Consistent Depth Maps Recovery from a Video Sequence

Guofeng Zhang, *Student Member*, *IEEE*, Jiaya Jia, *Member*, *IEEE*, Tien-Tsin Wong, *Member*, *IEEE*, and Hujun Bao, *Member*, *IEEE*

Abstract—This paper presents a novel method for recovering consistent depth maps from a video sequence. We propose a bundle optimization framework to address the major difficulties in stereo reconstruction, such as dealing with image noise, occlusions, and outliers. Different from the typical multiview stereo methods, our approach not only imposes the photo-consistency constraint, but also explicitly associates the geometric coherence with multiple frames in a statistical way. It thus can naturally maintain the temporal coherence of the recovered dense depth maps without oversmoothing. To make the inference tractable, we introduce an iterative optimization scheme by first initializing the disparity maps using a segmentation prior and then refining the disparities by means of bundle optimization. Instead of defining the visibility parameters, our method implicitly models the reconstruction noise as well as the probabilistic visibility. After bundle optimization, we introduce an efficient space-time fusion algorithm to further reduce the reconstruction noise. Our automatic depth recovery is evaluated using a variety of challenging video examples.

Index Terms—Consistent depth maps recovery, multiview stereo, bundle optimization, space-time fusion.

1 Introduction

Stereo reconstruction of dense depth maps from natural video sequences is a fundamentally important and challenging problem in computer vision. The reconstructed depths usually serve as a valuable source of information, and facilitate applications in various fields, including 3D modeling, layer separation, image-based rendering, and video editing. Although the stereo matching problem [31], [19], [32], [52] has been extensively studied during the past decades, automatically computing high-quality dense depths is still difficult on account of the influence of image noise, textureless regions, and occlusions that are inherent in the captured image/video data.

Given an input video sequence taken by a freely moving camera, we propose a novel method to automatically construct a view-dependent depth map for each frame with the following two objectives. One is to make the corresponding depth values in multiple frames *consistent*. The other goal is to assign *distinctive* depth values for pixels that fall in different depth layers. To accomplish these goals, this paper contributes a global optimization scheme, which we call *bundle optimization*, to resolve most of the aforementioned difficulties in disparity estimation. This framework allows us to produce sharp and temporal consistent object boundaries among different frames.

- *G. Zhang and H. Bao are with the State Key Lab of CAD&CG, Zijingang Campus, Zhejiang University, Hangzhou, 310058, P.R. China. E-mail: {zhangguofeng, bao}@cad.zju.edu.cn.*
- *J. Jiya and T.-T. Wong are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: {leojia, ttwong}@cse.cuhk.edu.hk.*

Manuscript received 8 Aug. 2008; revised 3 Jan. 2009; accepted 19 Feb. 2009; published online 24 Feb. 2009.
Recommended for acceptance by S.B. Kang.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-08-0474.
Digital Object Identifier no. 10.1109/TPAMI.2009.52.

Our method does not explicitly model the binary visibility (or occlusion). Instead, it is encoded naturally in a statistical way with our energy definition. Our model also does not distinguish among image noise, occlusions, and estimation outliers, so as to achieve a unified framework for modeling the matching ambiguities. The photo-consistency and geometric coherence constraints associating different views are combined in a global energy minimization framework. They help reliably reduce the influence of image noise and occlusions with the multiframe data, and consequently, make our optimization free from the oversmoothing or blending artifacts.

In order to get an accurate disparity estimate in the textureless region and reduce the problem of false segmentation especially for the fine object structures, we confine the effect of color segmentation only in the disparity initialization step. Then, our iterative optimization algorithm refines the disparities in a pixelwise manner.

We have conducted experiments on a variety of challenging examples and found that our method is robust against occlusions, noise, and estimation outliers. The automatically computed depth maps contain very little noise and preserve fine structures. One challenging example is shown in Fig. 1, in which the scene contains large textureless regions, objects with strong occlusions, grassplot with smooth depth change, and a narrow bench. Our method faithfully reconstructs all these structures. Readers are referred to our supplementary video (http://www.cad.zju.edu.cn/home/gfzhang/projects/videodepth) for inspecting the preserved temporal consistency among the recovered dense depth maps.

2 Related Work

Since our system contains several components, such as global optimization, image segmentation, bundle optimization, and

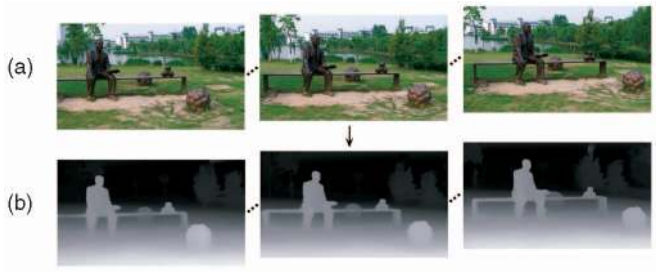

Fig. 1. High-quality depth maps recovered from the "Lawn" sequence. (a) An input video sequence taken by a moving camera. (b) The depth maps automatically estimated by our method. The sharp boundary of the statue, as well as the grassplot with smooth depth transition, are accurately constructed in the depth maps.

space-time fusion, we separately discuss the relevant previous work in the following sections.

2.1 Global and Local Optimization in Multiview Stereo

Multiview stereo algorithms [28], [6], [19], [52], [16] estimate depth (or disparity) with the input of multiple images. Early approaches [28], [6] used local and window-based methods, and employed a local "winner-takes-all" (WTA) strategy in depth estimation. Later on, several global methods [22], [39], [19] formulate the depth estimation as an energy-minimization problem and use graph cuts or belief propagation to solve it. Most of these methods adopt the first-order smoothness priors. For the slanted and curved 3D surfaces, methods in [44], [26], [46] incorporate the second-order smoothness prior for stereo reconstruction. Recently, Woodford et al. [46] proposed an effective optimization strategy that employs triple cliques to estimate depth.

However, it is known that the global optimum is not always computationally reachable. Even the state-of-the-art numerical optimizers, such as loopy belief propagation and multilabel graph cuts, cannot guarantee to produce the globally optimal solution in energy minimization [4], [23], [43]. In addition, given the matching ambiguity in the textureless regions or occlusion boundaries, the key to improving the depth estimates is an appropriate energy definition. For an oversimplified (or problematic) definition, even using the method that can yield the global optimum cannot improve much the depth estimates. With this observation, in this paper, we introduce a novel data term that combines the photo-consistency and geometric coherence constraints in a statistical way. Our experiments demonstrate that it is rather effective to improve the depth estimation around the occlusion boundaries and in the textureless regions.

2.2 Segmentation-Based Approaches

By assuming that the neighboring pixels with similar colors have similar depth values, segmentation-based approaches [42], [8], [47], [21], [40] were proposed to improve the depth estimation for large textureless regions. These methods typically model each segment as a 3D plane and estimate the plane parameters by matching small patches in neighboring views [47], [21], or using a robust fitting algorithm [42]. In [2], non-fronto-parallel planes are constructed on sparse 3D points obtained by structure-from-motion. Gallup et al. [13] used the sparse points to determine the plane directions for the three orthogonal sweeping directions. Zitnick and Kang [52] proposed an oversegmentation method to lower the risk of spanning a segment over multiple layers. However, even with over-segmentation or soft segmentation, accurate disparity estimate is still difficult to obtain especially in the textured regions and along the segment boundaries.

2.3 Occlusion Handling

Occlusion handling is another major issue in stereo matching. Methods in [20], [19], [35], [38], [36] explicitly detect occlusions in disparity estimation. Kang and Szeliski [19] proposed a hybrid method that combines shiftable windows, temporal selection, and explicit occluded-pixel labeling, to handle occlusions in dense multiview stereo within a global energy minimization framework.

Visibility maps are commonly used to indicate whether a pixel in one image is also visible in another. Each pixel in the map has a value of 0 or 1, indicating being occluded or not, respectively. Several algorithms [35], [19], [38] iteratively estimate the disparities (or depths) and visibilities. This strategy is effective if the amount of occlusions or outliers is relatively small. Strecha et al. [36] jointly modeled depth and visibility in a hidden Markov random field, and solved the problem using an expectation-maximization algorithm. The state of each pixel is represented as a combination of discrete depth and visibility. This method yields a good performance given a small set of wide-baseline images. However, for a video sequence containing many frames, a large amount of state variables makes the inference intractable.

2.4 Multiview Stereo Methods for Reconstructing 3D Models

Multiview stereo (MVS) methods were developed to reconstruct 3D object models from multiple input images. A survey can be found in [32]. Many of these methods (e.g., voxel-based approaches [33], [45]) aim to build a 3D model for a single object and are usually not applicable to large-scale sceneries due to the high computational complexity and memory space requirement. The approaches based on multiple depth maps [35], [36], [5] are more flexible, requiring fusing view-dependent depth maps into a 3D model. In these methods, the visibility or geometric coherence constraint is typically used only for fusion. To obtain a 3D surface representation of an object, Hernández et al. [18] proposed a probabilistic framework to model geometric occlusion in a probabilistic way. Recently, Merrell et al. [27] described a quick depth map fusion method to construct a consistent surface among multiple frames. They introduced two fusion strategies, namely, the stability-based and confidence-based fusions, based on the visibility constraint and confidences. Zach et al. [48] proposed a range image integrating method based on minimizing an energy functional incorporating a total variation (TV) regularization term and an L^1 data fidelity term. This method is globally convergent. For some MVS methods using level-set or deformable polygonal meshes [9], [49], the geometric coherence constraint is incorporated and formulated in 3D.

However, these methods typically need a good starting point (e.g., a visual hull model [25]).

2.5 Recovering Consistent View-Dependent Depth Maps

Instead of reconstructing a complete 3D model, we focus on recovering a set of consistent view-dependent depth maps from a video sequence in this paper. It is mainly motivated by applications such as view interpolation, depth-based segmentation, and video enhancement. Our work is closely related to that of [19], [15], which also aims to infer consistent depth maps from multiple images. Kang and Szeliski [19] proposed simultaneously optimizing a set of depth maps at multiple key frames by adding a temporal smoothness term. This method makes the disparities across frames vary smoothly. However, it is sensitive to outliers and may cause the blending artifacts around object boundaries. Gargallo and Sturm [15] formulated 3D modeling from images as a Bayesian MAP problem, and solved it using the expectation-maximization (EM) algorithm. They use the estimated depth map to determine the visibility prior. Hidden variables are computed in a probabilistic way to deal with occlusions and outliers. A multiple-depth-map prior is finally used to smooth and merge the depths while preserving discontinuities. In comparison, our method statistically incorporates the photo-consistency and geometric coherence constraints in the data term definition. This scheme is especially effective for processing video data because it can effectively suppress temporal outliers by making use of the statistical information available from multiple frames. Moreover, we use efficient loopy belief propagation [10] to perform the overall optimization. By combining the photo-consistency and geometric coherence constraints, the distribution of our data cost becomes distinctive, making the BP optimization stable and converge quickly.

The temporal coherence constraints were also used in optical flow estimation [1] and occlusion detection [30], [37]. Larsen et al. [24] presented an approach for 3D reconstruction from multiple synchronized video streams. In order to improve the final reconstruction quality, they used optical flow to find corresponding pixels in the subsequent frames of the same camera, and enforced the temporal consistency in reconstructing successive frames. With the observation that the depth error in conventional stereo methods grows quadratically with depth, Gallup et al. [14] proposed a multibaseline and multiresolution stereo method to achieve constant depth accuracy by varying the baseline and resolution proportionally to depth.

In summary, although many approaches have been proposed to model 3D objects or to estimate depths using multiple input images, the problem of how to appropriately extract information and recover consistent depths from a video remains challenging. In this paper, we show that by appropriately maintaining the temporal coherence, surprisingly consistent and accurate dense depth maps can be obtained from the video sequences. The recovered depth maps have high quality and are readily usable in many applications such as 3D modeling, view interpolation, layer separation, and video enhancement.

1. **Structure from Motion**:
   1.1 Recover the camera parameters.
2. **Disparity Initialization**:
   2.1 For each frame, apply loopy belief propagation to minimize (5).
   2.2 Use image segmentation to improve the initial disparity estimate.
3. **Bundle Optimization**:
   3.1 Process frames from 1 to n:
   For each frame t, fix disparities in other frames and refine D_t by minimizing (1).
   3.2 Repeat step 3.1 for two passes.
4. **Space-Time Fusion**:
   4.1 Perform space-time fusion.

Fig. 2. Overview of our method.

3 Framework Overview

Given a video sequence $\hat{I}$ with n frames taken by a freely moving camera, we denote $\hat{I} = \{I_t \mid t = 1, \ldots, n\}$, where $I_t(\mathbf{x})$ represents the color (or intensity) of pixel $\mathbf{x}$ in frame t. It is either a 3-vector in a color image or a scalar in a grayscale image. In our experiments, we assume it is an RGB color vector. Our objective is to estimate a set of disparity maps $\hat{D} = \{D_t \mid t = 1, \ldots, n\}$. By convention, disparity $D_t(\mathbf{x})$ ($d_\mathbf{x}$ for short) is defined as $d_\mathbf{x} = 1/z_\mathbf{x}$, where $z_\mathbf{x}$ is the depth value of pixel $\mathbf{x}$ in frame t. For simplicity, the terms "depth" and "disparity" are used interchangeably in the following sections.

The set of camera parameters for frame t in a video sequence is denoted as $\mathbf{C_t} = \{\mathbf{K_t}, \mathbf{R_t}, \mathbf{T_t}\}$, where $\mathbf{K}_t$ is the intrinsic matrix, $\mathbf{R}_t$ is the rotation matrix, and $\mathbf{T}_t$ is the translation vector. The parameters for all frames can be estimated reliably by the structure from motion (SFM) techniques [17], [29], [50]. Our system employs the SFM method of Zhang et al. [50].

In order to robustly estimate a set of disparity maps, we define the following energy in a video:

$$E(\hat{D}; \hat{I}) = \sum_{t=1}^{n} (E_d(D_t; \hat{I}, \hat{D} \backslash D_t) + E_s(D_t)), \qquad (1)$$

where the data term E_d measures how well disparity $\hat{D}$ fits the given sequence $\hat{I}$ and the smoothness term E_s encodes the disparity smoothness. For each pixel in disparity map D_t, because it maps to one point in 3D, there should exist corresponding pixels in other nearby frames. These pixels not only satisfy the photo-consistency constraint, but also have their geometric information consistent. We thus propose a *bundle optimization* framework to model the explicit correlation among the pixels and use the collected statistics to optimize the disparities jointly.

Fig. 2 gives an overview of our framework. With an input video sequence, we first employ the SFM method to recover the camera parameters. Then, we initialize the disparity map for each frame independently. Segmentation prior is incorporated into initialization for improving the disparity estimation in large textureless regions. After initialization, we perform bundle optimization to iteratively

refine the disparity maps. Finally, we use a space-time fusion to further reduce the reconstruction noise.

4 Disparity Initialization

With a video sequence input, we first initialize the disparity map for each frame independently. Denoting the disparity range as $[d_{\min}, d_{\max}]$, we equally quantize the disparity into $m+1$ levels, where the kth level $d_k = (m-k)/m \cdot d_{\min} + k/m \cdot d_{\max}, k = 0, \ldots, m$. So, the task in this step is to estimate an initial disparity d for each pixel. Similar to the traditional multiview stereo methods, using the photo-consistency constraint, we define the disparity likelihood as

$$L_{init}(\mathbf{x}, D_t(\mathbf{x})) = \sum_{t'} p_c(\mathbf{x}, D_t(\mathbf{x}), I_t, I_{t'}),$$

where $p_c(\mathbf{x}, d, I_t, I_{t'})$ measures the color similarity between pixel $\mathbf{x}$ and the corresponding pixel $\mathbf{x}'$ (given disparity d) in frame t'. It is defined as

$$p_c(\mathbf{x}, d, I_t, I_{t'}) = \frac{\sigma_c}{\sigma_c + \|I_t(\mathbf{x}) - I_{t'}(l_{t,t'}(\mathbf{x}, d))\|}, \quad (2)$$

where σ_c controls the shape of our differentiable robust function. $\|I_t(\mathbf{x}) - I_{t'}(l_{t,t'}(\mathbf{x}, d))\|$ is the color L-2 norm. With these definitions, for each frame t, data term E_d^t is expressed as

$$E_d^t(D_t; \hat{I}) = \sum_{\mathbf{x}} 1 - u(\mathbf{x}) \cdot L_{init}(\mathbf{x}, D_t(\mathbf{x})), \quad (3)$$

where $u(\mathbf{x})$ is an adaptive normalization factor, and is written as

$$u(\mathbf{x}) = 1 / \max_{D_t(\mathbf{x})} L_{init}(\mathbf{x}, D_t(\mathbf{x})).$$

It makes the largest likelihood of each pixel always one, which is equivalent to imposing stronger smoothness constraint in the flat regions than in the textured ones.

The spatial smoothness term for frame t can be defined as

$$E_s(D_t) = \sum_{\mathbf{x}} \sum_{\mathbf{y} \in N(\mathbf{x})} \lambda(\mathbf{x}, \mathbf{y}) \cdot \rho(D_t(\mathbf{x}), D_t(\mathbf{y})), \quad (4)$$

where $N(\mathbf{x})$ denotes the set of neighbors of pixel $\mathbf{x}$, and λ is the smoothness weight. $\rho(\cdot)$ is a robust function:

$$\rho(D_t(\mathbf{x}), D_t(\mathbf{y})) = \min\{|D_t(\mathbf{x}) - D_t(\mathbf{y})|, \eta\},$$

where η determines the upper limit of the cost.

In order to preserve discontinuity, $\lambda(\mathbf{x}, \mathbf{y})$ is usually defined in an anisotropic way, encouraging the disparity discontinuity to be coincident with abrupt intensity/color change [11], [3], [4], [31], [35]. Our adaptive smoothness weight is defined as

$$\lambda(\mathbf{x}, \mathbf{y}) = w_s \cdot \frac{u_\lambda(\mathbf{x})}{\|I_t(\mathbf{x}) - I_t(\mathbf{y})\| + \varepsilon},$$

where $u_\lambda(\mathbf{x})$ is a normalization factor:

$$u_\lambda(\mathbf{x}) = |N(\mathbf{x})| \Big/ \sum_{\mathbf{y}' \in N(\mathbf{x})} \frac{1}{\|I_t(\mathbf{x}) - I_t(\mathbf{y}')\| + \varepsilon}.$$

w_s denotes the smoothness strength and ε controls the contrast sensitivity. Our adaptive smoothness term imposes smoothness in flat regions while preserving edges in textured ones.

Finally, the initial energy function for each frame t can be written as

$$E_{init}^t(D_t; \hat{I}) = \sum_{\mathbf{x}} \Bigg(1 - u(\mathbf{x}) \cdot L_{init}(\mathbf{x}, D_t(\mathbf{x})) + \sum_{\mathbf{y} \in N(\mathbf{x})} \lambda(\mathbf{x}, \mathbf{y}) \cdot \rho(D_t(\mathbf{x}), D_t(\mathbf{y})) \Bigg). \quad (5)$$

We minimize E_{init}^t to get the initial disparity estimates. Taking into account the possible occlusions, we employ the temporal selection method proposed in [19] to only select the frames in which the pixels are visible for matching. For each frame t, we then use loopy belief propagation [10] to estimate D_t by minimizing (5). Fig. 3b shows one frame result obtained in this step (i.e., step 2.1 in Fig. 2).

In order to better handle textureless regions, we incorporate the segmentation information into the disparity estimation. The segments of each frame are obtained by mean-shift color segmentation [7]. Similar to the nonfronto-parallel techniques [42], [38], we model each disparity segment as a 3D plane and introduce plane parameters $[a_i, b_i, c_i]$ for each segment s_i. Then, for each pixel $\mathbf{x} = [x, y] \in s_i$, the corresponding disparity is given by $d_{\mathbf{x}} = a_i x + b_i y + c_i$. Taking $d_{\mathbf{x}}$ into (5), E_{init}^t is formulated as a nonlinear continuous function w.r.t. the variables a_i, b_i, and c_i, $i = 1, 2, \ldots$. The partial derivatives over a_i, b_i, and c_i are required to be computed when applying a nonlinear continuous optimization method to estimate all 3D plane parameters. Note that $L_{init}(\mathbf{x}, d_{\mathbf{x}})$ does not directly depend on the plane parameters. We, therefore, apply the following chain rule:

$$\frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial a_i} = \frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}} \cdot \frac{\partial d_{\mathbf{x}}}{\partial a_i} = x \frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}}.$$

Similarly, $\frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial b_i} = y \frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}}$ and $\frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial c_i} = \frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}}$. In these equations, gradient $\frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}}$ is first computed on the quantized disparity levels:

$$\left.\frac{\partial L_{init}(\mathbf{x}, d_{\mathbf{x}})}{\partial d_{\mathbf{x}}}\right|_{d_k} = \frac{L_{init}(\mathbf{x}, d_{k+1}) - L_{init}(\mathbf{x}, d_{k-1})}{d_{k+1} - d_{k-1}},$$

where $k = 1, \ldots, m$. Then, a continuous version of $L_{init}(\mathbf{x}, d_{\mathbf{x}})$ (denoted as $L_{init}^c(\mathbf{x}, d_{\mathbf{x}})$) is constructed by cubic-Hermite interpolation. Finally, the continuous partial derivatives are calculated on $L_{init}^c(\mathbf{x}, d_{\mathbf{x}})$.

With the parametric form $d_{\mathbf{x}} = a_i x + b_i y + c_i$, estimating disparity $d_{\mathbf{x}}$ is equivalent to optimizing plane parameters $[a_i, b_i, c_i]$. It is thus possible to use a nonlinear continuous optimization method to minimize the energy in (5). Initial 3D plane parameters can be obtained by the nonfrontoparallel plane extraction method [42]. In experiments, we adopt a simpler method which can produce sufficiently satisfactory plane parameters with less computational time. Particularly, for each segment s_i, we first set $a_i = 0$ and $b_i = 0$ by assuming a frontoparallel plane. The disparity values in all other segments are fixed. Then, we compute a

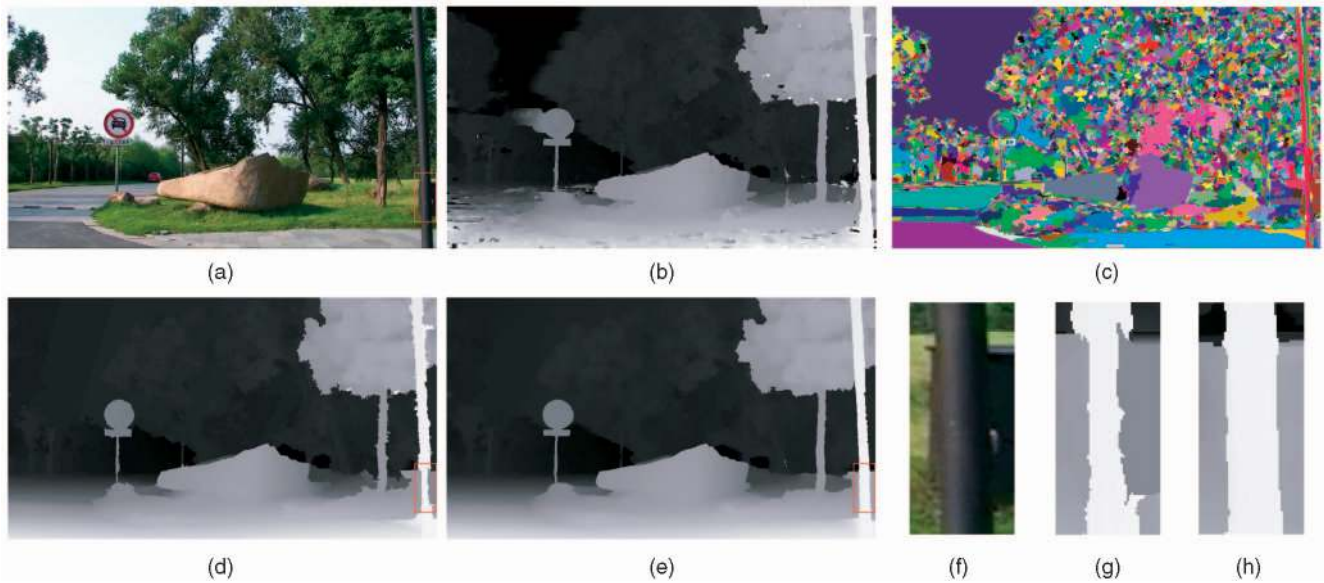

Fig. 3. Disparity estimation illustration. (a) One frame from the "Road" sequence. (b) The initial estimate after solving (5) by belief propagation without incorporating segmentation. (c) Segmentation prior incorporated in our initialization. (d) Disparity initialization with segmentation and plane fitting using a nonlinear continuous optimization. (e) Our refined disparities after bundle optimization. (f)-(h) Magnified regions from (a), (d), and (e), showing that our bundle optimization improves disparity estimate significantly on object boundaries.

set of c_i with different assignments of d_k, where $k = 0, \ldots, m$, and select the best c_i^* that minimizes (5). After getting c_i^*, we unfreeze a_i and b_i, for $i = 0, 1, 2, \ldots$, and use the Levenberg-Marquardt method to reestimate them by solving the function in (5). When all plane parameters are estimated, the disparities in each segment can be obtained accordingly. We show in Fig. 3 one frame from the "Road" example. Fig. 3c shows the incorporated segmentation in initialization. The disparity estimated from the initialization step is shown in Fig. 3d.

5 Bundle Optimization

In the disparity initialization step, we perform color segmentation and estimate the disparity map for each frame independently. It is widely known that segmentation is a double-edged sword. On one hand, segmentation-based approaches regularize the disparity estimate in large textureless regions. On the other hand, they inevitably introduce errors in textured regions and do not handle well the situation that similar-color pixels are with different disparity values. Figs. 3d and 3g show that there are visual artifacts along the occlusion boundaries. Our initialization independently estimates the disparity maps, which are not necessarily consistent among each other. This easily causes flicker during video playback.

In this section, we propose using the geometric coherence constraint to associate each video frame to others, and introduce bundle optimization to refine the disparity maps. The corresponding disparity estimate is iteratively refined by simultaneously imposing the photo-consistency and geometric coherence constraints.

5.1 The Energy Function

We define a new energy function for (1). Compared to (5), only the data term is largely modified. This is based on a common observation that data term usually plays an essential role in energy minimization. If the data costs for the majority of the pixels are not informative, the corresponding solution to the stereo problem will be ambiguous since the resultant minimal cost in (1) may refer simultaneously to multiple results that are quantitatively and visually quite different. For example, if the data term only measures color similarity, strong matching ambiguity for pixels in the textureless areas will be the result. One may argue that the smoothness term has an effect of regularizing the solver. However, this term only functions as compromising the disparity of one pixel to its neighborhood and does not contribute much to inferring the true disparity values.

One objective of defining the new data term is to handle occlusion. In our approach, we reduce the influence of occlusions and outliers by collecting both the color and geometry information statistically over multiple frames. More specifically, in a video sequence, if the disparity of a pixel in a frame is mistakenly estimated due to either occlusion or other problems, the projection of this pixel to other frames using this incorrect disparity has a small probability of satisfying both the photo-consistency and geometric coherence constraints simultaneously. With this intuition in mind, we define the data term in the following way.

Considering a pixel $\mathbf{x}$ in frame t, by epipolar geometry, the matching pixel in frame t' should lie on the conjugate epipolar line. Given the estimated camera parameters and the disparity $d_{\mathbf{x}}$ for pixel $\mathbf{x}$, we compute the conjugate pixel location in $I_{t'}$ by multiview geometry and express it as

$$\mathbf{x}'^h \sim \mathbf{K}_{t'}\mathbf{R}_{t'}^{\top}\mathbf{R}_t\mathbf{K}_t^{-1}\mathbf{x}^h + d_{\mathbf{x}}\mathbf{K}_{t'}\mathbf{R}_{t'}^{\top}(\mathbf{T}_t - \mathbf{T}_{t'}), \qquad (6)$$

where the superscript h denotes the vector in the homogeneous coordinate system. The 2D point $\mathbf{x}'$ is computed by dividing $\mathbf{x}'^h$ by the third homogeneous coordinate. We

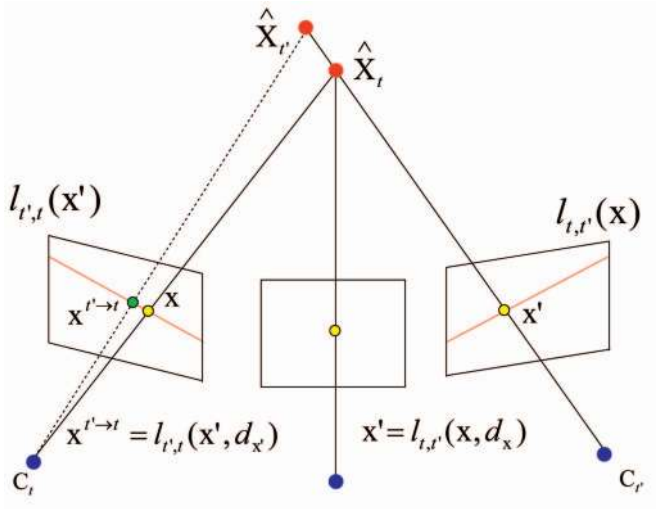

Fig. 4. Geometric coherence. The conjugate pixel of $\mathbf{x}$ in frame t' is denoted as $\mathbf{x}'$ and lies on the conjugate epipolar line. Ideally, when we project $\mathbf{x}'$ from frame t' back to t, the projected pixel should satisfy $\mathbf{x}^{t'\to t} = \mathbf{x}$. However, in disparity estimation, because of the the matching error, $\mathbf{x}^{t'\to t}$ and $\mathbf{x}$ are possibly in different positions.

denote the mapping pixel in frame t' of $\mathbf{x}$ as $\mathbf{x}' = l_{t,t'}(\mathbf{x}, d_{\mathbf{x}})$. The mapping $l_{t',t}$ is symmetrically defined. So, we also have $\mathbf{x}^{t'\to t} = l_{t',t}(\mathbf{x}', d_{\mathbf{x}'})$, as illustrated in Fig. 4.

If there is no occlusion or matching error, ideally, we have $\mathbf{x}^{t'\to t} = \mathbf{x}$. So, we define the likelihood of disparity d for any pixel $\mathbf{x}$ in I_t by combining two constraints:

$$L(\mathbf{x}, d) = \sum_{t'} p_c(\mathbf{x}, d, I_t, I_{t'}) \cdot p_v(\mathbf{x}, d, D_{t'}), \tag{7}$$

where $p_v(\mathbf{x}, d, D_{t'})$ is the proposed geometric coherence term measuring how close pixels $\mathbf{x}$ and $\mathbf{x}^{t'\to t}$ are, as shown in Fig. 4. It is defined as

$$p_v(\mathbf{x}, d, D_{t'}) = \exp\left(-\frac{\|\mathbf{x} - l_{t',t}(\mathbf{x}', D_{t'}(\mathbf{x}'))\|^2}{2\sigma_d^2}\right) \tag{8}$$

in the form of a Gaussian distribution, where σ_d denotes the standard deviation. The definition of p_c is given in (2). Our geometric coherence term is similar to the symmetric constraint used in two-view stereo [38] and the geometric visibility prior in [15].

Both the photo-consistency and geometric coherence constraints make use of the information of the corresponding pixels mapped from t' to t. But, they constrain the disparity from two different aspects. In the following paragraphs, we briefly explain why there is no need to explicitly model occlusion or visibility.

Our likelihood requires a correct disparity estimate to satisfy two conditions simultaneously, i.e., high photo-consistency as well as high geometric coherence for the corresponding pixels. We use the following example to explain how the data term ensures the reliable depth estimation. Suppose we compute the disparity likelihood of pixel $\mathbf{x}$ in frame t. A correct disparity d makes $p_c(\mathbf{x}, d, I_t, I_{t'}) \cdot p_v(\mathbf{x}, d, D_{t'})$ output a large value for several neighboring frames t'. An arbitrary d other than that has small chance to find similar consistent support from neighboring frames and, thus, can be regarded as noise. Combining the computed likelihood for all possible disparities, a highly nonuniform cost distribution for each pixel can be obtained favoring the correct disparity.

We also found that this model performs satisfactorily around depth discontinuous boundaries. The reason is similar to that given above. Specifically, we use color segmentation and plane fitting to initialize depths independently on each frame. So, the corresponding pixels in multiple frames are possibly assigned to the correct or incorrect depth segments. Even if we only obtain a few correct depth estimates for the corresponding pixels, it sufficiently makes $\sum_{t'} p_c(\mathbf{x}, d, I_t, I_{t'}) \cdot p_v(\mathbf{x}, d, D_{t'})$ output a relatively large value for the correct disparity d. Therefore, our data energy, in many cases, can form a highly nonuniform cost distribution where the likelihood of the correct depth is large.

In [19], an extratemporal smoothness term is introduced outside the data term, which functions similarly to the spatial smoothness constraint. It compromises the disparities temporally, but does not essentially help the inference of true disparity values.

To fit the energy minimization framework, our data term E_d is finally defined as

$$E_d(D_t; \hat{I}, \hat{D} \backslash D_t) = \sum_{\mathbf{x}} 1 - u(\mathbf{x}) \cdot L(\mathbf{x}, D_t(\mathbf{x})), \tag{9}$$

where $u(\mathbf{x})$ is an adaptive normalization factor, and is expressed as

$$u(\mathbf{x}) = 1 / \max_{D_t(\mathbf{x})} L(\mathbf{x}, D_t(\mathbf{x})).$$

It makes the largest likelihood of each pixel always one.

5.2 Iterative Optimization

With the above energy definition, we iteratively refine the depth estimate using loopy belief propagation. The segmentation prior is not used in this step and we, instead, perform pixel-wise disparity refinement to correct the error.

Each pass starts from frame 1. With the concern of computational complexity, in refining disparity map D_t, we fix the disparity values in all other frames. The data term only associates frame t with about 30-40 neighboring frames. One pass completes when the disparity map of frame n is optimized. In our experiments, after the first-pass optimization, the noise and estimation errors are dramatically reduced. Fig. 3e shows one depth map. Two passes are usually sufficient to generate temporally consistent depth maps in our experiments.

6 Space-Time Fusion

Bundle optimization can largely improve the quality of the recovered disparity maps in a video sequence. But, it does not completely eliminate the reconstruction noise. In this section, we describe a *space-time fusion* algorithm to reduce the remaining noise due to inevitable disparity quantization, video resolution, and other estimation problems. The disparity consistency error, after space-time fusion, can be decreased to an even lower fraction.

Our space-time fusion makes use of the sparse feature points in 3D computed by structure-from-motion and the depth correspondences from multiview geometry. Based on

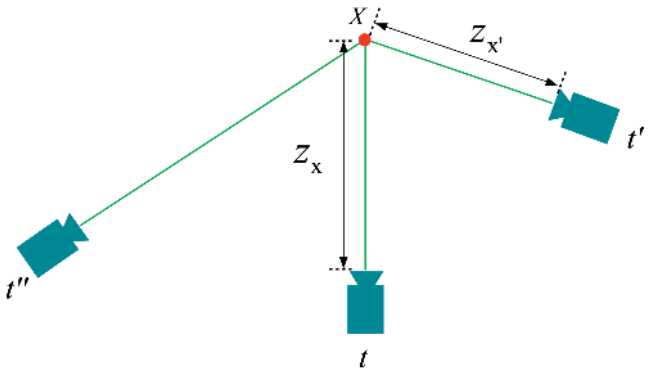

Fig. 5. Illustration of multiview geometry. X is a 3D point. $\mathbf{x}$ and $\mathbf{x}'$ are its projections in frames t and t', respectively. $z_{\mathbf{x}}$ and $z_{\mathbf{x}'}$ are the corresponding depth values.

the estimated $D_t(x,y)$ for each pixel $I_t(x,y)$ from the bundle optimization step, we attempt to compute the fused disparity maps $D^* = \{D_t^* \mid t = 1, \ldots, n\}$ with three groups of constraints.

6.1.1 Spatial Continuity

Depths computed by bundle optimization contain many correctly inferred depth structures, such as edges and smooth transitions. To preserve them in the final depth results, we require the first-order derivatives of the space-time fused depths to be similar to those from bundle optimization. So, the spatial constraints for every two neighboring pixels in D_t^* are defined as

$$\begin{aligned} D_t^*(x+1,y) - D_t^*(x,y) &= D_t(x+1,y) - D_t(x,y), \\ D_t^*(x,y+1) - D_t^*(x,y) &= D_t(x,y+1) - D_t(x,y). \end{aligned} \tag{10}$$

6.1.2 Temporal Coherence

Because depth values are view-dependent, one point in 3D is possibly projected to multiple frames. Using Fig. 5 as an example, if a 3D point X projects to $\mathbf{x}$ and $\mathbf{x}'$ in frames t and t', respectively, the corresponding depth values $z_{\mathbf{x}}$ and $z_{\mathbf{x}'}$ should be correlated by a transformation with the computed camera parameters. It is written as

$$(x_{\mathbf{x}'}, y_{\mathbf{x}'}, z_{\mathbf{x}'})^\top = z_{\mathbf{x}} \mathbf{R}_{t'}^\top \mathbf{R}_t \mathbf{K}_t^{-1} \mathbf{x}^h + \mathbf{R}_{t'}^\top (\mathbf{T}_t - \mathbf{T}_{t'}), \tag{11}$$

where $\mathbf{K}$ is the intrinsic matrix, $\mathbf{R}$ is the rotation matrix, and $\mathbf{T}$ is the translation vector. The transformation can be simplified to $z_{\mathbf{x}'} = A(\mathbf{x}) \cdot z_{\mathbf{x}} + B$, where $A(\mathbf{x})$ and B are determined by pixel $\mathbf{x}$ and the camera parameters.

Our temporal constraint is based on the above depth correlation in multiframes. Considering frames t and $t+1$, we denote the corresponding pixel in frame $t+1$ to $I_t(x,y)$ as $(x^{t\to t+1}, y^{t\to t+1})$. We accordingly define the *disparity consistency error* as

$$e = \left\| \frac{1}{A(x,y) + B \cdot D_t(x,y)} D_t(x,y) - D_{t+1}(x^{t\to t+1}, y^{t\to t+1}) \right\|,$$

which measures the disparity consistency error between D_t and D_{t+1}. We plot in Figs. 6a and 6b the average disparity consistency errors for different frames in the "Angkor Wat" and "Road" sequences. It shows that the recovered disparities after bundle optimization are already temporally consistent. The average error is only about $0.003 * (d_{\max} - d_{\min})$. By visual inspection, the pixels that cause abnormally large errors are mostly occlusions. If the error is above a threshold τ (i.e., $e > \tau$), we regard the correspondence as "unreliable."

Finally, the temporal constraint is defined for each *reliable* correspondence as

$$\alpha \cdot \left(\frac{D_t^*(x,y)}{A(x,y) + B \cdot D_t^*(x,y)} - D_{t+1}^*(x^{t\to t+1}, y^{t\to t+1}) \right) = 0, \tag{12}$$

where α is a weight, and is set to 2 in our experiments.

6.1.3 Sparse Feature Correspondences

Our SFM step has estimated a sparse set of 3D feature points $\mathbf{S}$. These 3D points are view-independent, and therefore, can be used as anchors to constrain different views with the geometric correlations.

For a 3D point $X \in \mathbf{S}$, its projection and the corresponding disparity in frame t are, respectively, denoted as $\mathbf{u}_X^t$ and d_t^X. $\mathbf{u}_X^t$ is given by

$$\mathbf{u}_X^t = \mathbf{K}_t (\mathbf{R}_t^\top X - \mathbf{R}_t^\top \mathbf{T}_t),$$

with the estimated camera parameters $\mathbf{K}_t$, $\mathbf{R}_t$, and $\mathbf{T}_t$ for frame t. We similarly define "reliable" projection from X to frame t if $\|D_t(\mathbf{u}_X^t) - d_t^X\| < \kappa$, where κ is a threshold. The feature correspondence requires, for all pixels that correspond to reliable 3D features, the refined disparity values should be similar to those of the features in each frame. The constraint is thus written as

$$\beta \cdot \left(D_t^*(\mathbf{u}_X^t) - d_t^X \right) = 0, \tag{13}$$

where $\beta = 100$ in all our experiments. It should be noted that the above three constraints are all necessary to make space-time fusion solvable. The spatial continuity constraint is to preserve depth structures, such as edges and depth details. The temporal coherence constraint is to make the disparity temporally consistent. The sparse feature correspondences help refine the depths making use of the reliable 3D point information.

Because (12) is nonlinear, to make the computation efficient, we employ an iterative optimization method and introduce a substitute for (12) that is defined as

$$\alpha \cdot \left(\frac{D_t^*(x,y)}{A(x,y) + B \cdot \tilde{D}_t^*(x,y)} - D_{t+1}^*(x^{t\to t+1}, y^{t\to t+1}) \right) = 0, \tag{14}$$

where $\tilde{D}_t^*(x,y)$ is the estimate of $D_t^*(x,y)$ from the previous iteration and is initialized as $D_t(x,y)$.

With (10), (13), and (14), in each iteration, we solve a linear system using the conjugate gradient solver. With the concern of the memory consumption, each time we perform space-time fusion in a slab of 5-10 frames. For example, with an interval of 5 frames, we first fuse frames 1 to 5, then we fix frames 1 to 4 and fuse frames 5 to 9, etc.

We analyze the disparity errors using the "Angkor Wat" and "Road" sequences, and plot them in Fig. 6. We introduce two measures—that is, the disparity consistency error between adjacent frames and the disparity error with respect to the sparse 3D feature points. Figs. 6a and 6b show

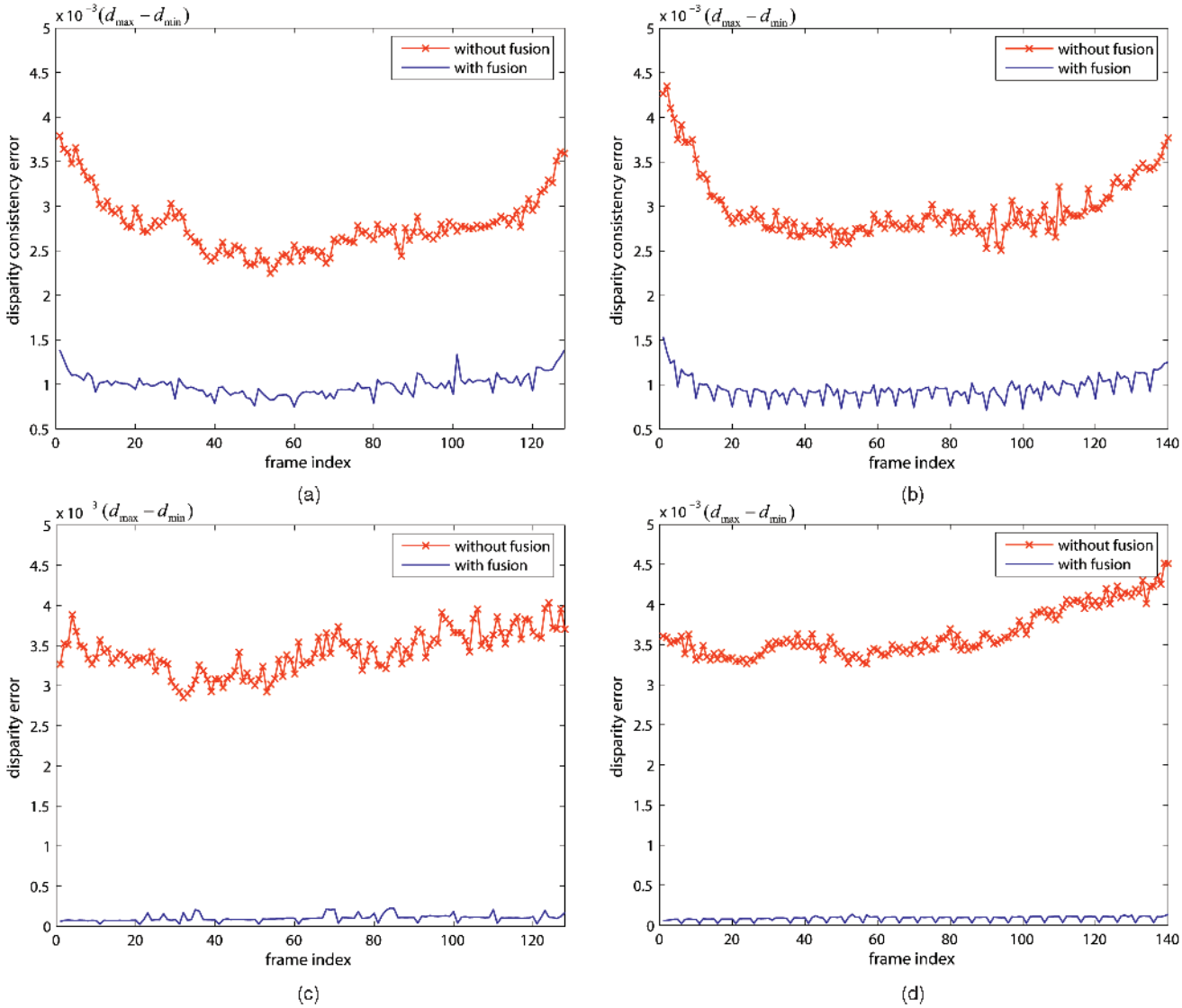

Fig. 6. Disparity error measures on the "Angkor Wat" and "Road" examples. The red/blue curve shows the average errors without/with space-time fusion. (a) and (b) Disparity consistency error. We compute the average error between consecutive frames. Without space-time fusion, the average disparity consistency error of these two examples is around $0.3\% \cdot (d_{\max} - d_{\min})$. After our space-time fusion, both of them are reduced to around $0.1\% \cdot (d_{\max} - d_{\min})$. (c) and (d) Disparity error w.r.t. the sparse 3D points obtained in the SFM step. Our space-time fusion also largely reduces it.

the average disparity consistency error for each frame. In computing the disparity errors, since we do not have the ground truth disparity maps, the computed sparse 3D points in the SFM step are regarded as "correct" coordinates. For all reliable projections from the 3D points to a frame, average disparity error $\|D_t(\mathbf{u}_X^t) - d_t^X\|$ is calculated. The plot is shown in Figs. 6c and 6d. The comparison of the average errors shows that the space-time fusion is effective. It reduces the reconstruction noise and makes the recovered depth temporally more consistent.

7 Results and Analysis

To evaluate the performance of the proposed method, we have conducted experiments on several challenging video sequences. Table 1 lists the statistics of the tested sequences. All our experiments are conducted on a desktop PC with Intel Core2Duo 2.0 GHz CPU. Most of the parameters in our system use fixed values. Specifically, $w_s = 5/(d_{\max} - d_{\min})$, $\eta = 0.05(d_{\max} - d_{\min}), \varepsilon = 50, \sigma_c = 10, \alpha = 2, \beta = 100$. We also found $2 \leq \sigma_d \leq 3$ works well in our experiments. Its default value is 2.5. For depth estimation from a video sequence, we

TABLE 1
The Statistics of the Tested Sequences Shown in This Paper

Sequences	Lawn	Road	Flower	Angkor Wat	Temple	Stair	Great Wall	Garden	Statue	Fountain-P11
Frames	201	141	229	129	121	125	156	150	11	11
Resolution	960×540	960×540	960×540	576×352	576×352	576×352	576×352	352×240	960×540	768×512

TABLE 2
Running Time of Three Tested Sequences

Sequence	Resolution	Running Time		
		Initialization	BO	Fusion
Garden	352×240	1 min./frame	3 min./frame	1 sec./frame
Angkor Wat	576×352	3 min./frame	8 min./frame	3 sec./frame
Flower	960×540	7 min./frame	20 min./frame	10 sec./frame

set the thresholds $\tau = 0.03 * (d_{\max} - d_{\min})$, $\kappa = 0.01 * (d_{\max} - d_{\min})$. For the "Statue" (Fig. 14) and "Fountain-P11" (Fig. 15) examples, since they only contain a sparse set of input images, we set $\tau = 0.005 * (d_{\max} - d_{\min})$ and $\kappa = 0.003 * (d_{\max} - d_{\min})$. The maximum disparity level m is usually with value 300.

The running time of our method for different steps on three tested sequences is shown in Table 2. It is approximately proportional to the video resolution. For a sequence with frame resolution 576×352, our initialization needs 3 minutes for each frame. Bundle optimization with two passes takes about 8 minutes per frame and the major computation is spent on the data cost estimation considering all pixels in multiple frames. Space-time fusion is quick and only consumes about 3 seconds per frame.

7.1 Workflow Illustration

We first illustrate how our system recovers the depths for the "Angkor Wat" sequence in Fig. 7. The "Angkor Wat" sequence contains textureless regions with complex occlusions. In initialization, we first solve the energy function in (5) without incorporating segmentation. The estimated disparity map is shown in Fig. 7b. Then, we use the mean shift algorithm to segment each frame independently. Fig. 7c shows the segmentation result of one frame. By incorporating the segmentation prior and using plane fitting, the disparities are refined as shown in Fig. 7d. To eliminate the erroneous disparities introduced by segmentation, we perform bundle optimization. The result is shown in Fig. 7e. Our supplementary video contains the recovered disparity maps for the whole sequence, in which the temporal consistency is maintained. To further reduce the reconstruction noise, we finally perform space-time fusion. The result is shown in Fig. 7f. Due to the limited 256 gray levels reproduced in the figure, the visual difference of the maps produced using and without using space-time fusion is not obvious. Readers are referred to Fig. 6 for a numerical comparison.

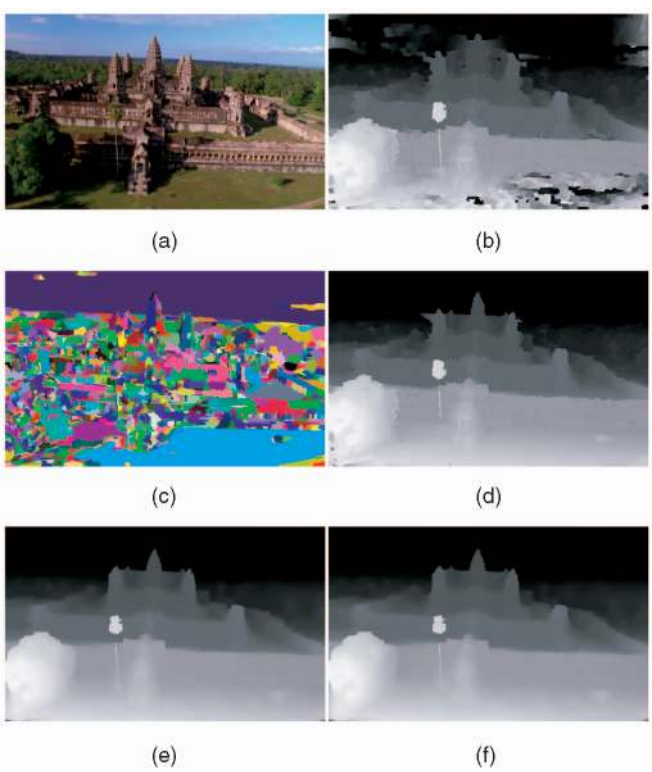

Fig. 7. Workflow illustration. (a) One frame from the "Angkor Wat" sequence. (b) Disparity initialization after only solving energy function (5). (c) Segmentation prior incorporated in our initialization. (d) Initialization result after segmentation and plane fitting. (e) The disparity result of bundle optimization. The estimate is improved significantly on object boundary. (f) The final disparity map after space-time fusion.

Fig. 8. "Road" sequence taken by a hand-held DV camera moving along a road. Row (a) shows a few frames, and row (b) shows the correspondingly estimated depth maps.

7.2 Results of Ordinary Video Sequences

The "Road" sequence shown in Fig. 8 is taken by a handheld video camera. The scene contains textureless sky and road. Different objects occlude each other and the road is with smooth depth change. The video even contains the thin posts of the traffic sign and street lamp. Our method faithfully reconstructs all these structures. To verify the quality of the recovered depth, we synthesize new images from different views using the 3D warping technique. Since the depth information is available for every pixel, we can create a dense grid where each grid point corresponds to a pixel. Then, we connect the neighboring grids for triangulation, excluding the discontinuous edges where the disparity contrast is larger than a threshold. With the grids, we build the texture-mapped scene surface, and render novel images by projecting the pixels in one frame to a new view. The synthesized images are shown in Fig. 9. They accurately preserve object boundary and the relation of occlusions.

Another "Flower" example is shown in Fig. 10. The sequence contains 229 frames. It is also challenging for depth estimation because the occlusion is complex and there exist narrow stems and small leaves in different depth layers. Our recovered depth maps are shown in Fig. 10b. Similar to the previous example, to demonstrate how accurate our depth estimates are, we construct the texture-mapped scene surface with the computed depth map, and synthesize novel views from different viewpoints, as shown in Fig. 11.

Fig. 9. Novel view synthesis with the recovered depth maps. (a) One frame extracted from the "Road" sequence. (b) The estimated depth map. (c) and (d) With the depth information, we build the texture-mapped scene surface, and synthesize new images with different view angles.

Fig. 10. "Flower" sequence. (a) Extracted frames from the input sequence. (b) The estimated depth maps.

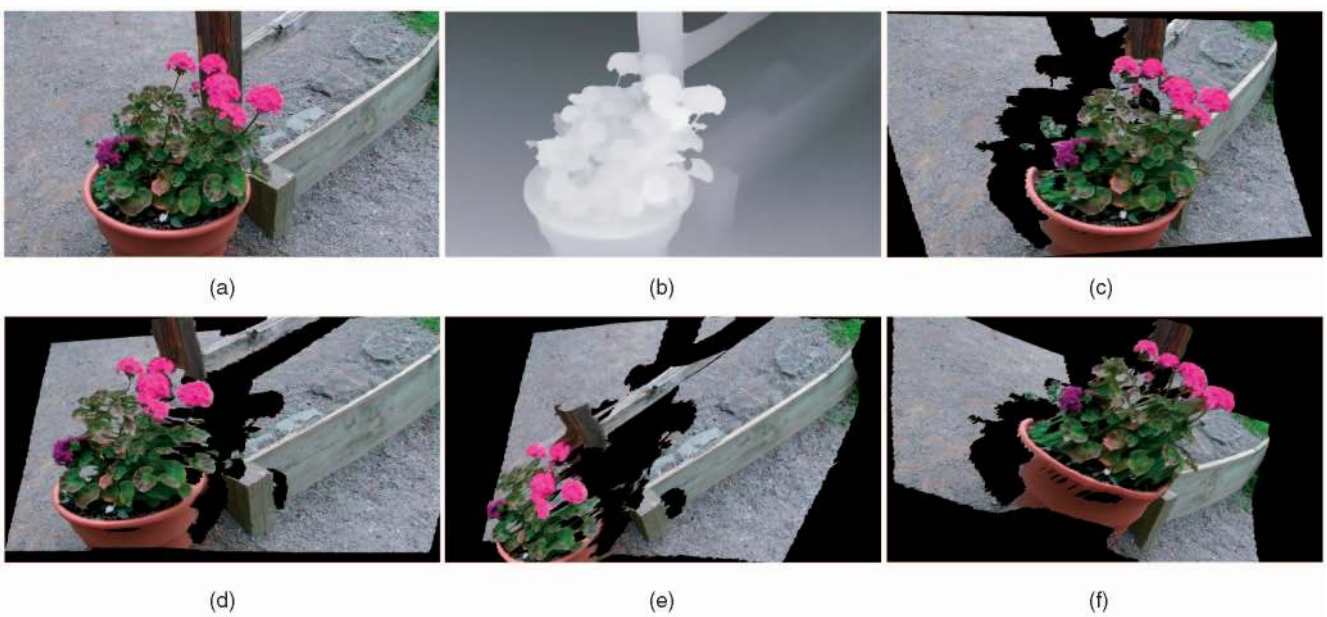

Fig. 11. Novel view synthesis. (a) and (b) One frame with the recovered depth map. (c), (d), (e), and (f) The synthesized views.

Fig. 12 shows the depth results of the "Angkor Wat" and "Temple" sequences. The image resolution is 576×352. Two more examples are shown in Fig. 13. They demonstrate how our method can robustly handle different types of camera motion besides sideways panning. The "Stair" sequence is taken by a vertically moving camera. In the "Great Wall" sequence, the camera moves surrounding the beacon on the mountain. Similar to all other examples, both of these sequences contain video noise and complex occlusions. Our recovered dense depth maps demonstrate the robustness of the proposed method.

7.3 Results of Low-Frame-Rate Sequences

Though our method is developed to solve the video depth estimation problem, it can also handle sequences that only contain a small number of frames and the baselines between consecutive frames are moderately wide. The "Statue"

Fig. 12. Video depth results of the (a) "Angkor Wat" and (b) "Temple" sequences.

Fig. 13. Video depth results of the "Stair" and "Great Wall" sequences. (a) "Stair" sequence with camera moving vertically. (b) "Great Wall" sequence with camera surrounding the beacon on the mountain.

sequence shown in Fig. 14 contains only 11 images. Three consecutive frames (i.e., frames 4-6) are shown in Figs. 14a, 14b, 14c. The small number of frames degrades the effectiveness of our method. However, the recovered depth maps still preserve sufficient fine structures as well as smooth depth transition, as shown in Figs. 14d, 14e, 14f. The reconstruction quality can be evaluated by synthesizing novel views. In experiments, with the computed depth maps for all frames, we project frames 4 and 6 onto frame 5, and linearly blend them to obtain the interpolated frame 5. The result is shown in Fig. 14j. It is very similar to the ground truth frame even on the discontinuous statue boundary.

7.4 Results of Standard Multiview Data

For quantitative evaluation on ground truth data, we test our method on the "Fountain-P11" example (http://cvlab.epfl.ch/~strecha/multiview/denseMVS.html) and show the result in Fig. 15. This example only contains 11 images and the baselines are relatively wide. Due to the limited memory space, we downsample the images to 768×512 (the original resolution is $3{,}072 \times 2{,}048$). The recovered depth maps preserve many fine structures. Note that our method is originally proposed to solve the video depth estimation problem. It does not perform similarly well given sparse image input. This is because the statistically computed data cost may not be sufficiently informative for obtaining good estimates.

Strecha et al. [34] provide quantitative evaluation for this scene dataset, given a single triangle mesh. However, for constructing a complete 3D model, we need to integrate individual depth maps. As model building is out of the scope of this paper, we simply construct a triangular mesh from an arbitrarily selected depth map (frame 5 in Fig. 15) and upload it to the evaluation Website to obtain the error histograms for this particular frame (Fig. 16). $Sigma$ denotes the standard deviation of the depth estimated using the laser range scanner [34]. After bundle optimization, about 41 percent of our depth estimates are within the $3 * sigma$ range of the LIDAR data. After space-time fusion, the percentage within the $3 * sigma$ range is further increased to about 48 percent. It indicates that the fusion step quantitatively reduces the depth reconstruction errors.

8 Discussion

We have demonstrated, with our experiments, that our algorithm can successfully and robustly handle different video sequences. However, if there is no sufficient camera motion, the recovered depths could be less accurate. This problem has been observed and widely studied in multiview geometry [17]. In addition, similar to most conventional stereo algorithms, our method assumes approximately view-independent pixel intensities, that is, with Lambertian surfaces. Therefore, if the scene contains reflection and translucency, the depth estimate in these regions may be erroneous.

Another limitation of our algorithm is that, if the scene contains extremely textureless regions, there exists inherent ambiguity for depth inference and our method could be stuck in a local optimum due to an unsatisfactory initialization. Our current initialization is by using color segmentation and plane fitting. Fig. 17 shows an example. The color in the

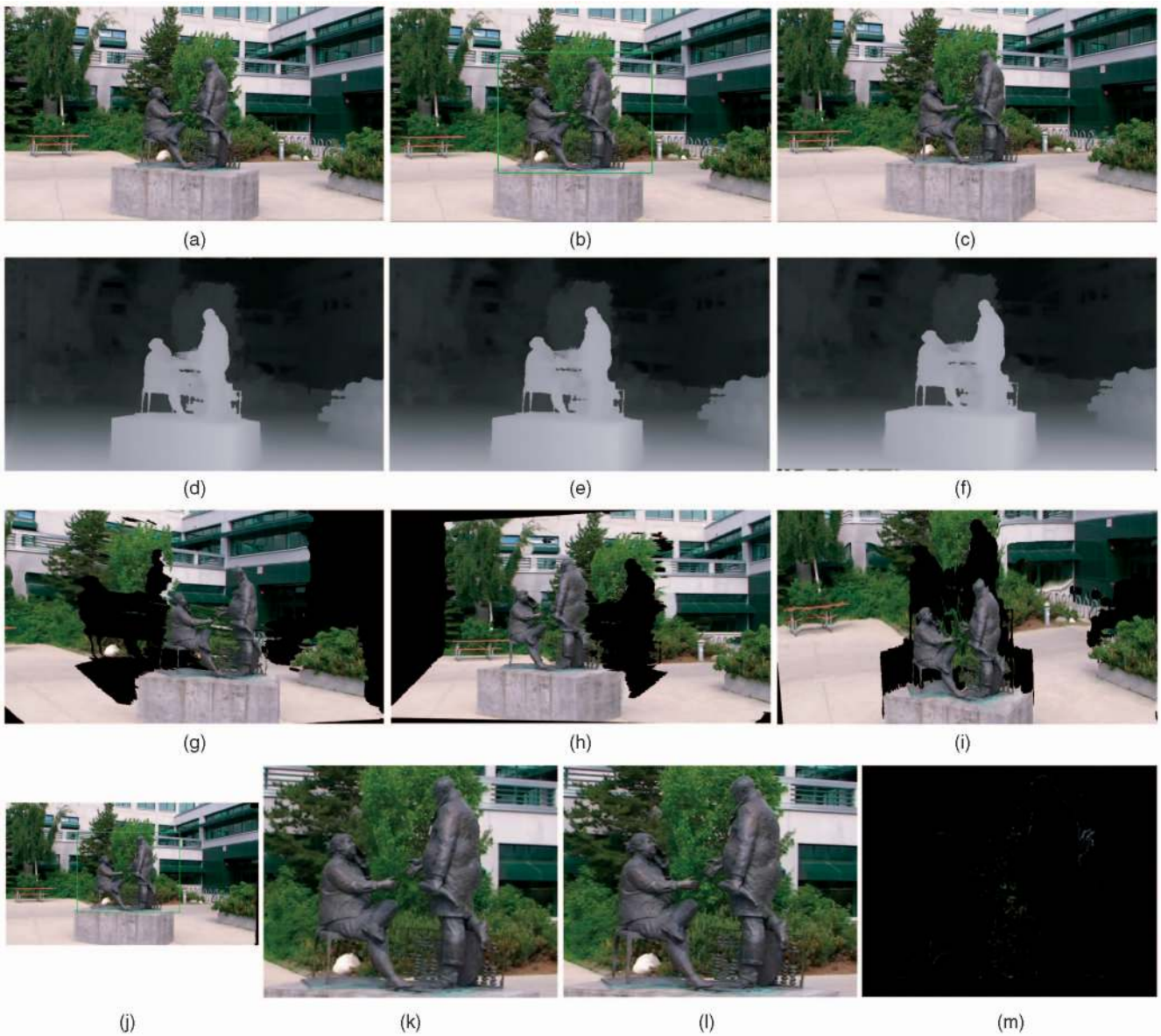

Fig. 14. "Statue" example. (a)-(c) Frames 4, 5, and 6 of the "Statue" sequence. (d)-(f) The estimated depth maps for (a)-(c). (g)-(i) The synthesized three different views from frame 5. The purely black pixels are the missing pixels. (j) Interpolated frame 5, using the depth information of frames 4 and 6. (k) Close-up of (j). (l) Close-up of (b). Our interpolation result, even near discontinuous object boundary, is natural. (m) The absolute difference image of (k) and (l).

background sky is almost constant. The depths around the tree branches have inherent ambiguity for inference. These regions can be interpreted as either in the background sky, or in a foreground layer with unknown disparities, both satisfying the multiview geometry constraint. So, without extra prior knowledge, inferring true depth values in these regions is extremely difficult.

In addition, our method is mainly developed for recovering consistent depth maps from a video sequence. For a small number of wide-baseline images, the effectiveness of our method could possibly be degraded.

9 Conclusions and Future Work

To conclude this paper, we have proposed a novel method for constructing high-quality depth maps from a video sequence. Our method advances multiview stereo reconstruction in a few ways. First, based on the geometry and photo-consistency constraints, we cope with visibility and reconstruction noise using the statistical information simultaneously from multiple frames. This model considers occlusions, noise, and outliers in a unified framework. Second, our method only uses segmentation in system initialization, and performs pixel-wise disparity refinement in the following iterative bundle optimization. By incorporating the photo-consistency and geometric coherence constraints, the depth estimate can be effectively improved. This process makes the optimization in both the textured and textureless regions reliable. Experimental results show that this process is rather effective in estimating temporally consistent disparities while faithfully preserving fine structures.

Our future work includes extending our method to estimating depths from sparse images. With a very small number of input images, occlusion handling and outlier rejection will be a more difficult problem. We expect to

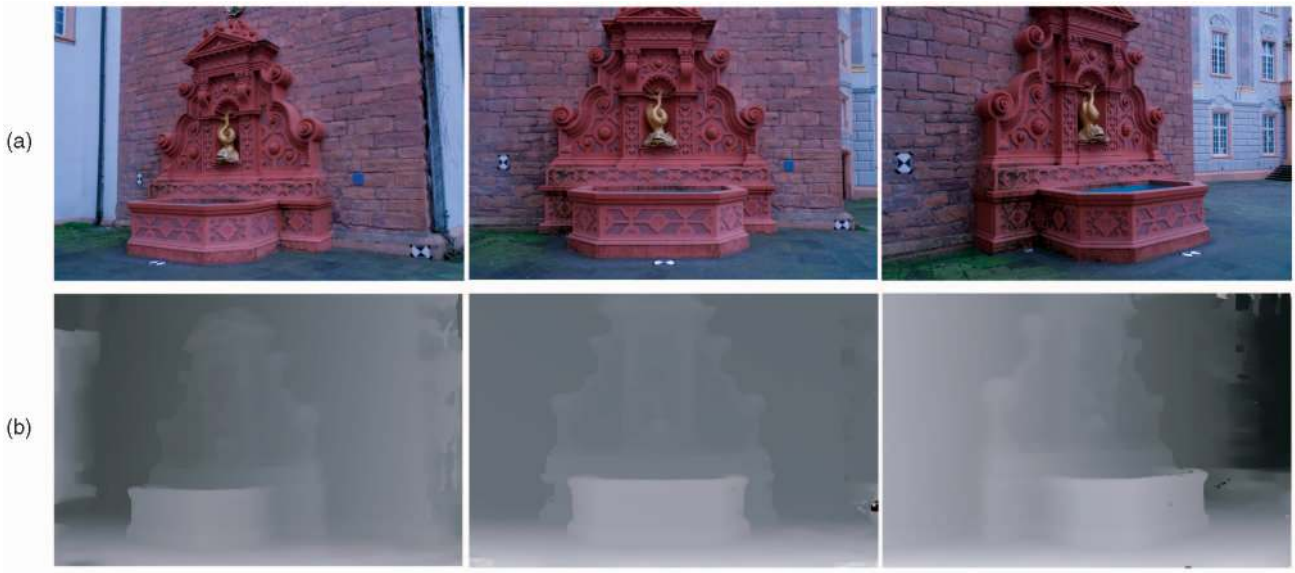

Fig. 15. "Fountain-P11" example. (a) Frames 2, 5, and 8. (b) The estimated depth maps.

tackle it by means of modifying our data cost definition and introducing a match confidence evaluation for the selected frames while statistically computing disparity likelihood for each pixel.

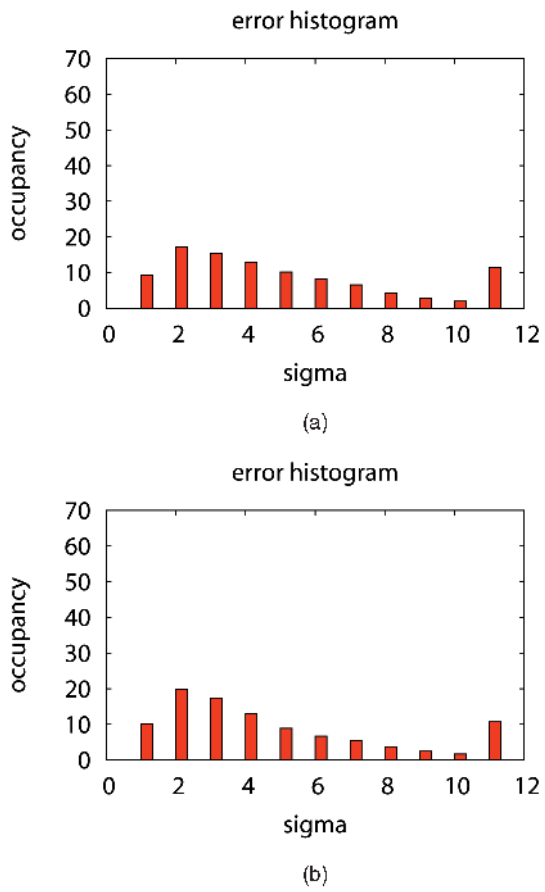

Fig. 16. The error histogram for the "Fountain-P11" example. (a) The relative error occurrence histogram for frame 5 after bundle optimization. (b) The relative error occurrence histogram for frame 5 after space-time fusion.

Another direction of our future work is to build complete 3D geometry models from our video depth estimates. As discussed in [32], reconstructing complete 3D models from real images is still a challenging problem. Many of the methods only aim to model a single object. They have inherent difficulties to model complex outdoor scenes. In comparison, our method can automatically estimate high-quality view-dependent depth maps that are temporally very consistent. We believe this work not only benefits 3D modeling, but also is applicable to video processing, rendering, and understanding. For example, for image/video segmentation, many existing algorithms only use the color information. If the depth estimates exist, layer separation could be done more effectively. For high-quality video-based rendering, obtaining accurate and temporally consistent depth maps is crucial.

Our algorithm is based on multiview geometry, and is restricted to videos of a static scene. The depths for the moving objects cannot be recovered since they do not satisfy the multiview geometry constraint. Recently, research has been conducted to deal with dynamic scenes by using multiple synchronized/unsynchronized video cameras [41], [51], [24], [12]. We believe it is possible to extend our bundle optimization framework to moving objects with multiple video streams. For example, for synchronized stereo video cameras, if we can exploit the respective temporal coherence and, at the same time, correlate neighboring frames in different streams, the depth estimation could be more robust.

ACKNOWLEDGMENTS

The authors would like to thank the associate editor and all the reviewers for their constructive comments to improve the manuscript. This work is supported by the 973 program of China (No. 2009CB320802), NSF of China (No. 60633070), the 863 program of China (No. 2007AA01Z326), and the Research Grants Council of the Hong Kong Special Administrative Region, under General Research Fund (Project No. 412307 and 417107).

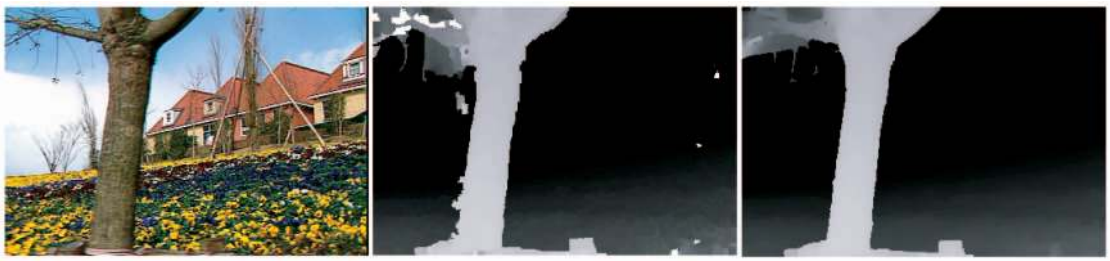

Fig. 17. Disparity result of the "Garden" sequence. (a) One frame from the input sequence. (b) The initialized disparity map. (c) The final disparity map from our system. The outliers and visual artifacts around discontinuous boundaries are dramatically reduced. However, because the disparities around the branches inherently have depth ambiguity regarding almost constant-color background sky, the disparity initialization is poor. This also makes the following optimization produce visually unsatisfying result in this region.

REFERENCES

[1] L. Álvarez, R. Deriche, T. Papadopoulo, and J. Sánchez, "Symmetrical Dense Optical Flow Estimation with Occlusions Detection," *Int'l J. Computer Vision,* vol. 75, no. 3, pp. 371-385, 2007.

[2] P. Bhat, C.L. Zitnick, N. Snavely, A. Agarwala, M. Agrawala, B. Curless, M. Cohen, and S.B. Kang, "Using Photographs to Enhance Videos of a Static Scene," *Rendering Techniques 2007: Proc. Eurographics Symp. Rendering,* J. Kautz and S. Pattanaik, eds., pp. 327-338, June 2007.

[3] A.F. Bobick and S.S. Intille, "Large Occlusion Stereo," *Int'l J. Computer Vision,* vol. 33, no. 3, pp. 181-200, 1999.

[4] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

[5] D. Bradley, T. Boubekeur, and W. Heidrich, "Accurate Multi-View Reconstruction Using Robust Binocular Stereo and Surface Meshing," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.

[6] R.T. Collins, "A Space-Sweep Approach to True Multi-Image Matching," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 358-363, 1996.

[7] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.

[8] Y. Deng, Q. Yang, X. Lin, and X. Tang, "A Symmetric Patch-Based Correspondence Model for Occlusion Handling," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1316-1322, 2005.

[9] O.D. Faugeras and R. Keriven, "Variational Principles, Surface Evolution, PDEs, Level Set Methods, and the Stereo Problem," *IEEE Trans. Image Processing,* vol. 7, no. 3, pp. 336-344, 1998.

[10] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient Belief Propagation for Early Vision," *Int'l J. Computer Vision,* vol. 70, no. 1, pp. 41-54, 2006.

[11] P. Fua, "Aparallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications,* vol. 6, pp. 35-49, 1993.

[12] Y. Furukawa and J. Ponce, "Dense 3D Motion Capture from Synchronized Video Streams," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.

[13] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2007.

[14] D. Gallup, J.-M.F.P. Mordohai, and M. Pollefeys, "Variable Baseline/Resolution Stereo," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.

[15] P. Gargallo and P.F. Sturm, "Bayesian 3D Modeling from Images Using Multiple Depth Maps," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 885-891, 2005.

[16] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S.M. Seitz, "Multi-View Stereo for Community Photo Collections," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[17] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* second ed. Cambridge Univ. Press, 2004.

[18] C. Hernández, G. Vogiatzis, and R. Cipolla, "Probabilistic Visibility for Multi-View Stereo," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2007.

[19] S.B. Kang and R. Szeliski, "Extracting View-Dependent Depth Maps from a Collection of Images," *Int'l J. Computer Vision,* vol. 58, no. 2, pp. 139-163, 2004.

[20] S.B. Kang, R. Szeliski, and J. Chai, "Handling Occlusions in Dense Multi-View Stereo," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 103-110, 2001.

[21] A. Klaus, M. Sormann, and K.F. Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, pp. 15-18, 2006.

[22] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions Via Graph Cuts," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 508-515, 2001.

[23] V. Kolmogorov and R. Zabih, "What Energy Functions Can Be Minimized via Graph Cuts?" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 147-159, Feb. 2004.

[24] E.S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs, "Temporally Consistent Reconstruction from Multiple Video Streams Using Enhanced Belief Propagation," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.

[25] A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 2, pp. 150-162, Feb. 1994.

[26] G. Li and S.W. Zucker, "Surface Geometric Constraints for Stereo in Belief Propagation," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2355-2362, 2006.

[27] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-Time Visibility-Based Fusion of Depth Maps," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[28] M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 4, pp. 353-363, Apr. 1993.

[29] M. Pollefeys, L.J. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual Modeling with a Hand-Held Camera," *Int'l J. Computer Vision,* vol. 59, no. 3, pp. 207-232, 2004.

[30] M. Proesmans, L.J. Van Gool, E.J. Pauwels, and A. Oosterlinck, "Determination of Optical Flow and its Discontinuities Using Non-Linear Diffusion," *Proc. European Conf. Computer Vision,* vol. 2, pp. 295-304, 1994.

[31] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision,* vol. 47, nos. 1-3, pp. 7-42, 2002.

[32] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 519-528, 2006.

[33] S.M. Seitz and C.R. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring," *Int'l J. Computer Vision,* vol. 35, no. 2, pp. 151-173, 1999.

[34] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.

[35] C. Strecha, R. Fransens, and L.J. Van Gool, "Wide Baseline Stereo from Multiple Views: A Probabilistic Account," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 552-559, 2004.
[36] C. Strecha, R. Fransens, and L.J. Van Gool, "Combined Depth and Outlier Estimation in Multi-View Stereo," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2394-2401, 2006.
[37] C. Strecha and L.J. Van Gool, "PDE-Based Multi-View Depth Estimation," *Proc. 3D Data Processing Visualization and Transmission,* pp. 416-427, 2002.
[38] J. Sun, Y. Li, and S.B. Kang, "Symmetric Stereo Matching for Occlusion Handling," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 399-406, 2005.
[39] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo Matching Using Belief Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 7, pp. 787-800, July 2003.
[40] Y. Taguchi, B. Wilburn, and L. Zitnick, "Stereo Reconstruction with Mixed Pixels Using Adaptive Over-Segmentation," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.
[41] H. Tao, H.S. Sawhney, and R. Kumar, "Dynamic Depth Recovery from Multiple Synchronized Video Streams," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 118-124, 2001.
[42] H. Tao, H.S. Sawhney, and R. Kumar, "A Global Matching Framework for Stereo Computation," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 532-539, 2001.
[43] M.F. Tappen and W.T. Freeman, "Comparison of Graph Cuts with Belief Propagation for Stereo, Using Identical MRF Parameters," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 900-907, 2003.
[44] D. Terzopoulos, "Multilevel Computational Processes for Visual Surface Reconstruction," *Computer Vision, Graphics, and Image Processing,* vol. 24, no. 1, pp. 52-96, 1983.
[45] G. Vogiatzis, P.H.S. Torr, and R. Cipolla, "Multi-View Stereo Via Volumetric Graph-Cuts," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 391-398, 2005.
[46] O.J. Woodfordy, P.H.S. Torrz, I.D. Reidy, and A.W. Fitzgibbon, "Global Stereo Reconstruction under Second Order Smoothness Priors," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2008.
[47] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation and Occlusion Handling," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2347-2354, 2006.
[48] C. Zach, T. Pock, and H. Bischof, "A Globally Optimal Algorithm for Robust TV-L^1 Range Image Integration," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.
[49] A. Zaharescu, E. Boyer, and R. Horaud, "Transformesh: A Topology-Adaptive Mesh-Based Approach to Surface Evolution," *Proc. Asian Conf. Computer Vision,* vol. 2, pp. 166-175, 2007.
[50] G. Zhang, X. Qin, W. Hua, T.-T. Wong, P.-A. Heng, and H. Bao, "Robust Metric Reconstruction from Challenging Video Sequences," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2007.
[51] C. Zhou and H. Tao, "Dynamic Depth Recovery from Unsynchronized Video Streams," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 351-358, 2003.
[52] C.L. Zitnick and S.B. Kang, "Stereo for Image-Based Rendering Using Image Over-Segmentation," *Int'l J. Computer Vision,* vol. 75, no. 1, pp. 49-65, 2007.

Guofeng Zhang received the BS degree in computer science from Zhejiang University, P.R. China, in 2003. He is currently working toward the PhD degree in computer science at the State Key Laboratory of CAD&CG, Zhejiang University. His research interests include camera tracking, 3D reconstruction, augmented reality, and video enhancement. He is a student member of the IEEE.

Jiaya Jia received the PhD degree in computer science from the Hong Kong University of Science and Technology in 2004. He joined the Department of Computer Science and Engineering at The Chinese University of Hong Kong in September 2004, where he is currently an assistant professor. His research interests include vision geometry, image/video editing and enhancement, image deblurring, and motion analysis. He has served on the program committees of ICCV, CVPR, ECCV, and ACCV. He is a member of the IEEE.

Tien-Tsin Wong received the BSci, MPhil, and PhD degrees in computer science from the Chinese University of Hong Kong in 1992, 1994, and 1998, respectively. Currently, he is a professor in the Department of Computer Science and Engineering at The Chinese University of Hong Kong. His main research interest is computer graphics, including computational manga, image-based rendering, natural phenomena modeling, and multimedia data compression. He received the *IEEE Transactions on Multimedia* Prize Paper Award 2005 and the Young Researcher Award 2004. He is a member of the IEEE.

Hujun Bao received the BS and PhD degrees in applied mathematics from Zhejiang University in 1987 and 1993, respectively. Currently, he is a professor and the director of State Key Laboratory of CAD&CG at Zhejiang University. His main research interest is computer graphics and computer vision, including real-time rendering technique, geometry computing, virtual reality, and structure from motion. He is a member of the IEEE.