Virtual Agents from 360° Video for Interactive Virtual Reality

Grégoire Dupont de Dinechin, Alexis Paljic

Virtual Agents from 360° Video for Interactive Virtual Reality

Grégoire Dupont de Dinechin
Alexis Paljic
gregoire.dupont_de_dinechin@mines-paristech.fr
alexis.paljic@mines-paristech.fr
Centre for Robotics, MINES ParisTech, PSL University
Paris, France

ABSTRACT

Creating lifelike virtual humans for interactive virtual reality is a difficult task. Most current solutions rely either on crafting synthetic character models and animations, or on capturing real people with complex camera setups. As an alternative, we propose leveraging efficient learning-based models for human mesh estimation, and applying them to the popular form of immersive content that is 360° video. We demonstrate an implementation of this approach using available pre-trained models, and present user study results that show that the virtual agents generated with this method can be made more compelling by the use of idle animations and reactive verbal and gaze behavior.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Computing methodologies** → *Machine learning*; *Computer vision*.

KEYWORDS

virtual agents, 360° video, interactive virtual reality, gaze behavior

ACM Reference Format:
Grégoire Dupont de Dinechin and Alexis Paljic. 2019. Virtual Agents from 360° Video for Interactive Virtual Reality. In *Computer Animation and Social Agents (CASA '19), July 1–3, 2019, PARIS, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3328756.3328775

1 INTRODUCTION

360° video is a popular and efficient way of capturing the real world for virtual reality (VR). (1) The recording process is easy, quick, and low-cost. (2) It produces high-fidelity visual results that are adapted to all-around VR viewing. (3) It is able to capture real-world motion, which includes - perhaps most importantly - the motion of people.

This last aspect is important, because although people are a fundamental part of our everyday lives, they remain quite complex to simulate convincingly in virtual worlds. Consequently, there is a need for an efficient, high-quality method able to quickly generate convincing virtual agents. We believe that this method could be to automatically estimate 3D characters from video recordings of people using learning-based models. One could thereby get the advantages of both video - that captures real-world motion with visual fidelity - and 3D characters - that are easy to animate and able to provide motion parallax for VR viewing.

In this paper, we therefore seek to evaluate the extent to which 360° video can help generate compelling virtual agents for VR. We provide two main contributions in this regard. First, we detail our method for applying learning-based human mesh generation models - which are generally trained on regular/planar images - to the 360°/spherical inputs typically used for VR. Second, we validate user perception of the generated character models by way of a study focused on agent responsiveness.

2 RELATED WORK

2.1 Human shape and pose from video

Our research leverages work from the field of human body shape and pose estimation from images. Within this field, we specifically study methods that satisfy three distinct constraints.

First, the method should be applicable to monoscopic, fixed-viewpoint video, since most 360° content takes this form. The method must therefore be able to estimate the output parameters from color images without depth or stereo, as was recently achieved quite convincingly e.g. by Xu et al [14] and Alldieck et al [1, 2].

Second, the output data must enable creating a textured, animated 3D character mesh. Methods that estimate parameters for the Skinned Multi-Person Linear (SMPL) model [12] are practical in this sense, especially for VR research, since the model is easy to understand and has been adapted for use with the Unity game engine. Notable approaches based on this model include those presented by Kanazawa et al [9], Güler et al [8], and Alldieck et al [1, 2].

Third, the method should be efficient and produce high-quality visual results, to encourage adoption by content creators. Learning-based models are likely more adapted in this sense than those relying heavily on optimization, as was reported by Alldieck et al [1] who thereby describe cutting processing time from hundreds of minutes to a few seconds for similar visual results.

2.2 Creating compelling virtual agents

We also seek to evaluate the extent to which the 3D human we generate by this method convincingly simulates a social presence in 3D space. This is likely to be impacted by the quality and realism of its mesh and animations, but also by the agent's ability to react to the user's presence and actions in a believable way [7].

Gaze behavior is a first important cue in this regard. Mutual gaze (between the user and the virtual agent) has notably been underlined by the literature as a crucial factor to increase perceived responsiveness and feeling of copresence [7, 10, 11]. Joint gaze (towards an object) was also recently shown to have a positive

CASA '19, July 1–3, 2019, PARIS, France

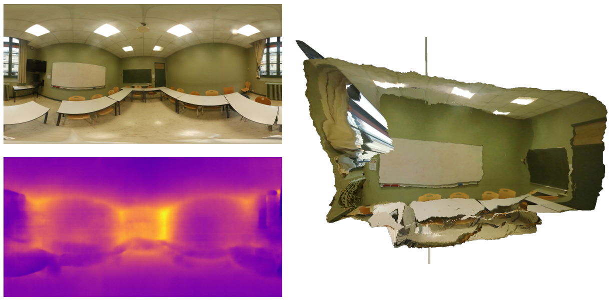

Figure 1: Background mesh estimation.

impact on user preference [10]. Extending gaze with specific facial expressions can also modify users' perception of the agent [4].

Interpersonal distance (between user and virtual human) is another factor to keep in mind. Indeed, proxemics can be used by virtual agents to elicit a feeling of intimacy, especially when coupled with eye contact [11]. Inversely, they can cause feelings of unease if users perceive an invasion of their personal space [3, 4].

Evaluating the extent to which users actually perceive the virtual agent as a social entity is then typically done using both questionnaire data and behavioral measures [7, 11]. Such measures include users' gaze direction and how close they get to the agent [3].

To compare to these works, our user study therefore includes mutual gaze as a factor, and measures several responses similar to those used by Garau et al [7] and Bailenson et al [3].

3 OUR APPROACH

3.1 Estimating a background mesh

A preliminary step in our approach is to create a 3D background mesh in which to place the virtual agent. Indeed, users will likely be required to move around to interact with the agent. If we simply project a 2D 360° image as background, this will be uncomfortable for users due to the lack of motion parallax [5].

A simple alternative is to estimate a background depth map, which can then easily be used to create a 3D mesh (see Figure 1). A requirement for this step is to have an image of the scene's background. This can easily be obtained either by capturing an image at acquisition, or by using background subtraction methods during processing (since our input is a fixed-viewpoint video). Using this background image, one can then immediately apply a pre-trained depth estimation model to infer the corresponding depth map. Note that this model must have been trained on images similar to the desired scene. For instance, to process the indoors scene we used for our user study, we applied the pre-trained UResNet model presented by Zioulis et al [15], designed for 360° images.

3.2 Obtaining a textured and animated character model

Very few learning-based models are trained to estimate characters' shape, pose or texture from 360° inputs. This causes many of them to - understandably - fail on such images, as we found was for example the case for the pre-trained Human Mesh Recovery (HMR) [9] and DensePose [8] models. Instead of re-training these models on 360° images, we propose a way to transform our spherical input images into planar ones that the pre-trained models can work with.

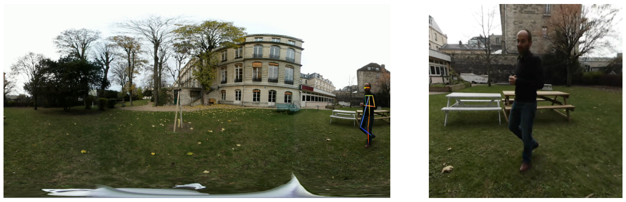

Figure 2: Character detection to obtain planar images.

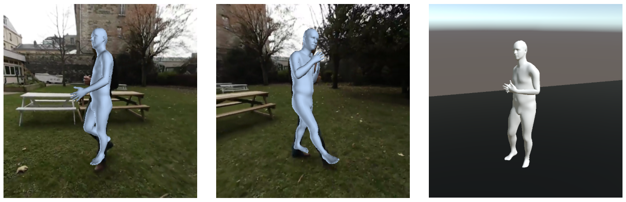

Figure 3: Human mesh estimation.

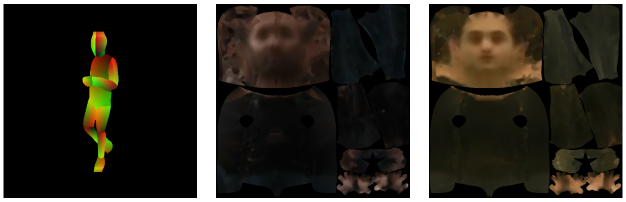

Figure 4: Texture map estimation.

First, we automatically detect where the character is located in the 360° image. In many cases, we could use background subtraction methods. However, we found that in practice an easier solution was to apply the pre-trained model for AlphaPose [6], a multi-person character detection model that we found worked consistently even on 360° inputs. Second, we use this character detection step to obtain, for every frame and person, a planar image containing entirely - and almost only - that person (see Figure 2). Specifically, we use the obtained bounding boxes to modify the direction and field-of-view of a virtual pinhole camera in order to capture images centered on the character in the 360° frame.

We then use the obtained planar image sequences as input to the pre-trained models for shape, pose and texture estimation. For our demo implementation, we estimated shape and pose parameters using the pre-trained model for HMR [9], which we converted into animation clips in Unity (see Figure 3). We then obtained the texture map by applying the pre-trained model for DensePose [9], the result of which we averaged and cleaned up automatically over a few dozen frames (see Figure 4). Note that this texture estimation step is necessary: pose estimation is not yet precise enough to allow simply projecting the image onto the posed character in order to obtain a texture map. Also note that we only use a few frames (e.g.

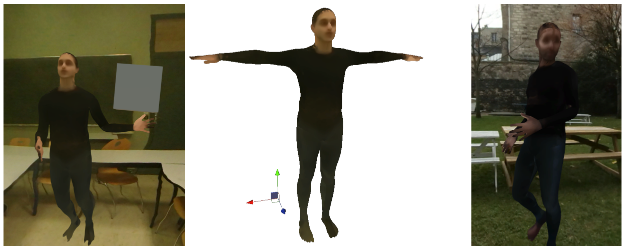

Figure 5: Generated 3D virtual agents.[1]

one in ten) to estimate pose, between which we interpolate the animation, in order to produce more natural results. Characters can then be positioned in absolute 3D space using the virtual camera parameters and the local estimated pose in this camera's view space.

3.3 Adding responsive behavior

A final step is to provide the virtual agents with lifelike behavior, for example by adding responses to user interaction.

For the character in our user study, we first implemented a reactive form of head-gaze: the agent turned its head towards the user when the user looked towards it or came close to it (< 2m).

Second, we gave our character idle animations. This allowed the agent not to stay completely frozen when waiting for the user to perform an action, despite us not having recorded additional video material. Specifically, we made the last few seconds of the previous animation loop at a slower speed, producing motion resembling the character breathing and making small body movements.

Third, we made the virtual agent verbally express discontent should the user attempt to throw one of the scene's virtual objects onto it. Note that users had no instruction to do so.

4 USER STUDY

4.1 Research question and hypotheses

Previous work has shown that participant response to virtual agents follows certain tendencies, e.g. users perceiving mutual gaze as a form of increased social presence. We want to know if these results hold when replacing the hand-made virtual agents of previous studies with our video-based character models.

We define two study hypotheses: **(H_1)** that agents made more interactive, e.g. given idle animations and reactive gaze behavior, will be perceived by participants as being more responsive, e.g. be more often interacted with and given more personal space; and **(H_2)** that participants shown a segment of the original video before being shown the 3D scene will also perceive the agent as more engaging, since this added information will help users relate the virtual human to the real-world person it is based on.

To test these hypotheses, we captured a 360° video using a tripod and a Samsung Gear 360 camera at 3840x1920 resolution, and transformed it into a VR scene for the user study. Total processing to transform the 99 second-long video into a 3D scene and animated character model took less than 3 minutes. During the scene, users were asked by a virtual agent to take or catch 3D objects appearing in its hand, thereby encouraging proximity and interaction.

Table 1: Group means for measured variables.

	R_-V_-	R_-V_+	R_+V_-	R_+V_+
Age (y)	25.92	26.67	25.38	25.38
Gender (# m/f)	9/3	7/5	11/2	8/5
Famil. VR (1-7)	3.25	4.17	3.15	3.00
Famil. 3D Char. (1-7)	2.92	4.08	2.92	4.31
Copresence (1-7)	4.90	5.00	4.98	4.81
Part. Behavior (1-7)	3.54	2.96	4.46	4.27
Perc. Awareness (1-7)	3.42	3.69	4.46	4.10
Comfort (1-7)	5.58	4.50	5.54	4.46
Min. Inter. Dist. (m)	0.87	0.90	0.77	0.78
Gaze Count (#)	11.92	12.67	13.31	13.23
Gaze Duration (s)	3.91	3.52	3.30	3.27
Throw Count (#)	*	*	0.38	1.07

* Not recorded due to coding error.

4.2 Independent variables

We used two active independent variables. To test H_1, the agent was either **(R_+)** given the aforementioned real-time responsive features or **(R_-)** not given responsive features. To test H_2, the VR scene started by displaying either **(V_+)** the first 10 seconds of the video or **(V_-)** only the background image with the video's audio.

Several attribute independent variables - specifically age, gender, level of familiarity with VR, and level of familiarity with 3D character models - were also recorded in a pre-test questionnaire, as some have been shown to be relevant in similar previous studies [3, 7].

4.3 Dependent variables

We measured user response using a post-test questionnaire with 7-point Likert-type questions. Referring to the study of Garau et al. [7], we measured **Copresence** (Did you have a sense that you were in the room with another person or did you have a sense of being alone? | To what extent did you respond to the character as if he was a person? | To what extent did you have a sense of being in the same space as the character? | Did you respond to the character more the way you would respond to a person, or the way you would respond to a computer interface?), **Participant Behavior** (To what extent did the presence of the character affect the way you explored the space? | Did you attempt to initiate any interaction with the character?), and **Perceived Awareness** (How much did the character seem to respond to you? | How much was the character looking at you? | How much did the character seem aware of your presence?). We also recorded **Comfort** (Did you feel comfortable moving around the space?). Additionally, we recorded several behavioral measures: minimum interpersonal distance, number of times that users looked towards the agent, duration of these looks, and number of times that users threw an object onto the agent.

[1] For more visuals, see the supplementary video.

Table 2: Significant results.

		ANOVA		ART	
Ind.	Dep.	$F_{(1,46)}$	p	$F_{(1,46)}$	p
R	Part. Behavior	7.36	0.03	5.96	0.06*
V	Comfort	7.11	0.03	8.95	0.01
R	Perc. Awareness	4.45	0.12*	4.77	0.10*
R	Min. Inter. Dist.	5.77	0.06*	5.74	0.06*

* Unadjusted p-value is below significance level.

4.4 Results

Results for each participant group are shown in in Table 1. 50 volunteers recruited on university campus took part in the study. We used a between-subjects experimental design to prevent bias linked to users understanding the purpose of the study: participants were randomly assigned to the four user groups, and were kept blind to the studied hypotheses.

Data analysis was performed in R. To better compare with results from previous research, we analyzed our data using a between-subjects two-way analysis of variance (ANOVA) with the aforementioned active independent variables as factors, significance set to α=0.05, and Bonferroni-adjusted p-values. Since we use Likert-type data in our study, we also complemented our examination with the nonparametric Aligned Rank Transform (ART) method specifically designed by Wobbrock et al. [13] to analyze such responses.

Table 2 details the results of this analysis. These results show that users rated Participant Behavior significantly higher when faced with more reactive virtual agents, and rated Comfort substantially lower when shown the actual video of the person. Moreover, perceived agent awareness increased and minimum interpersonal distance decreased when users were faced with more reactive agents, although these two trends were not found to be significant after adjustment of the p-values.

4.5 Discussion

The results underline that giving the virtual agent interactive features contributes to increasing users' feeling that it is a responsive entity. Users also seem encouraged to get closer to the agent in order to interact with it, while remaining far enough not to enter its perceived personal space. All of this supports H_1, and is consistent with results from previous research [3, 7, 11]. We can therefore validate a certain degree of similarity between hand-made virtual agents and our automatically-generated characters.

The influence of being shown the original video on perceived comfort is more unexpected. A possible interpretation is that the added motion caused by replacing the still 360° background image with a video is enough to be a source of discomfort in VR. In any case, there seems to be no impact on the perceived responsiveness of the virtual agent, and therefore H_2 cannot be validated.

Finally, user comments seem to indicate generally favorable response to the characters. A strong point is the quality of the animations, which were found to appear natural and go well with the corresponding audio. Conversely, the main point of concern seems to be the quality of the texture map: estimated facial features for our agents remain far too blurry, which prevents convincing gaze behavior and dynamic facial expressions. This is to be improved in future work, based on the availability of new models with better texture estimation (e.g. [1]).

5 CONCLUSION

In this paper, we presented a novel method for automatically generating virtual agents from 360° video for interactive virtual reality. Moreover, we demonstrated that users' perception of agents' responsiveness can be enhanced by implementing idle motion as well as reactive verbal and gaze behavior. This validates the idea that results from previous work on hand-made synthetic agents can also be applied to our automatic video-based agents.

REFERENCES

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. 2019. Learning to Reconstruct People in Clothing from a Single RGB Camera. (June 2019). https://arxiv.org/abs/1903.05885

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8387–8397. https://doi.org/10.1109/cvpr.2018.00875

[3] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2003. Interpersonal Distance in Immersive Virtual Environments. *Personality and Social Psychology Bulletin* 29, 7 (July 2003), 819–833. https://doi.org/10.1177/0146167203029007002

[4] Andrea Bönsch, Sina Radke, Heiko Overath, Laura M. Asché, Jonathan Wendt, Tom Vierjahn, Ute Habel, and Torsten W. Kuhlen. 2018. Social VR: How Personal Space is Affected by Virtual Agents' Emotions. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 199–206. https://doi.org/10.1109/vr.2018.8446480

[5] Grégoire Dupont de Dinechin and Alexis Paljic. 2018. Cinematic Virtual Reality With Motion Parallax From a Single Monoscopic Omnidirectional Image. In *Digital Heritage*. IEEE. https://hal-mines-paristech.archives-ouvertes.fr/hal-01915197

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2334–2343. https://doi.org/10.1109/iccv.2017.256

[7] Maia Garau, Mel Slater, David-Paul Pertaub, and Sharif Razzaque. 2005. The Responses of People to Virtual Humans in an Immersive Virtual Environment. *Presence: Teleoperators & Virtual Environments* 14, 1 (Feb. 2005), 104–116. https://doi.org/10.1162/1054746053890242

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7297–7306. https://doi.org/10.1109/cvpr.2018.00762

[9] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7122–7131. https://doi.org/10.1109/cvpr.2018.00744

[10] Kangsoo Kim, Arjun Nagendran, Jeremy Bailenson, and Greg Welch. 2015. Expectancy Violations Related to a Virtual Human's Joint Gaze Behavior in Real-Virtual Human Interactions. In *Proceedings of the International Conference on Computer Animation and Social Agents (CASA)*.

[11] Jan Kolkmeier, Jered Vroon, and Dirk Heylen. 2016. Interacting with Virtual Agents in Shared Space: Single and Joint Effects of Gaze and Proxemics. In *International Conference on Intelligent Virtual Agents (IVA)*. Springer International Publishing, 1–14. https://doi.org/10.1007/978-3-319-47665-0_1

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics* 34, 6 (Oct. 2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

[13] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. In *Conference on Human Factors in Computing Systems (SIGCHI)*. ACM Press, 143–146. https://doi.org/10.1145/1978942.1978963

[14] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human Performance Capture from Monocular Video. *ACM Transactions on Graphics* 37, 2 (May 2018), 1–15. https://doi.org/10.1145/3181973

[15] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. 2018. OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, 453–471. https://doi.org/10.1007/978-3-030-01231-1_28