Internship Report: 3D Human Reconstruction Methods using a Latent Representation of Human Form

Joel Huang

Singapore University of Technology and Design
NCS Product R&D, Singapore

September 9, 2018

Abstract

Summarize the latest work on recovering 3D spatial information about humans from a single photograph or video frame.

1 Introduction

Recovering spatial information about humans is a core problem in computer vision. Once we teach machines to see us as we do ourselves, we unlock a huge realm of possibilities, a future of which post-screen interfaces, vision-based analytics, and hyper-realistic simulation are a part of. Traditionally, much of the work in this field has focused on the problem of 2D **pose estimation**, defined as the localization of human joints. While this abstraction of human form is intuitive and has recently produced stunning results [1], it fails to retain information about the human form in three-dimensional space.

There has been a series of works focused on recovering the human form from photographs using a latent representation of the human body. Parametric deformable surface models such as the Skinned Multi-Person Linear (SMPL) model [2], a learned 3D model of human shape from thousands of body scans of different people, emulate reality where changes in human pose are coupled with body surface deformations. These models have been used in human **correspondence estimation**, in which a mapping between 2D pixels and the 3D model's surface is learned [3, 4].

It is possible that the human visual cortex also works with latent, learned representations of objects and forms that weight the estimations and decisions we make in our perception of form, depth and space. The authors of SMPLify [5] work along those lines, attempting the fitting of the 3D SMPL model to raw 2D keypoint detections. To do this, the parameters of the model which capture pose and shape must be optimized such that their combination results in the least error when reprojected to the plane of the original image. They introduce constraints in their iterative optimization approach, by favouring possible poses, penalizing impossible poses, and penalizing information from occluded body parts. Although iterative optimization has the least error among current methods, fitting for a single image is on the order of tens of seconds.

Once again, deep learning approaches have emerged to tackle the problem of efficiently estimating these parameters. Kanazawa *et al.* [6] propose an end-to-end encoder-regressor-discriminator framework which directly converts images into pose and shape estimates, while other approaches further decompose the problem into several smaller tasks, enabling training with samples that are easily obtained, such as 2D joints or shape segments. Omran *et al.* [7] decompose the problem this way, generating body part segmentation maps and using them to directly train a convolutional neural network, which then predicts the pose and shape parameters. Pavlakos *et al.* [8] separate the problem into modular tasks, using a stacked hourglass network to produce joint locations and silhouette masks, which are then processed separately by two networks to obtain the pose and shape parameters respectively.

2 Modelling Human Form

2.1 Principal Components

When trying to capture spatial information, we often fall back to a three dimensional worldview, force-fitting a Cartesian parametrization for ease of understanding: we describe objects as taller, wider, and closer than one another; objects rotate about a defined global axis; and

scale according to three familiar principal components, height, width and depth. In fact, it is sufficient to fully describe the spatial form of all objects, and is the most efficient way to describe regular shapes like cuboids and planes. But it does not mean anything useful when we describe human form using a set of three-dimensional vectors. We want to describe human form based on its principal components; features along the lines of stature, body mass, bone structure, mass distribution, fatness, skinniness, and others we subconsciously observe but do not have the words to describe.

2.2 Model

The Skinned Multi-Person Linear Model (SMPL) is an example of describing the human form in Euclidian space, but not a Cartesian representation. The model is parameterized by 72 pose and 10 shape variables.

Pose is modelled as the set of axis-angle representations of 23 pre-defined body parts in the skeletal rig, plus the root orientation. $\vec{\theta}$, the pose parameters, are defined as:

$$\vec{\theta} = [\vec{\omega}_0^T, \ldots, \vec{\omega}_k^T]^T \quad (1)$$

where $\vec{\omega}_k \in \mathbb{R}^3$ is the rotation of each part k in Euler angles. The total size of the pose vector is $|\vec{\theta}| = (23+1) \times 3 = 72$. By modelling pose as joint rotations, the model can easily be used in existing rigging and rendering software.

Shape is modelled as weighted changes in orthonormal principal components, which define the shape space for the particular model. The principal components, which have been learnt from registered training meshes using Principal Component Analysis (PCA), are reminiscent of concepts of human form, like stature, or fatness. Each shape coefficient, β, weights a particular element in the shape displacement matrix, allowing us to modify, for example, the stature of a model simply by changing the appropriate β in $\vec{\beta}$:

$$\vec{\beta} = [\beta_0, \ldots, \beta_{|\vec{\beta}|}]^T \quad (2)$$

3 Reconstruction Pipeline

3.1 Design

Modularity might be a guiding principle to approach the design of a real-time reconstruction pipeline, as it allows for modules to be improved separately, while allowing the use of the intermediate outputs to augment or process other information. For 3D human reconstruction in particular, the intermediate state is often some set of 2D joint confidence maps, body part segmentation masks, or silhouette masks.

In an end-to-end learning task where the sole objective is to learn the function mapping from an image to a 3D representation, the network might not be able to realise the significance of this set of outputs. Alldieck *et al.* [9] make impressive use of 2D silhouette segmentations, associating every point on the silhouette edge with a point on the SMPL model, then computing the inverse pose of this outline, finally obtaining a visual hull of the input human in T-pose, which massively simplifies the shape fitting of a T-posed SMPL model to this hull.

More obvious modular pipelines have been used with success, with division of labour among networks along the pipeline, each tackling a specific subproblem. In their recent paper [8], Pavlakos *et al.* decompose the 3D estimation problem into two steps, namely: (i) using a stacked hourglass network, return a simultaneous estimation of joint locations and a silhouette mask, and (ii), feeding this into two additional independent networks, estimate pose θ and shape β.

However, one of the downsides to modular design is the compounding of error. For example, if a pipeline consists of a 2D joint detector and a 3D regressor module, a possible failure in joint detection can completely change the behaviour of the 3D regressor module. For this reason, it might be advisable, though counterintuitive, to further separate the problem into simpler tasks where we can have more control over the stages of processing.

3.2 2D Feature Extraction

From a monocular input image, classical low-level image processing objectives can be met with an off-the- shelf ImageNet-trained model. They are typically included at the start of 2D pose estimation networks [1,7,10] to provide feature extraction capabilities. Fast, direct bounding box predictors like YOLOv3 [11] can be used to crop regions of interest at tens of frames per second or more.

The 2D module might output features like hard joint keypoints and segmentation masks, softer or more probabilistic predictions in heatmaps, or totally different representations of pose, like Part Affinity Fields [1].

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[2] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, Oct. 2015.

[3] R. A. Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "Densereg: Fully convolutional dense shape regression in-the-wild," *arXiv:1612.01202*, 2016.

[4] R. A. Guler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," *arXiv*, 2018.

[5] 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 *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer International Publishing, Oct. 2016.

[6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[7] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model-based human pose and shape estimation," (Verona, Italy), 2018.

[8] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.