Global Optimization For Spatio-Temporally Consistent View Synthesis

Hsiao-An Hsu, Chen-Kuo Chiang and Shang-Hong Lai
National Tsing Hua University, Hsinchu, Taiwan
E-mail: xiao-an@livemail.tw, ckchiang@cs.nthu.edu.tw, lai@cs.nthu.edu.tw

***Abstract*—We propose a novel algorithm to generate a virtual-view video from a video-plus-depth sequence. The proposed method enforces the spatial and temporal consistency in the disocclusion regions by formulating the problem as an energy minimization problem in a Markov random field (MRF) framework. In the system level, we first recover the depth images and the motion vector maps after the image warping with the preprocessed depth map. Then we formulate the energy function for the MRF with additional shift variables for each node. To reduce the high computational complexity of applying BP to this problem, we present a multi-level BPs by using BP with smaller numbers of label candidates for each level. Finally, the Poisson image reconstruction is applied to improve the color consistency along the boundary of the disocclusion region in the synthesized image. Experimental results demonstrate the performance of the proposed method on several publicly available datasets.**

I. Introduction

The revolution from 2D display to 3D display may be as significant as the revolution from monochrome to color display in the television history. several TV manufacturing companies have successively developed their advanced 3D display technologies that allow people to experience realistic 3D scenes without wearing 3D glasses. People can watch the 3D contents from different viewpoints by changing the viewing directions. Many companies have invested a considerable amount of resources in the related technology as the 3D industry has been foreseen to develop very rapidly in recent years.

Humans sense depth information due to parallax in the real world. Modern 3D display combines distinct image contents at different viewing directions to produce 3D effect. One practical way to create virtual views is from multiview images. This requires large bandwidth for multiview video transmission. An alternative is to generate virtual views from the image and corresponding depth map at a single view. This arises another problem that the regions covered by the foreground objects in the original view may be disoccluded when virtual views are generated. Thus, filling the disocclusions regions properly is very critical to achieve high-quality view synthesis results.

In this paper, we focus on the problem to generate the virtual videos of different viewpoints from a given video and associated depth maps at a viewpoint so that the viewers can realistically sense the depths. A spatio-temporal global optimization approach is proposed to synthesize the virtual views from a single video-plus-depth video sequence. The global energy minimization is formulated as the depth-based image completion problem for the video sequence to recover the disocclusion regions which are generated by the 3D image warping of different viewing directions. The disocclusion regions are recoverd from patches of the images by considering multiple important factors, such as depth maps, image structure and texture information. Then, the Belief Propagation (BP) [1] is applied to solve the energy minimization problem. Since belief propagation introduces high computational complexity which is proportional to the square of the total number of label candidates, we develop a multi-level optimization strategy with two BPs running with much smaller numbers of candidates to reduce the large number of label candidates. Affinity Propagation clustering (AP-clustering) [2] is used to cluster all the candidate patches in the image. Moreover, shift information is added in the Affinity Propagation clustering and the shift variables are considered into the energy function to increase the flexibility for the label candidates.

In the system framework, depth maps are pre-processed by a trilateral filter that jointly considers image intensity, depth and temporal consistency, followed by a bilateral filter that takes both depth information and spatial consistency into consideration. This not only reduces the noises but also enhances the coherence of the depth map in spatial and temporal domain. Finally, the Poisson image editing [3] is applied to maintain the color consistency in the disocclusion regions.

The main contribution of this paper can be summarized as follows. First, we propose a spatio-temporal global optimization approach that formulates an energy function considering depth map, image structure, texture information and patch shift. Second, we apply a multi-level BP by using AP-clustering in conjunction with patch shift variables to overcome the problem of the large number of label candidates in the optimization problem.

II. Related Works

More and more 3D display applications can be found in the high-tech products, including 3D LCD/LED displays, 3D laptops, 3D cameras, mobile devices and home video/games, etc. The suitable file formats to support these modern 3D devices have become one of the most important issues.

The video-plus-depth format is commonly used in the 3DTV community. It consists of the color intensity and the associated per-pixel depth map at one view. Based on this format, the DIBR (Depth-Image-Based Rendering) system [4] produced

virtual views based on the three steps: preprocessing of depth image, image warping and hole filling. However, the major problem in DIBR is how to fill the holes caused by the disocclusion regions in which the occluded pixels in the source view may become visible in the virtual views.

Under the video-plus-depth format, some research works focused on the preprocessing of the depth image [5] to reduce the disocclusion regions. Others developed hole filling methods based on image inpainting techniques [6] to fill in the disocclusion regions. For preprocessing of the depth image, the common approach is to apply the smoothness filters, (e.g. Gaussian filter and average filter) to the depth image. After image warping with the smoothed depth image, the disocclusion regions may be split into several small holes. Then the color interpolation can be used to fill in the small hole regions. Zhang et al. [5] extended the idea of the depth preprocessing for the hole filling from the symmetric smoothing filter to the asymmetric smoothing in order to reduce the geometric distortion. For image inpainting, Oh et al. [6] proposed a hole filling method by using depth-based inpainting. This method is designed by combining the depth-based hole filling and the image inpainting technique.

Because hole filling is the major problem in depth-image-based rendering. Image inpainting is a technique widely utilized to recover the disocclusion regions. The objective is to fill the unknown regions in a plausible way. Recent exemplar-based approaches are commonly used in image inpainting. In [7], a fast algorithm was presented to propagate texture and structure in a small patch. The success of structure propagation was dependent on the order in which the filling proceeds. The confidence value in the synthesized pixel values was propagated in a manner similar to the propagation of information in inpainting.

Contrary to the greedy methods, some approaches formulate the image completion as discrete global optimization problems [8] [9] [10]. In [8], image completion was automatically solved using an efficient BP algorithm. However, it did not consider structure information and thus the results may contain structure inconsistency. In [9], the image was completed with manually added structure information. Huang et al. [10] improved the hole filling method in [8] by adding the structure information into the global optimization formulation and solved the optimization problem with a two-step BP. In their method, only a single image was considered for the completion. Then, Liu et al. extended the method to the video completion [11] by adding the motion information to keep spatial and temporal coherence.

III. Proposed Disocclusion Region Recovering Method

Given the video frames and their corresponding depth maps, we aim to fill in the disocclusion regions by searching suitable patches from the video. Thus, the disocclusion region recovering problem can be reagrded as an exemplar-based labeling problem. In this section, we will show how this problem can be formulated considering spatial and temporal consistency of video frames in a Markov Random Field (MRF) framework. In addition, a multi-level BP with AP-clustering approach is proposed to solve the global optimization problem efficiently.

A. Problem Formulation

We first define the regions in a video frame. Let Φ be the source region and Ω represent the disocclusion region of a video frame I. Then, we have $\Phi+\Omega = I$. Similarly, in a video sequence, denote Φ^t and Ω^t by the source and disocclusion region of a video frame in time t.

To fill the disocclusion regions, each frame I is uniformly sampled to obtain sampled pixels. Let pixel position set $V = \{\{v_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ contain all the pixel positions sampled from the disocclusion region. The patch set $B = \{\{b_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ is the set of patches and each b_i^t is a patch centered at the position v_i in frame t. Note that patch b_i^t is likely to contain pixels in the source region even though the patch center v_i^t is in the disocclusion region. The goal of view synthesis is to fill in the disocclusion region by selecting patches from a set of patch candidates to the locations centered at the position in V. Define E as the set of edges connecting neighboring pixels, an undirected graph $g = \{V, E\}$ can be constructed for the MRF framework. Four neighbors are used within a video frame to constrain the spatial consistency whereas the four nearest neighboring pixels of the corresponding sampled pixel in the next frame (decided by motion estimation) are used for the temporal constraint.

B. Global Energy Minimization by MRF

Let $X = \{\{x_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ be the set of labels for V. For the set of patch candidates, let $M = \{m_l\}_{l=1}^{L}$ be the set of labels of all the patch candidates and p_l be the patch candidate. Our goal is to find the best label set X for V that minimizes the energy function under the spatial and temporal constraints, where $x_i^t \in M$. When $x_i^t = m_l$, it means the the label for the pixel position v_i in video frame time t is m_l. In other words, patch p_l is selected to fill in the region centered at v_i. In our method, we consider the patch candidates using $shift\ variable$ $(\Delta x, \Delta y)$ in the labeling problem (which will be discussed later in Section III-D). In short, by introducing $(\Delta x, \Delta y)$, patches are selected in a local region of (x, y) with a small amount of shift to reduce errors from spatial sampling for patch candidates. Since each patch has its own shift amount, let $\Delta X = \{\{\Delta x_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ and $\Delta Y = \{\{\Delta y_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ be the set of shift variables along x- and y-directions for V. Then, the global optimization problem can be formulated as follows:

$$E(X, \Delta X, \Delta Y) = E_s(X, \Delta X, \Delta Y) + \eta E_t(X, \Delta X, \Delta Y), \tag{1}$$

where the spatial term $E_s(.)$ and the temporal term $E_t(.)$ enforce the spatial and temporal constraint in this formulation, and η is a constant used to balance these two energy term.

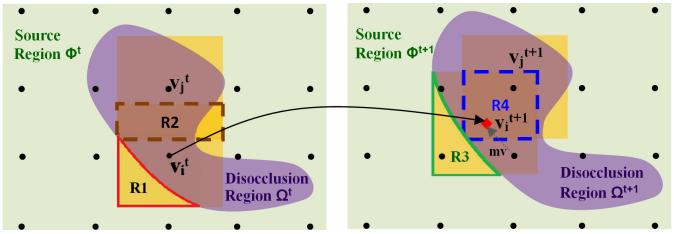

Fig. 1. Illustration of overlapping parts when calculating the spatial and temporal terms in the proposed MRF formulation. $R1$ shows the overlapping region of the patch centered at v_i^t and Φ^t. $R2$ covers the overlapping region from v_i^t and its neighboring patch centered at v_j^t. $R3$, $R4$ in video frame I^{t+1} are similar to $R1$, $R2$, respectively.

1) The Spatial Consistency: The spatial term $E_s(.)$ measures the consistency from color, depth and structure information. It is formed by combining the data term and the smoothness term as follows:

$$E_s(X,\Delta X,\Delta Y)=\sum_{v_i^t} E_s^{data}(x_i^t,\Delta x_i^t,\Delta y_i^t)+ \sum_{(v_i^t,v_j^t)\in e_s} E_s^{smo}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t), \tag{2}$$

where the data term $E_s^{data}(.)$ is the cost for labeling the pixel v_i^t as x_i^t with shift variable $\Delta x_i^t, \Delta y_i^t$, the smoothness term $E_s^{smo}(.)$ is the consistency cost for labeling the neighboring patch pair as x_i^t, x_j^t, and e_s is the set of a 4-neighborhood system. As shown in Figure 1, when a patch p_l is selected to fill in the location centered at v_i^t, the overlapping region $R1$ in the selected patch should be as similar as that in the patch b_i^t. On the other hand, the overlapping region from the patch candidate for position v_i^t and the patch candidate for its neighboring position v_j^t ($R2$ in this example) should be consistent as well.

Denote $p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)}$ to be the candidate patches for labeling pixel v_i^t adopting the shift variables. The spatial consistency is considered both in the color image and in the depth map. We use the pre-processed depth image which will be discussed later in Section IV-A. With the depth information, denote $q_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)}$ to be the candidate patches extracted from depth map. Similarly, let $D=\{\{d_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ be the set of patches, for each d_i^t is a patch centered at the position v_i in frame t of the depth map. The data cost for labeling x_i^t is defined as:

$$E_s^{data}(x_i^t,\Delta x_i^t,\Delta y_i^t)=D_{color}(p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},b_i^t) +D_{depth}(q_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},d_i^t), \tag{3}$$

where $D_{color}(.)$ measures the color difference between the overlapping region of candidate patch $p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)}$ and patch b_i^t in the color image. The difference can be obtained by the sum of the squared differences (SSD) or other metrics. In the case of using SSD, the function $D_{color}(.)$ with shift variables can be defined as:

$$D_{color}(p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},b_i^t)= \sum_{(x,y)\in W}|p_{x_i^t}(x+\Delta x,y+\Delta y)-b_i^t(x,y)|, \tag{4}$$

where W is a patch window. Similarly, $D_{depth}(.)$ calculates the depth difference from those regions from the depth map. If b_i^t is completely inside the disocclusion region, $E_s^{data}(x_i^t,\Delta x_i^t,\Delta y_i^t)$=0.

Human perception is sensitive to the discontinuity in the high contrast areas in the image, such as object boundary and strong texture. We measure the image continuity with the coherence between a pair of neighboring patches by incorporating the impact of texture and structure information in the smoothness term as follows:

$$E_s^{smo}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t)= \alpha E_s^{tex}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t) +\beta E_s^{str}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t) \tag{5}$$

where the texture term $E_s^{tex}(.)$ is used to constrain the texture consistency between the selected patches for the position v_i^t and its neighbor v_j^t. The structure term $E_s^{str}(.)$ enforces the consistency for structure propagation. α and β are two constants used to balance the two terms $E_s^{tex}(.)$ and $E_s^{str}(.)$. In our method, the color image and the depth map are also considered in $E_s^{tex}(.)$ which is defined by:

$$E_s^{tex}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t)= D_{color}(p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},p_{x_j^t}^{(\Delta x_j^t,\Delta y_j^t)}) +D_{depth}(q_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},q_{x_j^t}^{(\Delta x_j^t,\Delta y_j^t)}) \tag{6}$$

where the term $D_{color}(.)$ and $D_{depth}(.)$ are the SSD of the overlapping region of color and depth patches, respectively. The structure term is computed by:

$$E_s^{str}(x_i^t,\Delta x_i^t,\Delta y_i^t,x_j^t,\Delta x_j^t,\Delta y_j^t)= D_{grad_x}(p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},p_{x_j^t}^{(\Delta x_j^t,\Delta y_j^t)}) +D_{grad_y}(p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)},p_{x_j^t}^{(\Delta x_j^t,\Delta y_j^t)}) \tag{7}$$

where $D_{grad_x}(.)$ and $D_{grad_y}(.)$ are the gradients differences between the overlapping region of the patches $p_{x_i^t}^{(\Delta x_i^t,\Delta y_i^t)}$ and $p_{x_j^t}^{(\Delta x_j^t,\Delta y_j^t)}$ along the x- and y-directions, respectively. The gradient of a patch is defined as the maximum gradient of the pixels in the patch to represent the structure information.

2) The Temporal Consistency: The temporal term constrains that two corresponding patches in neighboring frames should have consistent color and depth information. Define the temporal corresponding point of pixel v_i^t in frame t as $\hat{v}_i^{t+1}$ in frame $t+1$. The patch centered of $\hat{v}_i^{t+1}$ is defined as $\hat{b}_i^{t+1}$. The temporal neighboring pixels of v_i^t are defined by

the four sampled pixels nearest to $\hat{v}_i^{t+1}$. We denoted them as $(v_i^t, \hat{v}_j^{t+1}) \in e_t$. In our method, the temporal correspondence can be found via motion estimation. Details for finding the temporal correspondence will be discussed in Section III-C.

Similar to the spatial term, when a patch is selected to fill in the location centered at v_i^t, the overlapping region in the source area $R3$ of the patch candidate should be similar to that in the temporal corresponding patch $\hat{b}_i^{t+1}$ in frame $t+1$. In addition, the overlapping region $R4$ should be consistent with its neighbors. Thus, the definition of the temporal term is expressed as the sum of two parts:

$$E_t(X, \Delta X, \Delta Y) = \sum_{v_i^t} E_t^{data}(x_i^t, \Delta x_i^t, \Delta y_i^t) + \sum_{(v_i^t, \hat{v}_j^{t+1}) \in e_t} E_t^{smo}(x_i^t, \Delta x_i^t, \Delta y_i^t, x_j^{t+1}, \Delta x_j^{t+1}, \Delta y_j^{t+1}), \tag{8}$$

where $E_t^{data}(.)$ represents the temporal cost of region consistency between patch candidate and its temporal corresponding patch, $E_t^{smo}(.)$ measures the cost between patch candidate and its temporal neighboring patches in frame $t+1$, and e_t represents the set of temporal neighbors. Similar to $E_s^{data}(.)$ and $E_s^{smo}(.)$, $E_t^{data}(.)$ and $E_t^{smo}(.)$ considering the temporal consistency both in color and depth image can be defined as:

$$E_t^{data}(x_i^t, \Delta x_i^t, \Delta y_i^t) = D_{color}(p_{x_i^t}^{(\Delta x_i^t, \Delta y_i^t)}, \hat{b}_i^{t+1}) + D_{depth}(q_{x_i^t}^{(\Delta x_i^t, \Delta y_i^t)}, \hat{d}_i^{t+1}) \tag{9}$$

$$E_t^{smo}(x_i^t, \Delta x_i^t, \Delta y_i^t, x_j^{t+1}, \Delta x_j^{t+1}, \Delta y_j^{t+1}) = D_{color}(p_{x_i^t}^{(\Delta x_i^t, \Delta y_i^t)}, p_{x_j^{t+1}}^{(\Delta x_j^{t+1}, \Delta y_j^{t+1})}) + D_{depth}(q_{x_i^t}^{(\Delta x_i^t, \Delta y_i^t)}, q_{x_j^{t+1}}^{(\Delta x_j^{t+1}, \Delta y_j^{t+1})}) \tag{10}$$

C. Constrained Optical Flow

In our energy function, the temporal constraint is applied to two corresponding patches in sequential frames. The correspondence can be found via the constrained optical flow.

Typically, optical flow describes apparent motion of objects, surfaces, and edges between images. The original optical flow formulation proposed by Horn and Schunck [12] minimizes the energy function of the difference of intensity values with a global smoothness constraint.

Later, Hsieh et al. [13] proposed to include optical flow constraints at some selected points and solve a constrained optical flow estimation problem. This constrained optimization problem is solved very efficiently with an Incomplete Cholesky Preconditioned Conjugate Gradient (ICPCG) algorithm [14]. We estimate the optical flow by the above method.

If dense motion is estimated, the correspondences for all patches in a video can be constructed. After image warping, the motion values of each pixel will also be warped to the new location. The disocclusion regions will not have the motion information. In order to obtain a completed motion map, we recover the motion map by using the same method as that of recovering the depth images, discussed in Section IV-A. In practical, the depth images and the motion maps can be recovered simultaneously.

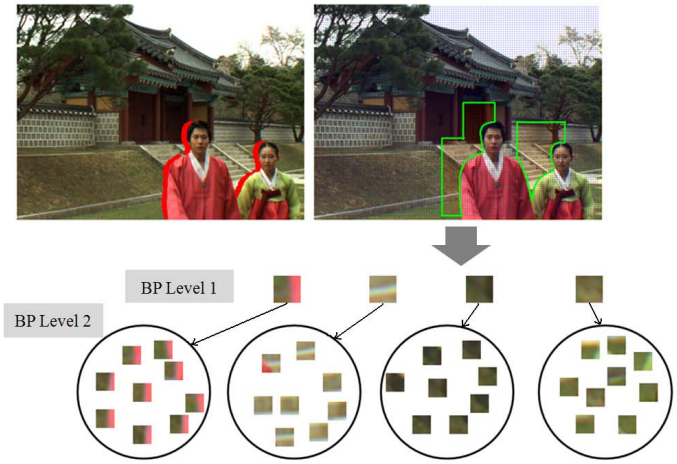

Fig. 2. An example for candidate patch selection and two-level BP.

D. Patch Selection and Shift Variables

The process of candidate patches selection in our method is described as follows: The regions for candidate patches are determined first. The minimal and maximal x- and y-coordinates of pixels in the disocclusion region are used to decide the candidate region. Since the patches for filling the disocclusion region are considered from the background area, the foreground region is excluded from the candidate region. The region is then extended by z pixels. We uniformly sample every r pixel in the candidate region. Patches centered from these sampled pixels are extracted. If there are multiple disocclusion regions, patches are extracted from the union of the candidate regions, as depicted in Figure 2. These sampled patches are used as the candidate patches in the MRF optimization framework. Note that the sampling is employed to reduce the total number of labels. Otherwise, the optimization problem becomes too large to solve in practice.

Since the patches are selected from the sampled pixels, this may decrease the accuracy for selecting the correct patch for the hole region. We introduce the shift variables $(\Delta x, \Delta y)$ associated with each selected patch. This makes the selection of the correct patch in the local neighborhood around sampled pixel (x, y). By setting $\Delta x = s$ and $\Delta y = s$, it searches all the patches centered at pixels in the $s \times s$ local window centered at pixel (x, y) when solving the proposed MRF energy function.

Although we used sampled pixels to extract patches, the total number of candidate patches, i.e. labels, is still too large in the MRF model for practical use. In the next section, we will introduce a multi-level BP algorithm to overcome this problem.

E. Multi-Level Belief Propagation

In our method, a multi-level BP optimization with Affinity propagation clustering method is proposed to reduce the computation for solving the optimization of MRF. The main idea is to perform BP in multiple levels, in which the the size of label candidates are much smaller than that of the original label candidates. Take two-level BP for example, BP is performed twice with patch set P_1 and P_2, respectively. P_1 and P_2 are much smaller than the original number of patch candidates P when BP is performed only once.

Affinity Propagation clustering (AP-clustering) [2] is first applied to cluster all the patches in P into C^1 clusters. Denote the center of each cluster $c_1^1, c_2^1, \ldots, c_{p^1}^1$. Thus we have the new label set $M^1 = \{m_1, m_2, \ldots, m_{p^1}\}$. Take the patches from C^1 cluster centers as the patch candidates and minimize the energy function using the standard BP with the label set M^1 to find the best label configuration $X^1 = \{x_1^1, x_2^1, \ldots, x_N^1\}$, where $x_i^1 \in M^1, 1 \leq i \leq N$.

Then we perform BP again in the second level. Suppose that after the first-level BP, the best label candidates for node v_1 is m_2. In the second round BP, the new label candidates for node v_1 are all elements belonging to the cluster with center c_2^1. Using associated label candidate sets for nodes of different clusters, the second-level BP is used to find the best label configuration to refine the patch filling result, as depicted in Figure 2.

There are two kinds of messages to be updated during each iteration of AP-clustering, namely responsibility and availability. Each of them accounts for a different kind of competition. Briefly speaking, responsibility updated lets all candidate exemplars compete for ownership of a data point while availability update collects evidence from data points reflecting the competence of each candidate exemplar.

We define the following negative real-valued similarity measure between any two patches, taking into account color difference and the shift information of each patch:

$$s(i,k) = d(p_i^t, p_k^t), \tag{11}$$

$$\bar{s}(i,k) = -\min_{\Delta x, \Delta y} s(p_i^{(\Delta x, \Delta y, t)}, p_k^t), \tag{12}$$

where d is Euclidean distance which measure the color difference of two patches p_i^t and p_k^t. And we add the shift information into this equation to find the minimized similarity of each shift. After some iterations, we can obtain the clusters of the label candidates.

Note that such a multi-level BP scheme may lead to a solution different from that obtained with the standard BP. However, the most important benefit of this scheme is that it significantly reduce the computational cost. This scheme can also be used to speed up the MRF optimization procedure for other MRF-based applications with a large number of labels.

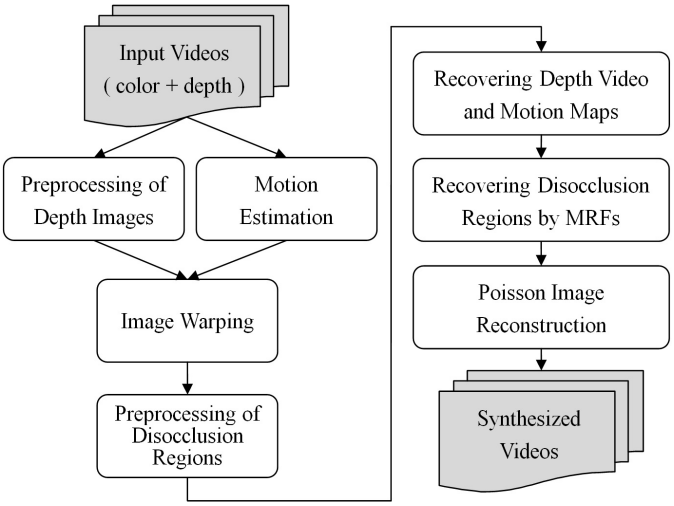

Fig. 3. Flow chart of system framework for the proposed spatio-temporally consistent view synthesis.

IV. System Framework Of Spatio-Temporally Consistent View Synthesis

The system framework of the proposed spatio-temporally consistent view synthesis contains several steps. The flow diagram is illustrated in Figure 3. Given the color video frames and their corresponding depth maps, the first three steps include preprocessing, motion estimation and image warping. The main focus of this work is the step of recovering the disocclusion regions in the color frames by the proposed MRF formulation. The other steps, including recovering the depth, are briefly discussed in the following subsections.

A. Preprocessing of Depth Images

In order to reduce the noise disturbance as well as to preserve the accuracy in the depth image, we apply a trilateral filter in the depth preprocessing procedure. In [15], a trilateral filter is extended from the bilateral filter [16]. Except for spatial filtering, it employs additional temporal information and color distance into the filter. Thus, it not only reduces the noise but also preserves the edge structure. Moreover, it enhances the temporal coherence. The adaptive filter weight $w_{T(u,v,t)}(\Delta u, \Delta v, \Delta t)$ for pixel (u, v) at time t in the trilateral filter is determined by the spatial and temporal displacement $(\Delta u, \Delta v, \Delta t)$ in the local window, as well as the corresponding color dissimilarity.

B. Image Warping

Image warping is usually required in depth-based view synthesis. In this problem, image warping is to map the pixel position to the corresponding location in the desired view based on associated depth map. In autostereoscopic display, image warping is generally degenerated to one-dimensional displacement along the horizontal scanline based on the assumption that the human eyes are in parallel to the screen at the same horizontal line when watching the display. Thus, we

simplify the description of image warping to one-dimensional displacement along the horizontal line in this work.

We warp the images and the associated depth maps to the position of the desired view in accordance with the user provided parameter setting. Figure 4 illustrated the warping results (c)(d) of the original color image and its depth map (a)(b) without depth preprocessing. (e)(f) shows the warping results when the depth images are preprocessed. Users should set the desired view and the relation between input disparities and the desired disparity in terms of pixel unit. A higher input disparity means it is closer to viewers and a lower input disparity indicates it is farther away from viewers. In our warping method, we assume the desired disparity is proportional to the input disparity. The linear relation can be obtained by two different points. Thus, the relation between the physical location u' and the origin location u for each pixel in horizontal direction can be described by the following equation:

$$u' = u + round(v(d_0 + (D(u,v) - g_0) \cdot \frac{d_1 - d_0}{g_1 - g_0})), \quad (13)$$

where v is the relative position of the desired view, (d_0, d_1) is the desired disparity range in terms of pixel unit on the display, and (g_0, g_1) is the input disparity range.

C. Preprocessing of Disocclusion Region

After the image warping based on the depth information, it contains internal empty regions in the warped images due to the depth discontinuities. These regions can be classified into two types, image cracks and disocclusion regions. Image cracks are generally caused by noises or digital numerical precision, whereas disocclusion regions come from the sharp depth discontinuity. As depicted In Figure 4 (e)(f), there are some image cracks in the left-half of the images. The red parts along two people in the right-half of the images shows the disocclusion regions. The procedure of disocclusion region preprocessing is mainly aimed at removing the image cracks. Here, the mean filter is applied to fill the cracks. Figure 4 (g)(h) depict the results after crack interpolation.

We can observe some mixed pixels appeared in the disocclusion regions around the background boundary. Usually, they may cause the color inconsistency during view synthesis. To remove the errors, we extended the holes boundaries by using image dilation in these regions.

D. Recovering Depth Images

After the preprocessing, we will recover the depth images first since the disocclusion regions of the depth images are easier to estimate than those of the color images. Most disocclusion regions are close to the background boundary. In our method, the disocclusion regions should be filled with the background regions.

Histograms are used to accumulate the depth values of pixels in the disocclusion regions. According to the histogram, we consider two conditions. In the first condition, there are only two peaks in the histogram. It means there are two main

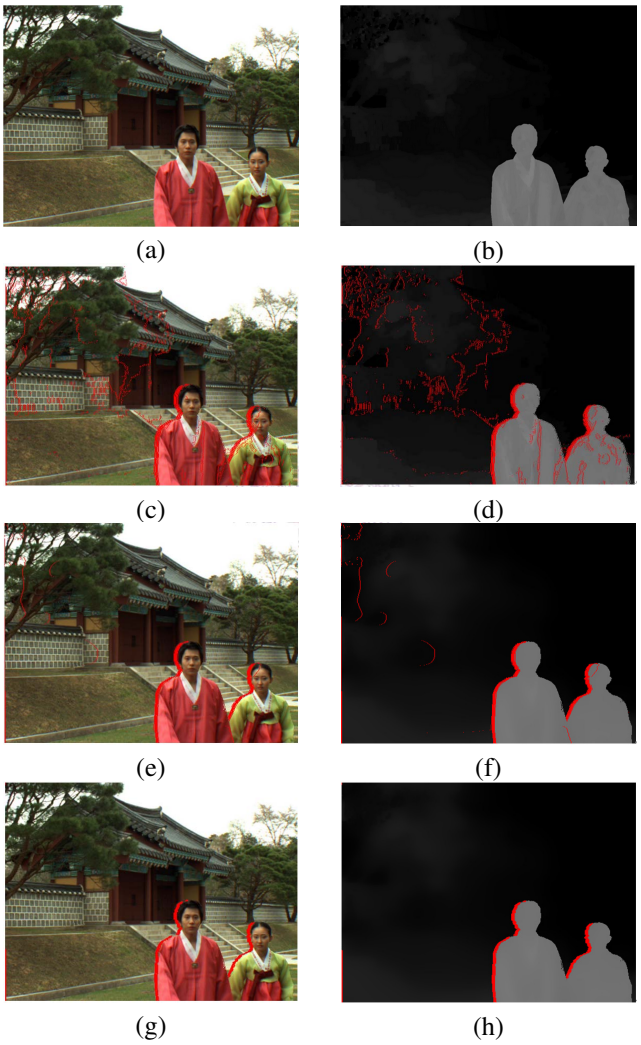

Fig. 4. (a) An original color frame and the associated depth map in (b). (c)(d) The warped color frame and the depth map without depth preprocessing (e)(f) The warped color frame and depth map with depth preprocessing (g)(h) The color and depth map after crack interpolation.

depth values in this disocclusion region. The higher intensity of depth value represents the foreground object and the lower one indicates the background depth value. Thus, we will fill the disocclusion region with the depth value of the low peak (background depth value). In the other situation, when there are more than two peaks in the histogram, multiple depth values are contained in the disocclusion region. We recover this disocclusion region considering the user-specified view direction. According to Section IV-B, users set the desired view and the view direction. We fill the disocclusion region of the depth image with its left side region or right side region based on the view direction.

E. Poisson Image Reconstruction

After recovering the disocclusion regions by the MRF optimization, the result may still contain the color inconsistency. We thereby reconstruct the image from the gradient fields by solving the Poisson equation. Poisson image reconstruction [3]

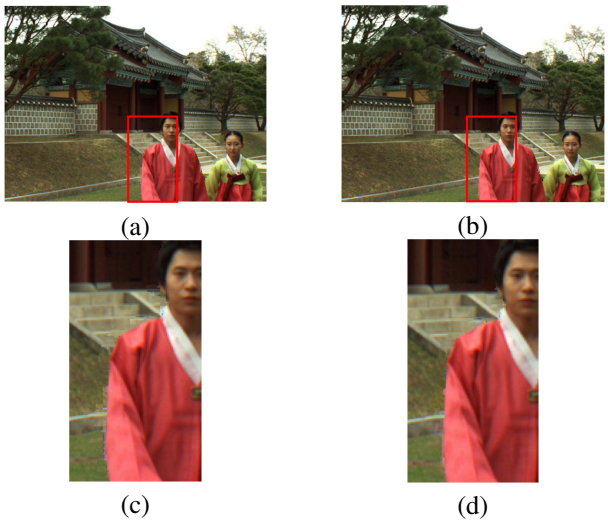

Fig. 5. The view synthesis results after hole filling in (a)&(c). (b)&(d) show the results after Poison reconstruction.

Fig. 6. (a) Warped color image. The synthesized results (b) without and (c) with shift variables $(\Delta x, \Delta y)$ in the energy function.

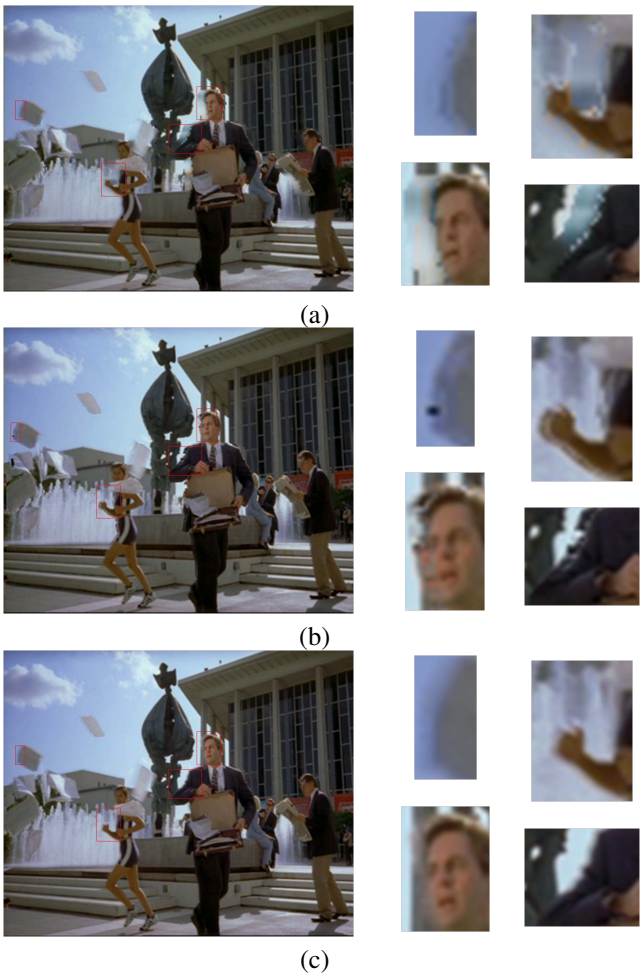

Fig. 7. The view synthesis results from *Samsung* by (a) Silva, (b) VSRS and (c) the proposed method.

was proposed to reconstruct the image to ensure the compliance of source and destination boundaries. For discrete images, the problem can be discretized naturally using the discrete pixel grid. Since the shape of boundary in the view synthesis problem can be arbitrary, we form the image reconstruction problem as solving a sparse, symmetric and positive-definite system. The Gauss-Seidel iterative method is applied to solve the linear system.

V. Experimental Results

In our experiments, we show several view synthesis results based on the proposed method. We used the following two publicly available video-plus-depth datasets in our experiments. The first is the *Statue dataset*. It is composed of a color video and an associated depth video provided from Zhang et al. [17]. It is a challenging video sequence since it was taken by a moving camera. It contains the sharp boundary of the statue as well as the smooth depth transition from the grassplot. The image size is 960 $\times$ 540. The second dataset is the *Lovebird1 dataset*. It is composed of color videos at eight different views and depth videos at three different views, which are provided by ETRI (Electronics and Telecommunications Research Institute) , MPEG-Korea Forum. The image size for each frame is 512 $\times$ 384. All of our experiments are conducted on a desktop PC with Intel Core2Duo 2.0 GHz CPU. The parameters used in our system include: η = 0.7, α = 0.3, β = 0.25 in the proposed MRF formulation. Shift variable δx and δy are set to 2. Level is set 2 in the multi-level BP. Patches are sampled every 5 pixels and the patch size is set to 9. We synthesized the color and depth images at a different view according to the user setting. In this case, we set (d_0, d_1) to (0, -15), (g_0, g_1) to (0, 255) and the desired view direction is -1. In our implementation, the two-level BP is used to speed up the MRF optimization. We utilized libDAI [18] to solve this global optimization problem.

Figure 5 demonstrates the results after Poisson image reconstruction. It can be observed that the color inconsistency at the occlusion boundary before the Poisson image reconstruction in Figure 5 (a) and (c). Figure 5 (b) and (d) show the correction of the color inconsistency after applying the Poisson reconstruction. Figure 6 shows the synthesized results with and without shift variables $(\Delta x, \Delta y)$. We can see the edge structure of the result with shift variables in Figure 6 (c) are better preserved than (b).

We compared our method with the DIBR-based method proposed by Silva et al. [19] and the MPEG View Synthesis Reference Software (VSRS) [20] for three video sequences

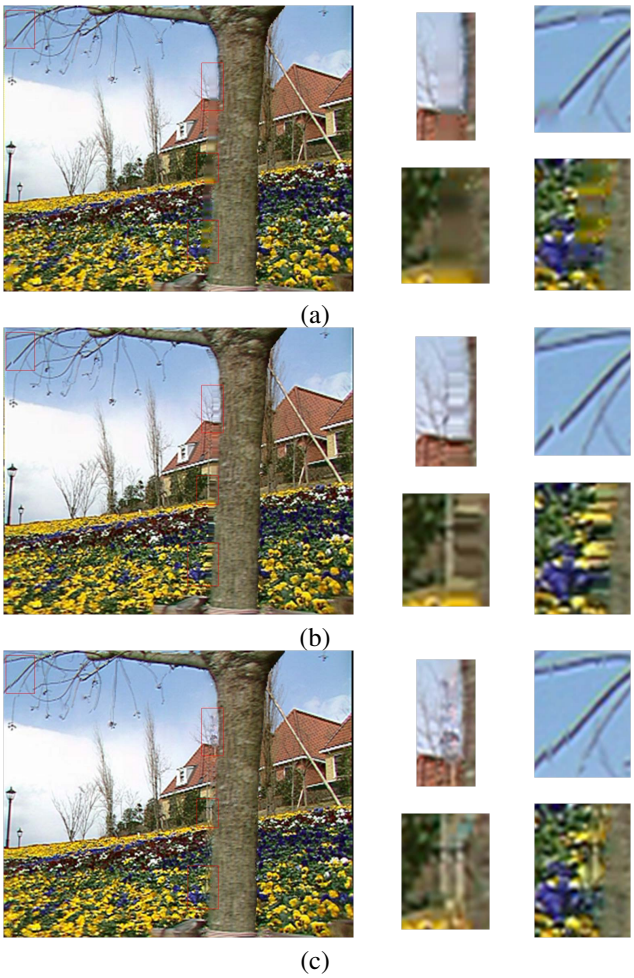

Fig. 8. The view synthesis results from $Tree$ by (a) Silva, (b) VSRS and (c) the proposed method.

$Samsung$, $Tree$ and $Temple$. In the Figure 7(a), we can notice obvious ghosting effect compared to the results of the proposed method. In the Figure 7(b), VSRS shows more artifacts around the object boundary of the man's face and arm than those of the proposed method. From the results of Figure 8, the disocclusion regions are kind of blurred which are recovered by Silva. Similarly, the proposed method provides more details in textured regions than those of VSRS.

VI. Conclusions

We proposed a novel view synthesis algorithm to generate spatio-temporally consistent videos from video-plus-depth information. The motion information is exploited for temporal term in the global exemplar-based optimization in an MRF framework. Based on the motion field, the images are recovered in a global exemplar-based scheme by minimizing an MRF energy function. The proposed energy function enforces both spatial and temporal consistency constraints in the recovery process. In addition, a two-level BP with AP clustering is proposed to solve the MRF minimization problem to reduce the computational complexity caused by the large number of the label candidates. Last, the Poisson image editing is applied to refine the reconstruction of disocclusion regions. Experimental results are shown to demonstrate the satisfactory view synthesis results from video-plus-depth data.

For the future work, the transformation on the label candidates can be formulated in the proposed MRF model. Thus, it can deal with not only translation motion with shift variables but also rotation and scaling. This will increase the variability for the label candidates (patches) and make the image synthesis results more visually plausible.

References

[1] J. Pearl, "Reverend bayes on inference engines: A distributed hierarchical approach," in *Conference on Association for the Advancement of Artificial Intelligence*, 1982, pp. 133–136.

[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[3] P. Perez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[4] W. R. Mark, L. McMillan, and G. Bishop, "Post-rendering 3d warping," in *Symposium on Interactive 3D graphics*, 1997, pp. 7–16, 180.

[5] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3d tv," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 191–199, 2005.

[6] K.-J. Oh, S. Yea, and Y.-S. Ho, "Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video," in *Proceedings of the 27th conference on Picture Coding Symposium*, 2009, pp. 233–236.

[7] A. Criminisi, P. Pérez, and K. Toyama, "Object removal by exemplar-based inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 721–728.

[8] N. Komodakis, "Image completion using global optimization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 442–452.

[9] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," *ACM Transactions on Graphics*, vol. 24, pp. 861–868, July 2005.

[10] H. Ting, S. Chen, J. Liu, and X. Tang, "Image inpainting by global structure and texture propagation," in *ACM International Conference on Multimedia*, 2007, pp. 517–520.

[11] M. Liu, S. Chen, J. Liu, and X. Tang, "Video completion via motion guided spatial-temporal global optimization," in *ACM International Conference on Multimedia*, 2009, pp. 537–540.

[12] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intellegence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[13] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "Expression-invariant face recognition with constrained optical flow warping," *IEEE Transactions on Multimedia*, vol. 11, pp. 600–610, June 2009.

[14] N. Foster and R. Fedkiw, "Practical animation of liquids," in *ACM SIGGRAPH*, 2001, pp. 23–30.

[15] S.-J. Lin, C.-M. Cheng, and S.-H. Lai, "Spatio-temporally consistent multi-view video synthesis for autostereoscopic displays," in *Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 532–542.

[16] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *International Conference on Computer Vision*, 1998, pp. 839–846.

[17] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, "Consistent depth maps recovery from a video sequence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 974–988, 2009.

[18] J. M. Mooij, "libdai: A free and open source c++ library for discrete approximate inference in graphical models," *J. Mach. Learn. Res.*, vol. 99, pp. 2169–2173, August 2010.

[19] D. V. S. X. D. Silva, W. A. C. Fernando, and H. K. Arachchi, "A new mode selection technique for depth maps of 3d video," in *ICASSP*, 2010, pp. 686–689.

[20] I. JTC1/SC29/WG11, "View synthesis reference software (vsrs), version 3.5," in *wg11.sc29.org*, 2009.