Deep 4D Shape Representation: Learning 4D Volumetric Video from Skeletal Motion

João Regateiro Marco Volino Adrian Hilton
{j.regateiro,m.volino,a.hilton}@surrey.ac.uk
Centre for Vision, Speech and Signal Processing
University of Surrey

1 Introduction

Volumetric video is an emerging medium that allows free-viewpoint replay and rendering of dynamic scenes with the realism of captured video [1, 5]. This has the potential to allow highly realistic content production for immersive virtual and augmented reality experiences. Human models are typically rendered using detailed, explicit 3D models, which consist of meshes and textures, and animated using tailored motion models to simulate human behaviour and activity. However, designing a realistic 3D human model is still a costly and laborious process. Recent work [2] has shown that it is possible to learn and animate natural human behaviour (e.g. walking, jumping, etc.) from human skeletal motion capture data (MoCap) of actor performance. Motivated by recent advances in generative networks [2, 3, 4, 6] we propose an architecture for learning to generate dynamic 4D shape. We show in this paper how to use a variational encoder-decoder to learn the mapping from 3D skeletal motion to the corresponding full 4D volumetric shape and motion.

2 Method

The network architecture, shown in Figure 1, maximises the probability distribution of the 3D skeletal joint positions, $p = \{\{p_t^s\}_{t=1}^{N_t^s}\}_{s=1}^{N_S}$, encoded in the latent space, $z = \{\{z_t^s\}_{t=1}^{N_t^s}\}_{s=1}^{N_S}$, and learns the generative mapping of the decoder to the corresponding 4D shape $\tilde{M}_t^s$ for sequence $s \in N_S$ at time instance t. Generative networks learn dependencies from the input data and capture them in a low-dimensional latent vector z_t^s, creating compact representations $z_t^s \in \mathbb{R}^d$, where $d = 128$ is the latent space dimension.

$$P(p) = \int P(p \mid z)\, P(z) dz \tag{1}$$

The distribution $P(p|z)$ denotes the maximum likelihood estimation of dependencies of p over the latent vector z, and $P(z)$ is the prior probability distribution of a latent vector z, and $P(p)$ is the probability density function for the 3D skeletal pose. Here to ensure a compact representation $P(p|z)$ is modelled as a Gaussian distribution with mean $\mu(z)$ and diagonal co-variance $\sigma(z)$ multiplied by the identity I, which implicitly assumes independence between the dimensions of z.

$$P(p \mid z) = \mathcal{N}(p \mid \mu(z),\, \sigma(z)^2 * I) \tag{2}$$

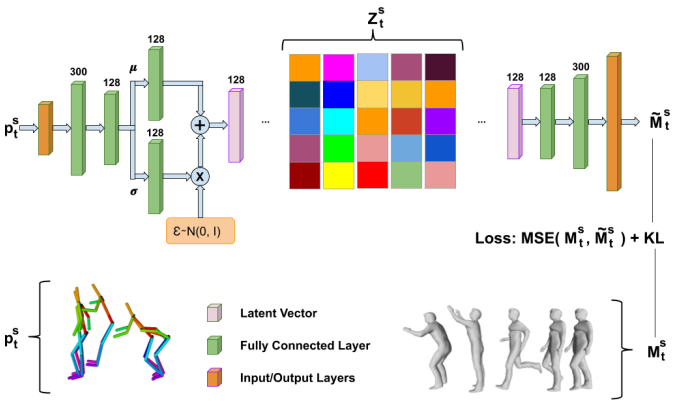

Figure 1: 4D shape representation network overview. The input is 3D skeletal motion and the output is 4D shape.

The variational encoder-decoder network architecture is composed of an encoder that receives 3D skeletal joints as input, and a decoder that generates high resolution 4D shape. It is trained over 10^4 epochs, which is optimised through validation data to avoid over-fitting with a learning rate of 0.001. The encoder is trained to map the posterior distribution of data samples p to the latent space z, meanwhile forcing the latent variables z to comply with the prior distribution of $P(z)$. However, both the posterior distribution $P(z|p)$ and $P(p)$ are unknown. Therefore, the variational encoder-decoder gives the solution that the posterior distribution is a variational distribution $Q(z_t^s|\tilde{M}_t^s)$, computed by a neural network. In order to make $Q(z_t^s|\tilde{M}_t^s)$ consistent with the distribution $P(z)$, we use the Kullback-Leibler (KL) divergence as follows:

$$\arg\min KL(Q(z_t^s \mid \tilde{M}_t^s) \mid\mid P(z_t^s)) \tag{3}$$

The decoder is trained to regress from any latent vector z_t^s in the latent space z to a 4D shape representation $\tilde{M}_t^s$. Equation 4 defines the loss function minimised by the network to achieve a compact latent space representation and generative network output.

$$L = (Q(P(p_t^s \mid z_t^s) \mid \tilde{M}_t^s) - M_t^s) + \frac{KL}{\omega} \tag{4}$$

This is an optimal approximation of the true samples M_t^s, where $\omega =$ batch size $\times$ input data size, and M_t^s is the ground truth 4D shape for the 3D skeletal pose p_t^s of sequence s at time t.

3 Results

The proposed generative network reconstruction error is ≈ 0.0072 m and ≈ 0.0036 standard deviation, for the results in Figure 2. The network achieves compact representation of 4D volumetric video sequences, see Figure 2, capable of two orders of magnitude compression compared to the captured 4D volumetric video. This allows applications to use less memory in run-time, making it more suitable for technologies with memory constraints. The network allows a mapping of skeletal motion capture data to generate novel 4D shape sequences. This gives the possibility to re-use MoCap datasets to generate novel 4D shapes, and creates opportunities for animation applications to easily incorporate novel motion sequences.

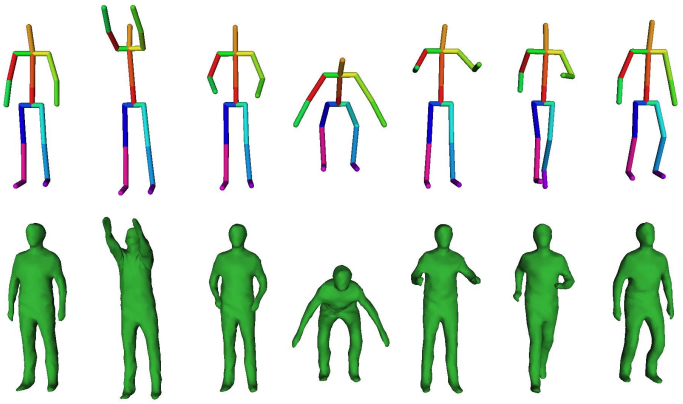

Figure 2: The top row represents the skeletal motion used to synthesis the 4D shape on the bottom row containing training and validation data.

Acknowledgements

This research was supported by the InnovateUK project Polymersive (105168) and EPSRC Audio-Visual Media Research Platform Grant (EP/P022529/1).

[1] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), 2015.

[2] D. Holden, T. Komura, and J. Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 36(4), July 2017.

[3] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 2018.

[4] Joao Regateiro, Adrian Hilton, and Marco Volino. Dynamic surface animation using generative networks. In *International Conference on 3D Vision (3DV)*, 2019.

[5] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3), 2007.

[6] Q. Tan, L. Gao, Y. Lai, and S. Xia. Variational autoencoders for deforming 3d mesh models. In *IEEE/CVF CVPR*, 2018.