Structure-aware Priority Belief Propagation for Depth Estimation

Kuanyu Ju, Botao Wang, Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai 200240, China
{jukuanyu, botaowang, xionghongkai}@sjtu.edu.cn

***Abstract*—2D to 3D video conversion has become a popular manner to produce 3D contents, in which the key task is to estimate the depth for the images. However, the main limitation of conventional motion-based depth propagation methods is that they are susceptible to inaccurate motion estimation and occlusion. To address this problem, we propose a two-stage structure-aware depth propagation method for semi-automatic 2D-to-3D conversion. In the first stage, an initial depth map is generated by shifted bilateral filtering over the temporally consistent region, which is determined by validating the forward and backward optical flow fields. In the second stage, the depth of the temporally inconsistent region will be estimated by solving a multi-label graph inference problem. In particular, an efficient priority belief propagation algorithm is developed, in which the priority of nodes to propagate messages depends on the structure saliency from tensor voting. Experimental results show that the proposed method outperforms existing depth estimation methods in public benchmark.**

***Index Terms*—Depth estimation, shifted bilateral filtering, priority belief propagation, tensor voting, optical flow**

I. Introduction

3D video can provide an enhanced visual experience with depth perception beyond conventional 2D content. With the growth of 3D display devices, the increasing demand for 3D contents has aroused a significant challenge to the 3D industry. One popular way to produce 3D videos is to perform 2D-to-3D conversion to 2D videos. Existing 2D-to-3D techniques can be divided into fully-automatic and semi-automatic approaches based on the involvement of human-computer interaction. Fully-automatic approaches estimate the depth of 2D videos by exploring monocular depth cues, e.g., texture gradients, linear perspective and motion parallax, and optimizing them with learning-based algorithms. On the other hand, semi-automatic approaches adopt human-computer interaction to guide the depth estimation process of key frames for better synthesis effect. To be specific, users are desired to mark only a few scribbles on the key frames to produce dense depth maps [1], and the depth maps of the non-key frames are estimated by propagating the depth from the key frames. In [2], depth is attained by bilateral filtering and refined through a block-based motion compensation from previous frames. In 2011, bilateral filtering based method was extended in [3], where the depth map is propagated by shifted bilateral filtering with motion information. Li et al. [4] suggested the depth propagation for non-key frames via bi-directional motion estimation. The bi-directional motion vectors are estimated to determine the depth propagation strategy that depth in areas with matched motion vectors will be copied and remaining areas will be estimated in pixel level by bilateral filtering. However, these motion-based depth propagation methods are sensitive to inaccurate motion estimation or object occlusion, which typically occur along the boundaries of objects.

This paper proposes a novel structure-aware depth propagation approach, which is robust against coarse motion estimation and occlusion. It is worth mentioning that it makes two technical contributions. The first contribution is to devise a two-stage depth propagation approach, which applies different depth estimation schemes to the frames based on the temporal consistency. To be specific, a frame is decomposed into temporally-consistent and temporally-inconsistent regions by validating the forward and backward optical flow fields. In the first stage, the depth of the temporally-consistent region, which tends to be uniform, is estimated by shifted bilateral filtering, and an initial depth map with black holes in the temporally-inconsistent region will be generated. In the second stage, the depth of the temporally-inconsistent region, where inaccurate motion and occlusion are more likely to occur, will be estimated by belief propagation, which is effective in preserving the discontinuities in the depth map. The second contribution is to develop a structure-aware *priority belief propagation* (priority-BP) algorithm to estimate the depth of the temporally-inconsistent region. To be concrete, the depth estimation of the temporally-inconsistent region is formulated as a multi-label Markov random field inference problem. An efficient priority belief propagation algorithm is developed to solve this problem by prioritizing the nodes in propagating the messages in a heuristic manner based on their structure saliency, so that the discontinuity in the depth map can be well preserved. In particular, the structure saliency of the frame is derived by tensor voting, which reflects the distribution of the edges and boundaries of the object in the frame.

II. Two-Stage Depth Propagation Algorithm

The framework of the proposed structure-aware depth propagation algorithm is illustrated in Fig. 1. The initial depth map of the current frame is generated by shifted bilateral

978-1-4673-7314-2/15/$31.00 ©2015 IEEE

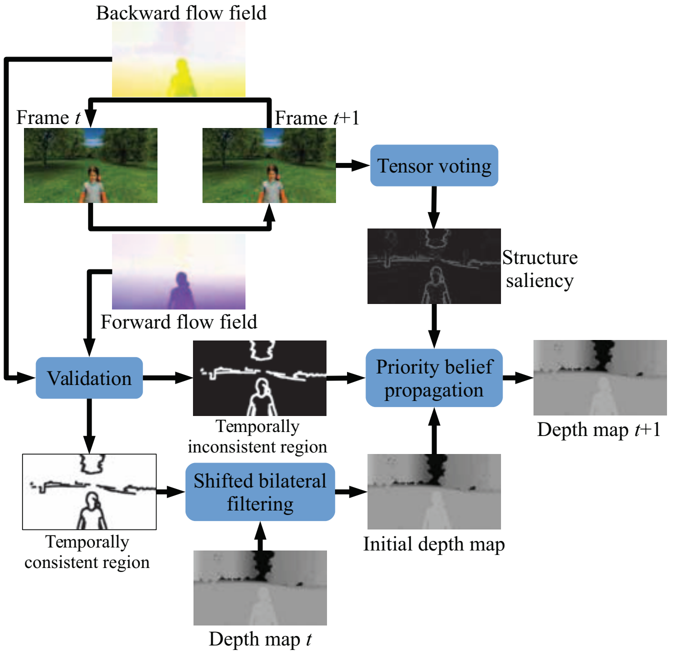

Fig. 1. Framework of the structure-aware depth propagation algorithm.

filtering over the temporally consistent region of the previous depth map, which is determined by validating the forward and backward optical flow fields. Since the initial depth map contains many black holes in the temporally inconsistent region, the depth estimation in these regions will be formulated as a multi-label MRF inference problem. Furthermore, a priority-BP algorithm is developed to solve this discrete optimization problem by organizing the priority of the nodes in passing messages.

A. Depth Estimation in the Temporally Consistent Region

In the first stage, the depth of the pixels that satisfy the temporal consistency condition will be estimated. Specifically, the temporal consistency is defined by cross-validating the forward and backward optical flow fields, which are computed by [5]. In the temporally consistent region, the backward optical flow vectors will be the opposite of the corresponding forward optical flow vectors. However, in the temporally inconsistent region, this condition does not hold, and the pixels may be occluded or falsely estimated in motion.

Consequently, the depth of the pixels in the temporally consistent region is initialized by the shifted bilateral filter:

$$d^{t+1}(x) = \frac{\sum_y f_s(x+v(x),y)\cdot f_r(C^{t+1}(x),C^t(y))\cdot d^t(y)}{\sum_y f_s(x+v(x),y)\cdot f_r(C^{t+1}(x),C^t(y))}, \quad (1)$$

where x is the two-dimensional coordinates of the pixel, $v(x)$ is the forward flow vector, $C(x)$ is the color vector, and $d(x)$ is the depth at x. In Eq. (1), f_s and f_r are the spatial and color filter kernels, which are defined as:

$$f_s(x,y) = \exp\left(-\frac{\parallel x-y \parallel^2}{2\rho^2}\right), \text{where } y \in N(x) \quad (2)$$

$$f_r(C(x),C(y)) = \begin{cases} \exp\left(-\frac{\|C(x)-C(y)\|^2}{2\sigma^2}\right), \|C(x)-C(y)\|^2 < \varepsilon \\ 0, \text{ otherwise} \end{cases} \quad (3)$$

As a result, the initial depth map can be generated with quite a few black holes in the temporally inconsistent region, which will be filled up in the second stage.

B. Depth Estimation in the Temporally Inconsistent Region

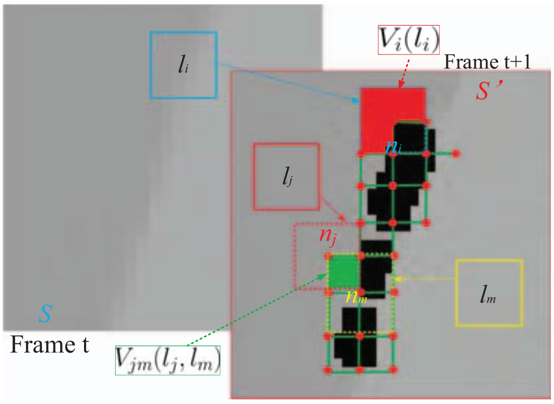

Fig. 2. An MRF model is established to estimate the depth in the temporally inconsistent region.

The second stage depth propagation estimates the depth of the temporally inconsistent region by solving an MAP approximation problem. To be concrete, with the source region S in the previous depth map and S' in the initial depth map, a spatio-temporal MRF model is established, which is illustrated in Fig. 2. The black holes in the initial depth map will be filled with suitable patches from S and S'. The candidate patches for each node should match both the missing region of the node and its neighborhood. Specifically, the matching cost of patch A and patch B is defined by:

$$D(A,B) = \|d(x_A)-d(x_B)\|^2 + \alpha\|I(x_A)-I(x_B)\|^2, \quad (4)$$

where x_A and x_B is the set of pixels in patch A and B, respectively, and I is the image intensities.

With the matching cost defined in Eq. 4, the data term $V_i(l_i)$ and the smoothness term $V_{ij}(l_i,l_j)$ in the MRF model can be defined. The data term $V_i(l_i)$ measures how well the patch l_i matches the source region R_c around the node n_i:

$$V_i(l_i) = \gamma D(l_i,R_c), \quad (5)$$

where γ is adaptive to inner node or boundary node. Similarly, the smoothness term $V_{ij}(l_i,l_j)$ measures the compatibility of patch l_i and patch l_j in the neighboring nodes n_i and n_j in the overlapping area:

$$V_{ij}(l_i,l_j) = D(l_i,l_j) = \sum_{x\in R_{n_i}\cap R_{n_j}} |l_i(x)-l_j(x)|^2. \quad (6)$$

Based on the data term and the smoothness term, the cost of assigning a suitable patch $\hat{l}_i \in L$ to each node in MRF will be minimized:

$$\min E(\hat{l}) = \sum_{n_i \in \mathcal{V}} V_i(\hat{l}_i) + \sum_{(n_i,n_j)\in\mathcal{E}} V_{ij}(\hat{l}_i, \hat{l}_j) \tag{7}$$

where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges in the MRF model, and L is the set of candidate patches.

Belief propagation (BP) [6] can solve Eq. (7), which finds a maximum a posteriori (MAP) estimator by propagating local messages along the nodes in the MRF model. Messages from node n_i to node n_j form a set $\{m_{ij}(l)\}_{l\in L}$, where $m_{ij}(l)$ indicates how likely node n_i considers that node n_j should be assigned with patch l. Thus, a set of beliefs $\{b_i(l)\}_{l\in L}$ can be obtained for each node:

$$b_i(l) = -V_i(l) - \sum_{(j,i)\in\mathcal{E}} m_{ji}(l), \tag{8}$$

where $b_i(l)$ approximates the max-marginal of the posterior at node n_i and indicates how likely label l will be assigned to that node. Finally, a node is assigned with the label of maximum belief, i.e., $\hat{l}_i = \text{argmax}_{l\in L} b_i(l)$.

III. Structure-Aware Priority Belief Propagation

Considering that BP is not appropriate for problems with a large number of labels, we develop an efficient structure-aware priority-BP algorithm to solve Eq. (7).

A. Structure Saliency via Tensor Voting

The tensor voting algorithm [7] is leveraged to estimate the geometric structures in unknown regions. As the discontinuities of depth always occur near the object boundaries, the edge map of the original frame will be computed by the Canny operator, which serves as the input to the tensor voting algorithm. First, the edge response at each point will be encoded into a second order symmetric tensor token, which represents the orientation and confidence of the point. Specifically, if a point is located on a curve, the token can be represented by its associated tangent or normal as a stick tensor: $\begin{pmatrix} n_x^2 & n_x n_y \\ n_y n_x & n_y^2 \end{pmatrix}$ where $(n_x, n_y)^\top$ is the normal of the point. Otherwise, the token is represented by a unit ball tensor $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

In the next step, each point communicates with its neighbors by casting and receiving votes, and updates the information it carries. To be concrete, each point collects votes from its receiving field, and sums them up into a covariance matrix, which is a generic second order symmetric tensor:

$$S = \begin{bmatrix} \sum_i v_{x,i}^2 & \sum_i v_{x,i} v_{y,i} \\ \sum_i v_{y,i} v_{x,i} & \sum_i v_{y,i}^2 \end{bmatrix}. \tag{9}$$

Since S is a second order symmetric tensor, its eigenvalues are always non-negative, which can be denoted by $\lambda_1 \geq \lambda_2 \geq 0$, and the corresponding eigenvectors form an orthonormal basis, which are denoted by e_1 and e_2. According to the spectrum theorem, Eq. (9) can be rewritten as:

$$S = (\lambda_1 - \lambda_2) e_1 e_1^\top + \lambda_2 (e_1 e_1^\top + e_2 e_2^\top), \tag{10}$$

where $e_1 e_1^\top$ is a two-dimensional stick tensor with e_1 indicating the normal direction of the curve, and $e_1 e_1^\top + e_2 e_2^\top$ is a two-dimensional plate tensor. As a result, the curve saliency s_i of node n_i is defined by:

$$s_i = \lambda_1 - \lambda_2. \tag{11}$$

Fig. 3 shows an example of the structure saliency in the black holes, which is consistent with the discontinuity in depth. The regions with salient structure have high priority to be estimated than other regions, so that the discontinuity and the structural consistency can be preserved.

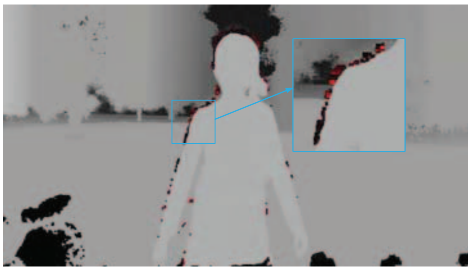

Fig. 3. The structure saliency, represented by the red points, of the temporally inconsistent region.

B. Priority Belief Propagation

To solve the MRF inference problem, and efficient priority-BP algorithm is developed, which can significantly reduce the computational complexity, and converges in a few iterations. In general, the priority of a node in propagating messages to its neighbors is determined by its confidence about label. In particular, the definition of confidence for node n_i depends on the current set of beliefs $\{b_i(l)\}_{l\in L}$, which have been estimated in Eq. (8). Based on the observation that $b_i(l)$ is roughly related to how likely label l is assigned to node n_i, we measure the confidence of a node by counting the number of likely labels that exceed a threshold b_{conf}: $P_i = |l \in L : b_i(l) \geq b_{conf}|$. Since discontinuity always exists at object boundary in the depth map, keeping the structure of the depth map is an essential task in depth propagation. Therefore, the structure-aware beliefs are defined as:

$$b_i(l) = -V_i(l) - \sum_{(j,i)\in\mathcal{E}} m_{ji}(l) + \beta s_i. \tag{12}$$

Obviously, the nodes with higher structure saliency have larger beliefs than others. As a consequence, regions with salient structure have high priority to be estimated. In this way, the stable marginal belief distribution of each node can be efficiently approximated by message propagation. In practice, the number of iterations required for convergence will be less than 4, which is much quicker than conventional BP algorithm.

To deal with large number of labels, we reduce the number of candidate labels by exploiting the beliefs calculated by BP. After sorting $\{b_i(l)\}_{l\in L}$ for node n_i, we discard the low-confident labels for each node, and only L_{max} candidate labels with $L_{max} \ll L$ are selected as “active labels”. The computation complexity of priority-BP can be significantly reduced in this manner.

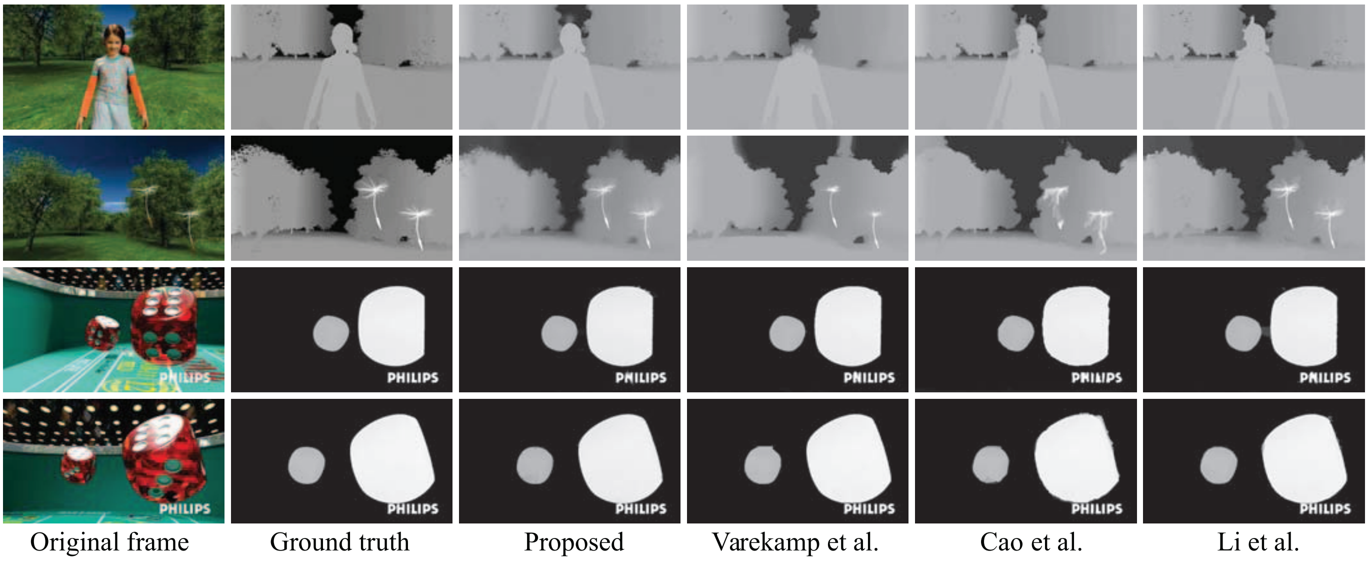

Fig. 4. Estimated depth maps of four testing sequences (from top to bottom): *Girl*, *Seed*, *Dice 1* and *Dice 2*.

TABLE I
AVERAGE MEAN SQUARED ERROR

	Girl	*Seed*	*Dice 1*	*Dice 2*	*Inner-gate*
Proposed	**38.87**	**150.30**	**68.21**	**39.38**	**129.63**
[2]	94.93	584,94	124.75	70.1	529.96
[3]	40.04	245.50	249.98	191.53	400.77
[4]	41.98	190.77	86.97	69.25	156.41

TABLE II
THE AVERAGE SSIM SCORES

	Girl	*Seed*	*Dice 1*	*Dice 2*	*Inner-gate*
Proposed	**0.981**	**0.941**	**0.988**	**0.992**	**0.970**
[2]	0.971	0.928	0.983	0.990	0.955
[3]	0.977	0.933	0.978	0.981	0.960
[4]	0.976	0.935	0.985	0.987	0.966

IV. EXPERIMENTAL RESULTS

The proposed method is evaluated on five video sequences, where *Girl*, *Seed*, *Dice 1* and *Dice 2* are collected from the Philips WowVc© project, and *Inner-gate* is obtained from [4]. In addition, three popular depth estimation methods, i.e., [2], [3], and [4], are also tested as comparison. The average mean squared error of the four approaches upon the five sequences are displayed in Table I, and the proposed method outperforms the other methods in all sequences. Furthermore, the relevant Structure Similarity (SSIM) index is also computed, which is demonstrated in Table II. Again, the proposed method achieves the highest SSIM scores in all sequences. Finally, some examples of the estimated depth maps are displayed in Fig. 4. It is clearly shown that the proposed method can preserve the structure in depth map effectively, especially near the boundary.

V. CONCLUSION

This paper proposes a structure-aware depth propagation algorithm, which is robust against inaccurate motion estimation and object occlusion. Specifically, the depth map is estimated in two stages. The first stage computes an initial depth map for the temporally consistent region using shifted bilateral filtering, and the second stage formulates the depth estimation in the temporally inconsistent region as a graph inference problem, which can be efficiently solved by the priority belief propagation algorithm, while preserving the structure in the depth map. Experimental results on various sequences clearly demonstrate the effectiveness of the proposed method.

ACKNOWLEDGMENT

The work was supported in part by the NSFC, under grants 61472234, 61271211, and 61425011.

REFERENCES

[1] M. Guttmann, L. Wolf, and D. Cohen-Or, “Semi-automatic stereo extraction from video footage,” in *Proc. Int’l Conf. Computer Vision (ICCV’09)*, Kyoto, Japan, Sept. 2009, pp. 136–142.

[2] C. Varekamp and B. Barenbrug, “Improved depth propagation for 2D-to-3D video conversion using key-frames,” in *Proc. IET 4th European Conf. Visual Media Production (CVMP’07)*, London, UK, Nov. 2007, pp. 1–7.

[3] X. Cao, Z. Li, and Q. Dai, “Semi-automatic 2D-to-3D conversion using disparity propagation,” *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 491–499, Jun. 2011.

[4] Z. Li, X. Cao, and Q. Dai, “A novel method for 2D-to-3D video conversion using bi-directional motion estimation,” in *Proc. Int’l Conf. Acoustics, Speech and Signal Processing (ICASSP’12)*, Kyoto, Japan, Mar. 2012, pp. 1429–1432.

[5] T. Brox and J. Malik, “Large displacement optical flow: descriptor matching in variational motion estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[6] P. Felzenszwalb and D. Huttenlocher, “Efficient belief propagation for early vision,” *Int’l J. Computer Vision*, vol. 70, no. 1, pp. 41–54, May 2006.

[7] G. Medioni, C. Tang, and M. Lee, “Tensor voting: Theory and applications,” in *Proc. RFIA*, Paris, France, vol. 3, 2000.