J. Vis. Commun. Image R. 56 (2018) 287–295

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

View synthesis using foreground object extraction for disparity control and image inpainting ☆

Dongxue Han [a], Hui Chen [a,*], Changhe Tu [b], Yanyan Xu [c,*]

[a] School of Information Science & Engineering, Shandong University, Jinan, PR China
[b] School of Computer Science & Technology, Shandong University, Jinan, PR China
[c] Department of Civil & Environmental Engineering, MIT, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 18 March 2018
Revised 26 July 2018
Accepted 6 October 2018
Available online 9 October 2018

Keywords:
Virtual view synthesis
DIBR
Disparity control
Exemplar-based inpainting
Foreground object extraction

ABSTRACT

Among the rapidly growing three-dimensional technologies, multiview displays have drawn great research interests in three-dimensional television due to their adaption to the motion parallax and wider viewing angles. However, multiview displays still suffer from dazzling discomfort on the border of viewing zones. Leveraging on the separability of scene via foreground segmentation, we propose a novel virtual view synthesis method for depth-image-based rendering to alleviate the discomfort. Foreground objects of interest are extracted to segment the whole image into multiple layers, which are further warped to the virtual viewpoint in order. To alleviate the visual discomfort, global disparity adjustments and local depth control are performed for specific objects in each layer. For the post-processing, we improve an exemplar-based inpainting algorithm to tackle the disoccluded areas. Experimental results demonstrate that our method achieves effective disparity control and generates high-quality virtual view images.

1. Introduction

Interests in three-dimensional technology have grown rapidly, particularly in autostereoscopic displays, which make glasses-free stereo perception possible. Crosstalk, Moire Fringe, resolution reduction and narrow viewing angle are the main technical obstacles of 3D flat panel displays [1]. The first three could be improved by utilizing advanced display devices. Aiming at the narrow viewing angle, there are two popular approaches accommodating wider viewing angle: eye-tracking and multiview displays [2]. Eye tracking has performed well in games and virtual reality applications but has the limitation of displaying to a single viewer. Multiview system separately redirects stereoscopic image sequences to multiple viewing zones, thereby allows multiple users to watch from different viewpoints simultaneously. Besides, it provides smooth motion parallax while the viewers are moving around. And the development of multiview video coding and transmission enables depth-image-based rendering (DIBR) of additional viewpoints and helps in the application scenario [3–5].

Visual comfort and virtual view synthesis are two important ingredients of multiview displays. The first problem can be compensated through disparity control. Previous works mostly focus on stereoscopic distortions of two-view displays that original images captured from cameras mismatch binocular perception [6–9], but few studies consider the strong visual discomfort and fatigue on the border of viewing zones in multiview displays. As illustrated in Fig. 1, the viewing zones in green indicate that viewers standing there could receive stereo image pair in the right order. In contrast, when the viewers move to the border of the viewing zones (the red zones), the perceived two images are non-adjacent in content, that is, the viewer receives a wrong stereo image pair. Accordingly, the 3D scene cannot be reconstructed by the human brain and causes the spectators' vertigo instead. Towards virtual view synthesis, DIBR [10] is an efficient solution to produce content for 3D television (3DTV), which also requires proper disparity maps to generate high-quality virtual views for the viewers.

Recent researches demonstrate that proper disparity remapping could find a balance between vivid 3D perception and visual comfort [7–9,11]. A simple implementation of remapping is shifting or scaling the horizontal disparity. Targeting at the visual comfort of regions of interest, Lei et al. [11] regarded geometric center of the salient regions as the key-point, where the disparity is set to zero by shifting the multiview images. The work in [8,9] introduced another remapping framework. They performed global linear disparity scaling and then local nonlinear refinement.

☆ This paper has been recommended for acceptance by Zhihai He.

* Corresponding authors.
E-mail addresses: huichen@sdu.edu.cn (H. Chen), yanyanxu@mit.edu (Y. Xu).

https://doi.org/10.1016/j.jvcir.2018.10.004
1047-3203/

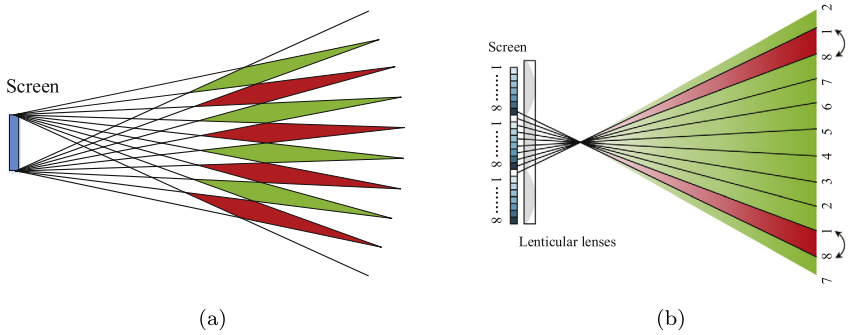

Fig. 1. Viewing zone (indicated in green) of (a) a two-view display, (b) a multiview display. Binoculus receive wrong view-image pair in red zone. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Coupled with the extracted features, support vector regressing is applied to predict remapping models in these approaches. But it leads to a high computational cost. Moreover, these methods uniformly control disparity on global images without the knowledge of specific objects. However, as different regions contribute to the discomfort to varying degrees, individually controlling the disparities of certain objects is a better strategy to enhance the visual comfort. Mangiat et al. [7] separately adjusted the disparities of face and background for mobile video, and they also utilized face detection and disparity threshold to identify the viewer's head ahead of the camera.

Previous studies have made great progress in DIBR. The Motion Picture Expert Group (MPEG) released View Synthesis Reference Software (VSRS) [12] as a reference. It takes two reference views plus the associated depth maps as inputs to produce vivid virtual views. With the emergence of depth cameras, depth maps have been widely utilized. Although the recorded depth images tend to be corrupted by missing data, there have been extensive sophisticated methods to overcome the limitation [13–16]. Furthermore, Wang et al. [17] put forward the 3D warping with depth-based pixel interpolation to remove cracks and background-based hole filling for disocclusions. Manap and Soraghan [18] separated the depth map into several layers to perform view interpolation independently. To eliminate the holes arising from limited sampling density, Mehrdad et al. [19] detected superpixel by segmentation before warping, which is not suitable for the disoccluded regions. Some works applied improved exemplar-based inpainting algorithms to address the disocclusion issue [20–22], which synthesize consistent and realistic texture in the patch level.

To alleviate the discomfort of viewers, this paper presents a disparity control method for multiview displays using foreground segmentation. To our knowledge, if the spatial thickness of the reconstructed 3D scene is within a small depth range, the viewers barely feel the visual discomfort even at the border of viewing zones. We hence propose to adjust z-dimensional depth of the designated objects to flatten the scene, which enables a gradual transition from a viewing zone to the adjacency. In the context of DIBR, we present a new segmentation-based view rendering algorithm. First, the foreground objects of interest are extracted to segment the whole image into several layers, which are warped to the virtual viewpoint in order. Secondly, with the labeled separated layers, an exemplar-based inpainting algorithm is proposed to handle the disocclusions. In summary, our contributions are in two aspects: (i) a novel disparity remapping method to overcome the visual discomfort induced by discontinuous viewpoints in multiview 3DTV; (ii) a new view synthesis algorithm based on foreground object extraction for disparity control and image inpainting.

The paper is organized as follows. In Section 2, we describe the main target problem to be solved and the outline of the proposed scheme. Section 3–5 present the details of our scheme. Experimental evaluations are presented in Section 6. Conclusions are drawn in Section 7.

2. Overview of the approach

2.1. Problem statement

Conventional 3DTV employs a two-view display for the viewer wearing a pair of stereo glasses. While for the recent glasses-free 3DTV, the two-view stereo image pair can be received correctly only when the viewer is at the ideal position and distance; nevertheless, there is a 50% chance that the viewer perceives the images in the wrong order, as shown in Fig. 1a. Such disordered image pair is named as pseudoscopic image [23], and causes dazzling discomfort especially when there are strong 3D effects with large disparities. In this context, multiview displays are designed to enlarge the viewing angle allowing more adjacent perceiving positions, but they still suffer from the abrupt view-image change at the border of viewing zones. Fig. 1 illustrates the difference in viewing zone and the dazzling discomfort caused by the two-view and 8-view displays, respectively. Multiview displays expand the viewing zones by mapping a number of images in a given order to the screen. The image sequence depicts the same stereo scene from slightly shifted viewpoints. When viewers are in the green viewing zone, binoculus receive a stereo pair, image k and $k+1$. Binocular parallax and motion parallax are enabled when viewers move around. Nevertheless, when the viewer moves to the border of the viewing zones (indicated in red), one eye receives the last view while the other receives the first, which results in an abrupt excessive change of disparities. Such discontinuity of viewpoints causes the dazzling discomfort of the scene with strong 3D effects.

Generally, the objects in a real scene at different depth level have different disparities. The objects with large disparity primarily contribute to the visual discomfort on the border of viewing zones. It is beneficial to individually adjust the disparities of the objects of interest. Depth layering and salient object segmentation are primarily explored in the literature [18,11], and yet rough segmentation commonly results in fracture or fold of complete objects. Hence, we extract semantically meaningful foreground objects, of which disparities can be freely adjusted.

2.2. Outline of the method

A flowchart of our multiview synthesis approach is shown in Fig. 2. Firstly, the RGB and preprocessed depth data from the reference viewpoint are utilized to segment the image to multiple layers, thus we could distinguish the foreground objects from the background. Secondly, the extracted objects and background are

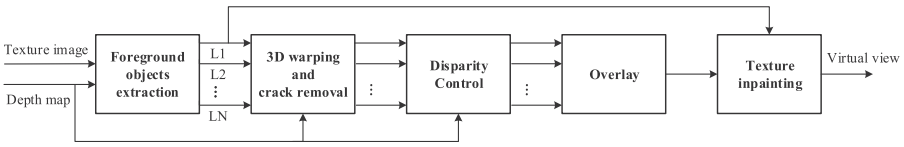

Fig. 2. Sketch of the proposed scheme.

warped to the virtual viewpoint in a specified order. In this step, we identify and remove the cracks. Thirdly, the global and local disparities are adjusted to improve the visual comfort. Then the layers are overlaid in sequence. Finally, the remaining holes are filled using the proposed inpainting method based on the segmentation results. It is noteworthy that we apply a reference view and an associated depth map to synthesize a virtual view.

3. Layered warping for designated objects

As aforementioned, it is reasonable to individually adjust the disparities of the objects of interest and background. Thus, we present to perform foreground object extraction and layer warping first. On the basis of that, we can remove the crack artifacts and freely adjust the disparity.

3.1. 3D Warping with foreground object extraction

Depth map provides additional information to enhance the segmentation accuracy. As such, an image segmentation method proposed by Xiao et al. [24] is adopted to semantically extract meaningful foreground objects from the background, as well as their silhouettes. Xiao et al. improved the graph-based method by combining the color and depth information to over-segment the color scene for region merging step. Next, a multi-threshold Otsu method is used to segment the depth map to multiple layers. An automatic depth layer selection scheme is designed to reduce user interactions. Ultimately, the regions are merged on the basis of regional continuity which is established under the constraints of depth layers, user-specified seed points and area threshold.

After the objects of interest are segmented from the background, all segments are ranked in descending order of the distance to the viewpoint and defined as layer 1, 2, . . ., N. Layer 1 is generally the background and layer 2 to N are the foreground regions. Each layer is warped to a virtual viewpoint and preliminarily completed with cracks interpolated.

3.2. Crack removal

There are usually cracks on the rendering foreground layers. The crack artifact is attributed to integer round offs of projected coordinates. Neighboring pixels are rounded off to non-neighboring integer values in the virtual view. When it occurs along with the foreground boundary, the crack is usually seeped through by the erroneously mapping of background pixels, which is referred to as translucent cracks [25]. As illustrated in Fig. 3, the crack on the foot of a baby is filled with background pixels by mistake as it is surrounded by the background. Unless translucent cracks are removed before warping, they will be restrained in post-processing filtering at the cost of image blurring. Thus it is necessary to identify translucent cracks ahead of hole filling. We utilize layer warping and perform separate interpolation by masking foreground.

Let E denote the set of empty pixels in a single warped layer I. Then the set difference $I - E$ is a group of mapped color pixels. The width of cracks is generally 1 pixel not only because it stems from round-off error, but because the cameras are supposed to be set in a parallel configuration. We hence define cracks with

$$crack = \left\{ p \,|\, \sum_{(i,j)\in N_8(p)} E(i,j) \geqslant 4 \right\} \tag{1}$$

crack is the set of empty pixel p in whose 3-by-3 neighborhood four or more pixels are nonempty. Next, *crack* is interpolated by non-missing neighboring pixels. Finally, all completed layers are overlapped in ascending order. The results are compared with full-image warping in Section 6.

Fig. 3. Example of translucent cracks.

4. Disparity control

The disparity of a stereo image pair is dependent on the setup of cameras. Stereo cameras are usually set in two fashions, toed-in and parallel [26]. The optical axes of toed-in cameras converge to a point in depth, while the convergence depth is placed at infinity in the parallel configuration. Previous researches show that toed-in configurations are impractical to set up and introduce keystone distortions [7,27]. Accordingly, this paper discusses disparity control in a parallel configuration.

For the stereo image pair captured from parallel cameras, the objects in a scene only appear in front of the display [7]. The disparity d is inversely proportional to the depth Z, with

$$d = f\frac{b}{Z} \tag{2}$$

where f is the camera focal length, and b denotes the baseline of stereo cameras. For 3DTV at far distance, Shibata et al. [28] found that the objects behind the screen are less comfortable than those in front of the screen. Therefore, it is better to keep the objects in front of the screen. Nevertheless, the excess disparities accumulated on the border of viewing zones still result in visual fatigue. In order to balance the stereoscopic perception and visual comfort, the objects farthest from the cameras should be placed on the image plane. The first step is to calculate the minimum disparity d_{min} with maximum depth using (2). We shift the multiview images with d_{min} to adjust the disparity of the farthest point to zero.

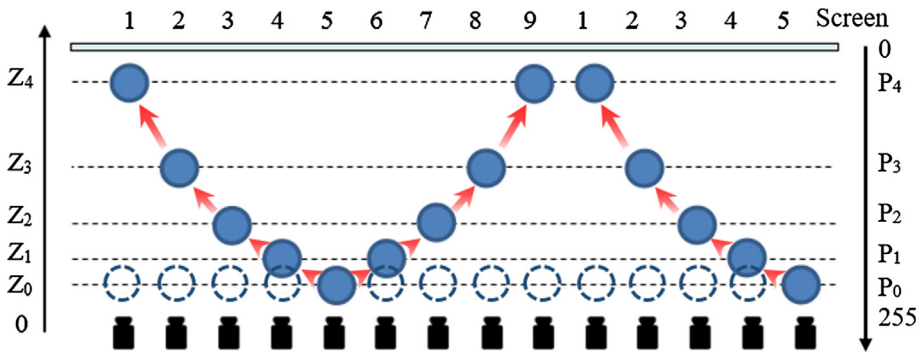

Fig. 4. Depth of the foreground object with (solid circle) and without (dashed circle) depth adjustment for 9-view foreground.

Besides, the foreground objects of interest are adjusted individually to generate smooth transition effect, from strong at the intermediate view to weak 3D perception on the border of viewing zones. Disparity control is in essence changing scene depth. Depth values are commonly quantized to intensity with the following function

$$Z = \frac{1.0}{\frac{P}{255.0}\left(\frac{1.0}{Z_{min}} - \frac{1.0}{Z_{max}}\right) + \frac{1.0}{Z_{max}}} \tag{3}$$

where Z_{min}, Z_{max} are the minimum and maximum actual depth of the scene, respectively; P is the depth value scaled to $[0, 255]$, which is equal to 255 at Z_{min} and 0 at Z_{max}.

Taking a 3D scene with 9 views for an example, we adjust the foreground depth closer to background according to the baseline between the reference and virtual view, as shown in Fig. 4. Supposing the intermediate view 5 is the reference view, we keep the background depth unchanged but pull the foreground objects backwards when synthesizing the virtual views. Let the scaled depth of an extracted foreground object be P_0, we decrease the depth to P_1 to synthesize adjacent views 4 and 6 in Fig. 4. Similarly, we generate view $i+1$ with view i for view 5–9 and generate view $i-1$ with view i for view 1–5, as illustrated by arrow in Fig. 4. The scaled depth $P_{|i-5|}$ of view i is decreased to $P_{|i-5|+1}$ by ΔP. ΔP depends on the depth of the object in the reference image and the total number of views n as

$$\Delta P = P_0 / \left\lfloor \frac{n-1}{2} \right\rfloor \tag{4}$$

Using (2) and (3), we calculate the disparity d_i of view i. Similarly, the change in disparity Δd is deduced from a change in intensity ΔP as

$$d_i = fb\left[\frac{P_{|i-5|}}{255}\left(\frac{1}{Z_{min}} - \frac{1}{Z_{max}}\right) + \frac{1}{Z_{max}}\right] \tag{5}$$

$$\Delta d = -fb\frac{\Delta P}{255}\left(\frac{1}{Z_{min}} - \frac{1}{Z_{max}}\right) \tag{6}$$

The final adjusted foreground disparity between view i and $i \pm 1$ is $d_i - d_{min} + \Delta d$. The views closer to either side have smaller disparities of foreground objects. The disparities are equal to zero on both sides of the view zone.

5. Disocclusion handling

In general, one object is of continuous depth, while the depths of foreground objects and background are discontinuous and obviously different. The discontinuity leads to translucent cracks and disocclusions. The slight cracks are removed in Section 3.2, but the large disoccluded areas are left to be solved. An exemplar-based inpainting method is proposed to handle the disocclusions in the post-processing, which is inspired by the pioneer work of Criminisi et al. [29]. The main contribution of Criminisi's algorithm lies in the isophote-driven priority term $P(p)$, which determines the filling order of patches along the boundary of the target region. It propagates the best matching texture elements to patches to be filled by a greedy method in decreasing priority order.

Image inpainting starts from tackling the boundary of holes. As illustrated in Fig. 5, given a patch Ψ_p centered at the pixel p on the border, the priority $P(p)$ is defined as the product of two terms

$$P(p) = C(p)D(p) \tag{7}$$

where $C(p)$ and $D(p)$ are the confidence term and data term, respectively. $C(p)$ gives higher priority to the patches containing more known pixels. $D(p)$ defines the strength of isophotes hitting the boundary, which encourages the propagation of local linear structure. $C(p)$ and $D(p)$ are updated in each iteration. The most similar patch $\Psi_{\hat{q}}$ is targeted within the source region Φ by minimizing their distance, that is, the sum of squared differences (SSD) between the pixels in the two patches:

$$\Psi_{\hat{q}} = \arg\min_{\Psi_q \in \Phi} d(\Psi_{\hat{p}}, \Psi_q) \tag{8}$$

Many extended methods have been developed to improve the priority term in the literature. Ahn et al. [22] employed structure tensor in $D(p)$ to strengthen the robustness of filling order. Under the circumstances of available depth information, Daribo et al. [20] added a depth variance term $L(p)$ to $P(p)$ in (7). $L(p)$ favors background patches overlaying similar depth values over foreground ones.

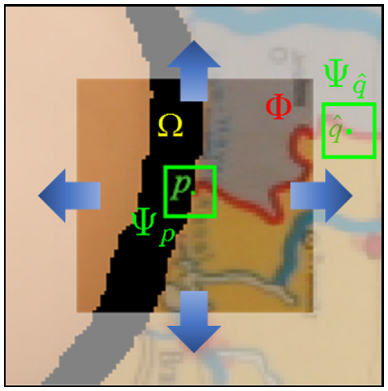

Fig. 5. Illustration of the improved exemplar-based texture synthesis algorithm.

In our approach, the confidence and data terms are retained in the priority term. We present a modification to the depth variance term $L(p)$ and a layer term $H(p)$ using layer information obtained in Section 3.1. Although Daribo et al. [20] give higher priority to background patches with similar pixel depth values, it does not work well when the filling proceeds to the center of holes. As the pixels in the outer layers are filled, updated $C(p)$ closer to background decays and plays a dominant part in the priority term, where $L(p)$ is less effective. To guarantee that the background patches are inpainted first, we replace the depth map with the layer label map obtained by foreground object extraction, and the updated layer mean term $H(p)$ is added to the original priority term. In conclusion, the revised priority term can be written as

$$P(p) = C(p)D(p)L(p) + (N - H(p)) \tag{9}$$

$$H(p) = \text{ave}H_p(q) \quad \forall q \in H_p \cap \Phi \tag{10}$$

where N is the number of layers, $H(p)$ denotes the mean value of non-empty pixels in the label patch H_p centered at p, which always favors the background patches over foreground ones. $L(p)$ is inversely related to the variance of H_p

$$L(p) = \frac{1}{1 + C * \text{var}H_p(q)} \quad \forall q \in H_p \cap \Phi \tag{11}$$

where the parameter C controls the importance assigned to the variance of layer label, and it is set to 5 in our experiments.

Once the target patch with the highest priority has been located, we search the most similar patch $\Psi_{\hat{q}}$ in the source region Φ. Disocclusions are mostly band-like and well matched with the surrounding background. We hence present a fast searching strategy within an expanded search window, coupled with a changing threshold of minimum distance, and a constraint to available patch candidates. See Fig. 5 for an illustration.

In detail, we begin with scanning the $L \times L$ pixel window $W_{\hat{p}}$ centered at the target pixel $\hat{p}$. The radius of L is initially set as the maximum disparity value $2d_{max} + 1$ (in pixel) derived from (2) using minimum depth value, which ensures valid source patches. The distance of two patches is computed when the layer labels of non-empty pixels H_q and $H_{\hat{p}}$ are the same. $\Psi_{\hat{q}}$ denotes the temporal matched patch

$$\Psi_{\hat{q}} = \left\{ \Psi_q \,\middle|\, \arg\min_{\Psi_q, H_q \in \Phi \cap W_{\hat{p}}} d(\Psi_{\hat{p}}, \Psi_q) \wedge H_q = H_{\hat{p}} \right\} \tag{12}$$

We initialize the distance threshold β_0 empirically. If the minimum distance β is less than β_0, which is normalized to a range of 0–1, $\Psi_{\hat{q}}$ is the final best matched patch. Otherwise the search window is dilated by 1 pixel and the matching threshold is added by a small value ϵ. We repeat this procedure until a patch satisfying above condition turns up or L reaches the maximum value, which is set as a quarter of image height H by default. Time efficiency is improved compared to the conventional exemplar-based inpainting methods since we begin searching from a reduced window. The pseudocode description of search steps is shown as follows.

Algorithm 1. Source patch search algorithm

Input: The target patch $\Psi_{\hat{p}}, H_{\hat{p}}$ and its coordinate
Output: $\Psi_{\hat{q}}$
1: Initialize β_0 and L
2: **while** $\beta > \beta_0$ or $L < H/4$ **do**
3: Find the best patch $\Psi_{\hat{q}}$ in the $L \times L$ window $W_{\hat{p}}$ (12);
4: $L \leftarrow L + 2, \beta_0 \leftarrow \beta_0 + \varepsilon$;
5: **end while**

6. Experimental results and discussion

6.1. Experimental setup

We implemented our method in Matlab R2013a, with the platform characterized by a PC with Intel Core i7 3.40 GHz CPU and 8 GB RAM memory. To demonstrate the subjective and objective quality of synthesized images, experiments were conducted with the MPEG test sequence *Shark* provided by NICT [30], *Ballet*, Lovebird1 from the 3DVC reference set [31,32] and the Middlebury's 2005, 2006 datasets [33]. The datasets contain three multiview image sequences captured by parallel cameras, Baby1, *Art* and *Reindeer*. The experiments were conducted in three scenarios, one for crack removal, one for disparity adjustments using segmentation, and another for virtual view quality, in which we skipped disparity control. In the third part, we compared our results against four competing schemes containing VSRS 3.5 [12,17,21,22] presented in Section 1. The first two methods used diffusion to fill holes, which are modified to use a single reference view for warping and their default inpainting scheme are adopted to fill in the missing pixels. The remaining ones, Joint Texture-Depth Inpainting (JTDI) algorithm and Ahn's method, are depth-aided exemplar-based inpainting methods for disocclusions. Since the input depth maps often have missing and inconsistent values,

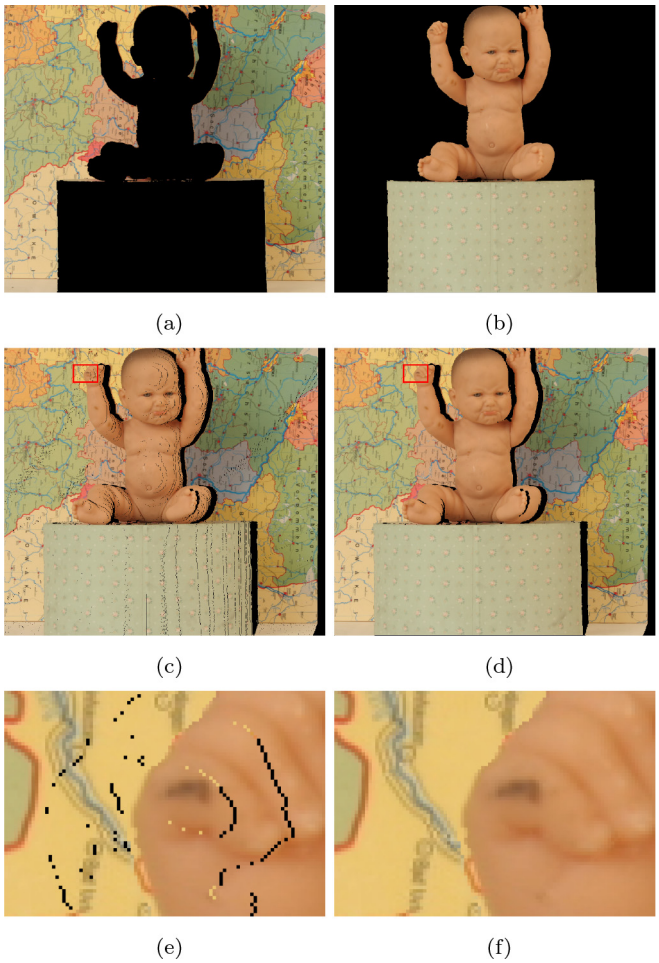

Fig. 6. Virtual view with and without crack removal. (a) layer 1: background; (b) layer 2: foreground; (c) warped image without segmentation and crack removal; (d) warped image with segmentation and crack removal; (e) and (f): comparison of rectangular portion in (c) and (d).

we preprocessed the depth data with the background-based hole filling in literature [17].

6.2. Results of crack removal

The subjective evaluation of translucent cracks is shown in Fig. 6. It illustrates the procedure of segmentation and warping. *Baby1* is segmented into 2 layers, as shown in Fig. 6a and b. Fig. 6c shows that translucent and empty cracks are removed in the warped image. The crack artifacts on the baby's right hand are completely removed by layering in Fig. 6e.

6.3. Results of disparity control

We compare the results of two disparity control methods, the one we proposed using foreground object extraction and the one using depth layering. Fig. 7a shows the depth and color maps of the reference viewpoint. As illustrated in Fig. 7b, the ventral is broken off in the virtual view generated by depth layering due to separate adjustments. While in Fig. 7c, the proposed foreground segmentation based disparity adjustments preserve the shape of the shark without visible distortion. The results demonstrate that foreground object extraction is appropriate for disparity adjustment.

Besides, *Art* has been used to test the proposed method due to multiple differentiated foreground layers. Fig. 8 utilizes red-blue anaglyph to show disparity control results. The original disparities in Fig. 8a are at the same level from the intermediate to rightmost view. After disparity control the objects at different depth levels are adjusted to different degrees in Fig. 8b This method adjusts both global and local disparities and makes a gradual transition from strong at the intermediate view to weak stereoscopic perception at the rightmost view.

6.4. Quantitative analysis

To evaluate the performance of the proposed scheme, we measure the similarity between the synthesized view and the existing original one. We adopted the commonly used evaluation methodology in the view synthesis context. Peak Signal-to-Noise Ratio (PSNR) and Mean Structural Similarity Index (SSIM) [34] are computed as objective metrics. Both metrics are applied on the full image. PSNR measures the absolute difference, while SSIM assesses the perceptual visual quality. The higher both metric values, the better the quality of the reconstructed image.

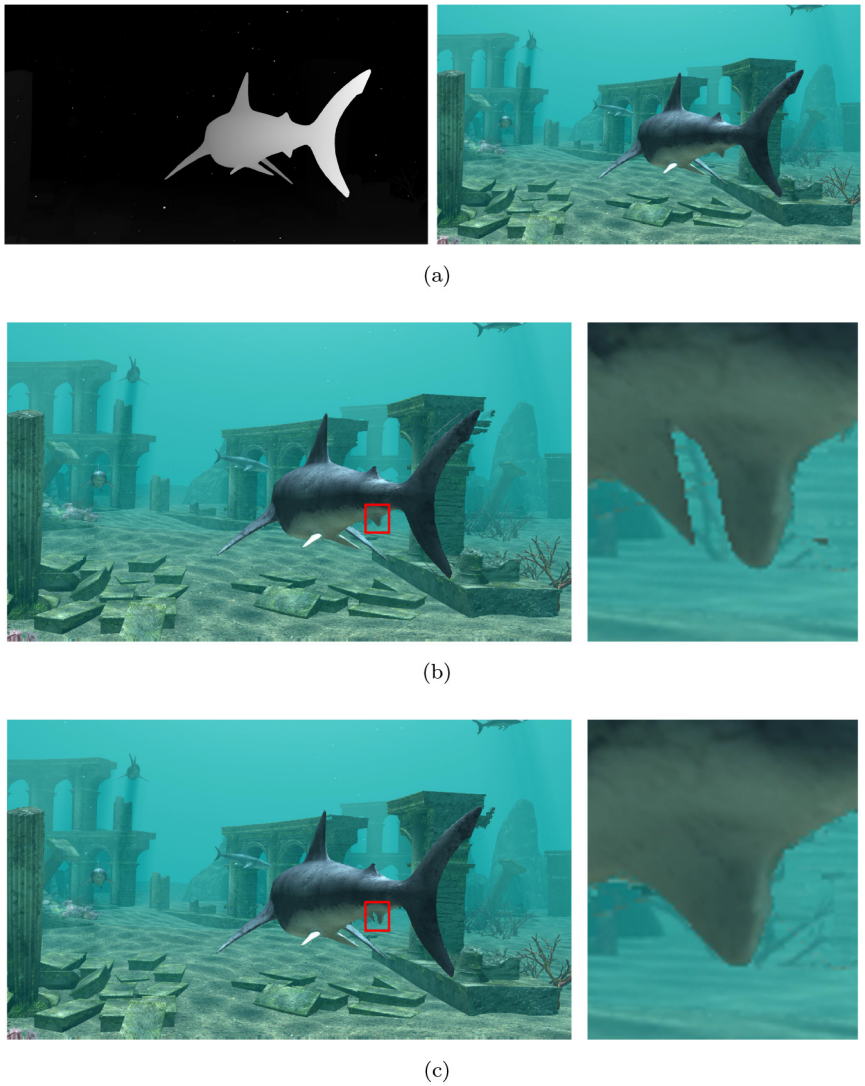

Fig. 7. (a) Reference view. Results comparison of disparity adjustments using (b) depth layering, (c) foreground segmentation.

(a) (b)

Fig. 8. Red-blue anaglyphs of adjacent views (intermediate to rightmost view from upper to lower row): (a) original views; (b) after disparity control. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

According to the rule that, in practice, the patch size is expected to be larger than distinguishable texture element [29], the optimal patch size is set as 9×9 for *Baby, Art, Reindeer*, 15×15 for *Ballet*, and 11×11 for Lovebird1 in JTDI, Ahn's and our method. The overall numerical results are presented in Tables 1 and 2. These results demonstrate that the proposed method yields better results than the other methods in both metrics. For the scenario *Baby*, our method surpasses the VSRS, Wang's, JTDI and Ahn's method by $7.89\%, 3.35\%, 1.67\%$ and 3.04% in terms of PSNR, respectively. From the perspective of SSIM, our method also produces better results, promoting the value of SSIM by $2.18\%, 0.65\%, 0.29\%$ and 0.51%, respectively. Likewise, there is evident promotion in *Art, Ballet* and Lovebird1. For the sequence *Reindeer*, the result of Ahn's method is approximate to ours and the SSIM value is slightly higher. Yet our method performs significantly better than VSRS and JTDI in *Reindeer*. We also observe that the objective quality of our method is slightly lower than Wang's method in the sequence *Reindeer*. The loss is due to the reference image's characteristic background: (i) the holes in the reference image are mostly surrounded by smooth background, so it appears consistent and natural when Wang's approach filled the missing regions with background pixels. It implies that the diffusion performs effective filling for low-structured texture background; (ii) The out-of-field area is just along the edge of the right box where there is a sharp color change. Our method filled it with inconsistent texture due to no similar information can be found in the source region.

6.5. Qualitative analysis

Fig. 9 depicts the visual quality of the virtual view generated by the proposed method in comparison with the three reference methods. The ground truth is shown in Fig. 9a, with two particular patch examples selected for illustration. On the edge of foreground objects, depth maps are usually not aligned with the color images due to inaccurate sampling and estimation. It leads to artifacts in exemplar-based inpainting, as illustrated in the first patch of the first row in Fig. 9d, e and f, which appear as a shadow of a few-pixel-width foreground. It also results in erroneous diffusion in Wang's method. As shown in Fig. 9c, the holes are supposed to be made up with the background, but partly filled by the foreground pixels by mistake. Besides, there are blurring artifacts in Fig. 9b and c, especially in the complex background. The incorrect fillings in the first, fourth and last rows of Fig. 9d might be caused by the disorder in priority. And JTDI suffers from the translucent cracks such as in the last two rows in Fig. 9d when the baseline distance is increased. The magnified parts of Fig. 9e shows that Ahn's method sometimes yields relatively inconsistent patches. Overall, it is observed that our algorithm better propagates texture and structure from background regions. In most cases, the synthesized textures in our method look more natural than those in JTDI and Ahn. In spite of inaccurate texture such as in the second row of Fig. 9f, the overall visual perception is acceptable and pleasing.

Table 1
PSNR Comparison for synthesized images (in dB).

	VSRS	Wang	JTDI	Ahn	Proposed
Baby v3 → v4	30.7024	32.0546	32.5809	32.1448	**33.1235**
Art v3 → v4	24.2976	27.5442	26.3946	27.1714	**27.8548**
Reindeer v3 → v4	26.6285	**30.0145**	27.0506	29.8121	29.9126
Ballet v5 → v6	25.3802	25.5850	26.5894	27.7599	**28.1823**
Lovebird1 v6 → v8	24.5695	24.3489	24.4970	23.8906	**24.7448**

Values in bold indicates the highest scores.

Table 2
SSIM Comparison for synthesized images.

	VSRS	Wang	JTDI	Ahn	Proposed
Baby v3 → v4	0.9558	0.9703	0.9738	0.9716	**0.9766**
Art v3 → v4	0.8781	0.9325	0.9229	0.9362	**0.9421**
Reindeer v3 → v4	0.9217	0.9653	0.9380	**0.9658**	0.9641
Ballet v5 → v6	0.8293	0.8827	0.8824	0.8977	**0.9046**
Lovebird1 v6 → v8	0.8883	0.8844	0.8858	0.8845	**0.8900**

Values in bold indicates the highest scores.

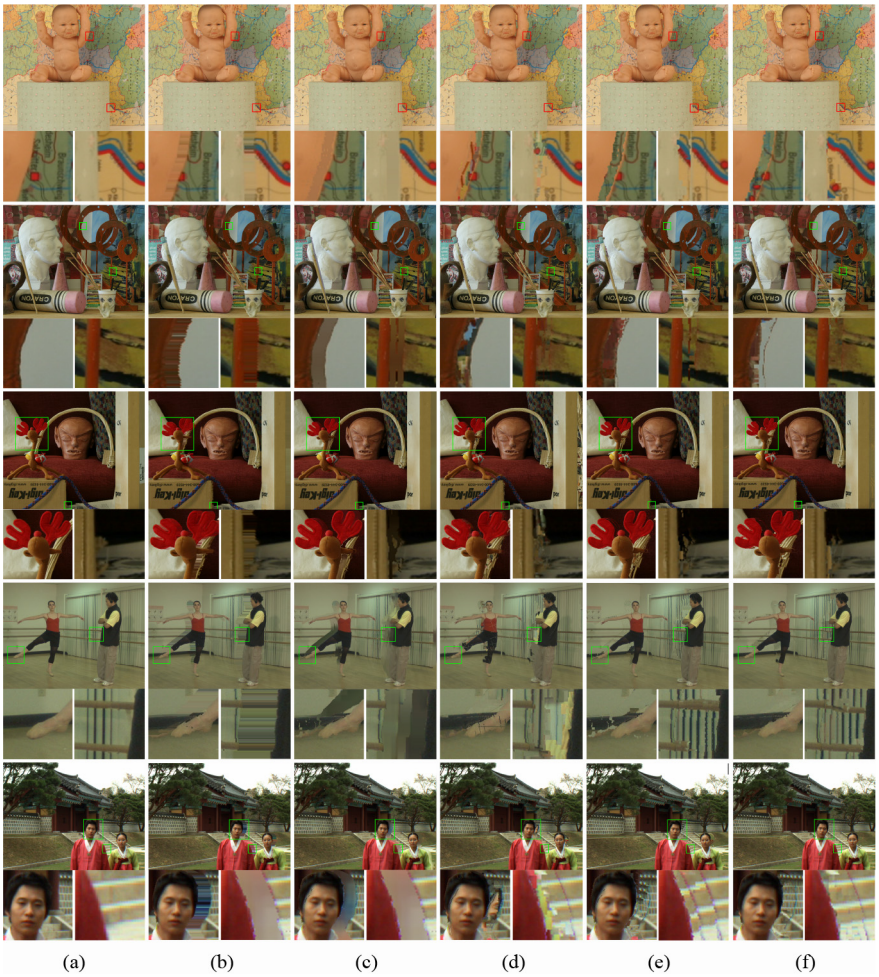

Fig. 9. Visual evaluation of synthesized images. (a) Original target view, (b) VSRS 3.5 [12], (c) Wang et al. [17], (d) JTDI [21], (e) Ahn et al. [22], (f) the proposed method.

7. Conclusions

This paper presents a new scheme of virtual view synthesis for multiview display system. We perform layered 3D warping after effective foreground object segmentation. Translucent cracks are identified and removed by morphological method. Then the proposed disparity control method alleviates the dazzling discomfort on the border of viewing zones and find a balance with stereoscopic perception and visual comfort. Semantically meaningful segmentation facilitates the disparity adjustments of foreground objects and background. Moreover, an improved exemplar-based inpainting is applied to fill the disocclusions. In the experiments, we adopted three different methods to validate the proposed method. The results demonstrate that our scheme surpasses the references in image quality. There is still some limitation that the missing regions are possibly erroneously filled when similar patches within the maximum searching radius are not available. Besides, the seed points are specified manually to locate the foreground objects. Our future work will focus on the automatic extraction of salient regions. Machine learning techniques have proved successful in image analysis and object detection [35,36], which can offer inspiration for this purpose.

Conflict of interest

There is no conflict of interest.

Acknowledgement

This work is supported by the Key Project of National Natural Science Foundation of China under Grant No. 61332015, and the Natural Science Foundation of Shandong Province of China under Grant Nos. ZR2013FM302 and ZR2017MF057. Thanks also goes to Dr. Weiping Huang and the Foundation of Hisense.

References

[1] W. Matusik, H. Pfister, 3d tv: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes, ACM Trans. Graph. (TOG) 23 (3) (2004) 814–824.
[2] N.A. Dodgson, Autostereoscopic 3d displays, Computer 38 (8) (2005) 31–36.
[3] Y. Chen, M.M. Hannuksela, T. Suzuki, S. Hattori, Overview of the MVC+ d 3d video coding standard, J. Vis. Commun. Image Represent. 25 (4) (2014) 679–688.
[4] C. Yan, Y. Zhang, J. Xu, F. Dai, L. Li, Q. Dai, F. Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors, IEEE Signal Process. Lett. 21 (5) (2014) 573–576.
[5] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient parallel framework for HEVC motion estimation on many-core processors, IEEE Trans. Circuits Syst. Video Technol. 24 (12) (2014) 2077–2089.
[6] U. Celikcan, G. Cimen, E.B. Kevinc, T. Capin, Attention-aware disparity control in interactive environments, Visual Comput. 29 (6) (2013) 685–694.
[7] S. Mangiat, J. Gibson, Disparity remapping for handheld 3d video communications, in: IEEE International Conference on Emerging Signal Processing Applications, 2012, pp. 147–150.
[8] H. Sohn, J.J. Yong, S.I. Lee, F. Speranza, M.R. Yong, Visual comfort amelioration technique for stereoscopic images: Disparity remapping to mitigate global and

local discomfort causes, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2014) 745–758.
[9] Y. Wang, M. Yu, H. Ying, G. Jiang, Visual comfort enhancement for stereoscopic images based on disparity remapping, J Image Graph 22 (4) (2017) 452–462 (in Chinese).
[10] C. Fehn, Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv, in: Electronic Imaging, 2004, pp. 93–104.
[11] J. Lei, S. Li, B. Wang, K. Fan, C. Hou, Stereoscopic visual attention guided disparity control for multiview images, J. Display Technol. 10 (5) (2014) 373–379.
[12] M. Gotfryd, K. Wegner, M. Domański, View synthesis software and assessment of its performance, ISO/IEC JTC1/SC29/WG11 MPEG/M15672.
[13] M. Kiechle, S. Hawe, M. Kleinsteuber, A joint intensity and depth co-sparse analysis model for depth map super-resolution, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 1545–1552.
[14] J. Shen, S.-C.S. Cheung, Layer depth denoising and completion for structured-light rgb-d cameras, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 1187–1194.
[15] Y.J. Chang, Y.S. Ho, Disparity map enhancement in pixel based stereo matching method using distance transform, J. Vis. Commun. Image Represent. 40 (2016) 118–127.
[16] Y. Tian, Y. Xian, Resolution enhancement in single depth map and aligned image, in: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1–9.
[17] L. Wang, C. Hou, J. Lei, W. Yan, View generation with DIBR for 3d display system, Multimedia Tools Appl. 74 (21) (2015) 9529–9545.
[18] N.A. Manap, J.J. Soraghan, Novel view synthesis based on depth map layers representation, in: 3dtv Conference: the True Vision – Capture, Transmission and Display of 3d Video, 2011, pp. 1–4.
[19] M.P. Tehrani, T. Tezuka, K. Suzuki, K. Takahashi, T. Fujii, Free-viewpoint image synthesis using superpixel segmentation, APSIPA Trans. Signal Inform. Process. 6 (2017) e5.
[20] I. Daribo, B. Pesquet-Popescu, Depth-aided image inpainting for novel view synthesis, in: Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on, 2010, pp. 167–170.
[21] S. Reel, G. Cheung, P. Wong, L.S. Dooley, Joint texture-depth pixel inpainting of disocclusion holes in virtual view synthesis, in: Signal and Information Processing Association Summit and Conference, 2013, pp. 1–7.
[22] I. Ahn, C. Kim, A novel depth-based virtual view synthesis method for free viewpoint video, IEEE Trans. Broadcast. 59 (4) (2013) 614–626.
[23] J. Arai, E. Nakasu, T. Yamashita, H. Hiura, M. Miura, T. Nakamura, R. Funatsu, Progress overview of capturing method for integral 3-d imaging displays, Proc. IEEE 105 (5) (2017) 837–849.
[24] C.H. Xiao Z, T. C, An effective graph and depth layer based rgb-d image foreground object extraction method, Comput. Visual Media 3 (4) (2017) 387–393.
[25] S.M. Muddala, M. Sjöström, R. Olsson, Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions, J. Vis. Commun. Image Represent. 38 (2016) 351–366.
[26] H. Yamanoue, The differences between toed-in camera configurations and parallel camera configurations in shooting stereoscopic images, in: IEEE International Conference on Multimedia and Expo, 2006, pp. 1701–1704.
[27] W.J. Tam, F. Speranza, S. Yano, K. Shimono, H. Ono, Stereoscopic 3d-tv: Visual comfort, IEEE Trans. Broadcast. 57 (2) (2011) 335–346.
[28] T. Shibata, J. Kim, D.M. Hoffman, M.S. Banks, The zone of comfort: Predicting visual discomfort with stereo displays, J. Vision 11 (8) (2011) 11.
[29] A. Criminisi, P. Perez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Trans. Image Process. 13 (9) (2004) 1200–1212.
[30] National institute of information and communications technology, ftp://ftp.merl.com.
[31] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, ACM Transactions on Graphics (TOG), vol. 23, ACM, 2004, pp. 600–608.
[32] G. Um, G. Bang, N. Hur, J. Kim, Y. Ho, 3d video test material of outdoor scene, ISO/IEC JTC1/SC29/WG11.
[33] D. Scharstein, C. Pal, Learning conditional random fields for stereo, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1–8.
[34] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[35] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai, Effective uyghur language text detection in complex background images for traffic prompt identification, IEEE Trans. Intell. Transport. Syst. 19 (1) (2018) 220–229.
[36] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised hash coding with deep neural network for environment perception of intelligent vehicles, IEEE Trans. Intell. Transport. Syst. 19 (1) (2018) 284–295.