x,y) coordinates for each keypoint in an image, while 3D human pose estimation aims to infer the (x,y,z) coordinates for each keypoint in the 3D space. Keypoints can be sparse body joints or densely sampled points on the body surface[16]. An illustration of estimating the 3D pose by lifting the 2D pose to 3D, and inferring the final pose and shape from the images is shown in Figure 1b. The sample image is from the Penn Action dataset[17].

Significant progress has been made in this area in recent years by leveraging the power of deep learning[18–20], which has shown remarkable performance especially in detecting 2D keypoints in RGB images[21,22]. While obtaining manual 2D pose annotations is easy, collecting accurate 3D pose annotations is difficult. The ambiguities in monocular 3D reconstruction, lack of annotated data to train 3D pose estimators, and the absence of benchmark datasets in real-world situations make the problem of 3D pose estimation even more challenging. Nevertheless, remarkable progress has been facilitated by the availability of accessible large-scale 3D pose datasets like HumanEva[1], Human3.6M[2], and CMU Panoptic[23]. These datasets are collected by MoCap systems in controlled lab environments, which show that it is possible to recover 3D poses by lifting from 2D poses, or by direct regression from images. A recent work also addressed the generalization problem of 3D pose estimation in the wild, such as training with mixed 2D and 3D data[24], synthesizing training images[25], and collecting accurate 3D pose annotations in the wild[15]. Moreover, with statistical body shape models like the skinned multi-person linear model (SMPL)[26], the full-body 3D shape[27–31] can be inferred from a single image with detailed expressions and gestures[32,33]. Consequently, interest in monocular 3D human pose and shape recovery has been increasing in computer vision and graphics. The emergence of increasing interest and remarkable advances in this field motivate us to survey the state-of-the-art approaches.

There have been some surveys and book chapters on human pose estimation. Readers can refer to Moeslund and Krüger[34] for an overview of early work in this field. Poppe et al. discussed the characteristics of human motion analysis[35]. Pons-Moll and Rosenhahn[36] provided an analysis of the kinematic parameterization of human motion, and representations of human shape among model-based pose estimation algorithms. Sminchisescu et al. reviewed the methods for feature-based pose estimation[37]. Perez-Sala et al. discussed the model-based approaches for pose estimation[38]. Liu et al. discussed body part parsing[39]. Gong et al. [40] and Holte et al. [41] surveyed literature on human pose estimation from monocular images and multi-view videos. The most recent survey on 3D human pose estimation is Sarafianos et al.[42], published in 2016. Since then, there has been significant progress in every aspect of this topic such as methods, datasets, and empirical results, which are summarized in this survey. We focus on 3D pose estimation from a monocular image or video, and provide a general taxonomy to cover existing approaches. An illustration of the taxonomy underpinning this survey is shown in Figure 2.

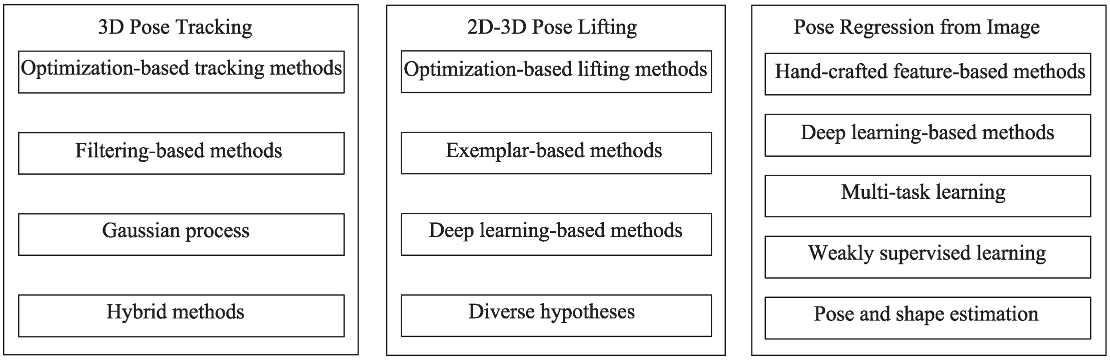

Figure 2 Taxonomy of 3D human pose estimation methods.

Existing methods of 3D human pose estimation can be classified into three categories: (1) 3D pose tracking, which covers most of the early works that are based on incremental frame-to-frame tracking. (2) 2D-3D pose lifting, which has two stages: detecting the 2D poses and lifting the 2D poses into 3D. (3) Pose regression from images, which directly infer 3D poses from raw image pixels. In the following sections, we present a more detailed discussion on each category of methods.

2 3D pose tracking

Early work on 3D human pose mainly concentrates on incremental frame-to-frame tracking[43−45] starting from the pose in the first frame. Due to the requirement of the initialization step and the inability to recover from tracking failures[46], later approaches have formulated the tracking problem as one of data association across frames, i.e., "tracking by detection"[47]. The studies in this category focus on tracking human joints on multiple frames. The publications discussed in this category are shown in Table 1.

2.1 Optimization-based tracking methods

For optimization methods, Bregler and Malik parameterized the human body locally, using twist and the product of exponential maps[43]. They obtained a least squares solution through the brightness constancy equation in the case of a scaled orthographic projection model. Their method can be extended easily to multi-view settings[58]. Gall et al. proposed an analysis-by-synthesis framework that combines patch-based and region-based matching in order to get 2D correspondences, where outliers caused by occlusion are removed by comparing the original image and the synthesized image[54]. These are then used in the twisted framework to solve linear equations based on the projection from 3D to 2D. These early approaches make use of the sufficient but not so robust correspondences such as pixel-level features or local descriptors[54] to compute the closed-form solution.

Table 1 Summary of tracking based methods

Authors	Method Highlights	Evaluation Datasets	Year
Leonardos et al.[48]	A second-order stochastic dynamical model for human motion. Deriving the equations of a Riemannian extended Kalman filter to perform the structure estimation.	CMU Mocap[1]	2016
Simo-Serra et al.[49]	Learning a joint Gaussian mixture model of both human pose and kinematics on a Riemannian manifold.	Human3.6M	2014
Hauberg et al.[50]	Generalizing the unscented transform and UKF to Riemannian manifold. Using this Riemannian UKF for articulated tracking.	Custom	2013
Sigal et al.[45]	Representing the body as an undirected graphical model. Using particle message passing to estimate human pose and motion.	HumanEva-I	2012
Pons-Moll et al.[12]	Using orientation cues derived from sparse inertial sensors to sample particles by inverse kinematics.	Custom	2011
Yao et al.[51]	Learning a Gaussian process latent variable model from data. Using stochastic gradient descent algorithm to track human subjects in the latent space.	HumanEva-I	2011
Andriluka et al.[47]	2D tracking and viewpoint estimation. Data association across frames. 3D pose recovery using tracklet-based observations.	HumanEva-II, TUD Stadtmitte[47]	2010
Gall et al.[52]	A multi-layer framework that combines stochastic optimization, filtering, and local optimization.	HumanEva-II	2010
Taylor et al.[53]	Using conditional restricted Boltzmann machine as a latent variable model for human pose tracking.	HumanEva-I	2010
Gall et al.[54]	An analysis-by-synthesis framework that combines patch-based and region-based matching to solve the drift problem.	HumanEva-II, Custom	2008
Wang et al.[55]	Gaussian process dynamical models for nonlinear dynamics analysis.	Custom	2006
Deutscher and Reid[56]	Annealed particle filtering for articulated body motion capture.	Custom	2005
Sminchisescu and Triggs[44]	Utilizing kinematic tree information to accelerate search and enforce online ambiguity rejection.	Custom	2003
Sminchisescu and Triggs[57]	A hybrid sample-and-refine search scheme guided by rescaled covariance with a particle filter.	Custom	2001

Some researchers have formulated the pose recovery problem as a search problem, and presented some efficient search methods to solve the optimization. In order to find pose parameters consistent with all views, Gavrila and Davis[59] used a search space decomposition approach and a chamfer matching-based best-first technique to search through the high dimensional pose parameter space. Sminchisescu and Triggs[57] designed a robust matching metric combining optical flow, edge energy and motion boundaries, and presented a hybrid sample-and-refine search scheme for more robust results with a particle filter. First, a distribution is generated by combining the previous posterior and the dynamical model through temporal propagation. Then the covariance is inflated to make sampling adequately. Finally, a local optimization based on the likelihood is used to refine the result. To accurately map the human structure, they further utilized kinematic tree information to change the sample strategy, covering both the forward and the backward flips to accelerate the search, and enforce online ambiguity rejection[44].

2.2 Filtering-based methods

Filtering approaches mainly stem from the Kalman and particle filters in the Bayesian framework. Kalman filter series were widely used owing to their explicit forms, simplicity and optimality[60,61]. For example, Wachter and Nagel[62] used the iterated extended Kalman filter (IEKF) to solve posterior estimation with a simple motion model of constant velocity. Kraft proposed a quaternion-based unscented Kalman filtering

[1] http://mocap.cs.cmu.edu/

(UKF) for orientation tracking[63]. Sidenbladh et al. analyzed the ambiguities and the non-linear dynamics, expecting the posterior probability distribution over model parameters to be multi-modal[64]. Thus, they defined a generative model by including the shape, appearance, and motion, and computed the posterior probability by a particle filter. Furthermore, the particle filter was combined with Markov chains using a hybrid Monte Carlo method to obtain samples in high dimensional spaces[65]. However, the prior knowledge about the motion model could not be obtained in most settings, which is essential in the Bayesian filter. Therefore, it is natural to learn the motion from the datasets. Howe et al. modeled plausible motions as a mixture of Gaussian probabilities in a high-dimensional space, and learned it from the training data[66].

Embedded manifolds are an alternative representation of Euclidean state spaces used for better description[67], as human motion can be naturally confined to some known Riemannian manifolds. Therefore, filtering on manifolds attracted much attention in this field. Deutscher and Reid developed a modified annealed particle filter to search for high-dimensional configuration of articulated human motion[56]. By using orientation cues derived from sparse inertial sensors, Pons-Moll et al. introduced an annealing particle-based optimization to sample particles from the manifold of valid poses[12]. Sigal et al. formulated body motion estimation as the inference in an undirected graphical model, and solved it by using particle message passing (PAMPAS)[45]. Hauberg et al. generalized the unscented transform and UKF to Riemannian manifolds[50]. They built an articulated tracking system using this Riemannian UKF, and achieved lower tracking errors as well as lesser computation times compared to particle filters. Leonardos et al. proposed a second-order stochastic dynamical model whose state space is an S^2 manifold, and derived the equations of a Riemannian extended Kalman filter to perform structure estimation from an image sequence captured by a perspective camera[48]. Simo-Serra et al. [49,68] modeled pose as a manifold, and learned a joint Gaussian mixture model of both the human pose and the kinematics on this manifold.

2.3 Gaussian process

Human pose parameter space is high dimensional and difficult to obtain, so over-fitting and limited generalization are significant issues in learning useful models. Dimensional reduction is one way to ease this problem. Researchers have presented many approaches to learn low-dimensional latent spaces from datasets. Gaussian process series are the most prominent latent variable models used to embed motion[69]. Yao et al. presented a stochastic gradient descent algorithm for the Gaussian process latent variable model (GPLVM) to learn probabilistic non-linear latent spaces composed of multiple activities. Because GPLVM does not model the dynamics in the latent space[51]. Wang et al. presented Gaussian process dynamical models (GPDM) using Gaussian process priors for both the dynamics and the observations mappings[55]. These can be learned from relatively small amounts of data. Urtasun et al. modified the GPDM by raising the dynamics density function to a scalar (the ratio between the dimensions of the observation space and the latent space), to balance the influence of the two stages in learning[70]. Since the learning and inference of GPLVM and GPDM are not linear, Taylor et al. presented implicit mixtures of conditional restricted Boltzmann machines (ImCRBM), which is linear in the number of training exemplars[53]. It is a kind of implicit dimensional reduction, i. e. the latent representation remains high dimensional, but the model learns to construct energy ravines in the latent space, eliminating the possibility of real data occasionally departing from the manifold.

2.4 Hybrid methods

Filtering methods have the advantage of maintaining temporal consistency, but they rarely improve the accuracy of the estimation. In contrast, local optimization methods could provide very accurate results given that the state vector is initialized near the global optimum, but they may suffer from jitters. Stochastic search methods have the ability to find the global solution, but the higher the state spatial dimension is, the more computation resources that are required. Latent space approaches have an easy structure, and hence it is natural to combine these types of approaches. Gall et al. introduced a multi-layer framework that combines stochastic optimization, filtering, and local optimization[52]. Andriluka et al.[47] proposed a three-stage hybrid method. The first two stages obtain 2D pose tracklets based on tracking-by-detection, and the third stage recovers 3D pose through hierarchical GPLVM. They can track 3D pose in realistic street conditions. Elhayek et al.[71] united CNN-based 2D joint detection with a generative motion-tracking algorithm based on the Sums of Gaussians (SoG), through a combined pose optimization energy.

3 2D-3D pose lifting

The process of 2D-3D lifting is a part of two-stage based cascaded frameworks that first performs 2D pose estimation to predict 2D joint positions or keypoints in the image with a 2D pose estimator[19,72], and then lifts these 2D joints to the 3D space[46,73]. The key idea is that 2D pose estimation can be performed easily owing to the availability of large-scale datasets with 2D annotations in the wild. 2D-3D lifting methods are usually generalizable across domains, benefiting from the reliable performance of state-of-the-art 2D pose detectors, and generally outperform the methods that directly regress 3D poses from images. The methods in this category are presented in Table 2. Note that MPII[21] and Leeds Sports Pose (LSP)[74] datasets are only annotated with 2D poses that are usually employed to show qualitative examples on real-world images by most works.

3.1 Optimization-based lifting methods

Early work in the field of optimization-based lifting methods employed annotated 2D landmarks of human joints to recover 3D poses by optimizing certain cost functions. For example, Ramakrishna et al. proposed a sparse representation based approach to estimate human 3D configuration from annotated landmarks in a single image[110]. They presented a projected matching pursuit algorithm for reconstructing 3D poses and camera settings by minimizing the re-projection error. Simo-Serra et al. employed a Bayesian framework to integrate a generative model with discriminative 2D part detectors, and performed inference using an evolutionary algorithm[108]. Wang et al. proposed to estimate the 3D pose by minimizing an L1-norm penalty between the projection of the 3D joints and the 2D detections to reduce the impact of inaccurate 2D pose estimations[105,115]. Instead of employing the joint positions or heatmaps as intermediates, Ionescu et al. predicted the 3D pose based on the descriptors computed from body part labels with an iterative scheme[107]. They proposed second-order pooling over a hierarchical region decomposition of the body to construct a global representation.

To address the issue of alternating minimization schemes being usually sensitive to initialization, Zhou et al. employed an augmented shape-space model to give a linear representation to intrinsic shape deformation and extrinsic viewpoint changes under a convex formulation[104]. Akhter and Black exploited the constraints on joint angles to avoid impossible poses[103]. Zhou et al. presented an expectation-maximization (EM) algorithm over the entire sequence to recover the 3D pose by combining CNN-based heatmaps with a sparse representation of 3D human pose[99]. Du et al. reinforced the pose-conditioned joint velocity and the temporal coherence constraints, and formulated an objective function to estimate 3D

Table 2 Summary of the 2D-3D pose lifting based methods

Authors	Method Highlights	Evaluation Datasets	Year
Rhodin et al.[75]	Dilated temporal convolution on 2D keypoint trajectories. A semi-supervised approach to exploit unlabeled videos.	Human3.6M, HumanEva-I	2019
Pavlakos et al.[76]	Extending SMPL with fully articulated hands and an expressive face.	Custom	2019
Zhao et al.[77]	Operating on tasks with graph-structured data for 2D to 3D human pose regression.	Human3.6M	2019
Wandt and Rosenhahn[78]	Learning a mapping from a distribution of 2D poses to a distribution of 3D poses using adversarial training.	Human3.6M, MPI-INF-3DHP, LSP	2019
Arnab et al.[79]	Applying bundle adjustment with the SMPL model by encouraging temporal consistency.	Human3.6M, HumanEva, 3DPW	2019
Chen et al.[80]	Using the self-consistency scheme: random projections of generated 3D skeletons are fed to a 2D pose discriminator to provide feedback to the 2D-3D lifting network.	Human3.6M, MPI-INF-3DHP	2019
Véges et al.[81]	A Siamese architecture that learns a rotation equivariant hidden representation.	Human3.6M, MPII[21]	2019
Li et al.[82]	Generating multiple 3D pose hypotheses with mixture density models to alleviate the ambiguity problem.	Human3.6M, MPI-INF-3DHP, MPII	2019
Cha et al.[83]	A multiple-partial-hypothesis-based scheme for single-image-based 3D human pose estimation.	Human3.6M, HumanEva-I, MPII	2019
Yang et al.[84]	Synthesizing virtual candidate poses, ensuring that the augmented exemplar set has more variety.	Human3.6M, LSP	2019
Fang et al.[85]	Designing a deep grammar network to explicitly encode a set of knowledge over human body configuration including kinematics, symmetry, and motor coordination..	Human3.6M, HumanEva-I, MPII	2018
Hossain and Little[86]	Layer-normalized LSTM units with shortcut connections to exploit temporal information.	Human3.6M, HumanEva-I	2018
Zhao et al.[87]	Using a set of fully connected layers to directly recover depth information.	Human 3.6M, CMU MoCap	2018
Lee et al.[88]	Propagating long short-term memory networks to estimate depth information based on the 2D pose.	Human3.6M, HumanEva-I,	2018
Drover et al.[89]	Random projection layer with adversarial training to enforce a prior on the 3D structure from 2D projections.	Human3.6M, MPII, LSP	2018
Wang et al.[90]	Learning the depth rankings between pairs of human joints from images using CNNs.	Human3.6M, MPII	2018
Park et al.[91]	Relational networks to capture the relations among the group configurations of different body parts.	Human 3.6M, MPII	2018
Martinez et al.[92]	Applying two fully connected layers with a residual connection on 2D to 3D keypoint regression.	Human3.6M	2017
Moreno-Noguer[93]	Performing 2D to 3D distance matrix regression between two Euclidean distance matrices.	Human3.6M, HumanEva-I	2017
Lassner et al.[94]	Predicting 31 segments and 91 landmark locations of the body and extending the SMPLify-based optimization method.	Human3.6M, HumanEva-I	2017
Nie et al.[95]	Two-level hierarchy of LSTM networks: a skeleton-LSTM that learns the depth information, and a patch-LSTM that utilizes the local image evidence.	Human3.6M, UCLA HHOI[96]	2017
Lin et al.[97]	A sequential prediction framework that refines the predicted poses with multiple recurrent stages in an implicit and comprehensive way.	Human3.6M, HumanEva-I	2017
Jahangiri and Yuille[98]	Generating diverse and valid human pose hypotheses in 3D, all consistent with the 2D detection of joints.	Human3.6M	2017
Bogo et al.[28]	Fitting a statistical body shape model by minimizing the error between the projected 3D model joints and the detected 2D joints.	HumanEva-I, Human3.6M, LSP	2016
Zhou et al.[99]	Leveraging the 3D geometric prior to help 2D joint localization, and to rigorously handle the 2D estimation uncertainty in a statistical framework.	Human3.6M, PennAction	2016
Yasin et al.[100]	Iterative update of the binaries of the pictorial structure model using the retrieved nearest 3D poses and updating the 2D pose.	HumanEva-I, CMU MoCap, Human3.6M	2016
Du et al.[101]	Combining RGB images and calculated height-maps to detect the landmarks of 2D joints. Reinforcing the temporal coherence constraints on both the camera and the 3D pose.	HumanEva-I, Human3.6M, MCAD[102]	2016

(To be continued on the next page)

(Continued)

Authors	Method Highlights	Evaluation Datasets	Year
Akhter and Black[103]	Learning pose dependent joint-angle limits. A multi-stage method to estimate 3D pose from 2D joint locations using an over-complete dictionary of poses.	CMU MoCap, LSP	2015
Zhou et al.[104]	Aligning a 3D shape-space model to 2D landmarks by solving a convex program.	Custom	2015
Wang et al.[105]	Estimating 3D pose by minimizing the L1-norm error between the projection of the 3D pose and the corresponding 2D detection.	CMU MoCap, HumanEva-I, UvA 3D pose[106]	2014
Ionescu et al.[107]	Label-sensitive pooling over a hierarchical region decomposition of the body. Continuous-valued pose regression.	Human3.6M	2014
Simo-Serra et al.[108]	A Bayesian framework that integrates a generative model and performs inference using evolutionary algorithms.	HumanEva-I, TUD Stadmitte	2013
Radwan et al.[109]	Regression based multiple view synthesis. Enforcing kinematic and orientation constraints on multiple synthetic views.	HumanEva-I, LSP	2013
Ramakrishna et al.[110]	A matching pursuit algorithm to estimate the sparse representation of 3D pose in an overcomplete dictionary.	CMU MoCap, Custom	2012
Simo-Serra et al.[111]	Stochastic exploration of ambiguous hypotheses and disambiguation.	HumanEva-I, TUD Stadmitte	2012
Jiang[112]	Searching upper body poses and lower body poses sequentially using kd-tree in the database	Custom	2010
Shakhnarovich et al.[113]	Learning a set of hashing functions that efficiently index examples.	Custom	2003
Mori and Malik[114]	Shape context matching from exemplar 2D views in conjunction with a kinematic chain-based deformation model.	Custom	2002

motion from the detected 2D joints in the monocular image sequence[101].

In contrast to previous methods that estimate 3D skeletons, Bogo et al. proposed SMPLify[28], to recover both the pose and the shape by fitting the statistical body shape model ,i.e., SMPL[26] to the 2D joints by minimizing an objective function that penalizes the error between the projected 3D and 2D model joints detected from a CNN-based estimator[116]. Lassner et al. predicted 31 segments and 91 landmark locations of the body, and applied an extended SMPLify-based optimization method[94]. Alldieck et al. unposed the cone defined by the projection rays to obtain a visual hull in a common reference frame, and generated the personalized blend shape model[117]. Unlike most approaches operating on single frames, Arnab et al. presented a bundle-adjustment-based algorithm to recover 3D human pose and meshes over the entire sequence to resolve the ambiguities[79]. Recent works also show the real-time performance of human pose and shape capture by applying optimization schemes on the 2D keypoints from CNN-based estimators. Habermann et al. presented a real-time human performance capture approach from monocular videos by designing a two-stage analysis-by-synthesis optimization[118].

3.2 Exemplar-based methods

As a non-parametric model, exemplar-based methods are effective in matching complex 3D poses from a given database. Mori and Malik proposed to estimate 3D human body configurations by matching their 2D projections with 2D exemplars from known viewpoints[114]. They used the correspondence between boundary points on the test image and the exemplar to estimate the 2D keypoints. 3D configuration of a body is further estimated with these keypoints. Shakhnarovich et al. proposed a parameter sensitive hashing method for efficiently matching the articulated pose of human upper body using the extracted image features.[113] Jiang constructed a database containing millions of different 3D poses in daily life[112]. To implicitly square the size of the exemplar database and reconstruct unconstrained poses efficiently, they split human pose into upper body and lower body poses. The 3D pose is reconstructed based on the kd-tree to achieve real-time search in the database.

The acquisition of large-scale training data with accurately annotated 3D poses is often an expensive and challenging task in 3D pose estimation. To overcome this difficulty, Yasin et al. proposed a dual-source approach, integrating annotated 2D poses and 3D motion capture data[100]. The 3D motion capture data is projected into a normalized 2D pose space. The image data with an annotated 2D pose is used to learn a pictorial structure model through random forest. Given a test image, the 2D pose is first estimated, and the final 3D pose is obtained by minimizing the projection error with the nearest retrieved 3D poses. Chen and Ramanan proposed a simple non-parametric method to encode high-level constraints and lift the predicted 2D poses to 3D using 2D-3D matching[119]. Specifically, they first generated a large number of 2D projections from a given 3D pose library. Using the prediction from a 2D pose estimation algorithm[72], they retrieved the most accurate 3D pose from the 3D pose library with a non-parametric nearest-neighbor model. Yang et al. generated an exemplar set including real and synthetic poses to ensure that the augmented exemplar set has a larger variety[84]. They introduced a two-step strategy to select the best exemplar from the candidate set to match the detected 2D pose.

3.3 Deep learning-based methods

With large amounts of 3D MoCap data captured in controlled environments, a deep neural network (DNN) can be trained to directly regress 3D human pose from a given 2D pose.

The multilayer perceptron is a straightforward way to implement mapping from a 2D to a 3D pose. Zhao et al. used a set of fully connected layers with hierarchical nonlinear transformations to recover depth information of a number of known 2D landmark points in a single image[87]. Martinez et al. also proposed a simple end-to-end network of two fully connected layers with a residual connection that directly regresses 3D keypoints from 2D keypoint detections[92]. Moreno-Noguer proposed to represent 2D and 3D poses with N×N matrices of Euclidean distances between every pair of joints, and formulated the 3D pose estimation problem as a 2D-to-3D distance matrix regression[93].

To better capture the kinematic relation of the human skeleton, Nie et al. designed a two-level tree-structure based LSTM network to integrate the features from global 2D skeletons and local image patches[95]. Véges et al. introduced a Siamese network with an equivariant embedding that provides regularization for cross-camera 3D human pose estimation[81]. Fang et al. developed a deep grammar network that extends bi-directional RNNs to encode a set of high-level knowledge of 3D human pose grammar, such as kinematics, symmetry, and motor coordination[85].

Temporally incoherent estimates usually cause jitters due to independent errors on individual frames. To address this problem, various attempts have been made by utilizing the temporal information across a sequence of 2D joint locations. Lin et al. presented a recurrent 3D pose sequence machine to automatically learn the image-dependent structural constraint and sequence-dependent temporal context[97]. They adopted multi-stage sequential refinement by assembling a 2D pose module, a 2D-to-3D adaption module, and a 3D pose recurrent module into a sequential prediction framework. Hossain and Little designed a layer-normalized LSTM unit based sequence-to-sequence network that imposed the temporal smoothness constraint to recover temporally consistent 3D poses even when the 2D pose detector failed[86]. In order to learn the spatial and temporal correlations, Lee et al. presented an LSTM-based multi-stage architecture based on joint interdependency to build a human body structure as a central-to-peripheral dimension extension, in accordance with natural human recognition over the temporal domain[88]. Rhodin et al. proposed a fully convolution model based on dilated temporal convolutions over the 2D pose sequence, and achieved state-of-the-art results[75].

To learn a latent 3D distribution indirectly via 2D poses, Drover et al. proposed a random projection layer to randomly project the generated 3D skeleton and then sent the resulting 2D pose to the discriminator[89].

Wang et al. designed a pairwise ranking CNN to learn the depth rankings between each pair of human joints from images, and then used it together with 2D joint locations to estimate 3D poses[90]. Park et al. adopted the structure of relational networks to capture the relations among different body parts[91]. 3D human pose was inferred from a group configuration of different body parts.

To tackle the overfitting problem, Wandt and Rosenhahn proposed a projection network (RepNet) to learn a mapping from the distribution of 2D poses to the distribution of 3D poses using an adversarial training approach[78]. Chen et al. proposed to learn 2D-3D lifting in a self-supervised manner by projecting the generated 3D skeletons to 2D, and supervised the training with a 2D pose discriminator[80]. Since 2D pose is a kind of graph-structured data, Zhao et al. proposed semantic graph convolutional networks (GCNs) for 2D to 3D human pose regression that learn semantic relationships via end-to-end training, without additional supervision or handcrafted rules[77].

3.4 Diverse hypotheses

Depth ambiguity is a major challenge when lifting 2D pose to 3D, as different 3D poses may have similar projections in 2D images. Some methods proposed to alleviate this ambiguity by generating multiple hypotheses. Simo-Serra et al. proposed a stochastic sampling strategy to propagate the noise from the image to the shape space[111]. It provides a set of plausible 3D shapes corresponding to the given 2D joint positions. An accurate 3D pose is then picked up from the sorted set by simultaneously imposing geometric and kinematic constraints. Radwan et al. employed a Twin-GP regression method to create synthetic views from the initial pose[109]. The ambiguity of the 3D pose is further reduced by enforcing kinematic and orientation constraints. Jahangiri and Yuille argued that generating a multiple pose hypotheses is more reasonable than generating only a single 3D pose, due to the occlusion and imperfection of 2D joints[98]. They proposed a method to generate multiple, diverse and valid 3D human pose hypotheses consistent with 2D joint detections. Cha et al. proposed to obtain a better pose by aggregating many weak estimations[83]. They generated the hypotheses from partial 2D pose. Li and Lee proposed using a mixture density network to solve the depth ambiguity and occlusion in 3D human pose estimation[82]. They introduced a hypotheses generator based on the features extracted from a 2D pose estimator to generate multiple feasible hypotheses.

4 Pose regression from images

Decoupling the task of 3D human pose estimation into 2D pose detection and 2D-to-3D lifting may lead to erroneous 3D poses due to the inherent ambiguity in single-view reconstruction. Directly inferring 3D human pose from images is another approach that may better leverage the rich information in raw RGB images such as shading and occlusion to resolve the ambiguity. The summary of these methods is listed in Table 3.

4.1 Hand-crafted feature-based methods

Many earlier works on human pose estimation from a single image relied on discriminatively trained models to learn a direct mapping from image features such as silhouettes[148-150] to 3D human poses without

Table 3 Summary of the pose regression based methods

Authors	Method Highlights	Evaluation Datasets	Year
Kocabas et al.[120]	Self-supervised learning. Using 2D pose estimation and epipolar geometry to obtain 3D poses.	Human3.6M, MPI-INF-3DHP	2019
Kolotouros et al.[121]	Directly regressing the 3D location of the mesh vertices while retaining the topology of the SMPL template.	Human3.6M, LSP	2019
Chen et al.[122]	Learning a shared 3D representation between viewpoints with synthesizing the human pose from one viewpoint to another.	Human3.6M, MPI-INF-3DHP, MPII	2019
Habibie et al.[123]	Comprising a disentangled hidden space encoding of explicit 2D and 3D features. Training jointly on images with 3D labels and images with only 2D labels.	Human3.6M, MPI-INF-3DHP	2019
Luo et al.[124]	Modeling 2D keypoint positions together with the orientation of limbs.	Human3.6M, MPI-INF-3DHP	2018
Kanazawa et al.[125]	Inferring 3D mesh parameters directly from image features in an end-to-end manner.	Human3.6M, MPI-INF-3DHP	2018
Luvizon et al.[126]	Designing multitask frameworks for joint 2D and 3D pose estimation from still images and action recognition from videos.	Human3.6M, MPII	2018
Yang et al.[127]	Designing a multi-source discriminator to distinguish the predicted 3D poses with adversarial learning.	Human3.6M, MPI-INF3DHP, MPII	2018
Pavlakos et al.[128]	Using a weaker supervision signal provided by the ordinal depths of human joints for end-to-end training.	Human3.6M, HumanEva-I, MPI-INF-3DHP, LSP+MPII Ordinal	2018
Rhodin et al.[129]	Considering multi-view constraints as weak supervision to train a deep network that predicts 3D pose from a single image.	Human3.6M, MPII-INF-3DHP, Ski-PosePTZ[129]	2018
Zanfir et al.[130]	Combining a single person model with additional constraints such as ground plane estimation, mutual volume exclusion, and joint inference for multiple persons in the scene.	Human3.6M, CMU Panoptic[23]	2018
Sun et al.[131]	Integral regression. Mixed usage of 3D and 2D data.	Human 3.6M, MPII, COCO Keypoint [132]	2018
Dabral et al.[73]	Exploiting temporal and structural cues present in predicted pose sequences to temporally harmonize the pose estimations.	Human3.6M, MPI-INF-3DHP	2018
Pavlakos et al.[133]	Predicting per voxel likelihoods of human joints under a coarse-to-fine framework.	Human3.6M, HumanEva-I, KTH Multiview Football-II[134]	2017
Tome et al.[135]	Multi-stage CNN architecture. Using the knowledge of plausible 3D landmark locations to refine the search for better 2D locations.	Human3.6M, MPII, LSP	2017
Rogez et al.[136]	Localization-classification-regression (LCR) architecture to estimate multiple human poses simultaneously.	Human3.6M, MPII	2017
Popa et al.[137]	Multitask-multistage architecture using multiple stages of recurrent feedforward processing.	Human3.6M, HumanEva-I, LSP	2017
Pavlakos et al.[138]	Geometry-driven 3D annotations from multi-view reconstruction to train single-view based pose estimation networks.	Human3.6M, KTH Multiview Football-II	2017
Zhou et al.[24]	Weakly supervised loss based on 2D annotation and the prior knowledge of human skeleton.	Human3.6M, MPII	2017
Sun et al.[139]	Unified 2D and 3D pose regression with a reparametrized pose representation using bones instead of joints.	Human3.6M, MPII	2017
Tekin et al.[140]	Designing a trainable fusion scheme that learns how to fuse the information optimally, instead of being hand-designed.	Human3.6m, HumanEva-I, KTH Multiview Football-II, LSP	2017
Coskun et al.[141]	Long short-term memory Kalman filter (LSTM-KF) network to yield an improved temporal regularizer.	Human 3.6M	2017
Tekin et al.[142]	Using overcomplete auto-encoders to learn high-dimensional latent pose representation.	Human3.6M	2016
Tekin et al.[143]	Directly regressing from a spatio-temporal volume of bounding boxes. Object-centric motion compensation.	Human3.6M, HumanEva-I, HumanEva-II KTH Multiview Football-II	2016
Zhou et al.[144]	Embedding a kinematic object model into the deep neutral network.	Human3.6M	2016
Sanzari et al.[145]	Hierarchical Bayesian non-parametric model. Dictionary based group 3D pose estimation.	Human3.6M	2016
Li et al.[146]	Learning image-pose embedding by a maximum-margin cost function.	Human3.6M	2015
Pons-Moll et al.[147]	Enforcing Boolean geometric relationships between body parts to resolve challenging pose ambiguities.	Annotated Posebit Database	2014

passing through 2D pose estimation.

Sanzari et al. introduced a hierarchical framework based on a Bayesian non-parametric model for 3D pose estimation[145]. Their model relies on a representation of the idiosyncratic motion of human body parts, which is captured by a subdivision of the human skeletal joints into groups. A dictionary of motion snapshots is constructed for each group, and used to estimate the likelihood of the group pose based on the extracted visual features. Tekin et al. proposed an approach to exploit the motion information from consecutive frames of a video sequence[143]. They argued that it is essential to align the successive bounding boxes of the spatio-temporal volume, so that the person inside them remains centered. To this end, they trained two CNNs to first predict large body shifts between consecutive frames, and then directly regressed from the histogram of oriented gradients (HOG) features in a spatio-temporal volume of bounding boxes to a 3D pose in the central frame.

4.2 Deep learning-based methods

The success of deep neural networks in visual tasks provides a feasible way to learn 3D pose estimation with large-scale training data.

Li et al. proposed a structured-output learning framework that takes images and 3D poses as inputs, and outputs a similarity score[146]. They designed separate sub-networks to transform the image features and poses into a joint embedding, and produced the similarity score for a given image-pose pair. Zhou et al. embedded a kinematic object model as prior knowledge into the network for general articulated object pose estimation[144]. The kinematic function is defined based on the appropriately parameterized object motion variables. Sun et al. proposed an integral approach sharing the merits of heatmaps generated by a 2D pose estimator and regression approaches[131]. Luo et al. proposed to use limb orientations to represent 3D poses, and bind the orientation together with the bounding box of each limb region to better associate images and predictions[124]. They used both 2D heatmaps and limb orientations as supervisions to train an hourglass based network.

To learn a high-dimensional latent pose representation and account for joint dependencies, Tekin et al. introduced a deep learning regression architecture for structured prediction of the 3D human pose from monocular images[142]. They trained an auto-encoder to encode 3D poses in a high-dimensional space, and designed a CNN to map the input image to this latent space. The decoding layers of the auto-encoder are then attached to project the latent space to the original pose space during the inferring period. In addition, temporal consistency is further exploited and enforced on 3D pose predictions[151].

Instead of representing the 3D pose with joint coordinates, efforts on other representations were also made. Pavlakos et al. proposed a volumetric representation for 3D human pose, which casts the problem of direct 3D coordinate regression to the form of a prediction in a discretized space[133]. They discretized the space around the joints, and used a CNN to directly predict per voxel likelihoods for each joint. Yang et al. extended the heatmap-based output strategies by predicting three two-dimensional marginal heatmaps per joint[152].

To exploit the underlying structure of human poses, Sun et al. proposed a structure-aware regression approach for both 2D and 3D pose estimation[139]. They adopted a reparametrized pose representation using bones instead of joints, and defined a compositional loss function that encodes the long-range interactions in the pose. Yang et al. proposed an adversarial learning framework to distill the 3D pose structures learned from the fully annotated datasets in constrained environments to in-the-wild images with only 2D pose annotations[127]. They designed a multi-source discriminator to distinguish the predicted 3D poses from the ground-truth. Dynamic representations of human motion are discussed in Cosfun et al.[141]. They designed an architecture built on Kalman filters and LSTM networks to yield the temporal regularization for common pose estimation tasks. The LSTM-KF adopts three LSTM modules to predict the internals of the Kalman filter, which can be end-to-end trained with backpropagation through time.

4.3 Multi-task learning

Recovering 3D human pose directly from images is a challenging task due to insufficient annotations and domain gaps between the 2D and 3D spaces. Therefore, many efforts have been made to impose 2D constraints as auxiliary supervisions for 3D human pose estimation.

Li and Chan first proposed a deep neural network for 3D human pose estimation, and designed a multi-task framework that jointly trains 3D pose regression and 2D body part detectors[153]. Tome et al. proposed a multi-stage network for joint 2D landmark and 3D human pose estimation[135]. The information captured by the 3D human pose model is embedded in the network as an additional layer that lifts 2D landmark coordinates into 3D. Tekin et al. designed a two-stream fusion framework to jointly infer 3D pose from 2D joint locations directly from the image[140]. Popa et al. presented a multitask-multistage architecture for joint 2D and 3D processing to exploit the complementary advantages of different datasets[137]. Brau and Jiang directly used a CNN for 3D human pose and camera estimation from an input image[154]. They supervised the CNN training with an additional output layer that projects the predicted 3D joints to 2D, and enforced constraints on the lengths of body parts in 3D. Habibie et al. proposed to learn 3D pose features represented as 2D heatmaps and additional 3D depth features in a hidden space[123]. Both outputs are used to predict a root-centered 3D pose and viewpoint parameters through fully connected networks.

Many works also exploited the latent complementarity of 2D and 3D tasks by designing integral frameworks. By utilizing additional information such as 2D classification results and the relative distance from multiple joints, Park et al. propagated 2D pose information to 3D pose regressors inside the CNNs[155]. Rogez et al. proposed an end-to-end localization-classification-regression (LCR) architecture for estimating multiple human poses simultaneously[136]. They obtain pose proposals by placing a fixed set of anchor-poses into the candidate regions. The pose proposals at different locations are then scored by a classification branch, and refined by a regression branch learned independently for each anchor-pose. Their extended work[156] showed that a considerable boost can be obtained by scoring the pose proposals. Mehta et al. presented a real-time method to capture the global 3D skeletal pose of a human using a single camera[157]. They integrated the 2D and 3D joint positions in a joint optimization framework along with temporal filtering and smoothing.

Except for jointly learning from 2D and 3D pose data, some works have also exploited multiple visual task learning to utilize the capabilities of deep neural networks. For instance, Luvizon et al. presented a multitask framework for simultaneous pose estimation and action recognition[126]. Both tasks share jointly learned visual features and benefit from each other.

4.4 Weakly supervised learning

In addition to the inherent difficulties, directly estimating the 3D pose generally suffers from the lack of large enough and diverse annotated 3D datasets. Existing 3D labeled datasets are usually collected in controlled studio environments, resulting in poor generalization of trained models with in-the-wild images. In addition to synthetic data[158,159], many recent works resort to diverse supervision from weak annotations or geometric constraints.

Geometric relationships of human body parts are commonly used constraints in weakly supervised methods. Pons-Moll et al. proposed posebits to represent Boolean geometric relationships between body parts, which can provide sufficient relative 3D information without requiring 3D annotation[147]. Zhou et al. incorporated the 2D pose estimation sub-network with a 3D joint depth module[24]. In order to deal with in-the-wild cases, they proposed a weakly supervised transfer learning method that mixes 2D and 3D labels to train the model. Pavlakos et al. proposed to add weak supervision provided with ordinal depths of joints to alleviate the need for accurate 3D pose annotation[128]. Ronchi et al. also employed relative depths as training data[160]. Dabral et al. proposed a structure-aware loss accounting for illegal angles and symmetries (e.g., certain body joints cannot bend beyond an angular range)[73]. The proposed geometric loss can be adapted to a weakly supervised setup to train images with only 2D annotations.

Multi-view geometry is effective in resolving the ambiguities in 3D[138,161]. Many works have used multi-view constraints as weak supervision signals. Pavlakos et al. proposed to collect 3D annotations from multi-view reconstructions to train single-view based pose estimation networks[138]. Rhodin et al. utilized a multi-view constraint as weak supervision to predict the correct pose in a small set of labeled images, and a regularization term to alleviate drift from initial predictions[129]. By learning a geometry-aware body representation from multi-view images without annotations, Rhodin et al. used an encoder-decoder to predict an image from one viewpoint, given an image from another viewpoint[162]. The representation encoding 3D geometry is then leveraged in a semi-supervised setting to learn a mapping to a 3D human pose. Chen et al. also propose a geometry-aware 3D representation by using multiple views in a simple auto-encoder model[122]. They introduced a representation consistency loss to further constrain the learning process of latent spaces. To handle the absence of camera parameters in multi-view settings, Kocabas et al. utilized epipolar geometry to obtain a 3D pose and camera geometry from the estimated 2D poses in multi-view images[120]. They then used this 3D pose to train a 3D pose estimator. Rhodin et al. proposed a self-supervision approach to train a network to produce a hierarchical scene representation from multi-view images, and regress the 3D pose from the latent representation with a simple decoder network[163].

4.5 Pose and shape estimation

Recent work has focused on the recovery of both the pose and the shape of the human body from images[28,30], benefitting from recent advances in statistical modeling of the human body[26], and more related datasets[94].

Benefiting from the task of establishing surface correspondences from depth sensors such as the Vitruvian manifold[5] and metric regression forests[8], Guler et al. established dense correspondences between an RGB image and a surface-based representation of the human body[16]. Based on the UV map representation of the human body, Yao et al. directly regressed dense 3D human poses and shapes, and trained an encoder-decoder network to map the input images and their 3D representations without solving the intermediate sub-tasks[164].

Based on the SMPL[26] model, many studies[28,30,125] proposed end-to-end CNN based frameworks to recover human meshes from monocular images. They trained the model by minimizing the re-projection loss of keypoints, which makes it applicable to in-the-wild images with only 2D annotations. In order to predict the poses and shapes of multiple people, Zanfir et al. proposed a top-down integrated approach including feedforward initialization, semantic segmentation, keypoints, and scene constraints[130]. Varol et al. also designed an end-to-end network for volumetric representation with a multi-view re-projection loss and intermediate supervision of 2D pose, 2D body part segmentation, and 3D pose[165]. By leveraging the bottom-up semantic body part segmentation and top-down constraints of the body model, Omran et al. integrated an SMPL implementation within a CNN to process the semantic part probability maps, and predicted the SMPL parameters[31]. Kolotouros et al. attempted to regress the 3D locations of the mesh vertices directly with graph convolutions[121]. To handle occlusion or truncation, Güler and Kokkinos introduced a part-based 3D parameterization, combined with 2D/3D keypoints and DensePose to reconstruct a holistic 3D human body[166]. By leveraging a dense semantic representation generated from the SMPL model, Zheng et al. introduced a volume-to-volume translation CNN for 3D human reconstruction from a single RGB image[167].

Some studies[32,33,76,168,169] also focus on capturing and modeling 3D human pose at multiple scales, including facial expressions, body and hand gestures, and personalized body shapes such as hair and clothing details, which has shown impressive results on in-the-wild images. Alldieck et al. proposed a CNN-based model to predict 3D human shapes in a canonical pose, given a few frames of a moving person[170]. Alldieck et al. also considered to turn the shape regression into an aligned image-to-image translation problem, and regressed the shape as UV-space displacements and normal maps from a partial texture map to infer the detailed human body shape[171].

5 Datasets and evaluation

5.1 Datasets

There are a number of 2D human pose datasets with high-quality annotated labels. However, the annotation for 3D human pose is far more challenging. Advanced by MoCap systems, large-scale benchmark datasets with accurate ground truth are available for researchers to train and validate their methods. The widely used datasets for 3D human pose estimation are listed in Table 4.

Table 4 Datasets for 3D human pose estimation

Dataset	Year	Characters
HumanEva[1]	2010	Marker-based MoCap in a controlled environment.
Human3.6M[2]	2014	Marker-based MoCap in a controlled environment.
TNT15[13]	2016	Indoor collection with multi-camera and wearable IMUs.
Total Capture[14]	2017	Marker-based MoCap along with IMU in a controlled environment.
MPI-INF-3DHP[172]	2017	Markerless MoCap in studio and outdoor scenes.
MuCo-3DHP[173]	2018	Composited dataset from MPI-INF-3DHP through data augmentation.
DIP-IMU[174]	2018	3D human poses captured with IMUs.
3DPW[15]	2018	3D human pose collected in the wild with IMUs and a moving camera.
MuPoTS-3D[173]	2018	Test set for in-the-wild multi-person 3D pose estimation.
AMASS[4]	2019	Unifying 15 different marker-based MoCap datasets as 3D human meshes.

In monocular 3D human pose estimation, HumanEva[1] and Human3.6M[2] are the two most widely-used indoor benchmarks in most of the works. HumanEva[1] contains two sets (HumanEva-I & HumanEva-II) with different numbers and types of video cameras, different numbers of MoCap cameras, and different types of synchronization. HumanEva-II provides only a test dataset, but no training data. A standard set of error measures are also defined for evaluating both 2D and 3D pose estimations and tracking algorithms. However, the "reconstruction error" is mostly used the evaluation protocol in this dataset, and it will be discussed in the next section. Human3.6M[2] is a large-scale dataset captured by a MoCap system including five female and six male subjects. This dataset is dedicated to be consistent with daily scenarios of human activities such as walking, eating, greeting, discussion, etc. Over 3.6 million different human poses were collected using an accurate optical MoCap system. A set of evaluation metric protocols are also defined by this dataset, and the most frequently used metrics are "mean per joint error" (MPJPE) and the "reconstruction error." A leaderboard for the evaluation set can be accessed from their website[2]. The Total Capture[14] dataset is another large-scale dataset designed for 3D pose estimation from markerless multi-camera capture. It employs multiple synchronized cameras and IMUs to obtain the accurate ground truth in a controlled environment.

However, accurate ground truth acquisition by a MoCap system is only available in indoor conditions. The generalization to in-the-wild images has been of significant concern in recent works. Newer datasets like MPI-INF-3DHP[172] and MuCo-3DHP[173], cover both indoor and outdoor scenarios. Even though the ground truth for outdoor scenes is obtained by triangulation of manually annotated 2D keypoints and data augmentation, which is not as perfect as MoCap data, these datasets still provide effective benchmarks to enable generalization of 3D pose estimation methods.

In addition, the recent 3DPW[15] dataset that captured accurate 3D human pose in-the-wild using IMUs and a moving camera, fills in the gap of a missing benchmark dataset with ground truth in the wild. This dataset includes more than 51000 frames of seven actors in 18 clothing styles and during various outdoor activities such as shopping, playing sports, discussing, capturing selfies, riding in buses, and relaxing. It also takes into account joint error, mesh error, and orientation error to evaluate the shape accuracy and pose accuracy. The realistically recorded dataset in cluttered outdoor scenes facilitates the quantitative evaluation of monocular-based methods in more realistic scenarios.

5.2 Evaluation metrics

The most common evaluation metrics used for evaluating 3D pose estimation accuracy are listed below.

MPJPE: The mean per joint position error (MPJPE) is the most common evaluation metric in 3D human pose estimation. MPJPE is defined as the mean Euclidean distance between the estimated joint position and the ground truth in 3D, after the root joints are aligned by a translation:

$$E\left(x, \tilde{x}\right) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \tilde{x}_i\|_2 \tag{1}$$

Here, E, N, x, and $\tilde{x}$ refer to the error function, the number of joints, the ground truth pose, and the estimated pose, respectively. x_i denotes the 3D position for the i-th joint.

Reconstruction error: Introduced by Simo-Serra et al. [111], the reconstruction error is the MPJPE after rigid alignment between the estimated poses and the ground truth.

3DPCK: It is a 3D extension of the percentage of correct keypoints (PCK) metric[175] used for 2D human pose evaluation, in which a detected joint is considered correct if the distance between the predicted and the true joint is within a certain threshold. 3DPCK is employed as an alternative metric in the MPI-INF-3DHP[172] dataset, account for the lack of accurate ground truth.

There are pros and cons among these metrics. For instance, 3DPCK is more robust against inaccurate ground truth, but cannot indicate the precisions of subtle pose details. A lower MPJPE may not necessarily mean a more realistic pose, as the joint position also depends on the predicted scale of the skeleton. The joint angle error[2] offers the possibility of evaluating the pose accuracy regardless of the size of the skeleton and bone length, which is crucial in some real-world applications like virtual reality and robotics. However, it is rarely used, as most joint-based pose estimation methods only output joint locations from which full joint angles cannot be recovered. Recently, the mesh error[15] between the predicted 3D mesh and

[2] http://vision.imar.ro/human3.6m/ranking.php

the ground truth has been typically used to evaluate the methods for shape estimation.

5.3 Quantitative comparison

In order to provide an overview of the performance of several representative 3D human pose estimation approaches, we summarize the reported results on the Human3.6M[2] and MPI-INF-3DHP[172] datasets. The evaluation results for all the methods are adopted from the original works. For Human3.6M, we consider the most commonly used protocol in which subjects S1, S5, S6, S7, and S8 are used for training, and subjects S9 and S11 are used for testing. The detailed results on Human3.6M are listed in Table 5.

Table 5 Detailed results on Human3.6M

Method	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Tekin et al.[143]	102.4	147.2	88.8	125.3	118	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou et al.[99]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Tome et al.[135]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Pavlakos et al.[133]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Tekin et al.[140]	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	74.3	69.7
Martinez et al.[92]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhou et al.[24]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	51.4	63.2	55.3	64.9
Yang et al.[127]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Pavlakos et al.[128]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Lee et al.[88]	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Zhao et al.[77]	47.3	60.7	51.4	60.5	61.1	49.9	47.3	68.1	86.2	55.0	67.8	61.0	42.1	60.6	45.3	57.6
Li and Lee[82]	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
Sun et al.[131]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
Moon et al.[176]	50.5	55.7	50.1	51.7	53.9	46.8	50.0	61.9	68.0	52.5	55.9	49.9	41.8	56.1	46.9	53.3
Kanazawa et al.[125]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	87.97
Güler et al.[177]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	64.28
Rhodin et al.[75]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Chen et al.[122]	41.1	44.2	44.9	45.9	46.5	39.3	41.6	54.8	73.2	46.2	48.7	42.1	35.8	46.6	38.5	46.3

Numbers are MPJPEs (mm). The results of all approaches are obtained from the original papers and supplementary materials.

Many recent works[24,128,172] reported results on the MPI-INF-3DHP dataset to illustrate the generalization for outdoor images. We summarize their reported results in Table 6.

Table 6 Detailed results on MPI-INF-3DHP

Method	Studio GS	Studio no GS	Outdoor	All	All
	3DPCK	3DPCK	3DPCK	3DPCK	AUC
Mehta et al.[172]	70.8	62.3	58.8	64.7	31.7
Zhou et al.[24]	71.1	64.7	72.7	69.2	32.5
Pavlakos et al.[128]	76.5	63.1	77.5	71.9	35.3
Martinez et al.[92](*)	–	–	–	68	34.7
Sun et al.[139](*)	–	–	–	68.4	29.4
Habibie et al.[123]	–	–	–	70.4	36.0
Chen et al.[122]	–	–	–	75.9	36.3

Numbers are PCKs. The results of all approaches are obtained from the original works, except for the (*) marked ones, which are obtained from literature[122].

In summary, 2D-3D lifting methods generally outperform direct regression of 3D pose from images, since the two decoupled stages can make use of the power of state-of-the-art 2D pose estimators.

Moreover, the 2D-3D pose lifting stage is generalizable, as it is independent of images, and can be trained with only the MoCap data, which is abundant. In contrast, direct regression from images requires image and 3D pose pairs, which can hardly be collected at a large scale except in MoCap studios. Nevertheless, direct pose regression from images also shows promising results and great potential, considering its capability to leverage more image information, and the fact that diverse training data with 3D annotations may be available in future.

6 Conclusion and future directions

In this review, we provided an overview of monocular 3D pose estimation. We summarized existing literature by a three-way classification: 3D pose tracking, 2D-3D pose lifting, and pose regression from images. We also introduced and discussed the commonly used datasets and metrics. Despite the remarkable achievements so far, recovering accurate 3D human pose remains difficult in general applications. A few possible research directions are discussed below.

The foremost challenge is the generalizability of 3D human pose estimation algorithms in natural scenes, as most of the existing methods report satisfactory results only on benchmark datasets with training and test sets from the same scene. High quality annotated 2D datasets have facilitated impressive performance in 2D detection and pose estimation tasks, but acquiring large scale 3D labels in-the-wild is not feasible. Recently collected datasets using portable capture systems (e. g., wearable IMUs[15]) may provide more diverse training and test data for future research. Leveraging geometric constraints or videos for self-supervised learning of 3D poses is also a promising approach to more generalizable pose learning.

The second challenge is 3D pose estimation of multiple people, where relative positions among people need to be inferred along with each 3D pose. Inter-person occlusion brings additional challenges, but provides cues for estimating relative positions of people. Algorithms for multi-person 2D pose tracking and estimation like OpenPose[178] provide inspiring ideas for the 3D counterpart. The analysis of consistent correspondences of individual motion over frames could be a promising solution to address the issue of heavy occlusion. The scene geometry is also an important cue for solving multi-person pose estimation, e.g., the feet of people should be on the same ground plane[130].

A limitation of existing methods on 3D pose estimation is only focusing on human bodies without considering human-object and human-scene interactions. Scene constraints and affordance have been studied in latest works[179,180]. Considering the interactions provides physical constraints and semantic cues for more precise pose estimation, which will enhance the visual realism and user experience in applications like virtual reality. In addition, current pose estimation methods mostly focus on single-frame estimation, and commonly used metrics like MPJPE do not necessarily favor smooth and realistic motion. Capturing more physically realistic motion from monocular videos and modeling human dynamics are promising directions that are worth further investigation.

Declaration of competing interest

We declare that we have no conflict of interest.

References

1 Sigal L, Balan A O, Black M J. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision, 2010, 87(1‒2): 4‒27
DOI:10.1007/s11263-009-0273-6

2 Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: large scale datasets and predictive methods for 3D human

sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339
DOI:10.1109/tpami.2013.248

3 Loper M, Mahmood N, Black M J. MoSh. ACM Transactions on Graphics, 2014, 33(6): 1–13
DOI:10.1145/2661229.2661273

4 Mahmood N, Ghorbani N, Troje N F, Pons-Moll G, Black M. AMASS: archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), IEEE, 2019
DOI:10.1109/iccv.2019.00554

5 Taylor J, Shotton J, Sharp T, Fitzgibbon A. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, IEEE, 2012
DOI:10.1109/cvpr.2012.6247664

6 Shotton J, Girshick R, Fitzgibbon A, Sharp T, Cook M, Finocchio M, Moore R, Kohli P, Criminisi A, Kipman A, Blake A. Efficient human pose estimation from single depth images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2821–2840
DOI:10.1109/tpami.2012.241

7 Han J G, Shao L, Xu D, Shotton J. Enhanced computer vision with microsoft kinect sensor: a review. IEEE Transactions on Cybernetics, 2013, 43(5): 1318–1334
DOI:10.1109/tcyb.2013.2265378

8 Pons-Moll G, Taylor J, Shotton J, Hertzmann A, Fitzgibbon A. Metric regression forests for correspondence estimation. International Journal of Computer Vision, 2015, 113(3): 163–175
DOI:10.1007/s11263-015-0818-9

9 Haque A, Peng B Y, Luo Z L, Alahi A, Yeung S, Li F F. Towards viewpoint invariant 3D human pose estimation// Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, 160–177
DOI:10.1007/978-3-319-46448-0_10

10 Yu T, Zheng Z R, Guo K W, Zhao J H, Dai Q H, Li H, Pons-Moll G, Liu Y B. DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00761

11 Pons-Moll G, Baak A, Helten T, Muller M, Seidel H P, Rosenhahn B. Multisensor-fusion for 3D full-body human motion capture. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA, IEEE, 2010
DOI:10.1109/cvpr.2010.5540153

12 Pons-Moll G, Baak A, Gall J, Leal-Taixe L, Muller M, Seidel H P, Rosenhahn B. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: 2011 International Conference on Computer Vision. Barcelona, Spain, IEEE, 2011
DOI:10.1109/iccv.2011.6126375

13 Marcard T V, Pons-Moll G, Rosenhahn B. Human pose estimation from video and IMUs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(8): 1533–1547
DOI:10.1109/TPAMI.2016.2522398

14 Trumble M, Gilbert A, Malleson C, Hilton A, Collomosse J. Total capture: 3D human pose estimation fusing video and inertial sensors. In: Procedings of the British Machine Vision Conference 2017. London, UK, British Machine Vision Association, 2017
DOI:10.5244/c.31.14

15 von Marcard T, Henschel R, Black M J, Rosenhahn B, Pons-Moll G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018, 614–631
DOI:10.1007/978-3-030-01249-6_37

16 Guler R A, Neverova N, Kokkinos I. DensePose: dense human pose estimation in the wild. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00762

17 Zhang W Y, Zhu M L, Derpanis K G. From actemes to action: a strongly-supervised representation for detailed action understanding. In: 2013 IEEE International Conference on Computer Vision. Sydney, Australia, IEEE, 2013
DOI:10.1109/iccv.2013.280

18 Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE, 2014
DOI:10.1109/cvpr.2014.214

19 Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation// Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, 483–499
DOI:10.1007/978-3-319-46484-8_29

20 Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G, Sun J. Cascaded pyramid network for multi-person pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00742

21 Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: new benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE, 2014
DOI:10.1109/cvpr.2014.471

22 Yang W, Li S, Ouyang W L, Li H S, Wang X G. Learning feature Pyramids for human pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.144

23 Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y. Panoptic studio: a massively multiview system for social motion capture. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, IEEE, 2015
DOI:10.1109/iccv.2015.381

24 Zhou X Y, Huang Q X, Sun X, Xue X Y, Wei Y C. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.51

25 Rogez G, Schmid C. Image-based synthesis for deep 3D human pose estimation. International Journal of Computer Vision, 2018, 126(9): 993–1008
DOI:10.1007/s11263-018-1071-9

26 Loper M, Mahmood N, Romero J, Pons-Moll G, Black M J. SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, 2015, 34(6): 248
DOI:10.1145/2816795.2818013

27 Guan P, Weiss A, Balan A O, Black M J. Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision. Kyoto, IEEE, 2009
DOI:10.1109/iccv.2009.5459300

28 Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black M J. Keep it SMPL: automatic estimation of 3D human pose and shape from a single image//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 561–578
DOI:10.1007/978-3-319-46454-1_34

29 Tung H Y, Tung H W, Yumer E, Fragkiadaki K. Self-supervised learning of motion capture. Advances in Neural Information Processing Systems, 2017, 5236–5246

30 Pavlakos G, Zhu L Y, Zhou X W, Daniilidis K. Learning to estimate 3D human pose and shape from a single color image. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00055

31 Omran M, Lassner C, Pons-Moll G, Gehler P, Schiele B. Neural body fitting: unifying deep learning and model based human pose and shape estimation. In: 2018 International Conference on 3D Vision (3DV). Verona, IEEE, 2018

DOI:10.1109/3dv.2018.00062

32 Joo H, Simon T, Sheikh Y. Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00868

33 Xiang D L, Joo H, Sheikh Y. Monocular total capture: posing face, body, and hands in the wild. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.01122

34 Moeslund T B, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006, 104(2/3): 90‒126
DOI:10.1016/j.cviu.2006.08.002

35 Poppe R. Vision-based human motion analysis: an overview. Computer Vision and Image Understanding, 2007, 108(1/2): 4‒18
DOI:10.1016/j.cviu.2006.10.016

36 Pons-Moll G, Rosenhahn B. Model-based pose estimation//Visual Analysis of Humans. London: Springer London, 2011, 139‒170
DOI:10.1007/978-0-85729-997-0_9

37 Sminchisescu C, Bo L F, Ionescu C, Kanaujia A. Feature-based pose estimation//Visual Analysis of Humans. London: Springer London, 2011, 225‒251
DOI:10.1007/978-0-85729-997-0_12

38 Perez-Sala X, Escalera S, Angulo C, Gonzàlez J. A survey on model based approaches for 2D and 3D visual human pose recovery. Sensors, 2014, 14(3): 4189‒4210
DOI:10.3390/s140304189

39 Liu Z, Zhu J K, Bu J J, Chen C. A survey of human pose estimation: the body parts parsing based methods. Journal of Visual Communication and Image Representation, 2015, 32: 10‒19
DOI:10.1016/j.jvcir.2015.06.013

40 Gong W J, Zhang X N, Gonzàlez J, Sobral A, Bouwmans T, Tu C H, Zahzah E H. Human pose estimation from monocular images: a comprehensive survey. Sensors, 2016, 16(12): 1966
DOI:10.3390/s16121966

41 Holte M B, Tran C, Trivedi M M, Moeslund T B. Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(5): 538‒552
DOI:10.1109/jstsp.2012.2196975

42 Sarafianos N, Boteanu B, Ionescu B, Kakadiaris I A. 3D Human pose estimation: a review of the literature and analysis of covariates. Computer Vision and Image Understanding, 2016, 152: 1‒20
DOI:10.1016/j.cviu.2016.09.002

43 Bregler C, Malik J. Tracking people with twists and exponential maps. In: Proceedings 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Santa Barbara, CA, USA, IEEE, 1998
DOI:10.1109/cvpr.1998.698581

44 Sminchisescu C, Triggs B. Kinematic jump processes for monocular 3D human tracking. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, WI, USA, IEEE, 2003
DOI:10.1109/cvpr.2003.1211339

45 Sigal L, Isard M, Haussecker H, Black M J. Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation. International Journal of Computer Vision, 2012, 98(1): 15‒48
DOI:10.1007/s11263-011-0493-4

46 Zhou X W, Zhu M L, Leonardos S, Daniilidis K. Sparse representation for 3D shape estimation: a convex relaxation approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1648‒1661
DOI:10.1109/tpami.2016.2605097

47 Andriluka M, Roth S, Schiele B. Monocular 3D pose estimation and tracking by detection. In: 2010 IEEE Computer

Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA, IEEE, 2010
DOI:10.1109/cvpr.2010.5540156

48 Leonardos S, Zhou X W, Daniilidis K. Articulated motion estimation from a monocular image sequence using spherical tangent bundles. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden, IEEE, 2016
DOI:10.1109/icra.2016.7487183

49 Simo-Serra E, Torras C, Moreno-Noguer F. Geodesic finite mixture models. In: Proceedings of the British Machine Vision Conference 2014. Nottingham, British Machine Vision Association, 2014
DOI:10.5244/c.28.91

50 Hauberg S, Lauze F, Pedersen K S. Unscented kalman filtering on Riemannian manifolds. Journal of Mathematical Imaging and Vision, 2013, 46(1): 103‒120
DOI:10.1007/s10851-012-0372-9

51 Yao A, Gall J, Gool L V, Urtasun R. Learning probabilistic non-linear latent variable models for tracking complex activities. Advances in Neural Information Processing Systems, 2011, 1359‒1367

52 Gall J, Rosenhahn B, Brox T, Seidel H P. Optimization and filtering for human motion capture. International Journal of Computer Vision, 2010, 87(1/2): 75‒92
DOI:10.1007/s11263-008-0173-1

53 Taylor G W, Sigal L, Fleet D J, Hinton G E. Dynamical binary latent variable models for 3D human pose tracking. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. SanFrancisco, CA, USA, IEEE, 2010
DOI:10.1109/cvpr.2010.5540157

54 Gall J, Rosenhahn B, Seidel H P. Drift-free tracking of rigid and articulated objects. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA, IEEE, 2008
DOI:10.1109/cvpr.2008.4587558

55 Wang J, Hertzmann A, Fleet D J. Gaussian process dynamical models. Advances in Neural Information Processing Systems, 2006, 1441‒1448

56 Deutscher J, Reid I. Articulated body motion capture by stochastic search. International Journal of Computer Vision, 2005, 61(2): 185‒205
DOI:10.1023/b:visi.0000043757.18370.9c

57 Sminchisescu C, Triggs B. Covariance scaled sampling for monocular 3D body tracking. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA, IEEE, 2001
DOI:10.1109/cvpr.2001.990509

58 Bregler C, Malik J, Pullen K. Twist based acquisition and tracking of animal and human kinematics. International Journal of Computer Vision, 2004, 56(3): 179‒194
DOI:10.1023/b:visi.0000011203.00237.9b

59 Gavrila D M, Davis L S. 3D model-based tracking of humans in action: a multi-view approach. In: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA, IEEE, 1996
DOI:10.1109/cvpr.1996.517056

60 Kalman R E. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 1960, 82(1): 35‒45
DOI:10.1115/1.3662552

61 Julier S J, Uhlmann J K. New extension of the Kalman filter to nonlinear systems. Signal processing, sensor fusion, and target recognition VI, 1997, 3068, 182
DOI:10.1117/12.280797

62 Wachter S, Nagel H H. Tracking persons in monocular image sequences. Computer Vision and Image Understanding, 1999, 74(3): 174‒192
DOI:10.1006/cviu.1999.0758

63 Kraft E. A quaternion-based unscented Kalman filter for orientation tracking. In: Proceedings of the Sixth International

Conference of Information Fusion. Cairns, Queensland, Australia, IEEE, 2003
DOI:10.1109/icif.2003.177425

64 Sidenbladh H, Black M J, Fleet D J. Stochastic tracking of 3D human figures using 2D image motion// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, 702–718
DOI:10.1007/3-540-45053-x_45

65 Choo K, Fleet D J. People tracking using hybrid Monte Carlo filtering. In: Proceedings Eighth IEEE International Conference on Computer Vision. Vancouver, BC, Canada, IEEE Comput. Soc, 2001
DOI:10.1109/iccv.2001.937643

66 Howe N R, Leventon M E, Freeman W T. Bayesian reconstruction of 3D human motion from single-camera video. Advances in Neural Information Processing Systems, 2000, 820–826

67 Elgammal A, Lee C S. Inferring 3D body pose from silhouettes using activity manifold learning. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA, IEEE, CVPR 2004
DOI:10.1109/cvpr.2004.1315230

68 Simo-Serra E, Torras C, Moreno-Noguer F. Lie algebra-based kinematic prior for 3D human pose tracking. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA). Tokyo, Japan, IEEE, 2015
DOI:10.1109/mva.2015.7153212

69 Andriluka M, Roth S, Schiele B. People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA, IEEE, 2008
DOI:10.1109/cvpr.2008.4587583

70 Urtasun R, Fleet D J, Fua P. 3D people tracking with Gaussian process dynamical models. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1. New York, USA, IEEE
DOI:10.1109/cvpr.2006.15

71 Elhayek A, de Aguiar E, Jain A, Tompson J, Pishchulin L, Andriluka M, Bregler C, Schiele B, Theobalt C. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, IEEE, 2015
DOI:10.1109/cvpr.2015.7299005

72 Wei S, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). LasVegas, NV, USA, IEEE, 2016
DOI:10.1109/cvpr.2016.511

73 Dabral R, Mundhada A, Kusupati U, Afaque S, Sharma A, Jain A. Learning 3D human pose from structure and motion// Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018, 679–696
DOI:10.1007/978-3-030-01240-3_41

74 Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference 2010. Aberystwyth, British Machine Vision Association, 2010
DOI:10.5244/c.24.12

75 Rhodin H, Constantin V, Katircioglu I, Salzmann M, Fua P. Neural scene decomposition for multi-person motion capture. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00789

76 Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman A A, Tzionas D, Black M J. Expressive body capture: 3D hands, face, and body from a single image. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). LongBeach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.01123

77 Zhao L, Peng X, Tian Y, Kapadia M, Metaxas D N. Semantic graph convolutional networks for 3D human pose regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00354

78 Wandt B, Rosenhahn B. RepNet: weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00797

79 Arnab A, Doersch C, Zisserman A. Exploiting temporal context for 3D human pose estimation in the wild. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00351

80 Chen C H, Tyagi A, Agrawal A, Drover D, Mv R, Stojanov S, Rehg J M. Unsupervised 3D pose estimation with geometric self-supervision. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00586

81 Véges M, Varga V, Lőrincz A. 3D human pose estimation with Siamese equivariant embedding. Neurocomputing, 2019, 339, 194‒201
DOI:10.1016/j.neucom.2019.02.029

82 Li C, Lee G H. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.01012

83 Cha G, Lee M, Cho J, Oh S. Deep pose consensus networks. Computer Vision and Image Understanding, 2019, 182: 64‒70
DOI:10.1016/j.cviu.2019.03.004

84 Yang J J, Wan L L, Xu W R, Wang S H. 3D human pose estimation from a single image via exemplar augmentation. Journal of Visual Communication and Image Representation, 2019, 59: 371‒379
DOI:10.1016/j.jvcir.2019.01.033

85 Fang H S, Xu Y, Wang W, Liu X, Zhu S C. Learning pose grammar to encode human body configuration for 3D pose estimation. Thirty-Second AAAI Conference on Artificial Intelligence, 2018

86 Hossain M R I, Little J J. Exploiting temporal information for 3D human pose estimation//Computer Vision‒ECCV 2018. Cham: Springer International Publishing, 2018, 69‒86
DOI:10.1007/978-3-030-01249-6_5

87 Zhao R Q, Wang Y, Martinez A M. A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3059‒3066
DOI:10.1109/tpami.2017.2772922

88 Lee K, Lee I, Lee S. Propagating LSTM: 3D pose estimation based on joint interdependency// Computer Vision‒ECCV 2018. Cham: Springer International Publishing, 2018, 123‒141
DOI:10.1007/978-3-030-01234-2_8

89 Drover D, V R M, Chen C H, Agrawal A, Tyagi A, Huynh C P. Can 3D pose be learned from 2D projections alone?// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, 78‒94
DOI:10.1007/978-3-030-11018-5_7

90 Wang M, Chen X P, Liu W T, Qian C, Lin L, Ma L Z. DRPose3D: depth ranking in 3D human pose estimation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden, California, International Joint Conferences on Artificial Intelligence Organization, 2018
DOI:10.24963/ijcai.2018/136

91 Park S, Kwak N. 3D human pose estimation with relational networks. British Machine Vision Conference (BMVC), 2018

92 Martinez J, Hossain R, Romero J, Little J J. A simple yet effective baseline for 3D human pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.288

93 Moreno-Noguer F. 3D human pose estimation from a single image via distance matrix regression. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017

DOI:10.1109/cvpr.2017.170

94 Lassner C, Romero J, Kiefel M, Bogo F, Black M J, Gehler P V. Unite the people: closing the loop between 3D and 2D human representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.500

95 Nie B X, Wei P, Zhu S C. Monocular 3D human pose estimation by predicting depth on joints. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.373

96 Shu T, Ryoo M S, Zhu S C. Learning social affordance for human-robot interaction. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, 3454–3461

97 Lin M D, Lin L, Liang X D, Wang K Z, Cheng H. Recurrent 3D pose sequence machines. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.588

98 Jahangiri E, Yuille A L. Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, IEEE, 2017
DOI:10.1109/iccvw.2017.100

99 Zhou X W, Zhu M L, Leonardos S, Derpanis K G, Daniilidis K. Sparseness meets deepness: 3D human pose estimation from monocular video. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). LasVegas, NV, USA, IEEE, 2016
DOI:10.1109/cvpr.2016.537

100 Yasin H, Iqbal U, Kruger B, Weber A, Gall J. A dual-source approach for 3D pose estimation from a single image. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). LasVegas, NV, USA, IEEE, 2016
DOI:10.1109/cvpr.2016.535

101 Du Y, Wong Y, Liu Y H, Han F L, Gui Y L, Wang Z, Kankanhalli M, Geng W D. Marker-less 3D human motion capture with monocular image sequence and height-maps//Computer Vision−ECCV 2016. Cham: Springer International Publishing, 2016, 20–36
DOI:10.1007/978-3-319-46493-0_2

102 Li W H, Wong Y K, Liu A A, Li Y, Su Y T, Kankanhalli M. Multi-camera action dataset (MCAD): a dataset for studying non-overlapped cross-camera action recognition. 2016

103 Akhter I, Black M J. Pose-conditioned joint angle limits for 3D human pose reconstruction. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, IEEE, 2015
DOI:10.1109/cvpr.2015.7298751

104 Zhou X W, Leonardos S, Hu X Y, Daniilidis K. 3D shape estimation from 2D landmarks: a convex relaxation approach. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, IEEE, 2015
DOI:10.1109/cvpr.2015.7299074

105 Wang C Y, Wang Y Z, Lin Z C, Yuille A L, Gao W. Robust estimation of 3D human poses from a single image. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE, 2014
DOI:10.1109/cvpr.2014.303

106 Hofmann M, Gavrila D M. Multi-view 3D human pose estimation combining single-frame recovery, temporal integration and model adaptation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, IEEE, 2009
DOI:10.1109/cvpr.2009.5206508

107 Ionescu C, Carreira J, Sminchisescu C. Iterated second-order label sensitive pooling for 3D human pose estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE, 2014
DOI:10.1109/cvpr.2014.215

108 Simo-Serra E, Quattoni A, Torras C, Moreno-Noguer F. A joint model for 2D and 3D pose estimation from a single image. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA, IEEE, 2013
DOI:10.1109/cvpr.2013.466

109 Radwan I, Dhall A, Goecke R. Monocular image 3D human pose estimation under self-occlusion. In: 2013 IEEE International Conference on Computer Vision. Sydney, Australia, IEEE, 2013
DOI:10.1109/iccv.2013.237

110 Ramakrishna V, Kanade T, Sheikh Y. Reconstructing 3D human pose from 2D image landmarks//Computer Vision–ECCV 2012. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 573–586
DOI:10.1007/978-3-642-33765-9_41

111 Simo-Serra E, Ramisa A, Alenya G, Torras C, Moreno-Noguer F. Single image 3D human pose estimation from noisy observations. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, IEEE, 2012
DOI:10.1109/cvpr.2012.6247988

112 Jiang H. 3D human pose reconstruction using millions of exemplars. In: 2010 20th International Conference on Pattern Recognition. Istanbul, Turkey, IEEE, 2010
DOI:10.1109/icpr.2010.414

113 Shakhnarovich G, Viola P, Darrell T. Fast pose estimation with parameter-sensitive hashing. In: Proceedings Ninth IEEE International Conference on Computer Vision. Nice, France, IEEE, 2003
DOI:10.1109/iccv.2003.1238424

114 Mori G, Malik J. Estimating human body configurations using shape context matching// Computer Vision–ECCV 2002. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, 666–680
DOI:10.1007/3-540-47977-5_44

115 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. In: CVPR 2011. Colorado Springs, CO, USA, IEEE, 2011
DOI:10.1109/cvpr.2011.5995741

116 Pishchulin L, Insafutdinov E, Tang S Y, Andres B, Andriluka M, Gehler P, Schiele B. DeepCut: joint subset partition and labeling for multi person pose estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). LasVegas, NV, USA, IEEE, 2016
DOI:10.1109/cvpr.2016.533

117 Alldieck T, Magnor M, Xu W P, Theobalt C, Pons-Moll G. Video based reconstruction of 3D people models. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00875

118 Habermann M, Xu W P, Zollhöfer M, Pons-Moll G, Theobalt C. LiveCap. ACM Transactions on Graphics, 2019, 38(2): 1–17
DOI:10.1145/3311970

119 Chen C H, Ramanan D. 3D human pose estimation = 2D pose estimation + matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.610

120 Kocabas M, Karagoz S, Akbas E. Self-supervised learning of 3D human pose using multi-view geometry. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00117

121 Kolotouros N, Pavlakos G, Daniilidis K. Convolutional mesh regression for single-image human shape reconstruction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.00463

122 Chen X P, Lin K Y, Liu W T, Qian C, Lin L. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.01115

123 Habibie I, Xu W P, Mehta D, Pons-Moll G, Theobalt C. In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019
DOI:10.1109/cvpr.2019.01116

124 Luo C, Chu X, Yuille A. OriNet: A fully convolutional network for 3D human pose estimation. British Machine Vision Conference(BMVC), 2018

125 Kanazawa A, Black M J, Jacobs D W, Malik J. End-to-end recovery of human shape and pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00744

126 Luvizon D C, Picard D, Tabia H. 2D/3D pose estimation and action recognition using multitask deep learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00539

127 Yang W, Ouyang W, Wang X L, Ren J, Li H S, Wang X G. 3D human pose estimation in the wild by adversarial learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00551

128 Pavlakos G, Zhou X W, Daniilidis K. Ordinal depth supervision for 3D human pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018
DOI:10.1109/cvpr.2018.00763

129 Rhodin H, Meyer F, Sporri J, Muller E, Constantin V, Fua P, Katircioglu I, Salzmann M. Learning monocular 3D human pose estimation from multi-view images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00880

130 Zanfir A, Marinoiu E, Sminchisescu C. Monocular 3D pose and shape estimation of multiple people in natural scenes: the importance of multiple scene constraints. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
DOI:10.1109/cvpr.2018.00229

131 Sun X, Xiao B, Wei F Y, Liang S, Wei Y C. Integral human pose regression//Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018, 536–553
DOI:10.1007/978-3-030-01231-1_33

132 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: common objects in context// Computer Vision – ECCV 2014. Cham: Springer International Publishing, 2014, 740–755
DOI:10.1007/978-3-319-10602-1_48

133 Pavlakos G, Zhou X W, Derpanis K G, Daniilidis K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.139

134 Kazemi V, Burenius M, Azizpour H, Sullivan J. Multi-view body part recognition with random forests. In: Proceedings of the British Machine Vision Conference 2013. Bristol. British Machine Vision Association, 2013
DOI:10.5244/c.27.48

135 Tome D, Russell C, Agapito L. Lifting from the deep: convolutional 3D pose estimation from a single image. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.603

136 Rogez G, Weinzaepfel P, Schmid C. LCR-net: localization-classification-regression for human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.134

137 Popa A I, Zanfir M, Sminchisescu C. Deep multitask architecture for integrated 2D and 3D human sensing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.501

138 Pavlakos G, Zhou X W, Derpanis K G, Daniilidis K. Harvesting multiple views for marker-less 3D human pose annotations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017
DOI:10.1109/cvpr.2017.138

139 Sun X, Shang J X, Liang S, Wei Y C. Compositional human pose regression. In: 2017 IEEE International Conference on

Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.284

140 Tekin B, Marquez-Neila P, Salzmann M, Fua P. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.425

141 Coskun H, Achilles F, DiPietro R, Navab N, Tombari F. Long short-term memory kalman filters: recurrent neural estimators for pose regularization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, IEEE, 2017
DOI:10.1109/iccv.2017.589

142 Tekin B, Katircioglu I, Salzmann M, Lepetit V, Fua P. Structured prediction of 3D human pose with deep neural networks. In: Proceedings of the British Machine Vision Conference 2016. York, UK, British Machine Vision Association, 2016
DOI:10.5244/c.30.130

143 Tekin B, Rozantsev A, Lepetit V, Fua P. Direct prediction of 3D body poses from motion compensated sequences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, IEEE, 2016, 991–1000
DOI:10.1109/cvpr.2016.113

144 Zhou X Y, Sun X, Zhang W, Liang S, Wei Y C. Deep kinematic pose regression//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, 186–201
DOI:10.1007/978-3-319-49409-8_17

145 Sanzari M, Ntouskos V, Pirri F. Bayesian image based 3D pose estimation//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, 566–582
DOI:10.1007/978-3-319-46484-8_34

146 Li S J, Zhang W C, Chan A B. Maximum-margin structured learning with deep networks for 3D human pose estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, IEEE, 2015, 2848–2856
DOI:10.1109/iccv.2015.326

147 Pons-Moll G, Fleet D J, Rosenhahn B. Posebits for monocular human pose estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE, 2014, 2345–2352
DOI:10.1109/cvpr.2014.300

148 Agarwal A, Triggs B. 3D human pose from silhouettes by relevance vector regression. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA, IEEE, 2004, 2: 882–888
DOI:10.1109/cvpr.2004.1315258

149 Mori G, Malik J. Recovering 3D human body configurations using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1052–1062
DOI:10.1109/tpami.2006.149

150 Agarwal A, Triggs B. Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 44–58
DOI:10.1109/tpami.2006.21

151 Katircioglu I, Tekin B, Salzmann M, Lepetit V, Fua P. Learning latent representations of 3D human pose with deep neural networks. International Journal of Computer Vision, 2018, 126(12): 1326–1341
DOI:10.1007/s11263-018-1066-6

152 Yang J J, Wan L L, Xu W R, Wang S H. 3D human pose estimation from a single image via exemplar augmentation. Journal of Visual Communication and Image Representation, 2019, 59: 371–379
DOI:10.1016/j.jvcir.2019.01.033

153 Li S J, Chan A B. 3D human pose estimation from monocular images with deep convolutional neural network//Computer Vision-ACCV 2014. Cham: Springer International Publishing, 2015, 332–347
DOI:10.1007/978-3-319-16808-1_23

154 Brau E, Jiang H. 3D human pose estimation via deep learning from 2D annotations. In: 2016 Fourth International Conference on 3D Vision (3DV). Stanford, CA, USA, IEEE, 2016, 582–591
DOI:10.1109/3dv.2016.84

155 Park S, Hwang J, Kwak N. 3D human pose estimation using convolutional neural networks with 2D pose information//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, 156–169
DOI:10.1007/978-3-319-49409-8_15

156 Rogez G, Weinzaepfel P, Schmid C. LCR-net++: multi-person 2D and 3D pose detection in natural images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 1
DOI:10.1109/tpami.2019.2892985

157 Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel H P, Xu W P, Casas D, Theobalt C. VNect. ACM Transactions on Graphics, 2017, 36(4): 1–14
DOI:10.1145/3072959.3073596

158 Rogez G, Schmid C. Mocap-guided data augmentation for 3D pose estimation in the wild. Advances in Neural Information Processing Systems, 2016, 3108–3116

159 Chen W Z, Wang H, Li Y Y, Su H, Wang Z H, Tu C H, Lischinski D, Cohen-Or D, Chen B Q. Synthesizing training images for boosting human 3D pose estimation. In: 2016 Fourth International Conference on 3D Vision (3DV). Stanford, CA, USA, IEEE, 2016, 479–488
DOI:10.1109/3dv.2016.58

160 Ronchi M R, Mac Aodha O, Eng R, Perona P. It's all Relative: Monocular 3D Human Pose Estimation from Weakly Supervised Data. British Machine Vision Conference (BMVC), 2018

161 Dong J T, Jiang W, Huang Q X, Bao H J, Zhou X W. Fast and robust multi-person 3D pose estimation from multiple views. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019, 7784–7793
DOI:10.1109/cvpr.2019.00798

162 Rhodin H, Salzmann M, Fua P. Unsupervised geometry-aware representation for 3D human pose estimation//Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018, 765–782
DOI:10.1007/978-3-030-01249-6_46

163 Rhodin H, Constantin V, Katircioglu I, Salzmann M, Fua P. Neural scene decomposition for multi-person motion capture. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019, 7695–7705
DOI:10.1109/cvpr.2019.00789

164 Yao P, Fang Z, Wu F, Feng Y, Li J. DenseBody: directly regressing dense 3D human pose and shape from a single color image. 2019

165 Varol G, Ceylan D, Russell B, Yang J M, Yumer E, Laptev I, Schmid C. BodyNet: volumetric inference of 3D human body shapes//Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018, 20–38
DOI:10.1007/978-3-030-01234-2_2

166 Güler R A, Iasonas K. HoloPose: Holistic 3D Human Reconstruction In-The-Wild Task-Specific Decoders. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019

167 Zheng Z R, Yu T, Wei Y X, Dai Q H, Liu Y B. DeepHuman: 3D human reconstruction from a single image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), IEEE, 2019, 7738–7748
DOI:10.1109/iccv.2019.00783

168 Kanazawa A, Zhang J Y, Felsen P, Malik J. Learning 3D human dynamics from video. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019, 5607–5616
DOI:10.1109/cvpr.2019.00576

169 Zhu H, Zuo X X, Wang S, Cao X, Yang R G. Detailed human shape estimation from a single image by hierarchical mesh deformation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019, 4486–4495
DOI:10.1109/cvpr.2019.00462

170 Alldieck T, Magnor M, Bhatnagar B L, Theobalt C, Pons-Moll G. Learning to reconstruct people in clothing from a single RGB camera. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019, 1175‒1186
DOI:10.1109/cvpr.2019.00127

171 Alldieck T, Pons-Moll G, Theobalt C, Magnor M. Tex2Shape: detailed full human body geometry from a single image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), IEEE, 2019, 2293‒2303
DOI:10.1109/iccv.2019.00238

172 Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W P, Theobalt C. Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 2017 International Conference on 3D Vision (3DV). Qingdao, IEEE, 2017, 506‒516
DOI:10.1109/3dv.2017.00064

173 Mehta D, Sotnychenko O, Mueller F, Xu W P, Sridhar S, Pons-Moll G, Theobalt C. Single-shot multi-person 3D pose estimation from monocular RGB. In: 2018 International Conference on 3D Vision (3DV). Verona, IEEE, 2018, 120‒130
DOI:10.1109/3dv.2018.00024

174 Huang Y H, Kaufmann M, Aksan E, Black M J, Hilliges O, Pons-Moll G. Deep inertial poser. ACM Transactions on Graphics, 2019, 37(6): 1‒15
DOI:10.1145/3272127.3275108

175 Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2878‒2890
DOI:10.1109/tpami.2012.261

176 Moon G, Chang J Y, Lee K M. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), IEEE, 2019, 10132‒10141
DOI:10.1109/iccv.2019.01023

177 Güler R A, Iasonas K. HoloPose: Holistic 3D Human Reconstruction In-The-Wild Task-Specific Decoders. IEEE Conference on Computer Vision and Pattern Recognition, 2019, 10884‒10894

178 Cao Z, Simon T, Wei shih-en, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, IEEE, 2017, 1302‒1310
DOI:10.1109/cvpr.2017.143

179 Monszpart A, Guerrero P, Ceylan D, Yumer E, Mitra N J. iMapper. ACM Transactions on Graphics, 2019, 38(4): 1‒15
DOI:10.1145/3306346.3322961

180 Hassan M, Choutas V, Tzionas D, Black M. Resolving 3D human pose ambiguities with 3D scene constraints. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), IEEE, 2019, 2282‒2292
DOI:10.1109/iccv.2019.00237