Prior-less 3D Human Shape Reconstruction with an Earth Mover's Distance Informed CNN

Jingtian Zhang
Northumbria University
Newcastle upon Tyne, UK
jingtian.zhang@northumbria.ac.uk

Hubert P. H. Shum*
Northumbria University
Newcastle upon Tyne, UK
hubert.shum@northumbria.ac.uk

Kevin McCay
Northumbria University
Newcastle upon Tyne, UK
kevin.d.mccay@northumbria.ac.uk

Edmond S. L. Ho
Northumbria University
Newcastle upon Tyne, UK
e.ho@northumbria.ac.uk

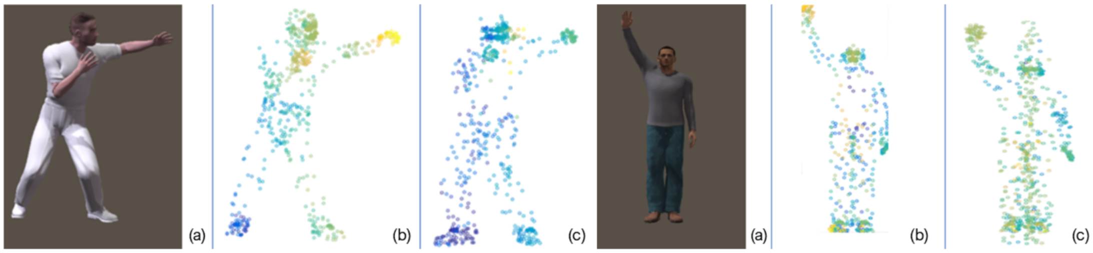

Figure 1: Two examples of reconstruction: (a) 2D images, (b) ground truth 3D point cloud, and (c) reconstructed 3D point cloud.

ABSTRACT

We propose a novel end-to-end deep learning framework, capable of 3D human shape reconstruction from a 2D image without the need of a 3D prior parametric model. We employ a "prior-less" representation of the human shape using unordered point clouds. Due to the lack of prior information, comparing the generated and ground truth point clouds to evaluate the reconstruction error is challenging. We solve this problem by proposing an Earth Mover's Distance (EMD) function to find the optimal mapping between point clouds. Our experimental results show that we are able to obtain a visually accurate estimation of the 3D human shape from a single 2D image, with some inaccuracy for heavily occluded parts.

CCS CONCEPTS

• **Computing methodologies**; • → Shape modeling; Reconstruction;

KEYWORDS

Human Surface Reconstruction, Deep Learning, CNN, Earth Mover's Distance

ACM Reference Format:
Jingtian Zhang, Hubert P. H. Shum, Kevin McCay, and Edmond S. L. Ho. 2018. Prior-less 3D Human Shape Reconstruction with an Earth Mover's Distance Informed CNN. In *MIG '19: ACM SIGGRAPH Conference on Motion, Interaction and Games, Oct 28–30, 2019, Newcastle upon Tyne, UK*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In this paper, we tackle the problem of 2D to 3D reconstruction. Previous work in this field typically makes use of strong prior knowledge of plausible 3D human shapes [1, 3, 7]. However, due to the complex geometry of the human body, and the large variety in human body size, using only a parametric model cannot precisely recover all of the details relating to human body shape.

We propose that the shape can be represented by directly generating a "prior-less" unordered point cloud. In order to facilitate the use of an unordered point cloud, we need to be able to measure the distance between the reconstructed points and the ground truths. Motivated by [2, 5], we propose a novel loss function that incorporates the Earth Mover's Distance (EMD). This method determines the optimal alignment between two point cloud distributions, and allows for evaluation of the reconstruction accuracy.

In this paper, we therefore present the following contributions:

- An end-to-end deep learning framework for 3D human shape reconstruction from a 2D image without the need for a 3D prior parametric model.
- A novel loss function based upon EMD [5] to evaluate the distances between unordered 3D point clouds representing human body shapes.
- A synthetic pairwise 2D and 3D dataset to train our deep learning framework inspired by [6].

*Corresponding Author

2 THE RECONSTRUCTION NETWORK

We propose a deep architecture that has strong representation ability and makes use of the statistics learned from the associated geometric data. Given an input image S and a random vector t, the network reconstructs a 3D point cloud M_r through a CNN encoder and a fully-connected regressor.

To model the uncertainty of the input image, we propose the incorporation of a random perturbation vector t as a part of the input, together with the input image S [4]. The core of the network consists of a CNN encoder and a fully-connected regressor. The encoder is able to understand the features of images, while the regressor can capture complex structures to generate the corresponding 3D point cloud. As we do not enforce prior knowledge, we use an unordered point cloud set $M = (x_i, y_i, z_i)_{i=1}^{N}$ to represent the 3D shapes, where N is a predefined constant that represents the number of points in the point cloud.

We define the ground truth as a probability distribution $P(\cdot|S)$ over the shapes conditioned on the input 2D image S to model the uncertainty from 2D to 3D. We train a deep neural network G as a conditional sampler from $P(\cdot|S)$:

$$M = G(S, t; \theta), \qquad (1)$$

where θ denotes the network parameter, and $t \sim N(0, I)$ is the aforementioned random vector used to perturb the input. During testing, multiple samples of t are used to generate different predictions.

The encoder is composed of a combination of ReLU and convolution layers. It maps a random vector t and the input image S into a subspace. By using MoN (min of N) to model the uncertainty, the network can change its prediction based upon different random vectors. The regressor generates the 3D shape as an $N \times 3$ matrix, where each row represents the coordinates of one vertex.

3 THE EMD-BASED LOSS FUNCTION

While the use of an unordered point cloud frees the system from relying upon any priors, it is challenging to compare two unordered point clouds due to the lack of correspondence. Such a comparison is required when we build the reconstruction loss function. Motivated by [2, 5], we propose the use of EMD in our deep learning loss. EMD evaluates the minimum overall distance between two point clouds by finding the optimal mapping between them. It optimizes a set of unidirectional flows to map the points.

The loss function is defined as:

$$L(M_r, M_{gt}) = d_{EMD}(M_r, M_{gt}), \qquad (2)$$

where M_r is the reconstructed 3D human shape, M_{gt} is the ground truth of each sample, d_{EMD} is the EMD calculated as:

$$d_{EMD}(M_1, M_2) = \min_{\phi: M_1 \rightarrow M_2} \sum_{x \in M_1, y \in M_2} ||y - \phi(x)||_2, \qquad (3)$$

where $M_1, M_2 \in R^3$ has equal size, $m = |M1| = |M2|$ and $\phi : M_1 \rightarrow M_2$ is a bijection (i.e. flows), $|| \quad ||_2$ represents the root mean square point to point distance.

With the EMD-based loss function, the system can effectively evaluate the distance between the synthesized human shape and the ground-truth one for backpropagation during training.

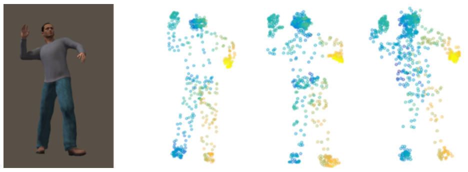

Figure 2: Different possible shapes for the same image.

4 PRELIMINARY EXPERIMENTAL RESULTS

Fig.1 shows the point cloud reconstructed by our system compared with the ground truth. Both examples show that our reconstructed point cloud resembles the body shape with the correct posture.

Fig. 2 shows that, due to the inclusion of a random vector, the same input image can have multiple plausible 3D shapes. Whilst there are small variations, the depth information is generally consistent and the overall posture provided is a good representation.

Our qualitative evaluations also suggest that the accuracy decreases as the amount of occlusion increases. We also observe that the presence of occluded body parts may affect other body parts which are not occluded. This is likely due to the EMD function attempting to find the optimal mapping for the whole body.

5 CONCLUSION AND DISCUSSIONS

In this paper, we propose an EMD-informed CNN framework for 2D to 3D point cloud reconstruction. Unlike the majority of previous works, we experiment with a setup in which there is no prior-knowledge. Our EMD function successfully solves the problem of using an unordered point cloud for prior-less human shape representation. Furthermore, to enable sufficient high-quality training data, we employ a computer graphics pipeline to generate synthetic training data. Our preliminary results suggest that the use of EMD demonstrates high potential in matching two prior-less point clouds in order to evaluate the reconstruction loss. However, when multiple joints are close together or occluded, the system has problems identifying which body parts the points should belong to, which results in poor quality during occlusion.

REFERENCES

[1] Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. 2010. Inferring 3D shapes and deformations from single views. In *European Conference on Computer Vision*. Springer, 300–313.

[2] Joseph Henry, Hubert P. H. Shum, and Taku Komura. 2012. Environment-aware Real-Time Crowd Control. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '12)*. Eurographics Association, 193–200.

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[4] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. 2017. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 86–94. https://doi.org/10.1109/CVPR.2017.17

[5] Y. Shen, L. Yang, E. S. L. Ho, and H. P. H. Shum. 2019. Interaction-based Human Activity Comparison. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. https://doi.org/10.1109/TVCG.2019.2893247

[6] Jingtian Zhang, Lining Zhang, Hubert PH Shum, and Ling Shao. 2016. Arbitrary view action recognition via transfer dictionary learning on synthetic training data. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 1678–1684.

[7] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. 2010. Parametric Reshaping of Human Bodies in Images. In *ACM SIGGRAPH 2010 Papers (SIGGRAPH '10)*. ACM, New York, NY, USA, Article 126, 10 pages. https://doi.org/10.1145/1833349.1778863