Project Proposal

Proposed topic

Reconstructing 3D Human Mesh Model from a Single Wild Image

Student Name: Ting Liu (UIN:730008149)

Problem description

All along, fast 3D human modeling has obtained great interest and been widely researched. In addition to 3D software modeling and scanning equipment, human body detection and part segmentation based on neural networks provides a new idea for reconstruction of 3D human body from 2D images. This project intends to realize the projection of detected 2D human body keypoints to 3D space, realizing the parametric reconstruction of 3D human mesh model from a single wild image.

Literature review

Traditional 3D human body modeling methods are mainly based on the manual operation of modeling software. Although some current modeling software allow designers to do parametric human modeling, which alleviates the heavy workload. In order to get a very refined human model, it is still quite time-consuming and labor-intensive, and also depends on the professional modeling knowledge of the designer to a great extent. Therefore, fast 3D human modeling methods, especially automatic 3D human reconstruction based on 2D images, have attracted substantial research interest.

Recent 3D human reconstruction approaches can be divided into 2 main categories, scanning based approaches and generation by utilizing neural networks. In terms of the representation for modeling 3D objects, the output models are mainly divided into point cloud data and volumetric models such as voxels and meshes.

1. 3D point cloud reconstruction

Early research focused on methods of fusing views captured by multiple cameras or sensors to generate 3D point cloud [1, 2]. However, scanning-based 3D point cloud generation requires demanding devices. Depth distortion and registration errors in the reconstructing process are difficult to deal with. But still there are 2 representative studies which addressed partial problems.

KinectFusion [3] provides a frame-to-model dense surfaces' generation system of indoor scenes including human bodies. It uses a hand-held Kinect sensor to track depth images in real time, obtaining point clouds by simultaneous localization and mapping, and then adopts TSDF model to fuse the captured depth image, integrating the point cloud of the current frame into a 3D volumetric representation, thereby produces denoised, detailed, and complete fused 3D models. Similar work has been done with DynamicFusion [4]. But it no longer requires prior scene model and achieves the reconstruction of moving objects.

It is worth mentioning that, instead of transfer 3D point clouds to other forms of data, Qi et al. [5] designed the PointNet which directly consumes point clouds for object classification and part segmentation applied to human bodies. This network well respects the permutation invariance of points in the input, which can be a potential research direction in human reconstruction.

2. Mesh model reconstruction

In image-based reconstruction approaches, a 3D mesh model is easier to manipulate and represent the complex geometry of 3D objects. In that case, Wang et al. [6] built the Pixel2Mesh, which is an end-to-end convolutional neural network using the perceptual features extracted from a single RGB image to produce correct geometry by progressively deforming an ellipsoid. However, Pixel2Mesh may not be applicable for the reconstruction of a human body since it has a high degree of asymmetry resulted from drastic poses.

2.1 SMPL model

A 3D mesh model for a human body is more complicated because of the body structure and its irregular surface. At that point, Loper et al. [7] created a skinned multi-person linear model (SMPL), which is a skinned vertex-based model that accurately represents various human body shapes in natural human poses. SMPL provides a richer and more useful mesh representation with shape and 3D joint angles. Since it was proposed in 2015, most human modeling-related researches [8, 10-15] have applied it. Almost all following studies involved in this review are based on this parametric statistical model. For fitting SMPL models, human shape and pose estimation in 2D images is the first task in human reconstruction.

2.2 Parametric fitting

SMPLify [8] automatically estimates the 3D pose of the human body as well as its 3D shape from a single unconstrained image, to produce a parameterized human SMPL model. For 3D pose estimation, it firstly uses DeepCut [9] to predict 2D human body joints and then project them to 3D model joints. They created 20 capsules to simulate the shapes of different human body parts and trained a regressor from model shape parameters to capsule parameters. However, the shape information are limited and are robustly fitted into SMPL models.

Pavlakos et al. [10] introduced another ConvNets framework named Human2D to predict 2D heatmaps and silhouette masks in images to generate both pose and shape parameter for SMPL models. It improved SMPLify fits by optimizing the surface using a 3D per-vertex loss and adding a differentialble renderer for the projection of 3D mesh to 2D keypoints and masks in images.

Instead of relying on 2D keypoints detection, Kanazawa et al. [11] described a Human

Mesh Recovery (HMR) framework which directly infers 3D pose and shape parameters from a single RGB image's pixels. HMR introduces an adversary learning module to discriminate whether the shape and pose parameters are ground truth data or not and out-performs previous optimization-based methods. The overview of the HMR framework is illustrated in Figure 1.

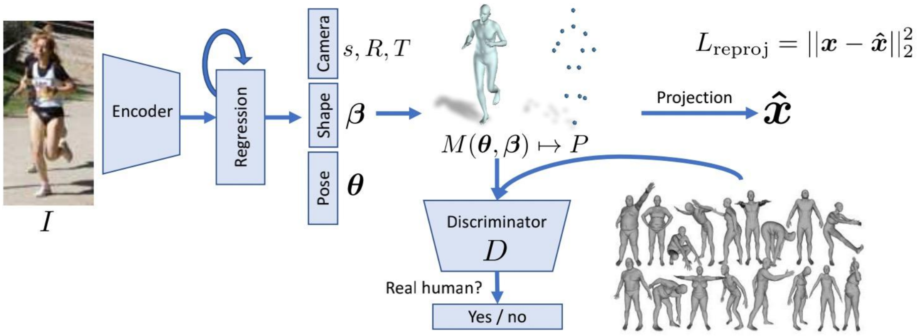

Figure 1: Overview of the HMR framework.

Guler et al. [12] introduced DensePose, a method for realizing the mapping from each human body pixel in 2D wild images to a 3D point on a SMPL model. As an extension, they proposed HoloPose [13] for holistic human body real-time reconstruction from a more than 10fps in-the-wild video. HoloPose formulate and iteratively optimize a misalignment loss comprised of 2D, 3D keypoints and Dense Pose estimation between top-down and bottom-up 3D model predictions, thereby largely improving the model's alignment with the input frame.

In order to represent more plausible geometry details on a reconstructed human mesh model, DeepHuman [14] exploits SMPLify [8] and HMR [10] and present a volume-to-volume approach. It first implements parametric body estimation, then performs full-body surface reconstruction and finally, refines the details on the visible areas of the surface utilizing volumetric feature transformation.

2.3 Latest research

Recently, Lin et al. [16] introduced a new MEsh TRansfOrmer (METRO) method to realize human pose and mesh reconstructions. It uses a transformer encoder to jointly model vertex-vertex and vertex-joint interactions, and outputs 3D joint coordinates and mesh vertices simultaneously. With learning non-local relationships among mesh vertices and joints, which can be shown in Figure 2, given an input image where the human body is partial occluded, the mesh model can still be effectively generated with less abnormal distortions. In addition, METRO does not rely on any parametric mesh models like SMPL, thus it can be easily extended to other objects.

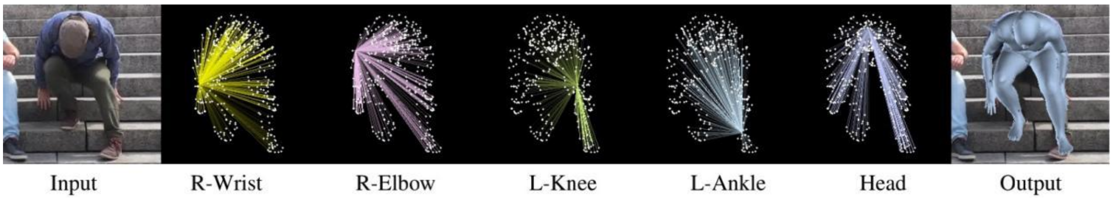

Figure 2: A METRO reconstruction result.

3. Discussion

Apart from software modeling and 3D scanning, most of existing human reconstruction methods are implemented based on 3D pose estimation, 2D silhouette and keypoints detection, combined with constraints on both local and non-local features between joints and meshes. However, it's more like a pose transfer by modifying a standard human model. The generated mesh model lack enough refined surface details of the human body to achieve automatic personalized modeling. For example, for obese people, the generated model has relatively low similarity to the real human body.

In the meanwhile, existing reconstructing approaches for human bodies aim merely at the recovery of the human body pose and shape, and do not involve the reconstruction of the human body skin from an image. A complete reconstruction of a 3D skinned human mesh model from a single color image can also of great research value.

Importance and Value

Compared with 2D human silhouettes and both 2D and 3D human skeletons, 3D human body mesh models provide more refined and personalized characteristics including various body figures. Reconstructing a 3D human body mesh model functions as a platform for the development of subsequent virtual and augmented reality applications, which is of great significance. In the meanwhile, the characteristics of human topological structure are rather complicated in both 2D image and 3D mesh model. Convolutional Neural Networks can quickly extract and learn features from a massive amount of data, which is an outstanding solution to this kind of problem. However, limited by network structure, hardware’s memory and computing capabilities, many existing reconstructing researches on human bodies based on neural network are still inadequate in end-to-end and real-time property.

It is of great research value in the fields of computer vision and virtual reality to reconstruct the 3D human body mesh from high-quality images in real time. First of all, 3D human body mesh model can be quickly generated and arbitrarily viewed from different angles. Secondly, the reconstruction based on mesh models goes beyond skeletons, and represents various poses and figures of human bodies in 2D images. Apart from having significant theoretical value, such fast human modeling approach is applicable to many practical applications like virtual fitting, movie special effects and so on. For instance, reconstructed 3D human body models can be used as a realistic substitute for real persons in certain special shots in the movie, avoid the laborious process of repeated modeling. Besides, it can be applied to public security monitoring systems, such as to help detect and analyze target people.

Proposed Plan

This project mainly exploits Res2Net [17] and Human Mesh Recovery [5]. Instead of ResNet [18], Res2Net makes full use of multi-scale features of images and provide contextual information of target objects by utilizing the boundary region of perceived objects, thereby achieving more accurate and stable detection of 2D human body keypoints. Based on that, 3D regression and adversarial learning are utilized to generate and optimize 3D human body parameters and produce a human SMPL mesh model.

1. Data preparation

Currently, there are some datasets collected for human body images and shape models. MS COCO [19] collects numerous wild images annotated with 2D keypoints. UP-3D [20] is an outdoor-image dataset whose annotations are created by model fitting. In addition, SURREAL [21] and Human 3.6M [22] provide large amounts of human SMPL mesh models, as well as realistic human body skin textures.

Input: Images with 2D ground-truth annotations, SMPL mesh models with 3D ground-truth annotations.

2. Implementation

According to HMR, 72 pose (involving 3 axis-angle rotation for each of 23 joints and a global rotation) and 10 shape parameters (first 10 coefficients computed by running PCA on shape registrations from multi-shape database) can be generated from an input image, which can be used thereby to reconstruct a parametric human SMPL mesh model. This project will substitute the encoder in HMR with Re2Net to test if better results can be attained.

The SMPL parameters can be defined as: $\Theta=\{\theta, \beta, R, t, s\}$, $\theta \in R^{3K}$, $\beta \in R^{10}$,$R\in R^3$,$t\in R^2$,$s\in R$. K=23, where θ is the axis-angle of the human body joints, β controls the human body shape, R is the global axis-angle, T is the translation of the camera on x-y plane, s is the zooming factor of the camera.

Goals

1. Environment set-up (Anaconda, PyTorch, Cuda, Python, OpenCV).
2. Encode images and extract features using ResNet-50.
3. Generate SMPL parameters using regression based on encoded images.
4. By 10/28, first update report.
5. Optimize shape and pose parameters based on adversarial learning.
6. Add Res2Net module.
7. By 11/18, second update report.
8. Training and testing on different datasets, parameter adjustment.
9. By 12/7, final report and project presentation.

REFERENCES

[1] Khoshelham, K., 2011, August. Accuracy analysis of kinect depth data. In ISPRS workshop laser scanning (Vol. 38, No. 5, p. W12).

[2] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud based multiview stereo algorithm for free-viewpoint video. IEEE Transactions on Visualization and Computer Graphics, 16(3):407–418, 2010

[3] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S. and Fitzgibbon, A., 2011, October. KinectFusion: Real-time dense surface mapping and tracking. In Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on (pp. 127-136). IEEE..

[4] Newcombe, Richard A., Dieter Fox, and Steven M. Seitz. "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[5] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3D classification and segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W. and Jiang, Y.G., 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. arXiv preprint arXiv:1804.01654.

[7] Loper, Matthew, et al. "SMPL: A skinned multi-person linear model." ACM transactions on graphics (TOG) 34.6 (2015): 1-16.

[8] Bogo, Federica, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image." In European conference on computer vision, pp. 561-578. Springer, Cham, 2016.

[9] Pishchulin, Leonid, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. "Deepcut: Joint subset partition and labeling for multi person pose estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4929-4937. 2016.

[10] Pavlakos, Georgios, Luyang Zhu, Xiaowei Zhou, and Kostas 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, pp. 459-468. 2018.

[11] Kanazawa, Angjoo, Michael J. Black, David W. Jacobs, and Jitendra Malik. "End-to-end recovery of human shape and pose." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7122-7131. 2018.

[12] Guler, Riza Alp, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7297-7306. 2018.

[13] Guler, Riza Alp, and Iasonas Kokkinos. "Holopose: Holistic 3d human reconstruction in-the-wild." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10884-10894. 2019.

[14] Zheng, Zerong, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. "Deephuman: 3d human reconstruction from a single image." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7739-7749. 2019.

[15] Varol, Gul, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. "Bodynet: Volumetric inference of 3d human body shapes." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 20-36. 2018.

[16] Lin, Kevin, Lijuan Wang, and Zicheng Liu. "End-to-end human pose and mesh reconstruction with transformers." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1954-1963. 2021.

[17] Gao, Shanghua, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr.

"Res2net: A new multi-scale backbone architecture." IEEE transactions on pattern analysis and machine intelligence (2019).

[18] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[19] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

[20] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In CVPR, 2017.

[21] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I. and Schmid, C., 2017, July. Learning from synthetic humans. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017) (pp. 4627-4635). IEEE.

[22] Ionescu, C., Papava, D., Olaru, V. and 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(7), pp.1325-1339.