Automatic 3D Avatar Generation from a Single RBG Frontal Image

Alejandro Beacco*
EventLab, Universitat de Barcelona

Jaume Gallego †
EventLab, Universitat de Barcelona

Mel Slater‡
EventLab, Universitat de Barcelona

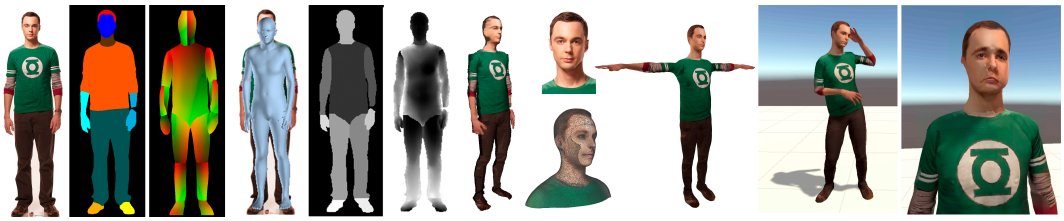

Figure 1: Our complete pipeline runs several deep learning techniques over a frontal RGB image of a single person and combines them to automatically retrieve a fully usable and recognizable 3D animated avatar.

ABSTRACT

We present a complete automatic framework to obtain a 3D avatar reconstruction from just a frontal RGB picture of the person to reconstruct. Our proposed workflow initially retrieves pose, shape and semantic information from the unconstrained input picture by using several existing deep learning methods. All that information is then combined into a skeleton and a 3D skinned and textured mesh which conform the final avatar. Since the head and face reconstruction is a central process to achieve a correct and realistic character modeling, we use an external head reconstruction method to properly adapt our final mesh to a recognisable head of the person. Our pipeline enables the recreation of 3D avatars focusing on three main aspects: automation (no input is required apart from a frontal image), recognition (the original subject can be identified) and usability (the obtained avatar is ready to use in any 3D application).

Index Terms: 3D character reconstruction—3D reconstruction from in the wild pictures—Animatable 3D reconstruction—Avatars

1 INTRODUCTION

Due to the rise of the metaverse, virtual reality (VR) is closer to being able to contain, among other things, a faithful reflection of reality. A critical aspect we will want to quickly recreate is our own avatar. Whether controlled by ourselves or by some artificial intelligence, we will want to have a true 3D representation in the virtual world. Since nowadays our smartphones come with RGB cameras, taking a frontal picture of a present subject with a specific pose should not be a problem anymore. But to recreate events from the past, like rock concerts [4], we might need to recreate avatars of others, famous personalities, distant friends, or long gone ones, from which obtaining new pictures of them in a specific pose will not be so straightforward. In this poster we address the challenge of automatically reconstructing a realistic 3D avatar of a person appearing with an unconstrained pose in a single RGB frontal image.

In recent years, computer vision and deep learning researchers have achieved incredible progress in the analysis and understanding of humans appearing in images. Once a human is segmented in the unconstrained input picture, even with some semantic information [13], we can roughly group the methods by the different human aspects they try to retrieve: shape, texture, and pose. Shape retrieval has mostly been solved by obtaining the shape parameters of a parametric human model that can approximate any human shape, like the widely used Skinned Multi-Person Linear Model (SMPL) [18] or the newest SMPL-X [19]. Although these methods only obtain a rough approximation of a naked human body, others like Tex2Shape [1] are able to retrieve displacement and normal maps to apply over the resulting mesh for a more accurate surface recovery of a dressed human. Shapes can also be recovered as implicit surface functions [21, 22], or by adapting the fitted SMPL model to the original input silhouette [3, 23]. Some of these last methods are also able to obtain a texture. The input image can be projected as a frontal texture, while the back part can be obtained by mirroring it and applying some post-processing [3, 23], or inferred by some neural network [21]. Pose detection has been addressed either by extracting joint positions in 2D [7, 10], by retrieving 3D rotations as parameters of the SMPL model [5, 15–17], or by establishing correspondences between human body parts [14]. More recent approaches are focusing on improving the results on challenging details such as face expressions and hand poses [11, 19].

Moreover, the head and the face of a person are critical in order to properly recognize the reconstructed subject. Thus, reconstructing a textured face or head from just a face picture is a more specific challenge that has also been widely explored [8,12]. But the obtained face or head is never attached to a reconstructed body in these methods. Actually, we found no complete method to obtain a full rigged avatar ready to use with all the desirable features. But we have all the necessary pieces to put everything together. In this work, we preprocess an input 2D RGB frontal picture with no specific pose requirements using some of the referenced methods, and define a pipeline that integrates and combines all the outputs to automatically obtain a recognizable 3D avatar that is ready to use and animate in any 3D virtual application.

2 WORKFLOW

Our goal is to generate a 3D model from an unconstrained picture of a person, therefore no specific pose or scenario is required. The generated 3D model of the character must be completely animatable to produce new content for example, for VR.

Our workflow starts by processing our input picture through a bunch of existing techniques to obtain different sets of data: human semantic segmentation [13], pose [16] and shape [1] SMPL parameters and extra information about pose [6, 14]. Additionally, given the semantic data from [13] we can crop the face of the appearing character and use it as input of a face or head generation method, such as AvatarSDK [2], and obtain a high quality mesh of this important part to use later.

*e-mail: abeacco@ub.edu
†e-mail: jgallego@ub.edu
‡e-mail: melslater@ub.edu

Given the obtained SMPL pose and shape parameters, we use them to render depth, normal and skin weight maps of such an SMPL model from a front and a back view. Following [23] and [3] we warp these renders, reconstruct a frontal and back mesh that we stitch together, and project the input image as texture. Again, like in [3], we perform inpainting on occluded parts and we synthesize a back texture using semantic information from the output of [13]. Using the joint positions of the fitted SMPL model, and applying inverse linear blend skinning, we set our reconstructed character to the canonical T-Pose. In such a pose, we do again like in [3] and wrap the SMPL neutral model to match our 3D reconstruction [20], obtaining a smoother version of our character sharing the same texture UV space as the original SMPL model.

While wrapping the template to our temporal reconstruction, we project the color texture on it. Analogously we are able to project the semantic information and visibility information, indicating which color zones have been inpainted, synthesized or directly retrieved from the original picture. Having all this data in the shared common UV space of the SMPL model, we run some post-processing to fix problems, specially on the hands level. For example, the hands pixels are known and always the same, so we can retrieve from the semantic texture a color skin patch, and synthesize a hands texture covering all that area. We later add some texture details like nails and blend everything to the original projected texture using poisson blending [9].

In our next step we wrap the head of our reconstruction to the one obtained by AvatarSDK. To do so we render the head mesh and use OpenPose [6] on it to retrieve the common face key points, which are always matched with the same corresponding points of the SMPL model. After wrapping, our model has a reconstructed body attached to a high quality head mesh. Please not that this process could be done with the head or face mesh obtained by any alternative method.

We repeat the wrapping process, but this time we use a human template that includes facial blendshapes. Doing so we are able to transfer such blendshapes to our resulting model. Finally, since the SMPL template model was basically nude, we trained a neural network to detect if shoes were present in the input image. If so, we use an additional template model of shoes adapted to the SMPL template, and apply the corresponding lattice transformation such that they are fitted to our wrapped model. This way, we obtain a separate adapted mesh for shoes which are equally skinned. We believe that this process could be applied to any kind of wearing accessories that could be easily detected, such as eyeglasses or hats, and maybe on general clothes and hairs. This will allow us to obtain more realistic avatars that are not only an adapted nude mesh with clothes projected on it.

3 Conclusions

Our results show that we have achieved a complete pipeline to automatically retrieve 3D animated avatars from RGB frontal pictures. The main limitation is that we need around 30 minutes to generate one character. As future work we would like to optimize that, carry on a proper qualitative and quantitative evaluation, and conduct a user study to validate that avatars are recognizable and agreeable.

Acknowledgments

This work is funded by the European Research Council (ERC) Advanced Grant Moments in Time in Immersive Virtual Environments (MoTIVE) number 742989.

References

[1] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proc. Int. Conf. on Computer Vision*, pp. 2293–2303. IEEE, 2019.

[2] AvatarSDK. https://avatarsdk.com.

[3] A. Beacco, J. Gallego, and M. Slater. Automatic 3d character reconstruction from frontal and lateral monocular 2d rgb views. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2785–2789, 2020. doi: 10.1109/ICIP40778.2020.9191091

[4] A. Beacco, R. Oliva, C. Cabreira, J. Gallego, and M. Slater. Disturbance and plausibility in a virtual rock concert: A pilot study. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 538–545, 2021. doi: 10.1109/VR50410.2021.00078

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pp. 561–578. Springer, 2016.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. Computer Vision and Pattern Recognition*, pp. 7291–7299. IEEE, 2017.

[8] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set, 2020.

[9] J. Di Martino, G. Facciolo, and E. Meinhardt-Llopis. Poisson Image Editing. *Image Processing On Line*, 6:300–325, 2016. `https://doi.org/10.5201/ipol.2016.163`.

[10] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *Proc. Int. Conf. on Computer Vision*, pp. 2334–2343. IEEE, 2017.

[11] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021.

[12] Y. Feng, H. Feng, M. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4), jul 2021. doi: 10.1145/3450626.3459936

[13] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *European Conf. on Computer Vision*, pp. 770–785, 2018.

[14] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pp. 7297–7306. IEEE, 2018.

[15] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

[16] M. Kocabas, C. P. Huang, O. Hilliges, and M. J. Black. PARE: part attention regressor for 3d human body estimation. *CoRR*, abs/2104.08527, 2021.

[17] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pp. 6050–6059. IEEE, 2017.

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. vol. 34, pp. 248:1–248:16. ACM, 2015.

[19] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2019.

[20] Russian3DScanner. R3DS Wrap, https://www.russian3dscanner.com/.

[21] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, 2019.

[22] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020.

[23] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pp. 5908–5917. IEEE, 2019.