3D Multi-View Stereoscopic Display and Its Key Technologies*

ZHANG Zhao-yang[1,2], *AN Ping*[1,2], *LIU Su-xing*[1,2]

(*1. School of Communication and Information Engineering, Shanghai University, Shanghai* 200072, *China*;
2. Key Laboratory of Advanced Displays and System Application, Ministry of Education, Shanghai 200072, *China*)

Abstract: As the next generation video display technique after 2D display based on DTV/HDTV, three-dimensional (3D) multi-view stereoscopic display has been one of the most popular research issues in the world. And for building a multi-view stereoscopic display system, related key technologies are detailed, which includes: Light field representation model and light field capturing system, high efficiency multi-view video coding and transmission method compatible with current video standard, high efficiency rendering method for arbitrary position view at the decoder, 3D display technologies and multi-view autostereoscopic display. Focusing on the key technologies above, the latest international development trends and existing problems is analyzed. Meanwhile a solution for implementing a 3D video processing system based on interactive auto-stereoscopic display is proposed.

Key words: stereoscopic display; light filed acquisition; multi-view video coding; view rendering
EEACC: **7260**

3D 多视点立体显示及其关键技术*

张兆杨[1,2],安 平[1,2],刘苏醒[1,2]

(1.上海大学通信与信息工程学院,上海 200072;
2.新型显示技术及应用集成教育部重点实验室,上海 200072)

摘 要:作为基于DTV/HDTV的二维(2D)显示之后的下一代视频显示技术,三维(3D)多视点立体显示已成为国际上的研究热点之一。为建立多视点立体显示系统,阐述了相关的关键技术,包括:光场表示模型和光场获取系统、高效的与现行视频标准兼容的多视点编码和传输方法、解码端任意位置视点的高效绘制方法、3D显示技术以及多视点自由立体显示。针对上述关键技术,分析了当前国际上的发展趋势及存在的问题,同时提出了一种基于交互式自由立体显示的3D视频处理系统的解决方案。

关键词:立体显示;光场获取;多视点视频编码;视点绘制
中图分类号:**TP399.41**　　**文献标识码**:**A**　　**文章编号**:**1005-9490(2008)01-0302-06**

Three-dimensional (3D) stereoscopic display can provides more natural and immersive visual effect which wins tremendous attention. As is known to all, two of the important cues for human to gain three-dimensional information are binocular parallax and motion parallax. Binocular parallax refers to seeing a different image of the same object with each eye, whereas motion parallax refers to seeing different images of an object when moving the head[1]. Today, although binocular stereopsis is still appreciated as the most impressive way of perceiving 3D scene, the classic two-view image format is no longer regarded to be the best way of representing 3D spatial information due to its limited view point. Therefore, the multi-view autostereoscopic display, which can provide binocular and

收稿日期:2007-04-30
基金项目:国家自然科学基金资助项目(60672052、60572127);上海市曙光计划项目资助(06SG43);上海市重点学科建设项目资助(T0102)

作者简介:张兆杨(1938-),男,教授,博士生导师,研究方向为图象及视频处理、编码与多媒体通信,zhyzhang@shu.edu.cn;
安 平(1968-),女,教授,研究方向为图象及视频处理、三维视觉,anping@shu.edu.cn;
刘苏醒(1980-),男,博士研究生,主要从事三维视频处理、编码与系统研究,liusuxing@sina.com.

motion parallax images for multiple observers from any viewpoint without special glasses, naturally is chosen to be the best solution for Three-dimensional (3D) stereoscopic display. To display the multi-view video images on LCD (Liquid Crystal Display) screen or thin film screen directly watched by man's bare eyes with strong immersive perception, following key technologies should be developed. ① Light field representation model and light field capturing system. ② High efficiency multi view video coding (MVC) and transmission method compatible with current video standard. ③ High efficiency rendering method for arbitrary position view at the decoder, and the comprehensive optimization strategy with the multi view coder. ④ Structure of the LCD screen for stereoscopic display and design of the optical characteristics of autostereoscopic display.

For key technologies above, this paper analyzes the latest international development trends, existing problems, and proposes a solution for implementing a 3D video processing system based on interactive auto-stereoscopic display, which consists of the light filed rendering, data format transformation, programming and compiling, coding and transmission, new view creation and interactive stereoscopic display. Light field rendering consists of three steps: capturing, compression and rendering. Since the distance and pose of cameras will have great impact on the rendering quality, camera array based capturing system is detailed. For MVC, spatial-temporal prediction structures and prediction tools are key factors influencing the compression ratio. New concepts and models should be developed to handle mass video data. Except for the scalability of signal-to-noise ratio, scalability for MVC also is measured by the scalability of temporal-spatial and scalability of view. In addition, for the aim of real time multi-view visual effect, efficient view rendering at decoder should fully exploiting and correcting the decoded information. Meanwhile efficient view prediction, interpolation, and fast post processing methods are needed. At last for stereoscopic display, thin film display technology of multi-view should focus on enhancing the stereoscopic effect in vertical direction compared to traditional stereoscopic display technology.

1 Capturing and Rendering

The interactivity in multi-view video application is realized by freely changing the viewpoint. Multi-view images are usually captured using a camera array, and the image-based rendering (IBR) is used to render the scene.

1.1 Camera Array System

A lot of camera arrays have been built for multiview imaging. Imaging methods of 4 cameras, 5 cameras and 16 cameras were proposed in succession. Several large arrays consisting of tens of cameras have been built, such as the Standford multi-camera array with 128 cameras, the MIT 64 distributed light field cameras and the CMU 3D room with 49 cameras. In the camera arrays, those with a small number of cameras can usually achieve real-time rendering. On the fly geometry reconstruction is widely adopted to compensate for lack of cameras, and the viewpoint is often limited. Large camera arrays, despite their increased viewpoint ranges, often have difficulty in achieving satisfactory rendering speed due to the large amount of data to be handled[2]. Another problem accompanying with the large camera arrays is the coding and transmission burden for the huge data. Therefore, there is certain relation between the camera number and the rendering performance. The balance of the two and the related 3D scene representation remains one of the hotspot research issues.

1.2 Representation and Rendering

How the light field data is representeded can have an impact on the rendering algorithms, as well as compression and transmission and hence the interactivity. An image-based representation implies using a dense camera setting. However, a relatively sparse camera setting would only give poor rendering results of virtual views but a broader field of view. Methods for 3D scene representation are often classified as a continuum in between two extremes shown as Fig. 1[3].

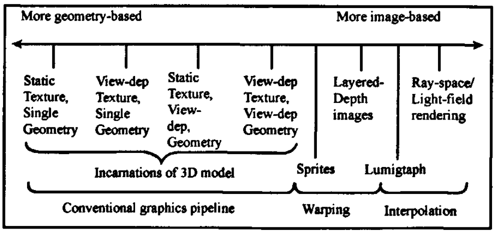
More geometry-based
More image-based
Static Texture, Single Geometry
View-dep Texture, Single Geometry
Static Texture, View-dep, Geometry
View-dep Texture, View-dep Geometry
Layered-Depth images
Ray-space/ Light-field rendering
Incarnations of 3D model
Sprites
Lumigtaph
Conventional graphics pipeline
Warping
Interpolation

Fig. 1　Categorization of scene representation

1.2.1 Representation based on total geometry information

The one extreme is represented by classical geometry-based modeling, which is usually described by 3D meshes. Real world objects are reconstructed using geometric 3D surfaces with an associated texture mapped onto them. More sophisticated attributes can be assigned as well. The main drawback of geometric is the high costs for content creation, and the system becomes even more complex when dynamic scene is created.

1.2.2 Representation Based on Total Information

The other extreme is given by scene representations of image based modeling without using any 3D geometry at all. Image-based representation mainly uses the ray space theory to render the virtual images from available real views by interpolation. Compared with geometric models that dominate the traditional 3D rendering pipelines, images are easier to obtain, simpler to handle and more realistic to render. It has attracted many researchers from different communities, including graphics, vision and signal processing. However, it needs large amount of images to achieve high-performance rendering. Typical image-based methods, such as light field rendering[4], panoramic configurations including concentric and cylindrical mosaics[5], do not make any use of geometry, but they either have to cope with an enormous complexity in terms of data acquisition or they execute simplifications restricting the level of interactivity. Two existing problems for the image-based representation are: ① capturing requires a tremendous effort and high quality acquisition of a dynamic Ray-Space is still a difficult task. ② a full dynamic Ray-Space results in an enormous data rate, therefore efficient compression is the second key technology to make a Ray-Space system feasible besides interpolation.

1.2.3 Representation Based on Partial Geometry Information

Between the two extremes there exits a continuum of methods that make more or less use of both approaches and combine the advantages in a particular manner. For examples, a Lumigraph use a similar representation as a light-field but adds a rough 3D model[6]. Other representations do not use explicit 3D models but depth or disparity maps such as DIBR (depth-image-based rendering)[7] and LDIBR (layered-depth-image-based rendering) methods[8]. Methods using view-dependent geometry and/or view dependent texture are closer to the geometry-based end of the spectrum[9]. Surface light-fields method and volumetric representations belong to this kind.

For multi-view capturing and rendering, we should do much work on following issues. ① Relationship between the camera array's size and the rendering quality. ② Optimal spatial layout of the camera array as well as the synchronization and connection manner. ③ High efficient methods for camera calibration and fast view rendering.

2 Multi-View Video Coding

2.1 MVC Background

Multi-view video coding consists of multiple views of the same scene in which there exists a high degree of correlation between the multiple views. Therefore, in addition to exploiting the temporal redundancy to achieve coding gains as in 2D video coding, spatial redundancy can also be exploited and achieved by performing spatial prediction across different views. The straight-forward solution for MVC would be to encode all the video signals independently using a current coding standard such as H.264/AVC. From the research results have been submitted to MPEG[10], it has been shown that specific MVC algorithms give significantly better results compared to the simple H.264/AVC simulcast solution. The basic idea in all of the submitted proposals is to exploit spatial and temporal redundancy for compression.

In MPEG document[11], some potential MVC techniques are described in detail, and document of reference [12] has given the related requirements on MVC. Accounting for these requirements, researchers have developed a number of multi-view coding schemes. For MVC, since spatial-temporal prediction structures and prediction tools are key factors influencing the compression ratio. New concepts and new models should be developed to suit to compressing great mass video data. For instance, besides measured by scalability of signal-to-noise ratio, scalability for MVC is also measured by scalability of temporal-spatial and scalability of view. On the other hand, efficient view rendering at decoder should satisfy the need of real time by fully exploiting and correcting the decoded information, efficient view prediction and interpolation, and fast post processing. The standardization is another problem. In MPEG-2, the multi-view profile can transmit two video channels. The disparity prediction and compensation is defined by using Temporal-Scalability, which can remove the inter-view redundancy. However, MPEG-2 does not support interactivity, and its compression efficiency is not high enough. In MPEG-4, the AFX (animation framework extension) gives definition of depth image and MAC (multiple auxiliary components) which can transmits depth or disparity data. However, these tools are not originally specified for 3D video that are not suitable for MVC. Therefore, developing high efficient MVC as well as its standardization attract high degree attention.

2.2 Related Work of MVC

MPEG document "N6909"[10] gives a survey of algorithms used for MVC. Several prediction structures of Group-of-GOP prediction, sequential view prediction and checkerboard decomposition prediction have been described. Several unique prediction and preprocessing tools also have been introduced. The prediction tools include illumination compensation, 2D direct mode, disparity/motion vector prediction and view interpolation.

Most systems compress the multi-view video off-line and focus on providing interactive decoding and display. Zitnick, et al. show that a combination of temporal and spatial encoding leads to good results. The Blue-C system converts the multi-view video into 3D "video fragments" that are then compressed and transmitted. However, all these current systems use a centralized processor for compression, which limits their scalability in the number of compressed views[1].

Another approach to MVC, promoted by the European ATTEST project[13], is to reduce the data to a single view with per-pixel depth map.

MVC based on Ray-Space is a kind of promising coding scheme. Ray-Space representation was first introduced by Fujii, et al. for Free-point TV system[14]. Scalable structure or hierarchical structure can also be introduced into Ray-Space to achieve more coding gain. Related work includes the 3D visual compression based on Ray-Space projection proposed by Stanford University[15]. Stanford University also proposed a Wyner-Ziv coding scheme of light-field for random access. The images are independently encoded by a Wyner-Ziv encoder. At the receiver, previously reconstructed images are used by the Wyner-Ziv decoder as side information to exploit similarities among images[16].

Multi-view video compression has mostly focused on static light-field[17]. There has been relatively little research on how to compress and transmit multi-view video of dynamic scene in real-time. However, the MERL of Cambridge presents a scalable system for real-time acquisition, transmission, and high-resolution 3D display of dynamic multiview TV content, which exploits a distributed coding architecture[1]. Yang, et al. also achieve a real-time distributed light field camera system[18].

For scalable MVC, in addition of MERL's scalable system, Yang, et al. proposed an MVC scheme based on wavelet, which can provide temporal, spatial, SNR as well as view scalabilities. This is the first time of wavelets used for MVC[19].

Another problem for MVC is the compatibility with current coding standard. Many researchers have noticed this problem and make their MVC schemes compatible with MPEG or H. 264/AVC.

Although, the above various MVC schemes have their unique characteristics, the coding efficiency still can be greatly improved in future work since the limited exploiting of the inter-view redundancy. Furthermore, MVC research towards multi-view video applications is encouraged.

3 3D Display

3D display technology had been developed about one hundred and fifty years ago. From 1850 to 1930, Brewster invented stereoscope and succeed in commercial. A stereoscopic named View-Master became more popular product from 1940 to 1950. Then, vectorgraph and earlier relief television had come into humans' life. Now, 3D display technology has been greatly improved and is in a flourish period along with the developing of computer technology.

3.1 3D Display Technology Classification

3D Display technology can be classified into several types shown as figure 2.

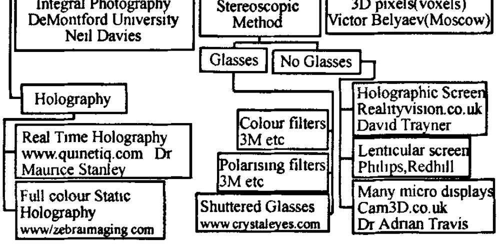
3D Displays
Integral Photography DeMontford University Neil Davies
Stereoscopic Method
3D pixels(voxels) Victor Belyaev(Moscow)
Glasses
No Glasses
Holography
Real Time Holography www.quinetiq.com Dr Maurice Stanley
Full colour Static Holography www/zebraimaging com
Colour filters 3M etc
Polarising filters 3M etc
Shuttered Glasses www crystaleyes.com
Holographic Screen Realityvision.co.uk David Trayner
Lenticular screen Philips,Redhill
Many micro displays Cam3D.co.uk Dr Adrian Travis

Fig. 2　Classification of 3D display technologies

(1) Holographic Display　Hologram was invented by Dennis Gabor in 1948. Holographic techniques were first applied to image display by

Leith and Upatnieks in 1962. The acquisition of holograms still demands carefully controlled physical processes and cannot be done in real-time. At least for the foreseeable future it is unlikely that holographic systems will be able to acquire, transmit, and display dynamic, natural scenes on large displays.

(2) Volumetric Display Volumetric displays use a medium to fill or scan a three-dimensional space and individually address and illuminate small volexs. Although there are some commercial application system, volumetric systems which can produce transparent images that do not provide a fully convincing three-dimensional experience. Furthermore, they cannot correctly reproduce the light-field of a natural scene because of their limited color reproduction and lack of occlusion. The design of large-size volumetric displays also poses some difficult obstacles.

(3) Parallax Displays Parallax displays emit spatial varying directional light. Much of the early 3D display research focused on improvements to Wheatstone. Earlier developed techniques include parallax stereogram and parallax panoramagrams that provide only horizontal parallax. Another parallax display technique is integral lens sheets which can be put on top of high-resolution LCDs. Integral photographys sacrifice significant spatial resolution in both dimensions to gain full parallax. Multi-projector lenticular display was invented in 1931 to improve the native resolution of display. Other research in parallax displays includes time-multiplexed and tracking-based. Today's commercial autostereoscopic displays use variations of parallax barriers or lenticular sheets placed on top of LCD or plasma screens. Parallax barrier generally reduce some of the brightness and sharpness of the image.

3.2 Multi-View Autostereoscopic 3D Display

Comparing with multi-view autostereoscopic 3D display, four cues are missing from 2D media, they are stereo parallax, movement parallax, accommodation and convergence. All 3D display technologies provide at least stereo parallax. Multi-view autostereoscopic provides both binocular and motion parallax for multiple observers, and autostereoscopic displays provide 3D perception without the need for special glassed or other headgear. Three technologies used in autostereoscopic display are spatial multiplex, multi-projector and time-sequential. In spatial multiplex technology, the resolution of a display device is split between the multiple views. Multi-projector technology makes a single projection display which is used for each view. A single fast display device is used for all views if time sequential technology is exploited. Drawing upon above three technologies, developers can make two different types of autostereoscopic displays: a two-view, head-tracked display for single-viewer systems or a multi-view display that supports multiple viewers. Detail description can be found in Ref. [20].

Compared to traditional stereoscopic display technology, thin film display technology of multi-view can focus on enhancing the stereoscopic effect in vertical direction. We look forward to the flat-panel multi-view autostereoscopic displays.

4 3D Video Processing System

We propose a solution for implementing a multi-view autostereoscopic display system with acquisition, MVC coding, view rendering at the decoder and interactive auto-stereoscopic display. We originally set a 4 ×4 camera arrays and an extension to more cameras will be achieved in the future. Figure 3 shows a schematic representation of it.

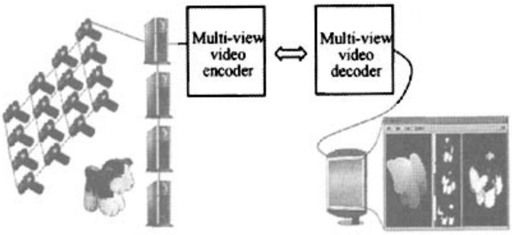

Fig. 3 A multi-view autostereoscopic display system

In the system, the key techniques should be explored to implement dynamically acquisition and rendering are summarized as follows.

(1) Acquisition We adopt a camera arrays both in line and in square layouts. The camera pose parameters including space between cameras and their direction should be determined, and the camera intrinsic parameters are computed by using a self-calibration based on a genetic algorithm. We exploit several DIBR algorithms to render the virtual viewpoint.

(2) Coding A H.264/AVC basic framework with joint estimation of motion and disparity is exploited to efficiently improve the inter-view prediction. A new scalable coding scheme will be explored to adapt the need of multi-view stereoscopic display. We also adopt histogram transformation

in piecewise linearity to normalize all view images as a preprocessing manner.

(3) Rendering at decoder　Aiming at real-time arbitrary view rendering, we focus on taking full advantages of the decoded information and view tracking varying with the view direction to implement the interactivity in processing and display. Besides that, efficient view prediction and interpolation are also crucial for real-time rendering.

(4) 3D display　This is a multi-view autostereoscopic display system by rear-projection on thin film screen using only one lenticular sheet with optical diffuser material. The projector projects the thin vertical stripes pictures onto the diffuser and each lenticular sheet acts as a light de-multiplexer and projects the view-dependent radiance back to the viewer. As a result, the stereoscopic effect is enhanced in vertical direction.

5 Conclusion

From multi-view acquisition, representation, coding, to autostereoscopic display, this paper outlines recent trends in the key technologies to implement a 3D multi-view stereoscopic display system. There is still much that we can do to improve the entire system, especially about dynamic light field data acquisition and compression, super wide-band transmission as well as multi-view autostereoscopic display with wider view angle and higher brightness.

Acknowledgments

This work is supported by Natural Science Foundation of China (grant No. 60672052 and 60572127), Shanghai Shu Guang Project (Grant No. 06SG43) and Shanghai leading Academic Discipline Project (Grant No. T0102).

References:

[1] Matusik W, ster H P, 3DTV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes[J]. ACM Transactions on Graphics, 2004, 23(3), 814-824.

[2] Zhang C, Chen Tsuhan , Multi-view Imaging: Capturing and Rendering Interactive Environments, Computer Vision for Interactive and Intelligent Environment[M]. 2005, 51-67.

[3] Smolic A, Kimata H, Vetro A, Development of MPEG Standards for 3D and Free Viewpoint Video, Communications[J]. Multimedia & Display Technologies, SPIE, Nov. 2005, Vol. 6014, 262-273.

[4] Levoy M, Hanrahan P, Light Field Rendering[J]. Proc. of ACM SIGGRAPH, August 1996, 31-42.

[5] Szeliski R, Shun H Y, Creating Full View Panoramic Image Mosaics and Texture-mapped Models[J]. Proc. ACM SIGGRAPH, August 1997, 251-258.

[6] Buehler C, et al. Unstructured Lumigraph Rendering[J]. Proceedings of SIGGRAPH, 2001, 425-432.

[7] Fehn C, A 3D-TV Approach Using Depth-Image-Based Rendering(DIBR)[C]//Proc. Visualization, Image and imaginf Processing, Benalmadena, Sept,2003, Spain,482-487.

[8] Shade J, Gortler S, He L W and Szeliski R, Layered Depth Images[C]//Proc. of SIGGRAPH'98, Jul. 1998, 231-242.

[9] Mueller K, et al. Reconstruction of a Dynamic Environment with Fully Calibrated Background for Traffic Scenes[J]. IEEE Trans. CSVT, 2005, 15(4),538-549.

[10] ISO/IEC JTC1/SC29/WG11, N6909, Survey of Algorithms used for Multi-view Video Coding[R]. Jan. 2005.

[11] ISO/IEC JTC1/SC29/WG11, N5878, Report on 3DAV Exploration[R]. July 2003.

[12] ISO/IEC JTC1/SC29/WG11, N5878, Requirements on Multi-view Video Coding v. 4[C]//July 2005.

[13] Fehn C, et al. . An evolutionary and Optimised Approach on 3D-TV[C]//Proceedings of International Broadcast Conference, 357-365.

[14] Fujii T, Kimoto T and Tanimoto M, Ray Space Coding for 3D Visual Communication[C]//PCS'96, 447-451.

[15] Ramanathan P and Gird B, Rate-Distortion Analysis Of Random Access For Compressed Light Fields[C]//2004 International Conference on Image Processing (ICIP),2463-2466.

[16] Aaron A, Ramanathan P, Girod B, Wyner-Ziv Coding of Light Field for Random Access[C]//IEEE 6th Workshop on Multimedia Signal Processing, 2004, 323-326.

[17] Magnor M, Ramanathan P and GIROD B, Multiview Coding for Image-Based Rendering Using 3-D Scene Geometry[J]. IEEE Trans. CSVT, 13(11), 1092-1106.

[18] Yang J C, Everett M, Buehler C and Mcmillan L, A Real-Time Distributed Light Field Camera[C]//Proc. of the 13th Eurographics Workshop on Rendering, 2002, 77- 86.

[19] Yang W, et al. , Scalable Multiview Video Coding Using Wavelet[C]//IEEE International Symposium on Circuits and System, May 2005, Vol 6, 6078-6081.

[20] Neil A. Dodgson, Autostereoscopic 3D Displays[C]//University of Cambridge Computer Laboratory, IEEE Computer Science, 2005, 31-36.