A Hole Filling Approach based on Background Reconstruction for View Synthesis in 3D Video

Guibo Luo[1] Yuesheng Zhu[1] Zhaotian Li[1] Liming Zhang[2]
[1]Commun. Inf. Secur. Lab, Shenzhen Graduate School, Peking University, China
[2]Faculty of Science and Technology, University of Macau, Macao, China
luoguibo@sz.pku.edu.cn, zhuys@pkusz.edu.cn, lizhaotian@sz.pku.edu.cn, lmzhang@umac.mo

Abstract

The depth image based rendering (DIBR) plays a key role in 3D video synthesis, by which other virtual views can be generated from a 2D video and its depth map. However, in the synthesis process, the background occluded by the foreground objects might be exposed in the new view, resulting in some holes in the synthetized video. In this paper, a hole filling approach based on background reconstruction is proposed, in which the temporal correlation information in both the 2D video and its corresponding depth map are exploited to construct a background video. To construct a clean background video, the foreground objects are detected and removed. Also motion compensation is applied to make the background reconstruction model suitable for moving camera scenario. Each frame is projected to the current plane where a modified Gaussian mixture model is performed. The constructed background video is used to eliminate the holes in the synthetized video. Our experimental results have indicated that the proposed approach has better quality of the synthetized 3D video compared with the other methods.

1. Introduction

As 3D TV and 3D movie become increasing popular recently, the technology of the production and communication for 3D video is a hot topic. How to acquire and transmit these 3D video data is a challenge. Depth image based rendering (DIBR) [9] technique is a practical way to generate multi-view video by using a reference 2D video and its corresponding depth map, which can reduce the storage and save much bandwidth. However, in the DIBR technique, the regions of background occluded by the foreground objects in the original views might become visible in the virtual views, resulting in some holes in the synthetized video. This is also known as "disocclusion".

Generally, there are two types of methods to fill the holes. One is to preprocess the depth map by a low-pass filter so that the hole regions are reduced. The symmetric Gaussian low-pass filter [9, 28] or asymmetric filter [14] is employed to smooth the whole depth map, which would lead to some geometrical distortions in the virtual view. These methods smooth not only the horizontal edge areas but also the areas that do not cause holes. To alleviate this problem, an edge-dependent Gaussian filter [4] is used to smooth the horizontal edge only and keep the non-hole areas unchanged, or an adaptive edge-oriented smoothing process [20] is utilized to preprocess the depth map with two types of smoothing filters. This type of methods would reduce 3D effect as the depth map is smoothed. Moreover, they are not suitable for the situation that the virtual camera is far away from the reference camera.

The other type of methods is to use the spatial or temporal correlation of the video to fill the holes. In the spatial domain, the view blending approaches [17, 21, 32, 22] can fill most of the holes by using multiple views, but they require more capturing devices and transmission bandwidth. Therefore, single view approaches are more practical and draw more attention. The hierarchical hole filling method [25] down-samples and then progressively up-samples the virtual view, in which no geometrical distortions are produced but blurry regions around the large holes may be introduced. The exemplar-based inpainting method is a popular solution to fill the large holes without introducing blur artifacts. Criminisi *et al*. [7] fill the holes by propagating both texture and structure simultaneously from non-hole regions, so blurry effect is not produced, but the foreground textures might be sampled to fill the holes. In order to alleviate this problem, some improved methods [8, 11, 1] employ the depth information to exclude the foreground textures in the filling process. Daribo and Saito [8] add the depth values to compute the priority and patch distance. Gautier *et al*. [11] apply 3D structure tensor of Di Zenzo matrix to compute the priority and add depth information for patch matching. But they [8, 11] both use the depth map of the virtual view, which is not practical. To overcome this problem, Ahn and

Kim [1] compute the depth values of the virtual view in the filling process, but the artifacts would be produced in the hole regions when the depth values are incorrect.

Spatial filling methods try to fill the disoccluded areas with visually plausible backgrounds according to some spatial correlation assumptions, but may not reflect the ground-truth textures occluded by the foreground objects. Temporal filling method is able to reveal the ground-truth textures in the disoccluded areas by using more frames.

In temporal domain, the occluded background in current frame might become visible in other frames when the foreground objects move away. Background reconstruction can exploit the temporal correlation information in both the 2D video and its corresponding depth map to generate a background video, which can be used to eliminate the holes in the synthetized video. Therefore, some background models are employed to recover the occluded background. The average background model [6] extracts the background from the scene and updates the background dynamically, which can get a stable background, but does not work for fast moving scene. The temporal background model [15] generates the uncovered background layer by median filtering its neighboring frames, so the background information is limited in the neighboring frames. The sprite update method [18] separates the background and foregrounds by depth values, and updates the background of 2D video and depth map respectively. The Switchable Gaussian Model (SGM) [27] builds an online background method, also reduces the computational complexity and increases the scene adaptability. The Gaussian Mixture Model (GMM) and Foreground Depth Correlation (FDC) [31] construct a stable background offline from several consecutive video frames and depth map. If the foreground object moves slowly or rotates, the GMM may regard this foreground object as part of the stable background since the real background is occluded by the foreground object in most of frames. If the depth map is imperfect, FDC may yield to some artifacts of the foreground textures.

Most of the background models based methods may bring some foreground textures in the constructed background or not suitable for moving camera scenario. In this paper, a hole filling approach based on background reconstruction is proposed, in which the foreground objects are removed, and then motion compensation is applied for moving camera scenario, finally a clean background video is generated by modified GMM. Our approach is suitable for motion scene, and can prevent blurry effect, or bringing artifacts from the foreground textures as the foreground objects are removed.

The rest of this paper is organized as follows. The proposed hole filling approach is presented in Section II. The foreground extraction is described in Section III, the new background models is described in Section IV, and the new disocclusions filling is given in Section V, and in Section VI, the experimental results are presented. The conclusions are given in Section VII.

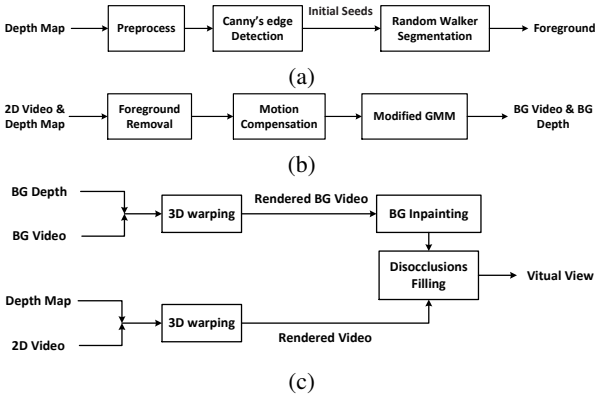


Figure 1. Block diagram of the proposed algorithm. (a) Foreground extraction. (b) BG video and BG depth map reconstruction. (c) DIBR module with BG for disocclusions filling.

2. Proposed hole filling approach

Due to the inaccuracy of depth map and the round-off error caused by 3D warping, there are some distortions in non-hole regions of the virtual view. So in our proposed approach, the background is reconstructed in the reference view.

The proposed approach consists of three parts, including the foreground extraction in Figure 1(a), the reconstruction of background (BG) video and BG depth map in Figure 1(b), and DIBR module with BG for disocclusions filling in Figure 1(c).

In Figure 1(a), the depth map is preprocessed by a cross-bilateral filter and morphological operations, and then the Canny's edge detection method is used to extract the initial seeds for the random walker, finally the foreground and background are separated in the depth map by the random walker segmentation.

In Figure 1(b), the foreground objects in the 2D video and depth map are removed, motion compensation is applied for moving camera scenario, and the modified GMM is applied to obtain the BG video and BG depth.

In Figure 1(c), the 2D video and its depth map, and the BG video and the BG depth map are warped by the 3D warping module to get the rendered video and rendered BG video, respectively; the holes of the rendered BG video are filled by inpainting-based method; and the rendered BG video are used to fill the holes in the rendered video.

The main processing modules will be described in the following sections.

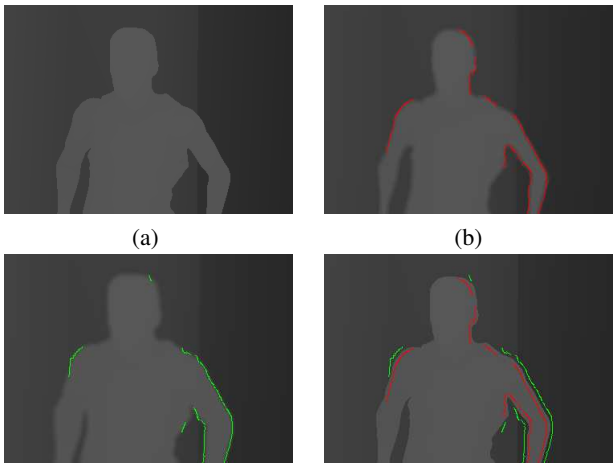


Figure 2. Initial seeds extraction for "Dancer". (a) Original depth map. (b) Z_i and *IBFO* (red). (c) Z_o and *OBFO* (green). (d) Initial seeds on the depth map, foreground label (red) and the background label (green).

3. Foreground extraction

In order to generate a clean background video, the foreground objects need to be removed from the 2D video and its depth map in the reference view. How to automatically extract the foreground in the video is still a challenging issue. Image segmentation methods in literature [23] and [12] can separate the foreground from background well, but they need to interact with users. In the literatures [26, 16, 19], GMM is employed to build a background model, and foreground objects can be extracted by background subtraction, but only the moving objects can be detected. In the literature [5], the foreground and background are separated in the depth map by the random walker segmentation, but it is usually used in the virtual views, in which the initial labels can be gotten by applying the Laplacian operator to the depth map.

In our proposed method, the foreground objects are automatically extracted in the reference view by random walker segmentation algorithm. One of the most important steps for random walks is to extract the initial seeds automatically.

3.1. Initial seeds

The initial seeds of the random walker are extracted by applying some characteristics of the depth map in the reference view. First, the edges of foreground objects and background need to be detected. Since there are some irregular depth discontinuities in the same object, these edges are not supposed to be extracted, so it is desirable to smooth them away. They can be smoothed away by using a low-pass filter, but the real edge regions will also be smoothed. In order to eliminate these unreal edges and preserve the real edges, a cross-bilateral filter [2] is applied as follows:

$$h(x) = k^{-1}(x) \iint_D Z(\xi) c(\xi, x) s(f(\xi), f(x)) d\xi \quad (1)$$

where D is the filtering window with size $W \times W$, the weight c and s both are Gaussian functions, and is given, respectively as:

$$c\left(\xi, x\right) = \exp\left(-\frac{1}{2}\left(\frac{|\xi - x|}{\sigma_d}\right)^2\right) \quad (2)$$

$$s\left(f\left(\xi\right), f\left(x\right)\right) = \exp\left(-\frac{1}{2}\left(\frac{f\left(\xi\right) - f\left(x\right)}{\sigma_r}\right)^2\right) \quad (3)$$

$Z(x)$ is the depth value at pixel x, $f(x)$ is the color value at pixel x, σ_d is the variance of Euclidean distance, σ_r is the variance of color space, k is the normalization factor, and is given by:

$$k(x) = \iint_D c(\xi, x) s(f(\xi), f(x)) d\xi \quad (4)$$

By introducing weight function s, large distance in color space would have small weight, which preserves real edge regions from smoothing.

Since the edges are the boundaries of foreground objects and background, they might locate in the background or foreground. So grayscale morphological erosion operation is conducted to the depth map to ensure the edges locate in the foreground objects, and grayscale morphological dilation operation is conducted to the depth map to ensure the edges locate in the background, their corresponding results are denoted as Z_i and Z_o, respectively.

$$Z_i = Z \odot B \quad (5)$$

$$Z_o = Z \oplus B \quad (6)$$

where Z is the depth map of reference view, $\odot$ is the operation of morphological erosion, $\oplus$ is the operation of morphological dilation, B is the structuring element with size $L \times L$.

After the preprocessing, Canny's edge detection method [30] is conducted to both Z_i and Z_o to extract the inner boundaries of foreground objects (*IBFO*) and the outer boundaries of foreground objects (*OBFO*) as shown in Figure 2(b) and Figure 2(c), respectively.

Based on the result of *IBFO* and *OBFO*, initial seeds can be automatically assigned to the random walker algorithm. Let us define a label set $S = \{s1, s2\}$, where labels $s1$ and $s2$ correspond to the foreground and background, respectively. Note that the points of *IBFO* (red line) are in the foreground, and the points of *OBFO* (green line) are in the background. So the points of *IBFO* are served as foreground label ($s1$), and the points of *OBFO* are served as the background label ($s2$) as shown in Figure 2(d).

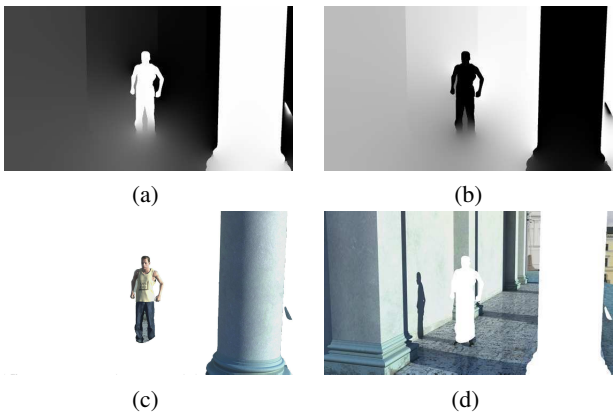


Figure 3. Foreground extraction for "Dancer". (a) and (b) Random walker's probability map of the foreground label (x^{s1}) and the background label (x^{s2}), respectively. (c) and (d) Foreground regions and background regions, respectively.

3.2. Random Walker Segmentation

To formulate the separation of foreground and background as a labeling problem, an undirected graph $G = (V, E)$ is constructed for random walks [12] formulation, where V is the set of all the points in the depth map, and E is the set of weighted edges. Let us define v_i represents the i^{th} point in the depth map , $v \in V$.

In order to solve this labeling problem, the weights between nodes are defined as typical Gaussian weighting function.

$$w_{ij} = \exp\left(-\beta(g_i - g_j)^2\right) + \varepsilon \tag{7}$$

where g_i indicates the depth value at pixel i , ε is a small constant (*e.g.*, 10^{-6}), β is a weighting factor to balance the sensitivity of the depth similarity cost (*e.g.*, β = 90 in our experiment), depth values are normalized prior to applying (7), so that $0 \leq |g_i - g_j| \leq 1$.

From the weight defined in (7), the combinatorial Laplacian matrix L is given as

$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -w_{ij} & \text{if } v_i \text{ and } v_j \text{ are adjacent nodes} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where L_{ij} is indexed by points v_i and v_j, and $d_i = \sum w(e_{ij})$ is the degree of node i for all edges e_{ij} incident on point v_i .

The vertices are partitioned into two sets, seeded nodes V_M and unseeded nodes V_U , and L can be decomposed as

$$L = \begin{bmatrix} L_M & B \\ B^T & L_U \end{bmatrix} \tag{9}$$

where L_M is the weights of seeded nodes and L_U is the weight of unseeded nodes.

Solving the unknown probabilities for the label is equivalent to solving the matrix equation of

$$L_U x_U = -B^T x_M \tag{10}$$

where x_M and x_U correspond to the probabilities of the seeded and unseeded nodes respectively.

Additionally, let us define the probability at node v_i for each label s by x_i^s . Define the set of labels for the seeded points as a function $Q(v_j) = s, \forall v_j \in V_M$, where $s \in S$, $S = \{s1, s2\}$. Define the $|V_M| \times 1$ vector for each label s at node $v_j \in V_M$ as

$$m_j^s = \begin{cases} 1 & \text{if } Q(v_j) = s \\ 0 & \text{if } Q(v_j) \neq s \end{cases} \tag{11}$$

Therefore for label s, the solution to the combinatorial Dirichlet problem can be found by solving

$$L_U x^s = -B^T m^s \tag{12}$$

With the initial seeds, random walker's probability map of the foreground label and the background label can be obtained by solving (12), whose results are shown in Figure 3(a) and Figure 3(b), respectively. The higher intensity value corresponds to higher probability. The label of each unseeded point is obtained from the label with the highest probability, then the segmented regions associated with the foreground label and the background label can be decided, as shown in Figure 3(c) and Figure 3(d), respectively. Notice that the still foreground object (the pillar) also can be extracted in the proposed method, as shown in Figure 3(c).

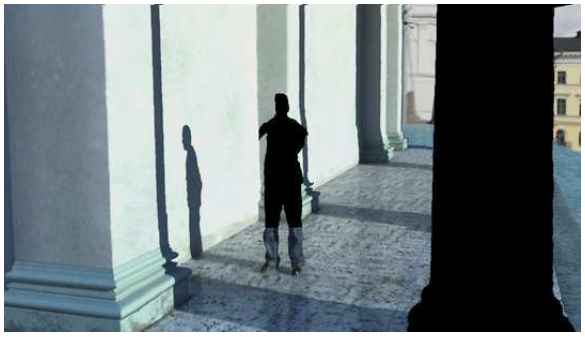
4. Dynamically background reconstruction

After the foreground objects are removed, the remaining part of the video can be used to reconstruct a clean background. Considering traditional background reconstruction methods are not suitable for moving camera scenario, the proposed dynamically background reconstruction is processed by two modules: motion compensation and modified GMM.

4.1. Motion compensation

In the case of non-stationary camera, the model learned until time $t - 1$ cannot be used directly in time *t*. All the parameters of the model need to be warped to the new positions by using motion compensation.

In the motion compensation process, SURF [3] is used to detect and describe the feature points in the reference frame and the current frame. To be more robust, the RANSAC algorithm [10] is utilized for optimally matched feature point-pair. The homography matrix $H_{t:t-1}$ can be gotten once the optimized feature point-pair are obtained. Then all the parameters of the model in time $t - 1$ are warped to time *t* through a perspective transformation.



(a)

(b)

Figure 4. Background reconstruction for 'Dancer'. (a) BG reconstruction by modified GMM. (b) The occluded regions recovered by (a).

4.2. Modified GMM

The Gaussian Mixture Model is usually used for detecting the moving objects [26], as it can be applied to model the stable background. GMM is performed at pixel level, where each pixel is modeled independently by a mixture of K Gaussian distributions (a common setting is K = 3) [19, 13]. In our proposed method, the foreground pixels are not used to build the models, and motion compensation is used for non-stationary scene. The modified Gaussian mixture distribution with K components can be written as:

$$p\left(I_{x,t}\right) = B\left(x_t\right) \cdot \sum_{i=1}^{K} w_{x,i,t} \cdot \eta\left(I_{x,t}, \mu_{x,i,t}, \sigma_{x,i,t}^2\right) \quad (13)$$

where $p\left(I_{x,t}\right)$ indicates the probability density of pixel x at time t, η is the Gaussian function with $I_{x,t}$ representing the value of pixel x at time t, $\mu_{x,i,t}$ and $\sigma_{x,i,t}^2$ denote the mean and variance of pixel x at time t, respectively, and $w_{x,i,t}$ is the i^{th} Gaussian distributions weight of pixel x at time t, with $\sum_{i=1}^{K} w_{x,i,t} = 1$, $B\left(x_t\right)$ is the background mask of pixel x at time t, with $B\left(x_t\right)$ = 0 when the models are empty, $B\left(x_t\right)$ = 1 when the models are not empty.

The detailed process of the proposed background model is described as follows:
(1) Firstly, an empty set of models is initialized at the time instant t_0 .

$$\mu_{x,i,t_0} = \begin{cases} I_{x,t_0} & \text{if } i = 1 \text{ and } F(x_{t_0}) = 0 \\ 0 & \text{others} \end{cases} \quad (14)$$

$$\sigma_{x,i,t_0} = \sigma_0 \quad (15)$$

$$w_{x,i,t_0} = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{others} \end{cases} \quad (16)$$

$$B\left(x_{t_0}\right) = \begin{cases} 1 & \text{if } F(x_{t_0}) = 0 \\ 0 & \text{others} \end{cases} \quad (17)$$

where σ_0 is a pre-defined large value, $F\left(x_t\right)$ is the foreground mask of pixel x at time t, if pixel x_t is detected as a foreground pixel, $F\left(x_t\right)$ = 1; otherwise, $F\left(x_t\right)$ = 0.
(2) For the next frame at the time instant t_1, all background models in time $t-1$ are warped to background models in time t through a perspective transformation. By using the homography matrix $H_{t;t-1}$, the coordinate x_t in the plane at time t is warped to $x_{t-1}^{'}$ in the plane at time $t-1$, correspondingly, the background models at time t of pixel x_t are updated from the background models at time $t-1$ of pixel $x_{t-1}^{'}$.

$$\mu_{x,i,t-1} = \mu_{x',i,t-1} \quad (18)$$

$$\sigma_{x,i,t-1}^2 = \sigma_{x',i,t-1}^2 \quad (19)$$

$$w_{x,i,t-1} = w_{x',i,t-1} \quad (20)$$

$$B\left(x_{t-1}\right) = B\left(x_{t-1}^{'}\right) \quad (21)$$

Then the background models will update if current pixel is not a foreground pixel ($F(x_t)$ = 0), the update process is described as the following role:

The current pixel is used to match with the K Gaussian models. For each model i, if the condition $|I_{x,t} - \mu_{x,i,t-1}| \leq 2.5\sigma_{x,i,t-1}$ is satisfied, the matching process will be stopped. The parameters of the matched Gaussian model will be updated as:

$$\mu_{x,i,t} = (1-\rho)\,\mu_{x,i,t-1} + \rho I_{x,t} \quad (22)$$

$$\sigma_{x,i,t}^2 = (1-\rho)\,\sigma_{x,i,t-1}^2 + \rho(I_{x,t} - \mu_{x,i,t})^2 \quad (23)$$

$$w_{x,i,t} = (1-\alpha)\,w_{x,i,t-1} + \alpha \quad (24)$$

And the parameters of the other Gaussian models will be updated as:

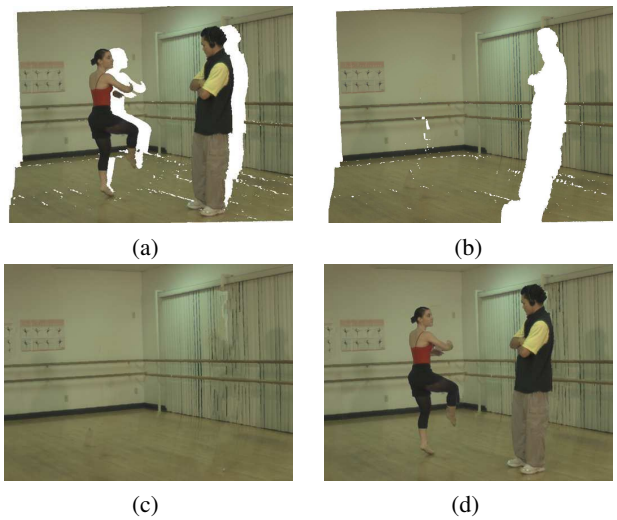
$$\mu_{x,i,t} = \mu_{x,i,t-1} \quad (25)$$

$$\sigma_{x,i,t}^2 = \sigma_{x,i,t-1}^2 \quad (26)$$

$$w_{x,i,t} = (1-\alpha)\,w_{x,i,t-1} \quad (27)$$

where $\rho = \alpha \cdot \eta\left(I_{x,t}, \mu_{i,t}, \sigma_{i,t}^2\right)$,α is the learning rate.

Whereas, if all of the Gaussian models fail to match the current pixel, then a new Gaussian model is introduced with $\mu_{x,t} = I_{x,t}$, $\sigma_{x,t} = \sigma_0$, $\omega_{x,t} = w_0$, where w_0 is a low weight value to evict the Gaussian model which



(a) (b) (c) (d)

Figure 5. Disocclusions filling for 'Ballet'. (a) Rendered video. (b) Rendered BG video. (c) Inpainting result of (b) . (d) (c) is used to fill the holes in (a).

has the smallest ω/σ value. The mean and variance value of the other Gaussian models remain unchanged, and the weight value of K Gaussian models are normalized to $\sum_{i=1}^{K} w_{x,i,t} = 1$.

(3) The remaining frames are processed by repeating the previous step (2). Finally the K Gaussian models are sorted based on ω/σ, and the value of the background pixel $bp\,(x_t)$ at the time instant t is obtained as

$$bp\,(x_t) = \mu_{x,1,t}, \text{if } B\,(x_t) = 1 \tag{28}$$

The setting of parameter σ_0 , α and w_0 have been discussed in literature [16], their typical values $\sigma_0 = 30$, $\alpha = 0.005$ and $w_0 = 0.001$ are used here.

One example of the BG video and BG depth map reconstruction by the modified GMM method is shown in Figure 4.

5. Disocclusions filling

After backgrounds are constructed, they are used to fill the holes in the virtual view.

In the virtual view, the rendered video and rendered BG video are generated by 3D warping module, which is shown in Figure 5(a) and (b), respectively. In the rendered BG video, some regions occluded by foreground objects cannot be reconstructed by background model, these regions are filled by Criminisi's inpainting method. This approach can prevent blurry effect and the penetration problem of the foreground textures since the foreground objects are removed. The BG video shown in Figure 5(c) is used to fill the holes in the rendered video, the result is shown in Figure 5(d).

Name	Resolution	Scene	Baseline
Ballet	1024×768	Stationary	380mm
Breakdancers	1024×768	Stationary	370mm
Dancer	1920×1088	Dynamic	2000mm

Table 1. Parameters of test dataset

6. Experimental results

6.1. Experiment setup

Three Multiview Video-plus-Depth (MVD) sequences ('Ballet', 'Breakdancers' [33], and 'Dancer' [24]) are used to evaluate the performance of the proposed approach in our experiment. They consist of stationary and moving camera scenario. The parameters of the test dataset are shown in Table 1.

The performance of visual quality are compared among the proposed method, the Criminisi's exemplar-based inpainting method [7], the Daribo's disocclusion filling method [8], the Ahn's depth-based image completion method [1], the MPEG view synthesis reference software (VSRS, version 3.5) [29] , and the Yao's GMM-based method [31]. In [7, 8, 1, 29, 31] and the proposed method, the virtual view generation need only single Video-plus-Depth (SVD) of reference view, but in [8], the depth map of the virtual view is also needed.

6.2. Visual quality evaluation

In our experiment, PSNR is used to measure the squared intensity differences of synthesized and reference image pixels, and SSIM (structural similarity) [30] is used to measure the structural similarity between synthesized and reference image. The average PSNR and SSIM values of proposed method and other methods [7, 8, 1, 29, 31] for the test sequences are shown in Table 2, where 'Test Seq.' denotes the dataset and projection information, for example, the sequence warped from view 5 to view 4 of 'Ballet' is named as BA54. The best results are highlighted in boldface. The results show that the proposed method yields the best overall results. The measured results on each frame of 'BA54','BR54',and 'DA15' are shown in Figure 6. The proposed method shows almost the best in both PSNR and SSIM measures compared with the other methods.

The visual quality comparisons of disoccluded areas for 'BA54','BR54',and 'DA15' are shown in Figure 7, the proposed method has more plausible results and preserves sharp edges, while other methods contain some artifacts or blurry results. In the Criminisi's method [7], foreground textures are sampled to fill a large part of hole regions as shown in Figure 7(b); in the Daribo's method [8], some artifacts occur along the foreground boundaries as shown in Figure 7(c); in the Ahn's method [1], some artifacts occurred along the image boundaries as shown in Figure 7(d);

Test Seq.	PSNR						SSIM					
	[7]	[8]	[1]	[29]	[31]	Ours	[7]	[8]	[1]	[29]	[31]	Ours
BA41	22.76	22.56	23.27	22.23	23.05	**24.15**	0.7391	0.7130	0.7478	0.7611	0.7496	**0.7759**
BA43	25.08	27.63	28.15	25.93	25.61	**28.86**	0.8388	0.8351	0.8465	0.8514	0.8429	**0.8570**
BA52	24.38	23.97	24.29	23.89	24.81	**25.76**	0.7422	0.7200	0.7385	0.7654	0.7563	**0.7768**
BA54	26.56	29.60	30.54	27.60	27.53	**32.00**	0.8468	0.8448	0.8545	0.8584	0.8524	**0.8665**
BR41	25.87	26.92	26.91	27.03	27.09	**27.30**	0.7639	0.7639	0.7737	**0.7814**	0.7694	0.7763
BR43	29.74	30.20	30.40	29.61	30.53	**30.60**	0.8197	0.8151	0.8223	0.8126	0.8227	**0.8248**
BR52	26.23	27.55	27.32	26.40	27.72	**27.85**	0.7660	0.7675	0.7737	0.7606	0.7712	**0.7776**
BR54	30.24	30.86	30.27	30.25	31.50	**31.73**	0.8217	0.8177	0.8225	0.8133	0.8255	**0.8275**
DA15	26.74	27.64	27.80	26.42	27.47	**29.76**	0.9412	0.9429	0.9446	0.9425	0.9408	**0.9471**
DA59	26.69	27.56	27.32	26.22	27.76	**29.53**	0.9405	0.9419	0.9434	0.9412	0.9415	**0.9463**

Table 2. The objective evaluations of the proposed and other methods using PSNR and SSIM

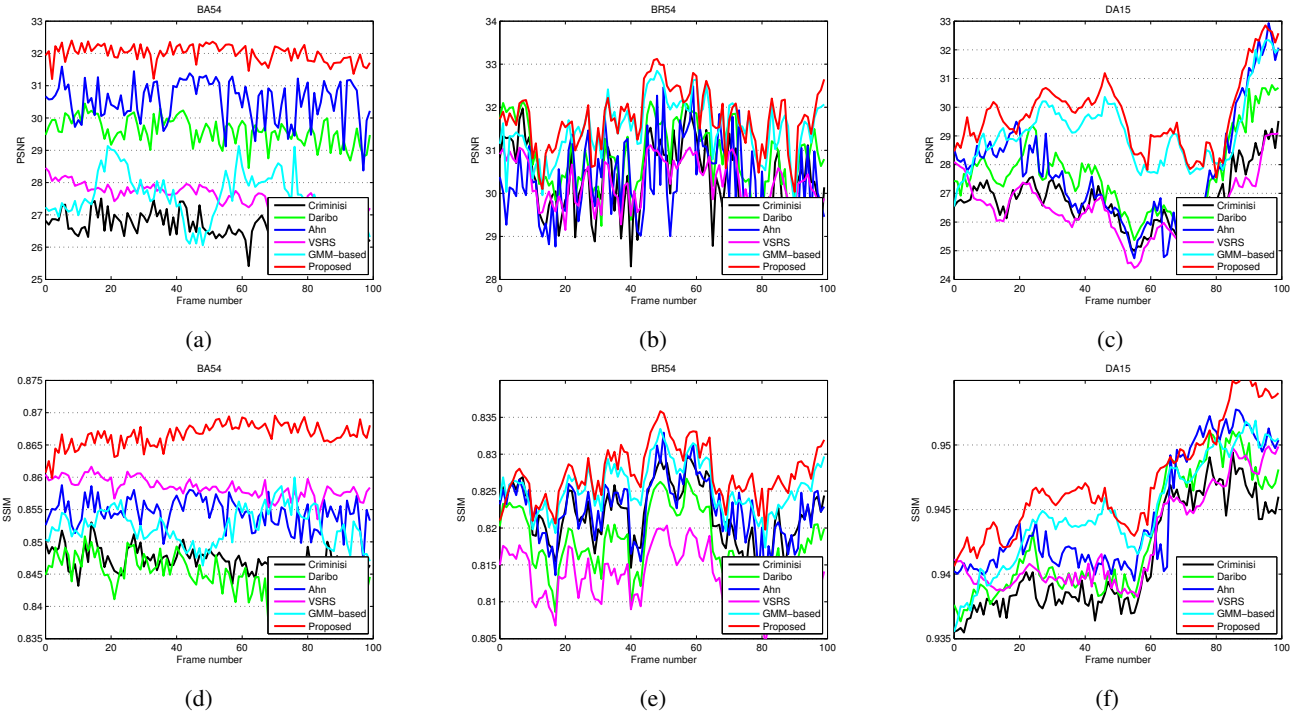
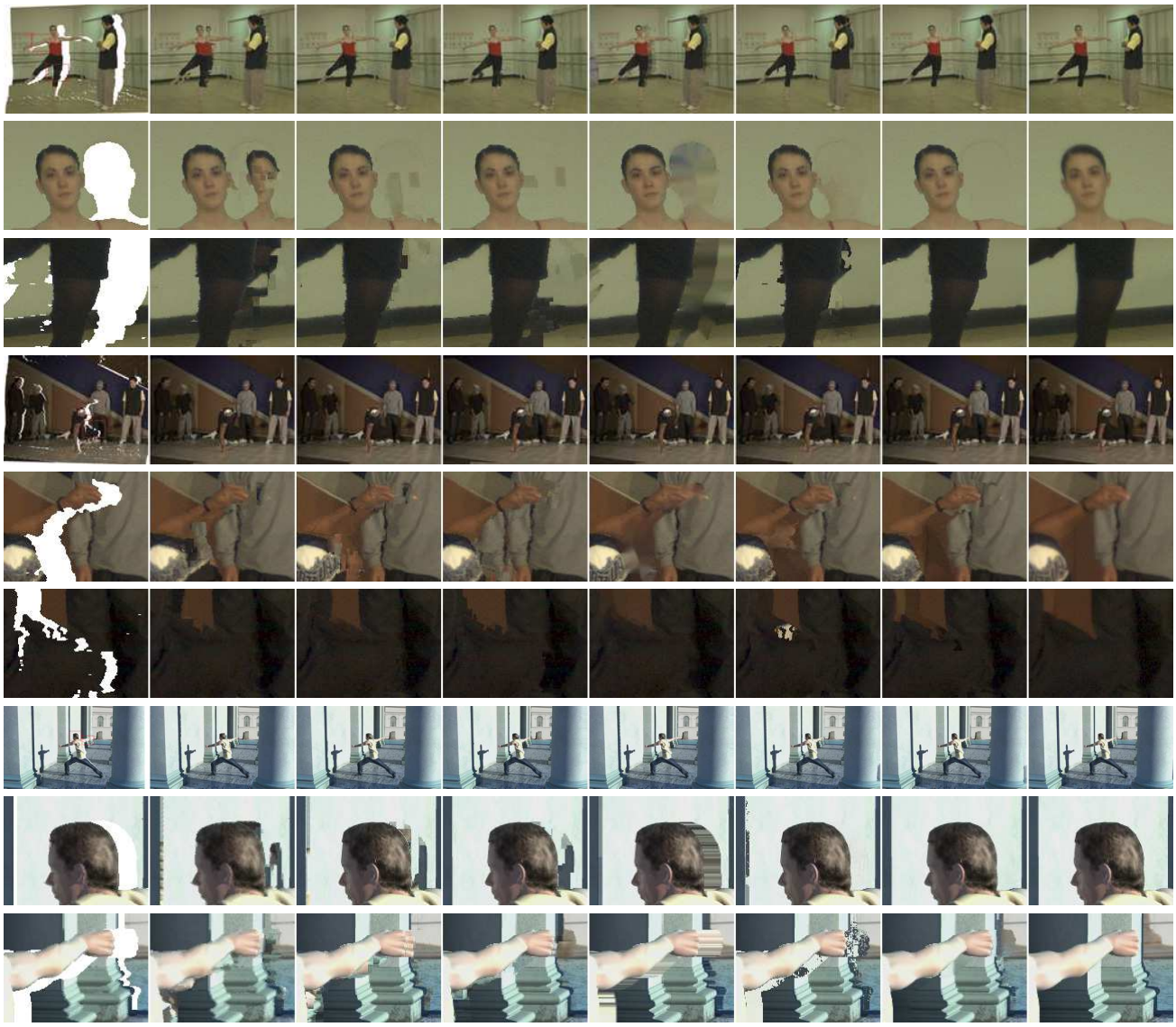


Figure 6. Objective quality measure results of the proposed and other methods. (a)(b)(c) PSNR of 'BA54','BR54' and 'DA15' respectively. (d)(e)(f) SSIM of 'BA54', 'BR54' and 'DA15'respectively.

in the VSRS [29] , some unrealistic regions or blurry artifacts are produced as shown in Figure 7(e); in the Yao's method [31], some artifacts of the foreground textures are produced as shown in Figure 7(f). The proposed method successfully preserves sharp edges along the foreground boundaries and shows realistic appearance in Figure 7(g).

7. Conclusion

In our hole filling approach, a constructed background video with modified GMM is used to eliminate the holes in the synthesized video. The foreground objects in 2D video and the depth map in the reference view are extracted and removed, and then motion compensation and modified GMM are applied to construct a stable background. Our investigation have indicated that a clean background without bringing the artifacts of foreground objects can be generated by using the proposed background model, so that the blurry effect or artifacts in the disoccluded regions can be eliminated and the sharp edges along the foreground boundaries with realistic appearance can be preserved as well.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 7. Results of the proposed method and competing algorithms in 'Ballet', 'BreakDancer' and 'Dancer'. (a) Hole regions. (b) Criminisi's method [7]. (c) Daribo's method [8]. (d) Ahn's method [1]. (e) VSRS [29]. (f) GMM-based method [31]. (g) Proposed. (h) Ground truth.

Acknowledgments

This work was supported by Shenzhen Engineering Laboratory of Broadband Wireless Network Security, and the Science and Technology Development Fund of Macao SARFDCT056/2012/A2 and UM Multi-year Research Grant: MYRG144(Y1-L2)- FST11-ZLM.

References

[1] I. Ahn and C. Kim. A novel depth-based virtual view synthesis method for free viewpoint video. *IEEE Transactions on Broadcasting*, 59(4):614–626, 2013.

[2] L. J. Angot, W.-J. Huang, and K.-C. Liu. A 2D to 3D video and image conversion technique based on a bilateral filter. In *IS&T/SPIE Electronic Imaging*, pages 75260D–75260D, 2010.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. 2006.

[4] W.-Y. Chen, Y.-L. Chang, S.-F. Lin, L.-F. Ding, and L.-G. Chen. Efficient depth image based rendering with edge dependent depth filter and interpolation. In *IEEE International Conference on Multimedia and Expo*, pages 1314–1317, 2005.

[5] S. Choi, B. Ham, and K. Sohn. Space-time hole filling with

random walks in view extrapolation for 3D video. *IEEE Transactions on Image Processing*, 22(6):2429–2441, 2013.

[6] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P. H. Torr. Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming. *International Journal of Computer Vision*, 71(1):89–110, 2007.

[7] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[8] I. Daribo and H. Saito. A novel inpainting-based layered depth video for 3DTV. *IEEE Transactions on Broadcasting*, 57(2):533–541, 2011.

[9] C. Fehn. Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3D-tv. In *Electronic Imaging 2004*, pages 93–104, 2004.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] J. Gautier, O. Le Meur, and C. Guillemot. Depth-based image completion for view synthesis. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. *IEEE*, 2011.

[12] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[13] M. Haque, M. M. Murshed, and M. Paul. Improved gaussian mixtures for robust object detection by adaptive multi-background generation. In *ICPR*, pages 1–4, 2008.

[14] Y.-R. Horng, Y.-C. Tseng, and T.-S. Chang. Stereoscopic images generation with directional gaussian filter. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2650–2653, 2010.

[15] Y. Huang and C. Zhang. A layered method of visibility resolving in depth image-based rendering. In *19th IEEE International Conference on Pattern Recognition*, pages 1–4, 2008.

[16] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. 2002.

[17] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. *Signal Processing: Image Communication*, 22(2):217–234, 2007.

[18] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand. Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 1809–1812, 2010.

[19] D.-S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005.

[20] P.-J. Lee et al. Nongeometric distortion smoothing approach for depth map preprocessing. *IEEE Trans.Multimedia Expo*, 13(2):246–254, 2011.

[21] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto. View generation with 3D warping using depth information for FTV. *Signal Processing: Image Communication*, 24(1):65–72, 2009.

[22] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View synthesis for advanced 3D video systems. *EURASIP Journal on Image and Video Processing*, 2008(1):1–11, 2008.

[23] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004.

[24] H. Schwarz, D. Marpe, and T. Wiegand. Description of exploration experiments in 3d video coding. *ISO/IEC JTC1/SC29/WG11 MPEG2010 N*, 11274, 2010.

[25] M. Solh and G. AlRegib. Hierarchical hole-filling for depth-based view synthesis in ftv and 3D video. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):495–504, 2012.

[26] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conference on CVPR*, volume 2, 1999.

[27] W. Sun, O. C. Au, L. Xu, Y. Li, and W. Hu. Novel temporal domain hole filling based on background modeling for view synthesis. In *19th IEEE International Conference on Image Processing (ICIP)*, pages 2721–2724, 2012.

[28] W. J. Tam, G. Alain, L. Zhang, T. Martin, and R. Renaud. Smoothing depth maps for improved steroscopic image quality. In *Optics East*, pages 162–172, 2004.

[29] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori. Reference softwares for depth estimation and view synthesis. *ISO/IEC JTC1/SC29/WG11 MPEG*, 20081:M15377, 2008.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[31] C. Yao, T. Tillo, Y. Zhao, J. Xiao, H. Bai, and C. Lin. Depth map driven hole filling algorithm exploiting temporal correlation information. *IEEE Transactions on Broadcasting*, 60(2):394–404, 2014.

[32] S. Zinger, L. Do, and P. de With. Free-viewpoint depth image based rendering. *Journal of visual communication and image representation*, 21(5):533–541, 2010.

[33] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 600–608. ACM, 2004.