3D video conversion system based on depth information extraction

Yingchun Zhang[1,a], Jianbo Huang[1,2], Siwen Duan[3]

[1]*Shanghai Film Academy, Shanghai University, Shanghai 200072 P.R. China*
[2]*Shanghai Engineering Research Center of Motion Picture Special Effects, Shanghai University, Shanghai 200072 P.R. China*

Abstract. 3D movies have received more and more attention in recent years. However, the investment in making 3D movies is high and difficult, which restricts its development. And there are many existing 2D movie resources, and how to convert it into 3D movies is also a problem. Therefore, this paper proposes a 3D video conversion system based on depth information extraction. The system consists of four parts: segmentation of movie video frame sequences, extraction of frame image depth information, generation of virtual multi-viewpoint and synthesis of 3D video. The system can effectively extract the depth information of the movie and by it finally convert a 2D movie into a 3D movie.

1 Introduction

In recent years, 3D movies have developed rapidly. The release of "Avatar" has made 3D movies come alive. In 2009, known as the "New Year of 3D Movies", there have been a lot of popular movies which also have excellent reputations. Although in the theater, 2D movies can also create an environment of experience with the plot, and bring the viewer corresponding touch. However, 3D movies can create a high-level content to help viewer experience the movie world, bring the immersive "participation feeling" of the viewing effect, enhance the impact of the plot on the consumer's mind, and the consumption satisfaction effect is not the same compared to 2D movies. 3D movies use the principle of human binocular stereo vision. During the formation of stereo vision, the brain automatically superimposes the images transmitted from the left and right eyes to synthesize a stereoscopic image, which makes people have a clear sense of depth.

However, in the production of 3D movies, special equipments are required for shooting, production, screening, etc., which determines the difficulty in 3D film production [1], and also restricts its development. At the same time, there are abundant 2D movie resources on the market, and converting 2D video into 3D video provides us with new ideas for making 3D movies. The conversion of 2D video into 3D video is currently the main method of making 2D video to 3D video. There is a key step in this technology: depth estimation - that's mean, extraction of depth information. Depth information perception is the premise of human stereo vision. By extracting depth, you can understand the depth(z-axis) information of the two-dimensional image and determine the distance of the object from the spatial depth [2]. The following steps is dependent on the depth information.

Based on these, a 3D video conversion system based on depth information extraction is proposed. The system consists of four parts: segmentation of movie video frame sequences, extraction of frame image depth information, generation of virtual multi-viewpoint and synthesis of 3D video. The system can effectively extract the depth information of a film, and through the above steps, converting into a 3D movie. It has a great application value.

2 Related work

At present, there are two main ways to make 3D movies. The first one is the active visual mode, which directly captures the three-dimensional information in the real scene through a three-dimensional device (such as a depth sensor), but this method requires high price and suitable devices for shooting. The shooting scene requirements are also higher. The second method is passive visual mode, which extracts three-dimensional information by estimating depth information. The 3D conversion technology adopts the passive mode. For depth estimation, extensive researches have been conducted on stereoscopic images because the depth of a stereoscopic image can be obtained by calculating the disparity between corresponding matching points in the binocular image. Scharstein et al. [3] proposed a research and evaluation method for matching, aggregating and optimizing two binocular images; Konda et al. [4] trained an automatic encoder to predict the depth of a binocular image sequence. However, these approaches depend on the disparity of the binocular images and are not suitable for monocular images. For monocular images, there are few studies at present, because there are an infinite number of three-dimensional scenes corresponding to

[a] Yingchun Zhang: maakun@163.com

monocular images, and light, shadow, and occlusion are all important factors that influence the depth estimation of objects in the image. Karsch et al. [5] used the KNN transfer mechanism based on SIFT flow [6] to estimate the static background depth of a single image, which could better estimate the information of the moving foreground in the video through motion. For the movie market, 3D movies have prospered and brought entertainment to the audience. However, the difficulty of its production has made some filmmakers discouraged. At the same time, there are many existing 2D movie resources. If you can convert it into 3D movies, it will greatly enrich the 3D movie library.

Therefore, this paper proposes a 3D video conversion system based on depth information extraction. The system consists of four parts: segmentation of movie videox frame sequences, extraction of frame image depth information, rendering of virtual multi-viewpoint and synthesis of 3D video. The system can effectively extract the depth information of the movie and convert it into a 3D movie, which has great application value.

3 3D video conversion system

3.1 Overview

The system consists of four parts: the segmentation of the movie video frame sequences, the extraction of the frame image depth information, the mapping of the virtual viewpoint and the synthesis of the 3D video. The flow is shown in Figure 1. First, the video is read by OpenCV, the image frame is extracted from it, and the extracted frame sequence is saved into the data set as the input image of the depth estimation. Then, the depth information is extracted from the frame image by using the deep learning neural network. Generating a virtual viewpoint by using the generated depth map and the original image, and then combine the virtual viewpoint map into a 3D movie. We present an interface to show the playback of the movie and preview the generated 3D movie.

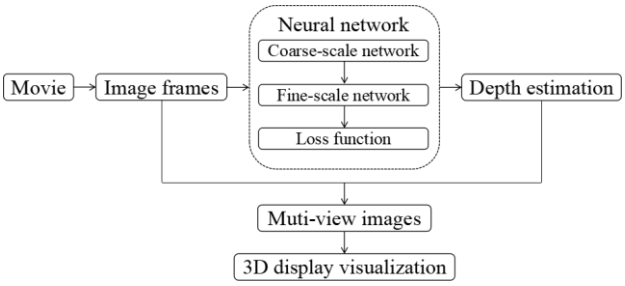

Figure 1. The flowchart of the system

3.2 Image frames acquirement

Video frame images extraction using OpenCV. The video is composed of frames one by one. Due to the visual residual characteristics of the human eye, when the image playback speed is higher than 16, the person thinks that the picture is coherent. Therefore, when processing an input movie, the video can be divided into frame images for processing. OpenCV is a cross-platform computer vision library based on BSD license (open source) distribution that runs on a variety of operating systems and implements many common algorithms for image processing and computer vision. The extracted frame image is shown in Figure 2.

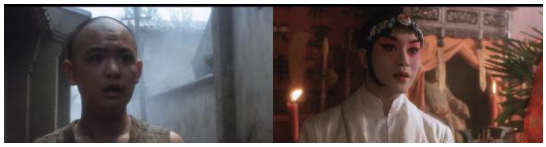

Figure 2. Image frames of movie

3.3 Depth information extraction

This paper cites David Eigen of New York University's bilayer neural network based monocular image depth estimation method [7] to extract depth information. Convolutional Neural Network is a research hotspot in the field of image processing [8]. It is a deep feedforward artificial neural network. Its artificial neural network can cover a surrounding unit in a corresponding part, mainly for processing image data. Compared with other neural networks, CNN can use the pooling layer to reduce the dimension of the feature map, and its weight sharing network can reduce the number of parameters in the model and reduce the complexity of the model. The basic structure of the CNN consists of an input layer, a convolution layer, a pooling layer, a fully connected layer, and an output layer, as shown in Figure 3. At present, there are many open source neural network learning systems. This paper uses the Tensorflow artificial intelligence learning system. TensorFlow is an open source software library for data calculation using data flow graphs for machine learning and deep neural networks, but the versatility of this system makes it widely used in other computing fields.

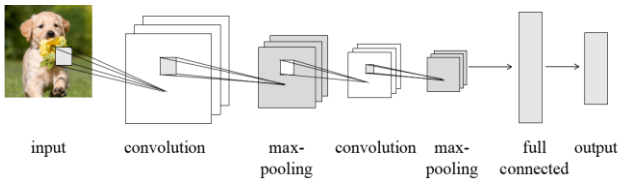

Figure 3. Convolutional Neural Network architecture

The model applied to the method of extracting depth information consists of two learning networks, as shown in Figure 4, which are divided into two parts: global coarse estimate and local refine estimate. The global coarse estimation network first predicts the depth of the scene in the global scope, and then refines the local area through the fine network. Both networks are applied to the input image, but at the same time, the output of the coarse estimation network is transmitted as an additional first layer image feature to the refine estimate network.

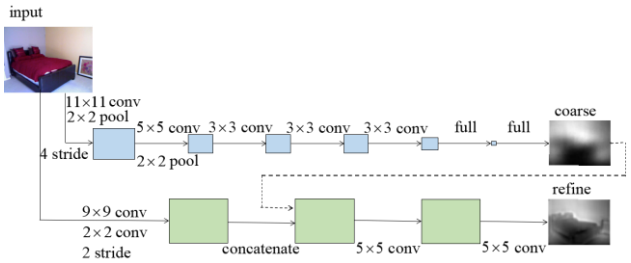

Figure 4. Model architecture

3.3.1 Network prediction

The overall coarse estimate of the CNN network is to predict the overall depth information of the image, including five feature extraction layers consisting of a convolutional layer and a pooling layer, and two fully connected layers. The final output image resolution becomes a quarter of original image. After using the global coarse estimation to predict the coarse depth information, we can obtain a depth map, but the map is fuzzy, and there is basically no edge information, so the edge of the object needs to be obtained through the second layer refined network. The refined network adds the output of the coarse network to the second layer as a feature map.

3.3.2 Scale-Invariant loss

Scale-invariant errors are used to measure the relationship between each point in the scene. For the predicted depth map y and the ground truth depth map $y*$, both graphs have n pixels indexed by i, and the scale-invariant mean square error is defined as Equation 1.

$$D(y,y*)=\frac{1}{n}\sum_{i=1}^{n}(\log y_i-\log y_i^*+\alpha(y,y^*))^2 \quad (1)$$

α is calculated by $\alpha(y,y*)=\frac{1}{n}\sum_i(\log y*_i-\log y_i)$

Using the scale invariant error as the loss function, and the calculation method is defined as Equation 3.

$$L(y,y*)=\frac{1}{n}\sum_i d_i^2-\frac{\lambda}{n^2}(\sum_i d_i)^2 \quad (3)$$

where, $d_i=\log y_i-\log y_i^*$ and $\lambda\in[0,1]$. Note that the output of the network is $\log y$, that's mean, the last linear layer predicts the logarithm of depth. In training, most target depth maps are subject to errors, especially at the edges of the object. Therefore, we only calculate the loss function of the effective point.

3.3.3 Model training

This method trains the model on the original versions of NYU Depth V2 and KITTI. The NYU dataset [9] consists of 464 indoor scenes captured by Microsoft Kinect cameras, trained using 249 scenes, and the training set consisting of 120,000 images. KITTI consists of outdoor scenes. These images are captured by car camera and depth sensor. 28 scenes of "City", "Residential" and "Road" are selected for training. The training set consists of 20,000 images.

3.3.4 The output of depth map

The depth information predicted by the neural network is effectively extracted. As shown in Figure 5, the depth estimation is a binary image, and the value of each pixel represents the estimated distance between the camera and the surface of the object. The brighter it is, the farther it is from the camera. However, deep learning prediction generates depth maps with instability, sometimes not obvious.

Figure 5. Original image and depth estimation map

3.4 Multi-viewpoint generation

When people are watching scenes from different angles, different viewpoints are generated. The difference in viewpoint causes the human eye to perceive the position of the object in space, far and near from itself. Depth map based rendering (DIBR) technology [10] can use the original image and depth estimation image to form different viewpoints. The main idea of DIBR algorithm is to revert the original image and depth map to the real scene according to the camera parameters, and then according to the camera parameters of the virtual viewpoint to inversely transformed to other viewpoints to form a multi-view image. However, in practice, camera parameters have often been lost, and a none-hole filled DIBR technique [11] provides a new approach.

3.5 Synthesis of 3D video

The data of the obtained virtual viewpoint image is arranged from left to right according to the viewpoint, sequentially reads data, and fuses the viewpoint data of the same frame image to generate a played 3D video.

3.6 Interface design

Interface design uses Qt Creator5.10.1 software. Qt is a cross-platform C++ graphical interface application development framework. The display effect is shown in Figure 6. The left side of the interface is the input movie. Click the button to import the movie and display the movie. On the right is the preview of generated 3D movie. If you want to observe the generated effect, click the button and watch the video. The button at the bottom is to generate 3D movie, click it and then a 3D movie is produced.

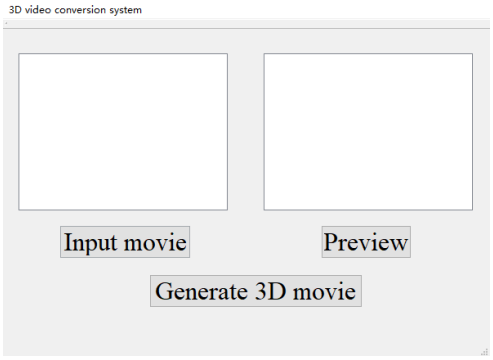

Figure 6. Interface

4 Conclusion

This paper proposes a 3D video conversion system based on depth information extraction. The system consists of four parts: segmentation of movie video frame sequence, extraction of frame image depth information, rendering of virtual viewpoint and synthesis of 3D video. The system can effectively extract the depth information of the movie and convert the 2D movie into a 3D movie through other key steps, which has great application value.

References

1. H. Shao, The Opportunities and Challenges Faced by the Development of 3D Films in China, Contemp. Cinema **6**, 175-178 (2017)
2. L. Zhu, Deep vision principle and stereo image characteristics of the human eye, J. B. Film. Aca **4**, 130-137 (2016)
3. D. Scharstein and R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision*, 131-140 (2001).
4. K. Konda and R. Memisevic, Unsupervised learning of depth and motion, Comput. Sci. (2013)
5. K. Karsch, C. Liu, S. B. Kang, and N. England, Depth extraction from video using nonparametric sampling, TPAMI (2014)
6. C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman, Sift flow: dense correspondence across difference scenes, (2008)
7. D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, *Proceedings of the Advances in Neural Information Processing Systems*, 2366-2374 (2014)
8. A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classfication with deep convlutional neural networks, *International Conference on Neural Information Processing Systems*, 2643-2651 (2013)
9. F. H. Sinz, J. Q. Candela, G. H. Bakır, C. E, Rasmussen, and M. O. Franz, Learning depth from stereo, 45-252 (2004)
10. Z. W. Liu, P. An, Z. Y. Zhang, Depth Image Based Rendering, 466-471 (2007)
11. Y. C. Fan, Y. C. Chen, S. Y. Chou, Vivid-DIBR Based 2D-3D Image Conversion System for 3D Display, J. Display. Technol **10**, 892-898 (2014)