LCR-SMPL: Toward Real-time Human Detection and 3D Reconstruction from a Single RGB Image

Elena Peña-Tapia *
Keio University
Cross-Compass Ltd.

Ryo Hachiuma †
Keio University

Antoine Pasquali ‡
Cross-Compass Ltd.

Hideo Saito§
Keio University

Abstract

This paper presents a novel method for simultaneous human detection and 3D shape reconstruction from a single RGB image. It offers a low-cost alternative to existing motion capture solutions, allowing to reconstruct realistic human 3D shapes and poses by leveraging the speed of an object-detection based architecture and the extended applicability of a parametric human mesh model. Evaluation results using a synthetic dataset show that our approach is on-par with conventional 3D reconstruction methods in terms of accuracy, and outperforms them in terms of inference speed, particularly in the case of multi-person images.

Index Terms: Artificial Intelligence—Computer vision—Computer vision problems—Shape inference; Artificial Intelligence—Computer vision—Computer vision problems—Reconstruction

1 Introduction

Accurate human 3D representation is a requirement for a growing number of applications, ranging from video-games to simulations for human-robot collaboration. Most state-of-the-art deep-learning based methods [5] focus on retrieving 3D joints, combining convolutional neural networks with 2D joint lifting. LCR-Net [5] repurposes Region Proposal Networks to provide end-to-end human detection and 3D skeleton prediction on multi-person images. However, 3D joints do not provide a detailed description of the morphology of the body, and fall out short for applications that require shape information such as integration in virtual reality (VR) simulated environments or augmented reality (AR) applications.

A wide range of models can be used to capture human volumetric information, from geometric primitives to more natural-looking parametric meshes. The recently proposed skinned multi-person linear model (SMPL) [3] allows to work with realistic mesh models parametrized using joint rotations and statistical shape coefficients.

Most existing SMPL-based 3D reconstruction methods are not standalone, they either perform 2D/3D skeleton fitting or depend on a tightly cropped bounding box, which require the use of an external detector. Nor are they multi-person, except for very recent contributions such as [1]. Finally, they depend on recurrent or iterative procedures that hinder their run-time performance.

2 Method

This paper proposes a novel approach for human 3D reconstruction from a single RGB image that is fast, accurate, multi-person and standalone. Inspired by real-time 3D joint prediction methods [5], our network generates SMPL parameter proposals that are simultaneously classified and regressed to obtain parameter variations.

*e-mail: ept@keio.jp
†e-mail: ryo-hachiuma@keio.jp
‡e-mail: antoine@cross-compass.com
§e-mail: hs@keio.jp

The LCR-SMPL method leverages the advantages of its two main components: the SMPL model formulation and the LCR pipeline.

2.1 SMPL Model

SMPL [3] is a realistic skinned vertex-based model represented by a statistical parametric function $M(\beta,\theta;\Phi)$ that maps 10 shape parameters β and 72 axis-angle pose parameters θ into $6,890$ vertices V, given a set of learned model parameters Φ. The mapping function represents a series of transformations performed to a mesh template based on linear blend skinning. The pose parameters represent 24 joint rotations following a kinematic tree in axis-angle representation, while the shape parameters are the 10 first element of the principal component analysis (PCA) of a diverse set of human scans. The model also includes a joint regressor $\mathcal{J}_R$, that outputs the 24 joint 3D coordinates: $J_{3D} = \mathcal{J}_R(\beta,\theta;\Phi)$.

2.2 Proposed LCR-SMPL Architecture

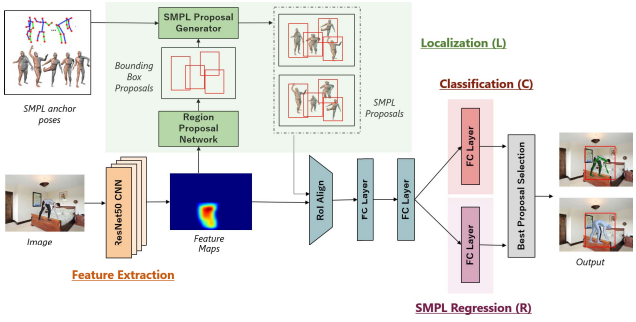

Figure 1: LCR-SMPL architecture.

Fig. 1 shows the four main components of the proposed network architecture: the convolutional feature extractor, the localization module (L), and two simultaneous classification (C) and SMPL regression (R) modules. A final processing step helps select the best model proposals to output a bounding box and a set of SMPL parameters associated to a 3D skeleton and 3D mesh. The final loss of the network is computed as a sum of the three losses: $\mathcal{L} = \mathcal{L}_{Loc} + \mathcal{L}_{Class} + \mathcal{L}_{SMPLreg}$. $\mathcal{L}_{Loc}$ $\mathcal{L}_{Class}$ are the same as in LCR-Net [5], while $\mathcal{L}_{SMPLreg}$ is the smooth L1 joint loss between the ground truth and network prediction. Shared convolutional features allow to train the network in an end-to-end manner.

The feature extractor uses a ResNet50 backbone to compute a feature map from the input RGB image. The localization module contains a Region Proposal Network (RPN) that hypothesizes candidate bounding boxes, and each of them is assigned a set of pre-computed candidate poses (SMPL anchors), forming *SMPL Proposals*. Differently to LCR-Net, each SMPL proposal represents the combination of a person's potential location, *shape* and pose in the image. The loss of this component is the loss of the RPN, which comprises bounding box classification and regression losses.

SMPL Proposals are later aggregated with image features using a RoI Align layer, after which two shared fully connected layers

978-1-7281-7675-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR-Adjunct51615.2020.00062

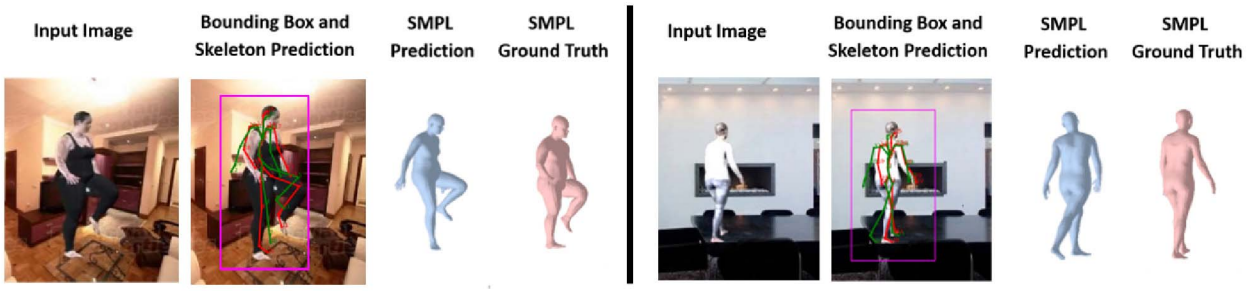

Figure 2: LCR-SMPL Qualitative results. Bounding box shown in magenta, predicted skeleton in green, ground truth skeleton in red.

then lead to a separation of the network into two parallel single-layer modules. In the classification branch, each anchor location receives a score according to how close anchor joints are to the ground truth pose, performing a coarse approximation to the desired reconstruction. The classification loss is a logistic regression loss over the probability distribution of the classes.

The SMPL regression module predicts the necessary variation of shape ($\Delta\beta$) and pose ($\Delta\theta$), with respect to the SMPL proposal parameters. To train this module, the distances between ground-truth and predicted joint positions in both 3D and 2D space are calculated. 3D joints can be obtained after applying the SMPL joint regressor $\mathcal{J}_R$, while 2D joints are computed by projecting 3D joints into the image plane. The selected distance metric is smooth L1 loss.

3 Experiments and Results

We evaluated our network in terms of 3D reconstruction accuracy and computational time using the SURREAL synthetic human dataset [6], which contains six million frames with annotated 2D joint, 3D joint and corresponding SMPL parameters. A subset of 5k images was used for training, and 1k images for test. 100 SMPL anchors were selected by performing k-means clustering on the shape and pose space of a randomized data subset of 200k SURREAL frames. A study of the trade-off accuracy/training-time according to the number of anchors suggested an optimal number between 80 and 200.

Table 1: Accuracy Results. $_{\text{Abs}}$ refers to the mesh before rigid alignment, while $_{\text{Align}}$ corresponds to the aligned mesh.

Method	MPJE (mm)	MVE (mm)
SHN [4]	40.8	−−
LCR-Net [5]	52.0	−−
Tung et al. [4]	64.4	74.5
SMPLR$_{\text{Abs}}$ [4]	50.6	66.0
SMPLR$_{\text{Align}}$ [4]	48.2	62.3
LCR-SMPL$_{\text{Abs}}$	**66.1**	**102.6**
LCR-SMPL$_{\text{Align}}$	**36.2**	**59.5**

Table 1 summarizes the results of measuring the Mean Per Joint Error (MPJE) and Mean Vertex Error (MVE), of both the resulting mesh after inference and the mesh after rigid alignment with the ground truth. Compared to results from recent literature, our approach is comparable to both 3D pose (top group) and SMPL methods (middle group) in terms of MPJE and MVE, and outperforms the state-of-the-art after rigid alignment. This observation is coherent with the fact that previous SMPL methods do not perform person localization, thus minimizing their absolute error. Fig. 2 shows a sample of network outputs, demonstrating its ability to capture shape and pose variations.

Table 2: Speed Results

Method	3D shape	FPS$_{\text{single}}$	FPS$_{\text{multi}}$
LCR-Net [5]	No	3.87	3.27
HMMR [2]	Yes	0.79	0.28
LCR-SMPL	**Yes**	**2.14**	**1.94**

As for time complexity, Table 2 shows a comparison of inference speeds with LCR-Net (3D pose) and HMMR (3D shape and pose) on sets of single and multi-person images. Our network outperforms HMMR in terms of achievable frames per second (FPS), showing inference times closer to those of 3D-pose only methods. This is particularly relevant in the case of multi-person images, where our network regresses all reconstructions once, while previous single-person 3D shape reconstruction methods must be run for each individual subject, reducing the total FPS count.

4 Conclusion

To sum up, we have introduced the novel LCR-SMPL approach for standalone human detection and 3D shape and pose reconstruction out of RGB images. Results on synthetic humans demonstrate an on-par performance with previous SMPL-based methods in terms of accuracy, and an increased performance in terms of inference speed, particularly in the case of multi-person images, showing potential for applications where fast inference is required, such as Mixed and Augmented Reality. Future works could make this advantage more palpable by switching to single-stage object-detection architectures such as YOLO or SSD, and focus on tackling failure case scenarios such as those involving inter-person occlusion or ambiguous poses.

References

[1] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of CVPR*, pp. 5579–5588, 2020.

[2] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of CVPR*, pp. 7122–7131, 2018.

[3] 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):1–16, 2015.

[4] M. Madadi, H. Bertiche, and S. Escalera. SMPLR: Deep SMPL reverse for 3d human pose and shape recovery. *arXiv preprint. doi:1812.10766*, 2018.

[5] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-classification-regression for human pose. In *Proceedings of CVPR*, pp. 3433–3441, 2017.

[6] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proceedings of CVPR*, pp. 109–117, 2017.