Using Stereo Camera System to Realize Realistic Video Avatar in Virtual Environment

CHEN Ling TIAN Mei-Hong CHEN Gen-Cai CHEN Chun
Department of Computer Science and Engineering, Zhejiang University, Hangzhou 310028, China
E-mail: chl6927@163.net

Abstract

With the appearance of different kinds of virtual environment systems, the technology of realistic video avatar becomes a hotspot of research. Our target is to create realistic video avatar with less hardware investment and limited computation. In this paper, we introduce a method that combines single stereo camera and 3D head model to engender realistic video avatar. To engender video avatar's most important part – face part, we use the 2D image and depth map, which is generated by the stereo camera. To engender the other parts of 3D head model, we map the head's texture to the 3D head model.

1. Introduction

With the development of computer technology and its related graphics and network technology, various virtual environments appear. The typical systems of these are as following: NPSNET [1] developed by the computer department of American navy postgraduate academy, DIVE [2] developed by Swedish computer science graduate school, MASSIVE [3] developed by British Nottingham University, and NICE [4] developed by Chicago Illinois University.

One research aspect of virtual environment is the realistic video avatar, where one or several cameras are used to generate a 3D avatar for one's head or body, which appears in the virtual environment instead of the people in the real world. The expressions or postures are represented by the avatar in the virtual environment. At present, the commonly used technologies of realistic video avatar are as following:

1. By using several cameras simultaneously to generate 3D avatar for head or body [5]. Namely, a group of 2D image captured by the cameras distributed around the people is used to generate people's 3D avatar. The advantage of this method is the high quality of presence for avatar, but the investment of hardware is large and it requires huge computation in generating avatars.
2. By shifting in several stereo cameras to generate 3D avatar for head or body [6]. Namely, we use 2D image and depth map which is engendered by the stereo camera to generate people's 3D avatar, and then according to different observers' viewpoint, shift the stereo cameras distributed around the people in order to generate 3D avatar for different people's viewpoints. The advantage of this method is the high quality of presence for avatar, but the investment of hardware for it is large and it requires huge computation in generating avatar.
3. Using single camera to generate 3D avatar for head or body [7]. Namely, using 2D image engendered by camera and people's 3D model, the people's 3D avatar is generated by texture-mapping method. The advantage of this method is that the investment of hardware is little and the computation load in generating the avatar is also light. But it has a poor quality of presence for avatar.

Generally speaking, the method 1 and 2 mentioned above are suitable for situation requiring high quality of presence of the avatar such as some virtual environments using immersive projection technology, e.g. CAVE [8] and CABIN [9]. The method 3 is suitable for situation not requiring high quality of presence for avatar such as desktop virtual environment using display, e.g., NetICE [10].

In this paper, we will introduce a method to generate realistic video avatar for head by combining single stereo camera with people's 3D head model. In this method, the key parts of face in head avatar are generated by using 2D image and depth map engendered by stereo camera. This method uses limited hardware and computation load to realize relatively realistic head video avatar. In the second part of the paper, we will introduce the stereo camera system. In the third part, we will describe the implementation of the system .In the last part, we will give the experimental system and some conclusions.

2. The stereo camera system

0-7803-7488-6/02/$17.00 © 2002 IEEE.

The working principle of stereo camera system is shown as Fig. 1. The whole system consists of the Left Camera, Right Camera and Top Camera. (x, y, z)is the coordinates of one point in the system. Pl, Pr, Pt are its corresponding points in the images generated by the Left Camera, Right Camera and Top Camera respectively. bh and bv are the horizontal and vertical translation from the Left Camera and Top Camera to Right Camera respectively, and f is the focus of the cameras. We can see from Fig. 1 that according to the values of Pl, Pr, Pt and combining these with bh, bv and f, we will get the coordinate value for point (x, y, z) by some triangular operations [11].

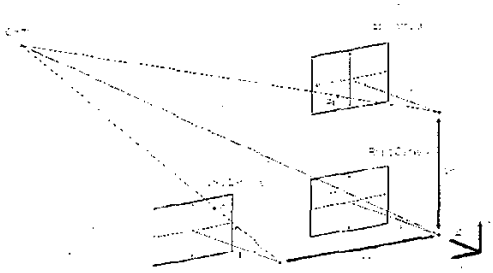

Fig. 1. The working principle of stereo camera system

In the implementation of the system, we use the stereo camera system made by Point Gray Research [12]. This system consists of two parts, one of it is the stereo camera module (digiclops) composed of three camera, as shown in Fig. 2, the other part is the software development package (triclops).

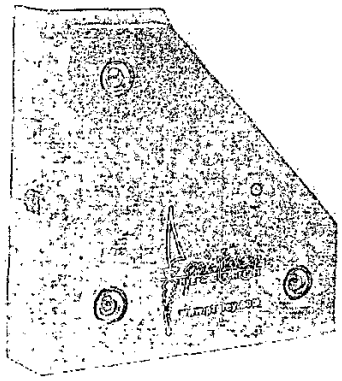

Fig. 2. Stereo camera system

When the system works, digiclops is connected with the host by IEEE1394 while the top level application can get the depth information for each pixel in the original captured image. Thus the system can give an original image for the captured object and its corresponding depth map showed as Fig. 3.

Fig. 3. Original image and depth map

3. Implementation of realistic video avatar

We introduce a method that combines single stereo camera and 3D head model to engender realistic video avatar. To engender video avatar's most important part – face part, we use the 2D image and depth map, which is engendered by the stereo camera. To engender the other parts of 3D head model, we map the head's texture to the 3D head model.

3.1. 3D modeling for head

Because we uses a method that combines 3D model with texture to engender the non-feature parts of head, a 3D Model for head is necessary for the system. Considering that the head features of each people are different, if we use a fixed 3D head model, then the video avatar generated will be distorted, not looking like the people filmed.

We use 3D model dynamic deformation [13] to resolve this problem. Namely, when the system begin to run, it use a predefined 3D head model, and in the process of running, the 3D head model is scaled in lateral x and vertical y according to the aspect ratio of head in the original image from video so that the resulting 3D head model has the same facial aspect ratio as the image from the video. Thus, the avatar engendered looks much more like the participant. The 3D head models before and after the deformation are shown in Fig. 4.

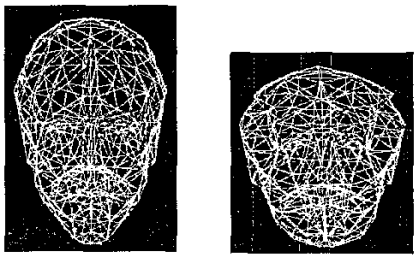

Fig. 4. 3D head model before and after deformation

Since in this system, the face of video avatar is generated through 2D image and depth map engendered by stereo camera, not through applying facial texture mapping to the 3D head model, therefore in the dynamic deformation of 3D model, we only regulate the aspect ratio of 3D head model and don't regulate the whole 3D head model in detail. If the facial avatar is generated through 3D head model and facial texture-mapping, then in the dynamic deformation of 3D model, we have to regulate the 3D head model in detail besides regulating the aspect ratio of 3D head model. We will set up feature points in facial image and 3D head model firstly, and then regulate the model's details by relativity matching for corresponding feature points.

3.2. Creation of facial avatar

Facial avatar is generated through 2D image and depth map engendered by stereo camera, namely, the system generates a triangular mesh from the depth map and then a facial avatar is generated by mapping each pixel in 2D image to facial triangular mesh.

The process for creation of avatar is as following:

1. Segmenting the facial depth map from the background. Since we need use only the facial depth map in the next step, facial depth map must be segmented from the background. In the system, threshold control method is used, namely we segment the pixels in background that their distance from the face are beyond the threshold and don't put them into the depth map. By this way, we resolve the problem of tracking and locating of user's head so that the user's head can move freely and don't need to be fixed in one place. By the method of segmentation, we can always locate the head.
2. We can determine the coordinates (xij, yij, zij) in the camera's coordinate system for each pixel pij in the depth map according to its depth data yij. Assume W is a horizontal pixel number, and H is a vertical pixel number for the depth image. If α is half of the horizontal viewing angle for camera and β is half of the vertical viewing angle, then xij and zij can be calculated by the following formulas:

$$xij = yij \times \tan\alpha \times \frac{(i - W/2)}{(W/2)}$$

$$zij = yij \times \tan\beta \times \frac{(H/2 - j)}{(H/2)}$$

3. A triangular mesh of the face is created from the point (xij, yij, zij) in the camera's coordinate system (see Fig. 5). In the process of creation, we finished it by connecting adjoining pixels. Because the shape of the triangular mesh generated by this process would be distorted unnaturally if the depth value include the noise, depth data must be dealt with before the creation, removing irregular data, so as to make the result more realistic and natural.

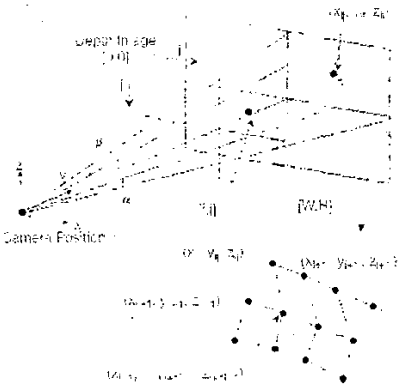

Fig. 5. Creation of facial avatar

4. Creation of realistic facial avatar. For each vertex of the triangular mesh generated in the top step, it has a texture coordinate (i/W, j/H) relating to the corresponding 2D image generated by the camera. The texture of the 2D image is assigned to the triangular mesh of this extracted surface model of the user, and so a realistic facial avatar is generated.

Using the original 2D image and depth map as shown in Fig. 3, adopting the process above, we will get the realistic facial avatar as shown as Fig. 6.

Fig. 6. Realistic facial avatar

3.3. Creation of head avatar

After getting the 3D model for head and facial avatar, then we can set about getting the avatar of the whole head. In order to get avatar of the whole head, firstly we have to get the textures for other parts of the head except the face, and then map these to the corresponding parts of the 3D model for head mentioned above.

Texture for head can be obtained in two methods now. One way is by using a laser scanner, the other way is by 3D to 2D mapping [14]. Under our system, a head texture map is generated for each user offline by combing several different views of the person [15], which in fact is the method of 3D to 2D mapping.

In creating avatar of head, we must have two things in mind. Firstly, because the size of the 3D head model and the texture is relatively unchanged, but the facial avatar is generated dynamically, namely, the size of the facial avatar is not fixed, the system must scale the facial avatar when it is running so that it can be embedded into the 3D head model fitfully. Secondly, because the head texture is static and the facial avatar is dynamically generated, the face may be in the different illuminate environment when it is generated, namely, the shininess of the facial avatar is not fixed, the system must manage the joint of facial avatar and the whole head avatar smooth when it is running, making the head avatar look more realistic and natural.

4. Conclusions

To prove the feasibility of the method that combines single stereo camera with 3D head model, we developed a experimental system of realistic video avatar for head. The experimental system has good performance in running and it has the following virtues compared o the previous method for implementing realistic video avatar:

- The video avatar generated has high quality of presence, and it can exhibit some small change in face.
- The hardware investment is comparatively little.
- The computation load is also comparatively little and it can run in PC.
- Provide the function of head tracking, which allows people's head not to be put into one fixed place.

We find several aspects should be improved in the future:

- Considering the system is used in virtual environment, in order to reduce the transmission amount, we can pick up and transmit change information only for the key parts of face, such as eyes and mouth.
- Regulating the 3D head model in detail according to the 2D image and the depth map, making the resulting avatar more realistic.
- In addition to smoothing the joints of the facial part of avatar and the whole avatar, we should dynamically regulating the luminance of the head texture according to the luminance of the facial avatar, making the video avatar obtained more realistic.

With the continually emergence and application of virtual environment, there are more and more people do some collaborative work depending on virtual environment. We have the confidence that the virtual environment will become more realistic and the communication between people will be more immersive, thereby virtual environment will become a good helper for people in collaborative work.

References

[1] Macedonia M R, Zyda M J. NPSNET: a network software architecture for large scale VE. Presence, 1994, 3(4): 265～287

[2] Gagsand O. Interactive multi-user VEs in the DVE system. IEEE Multimedia, 1996, 3(1): 30～39

[3] Greenhalgh C, Benford S. MASSIVE: a Distributed Virtual Reality System Incorporating Spatial Trading. DCS Conference, Vancouver, Canada, May 1995: 142～149

[4] Johnson, A Roussos, M Leigh, J Vasilakis, C Barnes, Moher T. The NICE Project: Learning Together in a Virtual World. In Proc of IEEE Virtual Reality Annual International Symposium (VRAIS) '98, Atlanta GA, 1998: 176～183

[5] T Kanade, P J Narayanan, P W Rander. Virtualized reality: Being mobile in a visual scene. In Proc of ICAT, 1995: 133～142

[6] Ken Tamagawa, Toshio Yamada, Tetsuro Ogi, Michitaka Hirose. Developing a 2.5-D video avatar. IEEE Signal Processing Magazine, 5, 2001: 35～42

[7] Wing Ho Leung, Tseng B L, Zon-Yin Shae, Hendriks F, Chen T. Realistic video avatar. In Proc of IEEE ICME, 2000: 631～634

[8] C Cruz-Neira, D Sandin, T Defante. Surround-screen projection-based virtual reality: The design and implementation of the CAVE. In Proc of ACM SIGGRAPH, 1993: 135～142

[9] M. Hirose, T Ogi, S Ishiwata, T Yamada. Development and evaluation of immersive multiscreen display 'CABIN' systems and computers in Japan. Scripta Technica, vol 30, no 1, 1999: 13～22

[10] Wing Ho Leung, Khalid Goudeaux, Sooksan Panichpapiboon, Sy-Bor Wang, Tsuhan Chen. Networked intelligent collaborative environment (NetICE). In Proc of IEEE ICME, 2000: 1645～1648

[11] Jong-Il Park*, Seiki Inoue. Acquisition of sharp depth map from multiple cameras. Signal Processing: Image Communication, 14, 1998: 7～19

[12] http://www.ptgrey.com/

[13] Peter Eisert, Bernd Girod. Analyzing Facial Expressions for Virtual Conferencing. IEEE Computer Graphics and Applications, 9, 1998: 70～78

[14] Won-Sook Lee, Marc Escher, Gael Sannier, Nadia Magnenat-Thalmann. MPEG-4 Compatible Faces from Orthogonal Photos. Computer Animation'99, 1999: 186～194

[15] Prem Kalra, Nadia Magnenat-Thalmann, Amaury Aubel, Daniel Thalmann. Real-Time Animation of Realistic Virtual Humans. IEEE Computer Graphics and Applications, 9, 1998: 42～56