3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates

Nikolaos Sarafianos[a], Bogdan Boteanu[b], Bogdan Ionescu[b], Ioannis A. Kakadiaris[a]

[a]*Computational Biomedicine Lab, Department of Computer Science*
University of Houston, 4800 Calhoun Rd. Houston, TX 77004
[b]*Image Processing and Analysis Lab, University Politehnica of Bucharest, 61071 Romania*

Abstract

Estimating the pose of a human in 3D given an image or a video has recently received significant attention from the scientific community. The main reasons for this trend are the ever increasing new range of applications (e.g., human-robot interaction, gaming, sports performance analysis) which are driven by current technological advances. Although recent approaches have dealt with several challenges and have reported remarkable results, 3D pose estimation remains a largely unsolved problem because real-life applications impose several challenges which are not fully addressed by existing methods. For example, estimating the 3D pose of multiple people in an outdoor environment remains a largely unsolved problem. In this paper, we review the recent advances in 3D human pose estimation from RGB images or image sequences. We propose a taxonomy of the approaches based on the input (e.g., single image or video, monocular or multi-view) and in each case we categorize the methods according to their key characteristics. To provide an overview of the current capabilities, we conducted an extensive experimental evaluation of state-of-the-art approaches in a synthetic dataset created specifically for this task, which along with its ground truth is made publicly available for research purposes. Finally, we provide an in-depth discussion of the insights obtained from reviewing the literature and the results of our experiments. Future directions and challenges are identified.

Keywords: 3D Human Pose Estimation, Articulated Tracking, Anthropometry, Human Motion Analysis

1. Introduction

Articulated pose and motion estimation is the task that employs computer vision techniques to estimate the configuration of the human body in a given image or a sequence of images. This is an important task in computer vision, being used in a broad range of scientific and consumer domains, a sample of which are: (i) Human-Computer Interaction (HCI): Human motion can provide natural computer interfaces whereby computers can be controlled by human gestures or can recognize sign languages [1, 2]; (ii) Human-Robot Interaction: Today's robots must operate closely with humans. In household environments, and especially in assisted living situations, a domestic service robot should be able to perceive the human body pose to interact more effectively [3, 4]; (iii) Video Surveillance: In video-based smart surveillance systems, human motion can convey the action of a human subject in a scene. Since manual monitoring of all the data acquired is impossible, a system can assist security personnel to focus their attention on the events of interest [5, 6]; (iv) Gaming: The release of the Microsoft Kinect sensor [7, 8] along with toolkit extensions that facilitate the integration of full-body control with games and Virtual Reality applications [9] are the most illustrative examples of how human motion capture can be used in the gaming industry; (v) Sport Performance Analysis: In most sports, the movements of the athletes are studied in great depth from multiple views and, as a result, accurate pose estimation systems can help in analyzing these actions [10, 11, 12]; (vi) Scene Understanding: Estimating the 3D human pose can be used in a human-centric scene understanding setup to help in the prediction of the "workspace" of a human in an indoor scene [13, 14]; (vii) Proxemics Recognition: Proxemics recognition refers to the task of understanding how people interact. It can be combined with robust pose estimation techniques to directly decide whether and to what extent there is an interaction between people in an

Email address: ioannisk@uh.edu (Ioannis A. Kakadiaris)

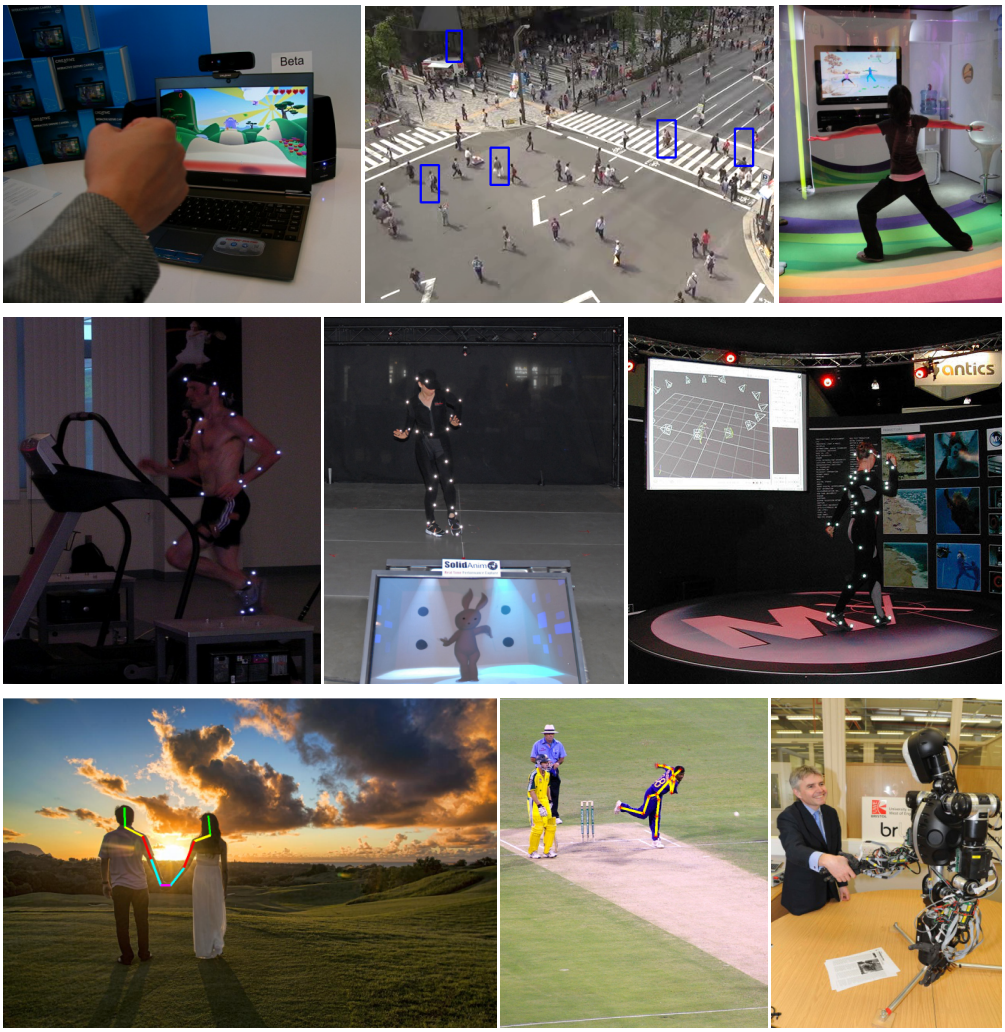


Figure 1: A summary of real-life applications of human motion analysis and pose estimation (images from left to right and top to bottom): Human-Computer Interaction, Video Surveillance, Gaming, Physiotherapy, Movies, Dancing, Proxemics, Sports, Human-Robot Interaction. Flickr image credits: The Conmunity - Pop Culture Geek, Intel Free Press, Patrick Oscar Boykin, Rae Allen, Christopher Prentiss Michel, Yuiseki Aoba, DIUS Corporate, Dalbra J.P., and Grook Da Oger.

image [15] and at the same time improves the pose estimation accuracy since it addresses occlusions between body parts; (viii) Estimating the anthropometry of a human from a single image [16, 17, 18]; (ix) 3D Avatar creation [19, 20] or controlling a 3D Avatar in games [21]; (x) Understanding the camera wearer's activity in an egocentric vision scenario [22]; and (xi) Describing clothes in images [23, 24] which can then be used to improve the pose identification accuracy.

In Figure 1 some of the aforementioned applications are depicted, which along with recent technological advances, and the release of new datasets have resulted in an increasing attention of the scientific community on the field. However, human pose estimation still remains an open problem with several challenges, especially in the 3D space.

Figure 2 shows the number of publications with the keywords: (i) "3D human pose estimation", (ii) "3D motion tracking", (iii) "3D pose recovery", and (iv) "3D pose tracking" in their title after duplicate and not relevant results are discarded. Note that, there are other keywords that return relevant publications such as "3D human pose recovery" [5] or "3D human motion tracking" [25]. Thus, Figure 2 does not cover all the methods we discuss but, even restricted to this particular search, still shows the increase of interest by the scientific community.

To cover the recent advances in the field and at the

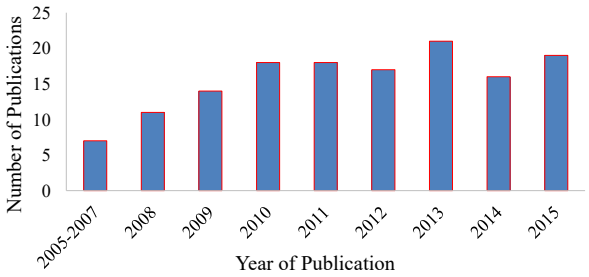


Figure 2: Depiction of the number of papers published during the last decade that include the keywords "3D human pose estimation", "3D motion tracking", "3D pose recovery", and "3D pose tracking" in their title after duplicate and irrelevant results are discarded.[1]

same time to be effective in our approach, we narrowed this survey to a class of techniques which are currently the most popular, namely the 3D human body pose estimation from RGB images. Apart from using RGB data, another major class of methods, which have received a lot of attention lately, are the ones using depth information such as RGB-D. Although an increasing number of papers has been published on this topic during the last few years with remarkable results [7], 3D Pose Estimation from RGB-D images will not be covered in this work because Helten *et al.* [26] and Ye *et al.* [27] published surveys on this topic recently which cover in detail the recent advances and trends in the field.

1.1. Previous Surveys and Other Resources

The reader is encouraged to refer to the early works of Aggarwal and Cai [28] and Gavrila [29] to obtain an overview of the initial methods in the field. The most recent surveys on human pose estimation by Moeslund *et al.* [30] and Poppe [31], date back to 2006 and 2007, respectively, and since they cover in great breadth and depth the whole vision-based human motion capture domain, they are highly recommended. However, they do not focus specifically on the 3D human pose estimation and are now outdated. Other existing reviews, focus on more specific tasks. For instance, a review on view-invariant pose representation and estimation is offered by Ji and Liu [32]. In the work of Sminchisescu [33], an overview of the problem of reconstructing 3D human motion from monocular image sequences is provided, whereas Holte *et al.* [34] present a 3D human pose estimation review, which covers only model-based methods in multi-view settings.

The primary goal of our review is to summarize the recent advances of the 3D pose estimation task. We conducted a systematic research of single-view approaches published in the 2008-2015 time frame. For multi-view scenarios, we focused on methods either published after the work of Holte *et al.* [34] or published before, but not discussed in their work. The selected time frames ensure that all approaches discussed in this survey are not referenced in previous reviews. However, for an incipient overview of this field, the reader is encouraged to refer to the publications of Sigal *et al.* [35, 36] where, inspired by the introduction of the HumanEva dataset, they present some aspects of the image- and video-based human pose and motion estimation tasks. In the recent work of Sigal [37], the interested reader can find a well-structured overview of the articulated pose estimation problem. Finally, Moeslund *et al.* [38] offer an illustrative introduction to the problem and provide a detailed analysis and overview of different human pose estimation approaches.

1.2. Taxonomy and Scope of this Survey

Figure 3 presents the pool of steps which apply to most 3D human pose estimation systems and illustrates all the stages covered in this review. Three-dimensional pose estimation methods include some of the action steps shown which are: (i) the use of a priori body model which determines if the approach will be model-based or model-free, (ii) the utilization of 2D pose information which can be used not only as an additional source of information but also as a way to measure the accuracy by projecting the estimated 3D pose to the 2D image and comparing the error, (iii) the use of pre-processing techniques, such as background subtraction,

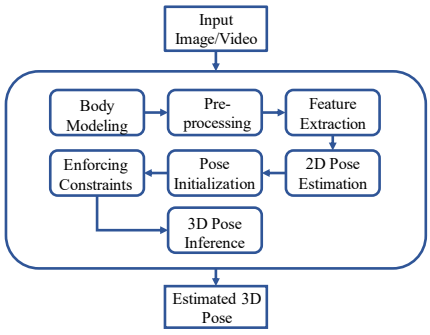


Figure 3: Pool of the stages of a common 3D human pose estimation system. Given an input signal the 3D pose is estimated by employing some or even all of the depicted steps.

[1] Results of the search on May 1st, 2016. We excluded searches related to patents or articles which other scholarly articles have referred to, but which cannot be found online.

(iv) feature extraction/selection approaches that obtain key features from the human subject which are fed to the estimation algorithms, (v) the process of obtaining an initial 3D pose which is used thereafter by optimization techniques that are employed to estimate the 3D pose and (vi) the pose estimation approach proposed each time that often is discussed along with constraints that are enforced to discard anthropometrically unrealistic poses, and finally how the final pose is inferred. A more specific categorization of the approaches wouldn't be practical since different approaches follow different paths according to the problem they are trying to address.

Despite the increasing interest from the scientific community, a well-structured taxonomy for the 3D human pose estimation task has not been proposed. To group approaches with similar key characteristics, we categorized the problem based on the input signal. We investigate articulated 3D pose and motion estimation when the input is a single image or a sequence of RGB frames. In the latter case approaches focus on capturing how the 3D human pose changes over time from an image sequence. A noteworthy amount of publications address the articulated 3D human pose estimation problem in multi-view scenarios. Since these approaches overcome some difficulties, while at the same time introducing new challenges to the pose estimation task, they are discussed separately in each case.

Similar to the aforementioned surveys and resources, we approach the pose estimation methods focusing on how they interpret the structure of the body: generative (model-based), discriminative (model-free), part-based which is a subcategory of generative models, and finally hybrid approaches. The taxonomy of 3D Pose Estimation methods is depicted in Figure 4.

Generative model approaches (also referred to as model-based or top-down approaches) employ a known model based on a priori information such as specific motion [39] and context [40]. The pose recovery process comprises two distinct parts, the modeling and the estimation [41]. In the first stage, a likelihood function is constructed by considering all the aspects of the problem such as the image descriptors, the structure of the human body model, the camera model and also the constraints being introduced. For the estimation part, the most likely hidden poses are predicted based on image observations and the likelihood function.

Another category of generative approaches found in the literature is *part-based* (also referred to as bottom-up approaches), which follows a different path by representing the human skeleton as a collection of body parts connected by constraints imposed by the joints within the skeleton structure. The Pictorial Structure Model (PSM) is the most illustrative example of part-based models. It has been mainly used for 2D human pose estimation [42, 43, 44] and has lately been extended for 3D pose estimation [45, 46]. It represents the human body as a collection of parts arranged in a deformable configuration. It is a powerful body model which results in an efficient inference of the respective parts. An extension of the PSM is the Deformable Structures model proposed by Zuffi *et al.* [47], which replaces the rigid part templates with deformable parts to capture body shape deformations and to model the boundaries of the parts more accurately. A graphical model which captures and fits a wide range of human body shapes in different poses is proposed by Zuffi and Black [48]. It is called Stitched Puppet (SP) and is a realistic part-based model in which each body part is represented by a mean shape. Two subspaces of shape deformations are learned using principal component analysis (PCA), independently accounting for variations in intrinsic body shape and pose-dependent shape deformations.

Discriminative approaches (also referred to as model-free) do not assume a particular model since they learn a mapping between image or depth observations and 3D human body poses. They can be further classified into learning-based and example-based approaches. Learning-based approaches learn a mapping function from image observations to the pose space, which must generalize well for a new image from the testing set [49, 50]. In example-based approaches, a set of exemplars with their corresponding pose descriptors is stored and the final pose is estimated by interpolating the candidates obtained from a similarity search [49, 51]. Such methods benefit in robustness and speed from the fact that the set of feasible human body poses is smaller than the set of anatomically possible ones [52]. The main advantage of generative methods is their ability to infer poses with better precision since they generalize well and can handle complex human body configurations with clothing and accessories. Discriminative approaches have the advantage in execution time because the employed models have fewer dimensions. According to Sigal and Black [35], the performance of discriminative methods depends less on the feature set or the inference method than it does for generative approaches.

Additionally, there are *hybrid approaches*, in which discriminative and generative approaches are combined to predict the pose more accurately. To combine these two methods, the observation likelihood obtained from a generative model is used to verify the pose hypotheses obtained from the discriminative mapping functions for pose estimation [53, 54]. For example, Salz-

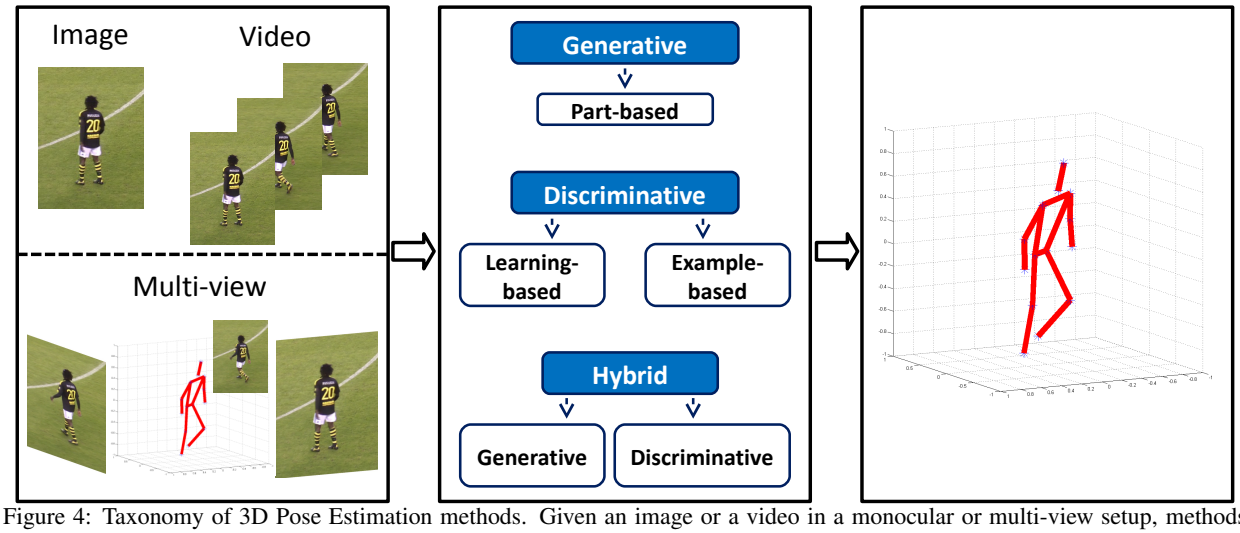


Figure 4: Taxonomy of 3D Pose Estimation methods. Given an image or a video in a monocular or multi-view setup, methods can be classified as generative (a subcategory of which are part-based approaches), discriminative (which can be classified into learning-based and example-based) and finally hybrid which are a combination of the previous two.

mann and Urtasun [55] introduced a unified framework that combines model-free and model-based approaches by introducing distance constraints into the discriminative methods and employing generative methods to enforce constraints between the output dimensions. An interesting discussion on generative and discriminative approaches can be found in the work of Bishop and Lasserre [56].

In the following, we present a detailed analysis of 3D pose estimation techniques in different setups. The rest of the paper is organized as follows. In Section 2, we discuss the main aspects of the body model employed by model-based methods and the most common features and descriptors used. In Section 3, we present the proposed taxonomy by discussing the key aspects of pose estimation approaches from a single image. Section 4 presents the recent advances and trends in 3D human pose estimation from a sequence of images. In both sections, we discuss separately single- and multi-view input approaches. In Section 5, we discuss some of the available datasets, summarize the evaluation measures found in the literature, and offer a summary of performance of several methods on the HumanEva dataset. Section 6 introduces a new synthetic dataset in which humans with different anthropometric measurements perform actions. An evaluation of the performance of state-of-the-art 3D pose estimation approaches is also provided. We conclude this survey in Section 7 with a discussion of promising directions for future research.

2. Human Body Model and Feature Representation

The human body is a very complex system composed of many limbs and joints and a realistic estimation of the position of the joints in 3D is a challenging task even for humans. Marinoiu *et al.* [57] investigated how humans perceive the pictorial 3D pose space, and how this perception can be connected with the regular 3D space we move in. Towards this direction, they created a dataset which, in addition to 2D and 3D poses, contains synchronized eye movement recordings of human subjects shown a variety of human body configurations and measured how accurately humans re-create 3D poses. They found that people are not significantly better at re-enacting 3D poses in laboratory environments given visual stimuli, on average, than existing computer vision algorithms.

Despite these challenges, automated techniques provide valuable alternatives for solving this task. Model-based approaches employ a human body model which introduces prior information to overcome this difficulty. The most common 3D human body models in the literature are the skeleton (or stick figure), a common representation of which is shown in Figure 5 along with its structure, and shape models. They both define kinematic properties, whereas the shape models also define appearance characteristics. The cylindrical and the truncated cone body models are illustrative examples of shape models. After constructing the body model, constraints are usually enforced to constrain the pose parameters. Kinematic constraints, for example, ensure that limb lengths, limb-length proportions, and

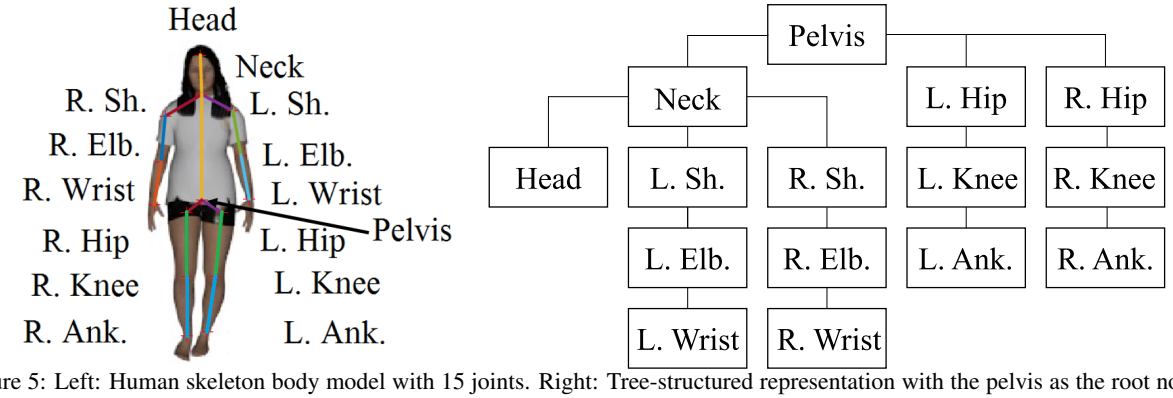


Figure 5: Left: Human skeleton body model with 15 joints. Right: Tree-structured representation with the pelvis as the root node (Sh. - Shoulder, Elb. - Elbow and Ank. - Ankle).

joint angles follow certain rules. Other popular constraints found in the literature are occlusion constraints that allow more realistic poses in which some body parts (legs or arms) are occluded by others and prevent double-counting phenomena, appearance constraints introduced by the symmetry of left and right body part appearances [58], and smoothness constraints in the angle of the joints which are used to avoid abrupt changes between sequential video frames.

Whether a body model is employed or not (model-based or model-free approaches), the next action step in the study of 3D human motion, is the accurate feature extraction from the input signal. Early approaches in the field used low-level features such as edges, color, optical flow or silhouettes which are obtained after performing background subtraction. Silhouettes are invariant to texture and lighting but require good segmentation of the subject, and can easily lose specific details of human parts. Image descriptors are then employed to describe these features and to reduce the size of the feature space. Common feature representations employed in the literature include the use of Scale Invariant Feature Transforms (SIFT) [59], Shape Context (SC) [60] and Appearance and Position Context (APC) descriptors [40]. APC is a sparse and local image descriptor, which captures the spatial co-occurrence and context information of the local structure as well as their relative spatial positions. Histograms of Oriented Gradients (HoG) have been used a lot lately [61, 62], because they perform well when dealing with clutter and can capture the most discriminative information from the image. Instead of extracting features from the image, some approaches [5, 49] select the most discriminative features. Moll *et al.* [63] proposed *posebits* which are semantic pose descriptors which represent geometrical relationships between body parts and can take binary values depending on the answer to simple questions such as "Left foot in front of the torso". Posebits can provide sufficient 3D pose information without requiring 3D annotation, which is a difficult task, and can resolve depth ambiguities.

3. Recovering 3D human pose from a single image

The reconstruction of an arbitrary configuration of 3D points from a single monocular RGB image has three characteristics that affect its performance: (i) it is a severely ill-posed problem because similar image projections can be derived from different 3D poses; (ii) it is an ill-conditioned problem since minor errors in the locations of the 2D body joints can have large consequences in the 3D space; and (iii) it suffers from high dimensionality [64]. Existing approaches propose different solutions to compensate for these constraints and are discussed in Section 3.1.

3.1. Three-dimensional Human Pose Estimation from a Single Monocular Image

The recovery of 3D human poses in monocular images is a difficult task in computer vision since highly nonlinear human motions, pose and appearance variance, cluttered backgrounds, occlusions (both from other people or objects and self-occlusions), and the ambiguity between 2D and 3D poses are common phenomena. The papers described in this category estimate the human pose explicitly from a single monocular image and are summarized in Table 1. Publications that fit into both the single image and the video categories are discussed in Section 4.

Deep-Learning Methods: Deep-learning methods are representation-learning approaches [83] composed of

Table 1: 3D human pose estimation from a single monocular RGB image. Wherever a second reference is provided, it denotes the availability of source code for the method. The Body Model column indicates whether a body model is employed. The Method Highlights column reflects the most important steps in each approach.

Year	First Author	Body Model	Method Highlights	Evaluation Datasets	Evaluation Metrics
2016	Yasin [65, 66]	Yes	Training: 3D poses are projected to 2D and a regression model is learned from the 2D annotations; Testing: 2D pose is estimated, the nearest 3D poses are predicted; final 3D pose is obtained by minimizing the projection error	HumanEva-I, Human3.6M	3D pose
2015	Li [67]	No	The input is an image and a potential 3D pose and the output a score matching value; ConvNet for image feature extraction; Two sub-networks for transforming features and pose into a joint embedding	Human3.6M	MPJPE
2014	Kostrikov [68]	Yes	Predict the relative 3D joint position using depth sweep regression forests trained with three groups of features; 3DPS model for inference	Human3.6M, HumanEva-I	3D 3D pose
2014	Li [69]	No	Train a deep ConvNet; and joint point regression to estimate the positions of joint points relative to the root position and joint point detection to classify whether one local window contains the specific joint	Human3.6M	MPJPE
2014	Wang [70, 71]	Yes	2D part detector and a sparse basis representation in an overcomplete dictionary; Anthropometric constraints are enforced and an L_1-norm projection error metric is used; Optimization with ADMM	HumanEva-I, CMU MoCap, UVA 3D	3D pose
2014	Zhou [72, 73]	Yes	Convex formulation by using the convex relaxation of the orthogonality constraint; ADMM for optimization	CMU MoCap	3D
2013	Radwan [74]	Yes	Employ a 2D part detector with an occlusion detection step; Create multiple views synthetically with a twin-GPR in a cascaded manner; Kinematic and orientation constraints to resolve remaining ambiguities	HumanEva-I, CMU MoCap	3D pose
2013	Simo-Serra [75]	Yes	Bayesian approach using a model with discriminative 2D part detectors and a probabilistic generative model based on latent variables; Inference using the CMA-ES	HumanEva-I, TUD Stadmitte	3D 3D pose
2012	Brauer [76]	Yes	ISM to obtain vote distributions for the 2D joints; Example-based 3D prior modeling and comparison of their projections with the respective joint votes	UMPM	MJAE, Orientation Angle
2012	Ramakrishna [77, 78]	Yes	Enforce anthropometric constraints and estimate the parameters of sparse linear representation in an overcomplete dictionary with a matching pursuit algorithm	CMU MoCap	3D
2012	Simo-Serra [79]	Yes	2D part detector and a stochastic sampling to explore each part region; Set of hypotheses enforces reprojection and length constraints; OCSVM to find the best sample	HumanEva-I, TUD Stadmitte	3D 3D pose
2011	Greif [80]	No	Train an action-specific classifier on improved HoG features; use a people detector algorithm and treat 3D pose estimation as a classification problem	HumanEva-I	3D
2009	Guo [81]	No	Pose tree is learned by hierarchical clustering; Multi-class classifiers are learned and the relevance vector machine regressors at each leaf node estimate the final 3D pose	HumanEva-I	3D
2009	Huang [49, 82]	No	Occluded test images as a sparse linear combination of training images; Pose-dependent (HoG) feature selection and L_1-norm minimization to find the sparest solution	HumanEva-I, Synthetic	3D, MJAE
2008	Ning [40]	No	Employ an APC descriptor and learn in a jointly supervised manner the visual words and the pose estimators	HumanEva-I, Quasi-synthetic	3D, MJAE

multiple non-linear transformations. Feature hierarchies are learned with features from higher and more abstract levels of the hierarchy formed by the composition of lower level features [84, 85]. Depending on the method used and how the architecture is set-up, it finds applications in both unsupervised and supervised learning as well as hybrid approaches [86]. After its early introduction by Hinton *et al.* [87, 88], employing deep architectures, is found to yield significantly better results in many computer vision tasks such as object recognition, image classification and face verification [89, 90, 91]. Following that, approaches which employ deep-learning techniques to address the 2D pose estimation task with great success, have been proposed [92, 93, 94, 95] and only recently the 3D pose estimation task was approached using deep learning. In the work of Li and Chan [69], deep convolutional networks (ConvNets) are trained for two distinct approaches: (i) they jointly train the pose regression task with a set of detection tasks in a heterogeneous multi-task learning framework and (ii) pre-train the network using the detection tasks, and then refine the network using the pose regression task alone. They show that the network in its last layers has an internal representation for the positions of the left (or right) side of the person, and thus, has learned the structure of the skeleton and the correlation between output variables. Li *et al.* [67] proposed a framework which takes as an input an image and a 3D pose and produces a score value that represents a multi-view similarity between the two inputs (i.e., whether they depict the same pose). A ConvNet for feature extraction is employed and two sub-networks are used to perform a non-linear transformation of the image and pose into a joint embedding. A maximum-margin cost function is used during training which enforces a re-scaling margin between the score values of the ground truth image-pose pair and the rest image-pose pairs. The score function is the dot-product between the two embeddings. However, the lack of training data for ConvNet-based techniques remains a significant challenge. Towards this direction, the methods of Chen *et al.* [96] and Rogez and Schmid [97] propose techniques to synthesize training images with ground truth pose annotations. Finally, the task of estimating the 3D human pose from image sequences has also been explored using deep learning, [98, 99, 100, 101, 102, 103] and the respective methods are going to be discussed individually in Sections 4.1 and 4.2.

Two-dimensional detectors for 3D pose estimation: To overcome the difficulty and the cost of acquiring images of humans along with their respective 3D poses, Yasin *et al.* [65] proposed a dual-source approach which employs images with their annotated 2D poses and 3D motion capture data to estimate the pose of a new test image in 3D. During training, 3D poses are projected to a 2D space and the projection is estimated from the annotated 2D pose of the image data through a regression model. At testing time, the 2D pose of the new image is first estimated from which the most likely 3D poses are retrieved. By minimizing the projection error the final 3D pose is obtained. Aiming to perform 3D human pose estimation from noisy observations, Simo-Serra *et al.* [79] proposed a stochastic sampling method. As a first step, they employ a state-of-the-art 2D body part detector [61] and then convert the bounding boxes of the parts to a Gaussian distribution by computing the covariance matrix of the classification scores within each bounding box. To obtain a set of ambiguous candidate poses from the samples generated in the 3D space by the Gaussian distribution, they use the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to simultaneously minimize re-projection and length errors. The most anthropometric pose between the candidates is determined by using a One-Class Support Vector Machine (OCSVM). To exploit the advantages of both generative and discriminative approaches, Simo-Serra *et al.* [75] proposed a hybrid Bayesian approach. Their method comprises 2D HoG-based discriminative part detectors which constrain the 2D location of the body parts and a probabilistic generative latent variable model which (i) maps points from the high dimensional 3D space to the lower dimensional latent space, (ii) specifies the dependencies between the latent states, (iii) enforces anthropometric constraints, and (iv) prevents double counting. To infer the final 3D pose they use a variation of CMA-ES. Brauer *et al.* [76] employ a slightly modified Implicit Shape Model (ISM) to generate vote distributions for potential 2D joint locations. Using a Bayesian formulation, 3D and 2D poses are estimated by modeling (i) the pose prior following an example-based approach and (ii) the likelihood by comparing the projected joint locations of the exemplar poses with the corresponding nearby votes.

Discussion of Norms and Camera Parameter Estimation: To resolve the ambiguities that arise when performing pose estimation from a single image, some methods also estimate the relative pose of the camera. The approaches of Ramakrishna *et al.* [77] and Wang *et al.* [70] belong to this category. Both methods require the locations of the joints in the 2D space as an input, use a sparse basis model representation, and employ an optimization scheme which alternatively esti-

mates the 3D pose estimation and the camera parameters. In the first case, the authors constrain the sum of the limb lengths and use a matching pursuit algorithm to perform reconstruction. Their method can also recover the 3D pose of multiple people in the same view. In the latter case, L_1-norm is used as a reprojection error metric that is more robust when the joint locations in 2D are inaccurate. This approach also enforces not only limb length constraints, which eliminate implausible poses, but also L_1-norm constraints on the basis coefficients. A discussion on why L_2-norms are insufficient for estimating 3D pose similarity is provided by Chen *et al.* [104]. However, Zhou *et al.* [72] argue that the solution to such alternating minimization approaches is sensitive to initialization. Using the 2D image landmarks as an input, they used an augmented shape-space model to give a linear representation of both intrinsic shape deformation and extrinsic viewpoint changes. They proposed a convex formulation that guarantees global optimality and solved the optimization problem with a novel algorithm based on the Alternating Direction Method of Multipliers (ADMM) and the proximal operator of the spectral norm. Their method is applicable not only to human pose but also to car and face reconstruction. An approach which also uses a sparse image representation and solves a convex optimization problem with the L_1-norm is proposed by Huang and Yang [49]. Aiming to estimate 3D human pose when humans are occluded, they proposed a method which exploits the advantages of both example-based and learning-based approaches and represents each test sample as a sparse linear combination of training samples. The background clutter in the test sample is replaced with backgrounds from the training images which results in pose-dependent feature selection. They use a Gaussian process regressor to learn the mapping between the image features (HoG from original or corrupted images and recovered features) and the corresponding 3D parameters. They observed that when a sparse linear representation of the training images is used for the probes, the set of coefficients from the corrupted (i.e., occluded) test image is recovered with minimum error via solving an L_1-norm minimization problem.

Discriminative Approaches: Ning *et al.* [40] proposed a discriminative bag of words approach. As a first step, they utilize an APC descriptor, and learn in a supervised manner a separate metric for each visual word from the labeled image-to-pose pairs. They use a Bayesian Mixture of Experts (BME) model to represent the multimodal distribution of the 3D human pose conditioned on the feature space and also a gradient ascent algorithm which jointly optimizes the metric learning and the BME model. Kostrikov and Gall [68] approached the pose estimation task from a different perspective, and proposed a discriminative depth sweep forest regression approach. After extracting features from 2D patches sampled from different depths, the proposed method sweeps with a plane through the 3D volume of potential joint locations and uses a regression forest that learns 2D-2D or 3D-3D mappings from the relative feature locations. Thus, they predict the relative 3D position of a joint, given the hypothesized depth of the feature. Finally, the pose space is constrained by employing a 3D pictorial structure model used to infer the final pose. Okada and Soatto [105] introduced a method comprising three main parts that estimates the 3D pose in clutter backgrounds. Given a test image with a window circumscribing a specific subject, (i) they extract a HoG-based feature vector of the window; (ii) they use a Support Vector Machine (SVM) classifier that selects the pose cluster which the current pose belongs to; and (iii) having taken into consideration that the relevance of features selected depends on the pose, they recover the 3D pose using a piecewise linear regressor of the selected cluster.

Both Guo and Patras [81] and Jiang [106] proposed exemplar-based approaches. In the first approach a tree is learned by hierarchical clustering on pose manifold via affinity propagation and the final 3D pose is estimated by applying the learned relevance vector machine regressor that is attached to the leaf node to which the example is classified. In the second method, the 3D pose is reconstructed by using a k-dimensional tree (kd-tree) to search in a database containing millions of exemplars of optimal poses for the optimal upper body and lower body pose. Another interesting approach is proposed by Urtasun and Darrell [107]. They developed an online activity-independent method to learn a complex appearance-to-pose mapping in large training sets using probabilistic regression. They use a (consistent in pose space) sparse Gaussian process model which: (i) forms local models (experts) for each test point, (ii) handles the mapping inaccuracy caused by multimodal outputs, and (iii) performs fast inference. The local regressors at each test point overcome the boundary problems that occur in offline (clustering) approaches. Finally, Greif *et al.* [80] treat pose estimation as a classification problem. They consider the full body pose as a combination of a 3D pose, and a viewpoint, and define classes that are then learned by an action specific forest classifier. The input of the classification process are lower-dimensional improved HoG [108] features. The proposed method does not require labeled viewpoints

and background subtracted images, and the action performed by the subject does not need to be cyclic.

The approach of Radwan *et al.* [74] differentiates itself from the rest since it performs pose estimation from a single image by utilizing information from multiple synthetically created views. First, they employ the 2D part detector of Yang and Ramanan [61] to which they add an extra occlusion detection step to overcome self-occlusions. Then, they use the twin Gaussian Process Regression (GPR) in a cascaded manner to generate synthetic views from different viewpoints, and finally impose kinematic and orientation constraints on the 3D ambiguous pose resulting from the projection of a 3D model onto the initial pose.

3.2. Three-dimensional Human Pose Estimation from a Single image in a Multiple Camera View Scenario

Resolving the ambiguities that arise in the 3D space would be a much easier task if depth information obtained from a sensor such as the Microsoft Kinect [7] was used. However, Kinect has a specific range within which it operates successfully and it cannot be used for outdoor applications. Two approaches that use multiple view images to overcome these difficulties and to construct a more realistically applicable pose estimation system are presented in Table 2 and described below.

Burenius *et al.* [45] implemented a framework for 3D pictorial structures for multi-view articulated pose estimation. First, they compute the probability distribution for the position of body parts with 2D part detectors based on HoG features. The parts are connected in a tree graph and the dependency structure of the variables of the model is represented through a Bayesian network. A weak pose prior (translation and rotation) is imposed to the pose and dynamic programming is used to discretize the state space. For the translation prior, they use a max-product algorithm with two variations according to the constraints imposed. A two-step algorithm is finally employed to deal with the double counting phenomenon which is a typical problem in tree structures. An approach which employs a 2D pictorial structure model in multi-view scenarios is proposed by Amin *et al.* [60]. Instead of using a 3D body model, they infer the 3D pose over a set of 2D projections of the 3D pose in each camera view. The 2D pictorial structures model is extended with flexible parts, color features, multi-modal pairwise terms, and mixtures of pictorial structures. Appearance and spatial correspondence constraints across views are enforced to take advantage of the multi-view setting. The final 3D pose is recovered from the 2D projection by triangulation.

4. Recovering 3D Human Pose from a Sequence of Images

Besides high dimensionality, which is always an issue, a difficulty that arises when trying to locate the 3D position of the body joints from a sequence of images is that the shape and appearance of the human body may change drastically over time due to: (i) background changes or camera movement, especially outside of controlled laboratory settings; (ii) illumination changes; (iii) rotations in-depth of limbs; and (iv) loosely fitting clothing.

4.1. Three-dimensional Human Pose Estimation from a Sequence of Monocular Images

Most of the video data used nowadays are captured from a single camera view. Even in multi-view scenarios (e.g., surveillance systems) the person is not always visible from all the cameras at the same time. As a result, estimating the 3D pose of a human from monocular images is an important task. According to Sigal [37], accurate pose estimation on a per frame basis is an ill-posed problem and methods that exploit all available information over time [109, 110] can improve performance. The papers in this category focus on estimating the 3D human pose from a sequence of single-view images and are presented in Table 3.

Discriminative approaches: In the work of Tekin *et al.* [103] spatiotemporal information is exploited to reduce depth ambiguities. They employ 2 ConvNets to first align (i.e., shifting to compensate for the motion) the bounding boxes of the human in consecutive frames and then refine them so as to to create a data volume. 3D HoG descriptors are computed and the 3D pose is reconstructed directly from the volume with Kernel Ridge Regression (KRR) and Kernel Dependency Estimation (KDE). They demonstrated that (i) when information from multiple frames is exploited, challenging ambiguous poses where self-occlusion occurs can be estimated more accurately and (ii) the linking of detections in individual frames in an early stage, followed by enforcing temporal consistency at a later stage improves the performance significantly. ConvNets were also employed in a deep learning regression architecture work of Tekin *et al.* [102]. To encode dependencies between joint locations, an auto-encoder is trained on existing human poses to learn a structured latent representation of the human pose in 3D. Following that, a ConvNet architecture maps through a regression framework the input image to the latent representation and the decoding layer is then used to estimate the 3D pose from the latent to the original 3D space.

Table 2: 3D human pose estimation from a single RGB image in a multi-view setup. The Body Model column indicates whether a body model is employed. The Method Highlights column reflects the most important steps in each approach.

Year	First Author	Body Model	Method Highlights	Evaluation Datasets	Evaluation Metrics
2013	Amin [60]	Yes	Infer 3D pose over a set of 2D projections of the 3D pose in each camera view; Enforce appearance and spatial correspondence constraints across views; Recover final pose by triangulation	HumanEva-I, MPII Cooking	3D
2013	Burenius [45]	Yes	Use a tree graph (3DPS) to connect the parts extracted from 2D part detectors; Discretize state space and use the max-product algorithm with view, skeleton, and joint angle constraints	KTH Multiview Football II	3D PCP

An interesting approach from Yamada *et al.* [115] addresses the problem of dataset bias in discriminative 3D pose estimation models. Under covariate shift setup, a mapping is learned based on a weighted set of training image-pose pairs. The training instances are reweighted by the importance weight to remove the training set bias. They finally propose weighted variants of kernel regression and twin Gaussian processes to illustrate the efficacy of their approach. Chen *et al.* [5] proposed an example-based approach which focuses on the efficient selection of features by optimizing a trace-ratio criterion which measures the score of the selected feature component subset. During pose retrieval, a sparse representation is used which enforces a sparsity constraint that ensures that semantically similar poses have a larger probability to be retrieved. Using the selected pose candidates of each frame, a sequential optimization scheme is selected which employs dynamic programming to get a continuous pose sequence. Sedai *et al.* [50] introduced a learning-based method that exploits the complementary information of the shape (histogram of shape context) and appearance (histogram of local appearance context) features. They cluster the pose space into several modular regions and learn regressors for both feature types and their optimal fusion scenario in each region to exploit their complementary information [122].

Latent variable models: Latent variables are often used in the literature [123, 124], because it is often difficult to obtain accurate estimates of part labels because of possible occlusions. To alleviate the need for large labeled datasets, Tian *et al.* [114] proposed a discriminative approach that employs Latent Variable Models (LVMs) that successfully address over-fitting and poor generalization. Aiming to exploit the advantages of both Canonical Correlation Analysis (CCA) and Kernel Canonical Correlation Analysis (KCCA), they introduced a Canonical Local Preserving Latent Variable Model (CLP-LVM) that adds additional regularized terms that preserve local structure in the data. Latent spaces are jointly learned for both image features and 3D poses by maximizing the non-linear dependencies in the projected latent space while preserving local structure in the original space. To deal with multi-modalities in the data, they learned a multi-modal joint density model between the latent image features and the latent 3D poses in the form of Gaussian mixture regression which derives explicit conditional distributions for inference. A latent variable approach is also proposed by Andriluka *et al.* [110]. Their objective is to estimate the 3D human pose in real-world scenes with multiple people present where partial or full occlusions occur. They proposed a three-step hybrid generative/discriminative method using Bayesian formulation. They started by employing discriminative 2D part detectors to obtain the locations of the joints in the image. During the second stage, people tracklets are extracted using a 2D tracking-by-detection approach which exploits temporal coherency already in 2D, improves the robustness of the 2D pose estimation result, and enables early data association. In the third stage, the 3D pose is recovered through a hierarchical Gaussian process latent variable model (hGPLVM) which is combined with a Hidden Markov Model (HMM). Their method can track and estimate poses of a number of people that behave realistically in an outdoor environment where occlusions between individuals are common phenomena.

Ek *et al.* [125] introduced a method which also takes advantage of Gaussian process latent variable models (GPLVM). They represent each image by its silhouette, and model silhouette observations, joint angles and their dynamics as generative models from shared low-dimensional latent representations. To overcome the ambiguity that arises from multiple solutions, the latent space incorporates a set of Gaussian processes that give temporal predictions. Finally, by incorporating a

Table 3: 3D human pose estimation from a sequence of monocular RGB images. Wherever a second reference is provided, it denotes the availability of source code for the method.

Year	First Author	Body Model	Method Highlights	Evaluation Datasets	Evaluation Metrics
2016	Tekin [103]	No	Human detection in multiple frames and motion compensation to form a spatiotemporal volume with two ConvNets; 3D HoGs are employed and 3D pose is estimated with KRR and KDE regression	HumanEva-I&II, Human3.6M, KTH Multiview Football II	3D
2016	Zhou [98, 111]	Yes	If 2D joints are provided, a sparsity-driven 3D geometric prior and temporal smoothness model are employed; If not, 2D joints are treated as latent variables and a deep ConvNet is trained to predict the uncertainty maps; 3D pose estimation via an EM algorithm over the entire sequence	Human3.6M, CMU MoCap, PennAction	3D
2015	Hong [101]	Yes	A multimodal deep autoencoder is used to fuse multiple features by unified representations hidden in hypergraph manifold; Backpropagation NN learns a non-linear mapping from 2D silhouettes to 3D poses	HumanEva-I, Human3.6M	3D, MJAE
2015	Schick [112]	Yes	Discretize search space with supervoxels to reduce search space and apply them to 3DPS; Min-sum algorithm for pose inference	HumanEva-I, UMPM	3D 3D pose
2015	Wandt [113]	Yes	3D poses as a linear combination of base poses using PCA; Periodic weight is used to reduce the complexity and the camera parameters; 3D pose is estimated alternatively	KTH Multiview Football II, CMU MoCap, HumanEva-I	3D
2013	Sedai [53]	Yes	Hybrid approach for 3D pose tracking; Gaussian Process regression model for the discriminative part combined with the annealed particle filter to track the 3D pose	HumanEva-I & II	3D
2013	Tian [114]	No	Introduce a CLP-LVM to preserve local structure; Learn a multi-modal joint density model in the form of Gaussian Mixture Regression	CMU MoCap, Synthetic	3D
2012	Andriluka [109]	No	Exploit human context information towards 3D pose estimation; Estimate 2D Pose using a multi-aspect flexible pictorial structure model; Estimate 3D pose relying on a joint GPDM as a prior with respect to latent positions	Custom	3D
2012	Yamada [115]	No	Assume covariate shift setup and remove training set bias by re-weighting the training instances; Formulate two regression-based methods that utilize these weights	HumanEva-I, Synthetic	3D, MJAE
2011	Chen [5]	No	Visual feature selection and example-based pose retrieval via sparse representation; Sequential optimization via DP	HumanEva-I, Synthetic	Weighted 3D
2010	Andriluka [110]	Yes	Employ 2D discriminative part multi-view detectors and follow a tracking-by-detection approach to extract people tracklets; Pose recovery with hGPLVM and HMM	HumanEva-II, TUD Stadtmitte	3D
2010	Bo [116, 117]	No	Background subtraction, HoG extraction and TGP regression	HumanEva-I	3D
2010	Valmadre [118, 119]	Yes	Rigid constraints can be enforced only on sub-structures and not in the entire body; Estimation performed with a deterministic least-squares approach	CMU MoCap	Qualitative
2009	Rius [120]	Yes	Action-specific dynamic model discards non-feasible body postures; Work within a particle filtering framework to predict new body postures given the previously estimated ones	HumanEva-I, CMU MoCap	3D
2009	Wei [121]	Yes	Enforce independent rigid constraints in a number of frames and recover pose and camera parameters with a constrained nonlinear optimization algorithm	CMU MoCap	3D

back-constraint, they learn a parametric mapping from pose to latent space to enforce a one-to-one correspondence. Another interesting approach is provided by Andriluka and Sigal [109] who proposed a 3D pose estimation framework of people performing interacting activities such as dancing. The novelty of their approach lies in the fact that by taking advantage of the human-to-human context (interactions between the dancers) they estimate the poses more accurately. After detecting people over time in the videos and focusing on those who maintain close proximity, they estimate the pose of the two people in 2D by proposing a multi-person pictorial structure model that considers the human interaction. To estimate the 3D pose they use a joint Gaussian Process Dynamic Model (GPDM) as a prior, which captures the dependencies between the two people, and learn this model by minimizing the negative log of posterior with respect to the latent positions and the hyperparameters that they define. To perform inference of the final pose, a gradient-based continuous optimization algorithm is used. A pose prior for 3D human pose tracking which (i) has lower complexity than GPDM, (ii) can handle large amounts of data and, (iii) is consistent if geodesic distances are used instead of other metrics is proposed by Simo-Serra *et al.* [126].

Discussion of rigid constraints: Wei and Chai [121] proposed a 3D reconstruction algorithm that reconstructs 3D human poses and camera parameters from a few 2D point correspondences. Aiming to eliminate the reconstruction ambiguity, they enforced independent rigid constraints across a finite number of frames. A constrained nonlinear optimization algorithm is finally used to recover the 3D pose. Later, Valmadre and Lucey [118] explored the method of Wei and Chai [121] and contradicted some of its statements. They demonstrated that camera scales, bone lengths, and absolute depths cannot be estimated in a finite number of frames for a 17-bone body model and that rigid constraints can be enforced only on sub-structures and not in the entire body. They proposed a deterministic least-squares approach, which exploits the aforementioned results and estimates the rigid structure of the torso, the camera scale, the bone lengths and the joint angles. The method of Radwan *et al.* [74] that was mentioned in the previous section utilizes techniques of these two approaches and extends their findings by requiring only a single image.

Particle filter algorithm: The particle filter algorithm is effective for 3D human motion tracking [127], and the works of Sedai *et al.* [53] and Liu *et al.* [128] have provided some extensions that improve its accuracy. Liu *et al.* [128], proposed an exemplar-based conditional particle filter (EC-PF) in order to track the full-body human motion. For EC-PF, system state is constructed to be conditional to image data and exemplars in order to improve the prediction accuracy. The 3D pose is estimated in a monocular camera setup by employing shape context matching when the exemplar-based dynamic model is constructed. In the work of Sedai *et al.* [53], a hybrid approach is introduced that utilizes a mixture of Gaussian Process (GP) regression models for the discriminative part and a motion model with an observation likelihood model to estimate the pose using the particle filter. The discrete cosine transform of the silhouette features is used as a shape descriptor. GP regression models give a probabilistic estimation of the 3D human pose and the output pose distributions from the GP regression are combined with the annealed particle filter to track the 3D pose in each frame of the video sequence. Kinematic constraints are enforced and a 16-joint cylindrical model is employed. Promising results are reported on both single- and multiple-camera tracking scenarios.

Action-specific human body tracking: Works that belong to this category use a priori knowledge on movements of humans while performing an action. Jaeggli *et al.* [129] proposed a generative method combined with a learning-based statistical approach that simultaneously estimates the 2D bounding box coordinates, the performed activity, and the 3D body pose of a human. Their approach relies on strong models of prior knowledge about typical human motion patterns. They use a Locally Linear Embedding (LLE) on all poses in the dataset that belong to a certain activity to find an embedding of the pose manifolds of low dimensionality. The reduced space has mappings to both the original pose space and the appearance (image) space. The mapping from pose to appearance is performed with a Relevance Vector Machine kernel regressor and the min-sum algorithm is employed to extract the optimal sequence through the entire image sequence. Rius *et al.* [120] discuss two elements that can improve the accuracy of a human pose tracking system. First, they introduce an action-specific dynamic model of human motion which discards the body configurations that are dissimilar to the motion model. Then, given the 2D positions of a variable set of body joints, this model is used within a particle filtering framework in which particles are propagated based on their motion history and previously learned motion directions.

Finally, the interested reader is encouraged to refer to the publications of Sigal and Black [130], Bray *et al.* [131], and Agarwal and Triggs [64], all of which introduced seminal methods on the 3D pose estimation

problem from monocular images. Since all three are covered in previous surveys [30, 31] they will not be further analyzed in the present review.

4.2. Recovering 3D Human Pose from a Sequence of Multi-view Images

The publications discussed in this category are shown in Table 4. The approaches of Belagiannis *et al.* [46] and Sigal *et al.* [135] employed 3D human models to estimate the pose from a sequence of frames in multi-view scenarios. The method of Belagiannis *et al.* jointly estimates the 3D pose of multiple humans in multi-view scenarios. In such cases, more challenges arise, some of which are the unknown identity of the humans in different views and the possible occlusions either between individuals or self-occlusions. Similar to the method of Burenius *et al.* [45] the first obstacle that the authors wanted to overcome is the high dimensional complex state space. Instead of discretizing it, they used triangulation of the corresponding body joints sampled from the posteriors of 2D body part detectors in all pairs of camera views. The authors introduced a 3D pictorial structures (3DPS) model which infers the articulated pose of multiple humans from the reduced state space while at the same time resolving ambiguities that arise both from the multi-view scenario and the multiple human estimation. It is based on a Conditional Random Field (CRF) with multi-view potential functions and enforces rotation, translation (kinematic) and collision constraints. Finally, by sampling from the marginal distributions, the inference on the 3DPS model is performed using the loopy belief propagation algorithm. Belagiannis *et al.* [134] extended the previous method, by making the 3DPS model temporally consistent. In their method, they first recover the identity of each individual using tracking and afterwards infer the pose. Knowing the identity of each person results in a smaller state space that allows efficient inference. A temporal term for regularizing the solution is introduced which penalizes candidates that geometrically differ significantly from the temporal joint and ensures that the inferred poses are consistent over time. Furthermore, Belagiannis *et al.* [139] built upon their aforementioned works by using a 3DPS model under different body part parametrization. Instead of defining the body part in terms of 3D position and orientation they retained only the position parameters and implicitly encoded the orientation in the factor graph.

A 3DPS model is also employed by Amin *et al.* [133] to find an initial 3D pose estimation which is then updated by utilizing information from the whole test set, since key frames are selected based on the agreement between models in the ensemble or with an AdaBoost classifier which uses three types of features. Their method incorporates evidence from these keyframes and refines the 2D model and improves the final 3D pose estimation accuracy. Sigal *et al.* [135] proposed a probabilistic graphical model of body parts which they call *loose-limbed body model* [140] to obtain a representation of the human body. They use bottom-up part detectors to enhance robustness and local spatial (and temporal) coherence constraints to efficiently infer the final pose. The main differences from the aforementioned methods are that it applies to single human pose estimation and that the Particle Message Passing method they use to infer the pose from the graphical model, works in a continuous state space instead of a discretized one. The proposed method works both for pose estimation from a single image and for tracking over time by propagating pose information over time using importance sampling.

Elhayek *et al.* [100] proposed a novel marker-less method for tracking the 3D human joints in both indoor and outdoor scenarios using as low as two or three cameras. For each joint in 2D, a discriminative part-based method was selected which estimates the unary potentials by employing a convolutional network. Pose constraints are probabilistically extracted for tracking, by using the unary potentials and a weighted sampling from a pose posterior guided by the model. These constraints are combined with a similarity term, which measures for the images of each camera, the overlap between a 3D model, and the 2D Sums of Gaussians (SoG) images [141]. A new dataset called MPI-MARCOnI [99, 100] was also introduced which features 12 scenes recorded indoors and outdoors, with varying subjects, scenes, cameras, and motion complexity.

Finally, Daubney [136] introduced a method that permits greater uncertainty in the root node of the probabilistic graph that represents a human body by stochastically tracking the root node and estimating the posterior over the remaining parts after applying temporal diffusion. The state of each node, excluding the root, is represented as a quaternion rotation and all the distributions are modeled as Gaussians. Inference is performed by passing messages between nodes and the Maximum-a-Posteriori (MAP) estimated pose is selected.

Markerless Motion Capture using skeleton- and mesh-based approaches: Although such approaches form a different category since they may require a laser scan to extract the 3D mesh model, we will point out two methods for completeness that are illustrative and estimate the human body pose in 3D. Given an articulated tem-

Table 4: 3D human pose estimation from a sequence of multi-view RGB images. The Body Model column indicates whether a body model is employed. The Method Highlights column reflects the most important steps in each publication.

Year	First Author	Body Model	Method Highlights	Evaluation Datasets	Evaluation Metrics
2015	Elhayek [100]	Yes	Marker-less tracking of human joints; ConvNet-based 2D joint detection is combined with a generative tracking algorithm based on the Sums of Gaussians.	MARCOnI, HumanEva-I	3D
2015	Moutzouris [132]	Yes	*Training*: Activity Manifolds are learned by Hierarchical Temporal Laplacian Eigenmaps; *Testing*: Pose is constrained by the hierarchy of Activity Manifolds; Hierarchical Manifold Search explores the pose space with inputs the observation and the previously learned activity manifolds	HumanEva-II	3D
2014	Amin [133]	Yes	3DPS model for initial 3D pose estimation; Key frames are selected based on ensemble agreement and discriminative classification; 2D model is refined with that evidence to improve the final 3D pose estimation	MPII Cooking, Shelf	3D PCP
2014	Belagiannis [134]	Yes	Tracking for identity recovery and recover the pose by adding a temporal term to the 3DPS model	Campus, Shelf, KTH Multiview Football II	3D PCP
2014	Belagiannis [46]	Yes	Use 2D body part detectors and reduce pose space by triangulation; Introduce 3DPS model based on a CRF for inference	Campus, Shelf, HumanEva-I, KTH Multiview Football II	3D 3D PCP
2012	Sigal [135]	Yes	Loose-limbed body model representation with continuous-valued parameters and bottom-up part detectors; Belief propagation for inference	HumanEva-I	3D
2011	Daubney [136]	Yes	Increasing uncertainty in the root node of the probabilistic graph by stochastically tracking the root node and estimating the posterior over the remaining parts	HumanEva-I	3D
2011	Liu [137]	Yes	Use of shape priors for multi-view probabilistic image segmentation based on appearance, pose and shape information	Custom	3D
2009	Gall [138]	Yes	Local pose optimization and re-estimation of labeled bones; Positions of the vertices are refined to construct a surface which will be the initial model for the next frame	HumanEva-II, Custom	3D

plate model and silhouettes (extracted by background subtraction) from multi-view image sequences, Liu *et al.* [137] and Gall *et al.* [138] proposed methods that recover the movement not only of the skeleton but also the 3D surface of a human body. In both cases, the 3D triangle mesh surface model of the tracked subject in a static pose and the skinning weights of each vertex which connect the mesh to the skeleton can be acquired either by a laser scan or by shape-from-silhouette methods. In the work of Gall *et al.* [138] the skeleton pose is optimized locally and labeled bones (misaligned or with less than three DOF) are re-estimated by global optimization such that the projection of the deformed surface fits the image data in a globally optimal way. The positions of all vertices are then refined to fit the image data and the estimated refined surface and skeleton pose serve as initialization for the next frame to be tracked. Liu *et al.* [137] used shape priors to perform multi-view 2D probabilistic segmentation of foreground pixels based on appearance, pose and shape information. Their method can handle occlusions in challenging realistic human interactions between people.

5. Existing Datasets and Evaluation Measures

5.1. Evaluation Datasets

Discussing the advances in human pose estimation, Moeslund *et al.* [30] pointed out in their survey in 2006, that a limitation of existing research was the comparison of different approaches on common datasets and performance evaluation of accuracy against ground truth. The HumanEva dataset created by Sigal *et al.* [36] addresses exactly these issues. It contains multiple subjects performing a set of predefined actions with several repetitions. It is a comprehensive dataset that contains synchronized video from multiple camera views, associ-

Table 5: Evaluation datasets for the 3D pose estimation task and their key characteristics.

Year	Dataset	No. of Videos	No. of subjects	Characteristics
2015	MARCOnI [100]	12	Varies	Multi-view, Indoor & Outdoor, Varying number and types of cameras, and conditions
2015	PosePrior [142]	N/A	N/A	Prior based on Pose-Conditioned Joint Angle Limits
2014	Human 3.6M [143]	1,376	11	15 actions, 3.6×10^6 poses
2014	Campus [46]	1	3	Multiple People, Outdoor
2013	KTH Multiview Football II [144]	4	2	4 actions
2012	MPII Cooking Activities [145]	44	12	65 actions
2011	Shelf [146]	1	4	Multiple People
2011	UMPM [147]	36	30	Multiple People
2010	HumanEva-I&II [36]	56	4	6 actions
2010	TUD Stadtmitte [110]	1	N/A	Multiple People, Outdoor, Qualitative
2009	TUM Kitchen [148]	20	4	4 actions
2009	UVA 3D [149]	12	N/A	Outdoor
N/A	CMU Mocap [150]	2,605	109	23 actions

ated 3D ground truth, quantitative evaluation measures, and a baseline human tracking algorithm. It is divided into two sets (I & II) with different number and types of video cameras, a different numbers of motions, different types of data and synchronization. The datasets are broken into training, validation, and test sub-sets. For the testing subset, the ground truth data are withheld and a web-based evaluation system is provided. It is by far the most widely used dataset in the literature since it allows different methods to be fairly compared using the same data and the same error measures. A recognition-based motion capture baseline on the HumanEva-II dataset is provided by Howe [151]. Nevertheless, there are other significant datasets available for benchmarking (e.g., the CMU Graphics Lab Motion Capture database [150], the Human 3.6M dataset [143, 152] or the KTH Multiview Football Dataset II [144]). We present a summary of them in Table 5.

Limitations of the available datasets: Existing research, constrained by the available datasets, has addressed mainly the 3D pose estimation problem in controlled laboratory settings where the actors are performing specific actions. This is due to the fact that collecting videos in unconstrained environments with accurate 3D pose ground truth is impractical. Following the example of the HumanEva and the Human3.6M datasets, which have contributed to advances in the field over recent years, there is a need for realistic datasets which should be captured (i) in as unconstrained and varying conditions as possible where occlusions can occur, (ii) with several people with varying anthropometric dimensions, (iii) with not necessarily only one actor per video (iv) with actors that wear loosely fitting clothing, and (v) with human-human and human-object interactions.

5.2. Evaluation Metrics

The variety of challenges that arise in the human pose and motion estimation task result in several evaluation metrics adapted to the problem that the authors are trying to address each time. Thus, a fair comparison between the discussed methods would be impossible even for approaches that use the same dataset, since different methods train and evaluate differently. For the HumanEva dataset [36], the authors introduced the 3D Error ($\mathcal{E}$) which is the mean squared distance in 3D (measured in millimeters) between the set of virtual markers corresponding to the joint centers and limb ends:

$$\mathcal{E}(x, \hat{x}) = \frac{1}{M} \sum_{i=1}^{M} \|m_i(x) - m_i(\hat{x})\|, \tag{1}$$

where x represents the ground truth pose, $\hat{x}$ refers to the estimated pose, M is the number of virtual markers and $m_i(x)$ represents the 3D position of the i^{th} marker. It is also referred to as Mean Per Joint Position Error (MPJPE) [143]. Simo-Serra *et al.* [79] introduced a rigid alignment step using least squares to compare with methods that do not estimate a global rigid transformation. They refer to this error as 3D pose error.

A common 2D pose estimation error in the literature is the Percentage of Correctly estimated Parts (PCP) error which measures the percentage of correctly localized body parts [153]. The PCP error has recently been used in 3D [45, 46, 112, 133] and a part is classified as "correctly estimated" if:

$$\frac{\|s_n - \hat{s}_n\| + \|e_n - \hat{e}_n\|}{2} \leq \alpha \|s_n - e_n\|, \tag{2}$$

Table 6: 3D pose error (i.e., after performing rigid alignment) in *mm* of methods that employ information from a monocular single image on the HumanEva-I dataset. Results are reported for camera C1 and S1, S2 and S3 refer to the three subjects that perform the action.

Year	Method	Walking				Jogging			
		S1	S2	S3	Average	S1	S2	S3	Average
2016	Yasin *et al.* [65]	**35.8**	32.4	41.6	**36.6**	46.6	41.4	35.4	**41.1**
2014	Kostrikov *et al.* [68]	44.0	**30.9**	41.7	38.9	57.2	35.0	33.3	41.8
2014	Wang *et al.* [70]	71.9	75.7	85.3	77.6	62.6	77.7	54.4	64.9
2013	Radwan *et al.* [74]	75.1	99.8	93.8	89.6	79.2	89.8	99.4	89.5
2013	Simo-Serra *et al.* [75]	65.1	48.6	73.5	62.4	74.2	46.6	**32.2**	51.0
2012	Simo-Serra *et al.* [79]	99.6	108.3	127.4	111.8	109.2	93.1	115.8	106.0
2010	Bo and Sminchisescu [116]	38.2	32.8	**40.2**	37.1	**42.0**	**34.7**	46.4	**41.0**

where s_n and e_n represent the ground truth 3D coordinates of the start and end points of part n, $\hat{s}_n$ and $\hat{e}_n$ the respective estimations, and α is the parameter that controls the threshold.

Evaluation measurements are frequently used to capture error in degrees. Illustrative examples can be found in the early publications of Agarwal and Triggs [64, 154] and in the work of Ning *et al.* [40]. The Mean Joint Angle Error (MJAE) is the mean (over all angles) absolute difference between the true and estimated joint angles in degrees, and is given by:

$$MJAE = \frac{\sum_{i=1}^{M} |(y_i - y'_i) \; mod \; \pm 180^o|}{M}, \qquad (3)$$

where M is the number of joints and y_i and y'_i are the estimated and ground truth pose vectors respectively and mod is the modulus operator. The $mod \; \pm 180^o$ term reduces angles to the $[-180^o, +180^o]$ range.

5.3. Summary of performance on HumanEva-I

Aiming to better understand the advantages and limitations of various 3D human pose estimation approaches we focused on the HumanEva-I dataset, grouped the respective methods based on the input (e.g., single image or video, monocular or multi-view), and report performance comparisons in each category. In Table 6 we report the 3D pose error (i.e., after performing rigid alignment as suggested by Simo-Serra *et al.* [79]) of methods that employ information from a single monocular image. Tables 7 and 8 summarize the results (3D error in *mm*) of methods, the input of which is a single multi-view image or a video respectively. However, this comparison is just meant to be indicative, and cannot be treated as complete, since: (i) it covers a subset of the state of the art, and (ii) different methods are trained and evaluated differently depending on the problem that are trying to address.

From the approaches that use a single monocular image as an input, the best results for each action are obtained by the methods of Yasin *et al.* [65], Kostrikov *et al.* [68] and Bo and Sminchisescu [116]. The key characteristic of Yasin *et al.* [65] is that they employ information from 3D motion capture data and project them in 2D so as to train a regression model that predicts the 3D-2D projections from the 2D joint annotations. After estimating the 3D pose of a new test image, the 2D pose can be refined and iteratively update the 3D pose estimation.

For the results reported in Table 6, motion capture data from the HumanEva-I dataset is used to train the regressor. When the motion capture data are from a different dataset (e.g., CMU Mocap [150]), the 3D pose error is 55.3 *mm* for the walking action, which is still better than most of the methods. However, for the jogging action, the respective average 3D pose error is 67.9 *mm* from which we conclude that estimating the 3D pose for more complicated and not cyclic actions without constrained prior information (i.e., 3D motion capture data from the same dataset) still remains a challenging task. Kostrikov *et al.* [68] proposed a method that uses regression forests to infer missing depth data of image features and 3D pose simultaneously. They hypothesize the depth of the features by sweeping with a plane through the 3D volume of potential joint locations and employ a 3D PSM to obtain the final pose. Unlike Yasin *et al.* [65], a limitation of this approach is that it requires the 3D annotations during training to learn the 3D volume. Finally, Bo and Sminchisescu [116] followed a regression-based approach that uses Gaussian process prior to model correlations among both inputs and outputs in multivariate, continuous valued supervised learning problems. The advantages of their approach are that it does not require an initial pose to

Table 7: 3D Error ($\mathcal{E}$) in *mm* of single-image, multi-view methods of subject S1 on the HumanEva-I dataset.

Year	Method	Walking	Boxing
2013	Amin *et al.* [60]	54.5	**47.7**
2012	Sigal *et al.* [135]	89.7	N/A
2011	Yao *et al.* [123]	**44.0**	74.1
2010	Taylor *et al.* [124]	55.4	75.4

perform an optimization scheme or a 3D body model such as the method of Wang *et al.* [70]. However, background subtraction is utilized, which is a strong assumption that prevents this method of dealing with real-life scenarios with a changing background.

An interesting observation arises from the reported results of methods that employ information from multiple views whether it's from a single image (Table 7) or from a sequence of images (Table 8). That is, the 3D error of the method of Amin *et al.* [60] that uses a single multi-view image is significantly lower than the 4 methods reported in Table 8 the input of which are multi-view videos. An explanation for this is that video-based methods do not exploit fully the temporal information (e.g., by following a tracking-by-detection approach [110]). Finally, the 3D error reported by Tekin *et al.* [103] which is based on a monocular sequence of images is lower in the walking sequence than the rest of the methods regardless of the number of views. The key characteristic of this regression-based approach is that they exploit temporal information very early in the modeling process, by using a Kernel Dependency Estimation (in the case of HumanEva-I) to shift the windows of the detected person so as the subject remains centered. Then 3D HoG features are extracted to form a spatiotemporal volume of bounding boxes, and a regression scheme is employed to predict the final 3D pose.

In summary, although model-based approaches have demonstrated significant improvements over the past few years, regression-based approaches, despite their own limitations which are described in detail in Section 1.2 and in the review of Sigal [37], tend to outperform the rest, at least in the HumanEva-I dataset.

Table 8: 3D Error ($\mathcal{E}$) in *mm* of video-based methods of subject S1 on the HumanEva-I dataset. The first two approaches employ information from a sequence of monocular images whereas the last four are multi-view.

Year	Method	Walking	Boxing
2016	Tekin *et al.* [103]	**37.5**	50.5
2010	Bo and Sminchisescu [116]	45.4	**42.5**
2015	Elhayek *et al.* [100]	66.5	**60.0**
2014	Belagiannis *et al.* [46]	68.3	62.7
2012	Sigal *et al.* [135]	**66.0**	N/A
2011	Daubney and Xie [39]	87.3	N/A

6. Experimental Evaluation

In order to provide the reader with an overview of the current capabilities of 3D human pose estimation techniques we conducted extensive experimentations with three state-of-the-art approaches, namely the methods of Wang *et al.* [70], Zhou *et al.* [72] and Bo and Sminchisescu [116] (presented in detail in Section 3.1). Instead of evaluating these methods in existing real-world datasets, we decided to go for a more generic evaluation and specifically developed a 3D synthetic dataset that simulates the human environment, i.e., the SynPose300 dataset. The idea is to be able to fully control the testing environments and human poses and ensure a common evaluation setup, which would have not been possible using actual people/actors (e.g., due to human and time constraints). SynPose300 can be used by the research community as supplementary to the existing datasets when the goal is to test the robustness of 3D pose estimation techniques to (i) different anthropometric measurements for each gender; (ii) the viewing distance and the angle; (iii) actions of varying difficulty; and (iv) larger clothes. The SynPose300 dataset and the ground truth are released publicly for the reproducibility of the results and are available online at [155]. Various experiments were conducted using this open framework and insights are provided.

Limitations: When designing the conditions of the proposed dataset we focused only on specific parameters (anthropometry, action, clothes, distance and angle from the camera). Thus, the conditions in which the synthetic human models act, such as the background and lack of noise, are not realistic. The available options for the clothes of the human models were limited and, as a result, we used jeans and jackets in the large clothes category since loosely fitting garments such as dresses were not available.

The rest of this section is structured as follows. Section 6.1 provides the description of SynPose300 and in Section 6.2 the experimental investigation of the chosen approaches is presented.

6.1. Description of the Proposed Synthetic Dataset

As mentioned in the introduction of this section, we evaluated the state-of-the-art 3D pose estimation approaches of Wang *et al.* [70], Zhou *et al.* [72] and Bo

and Sminchisescu [116] - the code of which is publicly available - on a synthetic dataset (SynPose300) we created for this paper. Existing datasets cover in great depth pose variations under different actions and scenarios. However, they contain a small number of humans of unknown anthropometric measurements, in controlled environments, who wear tight clothes which facilitate the pose estimation task. For example, all actors in the Human3.6M dataset [143] wear shorts and t-shirts, whereas in the HumanEva dataset [36] only one out of four subjects is female and only one wears clothes which are not tight.

To provide an even more challenging evaluation framework, SynPose300 provides a controlled environment where the aforementioned constraints are taken into consideration. SynPose300 comprises 288 videos, their respective 3D ground truth joint locations (25-limb models) and the parameters of the camera (focal length, location and translation matrices). Each video is five seconds long (24 fps), encoded using Xvid encoder (800x600 resolution, RGB images). The videos were generated using the open source software tools MakeHuman [156] and Blender [157]. The ground truth was computed for each video using the .bvh files. First, each video was exported from Blender in Biovision Hierarchy format. The resulting files were further processed in MATLAB and for each frame of the video, we parsed the 3D coordinates for all 26 joints. Its summary can be found in Table 9.

Table 9: SynPose300 Dataset summary

Subjects	8 (4 female & 4 male)
Percentiles (%)	20, 40, 60, 80
Actions	3
Distances	2 (close, far)
Points of View	3 (0^o, 45^o, 90^o)
Types of Clothes	2 (Tight, Large)

It contains videos from eight virtual humans (four female and four male) all of which follow specific anthropometric measurements in percentiles obtained from the CAESAR anthropometric database [158]. For example, a 20^{th} percentile male, is a male with 20^{th} percentile segments. The measurements we used are the stature, the spine to shoulder distance, the shoulder to elbow distance, the elbow to wrist distance, the knee height, the hip to knee distance, the hip circumference, the pelvis to neck distance and the neck to head distance. The humans perform three actions ("walking", "picking up a box" and "gymnastics") that were selected based on the 3D pose estimation difficulty they present. Each video was captured with both tight and larger clothes, from three points of view (0^o, 45^o, 90^o) and from two camera distance ranges (close and far). In the "close" scenario, the distance of the human model from the camera in the first frame ranges from 3 *m* to 5 *m* and was selected so the human was as close as possible to the camera, but without getting out of the camera's field of view while performing the action. In the "far" case, the distance is twice as much.

In general, background subtraction methods are employed as a pre-processing step to isolate humans from the background [159]. Therefore, in our tests we do not employ background. Illustrative examples of the proposed dataset are depicted in Figure 6.

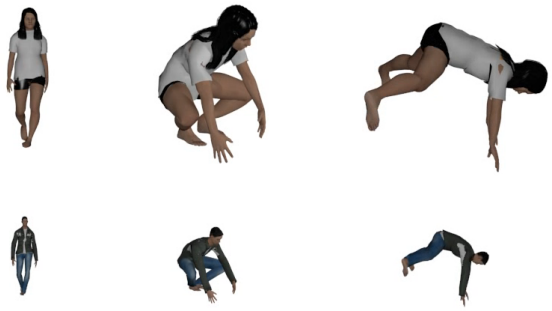


Figure 6: Illustrative frames of the SynPose300 dataset. The first row corresponds to 20^{th}, 40^{th} and 80^{th} percentile females wearing tight clothes, performing three actions from a "close" camera distance, and under (0^o, 45^o, 90^o) points of view. The second row corresponds to males of the same percentiles wearing large clothes, at the same frame of the video, performing the same actions from a "far" camera distance.

6.2. *Experimental Results*

The methods of Wang *et al.* [70] and Zhou *et al.* [72] represent a 3D pose as a combination of a set of bases (i.e., 3D human poses) which are learned from the whole SynPose300 dataset. In both methods, the dictionary of poses is learned using sparse coding. Once the set of bases is computed, 2D pose estimation is performed to obtain the joint locations in the image plane. Then, starting from an initial 3D pose, an Alternating Direction Method of Multipliers (ADMM) optimization scheme is followed to estimate the joint locations in 3D. A more extensive investigation on the impact of the dictionary of bases, on the 3D pose estimation accuracy is offered in Experiment 3. Other parameters required as prior information for both methods are (i) the structure of the 2D skeletal model and (ii) the structure of the 3D skeleton both of which remained unchanged throughout

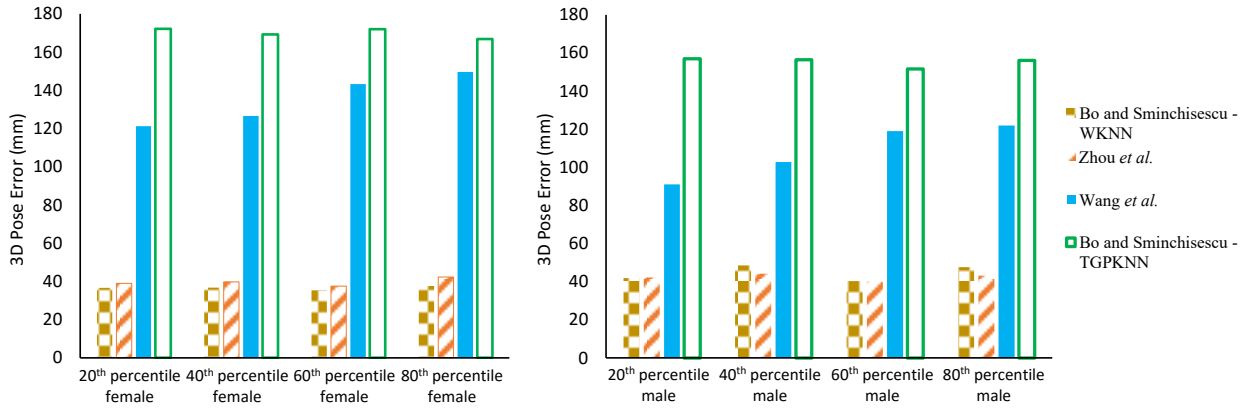


Figure 7: Comparison of the performance of four state-of-the-art approaches for female (left) male (right) body models for the walking action with varying percentiles of anthropometric measurements.

the experimental investigation. For the 2D pose estimation task we employed the approach of Yang and Ramanan [61], which uses a pre-trained 2D human skeletal model trained on the PARSE dataset [160]. To evaluate the regression-based method of Bo and Sminchisescu [116], background subtraction is first performed so as to keep only the region where the human is located in each frame, and HoG features are extracted. Aiming to keep the length of the feature vector consistent throughout the video sequence, we first resized the obtained region to the size of the bounding box of the first frame and then computed the HoG features. From the provided techniques, we evaluated the Twin Gaussian Processes with K Nearest Neighbors and the Weighted K-Nearest Neighbor Regression methods.

To speed-up the experimental procedure and since temporal information from the video is not exploited, we processed one every five frames for each video. Thus, 3D pose estimation was performed in 24 monocular images per video. Finally, the evaluation metric we chose is the 3D pose error which is the mean Euclidean distance between the estimation and the ground truth after performing a rigid alignment of the two shapes as proposed by Simo-Serra *et al.* [79].

Experiment 1 - Robustness to different anthropometric measurements: The objective of this experiment is to assess the impact of different anthropometric measurements on the 3D pose estimation task. The obtained results are summarized in Figure 7. The anthropometry of the human whose pose we want to estimate affects the 3D pose estimation accuracy when the model-based approach of Wang *et al.* [70] is tested. In that case, the error is increasing more than 20% when we use a 40^{th} instead of a 20^{th} percentile female. For the rest of the techniques, the anthropometry of the human does not affect the pose estimation accuracy.

Experiment 2 - Robustness to viewing distance and point of view: The objective of this experiment is to assess the robustness of the 3D pose estimation approaches to different points of view and distances from the camera. A summary of the results is offered in Table 10. The camera view affects the pose accuracy, as estimating the 3D pose from a wider angle is a more challenging task since occlusions between different parts take place. When the camera is placed at a close distance from the human, the approaches of Zhou *et al.* [72] and the Weighted KNN of Bo and Sminchisescu [116] have an average 3D pose error of 35.3 and 26.1 *mm* over all videos of the walking action when the point of view is 0^o. When the angle of the camera is at 90^o, all four methods performed significantly worse. A similar pattern is followed when the distance from the camera is increased, since the average error over all angles increased by 12.33% for the method of Wang *et al.* [70] and 15.32% for the Weighted KNN method of Bo and Sminchisescu [116]. For both model-based approaches, the change of the angle of the camera from 0^o to 45^o does not affect significantly the pose estimation accuracy since the 3D shape can be reconstructed from the available poses in the dictionary with approximately the same error. Furthermore, we observed that the approach of Zhou *et al.* [72] was more robust to angle and distance variations, and that at 90^o, for all four methods, the change of distance of the camera does not affect significantly the 3D pose error, since the pose is already difficult to be estimated accurately.

Experiment 3 - Robustness to the error imposed by the dictionary of poses: For the model-based approaches of Wang *et al.* [70] and Zhou *et al.* [72] the

Table 10: 3D pose error in *mm* for four state-of-the-art methods under different distances from the camera and points of view for the walking action.

	Close			Far		
	0^o	45^o	90^o	0^o	45^o	90^o
Wang *et al.* [70]	106.7	108.6	131.7	126.4	125.3	135.9
Zhou *et al.* [72]	35.3	**37.6**	**41.2**	38.4	**38.8**	**43.0**
Bo and Sminchisescu [116] - WKNN	**26.1**	38.8	52.5	**26.4**	53.3	56.4
Bo and Sminchisescu [116] - TGPKNN	154.4	180.8	167.2	139.2	164.2	170.1

dictionary was comprised of poses of both different anthropometric measurements and 3D pose stances than the ground truth. Aiming to investigate the sensitivity of the pose estimation task to the initialization of the 3D pose and the set of poses in the dictionary we experimented under two scenarios. In the first case, the provided initial 3D pose has varying anthropometric measurements, but the rest of the conditions are kept the same and the pose stance of the humans is 100 % accurate. For example, for a 20^{th} percentile female in the j^{th} frame of a video, we provide each time the pose of the j^{th} frame of similar videos of a different percentile female. Note that in this experiment we investigated only the method of Wang *et al.* [70], since its accuracy depends more on the conditions of experimental setup. A summary of the results can be found in Tables 11 and 12. In both female and male cases, the 3D pose error increases when the provided initial anthropometry is far from the real one. However, the errors obtained in this experiment are lower than all the other scenarios from which we can conclude that finding the correct pose stance - even when the anthropometry of the estimated human is wrong - is the most challenging task in the 3D pose estimation problem.

Table 11: 3D pose error in *mm* for females when the pose stance of the initial pose is correct and the anthropometric measurements vary. The rows correspond to the anthropometric measurements of the ground truth whereas the columns to the anthropometric measurements of the initial pose provided as an input.

	F_{20}	F_{40}	F_{60}	F_{80}
F_{20}	-	37	75	103
F_{40}	53	-	34	66
F_{60}	70	34	-	32
F_{80}	101	65	40	-

In the second case, the dictionary of poses is comprised of poses only from the respective video, and thus it has the correct anthropometric measurements but an inaccurate pose stance which has to be estimated. The

Table 12: 3D pose error in *mm* for males when the pose stance of the initial pose is correct and the anthropometric measurements vary. The rows correspond to the anthropometric measurements of the ground truth whereas the columns to the anthropometric measurements of the initial pose provided as an input.

	M_{20}	M_{40}	M_{60}	M_{80}
M_{20}	-	37	75	110
M_{40}	41	-	51	75
M_{60}	83	44	-	82
M_{80}	109	73	45	-

method of Zhou *et al.* [72] was tested and the results obtained from the second scenario are presented in Figure 8. When the dictionary contains poses that belong only from the same video, there is a 31.4% decrease in the 3D pose error for the walking action and a 24.3% for the picking-up-box action which is of medium difficulty. The error in the third and most challenging action is reduced only by 6.9%, and the reason for this is the really challenging nature of the gymnastics action.

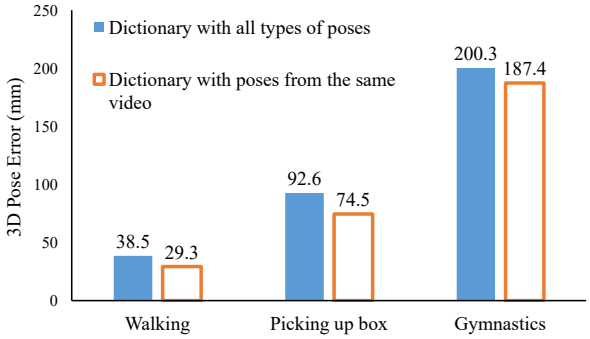


Figure 8: The impact of the bases in dictionary on the 3D pose error for actions of varying difficulty for the method of Zhou *et al.* [72]

Experiment 4 - Robustness to actions with different levels of difficulty and to large clothing: The objectives of these experiments are to test the importance of

Table 13: The impact of the difficulty of the action and the type of clothes is investigated for four state-of-the-art approaches.

	Actions			Clothes	
	Walking	Picking up box	Gymnastics	Tight	Large
Wang *et al.* [70]	105.4	295.2	331.4	129.1	118.6
Zhou *et al.* [72]	**38.5**	92.6	**200.3**	**38.3**	**35.4**
Bo and Sminchisescu [116] - WKNN	40.6	**90.6**	214.3	39.3	41.8
Bo and Sminchisescu [116] - TGPKNN	162.6	287.0	318.9	159.8	165.5

the impact of actions with varying difficulty and clothing of a human on the 3D pose estimation performance. The results for different types of actions and clothes are depicted in Table 13. As the difficulty of the action increases (from walking to pick up box and then to gymnastics), the error increases in all cases. The approaches of Zhou *et al.* [72] and the Weighted KNN of Bo and Sminchisescu [116] demonstrated small 3D pose errors for the first two actions compared to the other two techniques. In the gymnastics action which demonstrates high variance and challenging poses, the major source of error is in the estimated pose and not in depth. When different types of clothes model-based methods performed better with larger clothes whereas regression-based approaches demonstrated better results with tighter clothes. This behavior can be attributed to the fact that regression-based techniques like the Weighted KNN or the Twin Gaussian Processes KNN rely on image-based features (i.e., HoGs), whereas the methods of Wang *et al.* [70] and Zhou *et al.* [72] require an initial model and a skeleton structure that tend to generalize better when the human wears larger clothes and thus, the human silhouette covers a larger region.

Comparison with the originally reported results: For completeness we present the evaluation results of the two state-of-the-art methods as reported in the respective publications. We compare the results obtained from the "walking" action in our investigation with the method of Wang *et al.* [70] and Bo and Sminchisescu [116] which are tested in the walking action of the HumanEva-I dataset. The reported results are presented in Table 14. The results of both methods are better than those obtained from our experimental investigation for the walking action, and the reason for this is that the conditions under which we tested the robustness of the 3D pose estimation accuracy are more challenging.

Table 14: 3D pose error in *mm* of the methods of Wang *et al.* [70] and Bo and Sminchisescu [116] in the walking actions for camera C1 of the HumanEva-I dataset. S1, S2 and S3 refer to the three subjects that perform the action.

Dataset	[70]	[116]
HumanEva-I S1	71.9	38.2
HumanEva-I S2	75.7	32.8
HumanEva-I S3	85.3	40.2
SynPose300	105.4	40.6

Aiming to identify which joints contribute the most towards the final error, we also compute the average 3D pose error per joint throughout the whole dataset and the results are depicted in Figure 9. Joints that belong to the main body, such as the neck, pelvis and hips, contribute relatively little towards the total error compared to the wrists or the feet.

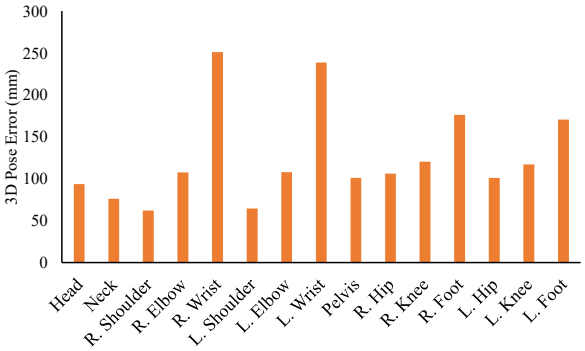
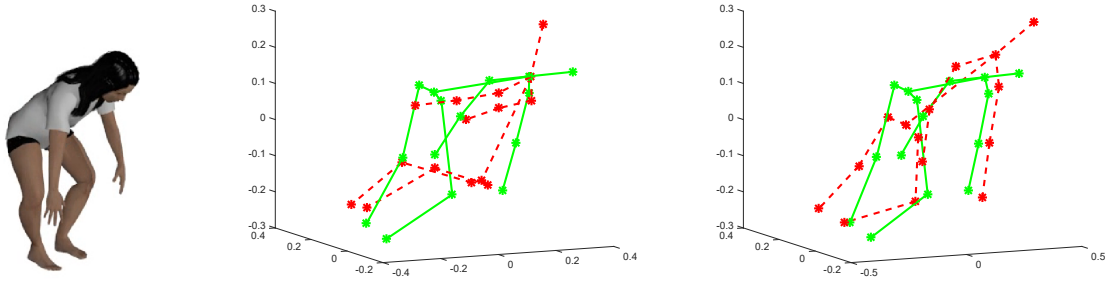


Figure 9: Mean 3D pose error per joint throughout the whole dataset for the Weighted KNN method of Bo and Sminchisescu [116]. L corresponds to left and R to right.

Finally, we present examples of failures and successful estimations of the method of Wang *et al.* [70] in images of the SynPose300 dataset. In Figure 10, given the same frame (first image) and the same 2D joint locations, by applying the method of Wang *et al.* [70] twice, different results are obtained. In the second image the pose estimation algorithm fails since the estimated pose is presented facing the opposite direction. The reasons for this behavior are: a) the ill-posed nature of the problem and b) the human is bending and thus, the 2D locations are misleading for the recovery of the 3D joints.



(a) (b) (c)

Figure 10: For the same image frame (a), we present examples where the method of Wang *et al.* [70] fails (b), and successfully estimates (c), the 3D pose of the human subject. The green solid line human skeleton corresponds to the ground truth and the red dashed line to the estimated 3D pose.

7. Conclusion

In this paper, we offer an overview of recently published papers addressing the 3D pose estimation problem from RGB images or videos. We proposed a taxonomy that organizes pose estimation approaches into four categories based on the input signal: (i) monocular single image, (ii) single image in multi-view scenario, (iii) sequence of monocular images, and (iv) sequence of images in multi-view settings. In each category, we grouped the methods based on their characteristics and paid particular attention to the body model when employed, to the pose estimation approach used, how inference is performed, and to other important steps such as the features or the pre-processing methods used. We created a synthetic dataset of human models performing actions under varying conditions and assessed the sensitivity of the pose estimation error under different scenarios. The parameters we investigated were the anthropometry of the human model, the distance from the camera and the point of view, the clothing and the action performed.

Articulated 3D pose and motion estimation is a challenging problem in computer vision that has received a great deal of attention over the last few years because of its applications in various scientific or industrial domains. Research has been conducted vigorously in this area for the past few years, and much progress has been made in the field. Encouraging results have been obtained, strong methods have been proposed to address the most challenging problems that occur and current 3D pose estimation systems have reached a satisfactory maturity when operating under constrained conditions. However, they are far from reaching the ideal goal of performing adequately in the conditions commonly encountered by applications utilizing these techniques in practical life. Thus, 3D pose estimation remains a largely unsolved problem and its key challenges are discussed in the rest of this section.

The first challenge is the ill-posed nature of the 3D pose estimation task especially from a single monocular image. Similar image projections can be derived from completely different 3D poses due to the loss of 3D information. In such cases, self-occlusions are common phenomena, which result in ambiguities that prevent existing techniques from performing adequately. A promising solution towards this direction is utilizing temporal information or multi-view setups which can resolve most of the ambiguities that arise in monocular scenarios [110, 134].

The second issue is the variability of human poses and shapes in images or videos in which the subjects perform complicated actions such as the gymnastics action in the proposed SynPose300 dataset. To address this issue future approaches can benefit from the recent release of the PosePrior dataset [142] which includes a prior that allows anthropometrically valid poses and restricts the ones that are invalid.

A third challenging task is the estimation of the 3D pose of multiple people who interact with each other and with the environment. In such cases handling occlusions is a difficult task since, besides self-occlusions, occlusions of limbs can also occur from other people or objects. Methods that employ a tracking-by-detection approach in a multi-view setup [46, 110, 134] can overcome most of these difficulties and address this problem successfully.

Finally, 3D pose estimation cannot be successfully incorporated in real-life applications unless future approaches are able to perform sufficiently in outdoor environments where the lighting and background condi-

tions as well as the behavior of the humans subjects are unconstrained. Although datasets in unconstrained outdoor environments with accurate 3D pose ground truth, would result into future approaches that would tackle these limitations, their creation is almost unrealistic due to the size of hardware equipment that is required to capture 3D data. A few methods [46, 149] have tried to address this limitation, by providing datasets (along with 3D annotation of the joints) in outdoor environments but they are far from simulating effectively real-life conditions.

Acknowledgments

This work has been funded in part by the Ministry of European Funds through the Financial Agreement POSDRU 187/1.5/S/155420 and the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

[1] Y. Song, D. Demirdjian, R. Davis, Continuous body and hand gesture recognition for natural human-computer interaction, ACM Transactions on Interactive Intelligent Systems 2 (1) (2012) 5. 1

[2] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, X. Twombly, Vision-based hand pose estimation: A review, Computer Vision and Image Understanding 108 (1) (2007) 52–73. 1

[3] D. McColl, Z. Zhang, G. Nejat, Human body pose interpretation and classification for social human-robot interaction, International Journal of Social Robotics 3 (3) (2011) 313–332. 1

[4] D. Droeschel, S. Behnke, 3D body pose estimation using an adaptive person model for articulated ICP, in: Intelligent Robotics and Applications, Springer, 2011, pp. 157–167. 1

[5] C. Chen, Y. Yang, F. Nie, J. Odobez, 3D human pose recovery from image by efficient visual feature selection, Computer Vision and Image Understanding 115 (3) (2011) 290–299. 1, 2, 6, 11, 12

[6] S. Sedai, M. Bennamoun, D. Huynh, Context-based appearance descriptor for 3D human pose estimation from monocular images, in: Proc. IEEE Digital Image Computing: Techniques and Applications, Melbourne, VIC, 2009, pp. 484–491. 1

[7] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124. 1, 3, 10

[8] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, B. A., Efficient human pose estimation from single depth images, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12) (2013) 2821–2840. 1

[9] E. Suma, B. Lange, A. S. Rizzo, D. M. Krum, M. Bolas, Faast: the flexible action and articulated skeleton toolkit, in: IEEE Conference on Virtual Reality, Singapore, 2011, pp. 247–248. 1

[10] M. Fastovets, J.-Y. Guillemaut, A. Hilton, Athlete pose estimation from monocular TV sports footage, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, Oregon, 2013, pp. 1048–1054. 1

[11] L. Unzueta, J. Goenetxea, M. Rodriguez, M. T. Linaza, Viewpoint-dependent 3D human body posing for sports legacy recovery from images and video, in: Proc. 22^{nd} IEEE European Signal Processing Conference, Lisbon, Portugal, 2014, pp. 361–365. 1

[12] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: Proc. British Machine Vision Conference, Aberystwyth, Wales, 2010, pp. 12.1–12.11. 1

[13] A. Gupta, S. Satkin, A. A. Efros, M. Hebert, From 3D scene geometry to human workspace, in: Proc. 24^{th} IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011, pp. 1961–1968. 1

[14] Y. Zheng, H. Liu, J. Dorsey, N. J. Mitra, Ergonomics-inspired reshaping and exploration of collections of models, IEEE Transactions on Visualization and Computer Graphics (2015) 1–14. 1

[15] Y. Yang, S. Baker, A. Kannan, D. Ramanan, Recognizing proxemics in personal photos, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 3522–3529. 2

[16] C. Barron, I. Kakadiaris, Estimating anthropometry and pose from a single image, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, 2000, pp. 669–676. 2

[17] C. Barron, I. Kakadiaris, On the improvement of anthropometry and pose estimation from a single uncalibrated image, Machine Vision and Applications - Special Issue on Human Modeling, Analysis and Synthesis 14 (4) (2003) 229–236. 2

[18] I. A. Kakadiaris, N. Sarafianos, N. Christophoros, Show me your body: Gender classification from still images, in: Proc. 23^{rd} IEEE International Conference on Image Processing, Phoenix, AZ, 2016. 2

[19] Y. Zhang, T. Han, Z. Ren, N. Umetani, X. Tong, Y. Liu, T. Shiratori, X. Cao, Bodyavatar: creating freeform 3D avatars using first-person body gestures, in: Proc. 26^{th} annual ACM symposium on user interface software and technology, ACM, St. Andrews, Scotland, United Kingdom, 2013, pp. 387–396. 2

[20] A. Barmpoutis, Tensor body: real-time reconstruction of the human body and avatar synthesis from RGB-D, IEEE Transactions on Cybernetics 43 (5) (2013) 1347–1356. 2

[21] R. Pugliese, K. Förger, T. Takala, Game experience when controlling a weak avatar in full-body enaction, in: Proc. Intelligent Virtual Agents, Springer, Delft, The Netherlands, 2015, pp. 418–431. 2

[22] H. Jiang, K. Grauman, Seeing invisible poses: Estimating 3D body pose from egocentric video, arXiv preprint arXiv:1603.07763. 2

[23] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: Proc. 12th European Conference on Computer Vision, Springer, 2012, pp. 609–623. 2

[24] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, T. L. Berg, Parsing clothing in fashion photographs, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Providence, Rhode Island, 2012, pp. 3570–3577. 2

[25] I. Kakadiaris, D. Metaxas, Model-based estimation of 3D human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1453–1459. 2

[26] T. Helten, A. Baak, M. Müller, C. Theobalt, Full-body human motion capture from monocular depth images, in: Time-of-

Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer, 2013, pp. 188–206. 3
[27] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer, 2013, pp. 149–187. 3
[28] J. K. Aggarwal, Q. Cai, Human motion analysis: A review, in: Proc. IEEE Nonrigid and Articulated Motion Workshop, San Juan, Puerto Rico, 1997, pp. 90–102. 3
[29] D. M. Gavrila, The visual analysis of human movement: A survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98. 3
[30] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2) (2006) 90–126. 3, 14, 15
[31] R. Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding 108 (1) (2007) 4–18. 3, 14
[32] X. Ji, H. Liu, Advances in view-invariant human motion analysis: A review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40 (1) (2010) 13–24. 3
[33] C. Sminchisescu, 3D human motion analysis in monocular video: techniques and challenges, in: Human Motion, Springer, 2008, pp. 185–211. 3
[34] M. B. Holte, C. Tran, M. M. Trivedi, T. B. Moeslund, Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments, IEEE Journal of Selected Topics in Signal Processing 6 (5) (2012) 538–552. 3
[35] L. Sigal, M. J. Black, Guest editorial: State of the art in image- and video-based human pose and motion estimation, International Journal of Computer Vision 87 (1) (2010) 1–3. 3, 4
[36] L. Sigal, A. O. Balan, M. J. Black, HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (1-2) (2010) 4–27. 3, 15, 16, 19
[37] L. Sigal, Human pose estimation, Computer Vision: A Reference Guide (2014) 362–370. 3, 10, 18
[38] T. B. Moeslund, A. Hilton, V. Krüger, L. Sigal, Visual analysis of humans: looking at people, Springer, 2011. 3
[39] B. Daubney, D. Gibson, N. Campbell, Estimating pose of articulated objects using low-level motion, Computer Vision and Image Understanding 116 (3) (2012) 330–346. 4, 18
[40] H. Ning, W. Xu, Y. Gong, T. Huang, Discriminative learning of visual words for 3D human pose estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1–8. 4, 6, 7, 9, 17
[41] C. Sminchisescu, Estimation algorithms for ambiguous visual models. Three-dimensional human modeling and motion reconstruction in monocular video sequences, Ph.D. thesis, Institut National Polytechnique de Grenoble, France (July 2002). 4
[42] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79. 4
[43] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, 2D articulated human pose estimation and retrieval in (almost) unconstrained still images, International Journal of Computer Vision 99 (2) (2012) 190–214. 4
[44] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet conditioned pictorial structures, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 2013, pp. 588–595. 4
[45] M. Burenius, J. Sullivan, S. Carlsson, 3D pictorial structures for multiple view articulated pose estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 2013, pp. 3618–3625. 4, 10, 11, 14, 16
[46] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3D pictorial structures for multiple human pose estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1669–1676. 4, 14, 15, 16, 18, 23, 24
[47] S. Zuffi, O. Freifeld, M. J. Black, From pictorial structures to deformable structures, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 3546–3553. 4
[48] S. Zuffi, M. Black, The stitched puppet: A graphical model of 3D human shape and pose, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, 2015, pp. 3537–3546. 4
[49] J.-B. Huang, M.-H. Yang, Estimating human pose from occluded images, in: Proc. 9^{th} Asian Conference on Computer Vision - Volume Part I, Springer, Xian, China, 2009, pp. 48–60. 4, 6, 7, 9
[50] S. Sedai, M. Bennamoun, D. Huynh, Localized fusion of shape and appearance features for 3D human pose estimation, in: Proc. British Machine Vision Conference, Aberystwyth, Wales, 2010, pp. 51.1–51.10. 4, 11
[51] K. Grauman, G. Shakhnarovich, T. Darrell, Inferring 3D structure with a statistical image-based shape model, in: Proc. 9^{th} IEEE International Conference on Computer Vision, IEEE, Nice, France, 2003, pp. 641–647. 4
[52] M. Van den Bergh, E. Koller-Meier, R. Kehl, L. Van Gool, Real-time 3D body pose estimation, Multi-Camera Networks: Principles and Applications 335 (2). 4
[53] S. Sedai, M. Bennamoun, D. Q. Huynh, A Gaussian process guided particle filter for tracking 3D human pose in video, IEEE Transactions on Image Processing 22 (11) (2013) 4286–4300. 4, 12, 13
[54] R. Rosales, S. Sclaroff, Combining generative and discriminative models in a framework for articulated pose estimation, International Journal of Computer Vision 67 (3) (2006) 251–276. 4
[55] M. Salzmann, R. Urtasun, Combining discriminative and generative methods for 3D deformable surface and articulated pose reconstruction, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 647–654. 5
[56] C. M. Bishop, J. a. Lasserre, Generative or discriminative? getting the best of both worlds, Bayesian Statistics 8 (2007) 3–23. 5
[57] E. Marinoiu, D. Papava, C. Sminchisescu, Pictorial human spaces: how well do humans perceive a 3D articulated pose?, in: Proc. IEEE International Conference on Computer Vision, Sydney, Australia, 2013, pp. 1289–1296. 5
[58] A. Gupta, A. Mittal, L. S. Davis, Constraint integration for efficient multiview pose estimation with self-occlusions, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008) 493–506. 6
[59] J. Müller, M. Arens, Human pose estimation with implicit shape models, in: Proc. 1st ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, ACM, Firenze, Italy, 2010, pp. 9–14. 6
[60] S. Amin, M. Andriluka, M. Rohrbach, B. Schiele, Multi-view pictorial structures for 3D human pose estimation, in: Proc. 24^{th} British Machine Vision Conference, Vol. 2, Bristol, United Kingdom, 2013. 6, 10, 11, 18

[61] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: Proc. 24th IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011, pp. 1385–1392. 6, 8, 10, 20

[62] G. Gkioxari, P. Arbeláez, L. Bourdev, J. Malik, Articulated pose estimation using discriminative armlet classifiers, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 2013, pp. 3342–3349. 6

[63] G. Pons-Moll, D. Fleet, B. Rosenhahn, Posebits for monocular human pose estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 2345–2352. 6

[64] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58. 6, 13, 17

[65] H. Yasin, U. Iqbal, B. Krüger, A. Weber, J. Gall, A dual-source approach for 3D pose estimation from a single image, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 7, 8, 17

[66] H. Yasin, U. Iqbal, B. Krüger, A. Weber, J. Gall, Source code for A dual-source approach for 3D pose estimation from a single image, (available online at https://github.com/iqbalu/3D_Pose_Estimation_CVPR2016). 7

[67] S. Li, W. Zhang, A. B. Chan, Maximum-margin structured learning with deep networks for 3D human pose estimation, in: Proc. IEEE International Conference on Computer Vision, Santiago, Chile, 2015, pp. 2848–2856. 7, 8

[68] I. Kostrikov, J. Gall, Depth sweep regression forests for estimating 3D human pose from images, in: Proc. British Machine Vision Conference, Nottingham, United Kingdom, 2014. 7, 9, 17

[69] S. Li, A. B. Chan, 3D human pose estimation from monocular images with deep convolutional neural network, in: Proc. 12th Asian Conference on Computer Vision, Singapore, 2014, pp. 332–347. 7, 8

[70] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, W. Gao, Robust estimation of 3D human poses from a single image, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 2361–2368. 7, 8, 17, 18, 19, 20, 21, 22, 23

[71] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, W. Gao, Source code for Robust estimation of 3D human poses from a single image, (available online at http://idm.pku.edu.cn/staff/wangyizhou/WangYizhou_Publication.html). 7

[72] X. Zhou, S. Leonardos, X. Hu, K. Daniilidis, 3D shape estimation from 2D landmarks: A convex relaxation approach, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 4447–4455. 7, 9, 18, 19, 20, 21, 22

[73] X. Zhou, S. Leonardos, X. Hu, K. Daniilidis, Source code for 3D shape reconstruction from 2D landmarks: A convex formulation, (available online at https://fling.seas.upenn.edu/~xiaowz/dynamic/wordpress/3d-shape-estimation/). 7

[74] I. Radwan, A. Dhall, R. Goecke, Monocular image 3D human pose estimation under self-occlusion, in: Proc. IEEE International Conference on Computer Vision, Sydney, Australia, 2013, pp. 1888–1895. 7, 10, 13, 17

[75] E. Simo-Serra, A. Quattoni, C. Torras, F. Moreno-Noguer, A joint model for 2D and 3D pose estimation from a single image, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 2013, pp. 3634–3641. 7, 8, 17

[76] J. Brauer, W. Hübner, M. Arens, Generative 2D and 3D human pose estimation with vote distributions, in: Advances in Visual Computing, Springer, 2012, pp. 470–481. 7, 8

[77] V. Ramakrishna, T. Kanade, Y. Sheikh, Reconstructing 3D human pose from 2D image landmarks, in: Proc. 12th European Conference on Computer Vision, Springer, Firenze, Italy, 2012, pp. 573–586. 7, 8

[78] V. Ramakrishna, T. Kanade, Y. Sheikh, Source code for Reconstructing 3D human pose from 2D image landmarks, (available online at https://github.com/varunnr/camera_and_pose). 7

[79] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, F. Moreno-Noguer, Single image 3D human pose estimation from noisy observations, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 2673–2680. 7, 8, 16, 17, 20

[80] T. Greif, R. Lienhart, D. Sengupta, Monocular 3D human pose estimation by classification, in: Proc. IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 2011, pp. 1–6. 7, 9

[81] W. Guo, I. Patras, Discriminative 3D human pose estimation from monocular images via topological preserving hierarchical affinity clustering, in: Proc. IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 9–15. 7, 9

[82] J.-B. Huang, M.-H. Yang, Source code for Estimating human pose from occluded images, (available online at https://sites.google.com/site/jbhuang0604/publications/pose_accv_2009). 7

[83] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828. 6

[84] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. 8

[85] Y. Bengio, Learning deep architectures for AI, Foundations and trends in Machine Learning 2 (1) (2009) 1–127. 8

[86] L. Deng, D. Yu, Deep learning: Methods and applications, Foundations and Trends in Signal Processing 7 (3–4) (2014) 197–387. 8

[87] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507. 8

[88] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554. 8

[89] C. Szegedy, A. Toshev, D. Erhan, Deep Neural Networks for Object Detection, in: Proc. Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 2553–2561. 8

[90] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2012, pp. 1097–1105. 8

[91] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the gap to human-level performance in face verification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1701–1708. 8

[92] A. Toshev, C. Szegedy, DeepPose: Human pose estimation via deep neural networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1653–1660. 8

[93] J. Charles, T. Pfister, D. Magee, D. Hogg, A. Zisserman, Personalizing human video pose estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 1653–1660. 8

[94] X. Chen, A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in: Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1736–1744. 8

[95] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1799–1807. 8
[96] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen, et al., Synthesizing training images for boosting human 3D pose estimation, arXiv preprint arXiv:1604.02703. 8
[97] G. Rogez, C. Schmid, Mocap-guided data augmentation for 3D pose estimation in the wild, arXiv preprint arXiv:1607.02046. 8
[98] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, K. Daniilidis, Sparseness meets deepness: 3D human pose estimation from monocular video, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 8, 12
[99] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, C. Theobalt, MARCOnI - ConvNet-based MARker-less motion capture in outdoor and indoor scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence (99) (2016) 1. 8, 14
[100] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, C. Theobalt, Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 3810–3818. 8, 14, 15, 16, 18
[101] C. Hong, J. Yu, D. Tao, J. Wan, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Transactions on Image Processing (2015) 5659 – 5670. 8, 12
[102] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, P. Fua, Structured prediction of 3D human pose with deep neural networks, in: Proc. 27^{th} British Machine Vision Conference, York, UK, 2016. 8, 10
[103] B. Tekin, A. Rozantsev, V. Lepetit, P. Fua, Direct prediction of 3D body poses from motion compensated sequences, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 8, 10, 12, 18
[104] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, J. Xiao, Learning a 3D human pose distance metric from geometric pose descriptor, IEEE Transactions on Visualization and Computer Graphics 17 (11) (2011) 1676–1689. 9
[105] R. Okada, S. Soatto, Relevant feature selection for human pose estimation and localization in cluttered images, in: Proc. 10^{th} European Conference on Computer Vision, Springer, Marseille, France, 2008, pp. 434–445. 9
[106] H. Jiang, 3D human pose reconstruction using millions of exemplars, in: Proc. 20^{th} IEEE International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 1674–1677. 9
[107] R. Urtasun, T. Darrell, Sparse probabilistic regression for activity-independent human pose inference, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1–8. 9
[108] P. F. Felzenszwalb, . B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645. 9
[109] M. Andriluka, L. Sigal, Human context: modeling human-human interactions for monocular 3D pose estimation, in: Articulated Motion and Deformable Objects, Springer, 2012, pp. 260–272. 10, 12, 13
[110] M. Andriluka, S. Roth, B. Schiele, Monocular 3D pose estimation and tracking by detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 623–630. 10, 11, 12, 16, 18, 23
[111] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, K. Daniilidis, Source code for Sparseness meets deepness: 3D human pose estimation from monocular video, (available online at https://fling.seas.upenn.edu/~xiaowz/dynamic/wordpress/). 12
[112] A. Schick, R. Stiefelhagen, 3D pictorial structures for human pose estimation with supervoxels, in: Proc. IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, 2015, pp. 140–147. 12, 16
[113] B. Wandt, H. Ackermann, B. Rosenhahn, 3D human motion capture from monocular image sequences, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 1–8. 12
[114] Y. Tian, L. Sigal, F. De la Torre, Y. Jia, Canonical locality preserving latent variable model for discriminative pose inference, Image and Vision Computing 31 (3) (2013) 223 – 230. 11, 12
[115] M. Yamada, L. Sigal, M. Raptis, No bias left behind: covariate shift adaptation for discriminative 3D pose estimation, in: Proc. 12th European Conference on Computer Vision, Springer, Firenze, Italy, 2012, pp. 674–687. 11, 12
[116] L. Bo, C. Sminchisescu, Twin gaussian processes for structured prediction, International Journal of Computer Vision 87 (1-2) (2010) 28–52. 12, 17, 18, 19, 20, 21, 22
[117] L. Bo, C. Sminchisescu, Source code for Twin gaussian processes for structured prediction, (available online at http://www.maths.lth.se/matematiklth/personal/sminchis/code/TGP.html). 12
[118] J. Valmadre, S. Lucey, Deterministic 3D human pose estimation using rigid structure, in: Proc. 11th European Conference on Computer Vision, Springer, Crete, Greece, 2010, pp. 467–480. 12, 13
[119] J. Valmadre, S. Lucey, Source code for Deterministic 3D human pose estimation using rigid structure, (available online at http://jack.valmadre.net/papers/). 12
[120] I. Rius, J. Gonzàlez, J. Varona, F. Xavier Roca, Action-specific motion prior for efficient bayesian 3D human body tracking, Pattern Recognition 42 (11) (2009) 2907–2921. 12, 13
[121] X. K. Wei, J. Chai, Modeling 3D human poses from uncalibrated monocular images, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 2009, pp. 1873–1880. 12, 13
[122] S. Sedai, M. Bennamoun, D. Q. Huynh, Discriminative fusion of shape and appearance features for human pose estimation, Pattern Recognition 46 (12) (2013) 3223–3237. 11
[123] A. Yao, J. Gall, L. V. Gool, R. Urtasun, Learning probabilistic non-linear latent variable models for tracking complex activities, in: Proc. Advances in Neural Information Processing Systems, Granada, Spain, 2011, pp. 1359–1367. 11, 18
[124] G. W. Taylor, L. Sigal, D. J. Fleet, G. E. Hinton, Dynamical binary latent variable models for 3D human pose tracking, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 631–638. 11, 18
[125] C. H. Ek, P. H. Torr, N. D. Lawrence, Gaussian process latent variable models for human pose estimation, in: Machine Learning for Multimodal Interaction, Springer, 2008, pp. 132–143. 11
[126] E. Simo-Serra, C. Torras, F. Moreno-Noguer, Lie algebra-based kinematic prior for 3D human pose tracking, in: Proc. International Conference on Machine Vision Applications, Tokyo Japan, 2015, pp. 394–397. 13
[127] P. Peursum, S. Venkatesh, G. West, A study on smoothing for particle-filtered 3D human body tracking, International Journal of Computer Vision 87 (1-2) (2010) 53–74. 13
[128] J. Liu, D. Liu, J. Dauwels, H. S. Seah, 3D human motion tracking by exemplar-based conditional particle filter, Signal Processing 110 (2015) 164–177. 13
[129] T. Jaeggli, E. Koller-Meier, L. Van Gool, Learning generative models for multi-activity body pose estimation, International

Journal of Computer Vision 83 (2) (2009) 121–134. 13
[130] L. Sigal, M. J. Black, Predicting 3D people from 2D pictures, in: Articulated Motion and Deformable Objects, Springer, Berlin, Heidelberg, 2006, pp. 185–195. 13
[131] M. Bray, P. Kohli, P. H. S. Torr, POSECUT: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts, in: Proc. 9th European Conference on Computer Vision, Graz, Austria, 2006, pp. 642–655. 13
[132] A. Moutzouris, J. Martinez-del Rincon, J. Nebel, D. Makris, Efficient tracking of human poses using a manifold hierarchy, Computer Vision and Image Understanding 132 (2015) 75–86. 15
[133] S. Amin, P. Müller, A. Bulling, M. Andriluka, Test-time adaptation for 3D human pose estimation, in: Pattern Recognition, Springer, 2014, pp. 253–264. 14, 15, 16
[134] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, N. Navab, Multiple human pose estimation with temporally consistent 3D pictorial structures, in: Proc. 13th European Conference on Computer Vision, ChaLearn Looking at People Workshop, Zurich, Switzerland, 2014, pp. 742–754. 14, 15, 23
[135] L. Sigal, M. Isard, H. Haussecker, M. J. Black, Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation, International Journal of Computer Vision 98 (1) (2012) 15–48. 14, 15, 18
[136] B. Daubney, X. Xie, Tracking 3D human pose with large root node uncertainty, in: Proc. 24^{th} IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011, pp. 1321–1328. 14, 15
[137] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, C. Theobalt, Markerless motion capture of interacting characters using multi-view image segmentation, in: Proc. 24^{th} IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011, pp. 1249–1256. 15
[138] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, H.-P. Seidel, Motion capture using joint skeleton tracking and surface estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 2009, pp. 1746–1753. 15
[139] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3D pictorial structures revisited: Multiple human pose estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence (99). 14
[140] L. Sigal, S. Bhatia, S. Roth, M. J. Black, M. Isard, Tracking loose-limbed people, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, Washington, DC, 2004, pp. I–421 – I–428. 14
[141] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, C. Theobalt, Fast articulated motion tracking using a sums of gaussians body model, in: Proc. IEEE International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 951–958. 14
[142] I. Akhter, M. Black, Pose-conditioned joint angle limits for 3D human pose reconstruction, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, 2015, pp. 1446–1455. 16, 23
[143] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1325–1339. 16, 19
[144] V. Kazemi, M. Burenius, H. Azizpour, J. Sullivan, Multi-view body part recognition with random forests, in: Proc. 24^{th} British Machine Vision Conference, Bristol, United Kingdom, 2013. 16
[145] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 1194–1201. 16
[146] J. Berclaz, F. Fleuret, E. Tretken, P. Fua, Multiple object tracking using k-shortest paths optimization, IEEE Transactions on Pattern Analysis and Machine Intelligence (2011) 1806 – 1819. 16
[147] N. Van der Aa, X. Luo, G. Giezeman, R. Tan, R. Veltkamp, Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction, in: Proc. IEEE International Conference on Computer Vision Workshops, Barcelona Spain, 2011, pp. 1264–1269. 16
[148] M. Tenorth, J. Bandouch, M. Beetz, The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition, in: Proc. IEEE 12^{th} International Conference on Computer Vision Workshops, Kyoto, 2009, pp. 1089–1096. 16
[149] M. Hofmann, D. M. Gavrila, Multi-view 3D human pose estimation combining single-frame recovery, temporal integration and model adaptation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, 2009, pp. 2214–2221. 16, 24
[150] MoCap: Carnegie Mellon University Graphics Lab motion capture database. (available online at http://mocap.cs.cmu.edu). 16, 17
[151] N. R. Howe, A recognition-based motion capture baseline on the HumanEva II test data, Machine Vision and Applications 22 (6) (2011) 995–1008. 16
[152] C. Ionescu, F. Li, C. Sminchisescu, Latent structured models for human pose estimation, in: Proc. 13^{th} IEEE International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2220–2227. 16
[153] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Progressive search space reduction for human pose estimation, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Anchorage, AK, 2008, pp. 1–8. 16
[154] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, Washington, DC, 2004, pp. II–882. 17
[155] B. Boteanu, N. Sarafianos, B. Ionescu, I. Kakadiaris, Synpose300 dataset and ground truth, (available online at http://imag.pub.ro/~bionescu/index_files/Page12934.htm). 18
[156] MakeHuman, Makehuman open source software for the modelling of 3-dimensional humanoid characters (available online at http://www.makehuman.org/download_makehuman_102.php). 19
[157] Blender Foundation, Blender open source 3D creation suite (available online at http://download.blender.org/release/). 19
[158] SAE International, CAESAR: Civilian American and European Surface Anthropometry Resource database. (available online at http://store.sae.org.caesar). 19
[159] M. Hofmann, D. M. Gavrila, Multi-view 3D human pose estimation in complex environment, International Journal of Computer Vision 96 (1) (2012) 103–124. 19
[160] D. Ramanan, Learning to parse images of articulated bodies, in: Proc. Advances in Neural Information Processing Systems, Vancouver, Canada, 2006, pp. 1129–1136. 20