PAPER

A Robust Depth Image Based Rendering Scheme for Stereoscopic View Synthesis with Adaptive Domain Transform Based Filtering Framework*

Wei LIU[†a)], ***Nonmember***, **Yun Qi TANG**[††], ***Member***, **Jian Wei DING**[††], ***and*** **Ming Yue CUI**[†], ***Nonmembers***

SUMMARY Depth image based rendering (DIBR), which is utilized to render virtual views with a color image and the corresponding depth map, is one of the key procedures in the 2D to 3D conversion process. However, some troubling problems, such as depth edge misalignment, disocclusion occurrences and cracks at resampling, still exist in current DIBR systems. To solve these problems, in this paper, we present a robust depth image based rendering scheme for stereoscopic view synthesis. The cores of the proposed scheme are two depth map filters which share a common domain transform based filtering framework. As a first step, a filter of this framework is carried out to realize texture-depth boundary alignments and directional disocclusion reduction smoothing simultaneously. Then after depth map 3D warping, another adaptive filter is used on the warped depth maps with delivered scene gradient structures to further diminish the remaining cracks and noises. Finally, with the optimized depth map of the virtual view, backward texture warping is adopted to retrieve the final texture virtual view. The proposed scheme enables to yield visually satisfactory results for high quality 2D to 3D conversion. Experimental results demonstrate the excellent performances of the proposed approach.

key words: *depth image-based rendering, domain transform, adaptive smoothing, stereoscopic image generation*

1. Introduction

Three-dimensional (3D) films are becoming popular because of their higher realism over the customary two-dimensional (2D) ones. The thriving of 3D industry has prompted a growing demand for 3D contents. The conventional approach of constructing stereoscopic pictures utilizes two or multiple cameras [1] to capture at least two streams of images and transmit them to receivers. In fact, the most materials available for 3D TV broadcast today have been created in this way. Nonetheless, such factors as costly hardware, shot planning and stereoscopic camera control, render this procedure both unwieldy and costly.

Depth image based rendering (DIBR) techniques, which just utilize one color image and the related per-pixel depth information to produce virtual image of other views, is considered as an another solution by the European Information Society Technologies (IST) project “Advanced Three-Dimensional Television System Technologies” (ATTEST) [2]. The depth image is a grey scale image in which each pixel demonstrates the related depth value of the real scene. Depth image is also called depth map. There are numerous approaches to produce a depth map. Depth map can be caught by active methods with range devices such as Zcam [3], which measures the depth of a scene by recording the time of flight. Compared with a passive stereo camera method, a depth camera can deal with complex conditions more effectively. In another solution, the depth camera is substituted by a 2D to 3D converter, where depth information is obtained from a monoscopic frame sequence with computer vision algorithms [4]. This solution does not need the depth camera. More importantly, it helps to reuse the existing libraries of 2D contents. Accordingly, this DIBR-based method enables to solve the bottleneck of 3D contents creation. If the depth map is available, DIBR frameworks can create any number of virtual views without multi-camera systems, therefore the equipment cost of 3D cinema frameworks is decreased.

The DIBR framework has several advantages over the current 3D broadcasting framework. The transmission bandwidth required by the DIBR framework can be decreased no less than 33% in contrast with that of regular broadcasting framework [5]. Another advantage is that the parallax of generated virtual view images can be customized by viewers to accomplish different depth effects and 3D perceptions. Theoretically, the DIBR techniques can be used to synthesize any virtual views from the color image and the depth map. But there still exist some disturbing issues, such as disocclusion, resampling and boundary artifacts in the virtual view synthesis process.

The disocclusion is the most difficult issue in the DIBR process. A fundamental problem of DIBR is that all pixels in the virtual view do not necessarily exist in the original image. Because of sharp horizontal changes in the depth map, areas that are occluded in the first view may be revealed by image warping and become visible in other views. To deal with this issue and accomplish high quality 3D, these gaps ought to be filled. Lots of methods have been proposed to deal with the problem. Significant contributions comprise of two following classes. One answer

Manuscript received February 20, 2018.
Manuscript revised July 6, 2018.
Manuscript publicized August 31, 2018.

[†]The authors are with the College of Electromechanic Engineering, Nanyang Normal University, Nanyang Henan 473061, China.
[††]The authors are with the People’s Public Security University of China, Beijing 100038, China.

[*]This work was supported in part by National Natural Science Foundation of China (No. 61503202), Key scientific and technological project of Henan Province (No. 182102310760), Key scientific and technological project of Nanyang City (No. KJGG28).

a) E-mail: lw3171796@163.com

DOI: 10.1587/transinf.2018EDP7069

for filling the disoccluded regions is by picture inpainting methods to close holes after they have come in a texture picture of post-processing step [6]. This solution achieves hole-filling with information around predicted hole areas by either structure continuation or texture replication after DIBR. Since an exhaustive discussion of inpainting is out of the scope of this paper, readers can refer to a summary about these approachs [7]. Yet inpainting is a tough work, and it is more difficult with stereoscopic content as image features need to have consistent disparity across the two generated views. Generally speaking, for these methods, the computational cost would increase with the number of disoccluded regions. The other strategy is accomplished in pre-processing procedure, which reduces depth data discontinuities by filtering the depth maps. In this manner, holes are diminished in the first case rather filled later. This strategy mainly uses depth map filtering before DIBR to reduce hole regions. Previous study results [8] show that smoothing is helpful to lessen the level of disoccluded regions for depth maps which are required to be filled in the rendering procedure. To diminish sharp discontinuities from depth map image, Tam et al. [9] use symmetric Gaussian filter to preprocess depth maps. In this manner, the artifacts around boundary lines are diminished after warping procedure. Yet, the symmetric strategy may introduce distortions in some regions relying upon the depth information around, particularly for that over vertically straight boundary lines. Several approaches are proposed to solve the problem, for example, asymmetric smoothing filter [10] and some edge-oriented methods [11], [12]. The asymmetric filter keeps stronger smoothing along vertical direction, while the edge-oriented methods improve the smoothing effects by limiting the filtered regions. Even so, in the severe depth discontinuities of the filtered regions, geometric distortions are yet inevitable. In addition to this, for edge-oriented methods, the filters mainly focus on the minimized filtered regions around. This may bring about a few gaps which are far from areas being disregarded. So an additional hole filling procedure is always required.

The second problem we have to overcome is cracks which are introduced through resampling in the process of DIBR. The phenomenon of an integer pixel position in the reference view image being projected to a subpixel position in the virtual view is called resampling problem. In other words, the content in the planar scene is discrete, whereas the actual projection from objects in the real world to such a planar scene is continuous and may often fall at sub-pixel locations. This phenomenon appears after 3D warping, so image inpainting techniques in post-processing steps can be used to solve this problem. But as discussed above, inpainting techniques always need more computational costs than the depth map preprocessing methods. Instead, for more efficiency in the 2D to 3D conversion process, the cracks of this type can be coped with upsampling procedure or much more advanced algorithm such as backwards warping with interpolation [13].

The last problem, how to reduce boundary artifacts in the synthesized virtual view, is still a tough challenge. Visual quality of a synthesized view is critical in 3D TV systems. However, current depth estimation techniques in the 2D to 3D conversion process always introduce complex texture-depth misalignment, thus yielding annoying boundary artifacts in turn. Recent study on the artifacts of the DIBR view synthesis [14] found that the major cause of the annoying artifacts is the misalignment of the object boundaries between the depth map and the corresponding color image. In practice, the depth maps estimated by various cues, especially the automatic 2D to 3D conversion schemes [15], may not align with the corresponding color image correctly in limited constraints when estimating the stereoscopic information with just incomplete 2D cues. As a result, unprocessed depth map usually causes annoying artifacts after the stereoscopic rendering process. Unfortunately, the conventional depth map smoothing and hole filling methods cannot eliminate these artifacts well. So some additional correction steps, for example, calculations based on image or video segmentation [16] are constantly used to correct the misaligned edges for an improved depth map. However, these strategies will cost more computational times inevitably. Depth map refinement algorithms such as [17] and [18] can also be adopted. While it should be noted that in 2D to 3D conversion, besides noise reduction and boundary recovery, when optimizing a depth map, other specific factors such as disocculusion occurrences should to be considered simultaneously. The proposed methods in this paper can realize all these requirements in a united framework.

To better deal with the problems of disocclusion occurrences, resampling cracks and boundary artifacts in the DIBR process, this paper proposes a robust depth image based rendering scheme for stereoscopic view synthesis. Compared to other similar works, the main novel advantages of our works are:

- The presented scheme is based on domain transform and backward texture warping. Firstly, two domain transform based filters are conduced and used sequentially both before and after the depth map 3D warping to realize texture-depth boundary alignments, and meanwhile, directional smoothing for avoiding disocclusion. Then, with the optimized depth map on the virtual view, backward texture warping is adopted to retrieve the final texture virtual view. As discussed above, a backward DIBR method can effectively tackle the cracks due to resampling problem. So thanks to the well designed scheme, our method can deal with all three problems in one united scheme simultaneously and efficiently.
- The main idea of our smoothing strategy in this paper is to apply two different adaptive smoothing to the processed depth maps sequentially. Yet, in this work the two smoothing operations share a common domain transform based filtering framework. By adding different constraints, the smoothing strategy can be adjusted

adaptively to realize expected process effects at different steps. The proposed domain transform based filtering framework uses a dimensionality-reduction procedure and has some desirable advantages, such as great speed-ups over existing methods and computational cost not influenced by different filter parameters. So it is not only more efficient, but also more flexible than other similar works when hybrid constraints are considered. In this way, our method makes good performance both on computational cost and virtual view quality, and is more suitable for the 2D to 3D conversion applications.

The rest of the paper is organized as follows. Firstly, the technical scheme of the approach and the proposed domain transform based filtering framework will be introduced. Then, each step of the scheme, especially the two adaptive depth map filters at different steps, will be discussed respectively and elaborately. Finally the experimental results are reported and some concluding remarks are given. Experiments and comparisons show that our approach is excellent at both time saving and virtual view quality.

Note that in this paper, we made focus on stereoscopic view generation in the application of traditional 2D to 3D conversion process. As general 2D to 3D image conversion cannot depend on depth estimation with multiple cameras or a priori provided depth maps, depth has to be computed from a single view. Not any more complex multi-dimensional information just as that from the LDI format [19] can be utilized, so it is harder to deal with the problems mentioned above. This paper expects that only a texture image and its corresponding depth map image, which has been generated with some depth cues, are given. Study of stereoscopic view generation for videos would be performed in our further work.

2. The Proposed Approach

2.1 Framework of the Proposed Scheme

In this paper, based on the previous work [20], we further introduce the domain transform framework from the traditional forward warping scheme of DIBR to a new backward warping strategy. The backward warping scheme has been found to be more favorable than the traditional one if the warped depth map in the target view is available [21]. This can be realized easily by projecting each pixel in the source depth map to the destination with traditional DIBR systems. But as discussed above, several problems still remain. Even so, it is possible to handle the issues better for the warped depth map than for the warped texture image in the proposed scheme. Because from observation we can see that the characteristics of depth map and natural color image are very different. While texture images have many different patterns, depth maps have an obvious dual structure which comprises large homogeneous regions within scene objects and sharp changes at object boundaries [13]. Based on these characteristics, the filtering operation to a depth map has a minimal adverse effect and it is feasible to deal with the problems in the backward warping DIBR systems through specific filtering strategies to depth maps. In this paper, we propose a scheme to realize this thought and focus on depth map adaptive filtering at different steps. As discussed above, two smoothing operations, which share a common domain transform based filtering framework, are adopted in our scheme. And in the following, we will discuss this framework in details. Firstly, we will discuss the common domain transform based filtering framework.

Domain transform [22] uses a transform strategy to warp the input signal adaptively so one 2D signal can be filtered by iterating 1D-filtering operations in the row or column order efficiently rather than filtering it in a two dimensional manner, which is the main cause of heavy computational burden. Taking these advantages into consideration, domain transform has been used for some real-time image or video processing tasks efficiently and successfully, just like colorization, tone mapping, etc. In this paper, we try to build a framework using this technique in the application of backward warping DIBR systems.

The common filter framework for multichannel form of domain transformation can be indicated as:

$$ct(u) = \int_0^u 1 + \sum_{k=1}^{c} |I'_k(x)|\, dx \tag{1}$$

where $I_k(x)$ expresses different channel of the input signal. The domain transform can be further expressed in another form if we take the input's space and range into considerations.

$$ct(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \sum_{k=1}^{c} \frac{\sigma_H}{\sigma_{r_k}} |I'_k(x)|\, dx \tag{2}$$

where the parameters $\sigma_H, \sigma_s, \sigma_{r_k}(k = 1, \ldots c)$ are used to change the filtering effect adaptively. After the domain transform operation, we need design some filters in the transformed domain to perform image filtering. In this paper, hybrid constraints at different processing steps are viewed as different multiple channels of the input signal, and which will be explained deeply in each following filters.

The block diagram of the proposed scheme is shown in Fig. 1. It mainly consists of three processing steps. Firstly, a domain transform based filter with related constraints is carried out to fulfill the requirement of depth map boundary correction and disocclusion reduction smoothing simultaneously. Then, after the preprocessed depth maps are warped to the virtual viewpoint, another domain transform based filter, which share a common framework as the previous one, with another constraints delivered from the warped gradient scene structures is adopted to further optimize the projected depth map. Finally the virtual image is retrieved by performing an inverse warping from the optimized depth map to the reference image.

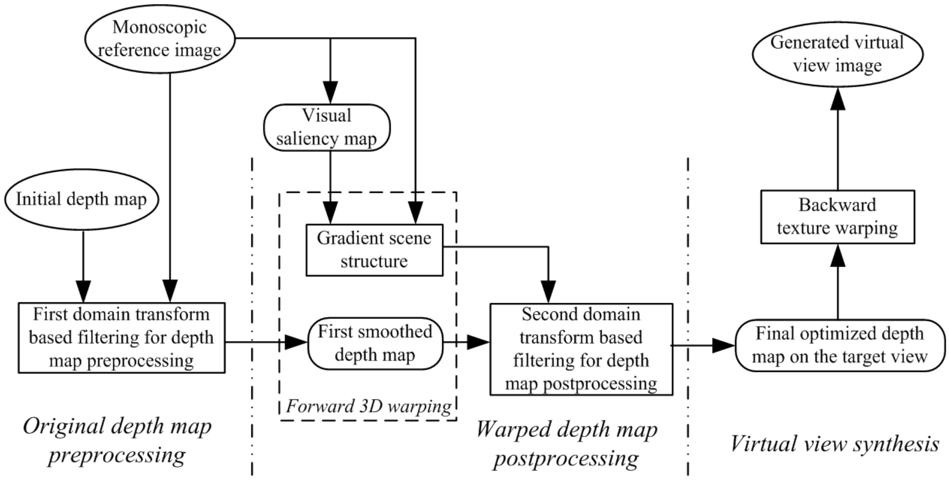

Monoscopic reference image
Visual saliency map
Generated virtual view image
Initial depth map
Gradient scene structure
Backward texture warping
First domain transform based filtering for depth map preprocessing
First smoothed depth map
Second domain transform based filtering for depth map postprocessing
Final optimized depth map on the target view
Forward 3D warping
Original depth map preprocessing
Warped depth map postprocessing
Virtual view synthesis

Fig. 1 Flow chart of the proposed depth image based rendering scheme for stereoscopic view synthesis.

In this paper, domain transform based filtering technique is used to adaptively process a depth map at different steps. To make more related cues into considerations for different filters, in this framework, a pixel of an image is viewed in a high dimensional space, which is formed by coordinate, pixel texture, depth or scene structure, etc. Then in this paper, the extended domain transform with a recursive form can be defined as:

$$D_{k+1}[n] = (1 - \lambda^d)D_k[n] + \lambda^d D_{k+1}[n-1] \tag{3}$$

where $D_k[n]$ is a original pixel value in a row or column of depth map to be processed, and k indicate the iteration number. The smoothing is processed multiple times, and in each time firstly the filter is performed for a horizontal pass along each image row, and then a vertical pass along each image column. It should be noticed that to achieve symmetric impulse response, Eq. (3) is carried out with the positive order (left-to-right, top-to-bottom) for one iteration, then in the next turn with the reverse order (right-to-left, bottom-to-top). The total number of k relays on the image content. Four iterations are set in this paper. $d = ct(x_n) - ct(x_{n-1})$ indicates the smooth strength control factor between neighbor samples x_n and x_{n-1} in the transformed domain. $\lambda = e^{\frac{-\sqrt{2}}{\sigma_H}}$ is the feedback coefficient of the filter [22], where σ_H is the standard deviation of the designed filter. Since $\lambda \in [0, 1]$, the filter is stable, and has a complexity of $O(N)$.

In Eq. (3) smooth strength, which is delivered from transformed domain to target depth maps, is controlled by d. If d increases, λ^d decreases and goes to zero, so the propagation chain is stopped. In this way, we can use the related constraints which have been defined in the transformed domain to adjust smoothing effect adaptively. For notion convenience, the filter in the transformed domain can be further expressed as:

$$ct_{First/Second}(u) = \sum_m ct_m(u) \tag{4}$$

where $ct_m(u)$ denotes different constraints added to each filter in the transformed domain. In this paper, the filter used before depth map forward warping is denoted as ct_{First}, and the other one used after the depth map forward warping is denoted as ct_{Second}. We will discuss the details elaborately in the each following step.

2.2 Original Depth Map Preprocess Smoothing: First Smoothing Filter

With imperfect depth maps, the conventional view synthesis suffers from rendering artifacts, especially at object boundaries. Generally, incorrect depth values will render the associated pixels to wrong positions in the virtual view. In this part, as a first step, structure-aided related constraints are added to the designed domain transform based filter to align depth discontinuities in the original depth map to color discontinuities in the textured image and to further reduce estimation errors in the depth map.

$$ct_{First}(u) = ct_1(u) + ct_2(u) \tag{5}$$

$$ct_1(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \frac{\sigma_H}{\sigma_r}\left|I'_{ori}(x)\right| dx \tag{6}$$

$$ct_2(u) = \int_0^u \sigma_c e^{-\alpha I_e(x)} dx \tag{7}$$

$$I_e(p) = \begin{cases} \varepsilon_1, & if \quad p \in E(D_{ori}) \cap E(I_{ori}) \\ 1, & otherwise \end{cases} \tag{8}$$

As shown in Eq. (5), these constraints of the first filter include two terms. The first term $ct_1(u)$ is simplified from the basic domain transform described in Eq. (2). In this paper, since anaglyph images are synthesized for 3D view, a gray scale image is used as the original texture image I_{ori}. The second term $ct_2(u)$ is intended to realize depth correction in the smoothing process. Besides that, we use asymmetric smoothing to further reduce geometric distortions. In the first domain transform filter, the size of $3 \times \sigma_s$ is adopted to enhance smoothing in space for each vertical pass along

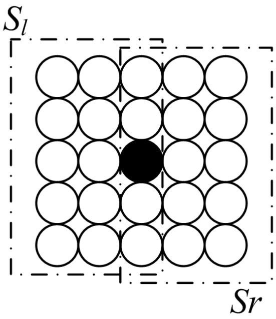


Fig. 2 Schematic diagram of window blocks around a pixel.

image columns.

As discussed above, inaccurate depth may cause boundary artifacts in the synthesized virtual view. Conventional 2D to 3D conversion methods usually estimated a scene depth from a single view. The depth cues used in these approaches are limited, so it is still hard to estimate accurate edge boundaries in depth maps, especially for the approaches with automatic processes with certain single cues [15]. The second term in this filter can help to deal with this problem. σ_c and α in Eq. (7) are positive parameters ,and they can be adjusted when the system is operating. $I_e(p)$ in Eq. (8) is used to describe the degree of accuracy between the estimated depth map and the corresponding texture image, where $E(p)$ is a edge extraction function, and D_{ori} indicates the original depth map to be smoothed. ε_1 is preset to 0.2 in this paper. A pixel p on one object edge of an accurate depth map should be also on one edge of the corresponding texture image. In that case, $I_e(p)$ gets a small preset value ε_1, then $ct_2(u)$ in Eq. (7) and $ct_{First}(u)$ in Eq. (5) would increase, and finally the filtering strength in Eq. (3) decreases. In this way, the original depth values are preserved. Contrarily, the false edges would be filtered to eliminate. As a result, the constraint in $ct_2(u)$ realizes depth correction.

Other than that, in this part, another cue is considered in the filter for directional smoothing. This cue originates from the observation that the newly generated hole areas are mainly distributed around one boundary side of the objects in the synthesized view. When the virtual view is a left view, they are along the left side. Conversely, for a right view, they are along the right side. This cue can further limit depth smoothing around the hole areas. Adding this constraint into the filter, the first term in Eq. (6) is modified as follows.

Suppose in the original depth image p is a pixel at location (x, y), then $N(p)$ indicates the neighborhood areas around p. Schematic diagram of window blocks for the neighborhood around p is showed in Fig. 2, which include a left block S_l with average depth value $V_{sl}(p)$ and a right one S_r with average depth value $V_{sr}(p)$. Then, the directional smoothing cue can be added to the filter and Eq. (6) is updated as:

$$ct_1(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \frac{\sigma_H}{\sigma_r}\beta W_{direction}^{-1} \left| I'_{ori}(x) \right| dx \tag{9}$$

$$W_{direction} = e^{sgn(p)\times(V_{sl}(p)-V_{sr}(p))} \tag{10}$$

where β is the directional reference factor, and is set to be 1 in this paper. $sgn(p)$ is a pre-defined symbolic function. In this paper, for a right virtual view, $sgn(p) = 1$, and for a left one, $sgn(p) = -1$.

Since the distribution of the newly exposed holes is directly related to direction of the synthesized visual view, the updated filter can increase the strength adaptively on one side of the objects in the depth map where the disocclusion mainly occurs. Through this way, the disocclusion problem in the generated virtual view will be overcome more effectively.

2.3 Warped Depth Map Postprocess Smoothing: Second Smoothing Filter

Since our proposed depth image based rendering scheme is based on backward texture warping for stereoscopic view synthesis, depth maps of the target view are required. In this step, forward warping is firstly used to construct depth maps of the virtual view. Generally speaking, forward warping, which proceeds by projecting each pixel in the source image to the destination, always suffer from disoccluded holes and rendering artifacts due to incorrect depth values. However the preprocess smoothing in the last step eliminates most boundary noises and closes large disoccluded holes. To deal with the remaining small holes and cracks at resampling in the 3D warping, in this step, we conduct a second smooth operation to the warped depth maps, which is still based on the proposed domain transform framework.

The gradient scene structures warped from the original view are brought in as the related constraints to this filter, and which can be expressed as:

$$\begin{aligned} G_t &= \varphi_{warp}(G_{ori}) \\ G_{ori} &= \left| I'_{ori}(p) \right| + \gamma \left| S'_{ori}(p) \right| \end{aligned} \tag{11}$$

where φ_{warp} is a 3D warping operation, G_t denotes the warped gradient scene structure on the target view. The corresponding scene structure G_{ori} from the original view includes two parts. The first term is the texture structure I'_{ori}, which is the same as the first filter defined in Eq. (6). The additional term is the saliency structure S'_{ori}. In Eq. (11), $S_{ori}(p)$ is the generated saliency map of texture image $I_{ori}(p)$ with the existing state of the art method from [23]. γ is a user controllable weight to determine the effects between visual saliency structure and texture structure, and it is set to 2 in this paper.

Then, the second filter is realized as:

$$ct_{Second}(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \frac{\sigma_H}{\sigma_r} |G_t(x)| dx \tag{12}$$

where $ct_{Second}(u)$ denotes the constraint set of this filter in the transformed domain. The parameters $\sigma_H, \sigma_s, \sigma_r$ can be set the same as the first one.

Two points need to be illustrated:

- Firstly, the saliency structure constraint is only added to this filter rather than the first one. The reason behind is that more smoothes are needed for depth correction and depth boundary discontinuity diminution to avoid artifacts and disoccluded holes in the first step. While in this step, after 3D warping operations, only small rendering noises are left. So more structure-related constraints can be used to further suppose noises and optimize the warped depth maps.
- Secondly, it should be noted that unlike 3D warping for a whole image, the influence of the warping working on the gradient scene structures is limited. For gradient scene structures, the main available information is the discretely distributed scene layer boundary lines rather than large homogenous regions, so the disocclusion and occlusion phenomenon caused by adjacent region layers of different depth values can be avoided naturally.

2.4 Virtual View Synthesis with Backward Texture Warping

There are two basic approaches to image projection-forward warping and backward warping. Forward warping proceeds by projecting each pixel in the source image to the destination, while backward warping populates each pixel in the destination image by finding the corresponding pixel position in the source. Backward warping was found to be more favorable than forward warping in terms of handling the types of artifacts introduced in synthesized images [21]. In this step, a well optimized depth map of the target view is available after the previous processes. Then, texture information can be reconstructed via backward warping using the synthesized depth maps.

Suppose D_t denotes the final warped depth map of target view. A pixel $m(u, v)$ in D_t is reprojected to pixel $p(x, y)$ in image plane of original view to get the texture image of virtual view via interpolation operation. The texture image of virtual view I_t is calculated by:

$$I_t(m) = \frac{\sum\limits_{n=1}^{4} \omega_n I_{ori}(p_n)}{\sum\limits_{n=1}^{4} \omega_n} \tag{13}$$

where $I_{ori}(p_n)$ presents the intensity value of pixel (x_n, y_n) whose distance to (x, y) is less than one pixel either in horizontal or vertical direction. ω_n is the weight factor of distance, which is expressed as:

$$\omega_n = \frac{1}{\|p - p_n\|} \tag{14}$$

where $\|\cdot\|$ denotes the Euclidean distance.

3. Experimental Results and Discussions

In this section, some experiments are carried out to further evaluate the performance of the proposed approach. As shown in Table 1, five test sets, which cover a wide range of images with indoor and outdoor scenes, are used for evaluation. These test sets use different depth cues to generate the initial depth maps. And from the following experiments, we can see that in most situations, it is not easy to get accurate depth information directly from these automatically generated initial depth maps. The experiments were implemented on a commodity PC with an Intel Core2 Quad CPUQ9400 2.66GHz. We implemented our proposed scheme in Microsoft Visual Studio C++ 2008 platform combining this implementation with the domain transform module running in MATLAB. Experimental results are discussed in three subsections. The first subsection is designed to show the experimental details at each of the core steps. In the second subsection, our results are comprehensively compared with other similar works. In the third subsection, these results are further evaluated with subjective criteria. It should be pointed out that all constants in the models of this paper are found in experiments. The parameter presets for the domain transform filters used in testing are given in Table 2.

Table 1 Test sets used in the experiments

Test set	Spatial resolution	Depth cues used for initial depth map
Flower	328 × 242	Depth from motion [24]
Castle	768 × 576	Structure from motion [25]
Orbi	358 × 248	H.264/AVC estimated motion [26]
Desktop	644 × 472	Depth from focus [27]
Sculpture	476 × 350	Machine learning [28]

Table 2 Parameter presets for the domain transform filters used in testing

$ct_1(u)$	$ct_2(u)$
$\sigma_c = 10$ $\alpha = 1, \beta = 1$ $\varepsilon_1 = 0.2$	$\gamma = 2$
$\sigma_s = \sigma_H = 350, \sigma_r = 1$	

3.1 Analysis of Experimental Results and Depth Map Evaluation

Without loss of generality, for showing implementation details of our approach, the test sets *Flower* and *Castle* are mainly analyzed in the following. In this part, processing details of our approach are firstly displayed in Fig. 3. Images in Fig. 3 (b) are the corresponding initial depth maps of the original texture images shown in Fig. 3 (a). For test sets *Flower*, initial depth maps are generated by the cues from *Depth From Motion* [24], and for the second test sets *Castle*, initial depth maps are generated through simple *Delaney Triangulation* by the cues from *Structure From Motion* [25]. We can see that the generated initial depth maps were not perfect, especially around the transition regions where the depths become discontinuous. To eliminate these severe boundary noises, stronger smoothness can be adopted in depth preprocessing. However, some resulting side-effects, such as geometric distortions, may also increase, and that

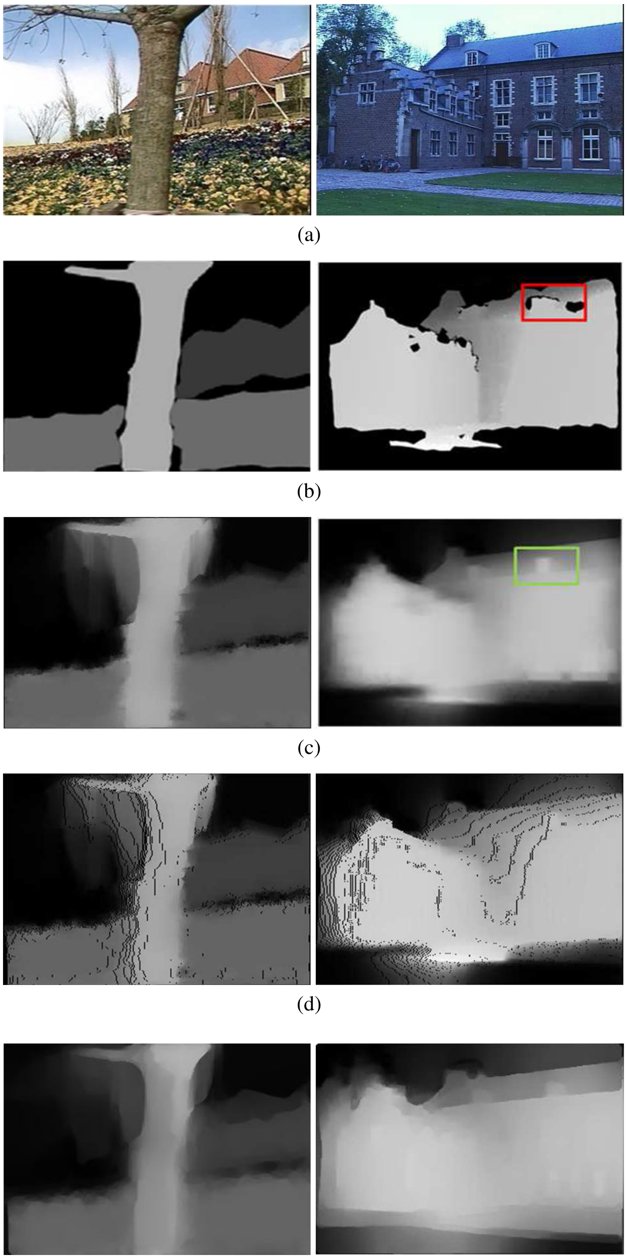


Fig. 3 Experimental results of processed depth maps with our proposed approach. Left: *Flower* set, right: *Castle* set. (a) Original texture images, (b) initial depth maps, (c) first filtered depth maps, (d) projected depth maps after 3D warping, (e) second filtered depth maps on the target views

can be seen in the following virtual view evaluation experiments. Experimental results in Fig. 3 (c) are the first filtered depth maps in our proposed scheme. As discussed above, in this step, on the one side, we diffuse initial depth maps by domain transform based filter with scene structure consistence constraints. It can be seen that the qualities of the processed depth maps have been improved greatly when comparing to the initial depth maps in Fig. 3 (b), especially for test set *Castle* as shown in the image parts color boxed. On the other side, the filter smoothes the depth maps directionally according to the estimated hole distributions. For both sets, left virtual views are needed, so the smooth operations strengthen especially on the left sides of object boundaries in the process. The warped depth maps are shown in Fig. 3 (d). We can see that there are still small holes and cracks left over. Figure 3 (e) illustrates the second filtered depth maps. Not only all the holes and cracks are closed or eliminated, but also more details are presented. Such as test set *Flower* shown in Fig. 3 (e), more depth layers can be seen in the background regions around the houses behind.

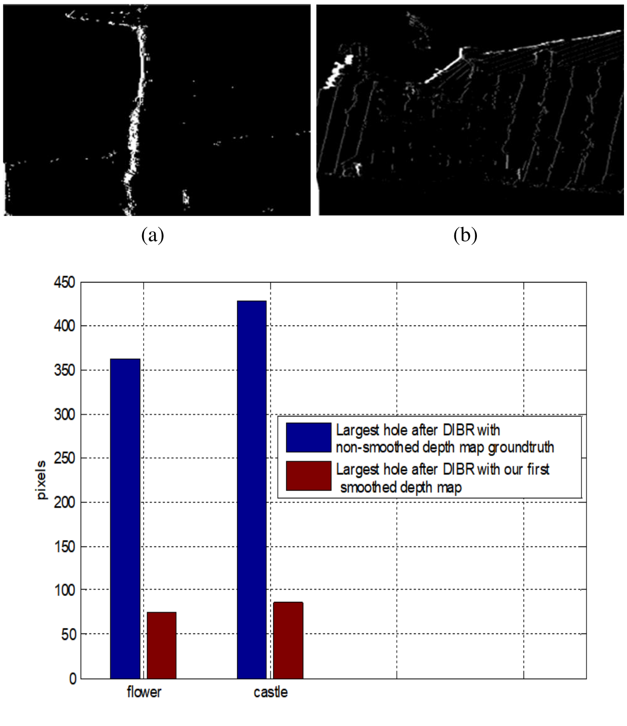


Fig. 4 Detected hole regions after 3D warping with non-smoothed depth map groundtruth for (a) *Flower* set and (b) *Castle* set, (c) comparison of generated largest holes after 3D warping with non-smoothed depth map groundtruth and our first smoothed depth maps

In this part, another experiment is carried out to further verify the disocclusion reduction ability of the proposed filter at depth map preprocess step. A novel approach in our previous works [29], which needs more computational costs but can generate more precise depth maps, is adopted to test sets *Flower* and *Castle* for new reference depth maps. These new reference depth maps are regarded as depth map groundtruth in this experiment. Figure 4 (a) and Fig. 4 (b) show the experimental results of detected hole regions after 3D warping with these non-smoothed new reference depth maps for the two test sets. Comparing with Fig. 3 (d), which illustrates the corresponding results with our first smoothed depth maps, we can see the obvious improvements. In the proposed approach, the depth map preprocessing allows the reduction of the number and especially the size of the disoccluded areas by smoothing depth discontinuities directionally with the first domain transform based filter. Small and

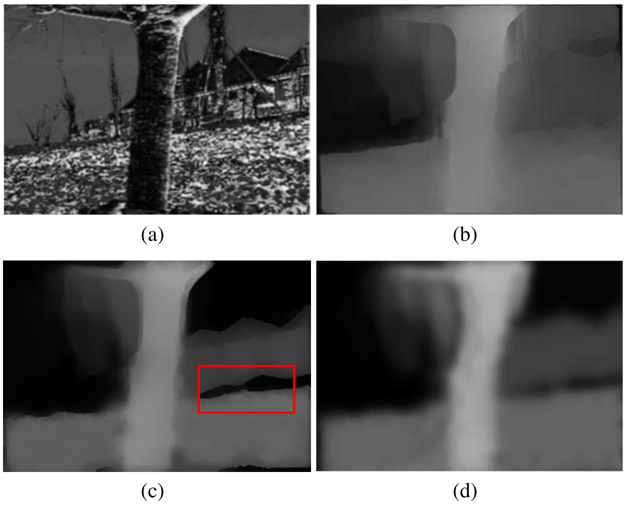


Fig. 5 Experimental results with different constraints in the proposed framework. (a) Saliency map for test sets *Flower*, (b) filtered depth map in *Test 1*, (c) filtered depth map in *Test 2*, (d) filtered depth map in *Test 3*

gradually changing of depth values does not give annoying artifacts because small depth discontinuities only generate very small holes in the warped virtual view. Figure 4 (c) gives a more directly quantitative comparison.

To better illustrate the effects of different constraints in our proposed framework, more experiments are implemented for test sets *Flower*, and which include:

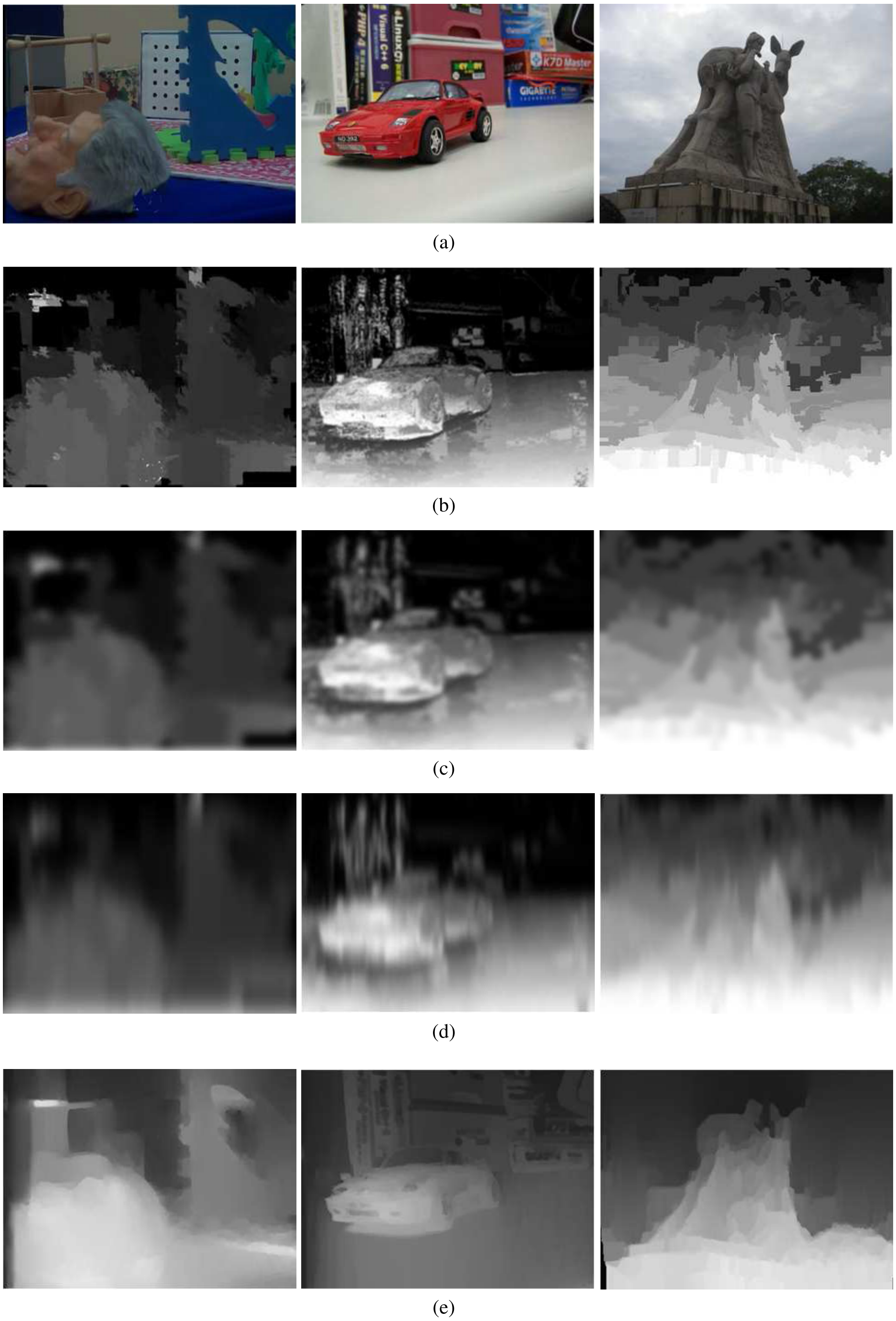
- *Test 1*: Proposed scheme without saliency structure constraints added in the second smoothing for warped depth
- *Test 2*: Combination of symmetric Gaussian for original depth and proposed second smoothing for warped depth
- *Test 3*: Combination of proposed first smoothing for original depth and symmetric Gaussian for warped depth

The experimental results are shown in Fig. 5. From Fig. 5 (a), it can be seen that the primary structures of a scene are highlighted in a saliency map. Without these constraints, the depth map is over-smoothed partially as shown in Fig. 5 (b). As we discussed in Sect. 2.2, the proposed first smoothing can not only reduce estimation errors in the depth map like a Gaussian filter, but also can align depth discontinuities in the original depth map to color discontinuities in the textured image simultaneously. It is substituted by a symmetric Gaussian filter in Fig. 5 (c), and we can see that the original imperfect parts in the red boxed region are not processed well when compared with the corresponding parts shown in Fig. 3 (e). In Fig. 5 (d), the proposed second filter is substituted by a symmetric Gaussian filter, and the experimental results are also not as good as the results in Fig. 3 (e).

Besides of above aforementioned intermediate results, for the other three test sets, the final depth maps filtered by the proposed scheme and some other smoothing methods are presented in Fig. 6. Similar to the results shown in Fig. 3, from Fig. 6 (b) it can be seen that the initial depth maps generated by various automatic 2D to 3D conversion methods suffer from depth edge misalignments of different levels. To figure out the problem, Gaussian filters in Fig. 6 (c) and Fig. 6 (d) have to increase the smoothing kernel globally. Unfortunately, these traditional smoothing strategies destroy not only the texture in the color image but also the geometrical depth information, degrading the 3D perception in the reconstructed 3D images. With the sequentially introduced domain transform based filters in our scheme, multiple constraints are considered adaptively instead of the unilateral one. Comparing with the results in Fig. 6 (c) and Fig. 6 (d), great improvements can be seen in Fig. 6 (e). Without stronger smoothing, the depth maps generated with our scheme still display better visual quality with more perfect texture-depth alignments. The reason behind is that, in the first smoothing, special constraints are added to correct the depth maps. Besides that, after the first smoothing with our filter, sharp depth edges are reduced greatly, then the alterations of depth structures in the first filtered depth maps can further lead to the alleviation of hole occurrences. For these reasons, the smooth strengths in the second filtering are limited adaptively. As a next step towards a more comprehensive evaluation, the visual qualities of the synthesized virtual view images for these test sets are displayed in the following.

3.2 Virtual View Evaluation

For image quality evaluation, virtual view images are illustrated in Fig. 7. From these results, several observations can be made. First, for scene only containing simple structures, such as test set *Sculpture*, it seems that both the traditional Gaussian filtering methods and our scheme can gain similar satisfactory visual effects on the virtual view. But it should be noted that our approach rather extends the range of distortion in depth maps and can provide more immersive 3D viewing when a user experiences 3D service. These factors are more important for virtual view quality, while they cannot be demonstrated directly in this experiment. The subjective evaluation in the next section will verify the observation. Second, our approach emerges its advantages when dealing with the other two test sets, which contain more complex scene structures. The most significant improvement is found for test set *Desktop*. As shown in the red rectangles of Fig. 7 (a), severe geometric distortions are introduced along the bookshelf edges vertically when using symmetric Gaussian filter with only unilateral smoothing kernel. In the red rectangle of Fig. 7 (b), asymmetric Gaussian filter with separate bilateral smoothing kernels can alleviate the problem. However, when using our approach with adaptive structure-aided setting instead of the global setting, more improvements can be seen in Fig. 7 (c). Third, in these figures, we notice that our approach can keep more details in the homogenous regions than the traditional Gaussian filtering methods. For test sets *Orbi* , thanks to the adaptive smoothing of our scheme, the texture details of the back-



(a)

(b)

(c)

(d)

(e)

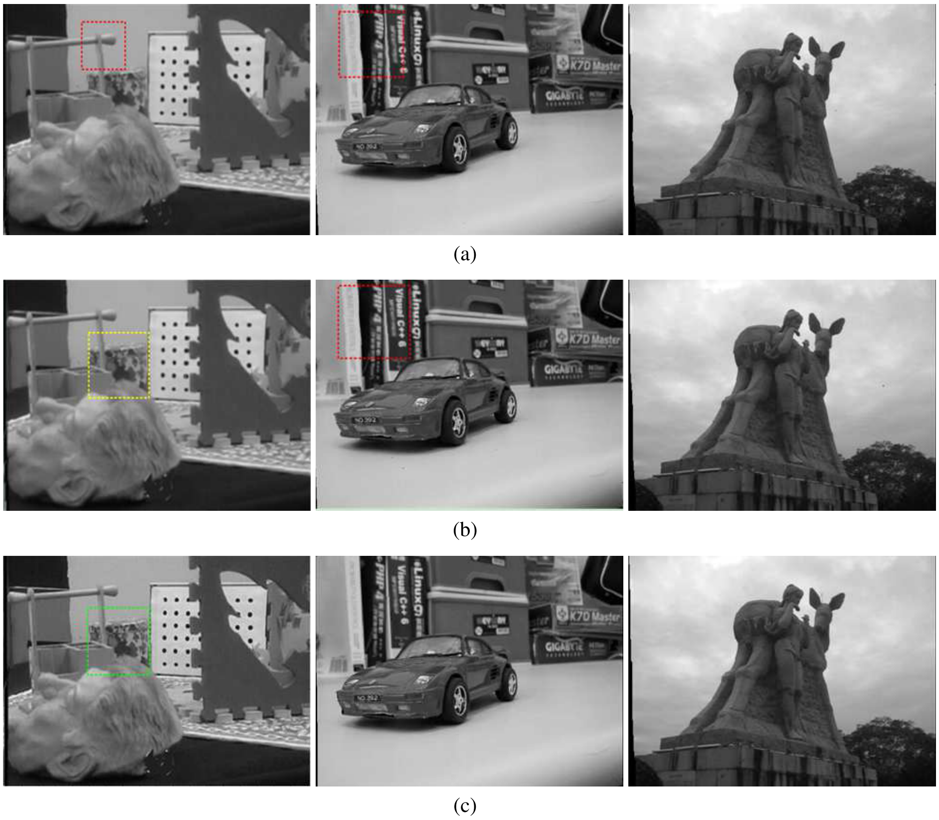
Fig. 6 Comparison of depth maps with different approaches using test sets including (from left to right) *Orbi*, *Desktop* and *Sculpture*. (a) Texture images, (b) initial depth maps, (c) symmetric filter, (d) asymmetric filter, (e) our approach.

grounds in the green rectangle boxed region of Fig. 7 (c) are more clear than the corresponding results generated by asymmetric Gaussian filtering method in Fig. 7 (b). It is of great benefit for high quality 2D to 3D conversion.

Table 3 shows the computation times of different filters. For traditional Gaussian filtering methods, only one smooth operation is carried out to the corresponding depth map for rendering a virtual view image. While in our scheme, two

Table 3 Computation time(s) comparison among different filters

Test set	Symmetric filter	Asymmetric filter	Proposed first filter	Proposed second filter
Orbi	0.81	1.08	0.21	0.23
Desktop	1.12	1.37	0.24	0.27
Sculpture	0.88	1.19	0.23	0.24



(a)

(b)

(c)

Fig. 7 Virtual view images with different approaches using test sets including (from left to right) *Orbi*, *Desktop* and *Sculpture*. (a) Symmetric filter, (b) asymmetric filter, (c) our approach.

filtering operations are needed sequentially. Even so, from Table 3, it can be seen that any filter of our scheme takes great advantage on computation times against other methods, the high efficiency mainly benefits from the proposed domain transform framework. Besides that, even the total times of the two filters spending together are less than any other method. Considering the good results of image quality evaluation above, our proposed scheme in this paper got a good tradeoff between time saving and virtual view quality.

Table 4 Results of subjective quality evaluation

	Test set	**Symmetric filter**	**Asymmetric filter**	**Proposed approach**
TEST1	Flower	4.1	4.3	***4.5***
	Castle	4.0	4.2	***4.4***
	Orbi	4.0	4.2	***4.5***
	Desktop	3.8	4.0	***4.4***
	Sculpture	4.3	***4.4***	***4.4***
TEST2	Flower	3.7	3.8	***4.1***
	Castle	3.7	3.8	***4.1***
	Orbi	3.9	3.9	***4.2***
	Desktop	3.5	3.7	***4.2***
	Sculpture	3.8	3.9	***4.1***

3.3 Results of Subjective Quality Evaluation

Subjective viewing tests are also performed with these test sets by 15 individuals with normal or correct-to-normal visual acuity and stereo acuity. We made two tests to evaluate image quality of the newly generated virtual views and stereoscopic feeling of the synthesized 3D anaglyph images separately. Both of the scores were from 0 to 5, and a higher score illustrates higher image quality or stereoscopic feeling. In the first test, the participants watched the original images in each test set to have a high reference before formal evaluation. Also in the second test, before the formal evaluation, training is given to the participants using true 3D anaglyph images to help them gain a better understanding of the stereoscopic feeling. In the tests, test images of each set are displayed in a random order. The average score was obtained and used as a measure of the subjective evaluation as shown in Table 4. The best results are high lighted with bold face type. From this table, we can see that the proposed scheme presents better performances in both image quality on the newly generated virtual views and stereoscopic feeling on the synthesized 3D anaglyph images. We also notice that for test set *Sculpture*, all methods get similar scores in *Test 1*, while in *Test 2*, our approach gains higher score than other methods. In fact, the evaluation results are consistent with the previous experimental analysis in Fig. 7. The reason behind is that: for Gaussian filtering methods, although the visual quality of the newly generated virtual views is improved, the depth perception of the scene at important ob-

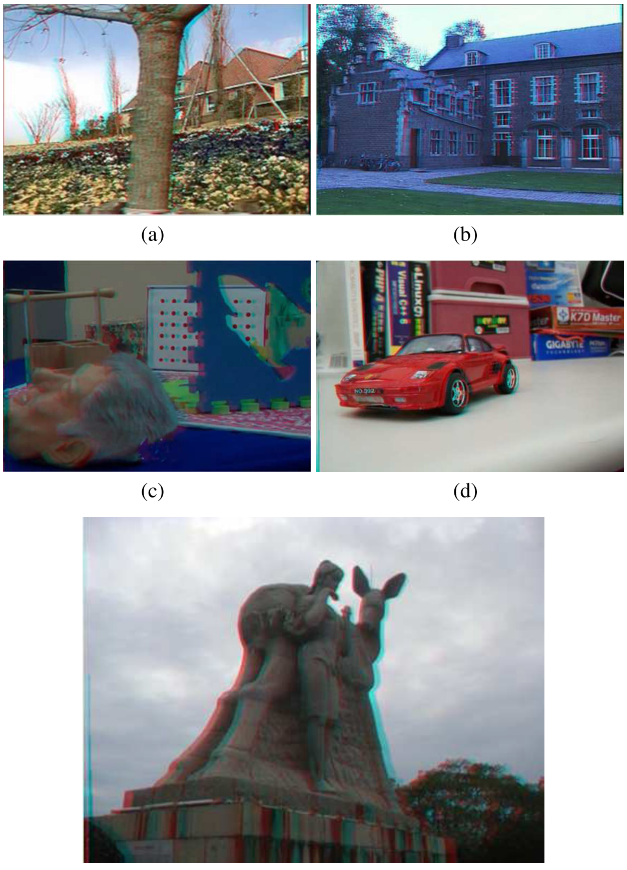


Fig. 8 Some synthesized 3D anaglyph images of the test sets, (a) *Flower*, (b) *Castle*, (c) *Orbi*, (d) *Desktop* and (e) *Sculpture*.

ject edges is still distorted due to additional disparity shifts when a filter smoothes the depth map. In this paper, our proposed scheme has special constraints to reduce distortions for these situations. Figure 8 shows some examples of the synthesized 3D anaglyph images in the evaluation test sets.

4. Conclusions

In this paper, a novel robust depth image based rendering scheme for stereoscopic view synthesis is proposed. This scheme is based on backward texture warping, which requests the depth map on the target view to be available for retrieving the corresponding texture virtual image. This expectation has sufficiently been met mainly by two depth map filters, which share a common domain transform based filtering framework and are carried out sequentially both before and after the depth map 3D warping. At each step, different constraints are added to the related filter for realizing specific process effects. Thanks to this flexible smoothing strategy, our approach can handle the problems of depth edge misalignment, disocclusion occurrences and cracks at resampling in one united scheme simultaneously and efficiently. Experimental results have illustrated the high efficiency of the proposed approach, which got a good trade-off between time saving and virtual view quality. It should be pointed that at present, the proposed scheme in this paper synthesizes a stereoscopic image only with an original texture image and its corresponding depth map. In future work, temporal correlation between different frames would be considered to incorporate into the scheme for extending its applications to 3D video synthesis.

References

[1] M. Kim, J. Nam, W. Baek, J. Son, and J. Hong, "The adaptation of 3D stereoscopic video in MPEG-21 DIA," Signal Processing Image Communication, vol.18, no.8, pp.685–697, 2003.

[2] C. Fehn and K. Hopf, "Key technologies for an advanced 3D TV system," Proceedings of SPIE - The International Society for Optical Engineering, vol.5599, pp.66–80, 2004.

[3] G.J. Iddan and G. Yahav, "Three-dimensional imaging in the studio and elsewhere," Photonics West 2001 - Electronic Imaging, pp.48–55, 2001.

[4] J. Gil and M. Kim, "Motion depth generation using MHI for 2D-to-3D video conversion," Electronics Letters, vol.53, no.23, pp.1520–1522, 2017.

[5] Y.-C. Fan and T.-C. Chi, "The novel non-hole-filling approach of depth image based rendering," 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2008, pp.325–328, IEEE, 2008.

[6] N. Plath, S. Knorr, L. Goldmann, and T. Sikora, "Adaptive image warping for hole prevention in 3D view synthesis," IEEE Transactions on Image Processing, vol.22, no.9, pp.3420–3432, 2013.

[7] L.Y. Wei, S. Lefebvre, V. Kwatra, and G. Turk, "State of the art in example-based texture synthesis," Eurographics 2009, State of the Art Report, EG-STAR, pp.93–117, Eurographics Association, 2009.

[8] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," Electronic Imaging 2004, pp.93–104, International Society for Optics and Photonics, 2004.

[9] W.J. Tam, G. Alain, L. Zhang, T. Martin, and R. Renaud, "Smoothing depth maps for improved steroscopic image quality," Optics East, pp.162–172, International Society for Optics and Photonics, 2004.

[10] L. Zhang and W.J. Tam, "Stereoscopic image generation based on depth images for 3D TV," IEEE Trans. Broadcast., vol.51, no.2, pp.191–199, 2005.

[11] S.-B. Lee and Y.-S. Ho, "Discontinuity-adaptive depth map filtering for 3D view generation," Proceedings of the 2nd International Conference on Immersive Telecommunications, p.8, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[12] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Distance dependent depth filtering in 3D warping for 3DTV," 2007 IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007, pp.312–315, IEEE, 2007.

[13] S. Zinger, L. Do, and P.H.N. De With, "Free-viewpoint depth image based rendering," Journal of Visual Communication and Image Representation, vol.21, no.5-6, pp.533–541, 2010.

[14] Y. Zhao, C. Zhu, Z. Chen, D. Tian, and L. Yu, "Boundary artifact reduction in view synthesis of 3D video: From perspective of texture-depth alignment," IEEE Trans. Broadcast., vol.57, no.2, pp.510–522, 2011.

[15] N. Chahal, M. Pippal, and S. Chaudhury, "Automated conversion of 2d to 3d image using manifold learning," Computer Vision, Pattern Recognition, Image Processing and Graphics, pp.1–4, 2015.

[16] X. Xu, L.-M. Po, K.-W. Cheung, L. Feng, and C.-H. Cheung, "Watershed based depth map misalignment correction and foreground biased dilation for DIBR view synthesis," IEEE International Conference on Image Processing, pp.3152–3156, 2013.

[17] J. Park, H. Kim, Y.-W. Tai, M.S. Brown, and I. Kweon, "High qual-

ity depth map upsampling for 3D-TOF cameras," 2011 IEEE International Conference on Computer Vision (ICCV), pp.1623–1630, IEEE, 2011.

[18] T. Matsuo, N. Fukushima, and Y. Ishibashi, "Weighted joint bilateral filter with slope depth compensation filter for depth map refinement," International Conference on Computer Vision Theory and Applications, 2015.

[19] S.-U. Yoon and Y.-S. Ho, "Multiple color and depth video coding using a hierarchical representation," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.11, pp.1450–1460, 2007.

[20] W. Liu, L. Ma, B. Qiu, M. Cui, J. Ding, and Y. Wang, "An efficient depth map preprocessing method based on structure-aided domain transform smoothing for 3d view generation," Plos One, vol.12, no.4, p.e0175910, 2017.

[21] S. Kirshanthan, L. Lajanugen, P.N.D. Panagoda, L.P. Wijesinghe, D.V.S.X. De Silva, and A.A. Pasqual, "Layered depth image based HEVC multi-view codec," International Symposium on Visual Computing, pp.376–385, Springer, 2014.

[22] E.S.L. Gastal and M.M. Oliveira, "Domain transform for edge-aware image and video processing," Acm Transactions on Graphics, vol.30, no.4, pp.1–12, 2011.

[23] M.-M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.3, pp.569–582, 2015.

[24] T. Li, Q. Dai, and X. Xie, "An efficient method for automatic stereoscopic conversion," 5th International Conference on Visual Information Engineering, VIE 2008, pp.256–260, IET, 2008.

[25] P. Li, D. Farin, R.K. Gunnewiek, et al., "On creating depth maps from monoscopic video using structure from motion," Proceedings of 27th Symposium on Information Theory in the Benelux, pp.508–515, Citeseer, 2006.

[26] M.T. Pourazad, P. Nasiopoulos, and R.K. Ward, "An H.264-based scheme for 2D to 3D video conversion," IEEE Trans. Consum. Electron., vol.55, no.2, pp.742–748, 2009.

[27] F. Zheng and Z. Yuan, "Depth estimation from single image based on vanishing point," J. Info. Tech. Appl, vol.1, no.3, pp.229–235, 2006.

[28] A. Saxena, M. Sun, and A.Y. Ng, "Make3D: Learning 3D scene structure from a single still image," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.5, pp.824–840, 2009.

[29] W. Liu, Y. Wu, F. Guo, and Z. Hu, "An efficient approach for 2D to 3D video conversion based on structure from motion," Visual Computer, vol.31, no.1, pp.55–68, 2015.

Wei Liu received the B.S. and M.S. degrees from the Department of Automation, Zhengzhou University, Zhengzhou, China, in 2006 and 2009 respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2012. He is currently an Assistant Professor in Nanyang Normal University, Nanyang, China. His research interests include image/video processing, 3D vision and applications.

Yunqi Tang received his Ph.D. degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2013. He is currently an assistant professor in People's Public Security University of China. His research interests include computer vision, pattern recognition and machine learning.

Jianwei Ding received the B.S. and M.S. degrees from Nanjing University of Aeronautics and Astronautics. He obtained his Ph.D. degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2013. He is currently a lecturer in People's Public Security University of China. His research interests include computer vision, pattern recognition and machine learning.

Mingyue Cui received the B.S. degree in Automation Engineering from Luoyang University of Science and Technology, the M.S. degree in Control Theory and Automation Engineering from Lanzhou University of technology and the Ph.D. degree in Control Theory and Automation Engineering from Chonqging University, China, in 1997, 2009 and 2012, respectively. He is currently a lecturer of Control Science and Engineering in Nanyang Normal University, Nanyang, China. His research interests include computer vision, mobile robot control and vibration active control.