The Research on Applying Artificial Intelligence Technology to Virtual YouTuber

Si-han Xu
Faculty of Science and Engineering
University of Nottingham
Ningbo, China
chican@geekchi.com

2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI) | 978-1-6654-2532-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/RAAI52226.2021.9507778

***Abstract*—Nowadays, Virtual YouTuber with its unique characteristics has achieved high popularity in the world. But at the same time, the technology used by the Virtual YouTuber also has a lot of room to improve. For example, today's Live2D technology can't show all the details of characters and perform all kinds of actions just like 3D models, when the cost of 3D modeling of both characters and backgrounds is much higher than 2D. If artificial intelligence technology can be applied to Virtual YouTuber, it is possible to solve these problems. Therefore, this paper explores the possibility of applying artificial intelligence technology to Virtual YouTuber from three aspects: character, motion capture and environment. This paper first proposes to use PIFuHD to transform 2D characters into 3D, then discusses how to use a single camera to achieve motion capture, and finally studies the stylization of background animation. Although there is still a certain distance to realize real low-cost Virtual YouTuber, with the progress of technology, there is still a lot of space to apply artificial intelligence to this industry.**

Keywords—Artificial Intelligence, Virtual YouTuber, 3D model, Motion Capture, Computer Vision

I. Introduction

With the growth of animation industry in the market, the Virtual YouTuber (VTuber) has also been rapid development. Many VTubers show their commercial value in the live broadcast market, and have a lot of fans on many social platforms such as YouTube, Niconico and Bilibili. For example, Kizuna AI with the 3D image has won millions of fans around the world, and many 2D image VTubers also have high popularity. However, both 2D and 3D images have some problems. The details, expressions and actions of the 2D VTubers are less, while the 3D image needs a fine 3D model and expensive motion capture equipment. With the progress of artificial intelligence technology, these problems might be solved. This paper studied the following problems that can significantly reduce the operation cost of VTuber and make the virtual image more vivid by studying the existing technical achievements: to translate 2D images into 3D models, use the camera to realize high-precision motion capture and transform photos to animation style images. These are three independent technologies, but if they can be applied to the Virtual YouTuber together, we may get better results. At the same time, this paper will also analyze the existing problems and shortcomings in the application of existing technology in VTuber, and propose some solutions. It is also hoped that artificial intelligence technology can play a greater role in the field of Virtual YouTuber.

II. Converting 2D Images into 3D Models

3D images have better details and can show more actions (Figure 1), so as to show a more vivid and specific character and attract more fans. However, with the advantages of low cost and convenience, the number of Virtual YouTuber with 2D image is much larger than that with 3D image. Because of the high threshold of 3D model, only some Virtual YouTubers with high investment cost or high income can have 3D virtual image. Therefore, if 3D models can be generated from 2D images through artificial intelligence technology, the threshold of using 3D models can be reduced, so that more Virtual YouTuber can obtain 3D models at low cost.

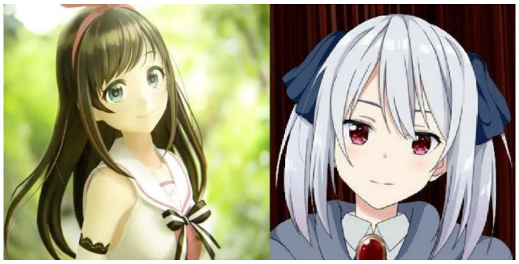

Fig. 1. Comparison of 3D and 2D characters

A. Existing Achievements

Compared with 2D images, 3D models have one more dimension, which provides more details. At the same time, compared with 2D image based on pixel features, 3D model is based on vector. Therefore, it is less difficult to flatten 3D model into 2D image, but it is very difficult to transform 2D image into 3D model. In the process of transforming 2D images into 3D models, it is necessary to expand the missing information in 2D images and integrate discrete pixels into vectors, which increases the difficulty of 2D translation into 3D models as well. By comparing different models (Figure 2), we find that PIFuHD has an excellent effect [1]. This is a new model proposed by Saito et al. of Facebook. As shown in Figure 3, the process of the model is as follows. First of all, they got the predicted back image of the characters by predicting the known positive images of the characters. After that, they used image translation to translate the front and back images into the normal map. Finally, they used the Multi-Level Pixel-Aligned Implicit Function to get the 3D model with the previous normal map.

Authorized licensed use limited to: ON Semiconductor Inc. Downloaded on October 16,2024 at 05:52:20 UTC from IEEE Xplore. Restrictions apply.

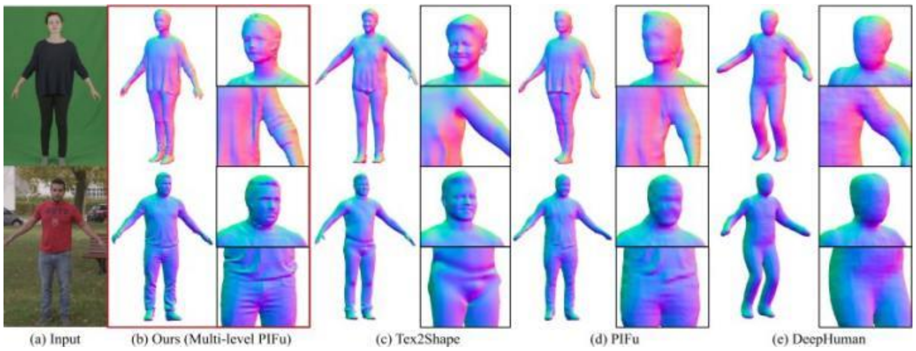

Fig. 2. Comparison of different models with PIFuHD [1]

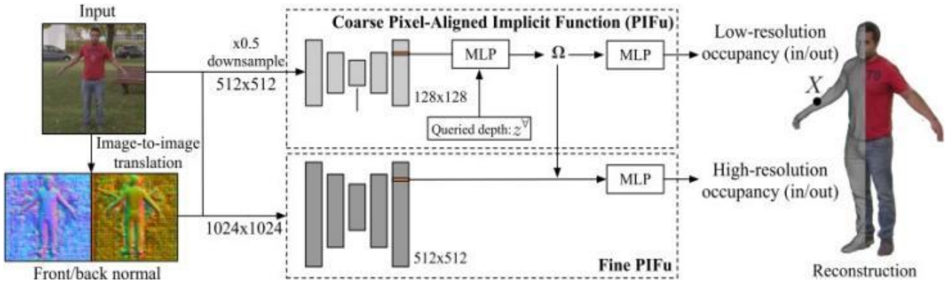
Input
x0.5
downsample
512x512
Coarse Pixel-Aligned Implicit Function (PIFu)
MLP
Ω
MLP
Low-resolution occupancy (in/out)
X
128x128
Queried depth: z^X
Image-to-image translation
1024x1024
MLP
High-resolution occupancy (in/out)
512x512
Fine PIFu
Front/back normal
Reconstruction

Fig. 3. The process of PIFuHD [1]

Pixel-Aligned Implicit Function is an efficient and spatially aligned representation of 3D surface proposed by Saito et al. with high storage efficiency [2]. It consists of convolution image, multi-layer perceptrons (MLPs) and an implicit function f. The implicit function f is defined as $f(F(x), z(X))$, where x is the 2D projection of X, $Z(X)$ represents the depth value at X, and $f(x)$ represents the image feature of x. The current surface s is defined as

$$s = f(F(x), z(X)) \qquad (1)$$

Through this function, we can realize the conversion between 2D image and 3D model, and PIFuHD is based on this function [1].

1) Front-to-Back Inference Predicting: Since the input 2D image has only the front, the back information about the character is completely missing. If we directly use the front image to generate a 3D model, it will produce a bad effect. In order to solve this problem, we need to predict the back of the characters first. Saito et al. provided a good method [1]. They first made a prediction of the 3D map and used it as a feature of pixel alignment. Then, these 3D maps are used to infer the geometric features in the space, so that the model can produce more details. Finally, pix2pixHD [3] is used to predict the normals in the space, so as to get reasonable back information.

2) Surface Sampling: The experimental data show that if the surface sampling is missing, the effect of the generated 3D model will be worse [1]. Therefore, it is necessary to sample the surface. Saito et al. used Gaussian perturbation of uniform volume sampling surface. This sampling scheme has better effect than other sampling schemes, because the sampling density near the high curvature region is higher. So this sampling method can have a high degree of restoration and more details.

3) Multi-Level Pixel-Aligned Implicit Function: Compared with PIFu, PIFuHD can be greatly improved because it uses two PIFu models. In the first model, low precision 3D models * are generated from 512 512 low precision * images. After that, 1024 1024 high-precision images are used to supplement the details of the model through the second model. The advantage of this method is that it can improve the efficiency without reducing the precision of the model. This model can also improve the efficiency of training, because only partial training is needed for high-precision details, that is, image content can be clipped. This also greatly reduces the memory requirement of the model.

B. Further Works

PIFuHD model has a good effect on the transformation of photo characters into 3D models, but it is not satisfactory in the transformation of animation characters into 3D models. The reason why it is difficult to transform animation pictures into 3D models is that the light and shadow of these pictures are far from reality. Therefore, it is difficult to get effective normal mapping through this model, so it is difficult to get a satisfactory 3D model. The other reason is that this model is based on the real picture training, and the adaptability of animation pictures is very poor. Such technology is still difficult to apply to the Virtual YouTuber industry. To solve the above problems, on the one hand, we need to replace the normal processing technology, on the other hand, we need a large number of animation pictures and matching 3D models for training. Although enough data can be obtained through public information, it is still very difficult to realize the function for predicting normal mapping form animation pictures. Only after solving this problem, it is possible to apply the technology of transforming 2D images into 3D models to the Virtual YouTuber.

Authorized licensed use limited to: ON Semiconductor Inc. Downloaded on October 16,2024 at 05:52:20 UTC from IEEE Xplore. Restrictions apply.

III. Motion Capture by Camera

3D image is mainly driven by motion capture technology, while traditional motion capture requires expensive equipment and cumbersome use methods, which greatly improves the threshold of using 3D model. At the same time, the Virtual YouTuber using live2d technology and 2D image must have at least one home camera, and the popularity of mobile phones also provides convenient video input devices. Therefore, if we can use artificial intelligence technology to achieve motion capture by the camera, it will have great convenience, so that more Virtual YouTuber can use 3D models.

A. Existing Achievements

Because the video camera can only provide a flat RGB picture, the information provided by the camera is insufficient to a large extent. At the same time, the video camera can return all the captured image information without processing, so there is a lot of interference information and invalid information in the image provided by the camera, such as the scene lighting, complex background and so on. All of these make it more difficult to capture motion with camera. So Mehta et al. proposed the VNect model [4]. This model can effectively recognize the person from the input signal of a single RGB camera, and bind the action skeleton for the person, so as to achieve motion capture, and can show good results. First, the model will track the bounding box to identify the bounding box of the human body. After that, they built a CNN model to convert the standard RGB image into a location-map. Then they identify the location map features through the temporary filter to get the skeleton information of the characters. Finally, the skeleton information in the picture is bound to the skeleton we need to achieve efficient motion capture. Figure 4 shows the flow of VNect in detail.

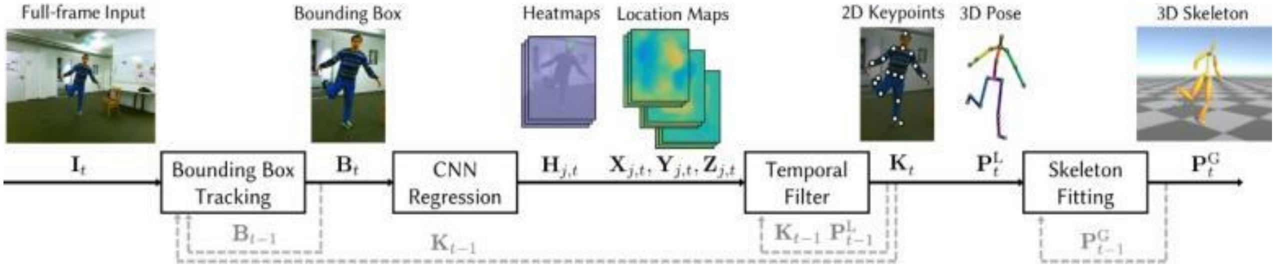

Fig. 4. The process of VNect [4]

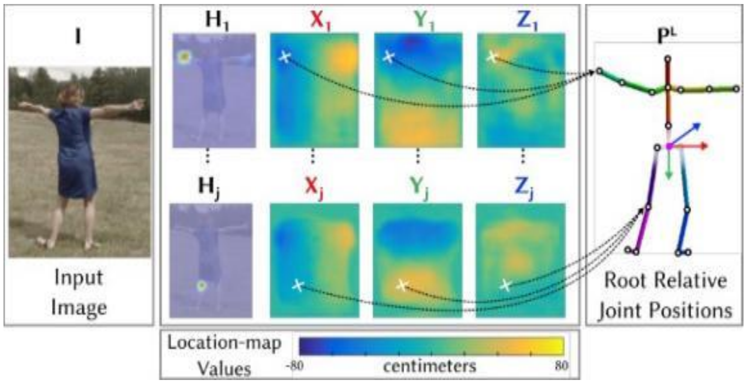

Fig. 5. Joint and location-map [4]

1) Bounding Box Tracking: As mentioned above, the video signal input by a single RGB camera has a lot of interference information and invalid information, but all we need is the skeleton information. Therefore, we first need to recognize the bounding box of the human body, so as to effectively filter out the information of environment, light and shadow. The previous technology can easily get the bounding box of the human body through image recognition technology. But because the input image is very large, it can only recognize each individual frame, and can not achieve the function of real-time recognition, so it can not achieve real-time motion capture. Mehta et al. moderately expanded the search scope of the next image through the position of the person in the previous image [4], thus reducing the image size of the tracking bounding box. When the input screen is reduced, the speed of the neural network can be greatly improved, so that the function of real-time recognition of the bounding box can be realized. This makes it possible to capture motion through a camera.

2) Predict location-map by CNN: Obviously, it is not enough to recognize the bounding box of the human body. We need to extract skeleton information from the human body to prepare for later motion capture. Previous techniques largely lack the connection between image and information, which leads to the prediction of bone information is not as good as expected. At the same time, the previous technology also has the problems of fixed input resolution ratio, complex process and low efficiency, so it is difficult to get bone information in real time. Mehta et al. proposed a new technology [4]. They generate three location maps through joints (Figure 5), and expand the two-dimensional information into three-dimensional information through location map, and store it. This technology breaks the limitation of image size and avoids

Authorized licensed use limited to: ON Semiconductor Inc. Downloaded on October 16,2024 at 05:52:20 UTC from IEEE Xplore. Restrictions apply.

clipping the image. This technology can get the information we want with high efficiency.

3) Skeleton Fitting: It is not enough to get the skeleton information of the character. In order to achieve motion capture, we need to bind the skeleton information of the character with the model skeleton. As long as the bone information is imported into 3D software, skeleton fitting can be realized. In this way, the complete motion capture is realized.

B. *Further Works*

Through the test, the technology of motion capture through a single camera has been very mature, and it is very easy to be applied to the virtual anchor of 3D model. But considering that most of the Virtual YouTuber will not expose the whole body in front of the camera, and this technology is for whole body motion capture, we still need to improve this technology. If only the upper body enters the camera, the technology will not be able to capture the motion to a certain extent, so we need to train the upper body motion capture separately, so that this technology can be better applied to the Virtual YouTuber.

IV. Background Animation

To enhance the sense of substitution of Virtual YouTuber, only virtual characters are not enough, but also need to create a virtual environment. Nowadays, the common methods of creating a virtual environment are background image and 3D modeling. If the background image is used, it will be very inconsistent with the 3D character, and the 3D model still has the problems of difficult production and high cost. Therefore, if we can directly animate the background of the camera input, we can achieve a better virtual environment at a lower cost, so as to reduce the cost of Virtual YouTuber.

A. *Existing Achievements*

Compared with the first two questions, it is relatively simple to convert background pictures into animation style. The existing image style transfer technology has been more mature, and the equipment requirements are not high. Now it is more convenient to transform the real map into an animation style. But if you want to achieve better results, there will still be some difficulty. Chen et al. provided a mature AnimeGAN model [5], which can achieve good results. This is a GAN model, through a large number of data training, can automatically realize the transformation of image style, so as to convert the real picture into an animation effect picture.

GAN is a neural network model proposed by Goodfellow et al., which is used to let computers learn and produce results by themselves, such as pictures, articles, etc. [6]. This network model proposes a generative scheme, which divides the model into two sides. One of them is used to generate results, and the other is used to check results. The producer is similar to the counterfeiter of the product and produces the counterfeited product. And the checking party is responsible for checking out these counterfeit products. In the continuous production process, through learning, the quality of products will be higher and higher. When the checking party can't tell the difference between the genuine and the fake, the Gan production result is considered successful.

1) Image style transfer: Image style transfer is a mature technology based on GAN. It can transform other pictures into the style of this picture by knowing the style of the picture. This technology is well matched with our destination, so we use image style migration technology to realize the animation function of real pictures.

2) AnimeGAN: AnimegGAN is an excellent GAN model for image style transfer [5]. This model is mainly composed of two convolutional networks (Figure 6), one for generation and the other for confrontation. The loss function of this convolutional neural network is as follows,

"Name" and "Paprika". The test data only contains realistic style images, because the model only needs a realistic image to generate animation style images. The resolution of the image is set to 256*256, which makes the training process highly efficient.

3) Model training and data processing: During the training, Chen et al. divided the data into two groups [5]. One for training and one for testing. The training data set contains realistic style images for processing, and animation style images for style standards, such as images in movies "Your improving the performance of Virtual YouTuber by applying artificial intelligence technology to virtual anchor, and this paper discusses how to apply artificial intelligence to Virtual YouTuber from three aspects of a character image, actor and environment. For actors, we propose to use the camera to achieve motion capture, so as to reduce the cost of equipment. For the environment, we propose to use image style transfer technology to transform the real environment into a virtual environment, so as to improve the sense of substitution of the virtual environment. But for character image, the technology of converting 2D image into 3D model is still not mature enough, and there is still a long way to go from the application of this technology. However, for the Virtual YouTuber who already has 3D models, the camera motion capture and image style migration technology can greatly reduce their operating costs and improve the virtual image representation. With the improvement of technology level, there is still a broad space to apply artificial intelligence technology to virtual anchor in the future.

B. *Further Works*

As shown in Figure 7, it is also a very mature technology to generate animation style pictures from real pictures. This technology can also be well applied to Virtual YouTuber. The only problem is that the style of the generated image is limited by the training data, and the style is less. This problem can be solved by training multiple AnimeGAN models and expanding the types of data styles. All in all, this technology has been able to be applied to Virtual YouTuber.

Authorized licensed use limited to: ON Semiconductor Inc. Downloaded on October 16,2024 at 05:52:20 UTC from IEEE Xplore. Restrictions apply.

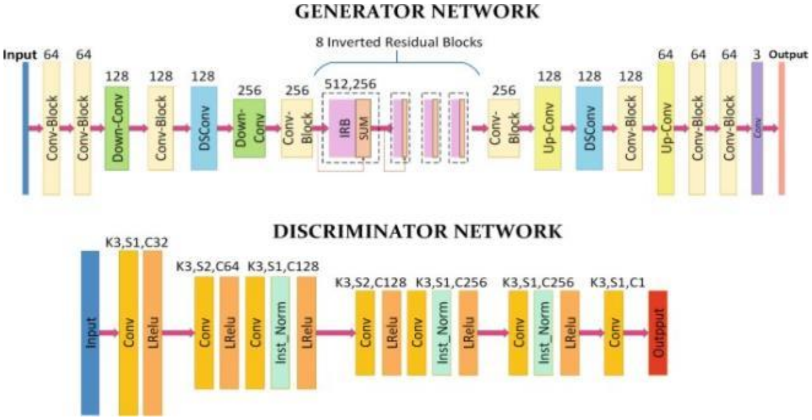

Fig. 6. The structure of AnimeGAN [5]

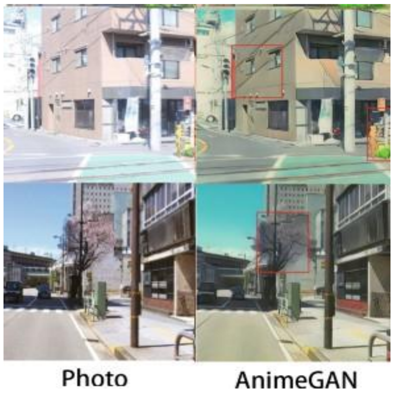

Fig. 7. The results of AnimeGAN [5]

$$L(G, D) = w_{adv}L_{adv}(G, D) + w_{con}L_{con}(G, D) + \\ w_{gra}L_{gra}(G, D) + w_{col}L_{col}(G, D) \qquad (2)$$

V. Conclusion

This paper discusses the possibility of reducing the cost, and through the experimental data, this model shows a good effect, and can successfully convert real pictures into animation style. At the same time, because this model has fewer layers and high efficiency, it can transform the image style immediately.

Considering that Virtual YouTube still has a huge market and development prospect in the future, we need to do more research. This paper puts forward the following solutions to the problems mentioned above. For the poor result of 2D to 3D, we need a database based on animation pictures and models. At the same time, because of the lack of light and shadow in animation pictures, we should study the function which is more suitable for the features of plane animation pictures. For single camera motion capture technology, just like most Virtual YouTube only show the upper body, we need to train the model with the data of the upper body of the characters. For animation style conversion technology, we need to expand the current data set, so that this technology can achieve more style conversion, so as to achieve the matching of environment and character painting style. These problems can be solved, and the future research direction of this application should focus on these aspects.

Acknowledgment

The reason why this paper can be published is largely due to the guidance of my teacher and instructor. Therefore, first of all, I would like to express my deepest thanks to them. Secondly, I also want to thank my classmates and family. Without their encouragement, I can’t finish this article either. Finally, I would like to thank all the people who have helped me in the process of writing. Without them, I will encounter more difficulties in the process of writing.

References

[1] Shunsuke Saito et al. “PIFuHD: Multi-Level Pixel Aligned Implicit Function for High-Resolution 3D Hu- man Digitization”. In: arXiv e-prints, arXiv:2004.00452 (Apr. 2020), arXiv:2004.00452. arXiv: 2004 . 00452 [cs.CV].

[2] Shunsuke Saito et al. “PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitiza- tion”. In: arXiv e-prints, arXiv:1905.05172 (May 2019), arXiv:1905.05172. arXiv: 1905.05172 [cs.CV].

[3] Ting Chun Wang et al. “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs”. In: (2017).

[4] Dushyant Mehta et al. “VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”. In: arXiv e-prints, arXiv:1705.01583 (May 2017), arXiv:1705.01583. arXiv: 1705.01583 [cs.CV].

[5] Jie Chen, Gang Liu, and Xin Chen. “AnimeGAN: A Novel Lightweight GAN for Photo Animation”. In: Ar- tificial Intelligence Algorithms and Applications. Ed. by Kangshun Li et al. Singapore: Springer Singapore, 2020, pp. 242–256. ISBN: 978-981-15-5577-0.

[6] Ian J. Goodfellow et al. “Generative Adversarial Net- works”. In: arXiv e-prints, arXiv:1406.2661 (June 2014), arXiv:1406.2661. arXiv: 1406.2661 [stat.ML].

Authorized licensed use limited to: ON Semiconductor Inc. Downloaded on October 16,2024 at 05:52:20 UTC from IEEE Xplore. Restrictions apply.