A 3D Display Parallel System: Light Field Re-rendering and Depth Sense Optimization

Renjing Pei [1,2,3], Kui Ma [1,2,3], Feiyue Wang[1,3]

1 State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, East Zhongguanchun Road, No. 95, Haidian District, Beijing.

2 School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing.

3 Parallel Optics Technology Innovation Center, Qingdao Academy of Intelligent Industries, Qingdao.

Abstract

The main benefit of 3D display over 2D display is the obvious ability to create a more lifelike character with high depth sense. However, the limitation of human eye's visual mechanism, unartful 3D scene structure design, or bad viewing condition always emerge a poor depth perception experience or even a physiological discomfort during the watching time, which are often sub-optimal for mass high-quality 3D display productions. To solve this problem, we propose a novel 3D display parallel system for depth sense optimization and it empirically guides how the light field should be re-rendered. Structurally, the parallel system consists of an artificial perception measurement system, a display evaluation model and a light field display rendering system, which includes the display calibration, scene capture, light field data process and display. Particularly, it systematically analyzes and models various factors affecting the depth sense that learned through the measurement system, like scene structure, objects' speeds in 3D video and so on. And those sense factors can be personally modified or increased according to the viewer's demands or technical improvement. Moreover, the light field could be real-time re-rendering, based on some image processing technology, optical flow analysis and object segmentation (or tracking) (especially the one-shot video segmentation). Theory and algorithms are developed and experimental validation results show a superior performance.

Author Keywords

3D Display Parallel System; Depth Sense; Light Field Re-rendering; 3D Video Processing; One-shot Segmentation.

1. Introduction

In recent years, the three-dimensional industries, such as 3D movies or 3D games, have driven a rapid development of the light field display products [1]. However, due to human visual mechanism, display content, display hardware, viewing conditions or other reasons, people cannot always experience a good depth perception when watching 3D display [2]. Besides, some viewers even have a physiological discomfort and such feelings varies from person to person, which limits the widespread use of three-dimensional display. In order to improve this problem, light field display needs a deep perception evaluation system guiding the display rendering and post-processing of the light field contents. There are many reasons that affect the depth perception of the three-dimensional display of light field, including scene parallax, content, dynamic objects, visual focus, and even some artistic methods such as scene color, light, shade, composition and so on.

The human visual system needs both physiological and psychological depth information to generate three-dimensional sensations in the brain [3]. Psychologically depth information can also be obtained through two-dimensional images, while physiological depth information can only be provided by true three-dimensional physics. The four physiological depth information needed by the human brain for three-dimensional perception include accommodation, convergence, motion parallax and binocular disparity. Through the two-dimensional images produced by the psychological depth of information processing, human brain can also get a sense of three-dimensional. Psychological depth information includes linear perspective, shading, texture, prior knowledge. At present, the light field depth perception control is usually based on the correction of the parallax. But the considered factor affecting the depth sense is relatively simple, and this method sometimes causes viewing discomfort. In this paper, a depth-sensing parallel system of light field display is proposed, which considers both the physiological and the psychological depth information, to render a better depth perception and comfortable light field display, through modifying the visual focus and re-processing the light field data correspond.

To determinate the position of the visual focus point for a light field display, we consider and capably control the depth sense affecting factors [5-8] through experiment and evaluation, such as the visual focus point's depth information (*depth*), the motion speed of the object (which the visual focus point focused on) (*speed*), the scene's brightness contrast (*brightness*), or the scene's ambiguity (*blur*). However, a better depth perception may bring uncomfortable viewing experience. Therefore, we also consider the influence of comfort when modeling the display system [9-12]. Moreover, this evaluation model can be adjusted according to the needs of different people. Particularly, our main contributions includes:

1) Various factors affecting the depth sense are systematically analyzed and used to determine the scene visual focus position.

2) A parallel learning approach is used to model the depth sense factors synthetically and it learns through an evaluation model how to choose a scene's visual focus personally.

3) Based on the free-viewpoint rendering method, optical flow analysis and object segmentation (or tracking), light field can be well re-rendered.

4) We use one-shot segmentation for 3D video focus position tracking process. Based on the parallel learning theory for adding "virtual" training data, a video processing parallel system for one-shot video segmentation is proposed and the segmentation accuracy is increased to 83.6%. It provides technical support for the realization of the 3D display parallel system.

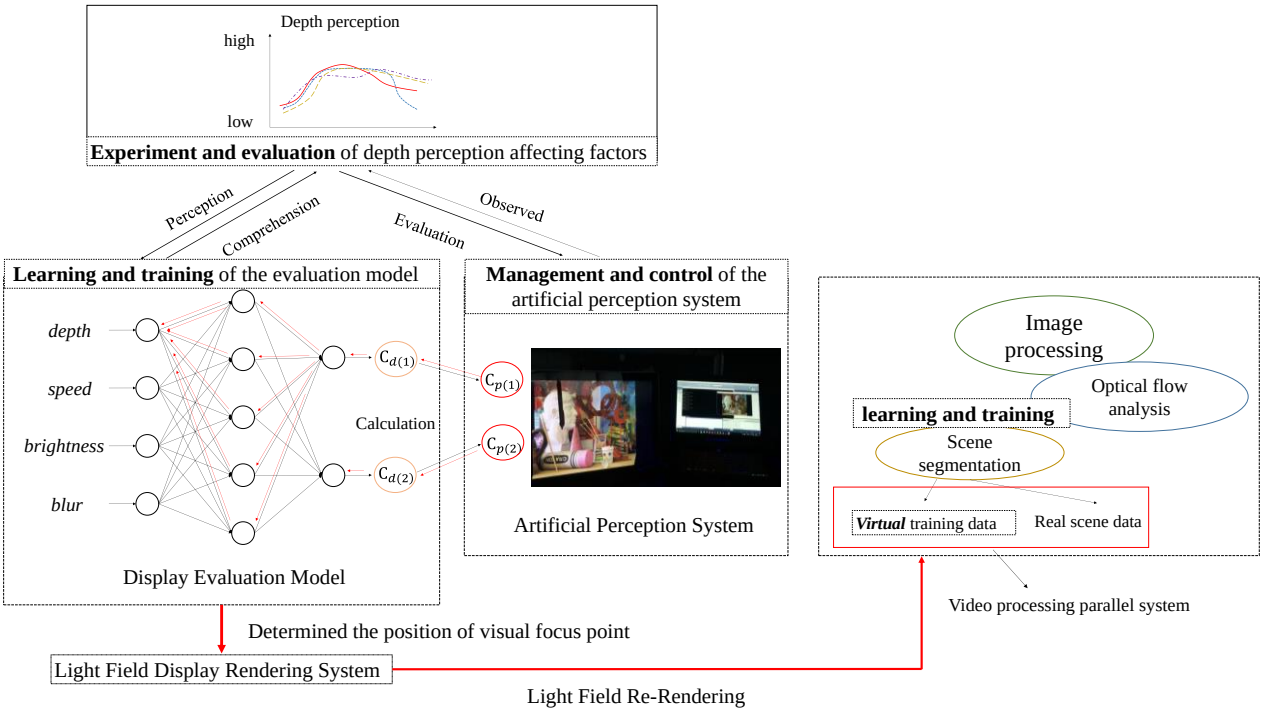


Figure 1. Overview of our proposed 3D display parallel system.

2. Technical Description

Overview of the proposed method: We propose a novel 3D display parallel system [13-14] for depth sense optimization and it empirically guides how the light field should be re-rendered. Structurally, the parallel system consists of a light field display rendering system, display evaluation model and an artificial perception measurement system. Through online learning, offline calculation and manual interaction with the artificial perception system, the display evaluation model provides reference, estimation and guidance of the visual focus' position for the light field display rendering system. That is the display evaluation model firstly learn how the human will feel at different light field renderings and then choose a better way to re-render. The main process of the parallel system includes experiment and evaluation of the depth perception affecting factors, learning and training of the evaluation model, and management and control of the artificial perception system (in Figure 1.).

Light Field Display Rendering System: In the previous work, we have systematically introduced the rendering process of the light field display [15-16], including the early light field data acquisition, processing [17] and display calibration [18-21]. Since the optical path is determined by the display hardware, we can only correct the color value of the light in the re-rendering process. And by changing the view of the data, we are able to correct the color of the light [22-23]. The correction process needs the parameter – the virtual focus which obtains through our depth perception parallel system.

Display Evaluation Model: The display evaluation model based on the supervised network is constructed. During the learning and training time, multi-3D videos are provided as the training videos. First, a visual focus is chosen by the manual interaction. After the visual focus determined,q depth-aware affecting factors' value (for example, the value of the *depth*, *speed*, *brightness* and *blur*. And the *speed* factor is only for 3D video not for 3D image) are fixed as the evaluation model's inputs. The outputs include a depth perception score $C_{d(1)}$ and a comfort score $C_{d(2)}$. After several back-propagation between the real scores ($C_{d(1)}$ and $C_{d(2)}$) and the artificial perception system's scores ($C_{p(1)}$ and $C_{p(2)}$), the evaluation model's weights can be well defined. The final scores of the display evaluation model are closer to that of the artificial perception system, which means the display evaluation model begins to learn how human feel at different light field renderings. When applying to the test 3D video, the evaluation model firstly selects n focus positions based on prior knowledge during the experiment and evaluation time and score the n focus positions through the evaluation model. Then, the light field display rendering system selects the point with the highest score, which is the best visual focus for re-rendering.

Artificial Perception Measurement System: Artificial perception measurement system are include the personnel participations for human's depth sense training and a manual interaction system for virtual focus setting. Through the interaction with mouse, the position of focus point can be determined and a series of scores under different rendering results can be obtained, including depth perception score $C_{p(1)}$ and a comfort score $C_{p(2)}$.

Let's discuss how to rendering scene at different focus during the manual interaction based on EPI. Schematic of the technique is shown in Figure 2. A 3D light filed, created from a set of multiview images, is presented as $\boldsymbol{LF}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{V})$. $\boldsymbol{V}$ denotes the discrete number of views and $\boldsymbol{V_{RGB}}$ represents the reference image $\boldsymbol{I_{RGB}}$'s sequence view number. A planar $\boldsymbol{x}$–$\boldsymbol{V}$ cut represents epi-polar plane image (EPI). $\boldsymbol{l}(\boldsymbol{y'},\boldsymbol{x'},\boldsymbol{V_{RGB}})$, which passes through the point $\boldsymbol{p}(\boldsymbol{y'},\boldsymbol{x'},\boldsymbol{V_{RGB}})$ on $\boldsymbol{I_{RGB}}$, denotes the linear structure in EPI. And the slopes $\boldsymbol{k}$ of all the linear structure is calculated as: $\boldsymbol{k} = f \cdot baseline/z$, where $baseline$ is the baseline of adjacent cameras (in Figure 3(a)), f denotes the camera focal length and z represents the depth for $\boldsymbol{I_{RGB}}$'s each pixel. And the depth scenes finally can be accurately and measurably obtained (in Figure 3(b)):

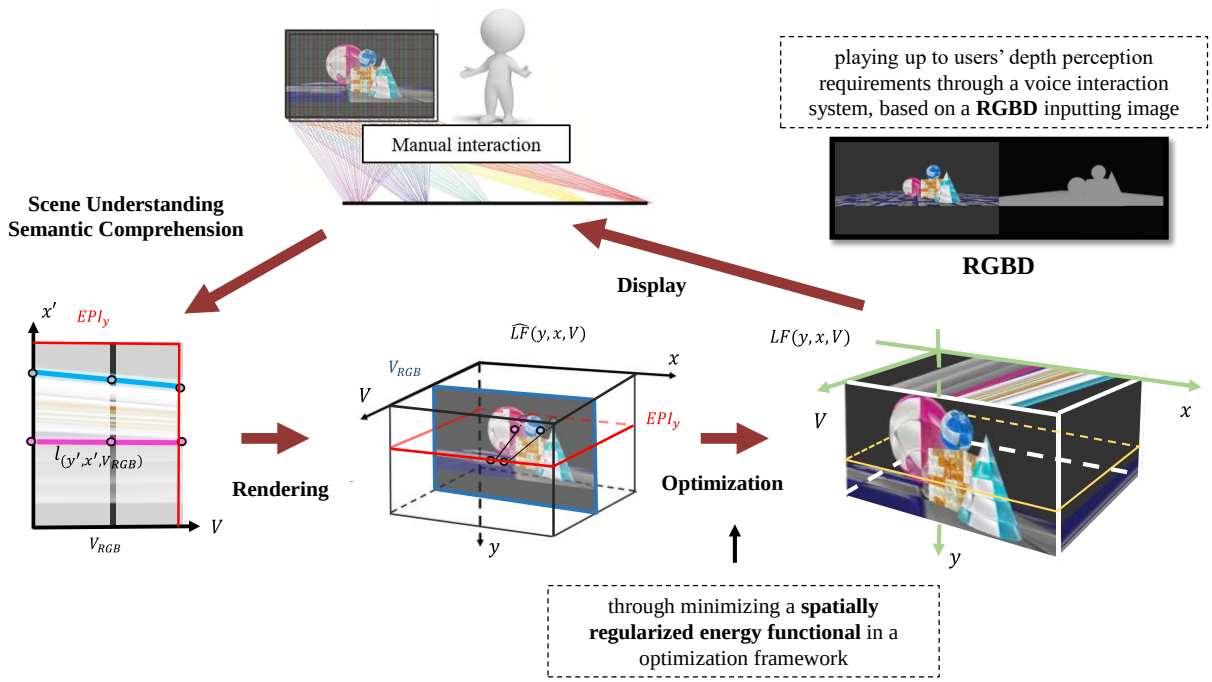


Figure 2. The schematic of the technique.

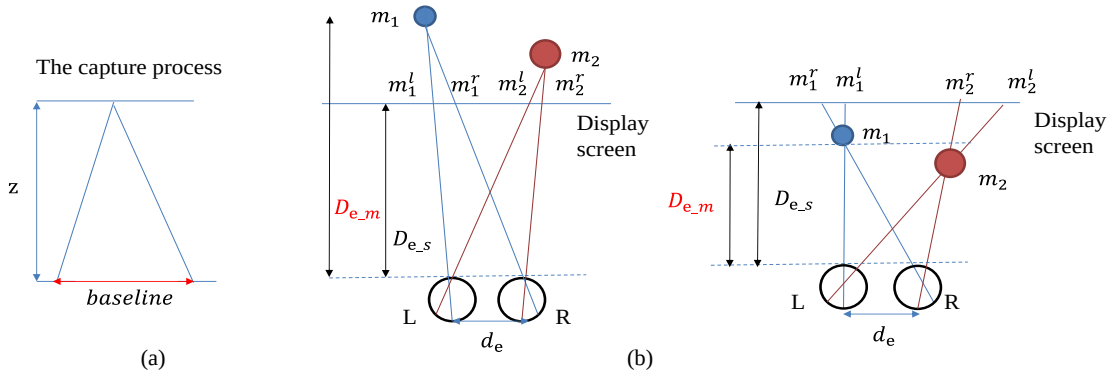


Figure 3. (a). The capture process when acquired the RGBD images; (b). Depth scenes' calculation: D_{e_s} and D_{e_m} represent the distances between viewer's eyes to screen and to displayed object m_1, respectively. And S is the display scale between the display image size and the original view image size.

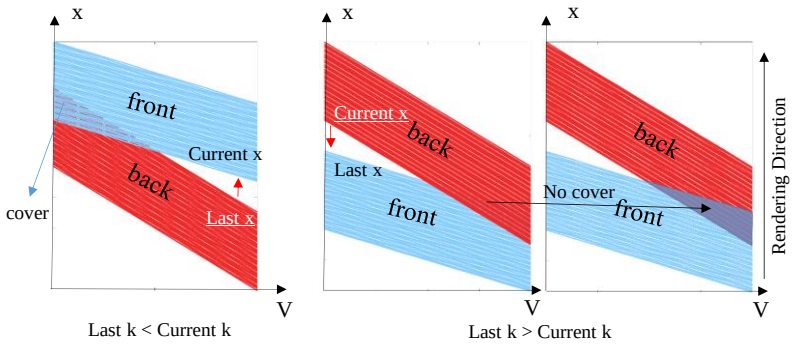


Figure 4. Holes' fill directions and pixels' cover strategies.

Figure 5. Rendering results with the artificial interactions.

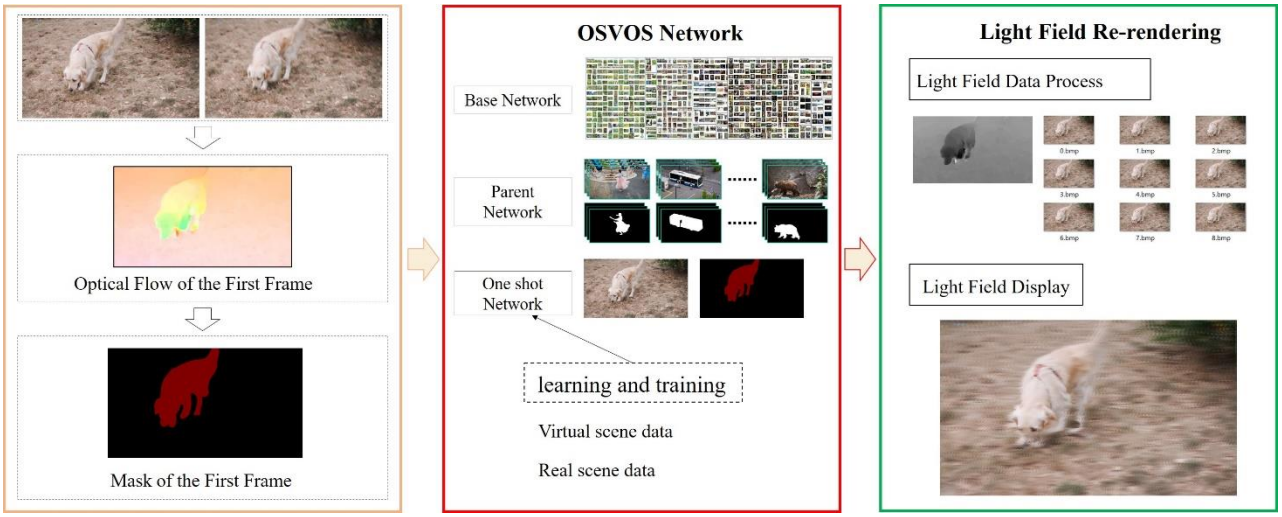


Figure 6. The object tracking and light field re-rendering process.

$$\text{depth scenes} = (1 - \frac{d_e}{d_e + S \cdot \boldsymbol{k}}) D_{e\_s} \quad (1)$$

When the point $\boldsymbol{p}$ is chose to be the focus, it's depth scenes sets 0 and other points' $\boldsymbol{k}$ are also changed according to *eq.2:*

$$\boldsymbol{k} = f \cdot baseline/z - k' \quad (2)$$

, where the k' is the $\boldsymbol{p}$'s original slope. By propagating the color of $\boldsymbol{p}$ to all the points on $\boldsymbol{l}(\boldsymbol{y}', \boldsymbol{x}', \boldsymbol{V_{RGB}})$, the initial light filed $\widehat{\boldsymbol{LF}}$ is rendered. During $\widehat{\boldsymbol{LF}}$ created time, when each EPI of the initial light filed rendered through the direction in Figure 4, we record the last $\boldsymbol{k}$, current $\boldsymbol{k}$, last rendered pixel and current rendered pixel respectively to determine the holes' fill directions and pixels' cover strategies.

We use a regulariser comprising a weighted Huber norm over the gradient of the final light filed $\boldsymbol{LF}$, $\omega \left|\left|\nabla_{y,x} LF(y,x,V)\right|\right|_{\varepsilon}$, which ensures un-twisted view images in $\boldsymbol{LF}$ for high quality 3D display, where $\omega \propto \nabla \boldsymbol{I_{RGB}}(y', x')$ $((y', x') \in \boldsymbol{l}(\boldsymbol{y}', \boldsymbol{x}', \boldsymbol{V_{RGB}})$ and $(y, x) \in \boldsymbol{l}(\boldsymbol{y}', \boldsymbol{x}', \boldsymbol{V_{RGB}})$). The resulting energy functional therefore contains a non-convex photometric error data term and a convex regulariser:

$$min_{LF} E = \iiint \left\{ \left(\boldsymbol{LF}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{V}) - \widehat{\boldsymbol{LF}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{V}) \right)^2 + \omega \left|\left| \nabla_{y,x} \boldsymbol{LF}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{V}) \right|\right|_{\varepsilon} \right\} dydxdV \quad (3)$$

In Figure 5, it is the rendering results during the artificial interactions at different virtual focus points (red points).

Video Processing Parallel System: In these part, we take the *speed* factor as an example to illustrate the light field re-rendering process. From the results of the artificial perception measurement, the focus point usually focused on the objects' with the medium speed. We can calculate the optical flow for each frame and determine for the medium speed objects for each frame, but it may be slow and the high frequency of switching the visual focus from one object to another object will reduce the viewing comfort. Usually, the 3D video's contents are not complicated and the objects in the video usually move evenly. So we only use the first frame's optical flow map to determine the medium speed object. As the following video frame, we tracking the object selected in the first frame.

We first calculate the moving speed of the objects in the 3D video's first frame by dense optical flow (u, v) [25]:

$$min_{u,v} E(u,v) = \int\int \left[\left(T(x,y) - I(x+u, y+v) \right)^2 + \alpha \left(u_x^2 + u_y^2 + v_x^2 + v_y^2 \right) \right] dxdy \quad (4)$$

After acquiring the optical flow, it can be easy to get the relative speed relationship of the objects' movements in the 3D video. For the object's tracking, we use one-shot video segmentation method in subsequent video frames, which aims to densely segment out the object(s) for all video frames, given only one frame (usually the first frame) mask of the required object(s). Figure 6 shows the object tracking and light field re-rendering process: by using the optical flow obtained in *eq.4* , the position of the most suitable

object in the first frame can be roughly determined. And through super pixel and network iteration, the mask of the first frame with high accuracy can be obtained. Then the OSVOS network is used to segment the object appearing in subsequent videos. The algorithm tracks and splits well even when the object is not rigidly deformed.

To improve the accuracy of OSVOS [26] network, we added a large number of virtual training data samples as shown in Figure 7 during the one-shot network training. The purpose is to simulate 1) the diversity of the video background of the following frames and 2) the rigid changes of the foreground objects, even the non-rigid changes. Experimental data show that our method is improved by about 4% compared with the original method, as shown in Part *3*.

As shown in Figure 8, taking the "basketball" video as an example, we first determine which object (the green one in Figure 8 (c)) a viewer tend to and track it with the improved OSVOS. Among them, 10 points are selected at random (Figure 8 (d)). Depth perception scores learned by the evaluation model are used to predict the visual focus' position. The point with the highest comprehensive score is selected as the focal point for scene rendering. According to the re-rendering process [22], the free-viewpoint rendering method based on the EPI (Figure 8 (b)) is performed. Figure 8 (e) shows the result of the final 3D display map.

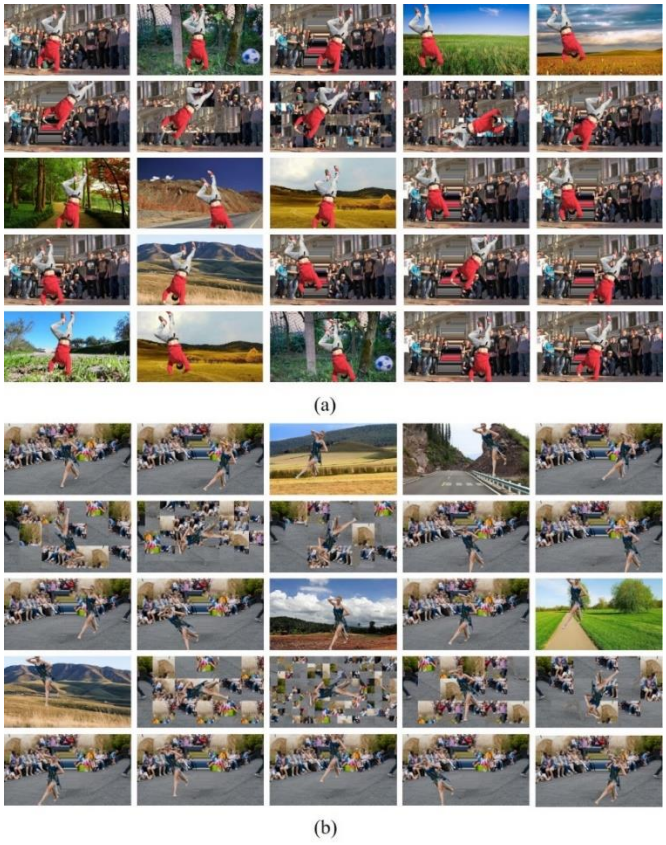


Figure 7. DAVIAS virtual training data (a) "breakdance", (b) " dance-twirl".

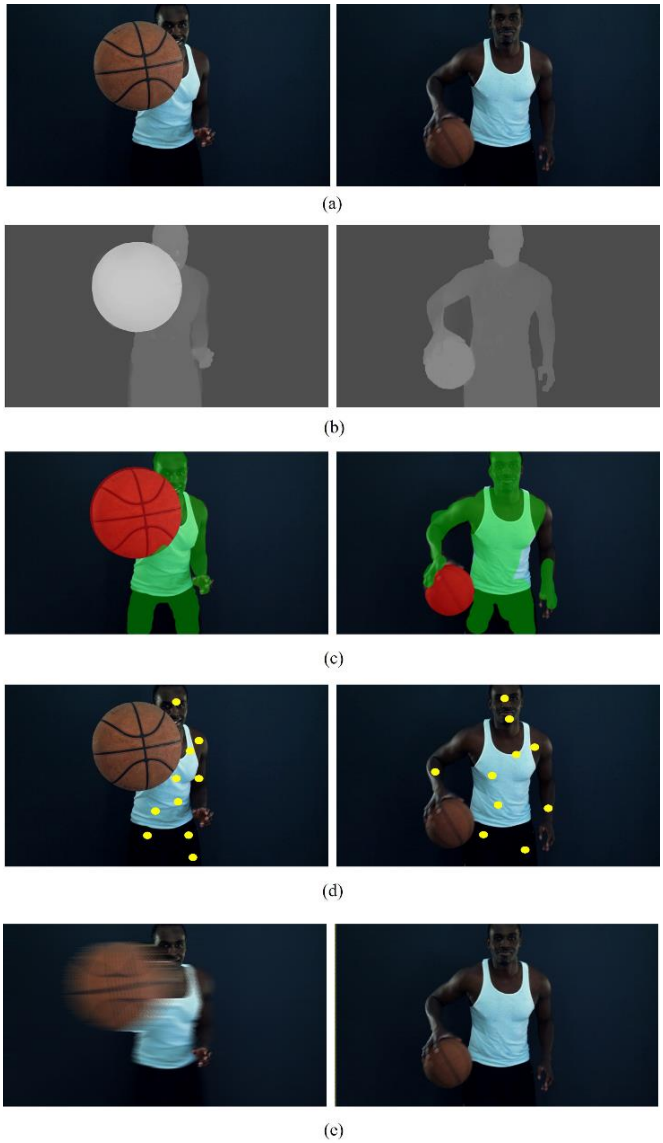


Figure 8. The results of the "basketball" video, the 30-th frame (left) and the 60-th frame (right). (a) The frame image;(b) The depth map;(c) The segmentation result;(d) Random selection of the candidate focus points;(e) The re-rendering 3D map.

3. Results and Discussion

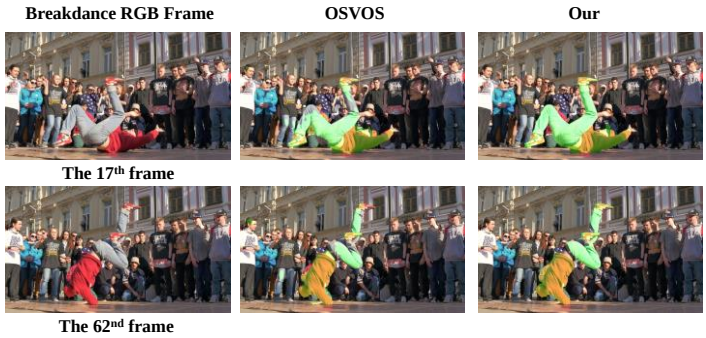
Display Evaluation Model Evaluation: The artificial perception measurement system provides the basis of selecting an effective visual focus for the light field rendering system or improving the current selected visual focus. With the deepening and expansion of the experiment, there may be a number of potential development trends. For example: 1) with new staff members joining in, the depth sense evaluation may change personally; 2) person's visual mechanism changes (like the myopia or emotional impact), which results in different evaluation; 3) As the technic improved, the display evaluation model can learn more factors or even the network model changes. At these time, when display evaluation model and the artificial perception measurement system produce difference or generate errors feedback signal, the evaluation modes or parameters need

to be corrected to reduce the errors and begin to analyze a new round of the optimization and evaluation. For example, we randomly selected testers to join the artificial perception test system and generated 1000 different visual focal light field display for evaluation. In a long-term training, light field display rendering system model obtains the depth sense factors' values, which are suitable for the tester and the prediction accuracy is 96.84%.

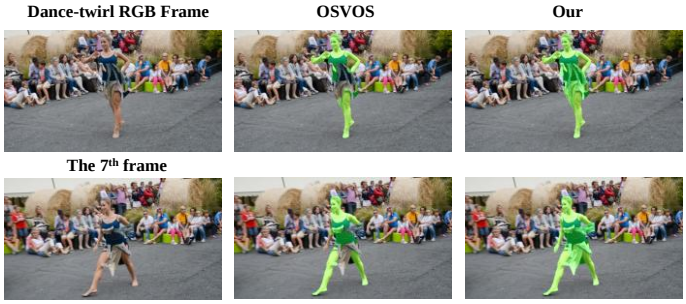
Video Processing Parallel System Evaluation: In order to improve the accuracy of the OSVOS network, we augment a large number of virtual training data samples which are expanded from the first frame while training on the one shot network. Experiment indicates that our video processing parallel system is able to improve accuracy compared with the original method (Table 3 and Figure 9).

Table 3. The Results of different virtual methods

	Accuracy
OSVOS	79.4%
Video processing parallel system	83.6%



(a)



(b)

Figure 9. Comparison the tracking results on DAVIS 2016. (a) Breakdance;(b) Dance-twirl.

3D Display Parallel System Evaluation: We measure the quality of the final re-rendering light field display by subjective test experiments. In order to avoid the interferences by the 3D video itself, we first select the best focal point position interactively and to score both the depth sense and the comfort ($\mathbf{C}_{o(1)}$ and $\mathbf{C}_{o(2)}$). Then, we rendered those test 3D videos by the method shown in Figure 8, and participants evaluates them by depth and comfort scores, $\mathbf{C}_{t(1)}$ and $\mathbf{C}_{t(2)}$. The final score is calculated as follow:

$$\text{Score} = \alpha \cdot \frac{\mathbf{C}_{t(1)}}{\mathbf{C}_{o(1)}} \times 5 + (1-\alpha) \cdot \frac{\mathbf{C}_{t(2)}}{\mathbf{C}_{o(2)}} \times 5 \qquad (5)$$

, where $\alpha = 0.6231$ in our experiments. The final average score result of the parallel system is 4.34 (out of 5), which was 29.82% higher than the focus randomly selections.

4. Conclusions

In this paper, we propose a depth sense parallel system of light field display, considering a variety of factors affecting the light field display when rendering the scene and modeling the guidance of re-rendering. The 3D display parallel system composites by a light-field display system, an evaluation model and an artificial depth measurement system. And with the video processing parallel system, the network can well train the first frame of the 3D video, so it tracks the best chosen object precisely. Thus it performs well on the light field re-rendering process.

5. Acknowledge

This work has been supported by project of National Natural Science Foundation of China (Grant No. 61605240) of Institute of Automation, Chinese Academy of Sciences (CASIA).

6. References

[1] Van Berkel, Cees. "Image preparation for 3D LCD." Electronic Imaging'99. International Society for Optics and Photonics, 1999.
[2] T Okoshi. 3 Dimensional Imaging Techniques (New York: Academic, 1976), ch. 2, 8–42.
[3] Geng, Jason. "Three-dimensional display technologies." Advances in optics and photonics 5.4 (2013): 456-535.
[4] M. Levoy and P. Hanrahan. "Light field rendering," in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH (1996), pp. 31–42.
[5] LI J, BARKOWSKY M, and CALLET P L. "Visual fatigue caused by viewing stereoscopic 3D video: Influence of 3D motion". Displays,2014,35(1):49-57.
[6] PARK J, LEE S, and BOVIK A C. 3D visual discomfort prediction: Vergence, foveation, and the physiological optics of accommodation[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(3): 415-427.
[7] LEE S, JUNG Y J, SOHN H, et al. Effect of stimulus width on the perceived visual discomfort in viewing stereoscopic 3-D-TV[J]. IEEE Transactions on Broadcasting, 2013, 59(4): 580-590.
[8] WOPKING M. Viewing comfort with stereoscopic pictures: an experimental study on the subjective effects of disparity magnitude and depth of focus[J]. Journal of the Society for Information Display, 1995, 3(3): 101-103.
[9] WANG Qin, WANG Qionghua, and LIU Chunling. Effects of parallax and spatial frequency on visual comfort in autostereoscopic display[J]. Journal of Optoelectronics ·Laser, 2012, 23(8): 1604-1608.
[10] SOHN H, JUNG Y J, LEE S, et al. Predicting visual discomfort using object size and disparity information instereoscopic images[J]. IEEE Transactions on Broadcasting, 2013, 59(1): 28-37.
[11] KIM H, LEE S, and BOVIK A C. Saliency prediction on stereoscopic videos[J]. IEEE Transactions on Image Processing, 2014, 23(4): 1476-1490.
[12] WILCOX L M and HESS R F. Dmax for stereopsis depends on size, not spatial frequency content[J]. Visual Research,1995, 35(9): 1061-1069.
[13] Fei-Yue Wang, "A Computational Framework for Decision Analysis and Support in ISI: Artificial Societies, Computational Experiments, and Parallel Systems", LNCS

3917, 2006, pp. 183-184.
[14] Fei-YueWang, “Parallel Control and Management for Intelligent Transportation Systems: Concepts, Architectures, and Applications”, IEEE Transactions on Intelligent Transportation Systems, 2010, Vol. 11 Issue: 3, pp. 630-638.
[15] R. Pei, Z. Geng, Z. Zhang, R. Wang. “A Novel Optimization Method for Lenticular 3D Display Based on Light Field Decomposion,” in Journal of Display Technology, vol. 12, no. 7, pp. 727-735, Jul. 2016.
[16] R. Pei, Z. Geng and Z. Zhang. “ Subpixel Multiplexing Method for 3D Lenticular Display,” in Journal of Display Technology, vol. 12, no.10, pp. 1197-1204, Oct. 2016.
[17] J Pei, Z Geng,Z Zhang. “A Real-Time Depth Map Refinement and Disparity Ranges Expansion System (DRDE) for Multiview Rendering,” Computer Graphics International, June 28-July 01, 2016, Heraklion, Greece. 2016 ACM.
[18] R.Pei, Z. Geng, K. Ma, M. Zhang, R. Wang. "A 3D Dimensional Lenticular Display Rendering Based on Light Field Acquisition", 2017 display week.
[19] R.Pei, Z. Geng, K. Ma, M. Zhang, R. Wang. "Three Dimensional Lenticular Display Synthetic Image Rendering Based on Light Field Acquisition", Journal of the Society for Information Display. 2017.
[20] K. Ma, R.Pei, Z. Geng. "A Turnkey Solution to Automatic Calibration and Crosstalk Reduction for Mobile Three-dimensional Display", 2017 display week.
[21] R.Pei, Z. Geng, K. Ma, M. Zhang, “Light Field Subsample Method for Three Dimensional Lenticular Display Rendering”, EuroDisplay 2017.
[22] R. Pei. "Light Filed Render and Optimization for Measurable 3D Depth Perception Interaction", 2017 ICDT.
[23] R.Pei, Z. Geng, K. Ma, M. Zhang, “Immersive three-dimensional environment Interaction method for the virtual reality (VR) system”, 2017 IDMC.
[24] Newcombe, Richard A., Steven J. Lovegrove, and Andrew J. Davison, "DTAM: Dense tracking and mapping in real-time." 2011 international conference on computer vision. IEEE, 2011.
[25] Brox T, Bruhn A, Papenberg N, et al．High accuracy optical flow estimation based on a theory for warping：Computer Vision-ECCV，2004：25-36.
[26] S. Caelles*, K.K. Maninis*, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. “OSVOS: One-Shot Video Object Segmentation State-of-the-Art Results in Accuracy and Speed”, CVPR 2017.