Review of estimating 3D human pose and shape

Yan Qianqian
21835037
School of Mathematical Sciences
qqyan_math@163.com

Abstract

This review analysises the CVPR paper [12] about single color image and shape from a single color image. Pavlakos's work proposes an efficient and effective direct prediction method based on Convolutional Networks(ConvNets). Central part to their approach is the corporation of a parametric statistical body shape model(SMPL) within their end-to-end framework. This allows us to get very detailed 3D mesh results, while requiring estimation only of a small number of parameters, making it friendly for direct network prediction. These parameters can be predicted reliably only from 2D keypoints and masks.These are typical outputs of generic 2D human analysis ConvNets, allowing us to relax the massive requirement that images with 3D shape ground truth are available for training. A differentiable renderer is employed to project the 3D mesh to image, which enables further refinement of the network, by optimizing for the consistency of the projection with 2D annotations. The reason why I recommend this paper is that the proposed approach outperforms previous baselines on this task and offers an attractive solution for direct prediction of 3D shape from a single color image. This review introduces some previous works about estimating human pose and details about the paper of Pavlakos.

Keywords: 3D human pose and shape

1. Introduction

The aim of this review is to analysis the work of Pavlakos. This work addresses the problem of estimating the full body 3D human pose and shape from a single color image. And it demonstrate that ConvNets can indeed offer an attractive solution for this problem, by proposing an efcient and effective direct prediction approach, which is competitive and even outperforms iterative optimization methods.

Estimating the full body 3D pose and shape of humans from images has been a challenging goal of computer vision. However, estimating 3D pose and shape from single color images remains the ultimate goal for 3D human analysis. Most approaches rely on iterative optimization, attempting to estimate a full body 3D shape that is consistent with 2D image observations, like silhouettes, edges, shading, or 2D keypoints [13, 4]. In this paper of Sigal [13], they propose a method for automatically recovering a detailed parametric model of non-rigid body shape and pose from monocular imagery. And in this paper of Guan [4], their approach computes shape and pose parameters of a 3D human body model directly from monocular image cues and advances the state of the art in several directions. Despite the signicant runtime required to solve the complicated optimization problem, the common failures because of local minima, and the error-prone reliance on ambiguous 2D cues, optimization-based solutions remain the leading paradigm for this problem. Even the emergence of deep learning has not changed signicantly the landscape. Pavlakos *et al*. [12] proposed an efficient and effective direct prediction method based on ConvNets. And the work incorporation of a parametric statistical body shape model (SMPL [9]) within their end-to-end framework.Skinned Multi-Person Linear model (SMPL) is a skinned vertex based model that accurately represents a wide variety of body shapes in natural human poses. The parameters of the model are learned from data including the rest pose template, blend weights, pose-dependent blend shapes,identity-dependent blend shapes,and a regressor from vertices to joint locations. Unlike previous models, the pose-dependent blend shapes are a linear function of the elements of the pose rotation matrices.

Pavlakos *et al*. [12] propose to employ a differentiable renderer to project the generated 3D mesh back to the 2D image. This enables end-to-end netuning of the network by optimizing for the consistency of the projection with annotated 2D observations, i.e., 2D keypoints and masks. The complete framework offers a modular direct prediction solution to the problem of 3D human pose and shape estimation from a single color image and outperforms previous

approaches on the relevant benchmarks.

2. Related work

2.1. Previous work

It is crucial to get an accurate prediction of the 3D pose of the person to estimate a convincing 3D reconstruction of the human body. Many previous works have limitions. For example, some recent works follow the end-to-end paradigm, using images as input to predict 3D joint locations, regress 3D heatmaps, or classify the image in a particular pose class. However most of the these ConsNets require images with 3D pose ground truth for training, limiting the available training data sources.Other approaches commit to the 2D pose estimates provided by state-of-the-art ConvNets and focus on the 3D pose reconstruction. And Martinez [11] demonstrate state-of-the-art results using a simple multi-layer perceptron which regresses the 3D joint locations from 2D pose input. Pavlakos [12] estimate the whole surface geometry of the human body. To estimate human shape, most methods attempt to estimate the parameters of a statistical body shape like SCAPE [2] or SMPL [9]. The input is usually silhouettes. Knowledge of human shape is useful for biometric applications, however Pavlakos [12] argue that for 3D perception the potential and the challenges are signicantly greater when pose and shape are inferred jointly.

Joint 3D human pose and shape estimation makes the task signicantly harder.This has consistently fostered research in non single image scenarios, for more robust results. Some works propose methods to solve this problem. In the spirit of exploring defferent sensors, von Marcard [10] use a sparse set of IMUs on the subject to recover pose and shape jointly. The most challenging case is that using only a single color image as input, the work of Sigal *et al*. [13] is among the rst to estimate high quality 3D shape estimates, by tting the parametric model SCAPE [2] to ground truth image silhouettes. Bogo [3] use 2D keypoint detections from a 2D pose ConvNet and fit the parametric model SMPL to these 2D locations.Their 3D pose results are very accurate, but shape remains highly underconstrained.

Direct prediction approaches estimate 3D pose and shape in a discriminative way, without explicitly optimizing a specic objective during inference. In contrast, Pavlakos *et al*. [12] demonstrate that only a much smaller set of annotations are critical for the reconstruction, i.e.,2D joints and masks, which can be provided by human annotators and are abundant for in-the-wild images, while they also incorporate everything within a unied end-to-end framework. And they use a diffierentiable way to identify these parameters to avoid half a million of extra learnable weights. They focus their computational and learning effort in the image to 3D shape part of the framework. Their framework can be trained from scratch instead of relying on synthetic image data for pretraining, and they demonstrate state-of-theart results for model-based 3D pose and shape prediction.

2.2. Approach

This part will introduce the important part of Pavlakos's work. Statistical body shape models, like SCAPE [5] or SMPL [25], which provide signicant opportunities for an end-to-end framework. One of the important advantages is their low-dimensional parameter space, which is very suitable for direct network prediction. With this parameter representation, we can keep the output prediction space small, compared to voxelized or point cloud representations. And the low dimensional prediction does not sacrice the quality of the output, since we can still generate high quality 3D meshes from the estimated parameters. Furthermore, from a learning perspective, we bypass the problem of learning the statistics of the human body, and devote the network capacity at the inference of the model parameters from image evidence.

In this work, Pavlakos employ the more recent SMPL model, introduced by Loper*et al*. [9]. There are some essential notation here. SMPL denes a function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi)$, where $\boldsymbol{\beta}$are the shape parameters, $\boldsymbol{\theta}$ the pose parameters and Φ are xed parameters of the model. The direct output of this function is a body mesh $P \in \mathbb{R}^{N\times 3}$ with $N = 6890$ vertices $P_i \in \mathbb{R}^3$ The shape of the model uses a linear combination of a low number of principal body shapes which are learned from a large dataset of body scans. The *shape parameters* $\boldsymbol{\beta}$ are the linear coefcients of these base shapes. The pose of the body is dened through a skeleton rig with 23 joints. The *pose parameters* $\boldsymbol{\theta}$ are expressed in the axis angle representation and dene the relative rotation between parts of the skeleton. In total, 72 parameters dene the pose (3 for each of the 23 joints, plus 3 for the global rotation). Given the rest pose shape retrieved by the shape parameters $\boldsymbol{\beta}$, SMPL denes pose-dependent deformations and uses the pose parameters $\boldsymbol{\theta}$ to produce the nal output mesh. Conveniently, the *body joints* $\boldsymbol{J}$ are a linear combination of a sparse set of mesh vertices, making joints a direct outcome of the estimated body mesh.

Pavlakos leverage all the resources they have available and use their insights for the problem to build an effective framework. Firstly, from ndings of prior work, they identify that 3D pose can be estimated reliably from 2D pose estimates, while the shape can be inferred from silhouette measurements. This observation conveniently decomposes the problem in a) estimation of keypoints and masks from color images and, b) prediction of 3D pose and shape from the 2D evidence. The advantage of this practice is that the framework can be trained without requiring images with 3D shape ground truth.

2.2.1 Keypoints and silhouette prediction

The rst step of their framework focuses on 2D keypoint and silhouette estimation. This part is motivated by the availability of large-scale benchmarks with 2D joints and mask annotations. Considering the volume and the variability of this data, they leverage it to train a ConvNet for 2D pose and silhouette prediction, that is particularly reliable under various imaging conditions and poses. For a more elegant solution, they train a single ConvNet, which they denote as *Human2D*, that generates two outputs, one for keypoints and one for silhouettes.The keypoint output is in the form of heatmaps, where an MSE loss, $\mathcal{L}_{hm}$, between the ground truth and the predicted heatmaps is used for supervision. The silhouette output has two channels (body and background) and is supervised using a pixelwise binary cross entropy loss, $\mathcal{L}_{sil}$. For training, we combine the two losses: $\mathcal{L}_{hg} = \lambda\mathcal{L}_{hm} + \mathcal{L}_{sil}$, where $\lambda = 100$.

2.2.2 3D pose and shape prediction

The second step requires estimation of the full body 3D pose and shape from 2D keypoints and silhouettes. Here, this mapping can also be learned from data while it is possible to get a reliable prediction in a single estimation step. They train two network components: (a) the *PosePrior*, which uses 2D keypoint locations as input together with the condence of the detections (realised by the maximum value of each heatmap) and estimates the pose coefcients $\boldsymbol{\theta}$, and (b) the *ShapePrior*, which uses the silhouette as input and estimates the shape coefcients $\boldsymbol{\beta}$. Regarding the architecture, the *PosePrior* uses two bilinear units, where the input is the 2D keypoint locations and the maximum responses from each heatmap, and the output is the 72 SMPL pose parameters $\boldsymbol{\theta}$. The *ShapePrior* uses a simple architecture with ve 3 3 convolutional layers, each one followed by max-pooling, and an additional bilinear unit at the end with 10 outputs, corresponding to the SMPL shape parameters $\boldsymbol{\beta}$.

Pavlakos's approach entails the generation of the full body mesh at training time, where they optimize explicitly for the predicted surface by applying a 3D per-vertex loss. Since the function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi)$ is differentiable, we can backpropagate through it and handle this mesh generator as a typical layer of our network, without any learnable parameters. Given the predicted mesh vertices $\hat{P}_i$ and the corresponding groundturth vertices P_i, we can supervise the network with a *3D per-vertex loss*:

$$\mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \|\hat{P}_i - P_i\|_2^2, \tag{1}$$

which considers all the vertices equally and has better correlation with the 3D per-vertex error which is usually employed for evaluation. Alternatively, if the focus is mainly on 3D pose, we can also supervise the network considering only the M relevant 3D joints J_i, which are trivially exposed by the model as a sparse linear combination of the mesh vertices. In this case, denoting with $\hat{J}_i$ the estimated joints, the corresponding loss can be expressed as:

$$\mathcal{L}_{\mathcal{J}} = \sum_{i=1}^{M} \|\hat{J}_i - J_i\|_2^2. \tag{2}$$

Empirically, we found that the best training strategy is to initially get a reasonable initialization for the network parameters using an $\mathcal{L}_2$ parameter loss, and then activate also the vertex loss $\mathcal{M}$ (or the joints loss $\mathcal{L}_{\mathcal{J}}$ if the focus is on pose only), to train a better model.

To close the loop, their complete approach includes an additional step that projects the 3D mesh to the image and examines consistency with 2D annotations. More specically, for their implementation, they employ an approximately differentiable renderer, OpenDR, which projects the mesh and the 3D joints to the image space, and enables backpropagation. The projection operation gives rise to: (a) the silhouette $\Pi(\hat{\boldsymbol{P}}) = \hat{S}$, which is represented as a 64 64 binary image, and (b) the projected 2D joints $\Pi(\hat{\boldsymbol{J}}) = \hat{\boldsymbol{W}} \in \mathbb{R}^{M\times 2}$ In this case, the supervision comes from the comparison of these projections with the annotated silhouettes S, and the 2D keypoints $\boldsymbol{W}$, using $\mathcal{L}_2$ losses:

$$\mathcal{L}_{\Pi} = \mu \sum_{i}^{M} \|\hat{W}_i - \hat{W}_2\|_2^2 + \|\hat{S} - S\|_2^2, \tag{3}$$

where $\mu = 10$. The goal of this type of supervision is twofold: (a) it can be employed for end-to-end renement of the network, using only images with 2D keypoints and/or masks for training, and (b) it can be useful to mildly adapt a generic pose or shape prior to a new setting (e.g., new dataset), where only 2D annotations are available.

2.2.3 Empirical evaluation

This part we simplely analysis the results which Pavlakos evaluate with the proposed approach. First, they present the benchmarks that they employed for quantitative and qualitative evaluation. They employed two recent benchmarks that provide color images with 3D body shape ground truth, the UP-3D dataset [7] and the SURREAL [14] dataset. Additionally, they used the Human3.6M [5] dataset for further evaluation of the 3D pose accuracy.

UP-3D:It is a recent dataset that collects color images from 2D human pose benchmarks, like LSP [6] and MPII [1] and uses an extended version of SMPLify [3] to provide 3D human shape candidates. The candidates were evaluated by human annotators to select only the images with good 3D shape ts. It comprises 8515 images, where 7818 are used for training and 1389 for testing. They compare with two

state-of-the-art direct prediction approaches by Lassner and Tan.Their approach outperforms the other two baselines by signicant margins.

SURREAL: It is a recent dataset which provides synthetic image examples with 3D shape ground truth. The dataset draws poses from MoCap [5] and body shapes from body scans to generate valid SMPL instances for each image. The synthetic images are not very realistic, but the accurate ground truth, makes it a useful benchmark for evaluation. They compare with two state-of-the-art approaches, one based on iterative optimization, SMPLify, and one based on direct prediction. Their approach outperforms the other two baselines. For this dataset we observed that because of the challenging color images (low illumination, out-of-context backgrounds, etc), the 2D detections where more noisy than usual, providing some hard failures for the iterative optimization approach. In contrast, their approach was more resistant to these noisy cases recovering a coherent 3D shape in most cases.

Human3.6M: It is a large-scale indoor dataset that contains multiple subjects performing typical actions like Eating and Walking. For Human3.6M they evaluate only the estimated 3D pose, since there is no body shape ground truth available. their network is the same as before (Priors trained on CMU), although, they use the 3D joints error for supervision (equation 2), since the focus is on pose. Similarly to the other approaches they compare with, we do not use any data from this dataset for training. Their approach again outperforms the other baselines.

2.2.4 Running time

Pavlakos's approach requires a single forward pass from the ConvNet to estimate the full body 3D human pose and shape. This translates to only 50ms on a Titan X GPU. In comparison, SMPLify report roughly 1 minute for the optimization, while the publicly available (unoptimized) code runs on 3 minutes per image on average. When the number of landmarks increases to 91, their direct prediction approach more than three orders of magnitude faster than the state-of-theart iterative optimization approaches.

3. Summary

The reason why I choose this paper is that it presents a viable ConvNetbased approach to predict 3D human pose and shape from a single color image. The most important part of their solution was the incorporation of a body shape model, SMPL, in the end-to-end framework. Through the work of Pavlakos we enabled: a) prediction of the parameters from 2D keypoints and silhouettes, b) generation of the full body 3D mesh at training time using supervision for the surface with a per-vertex loss, and c) integration of a differentiable renderer for further end-to-end renement using 2D annotations. Their approach achieved state-of-the-art results on relevant benchmarks, outperforming previous direct prediction and optimizationbased solutions for 3D pose and shape prediction. Finally, considering the efciency of their approach, they demonstrated its potential to accelerate and improve typical iterative optimization pipelines. At the same time, we can see others also do similar work. Mengyuan [8] propose a novel method to recognize human action as the evolution of pose estimation maps. Instead of relying on the inaccurate human poses estimated from videos, they observe that pose estimation maps, the byproduct of pose estimation, preserve richer cues of human body to benet action recognition. More and more accurate methods will be proposed to estimate human pose.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SPACE:shape completion and animation of people. *ACM Transactions on Graphics(TOG)*, 24:408–416, 2005.

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL:Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016.

[4] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.

[5] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Unite the people: Closing the loop between 3D and 2D human representations. *PAMI*, 36(7):1325–1339, 2014.

[6] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[7] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *CVPR*, 2017.

[8] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, 2018.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL:A skinned multi-person linear model. *ACM Transactions on Graphics(TOG)*, 34(6):248, 2015.

[10] T. Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3D human pose estimation from sparse imus. In *Eurographics*, 2017.

[11] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[12] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.

[13] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *NIPS*, 2008.

[14] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.