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 (\( I = \{ I_{t,n} | t = 1, \ldots, T; n = 1, \ldots, N \} \)), our goal is to estimate the depth maps \( Z = \{ Z_{t,n} | t = 1, \ldots, T; n = 1, \ldots, N \} \) or disparity maps \( D = \{ D_{t,n} | t = 1/2 \} \). The global energy function composed of the data term \( E_{d} \) and smoothness \( E_{s} \) term is defined as Eq.(1):

\[
E(D; I) = \sum_{t=1}^{T} \sum_{n=1}^{N} \left[ E_{d}(D_{t,n}; I, D_{t,n}) + E_{s}(D_{t,n}) \right],
\]

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.

\[
E_{d}(D_{t,n}; I, D_{t,n}) = \sum_{x} \left( 1 - u(x_{n} \cdot D(x_{n}, D_{t,n})) \right),
\]

where \( u(x) = 1/\max_{y} D(x_{n}, D(x_{n})) \) is a normalization factor. The differences of \( E_{d} \) in three steps lie in different likelihood \( D(x_{n}, D_{t,n}) \).

\[
E_{s}(D_{t,n}) = \sum_{x} \sum_{y} \omega_{x,y} \cdot \lambda_{x,y} \cdot \min \{ |D(x_{n}) - D(y)|, \eta \},
\]

where \( N(x) \) denotes the set of neighboring pixels of \( x \), and \( \omega \) is the smoothness strength constant. \( \eta \) determines the upper limit of contribution of disparity difference, and smoothness weight \( \lambda_{x,y} = \varepsilon/\left(1 + ||I(x) - I(y)||_{2} \right) \) 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 at least one adjacent view. Therefore the likelihood \( L(x_{n}, D_{t,n}) \) in Eq.(2) is determined by the maximum of color similarity \( P_{c} \) (4) (denoted as \( L \sim P_{c} \)).

\[
P_{c}(x_{n}, D_{t,n}) = \frac{\alpha_{c}}{\sigma_{c} + \|I(x_{n}) - I(y_{n})\|_{2}},
\]

Here \( n' = n - 1 \) or \( n + 1 \), and \( x_{n'} \) in \( (t, n') \) is the projection of \( x_{n} \) with \( D(x_{n}, I_{t,n}) \). \( \alpha_{c} \) 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 \( x_{n} \in S_{k} \) is denoted as \( D(x_{n}) = a x_{k} + b y_{k} + c_{k} \). 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 \( D^{*}(x_{n'}) \) in step 2, with which \( x_{n'} \) will map to a same 3D point \( \tilde{X} \) as \( x_{n} \) with \( D(x_{n}) \). But its inaccurate initial disparity value \( D(x_{n'}) \) induces the deviation (marked as red line segment). Therefore we define the spatial coherence constraint \( P_{c} \) as a Gaussian distribution of the difference between \( D(x_{n'}) \) and \( D^{*}(x_{n'}) \).

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 (\( I = \{ I_{t,n} | t = 1, \ldots, T; n = 1, \ldots, N \} \)), our goal is to estimate the depth maps \( Z = \{ Z_{t,n} | t = 1, \ldots, T; n = 1, \ldots, N \} \) or disparity maps \( D = \{ D_{t,n} | t = 1/2 \} \). The global energy function composed of the data term \( E_{d} \) and smoothness \( E_{s} \) term is defined as Eq.(1):

\[
E(D; I) = \sum_{t=1}^{T} \sum_{n=1}^{N} \left[ E_{d}(D_{t,n}; I, D_{t,n}) + E_{s}(D_{t,n}) \right],
\]

where \( E_{d} \) 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 \( E_{s} \) is used to ensure the smoothness on disparities of neighboring pixels.

\[
E_{d}(D_{t,n}; I, D_{t,n}) = \sum_{x} \left( 1 - u(x_{n} \cdot D(x_{n}, D_{t,n})) \right),
\]

where \( u(x) = 1/\max_{y} D(x_{n}, D(x_{n})) \) is a normalization factor. The differences of \( E_{d} \) in three steps lie in different likelihood \( D(x_{n}, D_{t,n}) \).

\[
E_{s}(D_{t,n}) = \sum_{x} \sum_{y} \omega_{x,y} \cdot \lambda_{x,y} \cdot \min \{ |D(x_{n}) - D(y)|, \eta \},
\]

where \( N(x) \) denotes the set of neighboring pixels of \( x \), and \( \omega \) is the smoothness strength constant. \( \eta \) determines the upper limit of contribution of disparity difference, and smoothness weight \( \lambda_{x,y} = \varepsilon/\left(1 + ||I(x) - I(y)||_{2} \right) \) 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 at least one adjacent view. Therefore the likelihood \( L(x_{n}, D_{t,n}) \) in Eq.(2) is determined by the maximum of color similarity \( P_{c} \) (4) (denoted as \( L \sim P_{c} \)).

\[
P_{c}(x_{n}, D_{t,n}) = \frac{\alpha_{c}}{\sigma_{c} + \|I(x_{n}) - I(y_{n})\|_{2}},
\]

Here \( n' = n - 1 \) or \( n + 1 \), and \( x_{n'} \) in \( (t, n') \) is the projection of \( x_{n} \) with \( D(x_{n}, I_{t,n}) \). \( \alpha_{c} \) 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 \( x_{n} \in S_{k} \) is denoted as \( D(x_{n}) = a x_{k} + b y_{k} + c_{k} \). 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 \( D^{*}(x_{n'}) \) in step 2, with which \( x_{n'} \) will map to a same 3D point \( \tilde{X} \) as \( x_{n} \) with \( D(x_{n}) \). But its inaccurate initial disparity value \( D(x_{n'}) \) induces the deviation (marked as red line segment). Therefore we define the spatial coherence constraint \( P_{c} \) as a Gaussian distribution of the difference between \( D(x_{n'}) \) and \( D^{*}(x_{n'}) \).