Point Cloud Completion of Foot Shape from a Single Depth Map for Fit Matching using Deep Learning View Synthesis

Nolan Lunscher
University of Waterloo
200 University Ave W.
nlunscher@uwaterloo.ca

John Zelek
University of Waterloo
200 University Ave W.
jzelek@uwaterloo.ca

Abstract

In clothing and particularly in footwear, the variance in the size and shape of people and of clothing poses a problem of how to match items of clothing to a person. 3D scanning can be used to determine detailed personalized shape information, which can then be used to match against clothing shape. In current implementations however, this process is typically expensive and cumbersome. Ideally, in order to reduce the cost and complexity of scanning systems as much as possible, only a single image from a single camera would be needed. To this end, we focus on simplifying the process of scanning a person's foot for use in virtual footwear fitting. We use a deep learning approach to allow for whole foot shape reconstruction from a single input depth map view by synthesizing a view containing the remaining information about the foot not seen from the input. Our method directly adds information to the input view, and does not require any additional steps for point cloud alignment. We show that our method is capable of synthesizing the remainder of a point cloud with accuracies of 2.92±0.72 mm.

1. Introduction

In clothing and in particular in footwear, there are numerous brands and models that come in all shapes and sizes. Similarly, the shapes of individuals can be just as varied. In footwear, the complex shapes involved make pairing a person to a product challenging, something that is important as fit largely determines performance and comfort. Currently the primary and often only indicator used to specify fit is shoe size, which is not sufficient to fully characterize the profile of a shoe or a foot [8]. Additionally not every type of shoe, regardless of size, will fit every person the same way. In foot morphology, foot shape is complex and includes measures for various lengths, widths, girths and angles [7]. Due to this, it is not straightforward to determine how footwear will fit from the size alone, which for example poses a particular challenge in online shopping. Fit estimation could be improved by virtually fitting a precisely measured 3D foot shape model with a 3D shoe cavity. In this way, a single foot model could be compared against whole catalogues of footwear to find the best fit.

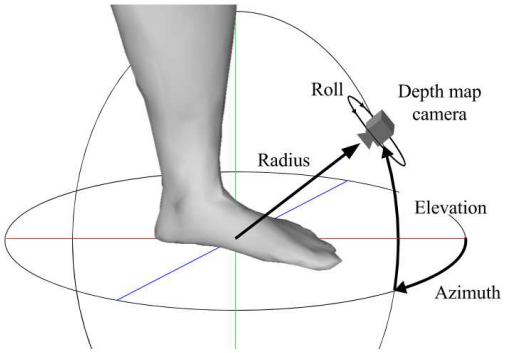

Figure 1. Depth camera pose configuration.

In order to achieve this, 3D scanning can be used to measure foot shape beyond a simple shoe size. Systems such as the Vorum Yeti[1] and the Volumental scanner[2] already exist, however they are not very common and tend to be expensive or cumbersome to operate. In recent years, RGBD cameras have generated a lot of interest in 3D scanning as they are affordable and easy to operate while providing sufficiently accurate depth maps. Various RGBD scanning techniques have been developed such as using a single moving camera's video [13], and using multiple stationary cameras [2, 10]. In either case, multiple images of the subject being scanned are required to be taken from various viewpoints covering all surfaces. When using RGBD cameras to scan a person however, artifacts from any movement, or from overlapping projector patterns from multiple cameras can complicate the process. Ideally, to maximize simplicity,

[1] vorum.com/footwear/yeti-3d-foot-scanner
[2] volumental.com

only a single camera image from a single instance in time would be needed. This would require that an overall object shape be captured from a single RGBD image from a single viewpoint.

Towards this goal, statistical models can be used to reduce the number of parameters needed to sufficiently reconstruct the overall foot shape [11, 12]. Similarly, a number of methods have been explored to create parameterized models of whole bodies [1, 16, 24]. These can be leveraged in learning methods to determine a mapping from an image or images to a set of parameters used to reconstruct a model from a template [5, 6]. The disadvantage to these methods however is that they rely on accurate measurements of predefined parameters that characterize the object being scanned. Accurate measurements of anthropomorphic body parts often require skill and patience, and are not as scalable as learning shape directly. In working with 3D structure directly, limited inputs can be used to build models in voxel volume representations [3, 4, 17, 20]. This 3D information can be operated on directly by deep neural networks through 3D convolution, however the complexity of these computations limits their resolutions, usually to 32x32x32 or less.

Another approach used to extrapolate information about an object is known as view synthesis, where the goal is to synthesize a novel view or views of an object or scene when given a single or set of arbitrary views. In deep learning, view synthesis systems can operate on 2D images, allowing for higher resolutions than voxel representations. A number of implementations exist that focus on RGB views [15, 22, 23], however it is difficult to extract 3D models using these techniques. In order to more easily work with 3D structure in view synthesis, a depth map view such as those provided by an RGBD camera can be used, as was done in work by Tatarchenko *et al*. [19]. Using a depth map allows for operations to take place on a point cloud, which can later be used to extract object shape.

We follow a deep learning view synthesis approach to capture full anthropomorphic body part shape from a single depth map input viewpoint. We used mesh data from MPII Human Shape [16] based on the CAESAR database [18], and focus on the application of foot scanning for virtual footwear fitting. In our approach we take advantage of the shape of a foot to select specific views that maximize our ability to reconstruct an overall foot point cloud from as few synthesized views as possible. We also introduce a way to synthesize a missing object view without the need to align the new view to that of the input when reconstructing the overall point cloud.

2. Proposed Method

We frame the problem of completing a limited foot scan point cloud as a depth map view synthesis problem. We leverage the power of deep learning to implicitly learn foot shape and the relationships in how it can appear between views. Our depth map view configuration is shown in Figure 1. A foot is placed at the origin, such that depth map images can be taken from camera poses at various azimuth, elevation and roll angles, as well as at varying radii.

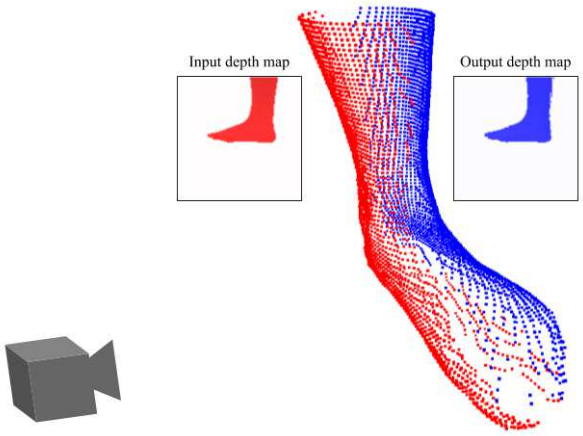

Figure 2. Depth map input/output configuration. Red: points from the input depth map, Blue: points from the synthesized output depth map.

Unlike general view synthesis problems, we restrict our input views to be only profile views of the foot rather than any arbitrary view. The profile of a foot contains a significant amount of information about overall shape [12], which a learning system can take advantage of. Additionally, our synthesized output views are always of the surface on the direct opposite side of the foot from the camera, produced as if it could be seen through the near side of the foot as shown in Figure 2. Due to how feet are shaped, these two opposite profile views of the foot contain the vast majority of points on the object's surface, as the foot has minimal self occlusions along this direction. In framing the problem in this way, the synthesized depth map can also be re-projected to the same coordinate system as the input view to produce a near complete point cloud of the foot surface without the need for any extrinsic parameters or alignment.

2.1. Dataset

Our network was trained using the meshed models from MPII Human Shape [16]. These meshes were created by fitting a statistical model of 3D human shape to full body scans from the CAESAR database [18]. We use the 4301 models that were fit using the posture normalization algorithm from Wuhrer *et al*. [21]. Each model consists of 6449 vertex points, with about 800 in each foot up to the knee. The MPII Human Shape models are technically a parameterized representation of shape, however these parameters were not used anywhere in our method, such that our net-

Figure 3. Meshed foot objects from MPII Human Shape [16].

Pose Parameter	Value Range	Step
Radius (mm)	640 to 700	15
Azimuth (deg)	70 to 110 or 250 to 290	1
Elevation (deg)	-20 to 20	1
Roll (deg)	-5 to 5	1

Table 1. Camera pose parameters for the network input.

work would learn its own representations of shape.

For each body model in the dataset, the points associated with the left and right feet up to the knee were separated. We then moved the origin to the center of the second toe and the heel. Following this process provided 8602 meshed foot objects for use in training our network, foot object samples are shown in Figure 3. We used Panda3D[3] to render depth map images of the foot objects at a size of 128x128, from the camera poses described in Table 1. The variations in camera poses used were intended to teach the network to handle cases of imperfect camera mounting rather than only dealing with perfect profile views. We additionally add random noise to the input depth maps using the Kinect noise model by Nguyen *et al*. [14].

2.2. Implementation Details

Our basic network architecture is similar to that by Tatarchenko *et al*. [19]. We use a deep convolutional deconvolutional neural network with fully connected layers in the middle to process an input depth map and synthesize an output depth map as shown in Figure 4. We train directly on a one channel depth map input, to produce a one channel depth map output. In this implementation, we do not incorporate any color information that would be present in a typical RGBD image.

When reconstructing the complete point cloud, we reproject the input depth map and the synthesized output depth map using the same camera parameters, as they are

[3]panda3d.org

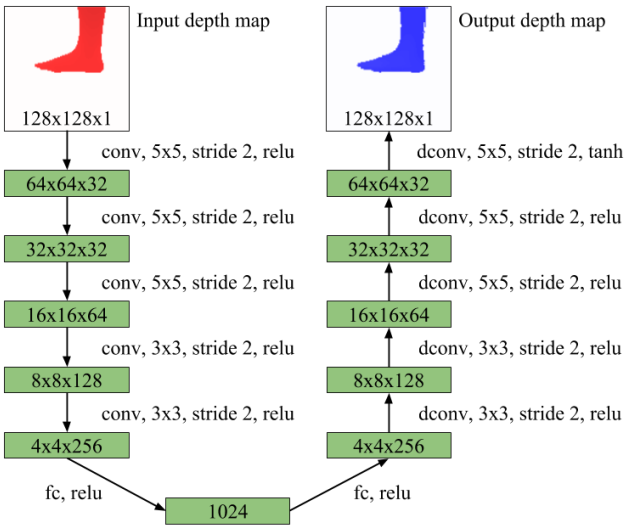

Figure 4. Network architecture.

already aligned. We also remove outliers and clean the point clouds using 3D cropping and MATLAB's *pcdenoise* function. Merging the points from the input and synthesized output produces a near complete foot point cloud.

Our dataset of 8602 feet was split into 80% train and 20% test, and was separated by individuals such that both of a persons feet would stay within the same set. We implemented our network in Tensorflow[4] on a Linux machine running an Nvidia K80 GPU. Training was done using a mini batch size of 64 and the Adam optimizer [9] with a learning rate of 5e-5. Our loss function was the mean L_1 distance between the output depth map pixel values and the ground truth pixel values.

3. Results

Our test set is comprised of 1720 foot objects that were not used in training. For each of the test objects, we generated 64 input-output pairs using the same camera pose parameters used in training. After 1,300,000 training iterations, our loss on the test set was 0.00541. Samples of synthesized depth map results are shown in Figure 5, along with the distribution of error within the depth maps. Looking at the error distributions, it can be seen that a large portion of the error is due to pixels on the foot outline. It appears that the network is uncertain whether pixels in these regions would be a part of the background or of the foot.

3.1. Point Cloud Results

We separately evaluate our network's synthesized depth maps in the context of generating a complete point clouds of entire feet. Each depth map from the test set is re-projected

[4]tensorflow.org

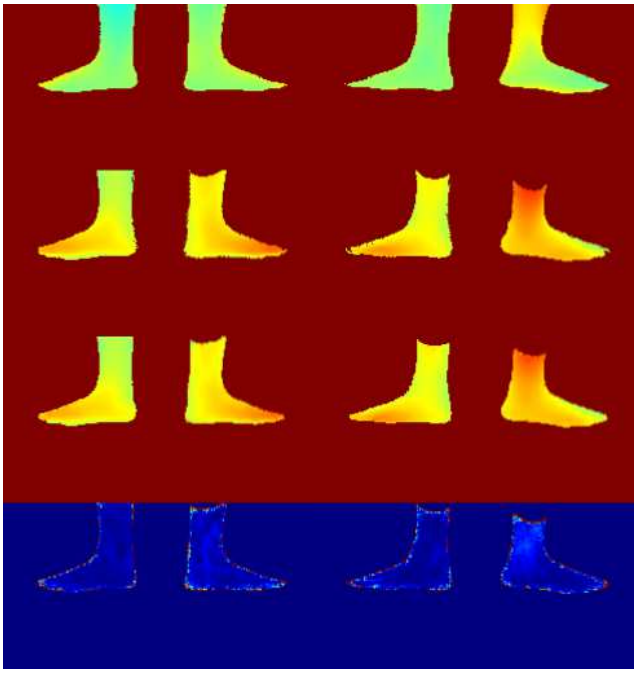

Figure 5. Sample depth map results. First row: input depth map, Second row: ground truth, Third row: synthesized output depth map, Fourth row: output depth map error.

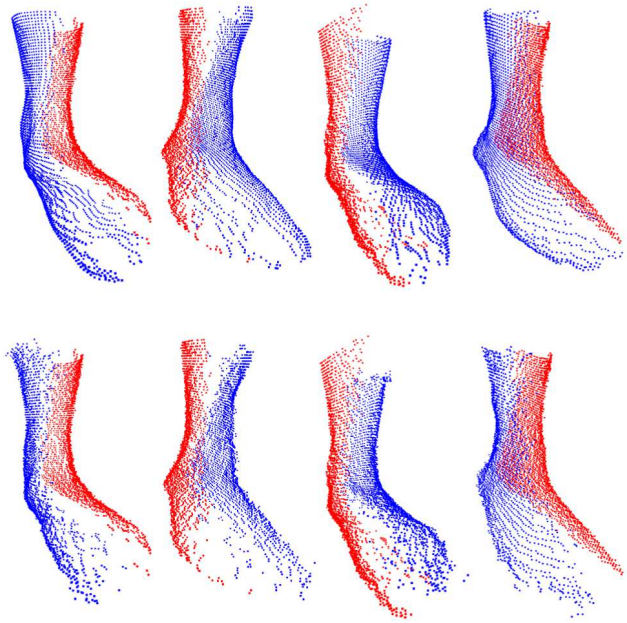

Figure 6. Sample point cloud results. First row: ground truth, Second row: synthesized point cloud. Red: input depth map points, Blue: ground truth/synthesized output depth map points.

to a point cloud in mm units. Sample point clouds are shown in Figure 6, compared with the ground truths.

We measure point cloud error in a method similar to that used by Luximon et al. [11], who parameterized a foot point cloud using a statistical model. Our measure is a two directional nearest neighbor euclidean distance metric that specifies the overall similarity between point clouds.

We calculate $e_{syn,i}$ as the error of point $\mathbf{p}_{syn,i}$ in the synthesized point cloud to the ground truth, by calculating its euclidean distance to the nearest point in the ground truth point cloud using the following equation:

$$e_{syn,i} = min_j \|\mathbf{p}_{syn,i} - \mathbf{p}_{gt,j}\|_2. \quad (1)$$

Similarly we calculate $e_{gt,j}$ as the error of point $\mathbf{p}_{gt,j}$ in the ground truth to the synthesized point cloud, by calculating its euclidean distance to the nearest point in the synthesized point cloud using the following equation:

$$e_{gt,j} = min_i \|\mathbf{p}_{gt,j} - \mathbf{p}_{syn,i}\|_2 \quad (2)$$

We normalized these measures using the number of points in each cloud, then averaged the normalized measures to form the overall error of our point cloud using the following equation:

$$e_{total} = \frac{\frac{1}{N}\sum_i e_{syn,i} + \frac{1}{M}\sum_j e_{gt,j}}{2} \quad (3)$$

where N and M are the number of points in the synthesized and ground truth point clouds respectively, and e_{total} is the total error reported for a point cloud.

Using this measure, we found that across all the images in our test set, our method produced points clouds with an average error of 2.92 mm with a standard deviation of 0.72 mm. Looking more closely at Figure 6, it can be seen that the synthesized point clouds perform accurately in areas that are dense with points, such as along the side of the foot, but has more difficulty producing points in sparse regions, such as along the top and bottom of the foot. These sparse regions corresponded to the high error points along the outlines of the depth maps and always occur in these areas as a by-product of how we chose our camera poses. We can additionally see the seam between the input and synthesized point clouds. This seam occurs along the surface of the foot that is at very high angles to the depth map camera, where reliable depth measurements cannot be taken by the input camera [14] and where the synthesis network becomes uncertain.

4. Discussions and Conclusions

We have presented a novel method for leveraging deep learning to perform 3D foot scanning from a single depth map input view. Our network was successfully able to learned 3D shape and determine the missing information

required to accurately generate a near complete point cloud representation of a foot from a single depth map. Our method was found to have a reconstruction accuracy of 2.92 ± 0.72 mm, which is more precise than the 4.23 mm half size increment used in the English and American shoe sizing systems, however it may not be accurate to the half sizes of 3.33 mm in the European sizing system [7].

Our method allows for foot scanners to avoid many of the complications associated with multi-camera and moving-camera systems such as extrinsic calibration, point cloud alignment, slow opperation, mechanical complexity and high cost. Due to its ability to require only a single input view, scanners can be made significantly cheaper and faster at capturing shape. The simplicity of our scanning method could make scanners far more accessible than current products on the market, and allow for more widespread use of scanning for footwear matching. This method also has potential in applications such as analysis of foot dynamics and loading from video, which can be useful in determining fit during motion and in footwear design.

When comparing with other forms of deep learning view synthesis for object reconstruction, our method is significantly simpler. We focused on how our network can supplement the existing data from the input depth map. By taking advantage of the general shape of a foot, we found that only a single additional view would be sufficient to reconstruct overall shape, requiring only a single forward pass of our network. Additionally, due to how we synthesize the additional view's depth map pixels from the same camera pose as the input, no additional steps to align the synthesized view with the original input are required to reconstruct the foot.

Despite our methods potential, there are limitations that make it less practical than traditional RGBD scanning methods in some aspects. The synthesized depth map is only an estimate and is not as accurate as a true scan taken from the same view, which would contain true information about shape. Our method will also not correctly reconstruct a foot in cases where for example there is for whatever reason some unique features on the side not seen by the camera with no indicating features on the surface that is seen by the camera. Our method also takes advantage of foot shape to select the two depth maps views used. For more complex objects with more self occlusions, a different implementation would be required to capture the whole surface. For these reasons, our method is not necessarily generally applicable and will not be practical as a replacement of traditional scanning techniques in all cases. In our application however, for most feet this method is sufficient to capture shape for use in virtual fitting.

In future works, we plan to investigate changes in network architecture as well as additional methods of preprocessing and post-processing the data to improve accuracy. We also plan to explore how color cameras could be used in single view scanning, as they are significantly cheaper, more readily available and often have higher resolutions than RGBD depth maps.

References

[1] F. Bogo, J. Romero, M. Loper, and M. J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

[2] Y. Chen, G. Dang, Z.-Q. Cheng, and K. Xu. Fast capture of personalized avatar using two kinects. *Journal of Manufacturing Systems*, 33(1):233–240, 2014.

[3] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[4] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. *arXiv preprint arXiv:1612.00101*, 2016.

[5] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 108–117. IEEE, 2016.

[6] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *European Conference on Computer Vision*, pages 88–104. Springer, 2016.

[7] R. S. Goonetilleke. *The science of footwear*. CRC Press, 2012.

[8] M. R. Hawes and D. Sovak. Quantitative morphology of the human foot in a north american population. *Ergonomics*, 37(7):1213–1226, 1994.

[9] D. P. Kingma and J. L. Ba. Adam: a Method for Stochastic Optimization. In *International Conference on Learning Representations 2015*, pages 1–15, 2015.

[10] S. Lin, Y. Chen, Y.-K. Lai, R. R. Martin, and Z.-Q. Cheng. Fast capture of textured full-body avatar with rgb-d cameras. *The Visual Computer*, 32(6-8):681–691, 2016.

[11] A. Luximon and R. S. Goonetilleke. Foot shape modeling. *Human Factors*, 46(2):304–315, 2004.

[12] A. Luximon, R. S. Goonetilleke, and M. Zhang. 3d foot shape generation from 2d information. *Ergonomics*, 48(6):625–641, 2005.

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

[14] C. V. Nguyen, S. Izadi, and D. Lovell. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 524–530. IEEE, 2012.

[15] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. *arXiv preprint arXiv:1703.02921*, 2017.

[16] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *CoRR*, abs/1503.05860, 2015.

[17] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems*, pages 4996–5004, 2016.

[18] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 380–386. IEEE, 1999.

[19] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016.

[20] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[21] S. Wuhrer, C. Shu, and P. Xi. Posture-invariant statistical shape analysis using laplace operator. *Computers & Graphics*, 36(5):410–416, 2012.

[22] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017.

[23] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301. Springer, 2016.

[24] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015.