Multiview video depth estimation with spatial-temporal consistency

Mingjin Yang
ymj04@mails.tsinghua.edu.cn
Xun Cao
xuncao@gmail.com
Qionghai Dai
qhdai@mail.tsinghua.edu.cn

Broadband Network and Digital Multimedia Lab,
Department of Automation,
Tsinghua University,
Beijing 100084,
China.

Abstract

In this paper, we present an approach to recover both spatially and temporally consistent depth maps from multiview synchronized and calibrated video streams. Depth maps are initialized by combining left-right view matching and color based segmentation. Then the color constancy and spatial coherence are integrated into the optimization framework in order to guarantee the spatial consistency at single time instant. Finally, we incorporate the depth and motion information in the form of spatial-temporal consistency constraint to refine and stabilize the depth video, without ruining the original spatial consistency in the estimation of each single instant. The experiments on different multiview sequences demonstrate the effectiveness of our method in providing both stable and accurate multiview depth estimation.

1 Introduction

As a fundamental problem in computer vision, stereo matching [5] has been an active research topic for many decades, facilitating applications such as 3D modeling, video editing, view synthesis and interpolation. Many methods have been proposed to estimate high quality depth maps, among which such as belief propagation (BP) [17, 19, 21]and Graph-Cuts [12] have demonstrated their advantages on achieving global optimization. To handle the challenging problem of textureless regions, segmentation-based approaches are applied with the assumption that areas of homogeneous color have smooth depth [1, 15, 20, 26]. Another major challenge of stereo matching is the occlusion problem, which is also investigated in many works [11, 20, 24]. They explicitly consider occlusion detection in their framework and are able to reduce artifacts at depth boundaries. Previous methods can be generally classified into two categories: recovering depth for the static scene and for the dynamic scene.

To extract static depth map from just one time instant, some research enforces the spatial consistency by combining depth maps of several views [8, 23]. Kang and Szeliki [23] simultaneously optimize a set of self-consistent depth maps at multiple key-views by adding a compatibility constraint to them. Zhang et al. [8] incorporate geometric coherence associating multiple views into photo-consistency constraint, and handle image noise, occlusions and outliers with a unified bundle optimization framework. Their recovered disparity maps

BMVC 2010 doi:10.5244/C.24.67

can maintain the spatial coherence without over smoothing. However, the accuracy relies on the number of views and degree of satisfaction in disparity initialization.

When estimating dynamic depth maps from multiview video, these methods can only perform depth estimation frame-independently, without considering the temporal consistency between adjacent frames. As a result, the generated depth video will appear obvious inconsistent and flickering. Some other research shows the improvement by incorporating temporal coherence constraint into their optimization model [6, 7, 9, 14].

Tao et al. [9] propose a method based on image segmentation which models each segment as a 3D planar surface patch. Temporal correspondences between segments of adjacent frames are determined with optical flow. Then parameters of the planar patch are estimated using initial spatial homography and temporal homography. Sharp depth boundaries are preserved by taking advantage of segmentation and plane fitting.

In Gong's approach [14] , the concept of disparity flow is used to model the 3D motion in the scene, which can also be considered view-dependent scene flow [22]. The disparity flow is used to predict the disparity maps of the next frame by cost adjustment with predicted values. To achieve real-time performance, they only use local WTA optimization in computation.

Scene flow is also used in Liu and Philomin's work [7] as a soft constraint to predict disparity map and to compute confidence map of the next frame. Over-segmentation based stereo and pixel-wise iterative stereo are applied first to initialize the disparity. Their method is able to estimate accurate and temporally consistent disparity maps with only two image streams. Nevertheless, there still exists some error propagation in disparity maps.

Larsen et al. [6] present an approach for temporally consistent depth reconstruction based on enhanced BP framework which extends traditional 4-neighborhood 2D image lattice to 6-neighborhood 3D lattice. Even when optical flow fails in finding corresponding pixels in adjacent frames, their algorithm can at least maintain the quality of single time reconstruction.

In this paper, a method to estimate spatio-temporally consistent depth maps from multiview video streams is presented. Our method is inspired by Zhang's approach [8] which can recover high quality spatially consistent depth maps of static scene from multiple images. The segmentation-based disparity initialization and integration of geometry coherence constraint and photo-consistency constraint by bundle optimization are adopted in our framework, but considering videos with sparse views in which occlusion problem cannot be well handled by temporal selection [23] , the initialization step applies left-right view matching (picking matching point from left or right view) instead. Besides, second-round segmentation which follows spatial consistence bundle optimization is used to remove the outliers introduced by left-right view matching and make the parameters in plane fitting more precise. Furthermore, optical flow and disparity maps of adjacent frames are combined to enforce the spacial-temporal consistency in target energy function, which is also optimized by bundle optimization to achieve more accurate and stable depth video.

2 Motivation

Our work is inspired by the observation that the state-of-the-art stereo methods would appear jittering when dealing with dynamic scenes. Although some recent literature [6, 7, 9, 14] report their results on maintaining smooth video without the jitter artifacts, the reconstructed depth maps still remain much to be improved. We first borrow the idea of geometry consis-

tence from [8] 's work to ensure the spatial consistency at each time instant. Then, temporal information is integrated into a unified bundle optimization framework to compute the multiview depth maps. The consideration of temporal information brings not only the stabilization on time axis, but also can be used to refine the inaccurate depth estimation by each single time instance. In the following sections, we will detail in how to incorporate both spatial and temporal consistency into the target function and demonstrate the results of our method on various multiview datasets.

The rest of this paper is organized as follows. In section 3, we explain the algorithm of spatio-temporally consistent depth maps estimation. In section 4, experimental results of several multiview videos are shown to demonstrate the effectiveness of our method. Finally we give the conclusions in section 5.

3 Proposed Approach

In this section, we describe in detail the framework of the proposed approach and the energy function constructed based on the Markov Random Field (MRF), which is optimized by loopy BP.

Given a multiview video sequence I with N views, and T frames for each view ($I = \{I_{t,n} | t = 1,\cdots,\mathrm{T}; \mathrm{n} = 1,\cdots,\mathrm{N}\}$), and the camera parameter set of view n ($P_n = \{K_n, R_n, T_n\}$), our goal is to estimate the depth maps Z ($Z = \{Z_{t,n} | t = 1,\cdots,\mathrm{T}; \mathrm{n} = 1,\cdots,\mathrm{N}\}$).

Let $I_{t,n}(x)$ and $Z_{t,n}(x)$ (also written as $I(x_{t,n})$ and $Z(x_{t,n})$) be the color and depth of pixel x in view n at time t ((t,n) for short), and here we use disparity $D_{t,n}(x) = D(x_{t,n}) = 1/Z_{t,n}(x)$ instead. Then the global energy function composed of the data term E_d and smoothness term E_s is defined as (1):

$$E(D;I) = \sum_{t=1}^{\mathrm{T}} \sum_{n=1}^{\mathrm{N}} (E_d(D_{t,n}; I, D \backslash D_{t,n}) + E_s(D_{t,n})), \tag{1}$$

where D is set of $D_{t,n}$, and $D \backslash D_{t,n}$ means all the disparity maps excluding $D_{t,n}$. So the data term gives the conditional probability of $D_{t,n}$ given I and $D \backslash D_{t,n}$.

For each step, we use unified form of E_d to measure the fitness of multiview disparitiy maps D to sequence I and refine the disparity maps reconstructed in previous step. Besides, we apply one single E_s form to ensure the smoothness on disparities of neighboring pixels.

3.1 Disparity Initialization

This step generates initial disparity maps of time t via loopy BP and segmentation based plane fitting.

Considering sequences with sparse views, especially when region of pixels are visible only in another one or two views, we use left and right views to generate the disparity of centre view instead of temporal selection [23], under the assumption that each pixel is visible in at least one neighbor view.

We quantize the disparity range $[D_{min}, D_{max}]$ into $L+1$ levels, and denote $x_{t,n'} | (x_{t,n}, D(x_{t,n}))$ ($x_{t,n'}$ for short) as the mapping pixel in (t,n') from pixel $x_{t,n}$ with disparity level $D(x_{t,n})$, which is computed by (2):

$$\tilde{x}_{t,n'} \cong K_{t,n'} R_{t,n'}^{\mathrm{T}} R_{t,n} K_{t,n}^{-1} \tilde{x}_{t,n} + D(x_{t,n}) K_{t,n'} R_{t,n'}^{\mathrm{T}} (T_{t,n} - T_{t,n'}), \tag{2}$$

where $\tilde{x}$ means the homogeneous coordinates of x, and $\cong$ means equal up to a scale factor.

The likelihood of disparity $D(x_{t,n})$ for pixel $x_{t,n}$ is defined as (3):

$$L^{initial}(x_{t,n}, D(x_{t,n})) = \max\{p_c(x_{t,n}, D(x_{t,n}), I_{t,n}, I_{t,n-1}), p_c(x_{t,n}, D(x_{t,n}), I_{t,n}, I_{t,n+1})\}, \quad (3)$$

where $p_c(x_{t,n}, D(x_{t,n}), I_{t,n}, I_{t,n'})$ is defined as (4) to measure the color similarity between $x_{t,n}$ and $x_{t,n'}$:

$$p_c(x_{t,n}, D(x_{t,n}), I_{t,n}, I_{t,n'}) = \frac{\sigma_c}{\sigma_c + \left\| I(x_{t,n}) - I(x_{t,n'}) \right\|_2}, \quad (4)$$

where σ_c is a constant to control the shape of color similarity function.

Then we have the data term E_d as (5):

$$E_d(D_{t,n}; I, D \backslash D_{t,n}) = \sum_{x_{t,n}} (1 - u(x_{t,n}) \cdot L^{initial}(x_{t,n}, D(x_{t,n}))), \quad (5)$$

where $u(x_{t,n}) = 1 / \max_{D(x_{t,n})} L^{initial}(x_{t,n}, D(x_{t,n}))$ is a normalization factor. For smoothness term E_s, we simply define it as (6):

$$E_s(D_{t,n}) = \sum_{x_{t,n}} \sum_{y_{t,n} \in N(x_{t,n})} \omega \cdot \lambda(x_{t,n}, y_{t,n}) \cdot \min\{|D(x_{t,n}) - D(y_{t,n})|, \eta\}, \quad (6)$$

where $N(x_{t,n})$ denotes the set of neighbor pixels of $x_{t,n}$, and ω is the smoothness strength constant. η determines the upper limit of contribution of disparity difference. The smoothness weight $\lambda(x_{t,n}, y_{t,n})$ is defined as (7) to preserve disparity discontinuity at color outliers:

$$\lambda(x_{t,n}, y_{t,n}) = \frac{\varepsilon}{\varepsilon + \| I(x_{t,n}) - I(y_{t,n}) \|_2}, \quad (7)$$

where constant ε controls the color sensitivity of smoothness term.

As we know, color information of only two neighbor views is not enough to handle the textureless regions and occlusions, so we employ the mean-shift color segmentation [4] and plane fitting to handle the textureless regions. Each segment S_k is modeled as a 3D plane and the disparity of $x_{t,n}(x_{t,n} = [x,y] \in S_k)$ is denoted as $D(x_{t,n}) = a_k x + b_k y + c_k$. After that, the plane fitting method proposed by Zhang et al. [8] is adopted to compute the plane parameters (a_k, b_k, c_k). They first fix $a_k = b_k = 0$ in S_k and the disparity values in other segments. Then a set of c_k which equal to $L+1$ disparity levels are used to compute a best c_k^* by minimize the energy function (1). After that, initial plane parameters $(0, 0, c_k^*)$ are further refined by again minimizing (1) using Levenberg-Marquardt method, getting final result (a_k^*, b_k^*, c_k^*). Plane parameters of other segments are computed in the same way one by one.

3.2 Spatial consistency

After the initialization of disparity, there still exist many outliers because of inaccurate plane parameters and the disturbance of similar-color pixels with different disparities. Meanwhile the disparity map of each view is estimated independently without considering their spatial correlation. So we incorporate spatial coherence constraint into the color constancy constraint to refine spatial errors and ensure the spatial consistency.

As shown in figure 1, by epipolar geometry, $x_{t,n'}$ in (t, n') is the projection of $x_{t,n}$ with $D(x_{t,n})$ in (t, n), and the initial disparity of it is $D(x_{t,n'})$. $\hat{X}$ is the corresponding 3D point.

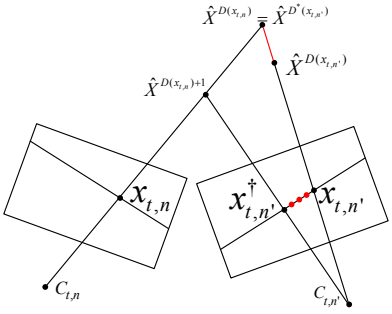

Figure 1: Spatial coherence constrain. The mapping pixel of $x_{t,n}$ is $x_{t,n'}$, which lies on the conjugate epipolar line. Its ideal disparity is $D^*(x_{t,n'})$, with which $x_{t,n'}$ will map to a same 3D point as $x_{t,n}$ with $D(x_{t,n})$. But the initial disparity of $x_{t,n'}$ is $D(x_{t,n'})$, so we use the difference between $D(x_{t,n'})$ and $D^*(x_{t,n'})$ to measure the spatial coherence. At the same time, we take those pixels between $x_{t,n'}$ and $x^{\dagger}_{t,n'}$ into consideration to handle large baseline problem.

Ideally, we have $\hat{X}^{D(x_{t,n})} = \hat{X}^{D^*(x_{t,n'})}$, but inaccurate initial disparity maps induce the deviation (marked as red line segment). Zhang et al. [8] propose the geometric coherence term to measure the deviation, but for large baseline, the difference of mapping coordinates $\|x^{\dagger}_{t,n'} - x_{t,n'}\|_2$ (where $x^{\dagger}_{t,n'}$ is short form of $x_{t,n'}|(x_{t,n}, D(x_{t,n})+1)$) results from neighbor disparity level will be larger than small baseline.

So we define the spatial coherence constraint (8) as a Gaussian distribution of the difference between $D(x_{t,n'})$ and its' ideal disparity $D^*(x_{t,n'})$ (disparity level corresponding to depth of $\hat{X}^{D(x_{t,n})}$ from $C_{t,n'}$, which can be computed with (11) as an instance of cross-line). We also take into account the pixels (marked as red point, whose set is noted as $l(x_{t,n'})$) between $x_{t,n'}$ and $x^{\dagger}_{t,n'}$ on the conjugate epipolar line. Spatial coherence constraint of $x_{t,n}$ with other disparity levels such as $D(x_{t,n})+1$ is defined in the same way.

$$p_s(x_{t,n}, D(x_{t,n}), D_{t,n'}) = \max_{x^l_{t,n'} \in l(x_{t,n'})} \exp\left(-\left(D(x^l_{t,n'}) - D^*(x^l_{t,n'})\right)^2 \Big/ 2\sigma_s^2\right) \tag{8}$$

Furthermore, in order not to extend the wrong disparity in occlusions to subsequent disparity maps, the occlusions detected by initial disparity are taken into consideration, so the likelihood (3) is modified to (9):

$$L(x_{t,n}, D(x_{t,n})) = \frac{N}{\sum_{\substack{n'=1 \\ n' \neq n}}^{N} (1 - O^{t,n'}_{t,n}(x_{t,n}))} \cdot \sum_{\substack{n'=1 \\ n' \neq n}}^{N} (1 - O^{t,n'}_{t,n}(x_{t,n})) \cdot p_c \cdot p_s, \tag{9}$$

where $O^{t,n'}_{t,n}(x_{t,n}) = 1$, if $x_{t,n}$ is occluded in (t,n'), else $O^{t,n'}_{t,n}(x_{t,n}) = 0$. We first map each point $x_{t,n}$ in (t,n) to (t,n') with corresponding disparity $D(x_{t,n})$, and order the points those map to a same $x_{t,n'}$ by considering their disparities. The points with larger disparity value are considered to be disoccluded, and others will be occluded points.

So the corresponding data term (5) is rewritten by substituting L (9) for $L^{initial}$ (3) as (10):

$$E_d(D_{t,n}; I, D \backslash D_{t,n}) = \sum_{x_{t,n}} (1 - u(x_{t,n}) \cdot L(x_{t,n}, D(x_{t,n}))), \tag{10}$$

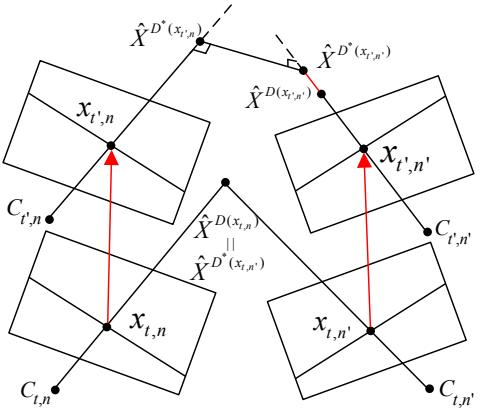

Figure 2: Spatial-temporal coherence constraint. The optical flow of $x_{t,n}$ and $x_{t,n'}$ is respectively $x_{t',n}$ and $x_{t',n'}$. Ideally, the optical rays $C_{t',n}\tilde{x}_{t',n}$ will cross $C_{t',n'}\tilde{x}_{t',n'}$ with the intersecting point $\hat{X}^{D^*(x_{t',n})} = \hat{X}^{D^*(x_{t',n'})}$. However, the occlusions, disparity and motion estimation errors cause the two optical rays noncoplanar and make $\hat{X}^{D^*(x_{t',n})}$ and $\hat{X}^{D^*(x_{t',n'})}$ the foot position of the rays' common perpendicular. So we use the difference between $D(x_{t',n'})$ and its' ideal value $D^*(x_{t',n'})$ to measure the spatial-temporal coherence.

where $u(x_{t,n}) = 1/\max_{D(x_{t,n})} L(x_{t,n}, D(x_{t,n}))$.

With the same E_s (6) as the initialization step, we apply bundle optimization [8] to refine disparity map of each view at time t one by one, and again plane fitting to handle remaining few occlusions, also make parameters of some textureless areas more accurate.

3.3 Spatial-temporal consistency

Incorporating segmentation to bundle optimization improves disparities of the occlusions and textureless regions, but also flaws originally refined object boundaries. If there's only one time instance or the purpose is to recover frame-independent disparity maps, we can apply bundle optimization another two passes without plane fitting, and then generate spatially consistent disparity maps at each time instance.

However, for a video sequence, different plane parameters of the same region in different (t,n) and the motion of scene inevitably lead to temporal inconsistencies and flickering effects, which are obviously noticed in depth videos. So we incorporate the spatial-temporal constraint into the data term to enforce temporal coherence.

Let $F_{t,n}^{t',n}(x_{t,n})$ be the optical flow of $x_{t,n}$ from time t to t' in view n. As shown in figure 2, the red arrows represent motion vectors in image from time t to t', which means $F_{t,n}^{t',n}(x_{t,n}) = x_{t',n}$, and the mapping pixel of $x_{t,n}$ is $x_{t,n'}$, which has $F_{t,n'}^{t',n'}(x_{t,n'}) = x_{t',n'}$. In Section 3.2 and here, we add subscript t to C_n because our method regards the moving of camera as the moving of scene. Even when cameras move at the next moment, there's no difference to our framework as long as the parameters of cameras are all known in every (t,n).

$\hat{X}^{D^*(x_{t',n})}$ is a 3D point on the optical ray $C_{t',n}\tilde{x}_{t',n}$ whose projection to (t',n') is closer to $x_{t',n'}$ than any other 3D points on the ray. $\hat{X}^{D^*(x_{t',n'})}$ is the same. If $D(x_{t,n})$ is the accurate disparity of $x_{t,n}$, and there are no occlusions and motion estimation errors, optical ray $C_{t',n}\tilde{x}_{t',n}$

will cross $C_{t',n'}\tilde{x}_{t',n'}$ and $\hat{X}^{D^*(x_{t',n})} = \hat{X}^{D^*(x_{t',n'})}$ should be the intersecting point. But the factors above cause the two optical rays noncoplanar, so $\hat{X}^{D^*(x_{t',n})}$ and $\hat{X}^{D^*(x_{t',n'})}$ respectively locate at the foot position of common perpendicular of the two optical rays. $D^*(x_{t',n})$ and $D^*(x_{t',n'})$ can be computed using camera parameters of (t',n) and (t',n') as (11):

$$(D^*(x_{t',n}), D^*(x_{t',n'})) = \underset{(d_{t',n}, d_{t',n'})}{\arg\min} \left(\frac{R_{t',n}K_{t',n}^{-1}}{d_{t',n}}\tilde{x}_{t',n} - \frac{R_{t',n'}K_{t',n'}^{-1}}{d_{t',n'}}\tilde{x}_{t',n'} + T_{t',n} - T_{t',n'} \right). \quad (11)$$

As a result, we define spatial-temporal coherence constraint as a function of difference between $D(x_{t',n'})$ and $D^*(x_{t',n'})$, using the same form of Gaussian distribution as the spatial coherence constraint (8), denoted as p_{st}.

What's more, we define the spatial-temporal color constancy constraint with the similar form of (4) by considering the color difference of $x_{t',n}$ and $x_{t',n'}$, denoted as p_{cst}.

So the likelihood (9) is modified to (12):

$$L(x_{t,n}, D(x_{t,n})) = \frac{\mathrm{N}}{\sum\limits_{\substack{n'=1,\\ n'\neq n}}^{\mathrm{N}} (1 - O_{t,n}^{t,n'}(x_{t,n}))} \cdot \sum_{\substack{n'=1,\\ n'\neq n}}^{\mathrm{N}} (1 - O_{t,n}^{t,n'}(x_{t,n})) \cdot (p_c \cdot p_s + p_{cst} \cdot p_{st}). \quad (12)$$

We still define E_d the same expression as (10), and E_s the same as (6). When there are optical flow errors or occlusions, $P_{cst} \cdot P_{st}$ hardly outputs a large value to contribute to the likelihood (12). But $P_{cs} \cdot P_s$ is still able to ensure the spatial consistency.

With the concern of computation complexity, here we only use the disparity maps of preceding time instance to refine those of current time-image. After processing one view (disparity maps of other views are fixed) with spatial-temporal optimization, we apply color and disparity based bilateral filter [2] to further improve the depth quality. And then the disparities of other views are refined one by one in the same way. Two passes are sufficient in this spatial-temporal optimization step.

4 Experimental results

Figure 3 illustrates the workflow of our framework with intermediate results of "book arrival" [10] (16 views). (a) is a color image, whose left-right matching result with BP is shown in (b). We can see many errors in textureless regions and occluded areas. After incorporating segmentation in (c), we get a better disparity initialization result shown in (d). However, there still exist inaccurate disparities in areas with wrong plane parameters and objects with similar color. So the spatial consistency optimization followed by segmentation is applied in (e) to correct some errors and enforce the spatial correlation of each view. Finally, as illustrated in (f), we apply spatial-temporal consistency optimization to stabilize and refine depth video. Averagely, it takes about 20 minutes to process all three steps for one frame-time with resolution 1024×768.

Figure 4 compares the disparity maps of "book arrival" with and without spatial-temporal optimization. The column (a) is the color images of the same view 8 at different time (frame 13 and 18). The column (b) shows the corresponding disparity maps generated frame-independently, where the marked regions illustrates their inconsistencies. The disparity errors of floor and wall around legs of stool result from inherent depth ambiguity for

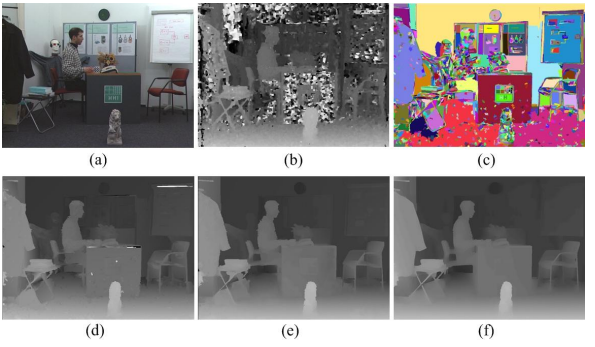

Figure 3: Workflow illustration of our method. (a) The 3rd frame of view 11 in "book arrival"; (b) Left-right matching with loopy BP; (c) Mean-shift segmentation; (d) Initial disparity map with segmentation and plane fitting; (e) Spatial consistency optimization followed by second-round segmentation; (f) Final disparity with spatial-temporal consistency optimization.

inference. And the wrong plane fitting parameters of right-side wall cause larger disparities when the man comes close to the chair. From column (c) of disparity maps estimated via spatial-temporal optimization, we can see the improvement in corresponding regions. The left column of (d) and (e) respectively show the magnified red and green regions of (b), while the right columns respectively illustrate those regions of (c). Even the two frames are not adjacent, which means they are not optimized together directly, our method can still propagate the information through frames between them.

Figure 5 shows other three examples. The top row shows one color image of sequence (a) "Akko&Kayo" [25] (5 views are used), (b) "Poznan Hall" [13] (9 views) and (c) "breakdancers" [3] (8 views) respectively. The second row illustrates the depth maps estimated frame-independently by other's work [3, 16, 18]. Without spatial-temporal coherence constraint (third row), there also exist segmentation errors in our disparity maps which are caused by homogeneous color of neighbor pixels. But the disparities can be effectively refined via our spatial-temporal optimization method (bottom row).

It is worth mentioning that the camera sets of moving camera sequence "Poznan Hall" are different from other sequences with static cameras. But the camera are all calibrated at different time instances, in which situation our method still works while some other methods [6] based on static cameras would fail. Note that the disparity of floor is estimated to be farther than the truth because of the glare there.

5 Conclusions and Future Work

In this paper, we present an approach to estimate depth maps from multiview synchronized video. The main contribution of the proposed method lies in that both the spatial and the temporal consistency are considered. Different from [8], the usage of temporal constraint benefits us for two aspects: the inaccurate result from single instant estimation can be further refined by taking temporal correspondences into account; on the other hand, the flickering of depth video is greatly alleviated by the spatial-temporal unified optimization.

In current method, the color constancy constraint only depends on pixel intensity, which

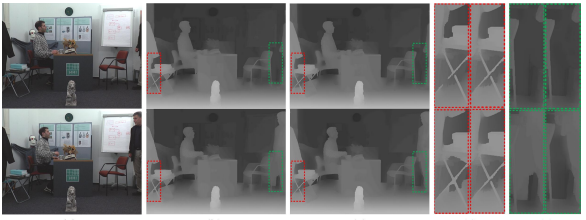

Figure 4: Disparity maps of "book arrival". (a) The 13th and 18th frame of view 8; (b) The corresponding disparity results estimated frame-independently; (c) Disparity maps generated via spatial-temporal optimization. (d)-(e) Magnified regions from (b) and (c) which demonstrate the improvement at corresponding regions.

is not robust enough when multiview videos are not color calibrated well. Combining texture and structure information into the data term will be more applicable. Secondly, motion estimation can be further refined in our framework, which leads to more reliable depth estimation. We may use disparity maps to cross-validate the motion, and use motion to refine disparities as in [14]. Another problem of our method is that the framework is complicated and time consuming, limiting its applications in real time situation. As a result, speeding up the approach, especially implementing it in a parallel fashion will also be our future work.

Acknowledgement

This work is supported by NSFC-Guangdong Joint Fund (No. U0935001), the National High Technology Research and Development Program of China (No. 2009AA01Z327) and NSFC (No. 60972013).

References

[1] C. L. Zitnick and S. B. Kang. Stereo for image-based rendering using image oversegmentation. *IJCV*, 75(1):49–65, 2007.

[2] C. Tomasi, R. Manduchi. Bilateral filtering for gray and color images. *ICCV*, 1998.

[3] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004.

[4] D. Comaniciu and P. Mee. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002.

[5] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[6] E.S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. *ICCV*, 2007.

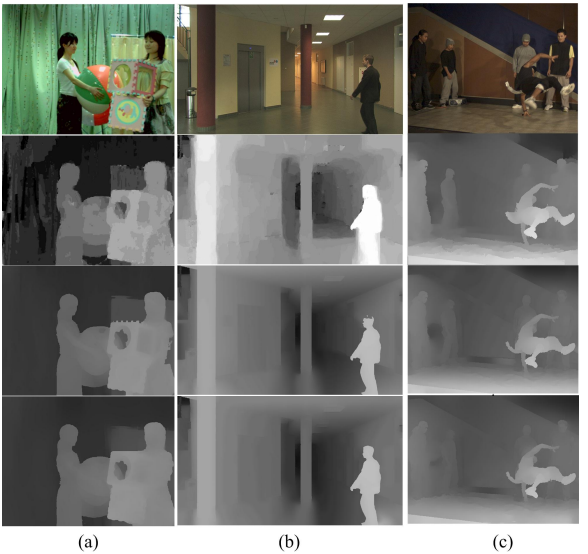

Figure 5: Disparity maps comparison of other three sequences. (top row) One color image of "Akko&Kayo", "Poznan Hall" and "breakdancers"; (second row) Disparity maps estimated frame-independently in other's work; (third row) Recovered disparity maps with only spatial consistency by our method; (bottom row) The final results via spatial-temporal optimization.

[7] F. Liu, V. Philomin. Disparity estimation in stereo sequences using scene flow. *BMVC*, 2009.

[8] G.F. Zhang, J.Y. Jia, T.T. Wong, H.J. Bao. Consistent depth maps recovery from a video sequence. *PAMI*, 31(6):974–988, 2009.

[9] H. Tao, H.S. Sawhney, and R. Kumar. Dynamic depth recovery from multiple synchronized video streams. *CVPR*, 2001.

[10] I. Feldmann, M. Mueller, F. Zilly, R. Tanger, K. Mueller, A. Smolic, P. Kauff, and T. Wiegand. HHI Test Material for 3D Video. *ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15413*, 2008.

[11] J. Sun, Y. Li, and S.B. Kang. Symmetric stereo matching for occlusion handling. *CVPR*, 2005.

[12] Kwatra V., Schödl A., Essa I., Turk G. and Bobick A. Graphcut textures: Image and video synthesis using graph cuts. *In Proceedings of ACM SIGGRAPH*, 2003.

[13] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner. Poznań. Multiview video test sequences and camera parameters. *ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050*, 2009.

[14] M. Gong. Enforcing temporal consistency in real-time stereo estimation. *ECCV*, 2006.

[15] M. Sormann A. Klaus and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. *ICPR*, 2006.

[16] M. Tanimoto, T. Fujii and K. Suzuki. First version of depth maps for poznan 3d/ftv test sequences. *ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15090*, 2008.

[17] N.N. Zheng J. Sun and H.Y. Shum. Stereo matching using belief propagation. *PAMI*, 25(7):787–800, 2007.

[18] O. Sankiewicz, K. Wegner and M. Domanski. First version of depth maps for poznan 3d/ftv test sequences. *ISO/IEC JTC1/SC29/WG11 MPEG 2010/M17176*, 2010.

[19] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006.

[20] R. Yang H. Stewénius Q. Yang, L.Wang and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *CVPR*, 2006.

[21] R. Yang S. Wang M. Liao Q. Yang, L. Wang and D. Nistér. Real-time global stereo matching using hierarchical belief propagation. *BMVC*, 2006.

[22] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *ICCV*, 1999.

[23] S.B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *IJCV*, 58(2):139–163, 2004.

[24] S.B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. *CVPR*, 2001.

[25] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto and Y. Suenaga. Multipoint measuring system for video and sound: 100-camera and microphone system. *ICME*, 2006.

[26] Y. Taguchi, B. Wilburn, and L. Zitnick. Stereo reconstruction with mixed pixels using adaptive over-segmentation. *CVPR*, 2008.