MultiView Mannequins for Deep Depth Estimation in 360°Videos

Barnabas Takacs
PanoCAST/Drukka, Budapest, Hungary
btakacs@panocast.com

Zsuzsanna Vincze
Drukka/MOME, Budapest, Hungary
vincze@mome.hu

Gergely Richter
Drukka, Budapest, Hungary
gergely.richter@gmail.com

ABSTRACT

We estimate depth maps from real-life monocular 360°VR videos using a Deep Neural Net architecture trained on 3D point-based renderings of people (called depth "mannequins") captured with a Multi36 camera setup and processed with a combined pipeline of AI Instance Segmentation, Structure from Motion and Multi View Stereo methods.

CCS CONCEPTS

• **Shape modeling**; • **Artificial intelligence**; • **Interactive systems and tools**;

KEYWORDS

6DOF video, Depth Estimation, Artificial Intelligence, Virtual Reality, SFM, MVS, Point-based Rendering

ACM Reference Format:
Barnabas Takacs, Zsuzsanna Vincze, and Gergely Richter. 2020. MultiView Mannequins for Deep Depth Estimation in 360°Videos. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '20 Posters), August 17, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3388770.3407410

1 PROBLEM

Estimating depth in monocular 360°VR videos is an important element of 6DOF productions and special effects. Segmenting human shapes, however, is particularly challenging especially when both camera and people may move. Deep Learning techniques can be efficiently used to address this problem for 2D unconstrained videos [Moving Camera and Moving People: A Deep Learning Approach to Depth Prediction n.d.], however there is no available 3D data and methodology to readily extend these solutions into the 360°equirectangular domain.

2 APPROACH

We address this problem by first capturing a large and varied human pose/shape point cloud data set visualized via point-based depth rendering under 360°panoramic camera distortions, and subsequently training a Deep Neural Net to estimate 3D distances in each pixel by transfer learning. The resulting 360°weights turn monoscopic VR video sequences into 360°depth map estimates (Fig. 1).

The capture process uses **36 cameras** to image our performers from all angles. Each recorded frame is first processed using **AI Instance Segmentation** to generate segmentation masks for a fully automated 3D reconstruction process based on **Structure from Motion** (SfM) and **Multi View Stereo** (MVS) methods [Schonberger and Frah 2016] (Fig 2/Left). The resulting **point clouds** are visualized using point-based rendering with **variable point sizes** and **depth coloring** as a function of distance from a **virtual panoramic camera** placed in the center (Fig. 2/right). Using this process we create a large and varied **training data set** for the neural network by placing our "Mannequin Figurines" at variable distances (translations) rotations (yaw-pith-roll), poses and shapes, effectively yielding characteristic equirectangular distortions in 360°camera space (implemented in Unity3D).

3 RESULTS

To create the training data set we recorded 7 people for an average of 3 minutes each and reconstructed 50 mannequins per person (0.5% reconstruction rate) rendered from 36 view angles with 29 rotation and 7 translation parameters and yielding 227K synthetically generated 4K panoramic images. The performance of the Hyper360 neural net was qualitatively evaluated on real life virtual reality production footage recorded with various camera rigs (*GoPro 6 rig* – Video1, Samsung *Gear360* Video2 – Fig. 3/Left) and also compared to ground-truth data with respect to reconstructed stereoscopic 360°videos [Takacs 2019] using *Insta360Pro* footage (Fig. 3/Right – Video3).

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme, grant no 761934, Hyper360 ("Enriching 360 media with 3D storytelling and personalisation elements"). http://www.hyper360.eu/

REFERENCES

Moving Camera, Moving People: A Deep Learning Approach to Depth Prediction, n.d., https://ai.googleblog.com/2019/05/moving-camera-moving-people-deep.html

Schonberger J.L., and J.M. Frah, Structure-from-Motion Revisited, in Proc Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Takacs, B. et. al. 2019. Hyper 360 – Towards a Unified Tool Set Supporting Next Generation VR Film and TV Productions in J. Software Engineering and Applications, 12,127-148, https://doi.org/10.4236/jsea.2019.125009

SIGGRAPH '20 Posters, August 17, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7973-1/20/08.
https://doi.org/10.1145/3388770.3407410

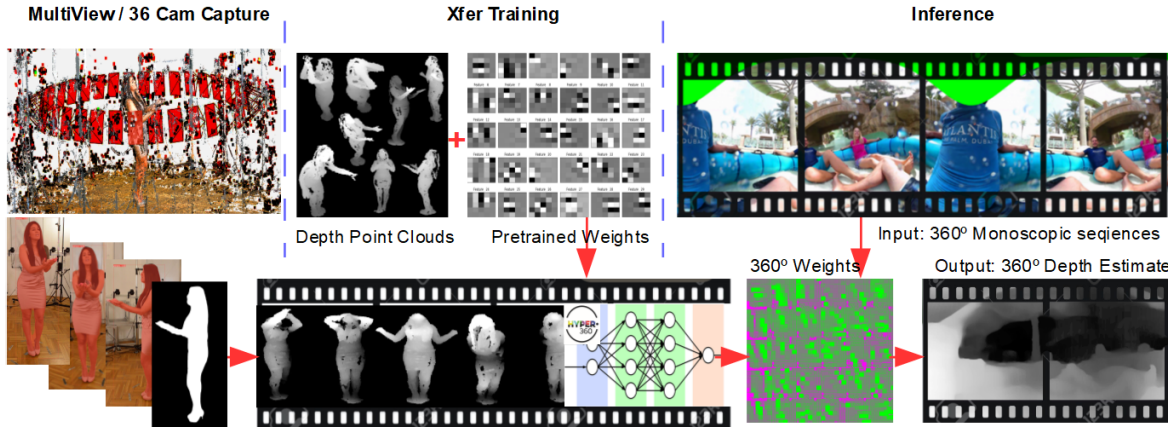

Figure 1: Overall processing pipeline.

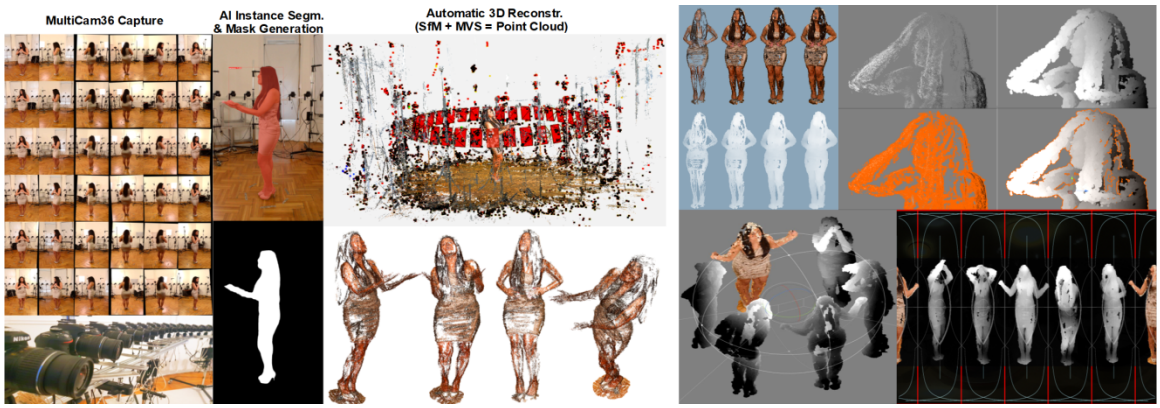

Figure 2: Multi-Camera capture and 3D reconstruction where AI instance segmentation generates masks for SfM/MVS reconstruction processes (left). Point-based rendering and 360ºdistortions used to create the training data set. (right).

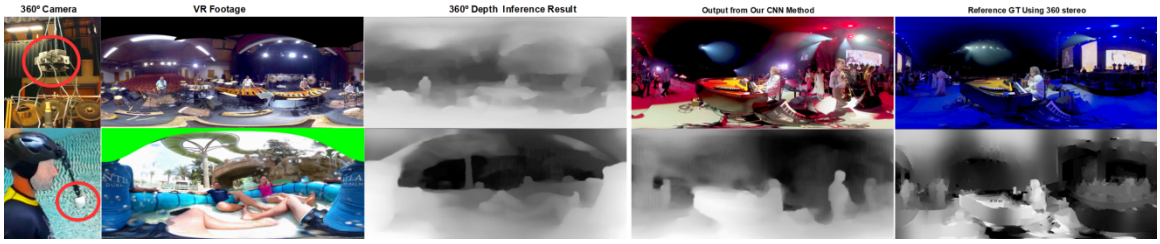

Figure 3: 360ºdepth reconstruction results from monoscopic VR video sequences and comparison with GT.