Human Point Cloud Generation using Deep Learning

Ryan Spick and James Walker
{ryan.spick,james.walker}@york.ac.uk
Tim Bradley and Nigel John Williams
{nigel.williams,tim.bradley}@sony.com

University of York,
York, Uk
Sony Interactive Entertainment,
London, UK

1 Introduction

Generative deep learning has been applied to a multitude of areas across many domains, each of these areas providing a different type of data from text, images, videos, and music [1, 3, 7, 8]. These examples use a variety of different network architectures, but the goal of each is to learn or exploit the underlying distribution of its training data. In this paper, a novel method of generating accurate pose and animation of human point cloud data using generative deep learning methods is presented, which uses dense correspondence based data, in which all points within the point cloud align with every other point in corresponding data points.

1.1 Point Clouds

Point clouds are sets of coordinates representing some multi-dimensional data, typically in a three-dimensional Cartesian coordinate frame, representing objects, surfaces, or shapes. In these cases, each point is represented with an x, y, and z component determining the geometric coordinate of each point in the cloud. These data points are usually the result of a type of 3D scanning such as LiDAR.

As of late, PointNet [5] type architectures, which facilitate the optimal consumption of point clouds directly by a neural network, have received a great deal of attention, though these approaches disregard potentially deep correspondences between points. PointNet++ [6] attempts to define weak correspondence through sampling overlapping regions/clusters, but this structure excludes any one to one point correspondence. If more information is known about the structure and layout of the point sets, then it is possible to derive well-defined correspondence that isn't specifically within the euclidean or geodesic space. PointNet architectures and their derivatives have been steadily applied in areas such as object detection [4], amongst many others.

2 Methodology

This work utilises the MPI-Faust data set [2] as the input data for the experiments and deep learning models in this paper. Each model file contains exactly 6890 points, where every point corresponds with every other point in the data-set. The data set consists of 10 unique models, of varying body structure and shape. There are 10 poses mirrored across every model. This dense formatting of the data allows for uniform learning of complex floating-point data through standard generative approaches, such as a Convolutional neural network or a static fully connected network. The MLP network is simply a fully connected network where every node in subsequent layers is connected to the previous layer's nodes. For the generator, the number of nodes increases with a factor of 2 each in each layer. The final output of the network increases to the size of the point cloud data, 6890, which is the size required for the input to the discriminator to match the real samples. The discriminator is a reverse of the generator, leading down to a node size of 1 with a Sigmoid activation, signaling if a sample passed in is determined real, or fake. A 1D convolutional network was also tested, but proved to have difficulties learning the symmetry of the data across different body shapes. We believe this was due to the inherent nature of the convolutions, having a lack of connection between those points that are close in space, but not in the data set.

2.1 Dot Based Loss Function

The idea of using the dot product is to add stability in the early stages of training. Because the data is in dense correspondence, a pre-computed dot product across the training data can be used to determine how well the generated samples conform to the original data distribution. A sample of the dot product of all of the training data was taken, where each dot product was taken for every point in each data point. Initially the calculation would take a point and its neighbouring point - where $D = Dot(N, N+1)$. Indexing all of the points proved far too inefficient to use during training, so random jittering was employed. This was changed to a stochastic approach where now - $D = Dot(N, N+step)$ with the step being a random value between roughly 0.8% and 1.2% of the data set size, the dot calculation was performed until N >= data length. This helped with performance without drastically reducing the quality of convergence to the pose shape.

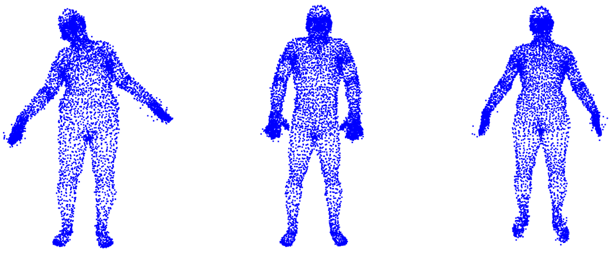

Figure 1: Generated examples from the MLP GAN, each model shows a different type of pose and body shape. Conditions were later added to control the type of pose and body shape through prior labelling.

3 Conclusion

This paper outlined a new method of loss calculation for ordered point cloud data, together with a deep parameter exploration of two neural network architectures has resulted in a robust method of generating new human pose and human poses animation from existing point cloud data. The idea of taking a prior dot product calculation and incorporating it into a weighted binary cross-entropy loss function provided a large stability increase when training the generator of the network. Subsequently improving the visual fidelity of human pose outputs and early training convergence.

[1] Adrián Barahona-Rıos and Sandra Pauletto. Synthesising knocking sound effects using conditional wavegan. In *17th Sound and Music Computing Conference, Online*, 2020.

[2] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. June 2014.

[3] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. Texture synthesis with spatial generative adversarial networks. *arXiv preprint arXiv:1611.08207*, 2016.

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[6] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[7] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.

[8] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.