A review of 3D human pose estimation algorithms for markerless motion capture

Yann Desmarais, Denis Mottet, Pierre Slangen, Philippe Montesinos

A review of 3D human pose estimation algorithms for markerless motion capture

Yann Desmarais[a], Denis Mottet[b], Pierre Slangen[a], Philippe Montesinos[a,**]

[a] *EuroMov Digital Health in Motion, Univ. Montpellier, IMT Mines-Ales, 30100 Ales, France*
[b] *EuroMov Digital Health in Motion, Univ. Montpellier, IMT Mines-Ales, 34090 Montpellier, France*

ABSTRACT

Human pose estimation is a very active research field, stimulated by its important applications in robotics, entertainment or health and sports sciences, among others. Advances in convolutional networks triggered noticeable improvements in 2D pose estimation, leading modern 3D markerless motion capture techniques to an average error per joint of 20 mm. However, with the proliferation of methods, it is becoming increasingly difficult to make an informed choice. Here, we review the leading human pose estimation methods of the past five years, focusing on metrics, benchmarks and method structures. We propose a taxonomy based on accuracy, speed and robustness that we use to classify de methods and derive directions for future research.

1. Introduction

Human Pose Estimation is the extraction of body configurations in images or videos. Typically, it is the inference of joint coordinates and the reconstruction of a human skeletal representation. In the last few years, 2D pose estimation reached detection rate above 90% on all different human joints Newell et al. (2016). This progress has been possible in great part because of the success of convolutional neural networks (CNN) and the appearance of accessible large scale datasets (Sigal et al. (2010); Ionescu et al. (2014)). However, it is only recently that these new architectures have been deployed to solve a similar problem in 3D. The challenge for these new 3D markerless pose estimation methods is to be competitive against classical techniques and marker-based motion capture systems. The ultimate goal would be a complete and accurate 3D reconstruction of an individual's motion from simple monocular images with tolerance to severe occlusion. As this ideal is unrealistic, results on similar tasks seem to indicate that it is possible to reach some of these conditions even if all are not fulfilled.

Traditionally, commercial motion capture systems track small reflective markers placed on the surface of subjects. While precise, traditional motion capture is heavily constrained by complex sensors and acquisition environment. In 2010, Microsoft released the Kinect sensor that captured human pose from RGB and depth images. However, it was mostly aimed at entertainment usage and was not suitable for outdoor acquisition.

However, pose and person detection algorithms based only on image features have existed since a long time. The former paradigm was the fitting of human defined features to complex part-based human models (representations of the silhouette with cylinders, stick-figures, meshes, cones or boxes). While some modern techniques still use that approach, the feature extraction part of the process is now realized using convolutional neural networks.

1.1. Other Surveys

Bray (2000) reviews optical markerless methods with a taxonomy based on commonly performed subtasks: Initialization, Tracking, Pose Estimation and Recognition (Fig.1). With this classification, they describe the different ways to extract visual features for the pose estimation process and then how to track them between frames.

**Corresponding author: Tel.: +33-(0)4-34-24-62-95
e-mail: philippe.montesinos@mines-ales.fr (Philippe Montesinos)

The survey of Moeslund and Granum (2001) explores twenty years of vision-based human pose estimation techniques from 1980 to the early 2000s, including marker-based and markerless methods. At that time, many assumptions were taken to facilitate the process of extracting the human silhouette: most methods functioned indoor with non-shifting lighting and nearly half of them used uniform static backgrounds. This survey provides a good history of the different families of pose estimation. The authors also carefully describe the different degrees of performances needed for applications using human pose estimation and how to quantify them. However, their functionality-based taxonomy is no longer adapted for today's techniques, because modern methods mostly do not use an initialization step and perform pose estimation and tracking at the same time. Finally, action recognition is nowadays a computer vision task with its own community.

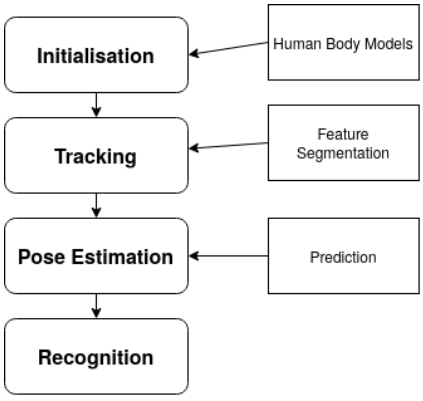

Fig. 1. Classical motion capture system

Sarafinos et al. (2016) wrote a review of more recent methods with an emphasis on input data. They suggest a taxonomy that divides pose estimation between monocular versus multi-view techniques and between static image versus video inputs. This distinction is still present in our review, but our analysis is more oriented towards the family of method employed (learning, model-based etc.). They also evaluate part-based and learning-based methods on modern benchmarks. They also propose a synthetic dataset (SynPose300) to evaluate robustness (erroneous initialization, viewing distance, difficult poses or actions). The focus is made on methods from 2008 until 2016. We chose to start our review from 2017, to present recent evolutions in 3D pose estimation.

Colyer et al. (2018) examined the historical methods used in biomechanical studies of human motion. The authors insist that training and validating markerless methods with marker-based ones is not completely correct, because the reflecting markers can modify the results (helping or degrading the results otherwise obtainable in an "in-the-wild" context). They also state that marker-based methods are not providing the accuracy of measurement to characterize real human movement for some sport science and biomechanical experiments. However, more precise methods are invasive and impractical, therefore most research is still conducted with optoelectronic commercial systems (which also have some inaccuracies). Furthermore, they compare the accuracy of markerless methods in the HumanEva benchmark Sigal et al. (2010) stating that the precision required for most analysis in sport science is not yet achieved with the current markerless algorithms. While giving a good overview of the field, this survey does not discuss more recent methods (Pavllo et al. (2019); Cheng et al. (2020)) that produce results with five times better accuracy of their best reviewed methods. It also gathers results from an older dataset and not on the more recent ones such as Human3.6M that contains more images with better resolution.

Finally, Chen et al. (2020) detailed recent monocular 2D and 3D pose estimation techniques that are based on deep learning. They explain in detail the different categories of methods that currently exist and the evaluation protocols and metrics for this task. More than twenty methods are reviewed from 2014 and forward. Our review differs as it focuses on 3D pose estimation, but also considers multi-view settings as well as techniques employing different sensors such as inertial measurement units (IMU).

1.2. Proposed Approach

Moeslund and Granum (2001) adopts three main criteria to evaluate pose estimation systems :

- Accuracy
- Speed
- Robustness

The relevance to different fields (surveillance, control or analysis) is then studied. Despite the needs of these domains of application having evolved, this analysis is still relevant. In the present review, our main objective is to guide the choices of developers, engineers and researchers who want to build upon the reviewed algorithms. We compare recently published academic methods with regard to their relevance to different application fields. First, we describe the metrics and benchmarks that are commonly used for method evaluation in section 2. Then, we detail and explain the families of architectures used for human pose estimation in section 3. Next, we present the current state-of-the-art techniques for 3D markerless pose estimation with an emphasis on carefully selected articles in 4. Finally an overall analysis of accuracy, robustness and speed performance indices is available in section 5.

2. Methods Evaluation

This section describes how to evaluate markerless pose estimation methods in 3D. To compare these methods, we use evaluations provided across multiple benchmarks datasets. The designers of these datasets often recommend different metrics. We start by describing how these metrics are computed and discuss their relevance in different contexts. For each of these benchmarks the quantity of images, the environments and acquisition modalities are also detailed.

2.1. Metrics

Several metrics measure the accuracy of pose estimation algorithms in 3D. Some are processing the average error, other a detection rate with a predefined threshold and finally some uses perceptual or structural criteria. In this subsection, these metrics are described and their strengths and fail cases are discussed in different contexts.

MPJPE: Mean Per Joint Position Error. It is one of the most frequently used metrics found in the literature. It is also sometimes referred to as mean reconstruction error or 3D error. MPJPE is the mean of the Euclidean distances between the estimated coordinate and ground truth coordinate over each joint :

$$MPJPE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} ||m_{\hat{x}}(i) - m_x(i)|| \tag{1}$$

N represents the number of processed joints, $m_{\hat{x}}(i)$ is the function estimating the ith joint coordinates and m_x is the ground truth position of the joint. Equation 1 is the measurement of MPJPE for one frame and one skeleton. To generalize for video frame the average of each MPJPE's frame of the sequence is calculated.

MPJPE is a good baseline metric that can be used to evaluate a wide variety of methods as long as they want to estimate coordinate position and overall skeleton structure. It can also be adapted to evaluate methods that do not estimate the same number of key points as the number of markers used in common datasets and methods estimating relative poses instead of absolute 3D position Sigal et al. (2010). This is performed by Procrustes alignment (adjustment to the ground truth poses) using a chosen root joint such as the pelvis. It can sometimes be referred as N-MPJPE or P-MPJPE depending if the alignment is by scale only or also by rotation and translation. The main drawbacks of such metrics, identified by Ionescu et al. (2014), is their low robustness to outlier errors and the fact that they can be influenced by perceptually irrelevant variations.

MPJVE: Mean Per Joint Velocity Error. Introduced by Pavllo et al. (2019), this metric can be used when a pose sequence is extracted from a video. Here, the absolute position

of joints obtained from the MPJPE is insufficient. For this reason, the author use "the MPJPE of the first derivative of the 3D pose sequences" to "measure the smoothness of predictions over time". This technique is useful to compare estimation models that are using temporal data.

Angular Metrics: Another approach would be to measure angle errors of joint segments. Ionescu et al. suggest the Mean Per Joint Angle Error (MPJAE) Ionescu et al. (2014) which is also sometimes referred as Mean Angular Error Agarwal and Triggs (2006):

$$MPJAE(x, \hat{x}) = \frac{1}{3N} \sum_{i=1}^{3N} |(m_{\hat{x}}(i) - m_x(i) mod \pm 180)|$$

Here m_x and $m_{\hat{x}}(i)$ refer to 3D angles for ground truth and prediction, respectively. These metrics can be used when the main analysis is performed on angles between two specific limbs rather than the whole body such as for rehabilitation, or sport motion. However, authors of Ionescu et al. (2014) report that this metric can have little perceptual meaning. It can also be difficult to interpret because angles are calculated locally and joints that are dependent from a faulty predicted one can yield no errors despite a globally misaligned skeleton. As a result, these metrics are less used in recent computer vision publications.

Thresholds Metrics: A common approach in 2D Human Pose Estimation and other detection tasks is to define a threshold where a key point is correctly detected. Then, statistics of correctly predicted joints over a set of images can be computed. The percentage of correct parts (PCP) uses half the size of the ground truth segment to determine if a prediction over a limb segment is correct. A 3D version of PCP can also be adapted. A limb is correctly estimated if the following expression is respected:

$$\frac{||\alpha - \hat{\alpha}|| + ||\omega - \hat{\omega}||}{2} \leq \theta ||\alpha - \omega||$$

α and ω are the two measured coordinates of the extremities of the limb and $\hat{\alpha}$ and $\hat{\omega}$ their predictions. θ is a chosen parameter to control the accuracy requirement for the threshold (commonly 0.5). This metric was used to evaluate model-based pose estimation using pictorial structure such as Burenius et al. (2013). The issue with this metric is that a shorter limb will be less likely to be considered detected as the threshold decreases.

Another threshold metric used in 2D pose estimation is the Percentage of Correct Keypoints (PCK). This metric does not have the issue of shorter limbs harder to detect as it uses a subject specific threshold for each individual joint instead of limbs. It is calculated with a portion of a fixed limb length (eg: 0.5 times the head bone length, often referred as PCKh@0.5). In this way, the metric is self-adapting to subjects with different proportions, without bias on specific size of the limbs of an individual. Mehta et al. (2016) propose a 3D version of the PCK used as the main evaluation metric for the MPI-INF-3DHP benchmark. A joint is considered detected with this condition:

$$||\alpha - \hat{\alpha}|| \leq \theta ||k - h||$$

α and $\hat{\alpha}$ are the target joint and its prediction. θ is a parameter controlling the fraction of a reference limb length, k and h are the coordinates of the extremities of this limb (head, torso...). Another solution is to choose a fixed threshold of 150 mm, which loses the specificity of the subject.

Using these threshold-based metrics is justified when comparing methods that could have a good overall accuracy, but produce errors in specific scenario (for singular joints or skeletons). However, this is done at the price of losing sensitivity that could be relevant when analyzing precise local coordinates (at the millimeter scale), as in biomechanical applications.

Volume & Surface Based Metrics: Some techniques for human pose estimation need a measurement over surfaces. This type of metric can be found in dense pose estimation Densepose (Güler et al. (2018)). This task aims at recovering the surface of

Dataset	#Frame #Video Sequence	#Subject	#View	Resolution	Frequency	Depth	IMU	Context and Main Characteristics
Human3.6M: Ionescu et al. (2014)	3.6 millions 1 376	11	4	1000x1000	50Hz	Yes	No	Lab environnement
HumanEva: Sigal et al. (2010)	80 000 56	4	7	660x500	60Hz	No	No	Lab environnement
Total Capture: Trumble et al. (2017)	1.9 millions N/A	5	8	1920x1080	60Hz	No	Yes	Lab environnement Inertial Measurement Units
MPI-INF-3DHP: Mehta et al. (2016)	1.3 millions 64	8	14	N/A	N/A	No	No	"In the wild" & Lab outdoor/indoor green screens. Markerless ground truthes
MuPoTS-3D (2018): Mehta et al. (2018)	8 000 20	8	1	2048x2048 1920x1080	30Hz - 60Hz	No	No	Multi-person, "In the wild" indoor/outdoor scenes. Markerless ground truthes
3DPW von Marcard et al. (2018)	> 50000 60	7	1	N/A	30Hz	No	Yes	"In the wild" outdoor single moving camera & IMUs Up to two subjects
Carnegie Mellon University Mocap	N/A 2 605	109	1	352x240	30Hz	No	No	Indoor environment Various actions and subjects
CMU-MMAC: De la Torre et al. (2008)	≈450 000 N/A	25	5	1024x768 (x3) 640x480 (x2)	30Hz (x3) 60Hz (x2)	No	Yes	Lab Environnement Subjects cooking 5 recipes
TNT15: von Marcard et al. (2016)	13 000 20	4	8	800x600	50Hz	No	Yes	Office environment No marker-based labeling, only IMU
AMASS: Mahmood et al. (2019)	N/A (>40 hours)	346	variable	variable	variable	No	No	Unified parametrization of 15 datasets Mesh body models
MoVi: Ghorbani et al. (2020)	N/A (17 hours)	90	4	800x600 1920x1080	30Hz	No	Yes	Synchronized MoCap shape, video and IMU data

Table 1. Popular datasets used to compare, train and test human pose estimation models. Video frames, the number of subjects and actions give an indication about the dataset diversity and the number of pose configurations. The number of views from RGB cameras, the resolution and acquisition frequency of cameras assess the quality and quantity of exploitable video information. Inertial Measurement Units (IMU) are sometimes used to refine results from the motion capture or single-image detection. If not specified, the motion capture method is marker-based.

the whole human body, not just a few joint key points. Geodesic distance-based metrics are often used in this context. An example would be the Geodesic Point Similarity described in Güler et al. (2018):

$$GPS(j) = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2k^2}\right)$$

P_j is a set of points representing the body surface of the jth one person. $g(i_p, \hat{i}_p)$ is the geodesic distance calculated between the estimated point and the ground truth one. A GPS score of 0.5 indicates that this distance is equal to half a predefined distance adjustable with the k parameter (often setup to be a fraction of a joint segment).

3D human shape tracking is another variant of the task that reconstruct and track the human body volume frame by frame in a video. A common approach uses the iterative closest point (ICP) algorithm to fit image data to a model. Huang et al. (2018) suggest using random forests and nearest-neighbor matching with two volumetric features based on voxels and centroïdal Voronoi tessellation instead of ICP.

Lastly, another popular family of methods is using multi-view data and shape-from-silhouette techniques to create volumetric representations of the human body. This shape can then be useful for joint location prediction. These methods produce probabilistic visual hulls (PVH) (Grauman et al. (2003))

in voxel grids. Even though human body shape estimation is not human pose estimation, it is a close task that can be used at different stages of a modular motion capture system. With multi-view datasets it is easy to obtain ground truths PVH that can then be used to evaluate 3D reconstructed volumes (e.g. Trumble et al. (2017)). Here Means Squared Error can be calculated from the voxel grid.

Each of these metrics can be used in specific variations or edge cases of the task. For "classical" human pose estimation, MPJPE seems more popular as it is simple and no extra parameters intervene in its computation. However, some published articles (Ionescu et al. (2014); Mehta et al. (2016)) claim that threshold metrics are better at identifying errors in specific joints and less prone to penalize perceptually irrelevant errors. Finally, to evaluate methods that process videos and produce 3D pose sequences, the MPJVE is a good alternative to highlight techniques that produce more realistic human motions.

Furthermore, the metrics described above only express physical accuracy in multiple ways, with threshold-based ones sometimes introducing perceptual parameters. However, depending on the use case, it might be pertinent to take into consideration more complex perceptual metrics Marinoiu et al. (2016), Marinoiu et al. (2013) or structural metrics Kocabas et al. (2019b). They can help when purely positional information produce the same error score for two different predicted poses. Significant work has been produced on the way human are perceiving what is a valid and realistic human body configuration. These metrics can be useful in fields that are not concerned about the biological and physical constraints, but more about pose semantic.

2.2. *Commonly used Benchmarks*

Collecting accurate data for human pose estimation is a long and complex process that is driven by progress in acquisition technologies. Moreover, several specific choices are needed concerning the sensor modality, quantity and the acquisition protocol.

The complexity of this task explains why it has taken time for the scientific community to create large benchmarks: today many variations exist between monocular versus multi-view, laboratory controlled versus in-the-wild environments (see Table 1) etc. With new commercial solutions starting to produce results similar to traditional motion capture some benchmarks are also starting to use markerless labeled ground truths (ie: Theia: Kanko et al. (2020), The Captury). They have the advantage of easily providing in-the-wild images. However, using this kind of data as ground truth can be questioned as it is itself obtained using methods that are not always available and transparent. Despite this diversity, there are still only a few openly accessible academic benchmarks containing more than millions of images.

Datasets have an important role as they are used to validate and test algorithms, but also to train and fine-tune deep learning models. Providing large, high-quality datasets with excellent labels and a wide variety of poses is a major challenge. Three references historically best meet these criteria: HumanEva I and II Sigal et al. (2010), Human3.6M Ionescu et al. (2014) and Total Capture Trumble et al. (2017). They contain video sequences with multiple view angles associated with ground-truth joint coordinates. However, these large-scale references are mainly captured in controlled laboratory environments with marker-based systems, hence with a limited variety of backgrounds, poses and subjects.

Other benchmarks focus on pose and subject diversity, or on in-the-wild environment acquisition. While interesting to experiment with for human pose estimation or its subtasks, they are not providing the same quantity and variation to conduct large-scale evaluations. However, a noticeable exception is MPI-INF-3DHP Mehta et al. (2016), which is more and more used to benchmark algorithms as it proposes a high variety

of contexts and subjects (using "green screen" and outdoor acquisition) with a significant amount of data.

Marker-based motion capture is the easiest way to obtain something close to ground-truth data. However, old datasets have a low image resolution and some have inaccurate annotation for some subjects (reported by Iskakov et al. (2019)). Additionally, Colyer et al. (2018) noted errors of about 10mm or 10° compared to intrusive methods closer to the real human anatomy (ie: intra-cortical bone pins). This is because the key points are reconstructed from groups of markers placed on the subjects' skin or clothing (i.e., soft surfaces).

Some datasets propose 3D mesh-based models of the human body (Mahmood et al. (2019); Ghorbani et al. (2020)). These representations can be used as learning targets for estimation algorithms that care about richer information than skeletal representation of the poses and joint positions. They are also useful for applications that need fully-rigged models such as animation. The AMASS dataset Mahmood et al. (2019) unifies several motion capture datasets by providing the 3D representation of subjects. This is done by computing body poses using the SMPL Loper et al. (2015) model with their regression method (MoSh++). Furthermore, Ghorbani et al. (2020) provide a new dataset with motion capture (MoCap) and inertial motion unit data, added to AMASS.

3. Architectures for Human Pose Estimation

In this section, the main 3D pose estimation families of methods will be described. They can be classified as methods using human body models, learning algorithms or geometric information. In the case of a neural network learning approach, backbone networks are employed and new loss functions are created. Table 2 is the complete taxonomy of all discussed methods according to these criteria. In the second part of this section, we summarize the most commonly used architectures for 2D and 3D pose estimation.

3.1. 3D Pose Estimation Taxonomy

Human Body Models

Historically, pose estimation algorithms were relying on part-based or **skeleton models** of the human body. Each node represented a joint and vertices limb length and orientation. One example of this kind of approach is the Pictorial Structure Model (PSM) introduced by Fischler and Elschlager (1973) and used for pose estimation in Felzenszwalb and Huttenlocher (2005), Marinoiu et al. (2013) and Belagiannis et al. (2014). Originally, this model is devised as the minimization of an energy function. The cost function combines an error term for joint location error and a penalty for segment length deformations (i.e. not corresponding to limb size).

Several methods adopt **kinematic** based human skeleton, where each linked joint pair is represented as a vector. With this technique, angular and length constraints can be applied to detect poses. Kinematic Chain Space proposed by Wandt and Rosenhahn (2019) is an example of such techniques.

Mesh models consists of complete reconstruction of the human body surface. These models offer a richer information than skeleton-based models and can be used to infer the spatial representation of a subject in a virtual scene or to render captured pose into fully-rigged meshes for animation. SMPL Loper et al. (2015) is the most frequently used human mesh model for pose estimation. The mesh is learned from numerous 3D body scans and can be adapted to a set of pose and shape parameters produced by pose estimation algorithms.

3.1.1. Geometric Information

When several cameras are available, multi-view geometry is frequently used for 3D pose estimation. One way to infer joint coordinates in three dimensions is to use **triangulation** with their 2D image coordinates in each view. Depending on the calibration and availability of camera extrinsic and intrinsic parameters, different reconstruction schemes are possible.

			Human Body Models			Neural Networks					
Method	Proxy Representation	Losses	Kinematic Chains	Skeleton (e.g. PSM)	Mesh (e.g. SMPL)	Backbone	GAN Adv.lea.	RNN LSTM	TCN	Attention	GCNN
Pavlakos et al. (2017)	2D Heatmaps	L2				SHNet					
Mehta et al. (2017)	2D heatmaps, "Location maps"	L2	•			Resnet50					
Zhou et al. (2017)	2D heatmaps	L2, "geometric loss"				SHNet					
Martinez et al. (2017)	2D pose	L2				SHNet (2D) + MLP					
Sun et al. (2018)	2D/3D heatmaps	any heatmap losses				Resnet models and SHNet tested					
Omran et al. (2018)	part segmentation map	3D and 2D joint loss (L2) 3D latent parameter loss (L1)			•	RefineNet Lin et al. (2016) + Resnet50					
Mehta et al. (2018)	"Occlusion Robust Pose Maps" Part affinity fields	L2	•			ResNet50					
Kolotouros et al. (2019)	N/A	L2			•	ResNet50					
Wandt and Rosenhahn (2019)	N/A	"Reprojection loss" Wasserstein loss, Camera loss	•			SHNet (2D)	•				
Xu et al. (2019)	pixel-to-surface maps	"render-and-compare loss" reconstruction loss (L2), parameter loss (L2)			•	Resnet50	•				
Kocabas et al. (2019b)	2D pose	smooth L1				Resnet50					
Mathis et al. (2018)	N/A	L2				Resnet50 Resnet101 tested					
Mehta et al. (2020)	"3D pose encoding" Part affinity fields	smooth L1	•			"SelecSLS Net" Fully connected					
Hossain and Little (2018)	2D pose sequence	L2 derivative loss on joint sets				SHNet (2D)		•			
Cai et al. (2019)	2D pose, ST-graph	L2, "symmetry loss" derivative loss on joint sets				CPN					•
Pavllo et al. (2019)	2D pose	trajectory and pose loss bone length L2 loss, 2D projection loss				SHNet, CPN and Mask-RCNN tested (2D)			•		
Cheng et al. (2019)	2D pose	L2, 2D projection loss		•		SHNet (2D)	•		•		
Cheng et al. (2020)	2D heatmaps	L2, "multi-view loss" 2D projection loss	•			HRNet (2D)	•		•		
Liu et al. (2020)	N/A	N/A				SHNet and CPN tested (2D)			•	•	
Wang et al. (2020)	2D pose, ST-graph	L2, "motion loss"				CPN and HRNet tested (2D)					•
Qiu et al. (2019)	2D heatmaps	L2		•		SimpleNet					
Iskakov et al. (2019)	2D heatmaps	soft L2, L1 regularized				SimpleNet					
He et al. (2020)	2D pose	L2				SimpleNet				•	
von Marcard et al. (2016)	multi-view silhouettes IMU orientations	N/A	•		•	N/A					
Trumble et al. (2017)	PVH, IMU orientations then 2D coordinates	L2				Classical 3D CNN		•			
von Marcard et al. (2018)	2D pose, IMU orientations	N/A	•		•	Cao et al. (2017) (multi-person)					
Huang et al. (2019)	"multi-channel volume"	L2				SHNet (3D Conv)					
Zhang et al. (2020)	2D heatmaps	N/A		•		SimpleNet					

Table 2. The taxonomy of reviewed methods. In case of multiple stages we indicate intermediate representations. For learning methods, loss functions and backbone architecture are indicated when they are present (for many two-stage methods these backbones concern only the first stage of 2D detection). Backbones are referred to as *SHNet*: Stacked Hourglass (Newell et al. (2016)), *CPN*: Cascaded Pyramid (Chen et al. (2018)), *HRNet*: High Resolution Network (Sun et al. (2019)) and *SimpleNet*: Simple Baselines (Xiao et al. (2018)).

Another approach consists of **fusing features** from different views along epipolar lines before inferring poses. The epipolar line is the image in one camera of a ray passing through the optical center of the other camera and a point in the scene. Considering a point in the first view, its corresponding point in the second view is guaranteed to lie on the epipolar line in the other image. Using this prior information about the different views, the 2D pose can be refined or multi-view features can be merged before the 3D pose itself is estimated.

Finally, other methods use shape-from-silhouette reconstruction to obtain the whole body shape before joint detection. These techniques first segment human shapes from each view and then reconstruct their volume (an example of these methods are probabilistic visual hulls from Grauman et al. (2003)). These volumes can then be used as intermediate features.

Learning Approaches

Since 2014 with Toshev and Szegedy (2014) and Tompson et al. (2014) convolutional neural networks were extensively used for human pose estimation. However, new architectures and training reformulations of their different building blocks are presented each year. These variants sometimes contribute to the improvement of the state-of-the-art for 3D pose estimation. Here, we review the most commonly used families of networks for this task.

The first thing to consider is the design choices regarding the convolution operations. For 3D pose estimation, many variants are employed according to the data representation that is used or the hypothesis of the authors. The choices range from **2D convolutions** on image data to convolutions on spatial-temporal graph representing a subject motion. Most monocular methods commonly use classic convolutions. For video sequences, multi-view or multimodal setups, richer information have inspired different techniques. Volumetric intermediate features computed from multi-view can be fed to **3D convolution** networks to refine or estimate the pose. **Temporal convolutions** can be used to reason on past and future frames and help better characterization of the pose.

Sometimes, the location of the pose and the trajectory in time are encoded by spatial-temporal graph that can be computed with **Graph Neural Network** (GNN). A spectral convolution can then be applied in the Fourrier domain using the eigenvalues of the Laplacian graph Kipf and Welling (2017).

Recurrent Neural Networks (RNN) are another way to process pose sequences. More specifically **Long Short-Term Memory** (LSTM) architectures have been used with success on 2D joint sequences (e.g. Hossain and Little (2018)) and other modalities (e.g. Trumble et al. (2017)). This technique showed success in text translation and other tasks to process sequences with long-term dependencies. From one sequence, LSTM can output another one keeping information about previous inputs passed successively (sequence-to-sequence model). The main difference between LSTMs and RNNs is the the state of their cells that is updated by different linear operations called ”gates.” These operations select and update information that are useful to remember (this website from Christopher Olah details LSTM functioning).

Attention mechanism aims to focus networks on the most important information for pose estimation in the input data or intermediate features. Cai et al. (2019) use attention to select the frames that contribute the most to the estimation, whereas He et al. (2020) describe an ”Epipolar Transformer” module taking advantage of multi-view to focus on learning across epipolar lines. Features in the paired image along the epipolar lines corresponding to the joint points are fused.

Finally, another interesting line of research concerns generative adversarial networks and **adversarial learning** (Goodfellow et al. (2014)). It has mainly been used for 3D pose estimation in two ways: unsupervised mapping of 2D to 3D poses Kudo et al. (2018) and more frequently as a pose validation

module (Cheng et al. (2020); Wandt and Rosenhahn (2019); Kocabas et al. (2019a)). In this case, a discriminator network improves pose consistency by being trained to recognize generated poses from the ones directly extracted from ground-truths. An adversarial loss component is then propagated to the generator network which estimates 3D human poses from visual information or 2D poses. Thus, poses that are inconsistent with known configurations are penalized.

3.2. Backbone Architectures

Before reviewing state-of-the-art algorithms it is important to consider the backbone architectures that are commonly used for 2D or 3D pose estimation. They are extensively used in "top-down" approaches before any computation reducing the problem to a mapping of 2D to 3D coordinates. They reason from low-level joint coordinates to infer high-level information about the human skeleton (see Fig. 3). In opposition, "bottom-up" techniques reason on human body models and extract features from the images after fitting them to the data. These backbone architectures are also employed directly for 3D pose estimation. For example, Huang et al. (2019) use Hourglass Networks including 3D convolution on volumetric representations. Space is missing to explain the backbones architectures methods in details, but the reader can refer to Xiao et al. (2018), Newell et al. (2016) or Chen et al. (2018). In the present paper, we limit ourselves to the general principles of commonly used backbone architectures and provide an overview of their structure and performance.

Two commonly used architectures are hourglass networks Newell et al. (2016) and cascaded pyramid networks Chen et al. (2018) (Fig. 2, (b) and (c)). They both compute image features at different resolution level with the key idea to encompass global and local discriminant features. Hourglass networks are composed of stacked hourglass modules : the first part of the module is reducing the resolution as it passes through convolutional layers. The second part up-samples the features while summing them with corresponding ones of the same dimension from the previous stage. Intermediate supervision is conducted at the end of each module. Cascaded pyramid networks are a two-step architecture predicting poses from a feature pyramid network. It then refine the results for hard-to-predict points. The pyramid network fuses the features at different resolutions to produce joint position heat maps. In the second stage, the refinement process is using intermediate features from the pyramid at different levels. They are up-sampled and summed before going through a final convolutional block. This process is done only for "difficult to predict key points" (chosen with the loss from the pyramid network). These two architectures are using residual connection modules from He et al. (2015) as "building blocks". Based on skip connection between convolutional layers, this technique is widely used in many computer vision tasks for feature extraction. Some 2D pose estimation systems are directly using one of the original variants of this network (ie: Resnet50, Resnet101 etc).

High resolution networks Sun et al. (2019) are considered as another architecture exploiting different resolutions of the same image. Here, it processes resolutions in parallel with convolution layers sharing weights through "exchange blocks."

The efficient 2D multi-person detector from Cao et al. (2017) is sometime used. It uses "part affinity fields" which is "a non-parametric representation of relationship between body parts". From these features and joint localization confidence maps, human poses are predicted with correct associations to multiple subjects. Another asset of this method is that it runs in real-time.

Finally, Xiao et al. (2018) propose a baseline method also using Resnet as its base. This technique is using deconvolution layers to produce heat maps from deep image features, without a specific procedure for difficult-to-predict points. Despite its simplicity, this model achieves competitive results at an efficient computational cost. It is therefore used for comparison evaluations but also sometimes as a backbone 2D detector.

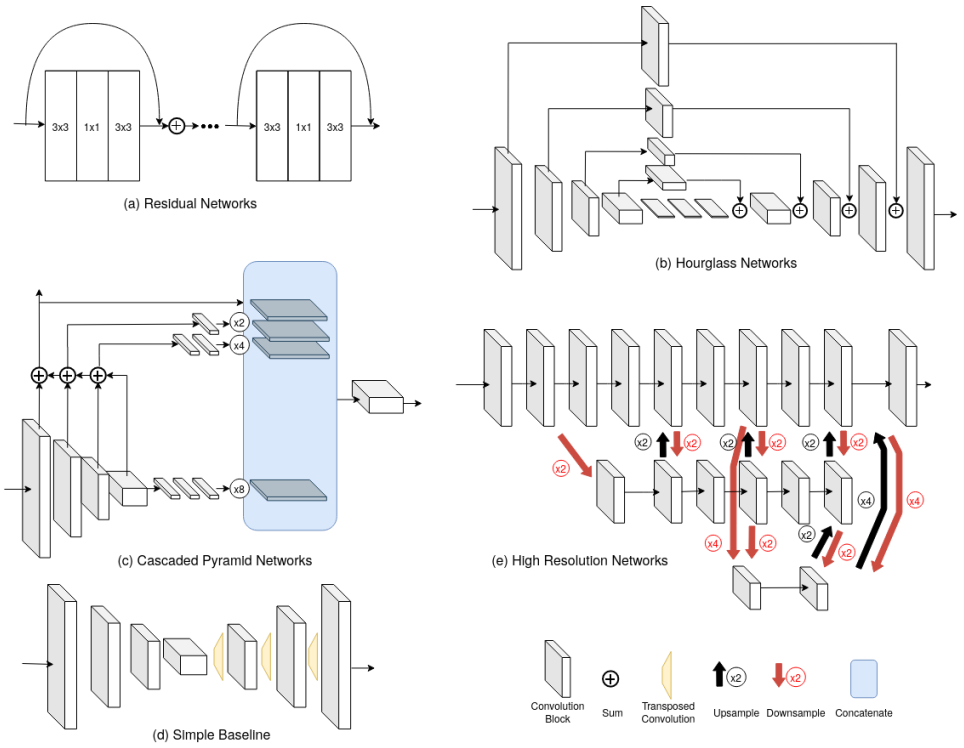

Fig. 2. Backbone architectures used for 2D pose estimation. (a) Residual blocks are the main characteristic of ResNet variants (Resnet101, Resnet50, Resnet152) He et al. (2015). They are also present in more "human pose estimation" specialized models: (b) Stacked Hourglass networks Newell et al. (2016) and (c) cascaded pyramid networks Chen et al. (2018). The simple baseline (d) network proposed by Xiao et al. (2018) uses transposed convolutions to recover the higher input resolution. (e) Higher resolution network (HRNet) Sun et al. (2019) processes high and low resolutions features in parallel within sub-networks that share information.

4. Methods Review

This section describes and compares the top performing methods for vision based 3D markerless pose estimation (see Tables 3, 4 and 5). The selection process was as the follows :

- Top methods from the state-of-the-art on each most popular benchmark
- Most cited methods in computer vision and fields of application for 3D human pose estimation (biomechanics, robotics, sport sciences, human-machine interaction etc.)
- Recent methods reporting interesting results or original approaches that can further advance research

The first observation regarding the evolution of accuracy in the state-of-the-art is that it is rapidly improving. This task is getting attention and, every year, new records are reached. The average error dropped from about 100mm to less than 20mm within 10 years. It is yet to be determined whether accuracy is still going to increase in the future. However, given the diversity of approaches and modalities, it can be argued that, to date, there is no consensus on the best method to use.

In this review, we focus on the best performing markerless techniques currently available, regardless of the sensors or algorithms they use. Methods using monocular images and video sequences are first presented. Second, we present methods using multiple views. Finally, we include methods using other modalities as a pre-processing step, or during the prediction, mainly inertial measurement units. An overall analysis of the performance indices of the different families of methods will be conducted in the following sections.

4.1. Monocular Images

Pavlakos et al. (2017) present one of the first methods to propose a one-stage end-to-end convolutional neural network to predict 3D human pose from a single RGB image. They do so by focusing on the 3D nature of the task. Their architecture uses stacked hourglass modules Newell et al. (2016) (see 2D architectures 3.2) that outputs volumes of voxel probability for each joint in a 3D discretized space around the target. They also propose a new intermediate supervision method inspired from success in 2D human pose estimation. This original method does not use a 2D joint estimation step. Instead, they employ a coarse-to-fine approach that leverage 2D heatmaps as ground truths for intermediate supervision, and then fuse them with image features as an output of the next modules. Further down, the network reconstructs 3D voxel maps. The supervision is also done using 3D Gaussian around the given 3D coordinates ground truths. This deep network provides accurate results (72mm) for a monocular method using purely 3D data representation. However, more recent methods showing better results use two-stages top-down architectures including 2D prediction as a first step for 3D detection. This suggests that image features are not rich enough for direct 3D inferences.

VNect from Mehta et al. (2017) is a framework for 3D root-relative human pose estimation in real-time. It consists of a similar to Resnet50 architecture that generates 2D heat maps for each joint as well as newly introduced ”location maps.” These location maps predict the relative X, Y, and Z positions of an articulation relative to their root joint. For each joint, the location is processed from the peek of the 2D heat maps and the root-relative coordinate is read in the location maps. Then, a kinematic model of the human skeleton is fitted to the predicted poses, to improve temporal consistency and reduce jitter. The strengths of this method are that it works in real-time and can be used in different outdoor contexts. However, the authors list several limitations to their approach, mainly that depth estimation errors might lead to erroneous 3D predictions. The method is also only capable of relative pose estimation and therefore requires accurate detection of a root joint (pelvis).

Mehta et al. (2018) following Vnect, this technique also uses location maps but modify them to become ”Occlusion-Robust Pose-Maps” (ORPM) that infers 3D pose from multiple subjects. These ORPMs are similar to location maps but also contain structural information about the pose. For each joint, the location in the 3D maps is stored at the position of the joint but also along a predefined set of joints (dividing the body into two arms and legs) and root joints (pelvis and neck). They also employ part affinity fields Cao et al. (2017), which are often used in 2D multi-person pose estimation. They represent ”2D vector fields pointing from a joint [...] to its parents.” Thus, starting from valid root joints, all joints can be inferred following the kinematic chain of the human body. The outstanding feature of this method is that it perform multi-person detection and in a single-shot manner without intermediate use of off-the-shelf 2D pose estimators or person detector. It also introduces two new dataset for training and evaluation: MuCo-3DHP (large-scale multi-person with occlusion dataset, composed from MPI-INF-3D images) and MuPoTS-3D (In-the-wild multi-person dataset, filmed in various environments with markerless ground truths).

Mehta et al. (2020) More recently, in continuation of the previous two methods, the Xnect framework was presented. It is athree-stages method that combines a convolutional network for feature extraction, a fully connected network for pose estimation and a fit of the previous results to a kinematic model to refine pose consistency. The first stageextracts 2D and subject association through part affinity fields (similar to Cao et al. (2017)) as well as 3D pose encoding in the same way as the previous methods. The key difference from Mehta et al. (2018) is that each joint only encodes information about its position relative to the parent joint and the position of its children. For this stage, a new backbone module is presented for network architectures to reduce computational costs: SelecSLS. It consists of successive 3 by 3 and 1 by 1 convolutions with inter- and intra-module skip connections. It achieves similar accuracy to Resnet50 while being 1.4 faster at inference. The second

Method	Human3.6M	MPI-INF-3DHP	HumanEva
Pavlakos et al. (2017)	71.90	-	**24.3**
Mehta et al. (2017)	80.5*	76.6	-
Zhou et al. (2017)	64.9	69.2	-
Martinez et al. (2017)	62.9/47.7*	-	24.6
Sun et al. (2018)	64.1/49.6+/40.6*+	-	-
Omran et al. (2018)	59.9*	-	64.0*
Mehta et al. (2018)	69.9	74.1	-
Kolotouros et al. (2019)	**41.1**	76.4	-
Wandt and Rosenhahn (2019)	50.9 / 38.2*	82.5	-
Xu et al. (2019)	82.4 / 53.9* /48.0*+	73.1 / 76.9+ / 89.0*+	-
Kocabas et al. (2019b)	51.83+/45.04*+	77.5	-
Mathis et al. (2018) (DeepLabCut)	-	-	-
Mehta et al. (2020)	63.6	**82.8**	-

Table 3. Accuracy comparison from several state-of-the-art monocular methods. Human3.6M and HumanEva results reported in absolute MPJPE (lower is better); Results from MPI-INF-3DHP are reported in 3DPCK (higher is better). Techniques with the annotation with + are using extra-training data to obtain the result; the others use the benchmark's recommended protocols. The * annotation indicates results published with procrustes alignment to ground truth poses before evaluation.

stage consists of 3D pose detection, for each subject in parallel, through fully connected networks trained on the MuCo-3DHP dataset incorporating examples with severe occlusions. Finally, kinematic skeleton fitting is performed using the minimization of an energy term consisting of position (through inverse kinematics for 3D features and 2D re-projection), orientation and temporal consistency. Clearly, this is a complex framework that involves many steps, as well as re-tracking of each subject before the last stage. However, the framework is robust to occlusions and adapted to multi-person. It is also computationally efficient since it can run in real-time at 30 fps on generic hardware configurations. However, it appears that the last stage of the framework decreases the overall accuracy while performing better on certain joints and producing a smoother orientation estimate.

Zhou et al. (2017) published a two-stage method using the hourglass network architecture of Newell et al. (2016) for 2D heat map generation, then regress depth for each joint. In addition, they apply a weakly-supervised process to exploit images that have only been labeled with 2D ground truths. The depth prediction is realized from features at different resolutions, which are extracted from several levels of the Hourglass network. The weakly-supervised training is applied using 3D and 2D labeled data. Euclidean loss is applied to predictions on images with 3D ground truths, whereas a geometric loss is used when only 2D labels are available. This lossadds constraints from the average limb length ratios among predefined bone-groups. The main contribution of this work is the weakly-supervised technique: the authors evaluated whether the contribution of 2D data improved results on the 2D portion of the framework or on the whole 3D estimation task. They state that at PCKh@0.5 (standard metric for 2D pose estimation see 2.1), the results are similar for 2D human pose estimation when 2D data is not used, but depth estimation is greatly improved. However, an analysis with a threshold smaller than 0.5 might be interesting, as a small increase in 2D accuracy is still observed. Nevertheless, their work confirms that using 2D detectors as part of the 3D detection task is

possible and yields accurate results (≃65mm MPJPE).

Martinez et al. (2017) presented simple baseline for 3D human pose estimation. This method differs from others in that it does not use image data or intermediate feature maps (ie. joint location heat maps) nor does it use optimization steps with model fitting. Instead, it infers 3D coordinates from 2D coordinates obtained with a state-of-the-art 2D human pose estimation architecture. Despite this simple design, it produces accurate results comparable to and sometimes better than some more complex contemporary techniques. The design choices for the model are the following: a simple 2-layer CNN with batch normalization, ReLU activation, and dropout. It takes as input the 2D predictions of the Hourglass Network Newell et al. (2016). The results obtained on Human3.6M reach 62mm average error on single images. This method is also one of the first to propose the direct lifting of 2D coordinates from efficient 2D detectors to 3D. The high accuracy obtained with such a simple approach without image data as an input lead the authors to hypothesize that the visual features used in contemporary methods were either not useful to 3D human pose estimation or still under-exploited. The latter hypothesis tends to be confirmed with new methods achieving increasing accuracy with clever exploitation of temporal features (Hossain and Little (2018); Pavllo et al. (2019); Cheng et al. (2019); Cheng et al. (2020)) or combination of direct regression and model fitting.Because it is simple, fast and is driven by the increasing performances of 2D detectors, this method and the two-stage technique inspired many recent studies.

Sun et al. (2018) introduced "integral regression" to extract 3D coordinates from 2D confidence heat maps. It is a function similar to the soft-argmax function commonly used in classification to normalize outputs. Here, it is used to switch from 2D pixel maps to differentiable coordinates, allowing direct regression within an end-to-end network. Their article describes extensive ablation studies on different training architectures losses and backbone (hourglass and residual networks 3.2) and presents good results for both 2D and 3D human pose estimation. The authors also adopt a training strategy Sun et al. (2017) allowing the usage of 2D labeled data with separate supervision for the xy coordinates and for the depth, achieving even higher accuracy on Human3.6M. Many approaches are now using 2D heat maps with the soft-argmax regression.

Omran et al. (2018) present the Neural Body Fitting approach that combines part segmentation with a convolutional neural network and a parameterized human body model (SMPL Loper et al. (2015)). The first stage predicts a part segmentation map using the RefineNet Lin et al. (2016) CNN architecture. In the second stage, these part masks are directly fed to a ResNet-50 network that estimates the pose and shape parameter of the SMPL model. The whole process is fully differentiable and can also be trained with 2D data in a weakly supervised manner. The authors demonstrate this by re-projecting the joint coordinate on 2D images and find that, with only 20% of the training data with 3D ground truths, the same accuracy is achieved as with complete annotations. This method is one of the first that describes the training of CNNs followed with the fitting to a parameterized human body model in a single integrated framework. The other characteristic that differentiates this method from others is that it uses the intermediate feature of the part segmentation map. According to the author analysis, using a 24-part segmentation as an input led to significantly better results than direct image data or joint coordinates. This result is particularly interesting and should be considered when designing architectures that want to reason on the 3D structure of the human body.

Similarly, Kolotouros et al. (2019) designed a system with joint usage of a CNN for direct key point regression and an iterative model optimization technique using a human volumetric model (SMPL Loper et al. (2015)). The neural network produces good initialization for the iterative fitting of the human model. Then, once the shape model position is refined, it is used to calculate a loss from the initial prediction, which increases the accuracy of the network. The originality of the method lies in using the best of the two techniques. Direct regression allows

fast initialization without a priori knowledge directly on the image data; iterative optimization from a human model produces a better shape for the image fitting. The two parts of the system complement and improve each other during each training cycle. The results obtained using this method on Human3.6M are the most accurate of monocular non temporal methods (on a single isolated RGB image). The average error is 41.1mm, which is close to the accuracy of temporal methods. These results show that with simple input data and a suitable method, high accuracy can be achieved. The question is: Could this direct regression/model fitting method using only visual features be suplemented with temporal data to reach a higher accuracy score?

Wandt and Rosenhahn (2019) uses the kinematic chain space to represent human 3D pose within their discriminator (or "critic") network. To project pose coordinates in the kinematic chain space, limbs (i.e., edges between detected joints) are described as directional vectors. It is then possible to map this representation back to point coordinates. The kinematic chain representation contains information about limb length, angles and body symmetry, while being easy to compute. Later, Cheng et al. (2020) used a temporal version of the kinematic chain space that is reporting the angle and length modification across frames in a video.

Xu et al. (2019) proposed the DenseRaC framework that converts pixel-to-surface maps of the human body (IUV) into parameters for statistical human body model. Similarly to Omran et al. (2018), this framework uses intermediate features but, instead of part segmentation, DenseRaC uses pixel-to-3D surface maps. In the first stage of the pipeline, these maps are computed using the Densepose-RCNN architecture (a network that is trained on Densepose-COCO, a dataset of manually annotated human body surface Güler et al. (2018)). To improve training, the authors present a large-scale dataset of synthetic human poses that can also easily produce IUVs. In the second stage, the IUVs are fed to a regression network that estimates the human body shape and pose parameters (similar to Loper et al. (2015)) as well as the camera parameters. Once reconstructed, the 3D mesh is re-projected using a differentiable renderer and rasterized in an IUV similar to those produced in the first step. Then, the computation of an adversarial loss with a discriminator network helps to eliminate impossible configurations.

Kocabas et al. (2019b) takes advantage of the multi-view setting that each major motion capture dataset provides. the authors propose to infer geometry from matching 2D detection in each view and then deduce the 3D coordinates. This technique renders self-supervision possible with completely unlabeled data. Unlike many multi-view techniques (Iskakov et al. (2019); He et al. (2020)) this method does not use camera parameters. It's important to note that the training part of the network uses a multi-view setting, but during prediction it becomes a monocular method. The architecture is composed of two networks using ResNet50 and a deconvolution layer as their backbone (see 2D architectures 3.2). Both produce spatial heat maps for each joint in each view. The difference is that, in one case, they are converted in 2D coordinates and in 3D in the other (both with a soft argmax). The 3D network is the one that will be used for inferences, and that has learnable weights. The 2D network is frozen and will be used for the self-supervision.Although the cameras are not calibrated, the authors advise using detected the joint key points in each view to obtain the camera parameters (using RANSAC and SVD). Then, triangulation can be performed to obtain the 3D coordinates, which are then used to supervise the 3D network with a smooth absolute loss. The accuracy score of the method is the current best when few or no labeled data are available (70mm average MPJPE on H3.6M and 64.7 3DPCK on MPI-INF-3DHP). This learning scheme is promising and allows for high accuracy by requiring only raw data (discounting pre-trained 2D detector), provided that multi-view images are accessible.

Finally, it is important to consider the methods being used in several research fields. The main one, DeepLabCut

Mathis et al. (2018), Mathis and Mathis (2019), Nath et al. (2018) is a generic markerless keypoint-tracking framework developed with the goal to obtain accuracy results similar to human annotation. Its original application target was the video tracking of predefined keypoints for different species. DeepLabCut achieves good results with little training data, which has made it popular and it is now cited in numerous articles in neuroscience and human movement research. For human gait analysis Moro et al. (2020)Fiker et al. (2020), its accuracy is similar to marker-based motion capture. Although DeepLabCut is not in the strict sense a HPE model, it can be used as one. In fact, it is based on an HPE architecture: It consists of the first layers of the DeeperCut network Insafutdinov et al. (2016), which is a multi-person 2D pose estimation model used here for feature extraction, followed by a deconvolution layer. This model and its inspiration do not directly fall within the scope of 3D human pose estimation. Yet, DeepLabCut has been used in a multi-view calibrated cameras context with simple triangulation (Nath et al. (2018); Sheshadri et al. (2020)). As this method seems popular in different fields for its flexibility and performance in a wide variety of contexts, it is important to note that more recent results on specific 2D keypoint detection have outperformed DeeperCut Newell et al. (2016). The main appeal for this method is the definition of personalized labels and its powerful generalization capability.

4.2. Temporal and Monocular Images

Hossain and Little (2018) used the sequence-to-sequence architecture that was initially applied in machine translation tasks for human pose estimation. The main idea is to use long-term temporal context to predict a new sequence, as for the translation of text into another language. Here, 2D coordinate sequences (predicted with the Hourglass Network architecture) are used to predict the 3D ones using Long Short-Term Memory units (LSTM). The first layer of the network encodes the sequences into a hidden space and the last layers decodes them into a 3D dimensional sequence using residual connections. Additionally, important choices are made for the training. First, the first derivative of the 3D coordinates is included in the loss to mitigate the effect of errors during the 2D prediction stage by enforcing temporal consistency. Second, the training combines layer normalization, recurrent dropout and finally the weighting of joint sets, to force the network to better predict challenging parts. To date, the method is one of the most accurate among methods using temporal data. Notably, an evaluation by the authors shows that the optimal sequence length for their model is 5 frames. Beyond this point, the accuracy slowly decreases, suggesting that either the long-term temporal context does not provide good information or the model does not properly exploit past image data.One problem with this architecture is the fixed length of the temporal data input, which was later solved using temporal convolutional networks (TCNs) Pavllo et al. (2019) Cheng et al. (2020). This paper further details and analyzes other useful techniques applied to temporal pose data, such as the important contribution of residual connections.

Cai et al. (2019) also uses temporal dependency within the pose sequence that composes the subject motion. However, in addition to time constraints, joint position constraints are also enforced using a graph structure. Edges are present to model spatial continuity and symmetry, but also connections with the same joint in future and past frames of the video. These graphs are computed from 2D skeletons that are then passed to a graph convolutional network. Depending on the type of neighboring (various spatial constraints and temporal connections), the authors propose a different type of convolution. Then, a local-to-global architecture is used similarly to Newell et al. (2016) to obtain 3D predictions. The graph structure is pooled and upsampled using lower level features from the previous layers in a hierarchical manner. Then, the method produces a 3D pose from a pose refinement module consisting of a fully connected layer. The authors employ a loss strategy combining different terms to enforce joint locations, limb symmetry and temporal smoothness. The results obtained on different benchmarks show a better accuracy than methods based purely on tempo-

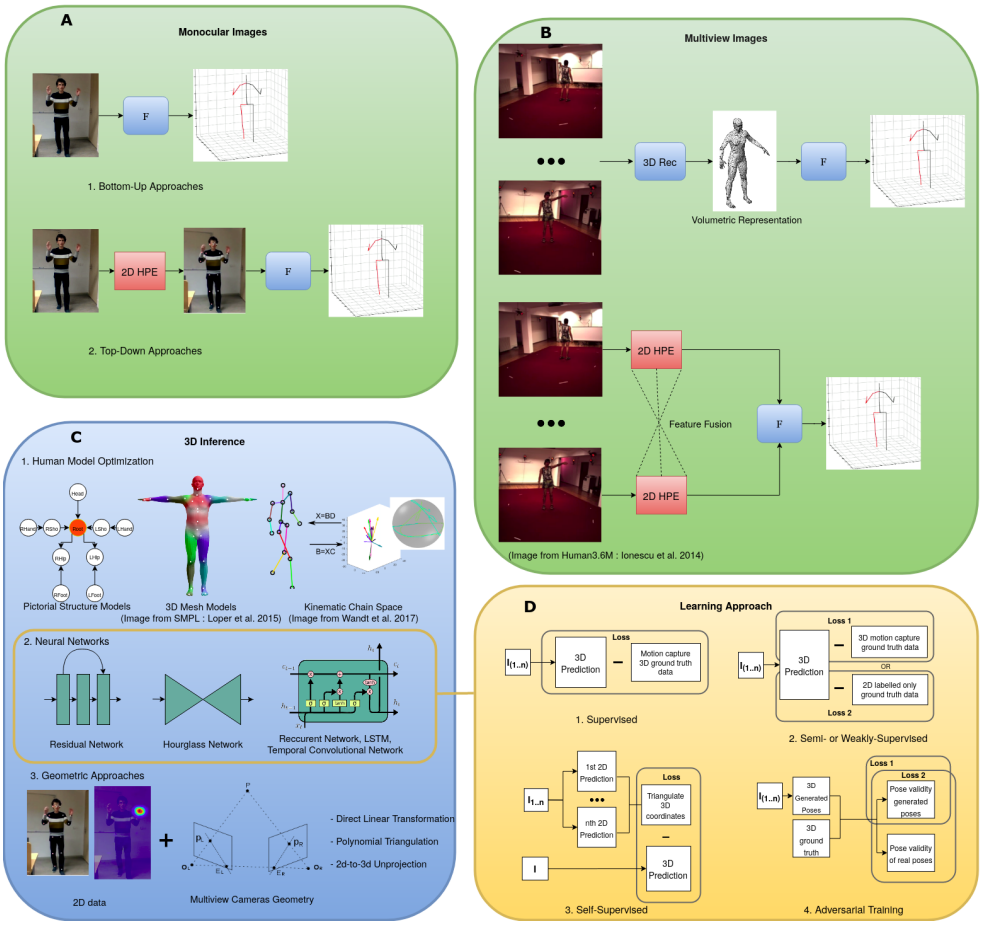

Fig. 3. Overview of the different levels of 3D markerless human pose estimation. A: Monocular approaches, commonly used 2D pose estimation backbone architectures are described in 3.2. B: 3D feature exploitation and multi-view 2D detection as an input for 3D detectors. C: The different families of 3D pose estimation. D: examples of learning approaches applied to human pose estimation.

ral sequences. However, the method is likely not suitable for sequences shorter than 7 frames on most actions. It is also important to note that the effect of the chosen 2D detector is not tested.

As mentioned previously Pavllo et al. (2019), the technique consists of a monocular method that leverages a fully convolutional network to process temporal data. The main advantage is that, compared to RNN and LSTM, they are more computationally efficient and do not need a fixed input size.In this paper, the authors present an architecture based on time-dilated convolutions that also apply semi-supervision to add non-labeled data to the training. The dilated convolutions capture the long-term dependencies, but also need less training parameters and have a better computation speed than sequence-to-sequence models Hossain and Little (2018). The authors present a comparison showing that about the same number of parameters and FLOPs are needed to yield better accuracy. The semi-supervised process they used is called ”back-projection”: an encoder-decoder architecture encodes a 2D joint prediction into 3D pose and decodes by projecting it back in 2D. The error is then computed between the original and back-projected 2D predictions.This method has a 10% increase in accuracy on two of the main benchmarks, leading to an average error of less than 50mm for monocular methods.The methods that follow this paper, based on video sequences, still use the Temporal Convolutional Networks (TCN) architecture (Cheng et al. (2019), Cheng et al. (2020), Liu et al. (2020)). Another interesting result is the impact of the 2D detector used. The tests conducted show that the CPN and the Mask-RCNN 2D detectors give better results in the end with their model. The

Method	Human3.6M	MPI-INF-3DHP	HumanEva
Hossain and Little (2018)	58.5	-	22.0
Cai et al. (2019)	48.8 / 39.0*	-	-
Pavllo et al. (2019)	46.8 / 36.5*	-	23.1/15.8*
Cheng et al. (2019)	42.9/32.8*	-	14.3*
Cheng et al. (2020)	**40.1**	**84.1**	**13.5***
Liu et al. (2020)	45.1 / 35.6*	-	15.4*
Wang et al. (2020)	42.6 / 32.7*	86.9(2d GT)	-

Table 4. Accuracy comparison from several state-of-the-art monocular temporal methods. Human3.6M and HumanEva results reported in absolute MPJPE (lower is better); Results from MPI-INF-3DHP are reported in 3DPCK (higher is better). Techniques with the annotation with + are using extra-training data to obtain the result; the others use the benchmark's recommended protocols. The * annotation indicates results published with procrustes alignment to ground truth poses before evaluation.

authors think that "it's due to the higher heatmap resolution, stronger feature combination" from these detectors.

Cheng et al. (2019) and Cheng et al. (2020) also use temporal convolutional networks with a specific occlusion training to improve prediction on challenging images.

In Cheng et al. (2019), they describe an occlusion-aware network where a "cylinder man model" is produces occlusion labels. Their architecture is composed of an end-to-end trainable human detector, a 2D pose estimator, a 3D pose estimator and finally a pose discriminator. The two pose estimation algorithms are respectively a 2D and a 3D temporal convolutional networks. They also use 2D-only data during training using a re-projection loss such as in Pavllo et al. (2019). The occlusion model intervenes at this stage where self-occluded points (computed with the cylinder model) are not taken into account as they are deemed unreliable.

Building on the previous work, Cheng et al. (2020) describe an end-to-end trainable model with several modules that reconstruct 3D joints from monocular videos. 2D confidence heat maps are estimated and used as a feature for 3D prediction. They use a multi-scale convolutional network (HRNet) that fuses spatial features Sun et al. (2019). Thus, the main characteristics of their method are the following :

- Learning of a multi-scale embedding obtained from those heat maps. 3D poses are then predicted with the embedding using a temporal convolutional network (TCN) Pavllo et al. (2019).
- Validation of the pose sequences with a discriminant model based on Spatio-Temporal Kinematic Chains (which enforces limbs angular and length constraints).
- Data augmentation using synthetic occlusion at different levels during TCN training.

Semi-supervised learning is used with the goal of including strictly 2D labeled data during the training process as in Pavllo et al. (2019). Finally, and only for the training stage, multi-view from Human3.6M dataset is used to enforce a good skeleton orientation prediction. To do so, the authors use a loss function comparing each inference from two pairs of different views at the same time in the video (after applying a camera rotation known from calibration data). Among the methods reviewed here, this method achieves the overall best results on HumanEVA and MPI-INF-3DHP. It also has the best results for a monocular method on Human3.6M (40mm MPJPE). However, it is a complex method with many submodules and parameters reducing errors from occlusion. It also exploits and expands the pose discriminator based on the kinematic

chain space from Wandt and Rosenhahn (2019), using it with temporal data. The authors address the contribution of each module to the final accuracy. However, they do so by adding the modules one at a time to the backbone, which does not provide any insight into whether one module improves or compensates for the errors or performance of another. Further cross-comparison could help determine which module has a greater impact and on which data or context.

Liu et al. (2020) Added an attention mechanism to the temporal approach for extracting 3D pose from video. Similarly to Pavllo et al. (2019), temporal dilated convolution extracts information in the 2D poses sequence. The added attention mechanism selects the frames and tensor outputs that are the most useful for the detection. The temporal attention modules are computed from the distribution of the tensors at each time step. The kernel attention modules are computed from the distribution of the channel outputs of each layer. Both attention modules are propagated within an attention matrix to the following layer. On top of this architecture, multi-scale dilation convolutions (see dilation convolution in the previous described method Pavllo et al. (2019)) with increasing receptive fields are employed to reduce the vanishing gradient issues. This method shows progress over the state-of-the-art methods. An ablation study also shows how attention modules associated with the multi-scale convolution strategy lead to better results, especially on difficult frames (fast motions or partially occluded subjects).

The two main contributions of Wang et al. (2020) are: *Motion loss*, a new loss function based on keypoint trajectory and *U-shaped Graph Convolutional Network* (UGCN), a network architecture. Motion loss is based on coordinate vectors. To make it differentiable, so it can be used in a learning architecture, any motion sequence needs to be encoded using a differentiable operator. The authors empirically chose the scalar product (among other tested). The final loss is composed of this motion loss computed with "pairwise motion encoding" and an absolute position reconstruction loss. UGCN uses spatial temporal graph to represent motions Cai et al. (2019). Then, spatial convolutions are applied on each skeleton for each frame before temporal convolutions are applied to the temporal dimension of each joint in the graph. The network architecture is similar to successful ones in semantic segmentation (e.g. Ronneberger et al. (2015). It consists in three stages: downsampling, upsampling and merging. This architecture captures features at different scales, which implies that UGCN explores temporal and spatial information at different scales. This architecture was tested adding upsampling and downsampling one by one, showing increasing accuracy. The addition of motion loss also drastically improves the results on two benchmarks (Human3.6M and MPI-INF3DHP), improving on the state-of-the-art results.

4.3. Multi-view

Qiu et al. (2019) uses a two-stage prediction process similar to Zhang et al. (2020) without taking IMU data as inputs. Instead, they opt to fuse data from multi-view images using projective geometry constraints. This process is done through a convolutional layer that merges pixel data from each view along the epipolar lines using weighted matrices. After this step, fused 2D joints heat maps are generated and 3D pose is inferred through the Pictorial Structure Model with a skeleton model. A recursive variation of this model is used to reduce quantization errors and complexity using a divide and conquer scheme for space discretization. Comparing results from methods providing absolute coordinates, the cross view approach improved the state-of-the-art results at the time and further improves them using additional 2D data during training. It also gives competitive results on the TotalCapture dataset without pre-training and without using the IMU data. It can also be used in different camera setup using pseudo-labeling from 2D pose estimator. The method can be applied to 3D human pose estimation in new contexts without the need of training with 3D ground truths. This method illustrates very well the main approach that consists of using improved results in 2D HPE and translating them into 3D. Here the refinement is done by adding multi-view to improve 2D predictions.

Method	Human3.6M	Total Capture	Input
Qiu et al. (2019)	31.17 / 26.21+	29	Multi-view
Iskakov et al. (2019)	**17.7+**/20.80*+	-	Multi-view
He et al. (2020)	26.9/19.0+	-	Multi-view
von Marcard et al. (2016)	-	-	Multi-view, IMU
Trumble et al. (2017)	87.3	77.0	Multi-view, Temporal, IMU
von Marcard et al. (2018)	-	26.0	Monocular, IMU
Huang et al. (2019)	37.5/13.4*	28.9	Multi-view, IMU
Zhang et al. (2020)	-	**24.6**	Multi-view, IMU

Table 5. Accuracy comparison from several state-of-the-art multi-view and multimodal methods. Human3.6M and TotalCapture results reported in absolute MPJPE (lower is better). Techniques with the annotation with + are using extra-training data to obtain the result; the others use the benchmark's recommended protocols. The * annotation indicates results published with procrustes alignment to ground truth poses before evaluation.

Iskakov et al. (2019) present a learnable way to triangulate human poses. This article proposes two geometric methods for triangulating 3D joint coordinates from multiple view joint heat maps. The first one is an algebraic method based on solving a system of vector equations with 3D coordinates. The second method is a triangulation from volumetric aggregation of re-projections of the 2D heat maps in a voxel grid. Both are weighting the information coming from different views with learnable coefficients. In the volumetric approach, each heat map corresponding to a joint from each different view is sampled into a voxel cube. These volumetric maps (for a specific joints) are then aggregated with a weighting of the impact from different views and fed into a 3D CNN that refines them. The final step is a soft-argmax operation on the resulting 3D heatmaps yielding computable 3D coordinates. Each of these steps is differentiable and the weights of the different convolutional layers at each stage are updated using an absolute loss. The results on Human3.6M for multiple view input are the best ones to date with a 17.7mm MPJPE accuracy (i.e., with the volumetric method and softmax aggregation during the learning stage as described above). On limit of this method is that it needs a correct cropped volume around the human skeleton to work well. Consequently, at least two camera must detect the pelvis joint. The given absolute MPJPE score is also calculated with the removal of several actions due to annotation errors on Human3.6M.Nevertheless, this method reaches the best accuracy for relative pose estimation among all the methods we reviewed.

Epipolar Transformers from He et al. (2020) are modules using the attention mechanism as well as knowledge of epipolar geometry. The authors noted that most 2D detection of keypoints did not use 3D features at all: this was the motivation for their "Epipolar Transformer". The goal is to fuse intermediate features from multiple views during 2D inference using projective geometry constraints Qiu et al. (2019). First, from a detected point in the source view, the module samples all points on the corresponding epipolar line in another view. Then, features across this line are fused according to a computed weighted similarity with the source point. In the end, the obtained feature maps have the same size as the input, which makes the module compatible with any two-stage multi-view system. Any triangulation algorithm can then be applied. On Human3.6M, the authors computed the 3D coordinates with the recursive pictorial structure model of Qiu et al. (2019) using their epipolar transformer: they obtained the best MPJPE score across state-of-the-art, and they did so without using external training data. They also compare their results with pre-training on MPII 2D dataset and get results close to the best methods (19mm against

17.7 from Iskakov et al. (2019)) with 10% less parameters and computation operations. The main limitation of the method is that it only works on a fully calibrated multi-camera system, because it needs camera parameters to compute epipolar lines.

4.4. Multimodal approaches

von Marcard et al. (2016) propose one of the first method combining multi-view video and IMU data for pose estimation. The authors claim that their method is less intrusive than marker-based MoCap, as only a few sensors are placed on the subjects. They use a representation of human body constraints based on kinematic chains. Using the silhouettes from multi-view extracted with background subtraction and limb orientation from IMU, they minimize a hybrid energy term obtained from orientation and contour consistency with a human mesh model (Loper et al. (2015)). They provide a thorough analysis of their method (with the TNT15 dataset presented in the same paper and some metrics on HumanEva) showing that video and orientation sensor complement each other. The idea is that IMUs accurately measure joint angles but tend to drift during the experiment, whereas video is better suited to obtain positional information.

Trumble et al. (2017) present the first large-scale motion capture dataset that also contains IMU data. The paper describes a 3D pose estimation technique fusing 3D data from multi-view and limb orientation from IMU, while maintaining the temporal context using a Long Short-Term Memory layer over the five past frames. A PVH Grauman et al. (2003) is computed from the multi-view video and inputted to a 3D convolutional network with 26 3D joint coordinates as an output. In parallel, kinematic solving provides the same joint coordinates from IMU data. The vectors from both sources are then passed through the LSTM layer and merged into an embedding representing the 3D pose. This way, the authors show that it is possible to learn a "mapping between the predicted joint estimates of the two data sources and the actual joint locations". They evaluate their method on the newly presented TotalCapture dataset. They also evaluate the multi-view part of their pipeline without IMU on Human3.6M.

Following their work with inertial units, von Marcard et al. (2018) present a monocular method with a moving camera in-the-wild and multiple subjects wearing IMUs. The method uses a 2D multi-person pose estimation algorithm Cao et al. (2017) and a module that fits the SMPL Loper et al. (2015) model to IMU data. Then, each 2D skeleton is paired with 3D pose and shape using graph optimization. Next, with the obtained associations model parameters, camera pose and orientation are optimized and fed back for further iterations. When evaluated on the TotalCapture dataset, the method outperforms previous ones by 44mm. This technique allows for the capture of a new dataset in outdoor environments and without marker ground-truths: the 3D Poses in the Wild Dataset.

Huang et al. (2019) developed an end-to-end trainable 3D convolutional network witha refinement module based on IMU data. The main idea is to process primitive 3D data from multi-view frames without any transformation. A multi-channel volume, constructed from the segmented human silhouettes and camera parameters, is used as input for the network. 3D voxel confidence heat maps are computed at this stage and can already be used for human pose estimation prediction. The refinement stage merges the volumes constructed from IMU and those constructed from heat maps, as well as the multi-channel volumes. IMU volumes are processed into a "bone cylinder" from quaternion orientation and the previously predicted joint position. Then, all this 3D information is fed into another 3D convolutional network. The architecture uses hourglass networks and residual network modules (see 2D architectures 3.2) and 3D soft-argmax to extract coordinates. MSE loss is computed at the different levels of the architecture. By randomly stopping information from a camera, the method allows for data augmentation during training, which significantly improves performance on partially captured images. With the vision-only module on a dataset without IMU, the method obtains good overall accuracy. The authors claim that their model can be used in a real-time system because it

does not use time sequences. However, they do not provide any performance or speed evaluation. They also point out that their technique does not use a complex human model and is therefore more likely to generalize to new subjects. However, this depends on the performance of the human shape segmentation algorithm at the preprocessing step.

Similarly, Zhang et al. (2020) fuse IMU with multi-view image data, but this time using a human skeleton model. The approach is to refine 2D joint confidence heat maps with IMU data to then use a part-based model. First, as in Qiu et al. (2019), they extract 2D joint heat maps. However, the authors use a simple baseline method Xiao et al. (2018) before merging the heat maps information and the IMU orientation to geometrically estimate the correct coordinates. They call this process Orientation Regularized Network (ORN). This technique can be used with any heat map-based prediction method and they use it to train an end-to-end network that produces more accurate results than state-of-the-art 2D methods not exploiting IMU data. The second part of their work consists of a variant of the Pictorial Structure Model (PSM) (Felzenszwalb and Huttenlocher (2005), Belagiannis et al. (2014)) which is often used for 3D and 2D HPE. This family of method calculates the most probable human pose within all possible ones in a discrete space. Typically, PSM-type methods uses limb length as the primary constraint. Here, the authors exploit the IMU data to also apply a limb orientation constraint, which improves accuracy. This framework includes a module that can deeply improve results from 2D estimators with inertial data. Note that the 3D part of this system does not use CNN, unlike many other current methods, although it gives excellent accuracy. The authors also use synthetic IMU data to predict 3D pose on the Human3.6M dataset to conclude that their use improves the results by an average of 10mm.

Finally, recent works using a single depth sensor Yu et al. (2017) and Yu et al. (2018) have shown good results for real-time motion capture. The first one uses the iterative closest point algorithm to reconstruct body shapes and the second optimize the SMPL body model. However, these methods are not tested for joint localization error on a pose estimation data set.

5. Analysis and discussion

As mentioned in the introduction, this article is structured around the specifications and requirements, which vary according to the fields of application. Today 3D pose estimation is employed in many fields :

Human Computer Interface: There is an increasing number of applications using human pose and gesture to interact with computers. 3D HPE is essential to help robots and machines better understand and respond to human motions.

Security: The classical applicationis to track people in the indoor and outdoor environment to ensure they do not commit theft or infractions.

Motion Analysis: This is a broad field that comprises sport and performance analysis, medical study, pose semantics or the study of inter-human interactions.

Entertainment: Pose estimation can be used for avatar control (e.g., Kinect) or VR and AR refinement and for avatar animation in games and movies.

Moeslund and Granum (2001) classify them into three categories : surveillance, analysis and control. Each category requires different performance regarding accuracy, speed or robustness. However, within the same category, there are variations on these criteria. For example, as the previously mentioned authors state, a control application can be constrained to highly controlled environment (avatar control) or within a generic outdoor scene with varying conditions.

For this reason, our review describes each criterion separately and explains in which use case it performs the best. Ta-

bles 6, 7 and 10 classify these methods with accuracy reported in MPJPE. The level of robustness corresponds to the number of assumptions or constraints necessary for correct detection. Lastly, to express the speed, we report whether the model can run in real-time. Each of these tables allow to cross-compare the different methods best suited to the most needed specifications.

5.1. Accuracy

Accuracy is the main criterion used to evaluate markerless human pose estimation methods within the computer vision community. However, it has some limitations that are important to keep in mind. First, most benchmarks compare markerless vision-based methods to results obtained from marker-based optoelectronic systems that are themselves not free from errors. As reported in Section 2.2, some old motion capture datasets contain videos with low resolution and inaccurate annotations.

The second limit in accuracy comparison is that, in many cases, it is not the local error rate that is important to the application but the semantics of the pose as whole (See 2.1). This can be the case for surveillance systems or human-machine interface for example. When the accuracy requirement (at least with conventional metrics) is low to minimal, see the two other criteria discussed in this review.

However, many other fields of research need high accuracy , such as medical science or biology. Typically, the accuracy prerequisite can also be different depending on the analysis. Sometimes, the highest accuracy possible is required regardless of the acquisition and computation procedure complexity (e.g., for research in biomechanics). However, many studies are conducted from human labeled videos with much simpler setups. We propose a solution based on machine vision for each of these scenarios.

5.1.1. Highest accuracy methods

When looking at results of contemporary methods, there is an average gap of 10mm error between monocular and multi-view methods. When multi-view methods are used properly, the combination of geometric knowledge of the scene and learning optimizations can results in errors as low as 20 mm. However, the error is higher in a general context (which is also true for monocular methods). For example, Iskakov et al. (2019) obtained 34mm MPJPE on a different dataset than the one used for training.

Accuracy				
Method	Type	Accuracy	Robustness Level	Real-time
Kolotouros et al. (2019)	Monocular	**41.1**	average	✗
Wandt and Rosenhahn (2019)	Monocular	<u>50.9</u>	average	✓
Iskakov et al. (2019)	Multi-view	**17.7**	average	✗
He et al. (2020)	Multi-view	<u>26.9</u>	low	✗
Cheng et al. (2020)	Temporal	**40.1**	average	✗
Wang et al. (2020)	Temporal	<u>42.6</u>	average	✗
Zhang et al. (2020)	Multimodal	*24.6**	low	✗

Table 6. Best accuracy methods. Methods reported with performance criteria of current real-life applications that use 3D pose estimation for four setups (monocular, temporal, multi-view). Accuracy is reported in MPJPE on Human3.6M dataset bold for best and underlined for second best. The best multi-modal approaches using IMU (marked with *) are evaluated with MPJPE on TotalCapture for comparison.

IMU sensors are also a good way to improve multi-view detection. However, for now, the results seem close to those obtained with image data alone, with Zhang et al. (2020) obtaining the most accurate results. Future evaluations on the TotalCapture dataset Trumble et al. (2017) with methods that do not use inertial information could help with the comparison.

The strength and weakness of the multi-view methods comes from the fact that they are based on camera parameters. This helps to the generalizability of the system as it can adapt to new camera views He et al. (2020), but also makes the process more complex as it requires calibration. A solution can be the one suggested by Kocabas et al. (2019b), with a system that computes camera parameters on-the-fly with the detected joint as calibration targets. Multiple images hold much more information than a single one and well-known stereo vision properties are applicable.It turns out that most

research has been concentrated on monocular images and 2D estimation, so many other avenues remain unexplored, such as the simultaneous exploitation of temporal and multi-view data.

Monocular methods based on video sequences that can be processed offline such as Cheng et al. (2020) and Pavllo et al. (2019) have merit. However, their accuracy is not less than 40 mm MPJPE on Human3.6M, and their ability to generalize is also unproven, as no comparative analysis is yet available for them.

5.1.2. *Simplest but accurate methods*

Monocular methods are the simplest, as they can be fed a simple video input. In many cases, they are a clear improvement over handcrafted annotation, as they have similar accuracy but provide richer 3D information with less input data. When video data is available, it is judicious to favor methods based on temporal data that provide better results (Cheng et al. (2020); Wang et al. (2020) or Pavllo et al. (2019)). Kolotouros et al. (2019) also propose a method based on single images that is competitive with image sequence techniques (41.1mm MPJPE on Human3.6M). This indicates that their approach of jointly optimizing a mesh model and training a neural network is an effective way to solve monocular pose estimation problems.

Although Wandt and Rosenhahn (2019) does not achieve the typical 40mm average error, it does so at a low computational cost. Thus, it is a good compromise between accuracy and speed for real-time pose estimation.

In many cases, another important factor is the flexibility of the method. Yet most algorithms were not designed with this in mind, as they are limited to the labels in the training data. For the study of human motion, it is common to have to define custom key points to track limbs, segments or joints. This problem can be solved using DeepLabCut Mathis et al. (2018) or other methods that are not human-specific Zhang and Park (2020). Obviously, these methods still require ground-truth examples for each new key point, but on a much smaller scale. Another possibility is to refine existing models with new custom labels.

5.1.3. *Overall conclusion on accuracy*

The table 6 shows that the most accurate methods have a medium to low level of robustness and that few of them work in real time. The explanation comes from the fact that they are often complex methods that tackle issues such as occlusion with specific modules that increase the overall computational cost. As explained above, the best methods are naturally multi-view techniques that exploit geometric constraints and therefore need calibrations.

Currently, the architectures that achieve the best accuracy are two-stage top-down algorithms. They achieve the best results on a variety of benchmarks in monocular image, monocular video, and multi-view configurations. They often build on the success of 2D pose estimation, which is a nearly solved problem (i.e., it achieves average PCK scores above 90%). Many different approaches are effective: direct regression from 2D to 3D, initialization of human parametric models, exploitation of temporal sequences, occlusion-aware modules, generative models, volumetric input representation, multi-view triangulation, to name the most important (See 4 and Fig. 3). An interesting point is that, logically, video sequence methods perform better for activities with temporal coherence, such as walking or running, while monocular methods best detect complex static poses better.

Another strong distinction can be made between model-based and model-free approaches. While the state of the art in 2D detection does not use human models, 3D detection successfully does. They are either based on Pictorial Structure Belagiannis et al. (2014) or on a 3D mesh model like SMPL Loper et al. (2015) or SCAPE Anguelov et al.. Recent successful propositions also use 2D detection from CNN as initialization for their models. However model-free techniques perform as well or even better in some cases. The detection error for the 3D human pose estimation task has decreased by nearly 70mm over the last decade, mainly due to convolutional networks in different architectures. Interestingly, these improvements were

Robustness				
Method	Type	Accuracy	Robustness Level	Real-time
Mehta et al. (2020)	Monocular	63.6	high	✓
Xu et al. (2019)	Monocular	82.4	high	✓
Hossain and Little (2018)	Temporal	58.5	high	✓
Cheng et al. (2020)	Temporal	40.1	average	✗
Liu et al. (2020)	Temporal	45.1	average	✗
Iskakov et al. (2019)	Multi-view	17.7	average	✗
von Marcard et al. (2018)	Multimodal	26.0*	high	✗

Table 7. Most robust methods. Robustness level is defined as follows: 1 - 3 assumptions: high level, 3 - 4 assumptions: average level and 5 or more: low level.

mainly due to better 2D modules, which may indicate that research into the 3D nature of the problem is a fruitful avenue.

5.2. Robustness

Robustness is usually assessed through changes in the accuracy during cross-dataset evaluation. Moeslund and Granum (2001) propose to express robustness as the number of assumptions required for a motion capture configuration to be operational. They define twenty assumptions relative to the acquisition protocol or environment and the subject's appearance. Some of them have been overcome by all current method (subject wearing specific clothes or static monochrome backgrounds), but others are still very much debated (occlusions, single person or no camera motion). Below are assumptions that remain in the state of the art:

Motion Constraints : No camera motion, no fast motion of the subject, no occlusion (no severe occlusion, no auto-occlusion). For temporal methods, the number of frames and the acquisition frequency may also be a limitation.

Environment Constraints: Extra hardware (IMU, Laser Scans), Multiple cameras (with or without camera parameters), etc.

Subject Constraints: A single person, a known first pose, a motion parallel to the camera plane (for a model-based approach)

5.2.1. Background, Lighting and Clothes

Most appearance constraints are no longer required, because convolutional networks do much better at identifying meaningful visual invariants than former hand-crafted feature extractors. Yet, it is difficult to evaluate generalization to in-the-wild situations, mainly because large motion capture datasets are still captured in indoor studios. Data augmentation can provide new scenes with background extraction or change the lighting. With the release of new commercial markerless solutions, new benchmarks such as MPI-INF-3DHP Mehta et al. (2016) also include real outdoor data.

Method	#Frames	Causal
Hossain and Little (2018)	5	✓
Cai et al. (2019)	7	✗
Pavllo et al. (2019)	243	✓
Cheng et al. (2019)	256	✗
Cheng et al. (2020)	128	✗
Liu et al. (2020)	243	✓
Wang et al. (2020)	96	✗

Table 8. Monocular temporal models assumptions. Needed number of frames to obtain the optimal accuracy and whether the system can be adapted to only use past frames (for real-time use)

5.2.2. Special Hardware & Calibrated Systems

Two constraints are still relevant: first, the need for specific hardware, second, the need for calibration in multi-view configurations. The two most commonly used extra hardware added for pose estimation are inertial motion units and depth sensors such as infrared or time-of-flight cameras. Inertial motion unit provide additional information on the limbs orientation but suffer from drift in their results after a short usage. They are also more practical than reflecting markers and motion capture suits but are still intrusive for the subject. Different depth sensors have also been used to infer the depth of joints directly or to construct more complex features as a pre-processing step of pose estimation. Less frequently some methods use the 3D scan of each subjects that is captured to fit shapes parameter of human body models. Logically, while most research is conducted on monocular methods, they are always outperformed by 10 to 20 mm MPJPE by multi-view techniques. In multi-view systems, the calibration step is a frequent requirement. Multi-view

Method	Calibration	IMU	#Views
Trumble et al. (2017)	✓	✓	4
Qiu et al. (2019)	✓		4
Kocabas et al. (2019b)			4
Iskakov et al. (2019)	✓		2
He et al. (2020)	✓		3
Huang et al. (2019)	✓	✓	4(8)
Zhang et al. (2020)	✓	✓	4(8)

Table 9. Multi-view models hardware-related assumptions. The number of camera view used to achieve less than the baseline 40mm MPJPE error (best results from monocular methods) on Human3.6M is also shown (and on TotalCapture under parenthesis).

methods that do not use or use partial calibration need more views to achieve acceptable accuracy, while others can produce good results with fewer cameras but need extrinsic parameters (see table 9).

5.2.3. *Single Person vs Multi-Person*

The strong assumptions that are still used for many 3D human pose estimation frameworks are related to camera or subject motion and acquisition protocol. One of them is the limitation to one person in the image. As top-down approaches are the most popular, many methods assume or even take as an input a single person. For this reason, some authors recommend the use of off-the-shelf person detectors to crop to the area of the image containing the individual subjects. In this way, the pose estimation algorithms can be applied to each area individually. However, this idea needs to be adapted in multi-view and temporal settings to track each different subject. The main limitation is when subject parts are overlapping in the image, hence indistinguishable by the person detectors. Some research addressed this issue in 2D, but it is still an open problem for 3D with few specific methods Mehta et al. (2018) Mehta et al. (2020).

5.2.4. *Motion Restriction*

Former markerless motion tracking systems were sometimes constrained to slow motion of only few limbs to perform good detection. It is less and less the case, but there is still some difficulty in predicting quick motions (e.g., in sports video). Cheng et al. (2020) suggest that temporal and spatial data at different resolutions can be a solution to this issue. A new assumption that can be added for methods based on temporal data : if the video footage is not long enough to provide sufficient information, this can be an issue for real-time inference and even for accuracy. Additionally, temporal methods often use information in the future frame, which is not suitable for real-time. Table 8 show that these methods perform best with varying temporal receptive fields. Some methods only need a few past and future images, while the accuracy of others saturates at more than 200. These methods can be adapted to shorter video clips and to real-time application using only past frames, but at the price of a decrease in accuracy. Another strong constraint is the use of video from moving cameras, but this is rarely addressed von Marcard et al. (2018). Many applications can work with the assumption of fixed cameras, but there is a substantial amount of video data produced with moving camera coordinate systems (e.g., in outdoor sports).

5.2.5. *Occlusions*

Recent solutions predict poses even in the presence of small occlusion, but this remains one of the main challenges in a monocular approach. In most real-world or multi-person scenarios, this needs to be addressed. To this end, many optimizations at training time are employed: data augmentation or occlusion-specific modules (Cheng et al. (2020); Cheng et al. (2019); Huang et al. (2019)). Another solution could be to consider the 3D scene around the subject. Hassan et al. (2019) show that combining 3D reconstruction of indoor scene and volumetric models of the human body can help overcome the occlusion issues. Their ”Proximal Relationships with Object eXclusion” method enforces physical constraints such as contacts with a static environment.

5.2.6. Generalization

Although neural network models are data-oriented algorithms, few analyses are performed on generalization to new contexts and in-the-wild situations. Protocols for benchmarks address generalization to different subjects, but the background scene and the actions rarely change. Multi-view models generalize best, with good performance in cross-validation between data sets, likely due to the geometric information provided by the camera projection matrices they often use. Yet, new benchmarks are necessary to properly assess generalization. Future research will likely provide carefully designed naturalistic datasets with new labeling solutions, or datasets augmented with new image processing techniques, and possibly benchmarks composed of synthetic poses, to challenge future methods.

5.2.7. Overall Conclusion on Robustness

Table 7 shows the less constrained methods and their performance. Robustness relates to the number of assumptions (the fewer the better). For monocular techniques, the multi-person methods trained on complex datasets containing severe occlusions logically impose fewer constraints. For temporal techniques, the ones that do not need future frames for inference perform best. Hossain and Little (2018) achieves maximum accuracy with the fewest images Liu et al. (2020). Cheng et al. (2020) address fast motion and occlusions but require a higher acquisition frequency and a wider temporal receptive field to produce better results. The most robust multi-view method is Iskakov et al. (2019) because it does not require any special hardware and can work with only two cameras while achieving acceptable accuracy (see table 9). Finally, von Marcard et al. (2018) can perform multi-person pose estimation in the wild with mobile cameras. Yet, the low constraints in terms of environment and subject come at the cost of high constraints in terms of additional hardware (IMU and scans).

Throughout this literature review, we have found that achieving the highest accuracy is the primary concern of novel methods. We also found that system robustness is regularly overlooked when evaluating algorithms and that the complexity of the novel technique is rarely considered, especially for fully engineered systems. Our impression is that the general emphasis on highest accuracy is somewhat misleading, and that the search for "good enough" systems depending on the application domain might be at least as important. For example, very high accuracy is mostly irrelevant for surveillance, but robustness and real time operation matters. For healthcare professionals monitoring patient recovery, robustness is paramount, provided that accuracy is enough (i.e., comparable to that of the human eye) but real-time operation only desirable.

Speed

Method	Accuracy	Robustness Level	Speed	GPU used
Mehta et al. (2017)	80.5	average	30fps	N/A
Wandt and Rosenhahn (2019)	50.9	average	20 000fps	Nvidia Titan X
Xu et al. (2019)	82.4	high	120fps	N/A
Mathis et al. (2018)	-	average	30fps	Nvidia 1080Ti
Mehta et al. (2020)	63.6	high	30fps	N/A
Hossain and Little (2018)	58.5	high	300fps	Nvidia Titan X
Liu et al. (2020)	45.1	average	3000fps	Nvidia Titan RTX
Pavllo et al. (2019)	46.8	average	150 000fps	Nvidia GP100
Trumble et al. (2017)	87.3	low	25fps	N/A
Huang et al. (2019)	37.5	low	25fps	Nvidia 1080Ti

Table 10. Real-time methods. Along with the other criteria for comparison, the speed (in frames inferred per second) and the graphical card used are reported. Note that methods such as Hossain and Little (2018) and Wandt and Rosenhahn (2019) are not considering the speed of the 2D detector in the first stage of their techniques when reporting fps.

5.3. Speed

It is easier to evaluate the performance in terms of speed by simply observing the complexity or the number of single operations needed for any one method (using floating-point operation *"FLOPs"*). When possible, knowing at what frame rate an inference can be made is also interesting for real-time application such as monitoring, surveillance or virtual reality. In these cases, the main constraint is the whole system latency, which should not exceed specified limits.

5.3.1. Real-time

Not all methods are complete motion capture systems such as Huang et al. (2019) or Mehta et al. (2017) that report 25 and 30 frames per second for detection. Some only provide the

inference timeper frame, without testing in a real life acquisition scenario such as Hossain and Little (2018) and Martinez et al. (2017) who each report 3ms per frame. The deepest architectures are often not suitable for real-time applications, as a forward pass in the network takes too long. To gain insight into the complexity of these models, one can look at the number of parameters that they learn.

5.3.2. *Training Time*

Most of the reviewed methods can be trained to adapt to new challenging contexts or refined on new data. It is also important to have an idea of the training time when an application might consider online training. The number of parameters is a good indication of the depth of the architecture (see Table 11), which can also be calculated using backbone 2D detection networks in many cases.

Method	#Parameters (in million)
Martinez et al. (2017)	4-5
Sun et al. (2018) Hourglass	26
Sun et al. (2018) Resnet50	26
Sun et al. (2018) Resnet101	45
Pavllo et al. (2019)	16.95
Qiu et al. (2019)	560
Iskakov et al. (2019) algebraic	80
Iskakov et al. (2019) volumetric	81
He et al. (2020)	69

Table 11. Number of reported parameters of different reviewed techniques.

5.3.3. *Overall Conclusion on Speed*

In the speed part of table 10, we report all methods that can run in real time. The inference speed is in frames per second and the GPU used is also specified. Robustness is variable in this collection of methods, and the accuracy seems a bit below average for the fastest methods. The most accurate is Huang et al. (2019), which achieves 37.5 MPJPE on Human3.6M without using its IMU component (it also performs well on TotalCapture with the addition of IMU data).

5.4. *Recommendations for users*

After this analysis according to the performance criteria, here are our suggestions for the different types of application :

Human Computer Interface: for this category of applications, the priority is good accuracy and real-time performance. Robustness depends on whether the system operates in a pre-defined environment or "in-the-wild". With these specifications, temporal methods that can run in real-time such as Pavllo et al. (2019) or Liu et al. (2020) correspond best. Hossain and Little (2018) requires fewer frames to be computed but its accuracy is lower. Other more robust methods could be multi-person ones, such as Mehta et al. (2020), but control applications usually only interact with one subject. Finally, Huang et al. (2019) is the most accurate method running in real-time, but it requires multiple views to become robust.

Security: Traditional monitoring systems generally need to be more robust to operate in varied and changing real-world environments and subjects. Speed is also important because infractions need to be identified in real-time. Finally, the higher precision is less important because it is the whole semantics of the movement that is important to identify the subject's actions. If you are looking for methods running in real-time with the highest robustness, consider monocular (Xu et al. (2019)) and temporal (Hossain and Little (2018)) methods. The methods proposed by Mehta et al. (2020) allows for an even greater robustness with multi-person detection in real-time, if necessary. Less robust methods can also be considered for higher accuracy (e.g. Pavllo et al. (2019) or Liu et al. (2020)) or speed (e.g. Wandt and Rosenhahn (2019)).

Motion Analysis: these applications typically run in lab-controlled environment and computation is performed offline. Accuracy is the most important criteria with less consideration for real-time or high robustness. However, human performance captured in "real-life" scenarios and the need of less intrusive techniques (e.g., medical diagnosis or rehabilitation) demands

better robustness than classical marker-based methods (Colyer et al. (2018)). Multi-view configurations now produce results with an average error per joint of less than 30 mm (Iskakov et al. (2019), He et al. (2015)). When multiple cameras or calibration are not possible, monocular temporal methods such as Kolotouros et al. (2019) or Cheng et al. (2020) can be considered. Finally, if simplicity and usability are required, easy-to-use and flexible systems based on good 2D detectors such as Mathis and Mathis (2019) or Martinez et al. (2017) produce efficient results.

Entertainment: Motion capture for animation or VR is usually done in a controlled environment, so robustness is not the main concern. Accuracy is important because the generated poses and movements must be realistic (see 2.1 for structural and perceptual metrics). Finally, speed is less important for offline processing that generates avatar or animated characters, whereas real-time is needed for controls in video games or VR. In the first case, multi-view markerless systems can replace classical motion capture systems at a lower cost (e.g. Iskakov et al. (2019) or He et al. (2020)). IIn the second case, real-time operation is possible with Huang et al. (2019), which also provides high accuracy in multi-view configurations (and even higher if adding IMUs to the subject is feasible).

5.5. *Challenge for future research*

Many problems concerning the estimation of the human pose in 3D have still not been solved. The most discussed is that the accuracy is still insufficient for some applications (e.g., in motion analysis). Currently, many different monocular and temporal approaches achieve an average joint position error of about 40 mm, while multi-view approaches achieve about 20-30 mm. This leads to important challenges:

For *monocular* techniques, the obstacle to their widespread use is the consistency of their detection over space and time. Difference in accuracy among joints needs to be addressed and temporal consistency (reducing jitter) of motions needs to be better enforced. Reaching 90% of joints accurately detected, as 2D pose estimators do, is a goal for the coming years (the average 3DPCK reaches about 85% with evaluations on MPI-INF-3DHP).

For *multi-view* setups, the issue is that they present results close to those of the best monocular methods despite the access to epipolar geometry and 3D information. The combination of recent advances in deep learning (e.g., recurrent network, attention, etc.) and strong prior information about scene structures from camera calibration could take this a step further. Another avenue for applications in more controlled environments could be the use of other modalities such as IMU or depth sensors. Multi-modal information fusion could then lead to even better results.

Real-time performances are achieved by many methods, but using powerful graphical cards. Porting real-time estimation to the average commercial equipment remains a challenge. Currently, most proposals rely on multi-stage computations, but future research could draw on single-shot object detectors (e.g. Liu et al. (2016) (SSD), Redmon et al. (2016) (YOLO)) to produce reasonable results faster. Similarly, research on real-time 3D human pose estimation with multi-view cameras is developing. Virtual reality could benefit from such systems for gesture-based control.

Finally, there are strong assumptions about occlusions and multi-person configurations that do not yet allow pose estimation to be applied to any video or image. Complex in-the-wild data sets, sometimes including difficult poses (generated or captured by motion), are beginning to emerge, which suggests that these are promising research questions.

6. Conclusion

Human pose estimation has been one of the focal points of computer vision in recent years. Deep learning has improved the results by a significant amount. However, the most accurate techniques use various architectures (temporal convolutional networks, 3D human body models or learnable triangulation) depending on the input data (single images, videos sequence or multi-view images). What these methods have in common is the use of 2D detection as an intermediate step. However,

the diversity of approaches in the new state-of-the art methods demonstrates that a consensus has not yet been established.

Improved results in the near future will require richer datasets, computational parsimony, and ease of use. Access to new and richer datasets, including a wide variety of poses, movements, and contexts, is essential for robustness. This could be facilitated by new learning processes working with partially annotated data based on monocular videos, but also by commercial markerless tracking tools, which could feed these repositories with better images in "natural conditions", without markers and in an outdoor environment. Parsimony in terms of computational cost paves the way for real-time operation, less expensive hardware architectures, and a welcome reduction in power consumption. Ease of use, as was the case with DeepLabCut, is a key to widespread adoption, which in turn challenges the algorithms to solve new questions and use cases.

Finally, we believe that important future developments could come from new ways of approaching the full richness of human pose estimation, by reasoning directly on the 3D nature of the problem (using voxel map representations, multi-channel volumes...), by employing less constrained multi-view approaches that can function with fewer cameras, or by transposing temporal methods (transformers architecture or temporal convolutional networks) from the monocular to the multi-view.

Acknowledgments

This project was supported by the LabEx NUMEV (ANR-10-LABX-0020) within the I-SITE MUSE. This research was partially supported by the HUT project co-financed by the European Regional Development Fund (ERDF) and the Occitanie Region. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Agarwal, A., Triggs, B., 2006. Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 44–58. doi:10.1109/TPAMI.2006.21. conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Davis, J., Rodgers, J., . SCAPE: Shape Completion and Animation of People , 9.

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2014. 3D Pictorial Structures for Multiple Human Pose Estimation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Columbus, OH, USA. pp. 1669–1676. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909612, doi:10.1109/CVPR.2014.216.

Bray, J., 2000. Markerless Based Human Motion Capture : A Survey.

Burenius, M., Sullivan, J., Carlsson, S., 2013. 3d pictorial structures for multiple view articulated pose estimation, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3625.

Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M., 2019. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South). pp. 2272–2281. URL: https://ieeexplore.ieee.org/document/9009459/, doi:10.1109/ICCV.2019.00236.

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1302–1310. doi:10.1109/CVPR.2017.143. iSSN: 1063-6919.

Chen, Y., Tian, Y., He, M., 2020. Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods. Computer Vision and Image Understanding 192, 102897. URL: http://arxiv.org/abs/2006.01423, doi:10.1016/j.cviu.2019.102897. arXiv: 2006.01423.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded Pyramid Network for Multi-Person Pose Estimation. arXiv:1711.07319 [cs] URL: http://arxiv.org/abs/1711.07319. arXiv: 1711.07319.

Cheng, Y., Yang, B., Wang, B., Tan, R.T., 2020. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. arXiv:2004.11822.

Cheng, Y., Yang, B., Wang, B., Yan, W., Tan, R.T., 2019. Occlusion-Aware Networks for 3D Human Pose Estimation in Video, pp. 723–732. URL: http://openaccess.thecvf.com/content_ICCV_2019/html/Cheng_Occlusion-Aware_Networks_for_3D_Human_Pose_Estimation_in_Video_ICCV_2019_paper.html.

Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.I.T., 2018. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Medicine - Open 4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986692/, doi:10.1186/s40798-018-0139-y.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial Structures for Object Recognition. International Journal of Computer Vision 61, 55–79. URL: https://doi.org/10.1023/B:VISI.0000042934.15159.49, doi:10.

1023/B:VISI.0000042934.15159.49.

Fiker, R., Kim, L.H., Molina, L.A., Chomiak, T., Whelan, P.J., 2020. Visual deep lab cut: A user-friendly approach to gait analysis. Journal of Neuroscience Methods , 108775URL: http://www.sciencedirect.com/science/article/pii/S0165027020301989, doi:https://doi.org/10.1016/j.jneumeth.2020.108775.

Fischler, M., Elschlager, R., 1973. The Representation and Matching of Pictorial Structures. IEEE Transactions on Computers C-22, 67–92. URL: http://ieeexplore.ieee.org/document/1672195/, doi:10.1109/T-C.1973.223602.

Ghorbani, S., Mahdaviani, K., Thaler, A., Kording, K., Cook, D.J., Blohm, G., Troje, N.F., 2020. MoVi: A Large Multipurpose Motion and Video Dataset. arXiv:2003.01888 [cs, eess] URL: http://arxiv.org/abs/2003.01888. arXiv: 2003.01888.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. arXiv:1406.2661 [cs, stat] URL: http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

Grauman, K., Shakhnarovich, G., Darrell, T., 2003. A Bayesian approach to image-based visual hull reconstruction, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., IEEE Comput. Soc, Madison, WI, USA. pp. I–187–I–194. URL: http://ieeexplore.ieee.org/document/1211353/, doi:10.1109/CVPR.2003.1211353.

Güler, R.A., Neverova, N., Kokkinos, I., 2018. DensePose: Dense Human Pose Estimation In The Wild. arXiv:1802.00434 [cs] URL: http://arxiv.org/abs/1802.00434. arXiv: 1802.00434.

Hassan, M., Choutas, V., Tzionas, D., Black, M., 2019. Resolving 3D Human Pose Ambiguities With 3D Scene Constraints, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South). pp. 2282–2292. URL: https://ieeexplore.ieee.org/document/9010321/, doi:10.1109/ICCV.2019.00237.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] URL: http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

He, Y., Yan, R., Fragkiadaki, K., Yu, S.I., 2020. Epipolar Transformers. arXiv:2005.04551 [cs] URL: http://arxiv.org/abs/2005.04551. arXiv: 2005.04551.

Hossain, M.R.I., Little, J.J., 2018. Exploiting temporal information for 3D pose estimation. arXiv:1711.08585 [cs] 11214, 69–86. URL: http://arxiv.org/abs/1711.08585, doi:10.1007/978-3-030-01249-6_5. arXiv: 1711.08585.

Huang, C.H., Allain, B., Boyer, E., Franco, J.S., Tombari, F., Navab, N., Ilic, S., 2018. Tracking-by-Detection of 3D Human Shapes: from Surfaces to Volumes. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 1994–2008. URL: https://hal.inria.fr/hal-01588272, doi:10.1109/TPAMI.2017.2740308.

Huang, F., Zeng, A., Liu, M., Lai, Q., Xu, Q., 2019. DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image. arXiv:1912.04071 [cs] URL: http://arxiv.org/abs/1912.04071. arXiv: 1912.04071.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. CoRR abs/1605.03170. URL: http://arxiv.org/abs/1605.03170, arXiv:1605.03170.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1325–1339. URL: http://ieeexplore.ieee.org/document/6682899/, doi:10.1109/TPAMI.2013.248.

Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., 2019. Learnable Triangulation of Human Pose. arXiv:1905.05754 [cs] URL: http://arxiv.org/abs/1905.05754. arXiv: 1905.05754.

Kanko, R., Strutzenberger, G., Brown, M., Selbie, S., Deluzio, K., 2020. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. preprint. engrXiv. URL: https://osf.io/j4rbg, doi:10.31224/osf.io/j4rbg.

Kipf, T.N., Welling, M., 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs, stat] URL: http://arxiv.org/abs/1609.02907. arXiv: 1609.02907.

Kocabas, M., Athanasiou, N., Black, M.J., 2019a. VIBE: Video Inference for Human Body Pose and Shape Estimation. arXiv:1912.05656 [cs] URL: http://arxiv.org/abs/1912.05656. arXiv: 1912.05656.

Kocabas, M., Karagoz, S., Akbas, E., 2019b. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. arXiv:1903.02330 [cs] URL: http://arxiv.org/abs/1903.02330. arXiv: 1903.02330.

Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K., 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. arXiv:1909.12828 [cs] URL: http://arxiv.org/abs/1909.12828. arXiv: 1909.12828.

Kudo, Y., Ogaki, K., Matsui, Y., Odagiri, Y., 2018. Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations. arXiv:1803.08244 [cs] URL: http://arxiv.org/abs/1803.08244. arXiv: 1803.08244.

Lin, G., Milan, A., Shen, C., Reid, I., 2016. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. arXiv:1611.06612 [cs] URL: http://arxiv.org/abs/1611.06612. arXiv: 1611.06612.

Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V., 2020. Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA. pp. 5063–5072. URL: https://ieeexplore.ieee.org/document/9156272/, doi:10.1109/CVPR42600.2020.00511.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: Single Shot MultiBox Detector. arXiv:1512.02325 [cs] 9905, 21–37. URL: http://arxiv.org/abs/1512.02325, doi:10.1007/978-3-319-46448-0_2. arXiv: 1512.02325.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J., 2015. SMPL: a skinned multi-person linear model. ACM Transactions on Graphics 34, 1–16. URL: https://dl.acm.org/doi/10.1145/2816795.2818013, doi:10.1145/2816795.2818013.

Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J., 2019. Amass: Archive of motion capture as surface shapes, in: The IEEE International Conference on Computer Vision (ICCV). URL: https://amass.is.tue.mpg.de.

von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G., 2018. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham. volume 11214, pp. 614–631. URL: http://link.springer.com/10.1007/978-3-030-01249-6_37, doi:10.1007/978-3-030-01249-6_37. series Title: Lecture Notes in Computer Science.

von Marcard, T., Pons-Moll, G., Rosenhahn, B., 2016. Human pose estimation from video and imus. Transactions on Pattern Analysis and Machine Intelligence 38, 1533–1547. URL: http://dx.doi.org/10.1109/TPAMI.2016.2522398, doi:10.1109/TPAMI.2016.2522398.

Marinoiu, E., Papava, D., Sminchisescu, C., 2013. Pictorial human spaces: How well do humans perceive a 3d articulated pose?, in: 2013 IEEE International Conference on Computer Vision, pp. 1289–1296.

Marinoiu, E., Papava, D., Sminchisescu, C., 2016. Pictorial human spaces: A computational study on the human perception of 3d articulated poses. International Journal of Computer Vision 119, 194–215. URL: http://dx.doi.org/10.1007/s11263-016-0888-3, doi:10.1007/s11263-016-0888-3.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. arXiv:1705.03098 [cs] URL: http://arxiv.org/abs/1705.03098. arXiv: 1705.03098.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience 21, 1281–1289. URL: https://www.nature.com/articles/s41593-018-0209-y, doi:10.1038/s41593-018-0209-y. number: 9 Publisher: Nature Publishing Group.

Mathis, M.W., Mathis, A., 2019. Deep learning tools for the measurement of animal behavior in neuroscience. arXiv:1909.13868 [cs, q-bio] URL: http://arxiv.org/abs/1909.13868. arXiv: 1909.13868.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2016. [1611.09813] Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. URL: https://arxiv.org/abs/1611.09813.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C., 2020. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. ACM Transactions on Graphics 39. URL: http://arxiv.org/abs/1907.00837, doi:10.1145/3386569.3392410. arXiv: 1907.00837.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., 2018. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. arXiv:1712.03453 [cs] URL: http://arxiv.org/abs/1712.03453. arXiv: 1712.03453.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C., 2017. VNect: real-time 3D human pose estimation with a single RGB camera. ACM Transactions on Graphics 36, 1–14. URL: http://dl.acm.org/citation.cfm?doid=3072959.3073596, doi:10.1145/3072959.3073596.

Moeslund, T.B., Granum, E., 2001. A Survey of Computer Vision-Based Human Motion Capture. Computer Vision and Image Understanding 81, 231–268. URL: http://www.sciencedirect.com/science/article/pii/S107731420090897X, doi:10.1006/cviu.2000.0897.

Moro, M., Marchesi, G., Odone, F., Casadio, M., 2020. Markerless gait analysis in stroke survivors based on computer vision and deep learning: A pilot study, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, Association for Computing Machinery, New York, NY, USA. p. 2097–2104. URL: https://doi.org/10.1145/3341105.3373963, doi:10.1145/3341105.3373963.

Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W., 2018. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. preprint. Neuroscience. URL: http://biorxiv.org/lookup/doi/10.1101/476531, doi:10.1101/476531.

Newell, A., Yang, K., Deng, J., 2016. Stacked Hourglass Networks for Human Pose Estimation. arXiv:1603.06937 [cs] URL: http://arxiv.org/abs/1603.06937. arXiv: 1603.06937.

Olah, C., . Understanding LSTM Networks – colah's blog. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B., 2018. Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation. arXiv:1808.05942 [cs] URL: http://arxiv.org/abs/1808.05942. arXiv: 1808.05942.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. arXiv:1611.07828 [cs] URL: http://arxiv.org/abs/1611.07828. arXiv: 1611.07828.

Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M., 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA. pp. 7745–7754. URL: https://ieeexplore.ieee.org/document/8954163/, doi:10.1109/CVPR.2019.00794.

Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W., 2019. Cross View Fusion for 3D Human Pose Estimation. arXiv:1909.01203 [cs] URL: http://arxiv.org/abs/1909.01203. arXiv: 1909.01203.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs] URL: http://arxiv.org/abs/1506.02640. arXiv: 1506.02640.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks

for Biomedical Image Segmentation. arXiv:1505.04597 [cs] URL: http://arxiv.org/abs/1505.04597. arXiv: 1505.04597.

Sarafinos, Boteanu, Ionescu, Kakadiaris, 2016. (PDF) 3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates. URL: https://www.researchgate.net/publication/307905073_3D_Human_Pose_Estimation_A_Review_of_the_Literature_and_Analysis_of_Covariates. library Catalog: www.researchgate.net.

Sheshadri, S., Dann, B., Hueser, T., Scherberger, H., 2020. 3D reconstruction toolbox for behavior tracked with multiple cameras. Journal of Open Source Software 5, 1849. URL: https://joss.theoj.org/papers/10.21105/joss.01849, doi:10.21105/joss.01849.

Sigal, L., Balan, A.O., Black, M.J., 2010. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. International Journal of Computer Vision 87, 4–27. URL: http://link.springer.com/10.1007/s11263-009-0273-6, doi:10.1007/s11263-009-0273-6.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. arXiv:1902.09212.

Sun, X., Shang, J., Liang, S., Wei, Y., 2017. Compositional Human Pose Regression. arXiv:1704.00159 [cs] URL: http://arxiv.org/abs/1704.00159. arXiv: 1704.00159.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral Human Pose Regression. arXiv:1711.08229 [cs] URL: http://arxiv.org/abs/1711.08229. arXiv: 1711.08229.

Tompson, J., Jain, A., LeCun, Y., Bregler, C., 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. arXiv:1406.2984 [cs] URL: http://arxiv.org/abs/1406.2984. arXiv: 1406.2984.

De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., Beltran, P., 2008. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. URL: https://www.ri.cmu.edu/publications/guide-to-the-carnegie-mellon-university-multimodal-activity-cmu-mmac-database/. library Catalog: www.ri.cmu.edu.

Toshev, A., Szegedy, C., 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition , 1653–1660URL: http://arxiv.org/abs/1312.4659, doi:10.1109/CVPR.2014.214. arXiv: 1312.4659.

Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors, in: Procedings of the British Machine Vision Conference 2017, British Machine Vision Association, London, UK. p. 14. URL: http://www.bmva.org/bmvc/2017/papers/paper014/index.html, doi:10.5244/C.31.14.

University, C.M., . Graphic lab motion capture database. http://mocap.cs.cmu.edu/. URL: http://mocap.cs.cmu.edu/. accessed: 2020-04-29.

Wandt, B., Rosenhahn, B., 2019. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. arXiv:1902.09868 [cs] URL: http://arxiv.org/abs/1902.09868. arXiv: 1902.09868.

Wang, J., Yan, S., Xiong, Y., Lin, D., 2020. Motion Guided 3D Pose Estimation from Videos, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham. volume 12358, pp. 764–780. URL: http://link.springer.com/10.1007/978-3-030-58601-0_45, doi:10.1007/978-3-030-58601-0_45. series Title: Lecture Notes in Computer Science.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. arXiv:1804.06208.

Xu, Y., Zhu, S.C., Tung, T., 2019. DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. arXiv:1910.00116 [cs, eess] URL: http://arxiv.org/abs/1910.00116. arXiv: 1910.00116.

Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y., 2017. BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice. pp. 910–919. URL: http://ieeexplore.ieee.org/document/8237366/, doi:10.1109/ICCV.2017.104.

Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y., 2018. DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor. arXiv:1804.06023 [cs] URL: http://arxiv.org/abs/1804.06023. arXiv: 1804.06023.

Zhang, Y., Park, H.S., 2020. Multiview Supervision By Registration, pp. 420–428. URL: http://openaccess.thecvf.com/content_WACV_2020/html/Zhang_Multiview_Supervision_By_Registration_WACV_2020_paper.html.

Zhang, Z., Wang, C., Qin, W., Zeng, W., 2020. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. arXiv:2003.11163.

Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y., 2017. Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach. arXiv:1704.02447 [cs] URL: http://arxiv.org/abs/1704.02447. arXiv: 1704.02447.