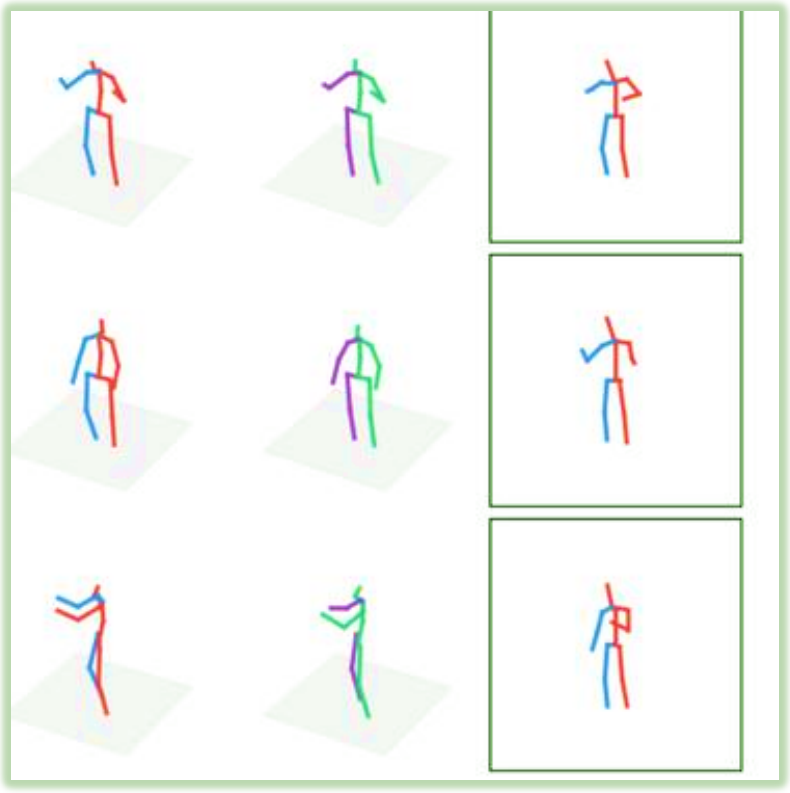


PROJECT “STICK ME”

YIELDING THE 3D HUMAN POSE FROM A MONOCULAR RBG IMAGE: A REVIEW

Samira Sriram | Satva | July 2020

SATVA

TABLE OF CONTENTS

INTRODUCTION……..2

PROCESS……...2

STAGE 1: JOINT ANNOTATIONS……..3

STAGE 2: OUTPUT 3D FIGURE……6

END-TO-END ARCHITECTURES……...8

RECOMMENDATION……..10

INTRODUCTION

Human pose estimation is a fundamental problem in computer vision. Computer's ability to recognize and understand humans in images and videos is crucial for multiple tasks including autonomous driving, action recognition, human-computer interaction, augmented reality and robotics vision.

In recent years, significant progress has been achieved in 2D human pose estimation. The crucial factor behind this success is the availability of large-scale annotated human pose datasets that allow training networks for 2D human pose estimation. At the same time, advances in 3D human pose estimation remain limited because obtaining ground-truth information on the dense correspondence, depth, motion, body-part segmentation, occlusions is a very challenging task.

Both two-stage and end-to-end architectures are explored below. Some of these methods include jointly training 2D and 3D ground truth, using 2D joint annotations as an intermediate, using spatial heatmaps and a geometric approach to obtain 3D information, and using k-nearest neighbors on pre-trained data sets are some of the methods researchers have used to optimize PCK (percentage of correct key points) and MJPE. Based on these key metrics, a recommendation is then proposed.

PROCESS

The process of converting a monocular RBG image to a 3D stick figure has two possible routes of completion.

- End-to-end architecture: This takes a natural image and outputs a 3D stick figure by jointly training 2D and 3D annotations, and/or uses 2D annotations as an intermediate factor.
- Two-stage: This uses two separate models. The former trains a 2D pose and outputs joint annotations to be used in step 2. The latter step uses the 2D ground truth as an input to then output a 3D stick figure.

CONSIDERATIONS

1. We are working with "natural" images. There is a distinction between "in-the-wild" images, "natural" images, and 3d pose datasets. Natural images provide the least information where both 2D and 3D ground truth are unknown. "In-the-wild" images have 2D joint annotations. Intrinsic camera parameters of these type of images may or may not be known.
2. Widely used 2D datasets: MPII, LSP, FLIC, FLIC plus are "in-the-wild," not natural as they have 14+ key points + joint coordinates already identified
3. Widely used 3D datasets: HumanEva, Human3.6M.

Under these considerations, our constraint is as follows: Finding the 3D figure output with no known camera parameters or ground truth and merging the 2D joint output with the input of many 3D pose architectures.

STAGE 1: JOINT ANNOTATIONS

Below is an analysis of popular joint annotation methodology:

DEEPCUT [APR 2016]

This joint formulation is in contrast to previous strategies, that address the problem by first detecting people and subsequently estimating their body pose. Deepcut[1] proposes a partitioning and labeling formulation of a set of body-part hypotheses generated with FR-CNN-based part detectors. Deepcut excels in multi-person detection and is tested on the LSP and MPII 2D datasets. Softmax 80.8-82.8% PCK.

TOMPSON NIPS'14 [2014]

The first stage of their detection pipeline is a deep ConvNet architecture for body part localization. The input is an RGB image containing one or more people and the output is a heat-map, which produces a per-pixel likelihood for key joint locations on the human skeleton. The network is slid over the input image to produce a dense heat-map output for each body-joint. This model incorporates a multi-resolution input with overlapping receptive fields. A Mean Squared Error (MSE) criterion is used to minimize the distance between the predicted output and a target heat-map. The target is a 2D Gaussian with a small variance and mean centered at the ground-truth joint locations.[2] This model was tested on the FLIC and extended-LSP dataset. >90% detection rate on torso, ankle, feet and hands.

DEEPPOSE [DEC 2017]

Deeppose represents a simplified DNN-based joint-regression problem. Each joint regressor uses the full image as a signal. Deeppose has no need to explicitly design feature representations and detectors for parts; no need to explicitly design a model topology and interactions between joints.[3] Deeppose is trained and tested on the LSP and FLIC 2D datasets. .61 PCP and ~92% detection rate.

DENSEPOSE [FEB 2018]

Meant for in-the-wild type data.[4] Densepose trains CNN-based systems that deliver dense correspondence 'in the wild', namely in the presence of background, occlusions and scale variations. They improve their training set's effectiveness by training an 'inpainting' network that can fill in missing ground truth values. They use the 2D COCO dataset, the

[1] https://arxiv.org/pdf/1511.06645
[2] https://arxiv.org/pdf/1406.2984
[3] https://arxiv.org/pdf/1312.4659
[4] https://arxiv.org/pdf/1802.00434

SMPL model, and train on ResNet50, as well as human annotators to establish those dense correspondences. .835 ratio of correct keypoints.

CONVOLUTIONAL POSE MACHINES [APR 2016]

The prediction and image feature computation modules of a pose machine can be replaced by a deep convolutional architecture allowing for both image and contextual feature representations to be learned directly from data. Convolutional architectures also have the advantage of being completely differentiable, thereby enabling end-to-end joint training of all stages of a CPM. The first stage of a convolutional pose machine predicts part beliefs from only local image evidence. Replacing the modules of a pose machine with the appropriately designed convolutional architecture provides a large boost of 42.4 percentage points over the previous approach in the high precision regime and 30.9 percentage points in the low precision regime. 84% PCK average and 90% PCKh on LSP dataset.[5]

SIMPLEPOSE [NOV 2019]

SimplePose performs the task in three steps: (1) predict the keypoint heatmaps and the body part heatmaps of all persons in a given image, (2) get the candidate keypoints and body parts by performing Non-Maximum Suppression (NMS) on the inferred heatmaps, and (3) the keypoint assignment algorithm is performed and collect all individual poses. Their models are trained and evaluated on the MS-COCO dataset.[6] 70% PCK.

METRICS OF SUCCESS

Bounding Boxes:

1. Percentage of Correct parts (PCP): measures detection rate of limbs, where a limb is considered detected if the distance between the two predicted joint locations and the true limb joint locations is at most half of the limb length.
2. Percent of Detected Joints (new and improved PCP) (PDJ): A joint is considered detected if the distance between the predicted and the true joint is within a certain fraction of the torso diameter.
3. For a given normalized pixel radius, the number of images in the test-set for which the distance of the predicted joint location to the ground-truth location falls within the given radius.

Joint Coordinates:

1. Percentage of Correct Keypoints (PCK): measures the accuracy of the localization of the body joints. The threshold for matching of the joint position to the ground truth is defined as a fraction of the person bounding box size.
2. Percent of Correct Keypoints (PCKh): is advanced iteration of PCK. It measures the ratio of estimated joints for which the distance to the ground-truth is below a

[5] https://arxiv.org/pdf/1602.00134

[6] https://arxiv.org/pdf/1911.10529

threshold. The standard threshold is set to half the size of the head, i.e., PCKh@0.5.

COMPARISON TABLE

METHOD	PCK/PCKH	PCP/PDJ
Deepcut	80.8	
Tompson NIPS'14		90
DeepPose	61	92
DensePose	83.5	
Convolutional Pose Machine	84	
SimplePose	70	

STAGE 2: OUTPUT 3D FIGURE

SMPL/ SMPLIFY-X [2016]

The full 3D mesh is estimated showing that 2D joints alone carry a surprising amount of information about body shape. They use a CNN-based method, DeepCut, to predict (bottom-up) the 2D body joint locations. They then fit (top-down) a recently published statistical body shape model, called SMPL, to the 2D joints. They do so by minimizing an objective function that penalizes the error between the projected 3D model joints and detected 2D joints. Because SMPL captures correlations in human shape across the population, they are able to robustly fit it to very little data. Tested on LSP, Human3.6m and HumanEva datasets.[7] MPJPE = 61.9 mm

MONOCAP [MAR 2018]

Treats 2D joint locations as latent variables whose uncertainty distributions are given by a deep fully convolutional neural network. Unknown 3D poses are modeled by a sparse representation and the 3D parameter estimates are realized via an Expectation-Maximization algorithm, where it is shown that the 2D joint location uncertainties can be conveniently marginalized out during inference. Surpasses benchmarks for "in-the-wild" data (annotated 2D data). MPII Dataset used. Uses MoCap dataset as a pose "directory". No intrinsic camera viewpoints: orthographic camera model is used to describe the dependence between a 3D pose and its imaged 2D pose. Tested on Human3.6M, Human Eva I, KTH Football II, and MPII.[8] MPJPE of 86.

3D HUMAN POSE MACHINES WITH SELF-SUPERVISED LEARNING [JAN 2019]

This model learns to integrate rich spatial and temporal long-range dependencies as well as 3D geometric constraints, rather than relying on specific manually defined body smoothness or kinematic constraints; ii) Developing a simple yet effective self-supervised correction mechanism to incorporate 3D pose geometric structural information is innovative in literature, and may also inspire other 3D vision tasks; iii) The proposed self-supervised correction mechanism enables the model to significantly improve 3D human pose estimation via sufficient 2D human pose data. Trained on the HumanEva and Human3.6M datasets.[9] 59.41 MPJPE.

MONOCULAR TOTAL CAPTURE: POSING IN THE WILD

An efficient representation called 3D Part Orientation Fields (POFs) is used to encode the 3D orientations of all body parts in the common 2D image space. POFs are predicted by a Fully Convolutional Network (FCN), along with the joint confidence maps. To train the network, a new 3D human motion dataset was collected capturing diverse total body

[7] https://people.eecs.berkeley.edu/~kanazawa/papers/SMPLify.pdf

[8] https://arxiv.org/pdf/1701.02354.pdf

[9] https://arxiv.org/pdf/1901.03798.pdf

motion of 40 subjects in a multi-view system. A 3D deformable human model was leveraged to reconstruct total body pose from the CNN outputs by exploiting the pose and shape prior in the model. A texture-based tracking method is also presented to obtain temporally coherent motion capture output.[10] MPJPE is 63 mm.

IN THE WILD HUMAN POSE ESTIMATION USING EXPLICIT 2D FEATURES AND INTERMEDIATE 3D REPRESENTATION

It has a network architecture that comprises a new disentangled hidden space encoding of explicit 2D and 3D features and uses supervision by a new learned projection model from predicted 3D pose. The algorithm can be jointly trained on image data with 3D labels and image data with only 2D labels. It achieves state-of-the-art accuracy on challenging in-the-wild data. Adds on to the SMPL model. Their camera network predicts the principal coordinate (cx, cy) and the focal length (αx, αy) parameters of a weak perspective camera model from the given input image. MPJPE is 65.4 and PCK is 91.3.[11]

METRICS OF SUCCESS

1. Average Precision (AP) at a number of geodesic point similarity (GPS) thresholds
2. Mean Per Joint Position Error (MPJPE). Per joint position error is the Euclidean distance between ground truth and prediction for a joint. Mean per joint position error is the Mean of per joint position error for all N joints (Typically, N = 16). Calculated after aligning the root joints (typically the pelvis) of the estimated and ground truth 3D pose. The joints are also normalized wrt root joints.

COMPARISON TABLE

METHOD	PCK/PCKH	MPJPE (MILLIMETERS)
SMPL		61.9
Monocap		86
Self-Supervised Learning		59.41
Monocular Total Capture		63
In the Wild Human Pose Estimation	91.3	65.4

[10] https://arxiv.org/pdf/1812.01598.pdf
[11] https://arxiv.org/pdf/1904.03289.pdf

END-TO-END ARCHITECTURES

SINGLE SHOT MULTI-PERSON 3D POSE ESTIMATION FROM MONOCULAR RGB [AUG 2018]

Dataset "MuCo3DHP" is proposed using ground truth from multi-view performance capture i.e., 3D pose annotations are available. The approach uses occlusion-robust pose-maps (ORPM) which enables full body pose inference even under strong partial occlusions by other people and objects in the scene. ORPM outputs a fixed number of maps which encode the 3D joint locations of all people in the scene.[12] MPJPE is 69 mm.

OPENPOSE [DEC 2018]

OpenPose uses PAF refinement to maximize accuracy. They increase the network depth but remove the body part refinement stages. This refined network increases both speed and accuracy by approximately 200% and 7%. They also present an annotated foot dataset with 15K human foot instances that has been publicly released and show that a combined model with body and foot keypoints can be trained preserving the speed of the body-only model while maintaining its accuracy.[13] MPJPE is 70.3.

3D HUMAN POSE ESTIMATION = 2D POSE ESTIMATION + MATCHING [APR 2017]

Uses nearest neighbors on the MoCap dataset (i.e., matching) with the 2D joint annotations from a DNN to get the figure. Their entire pipeline returns a 3D pose given a 2D image in under 200ms (160ms for 2D estimation by a CNN, 26ms for exemplar matching with a training library of 200,000 poses).[14] 5.95 MJPJE.

TOWARDS 3D HUMAN POSE ESTIMATION IN THE WILD: A WEAKLY-SUPERVISED APPROACH [JULY 2017]

This model is a weakly-supervised transfer learning method that uses mixed 2D and 3D labels in a unified deep neural network that presents a two-stage cascaded structure. The two stages are merged, obtaining 2D annotations of in the wild data, and obtaining 3D pose from existing MoCap data all in its' architecture. They propose a 3D geometric constraint for 3D pose estimation from images with only 2D joint annotations. They combine the 2D joint heatmaps and the intermediate feature representations in the 2D module as input to the depth regression module. These features, which extract semantic information at multiple levels for 2D pose estimation, provide additional cues for 3D pose recovery. It is trained on MPII and Human3.6M.[15] 64.9 MJPJE and 72.5 PCK.

---

[12] https://arxiv.org/pdf/1712.03453

[13] https://arxiv.org/pdf/1812.08008.pdf

[14] https://arxiv.org/pdf/1612.06524.pdf

[15] https://arxiv.org/pdf/1704.02447.pdf

LCR-NET ++ [JAN 2019]

Localization Classification-Regression Network: 1) the pose proposal generator that suggests candidate poses at different locations in the image; 2) a classifier that scores the different pose proposals; and 3) a regressor that refines pose proposals both in 2D and 3D. Uses 13 joints. The set of 2D and 3D joint annotations lend itself to pose proposals which consist of a set of candidate locations where the anchor-poses are hypothesized. Then iterative estimation combines the pose proposals with anchor-pose weights to output the true pose. Infer ground-truth 3D poses from 2D annotations using a nearest neighbor (NN) search performed on the annotated joints. Trained on MPII and LSP.[16] 54.2 mm MJPJE.

MEBOW [2020]

MEBOW is a large-scale high-precision human body orientation dataset. A simple baseline model for HBOE was established, which, when trained with MEBOW, is shown to significantly outperform state-of-the-art models trained on existing dataset. The first triple-source solution for 3-D human pose estimation was developed using this dataset as one of the three supervision sources, and it significantly outperforms a state-of-the-art dual-source solution for 3-D human pose estimation. This validates a new direction of improving 3-D human pose estimation by using significantly lower-cost labels (i.e., body orientation).[17] MPJPE is average of 50.9 average.

COMPARISON TABLE

METHOD	MPJPE (MILLIMETERS)
Single Shot	69
OpenPose	70.3
2D Pose Estimation + Matching	59.5
Towards 3D Human Pose Estimation	64.9
LCRNET ++	54.2
MEBOW	50.9

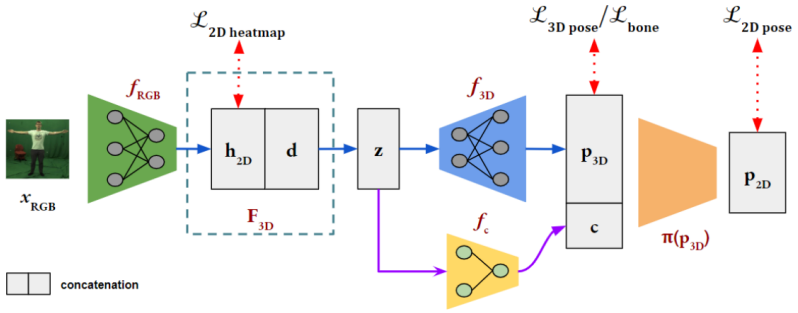
[16] https://arxiv.org/pdf/1803.00455.pdf

[17] http://infolab.stanford.edu/~wangz/project/imsearch/BODY/CVPR20/wu.pdf

RECOMMENDATION

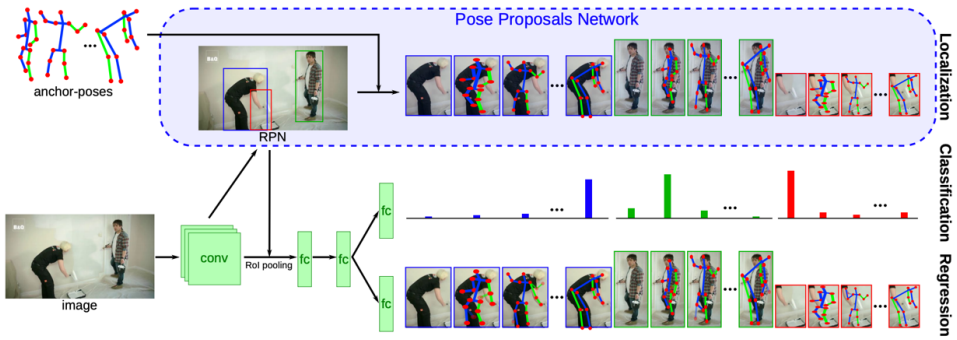
TWO STAGE

For a two-stage architecture DensePose with “In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations”. Both have competitive PCK and MPJPE values and have had the best success with “in-the-wild” images. A consideration is that stage two here is mostly used with the Deepcut stage one infrastructure so adjustments must be made to the DensePose output so it is readily used as an input.
The second stage architecture is as follows:[18]



END-TO-END

LCRNET++ is recommended. It has a very competitive MPJPE, and has tested best for natural images with lots of diversity and occlusion, fitting for the Yoga-82 dataset.
The architecture is as follows:[19]



[18] https://arxiv.org/pdf/1904.03289.pdf
[19] https://arxiv.org/pdf/1803.00455.pdf