

Multiview video depth estimation using spatial-temporal consistency

Mingjin Yang
ymj04@mails.tsinghua.edu.cn
Xun Cao
xuncao@gmail.com
QiongHai Dai
qh dai@mail.tsinghua.edu.cn

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

To extract smooth depth videos of dynamic scenes from multiview video sequences, several methods combining spatial consistency of depth maps with temporal consistency have been proposed. Generally, traditional algorithms first employ belief propagation (BP) and segmentation to generate single time depth maps, and then incorporate temporal correspondences found by optical flow or scene flow into different optimization models.

In this paper, we present an approach to recover spatio-temporally consistent depth maps from multiview synchronized and calibrated video streams. Depth maps are first initialized by combining left-right view matching with color based segmentation (step 1). Then the idea of geometry consistency in Zhang et al.'s work [2] is borrowed to ensure the spatial consistency at each time instant (step 2). Finally, temporal information is integrated into a unified bundle optimization framework to generate depth videos with spatial-temporal consistency (step 3).

Given a multiview video sequence I with N views, and T frames for each view ($I = \{I_{t,n} | t = 1, \dots, T; n = 1, \dots, N\}$), our goal is to estimate the depth maps Z ($Z = \{Z_{t,n} | t = 1, \dots, T; n = 1, \dots, N\}$) or disparity maps D ($D = 1/Z$). The global energy function composed of the data term E_d and smoothness E_s term is defined as Eq.(1):

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

where E_d (2) gives the conditional probability of $D_{t,n}$ given I and D excluding $D_{t,n}$, which measures the fitness of D to I . And E_s (3) is used to ensure the smoothness on disparities of neighboring pixels.

$$E_d(D_{t,n}; I, D \setminus D_{t,n}) = \sum_{x_{t,n}} (1 - u(x_{t,n})) \cdot L(x_{t,n}, D(x_{t,n})), \quad (2)$$

where $u(x_{t,n}) = 1 / \max_{D(x_{t,n})} L(x_{t,n}, D(x_{t,n}))$ is a normalization factor. The differences of E_d in three steps lie in different likelihood $L(x_{t,n}, D(x_{t,n}))$.

$$E_s(D_{t,n}) = \sum_x \sum_{y \in N(x)} \omega \cdot \lambda(x, y) \cdot \min\{|D(x) - D(y)|, \eta\}, \quad (3)$$

where $N(x)$ denotes the set of neighboring pixels of x , and ω is the smoothness strength constant. η determines the upper limit of contribution of disparity difference, and smoothness weight $\lambda(x, y) = \varepsilon / (\varepsilon + \|I(x) - I(y)\|_2)$ preserves disparity discontinuity at color outliers.

In step 1, we use left and right views to generate the disparity of centre view, under the assumption that each pixel is visible in at least one adjacent view. Therefore the likelihood $L(x_{t,n}, D(x_{t,n}))$ in Eq.(2) is determined by the maximum of color similarity P_c (4) (denoted as $L \sim P_c$):

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

Here $n' = n - 1$ or $n + 1$, and $x_{t,n'}$ in (t, n') is the projection of $x_{t,n}$ with $D(x_{t,n})$. σ_c is a constant to control the shape of color similarity function.

After minimizing the energy function (1) by loopy BP [3], we employ the mean-shift color segmentation [1] 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 $D(x_{t,n}) = a_k x + b_k y + c_k$. Then the best plane parameters (a_k, b_k, c_k) are estimated by using a non-linear continuous optimization method to minimize Eq.(1).

As shown in Figure 1, the ideal disparity of $x_{t,n'}$ is $D^*(x_{t,n'})$ in step 2, with which $x_{t,n'}$ will map to a same 3D point \hat{X} as $x_{t,n}$ with $D(x_{t,n})$. But its inaccurate initial disparity value $D(x_{t,n'})$ induces the deviation (marked as red line segment). Therefore we define the spatial coherence constraint P_s as a Gaussian distribution of the difference between $D(x_{t,n'})$ and $D^*(x_{t,n'})$.

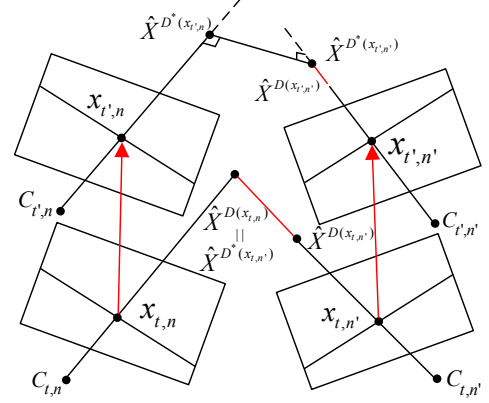


Figure 1: Spatial and temporal coherence constraint.

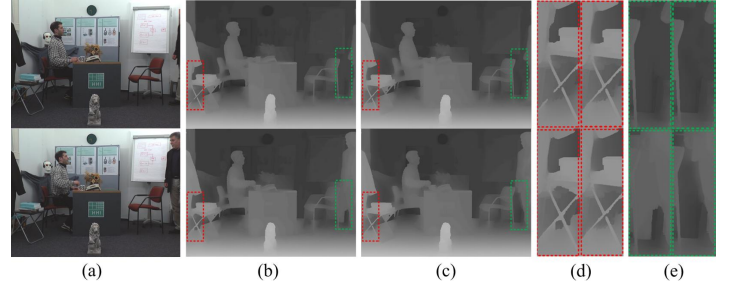


Figure 2: Disparity maps of "book arrival". (a) The 13th and 18th frame of view 8; (b) Disparity results estimated frame-independently; (c) Disparity maps generated via spatial-temporal optimization. (d-e) Magnified regions from (b) and (c) which demonstrate the improvement at corresponding regions.

Then we get the likelihood $L \sim P_c \cdot P_s$. Besides, we also take the occlusion detection into consideration, and compute disparity maps of single time instant one by one with other disparity maps fixed temporarily.

In step 3, optical flow is first used to get the motion vectors, which are represented by red arrows in Figure 1. Ideally, the optical rays $C_{t',n'} \hat{x}_{t',n'}$ will cross $C_{t,n} \hat{x}_{t,n}$ with the intersecting point $\hat{X}^{D^*}(x_{t,n}) = \hat{X}^{D^*}(x_{t,n'})$. However, 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. Then 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} . In addition, spatial-temporal coherence constraint P_{st} is defined as the same Gaussian distribution as P_s , which is a function of the difference between $D(x_{t',n'})$ and its ideal value $D^*(x_{t',n'})$. Finally we use loopy BP again to minimize Eq.(1) with $L \sim P_c \cdot P_s + P_{cst} \cdot P_{st}$.

The experimental results of "bookarrival" (Figure 2) demonstrate that the inaccurate results 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.

- [1] D. Comaniciu and P. Mee. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002.
- [2] 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.
- [3] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006.