Automatic Generation of Interactive 3D Characters and Scenes for Virtual Reality from a Single-Viewpoint 360-Degree Video

Gregoire Dupont de Dinechin* Alexis Paljic†

Centre for Robotics, MINES ParisTech, PSL Research University

ABSTRACT

This work addresses the problem of using real-world data captured from a single viewpoint by a low-cost 360-degree camera to create an immersive and interactive virtual reality scene. We combine different existing state-of-the-art data enhancement methods based on pre-trained deep learning models to quickly and automatically obtain 3D scenes with animated character models from a 360-degree video. We provide details on our implementation and insight on how to adapt existing methods to 360-degree inputs. We also present the results of a user study assessing the extent to which virtual agents generated by this process are perceived as present and engaging.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing methodologies—Computer graphics—Animation—Motion capture;

1 INTRODUCTION

Recreating a real-world scene in all its complexity for virtual reality (VR) would require acquiring considerable amounts of input data. A popular low-cost and easy-to-use alternative is to acquire the scene from a single viewpoint using a 360-degree camera, e.g. by mounting the camera on a fixed tripod, thereby capturing all the visual and audio information coming into that viewpoint and saving it as a video file that can directly be played back into a VR head-mounted display. However, in terms of interaction, such video scenes seem quite limited when compared to the 3D computer-generated (CG) scenes typically used for VR. Users attempting to move their head away from the captured viewpoint will notably feel a cruel lack of motion parallax - the feeling that, as one moves one's head, closer objects seem to move faster than objects further away. Users will not either be able to interact with characters in the video without these characters unnaturally being frozen in place whenever the scenario requires the user to perform an action to continue.

We provide two main contributions towards increasing the interactive potential of 360-degree video data. First, we tackle the problem using a novel approach based on adapting and combining existing deep learning models for depth and human pose estimation. We demonstrate that an efficient implementation can already be achieved by using pre-trained models from state-of-the-art literature, and detail how we selected and adapted them to obtain a complete working process. Second, we evaluate this approach, by conducting a study to assess whether the video-based animated characters obtained by our method are perceived by users as exhibiting convincing responsive behaviour. Our protocol is inspired by existing works studying user response to CG virtual agents, and aims to help determine whether the process is applicable to similar use cases.

*e-mail: gregoire.dupont_de_dinechin@mines-paristech.fr

†e-mail:alexis.paljic@mines-paristech.fr

2 RELATED WORK

Multiple recent works have proposed estimating depth information from 360-degree video data to create VR scenes with head motion parallax [2, 6]. Indeed, estimating depth enables reprojecting the input image data onto the user's viewpoint, thereby preventing important discomfort which would otherwise occur during movement [2]. This is increasingly being done using deep learning models for dense depth estimation from a single image, which have recently been adapted to work on 360-degree image data [8].

The next step is enhancing the dynamic elements of the video, and most prominent among these are the characters in the scene. In the literature, most works focus on studying user response to CG virtual agents: the impact of giving these agents certain responsive traits, such as idle behaviour and mutual gaze, is evaluated using standard questionnaire responses and behavioural measures [1,4]. To enhance the characters in our video scenes, it therefore seems natural to start by transforming them into similar 3D rigged character models - e.g. by estimating human body shape [7], pose [3] and texture [5] from our images - before similarly evaluating the obtained virtual agents.

3 APPROACH AND IMPLEMENTATION

In a first step, we seek to estimate a dense depth map for the static elements in our scene, in order to create a 3D background environment. For our implementation, we used the pre-trained UResNet model presented by Zioulis et al. [8], expected to produce satisfying results for indoor scenes resembling those of the training datasets. The background image can be obtained directly during acquisition for controlled scenes. In other cases, background/foreground subtraction methods can also be used, since our considered video is taken from a single viewpoint. We then generate a 3D mesh from the dense depth map directly in the Unity game engine, by projecting the image's pixels in 3D space based on their depth values and corresponding 360-degree projection.

The second step is detecting which parts of the 360-degree image contain people we want to transform into 3D models. Indeed, most of the methods estimating pose, shape and texture were designed and/or trained to work for planar images (i.e. images obtained by perspective projection on a pinhole camera), and thus often underperform when given a spherical image as input. We therefore need to reproject our images around the people of interest, which requires detecting them first. To do so, we applied the pre-trained AlphaPose [3] model for 2D pose estimation, which we found was one of the few models able to reliably detect people in our videos despite the images being 360-degree. We then reprojected the images onto a virtual pinhole camera, designed for each image with the right extrinsic and intrinsic parameters to view only - and fully - the bounding box in the image of the person of interest.

In a third step, we immediately apply pre-trained models for shape, pose and texture estimation. For our implementation, we used the model for Human Mesh Recovery (HMR) [7] to recover characters' shape and pose, and adapted the output to obtain an animated character model in Unity. Note that storing the virtual camera parameters in the second step allows us to reposition our characters in absolute 3D space, thanks to the input data being 360-degree. We then used the pre-trained model for DensePose [5] to

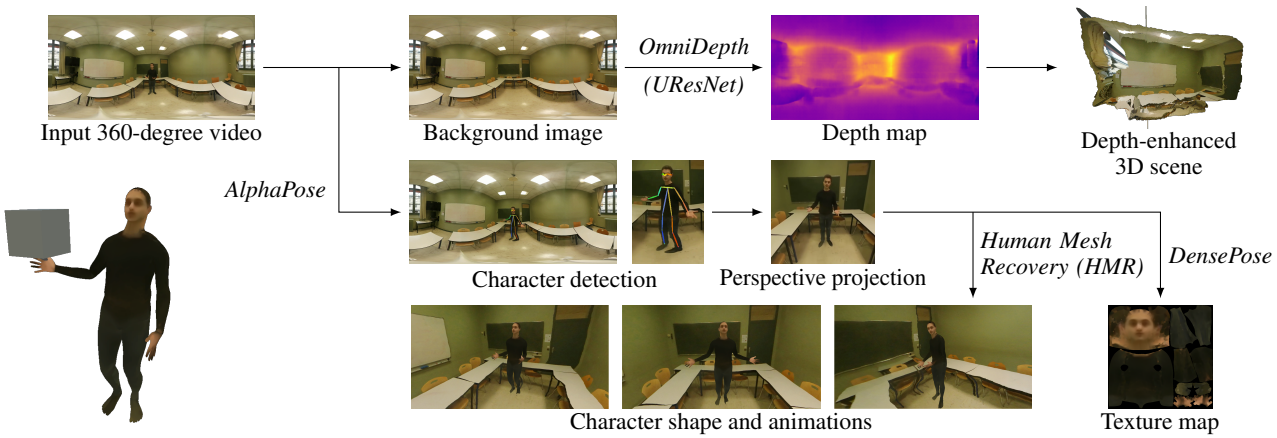

Figure 1: The different steps of the proposed approach, illustrated with results from our implementation.

generate the character's texture map, by selecting and projecting relevant color information based on the obtained dense 2D pose estimation for a small number of frames.

In a fourth step, we implement elements of responsive behaviour to transform our animated character model into a more convincing virtual agent. For our user study, we made the character rotate its head towards the user when the user looked in its direction or entered its personal space. We also added idle animations for sequences where the user was required to perform a specific action: more specifically, the agent looped over the last few seconds of the previous animation at a slower speed, producing motion resembling the character breathing and making small body movements. Finally, we made the agent in our user study verbally express discontent when users threw objects at it, because we expected this to be an interaction of interest for our study.

4 User Study

Our empirical research question was the following: to what extent do we obtain observations similar to those obtained by previous works studying CG virtual agents when we evaluate, using the same procedures, the video-based agents that result from our method? Our main hypothesis was that virtual agents made more interactive, e.g. given idle animations or the ability to look back at the user, should be perceived by participants as being more responsive, e.g. be more often interacted with or given more personal space. We used a between-subjects experimental design, with the agent's level of responsiveness as independent variable (the agent either exhibited all the elements of responsive behaviour described in Sect. 3 or exhibited none of them). Dependent variables included a number of self-report measures presented as a 7-point Likert-type post-test questionnaire, with many questions inspired by those presented by Garau et al. [4], as well as a number of behavioural measures, such as minimum interpersonal distance over the course of the experience. The 360-degree scene lasted 1 minute and 39 seconds, involved a single character, and included a sequence in which users had to get close to the character to grab a virtual object from its hand.

We recruited 50 volunteers (35 male, 15 female) on university campus to participate in the study, assigned randomly to each group. Data analysis was performed in R using a one-way analysis of variance (ANOVA) with Bonferroni's correction of p-values. Based on answers to the questionnaire, users felt that they initiated interaction more ($F(1,48) = 7.93$, $p < 0.01$) and felt more looked at by the agent ($F(1,48) = 5.17$, $p < 0.05$) when the character presented responsive traits. This is coherent with the work hypothesis and previous observations in the literature. More unexpected is that users also moved closer to the agent when it exhibited responsive behaviour ($F(1,48) = 5.99$, $p < 0.05$), whereas we would have expected the agent to be given more personal space.

Overall, user response to the enhanced scenes was quite positive, with particular praise being given to the quality of the agents' animations. Multiple users also reported that hearing audio of the person talking while seeing the agent move accordingly reinforced the sense of the agent being a person. In contrast, the quality of the agent's texture was often criticized, with the lack of dynamic facial features being particularly perceived as disturbing. The depth-enhanced background scene was also perceived as unconvincing. There therefore remains much room for future work.

5 Conclusion

We presented a novel approach to make scenes created from a single-viewpoint 360-degree video more interactive by way of automatically generating interactive 3D environments and character models, and demonstrated and evaluated an implementation of this approach.

References

[1] A. Bönsch, S. Radke, H. Overath, L. M. Asché, J. Wendt, T. Vierjahn, U. Habel, and T. W. Kuhlen. Social VR: How personal space is affected by virtual agents' emotions. In *Proc. Virtual Reality and 3D User Interfaces (VR)*. IEEE, Mar. 2018. doi: 10.1109/vr.2018.8446480

[2] G. Dupont de Dinechin and A. Paljic. Cinematic virtual reality with motion parallax from a single monoscopic omnidirectional image. In *Proc. Digital Heritage (DH)*. IEEE, Oct. 2018.

[3] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *Proc. International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. doi: 10.1109/iccv.2017.256

[4] M. Garau, M. Slater, D.-P. Pertaub, and S. Razzaque. The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators & Virtual Environments*, 14(1):104–116, Feb. 2005. doi: 10.1162/1054746053890242

[5] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018.

[6] J. Huang, Z. Chen, D. Ceylan, and H. Jin. 6-DOF VR videos with a single 360-camera. In *Proc. Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2017. doi: 10.1109/vr.2017.7892229

[7] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018.

[8] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras. OmniDepth: Dense depth estimation for indoors spherical panoramas. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 453–471. Springer International Publishing, Sept. 2018. doi: 10.1007/978-3-030-01231-1_28