Spatiotemporal Stereo Matching with 3D Disparity Profiles

Yongho Shin
yongho@gist.ac.kr
Kuk-Jin Yoon
kjyoon@gist.ac.kr

Computer Vision Laboratory
Gwangju Institute of Science and Technology (GIST), South Korea

Abstract

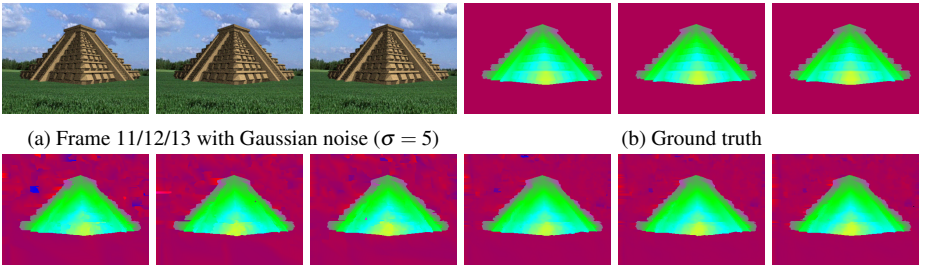
Adaptive support weights and over-parameterized disparity estimation truly improve the accuracy of stereo matching by enabling window-based similarity measures to handle depth discontinuities and non-fronto-parallel surfaces more effectively. Nevertheless, a disparity map sequence obtained in a frame-by-frame manner still tends to be inconsistent even with the use of state-of-the-art stereo matching methods. To solve this inconsistency problem, we propose a window-based spatiotemporal stereo matching method. We exploit the 3D disparity profile, which represents the disparities and window normals over multiple frames, and incorporate it into the PatchMatch Belief Propagation (PMBP) framework. Here, to make the 3D disparity profile more reliable, we also present the optical flow transfer method. Experimental results show the proposed method yields more consistent disparity map sequences than does the original PMBP-based method.

1 Introduction

Stereo matching is to identify correspondences between images captured at different viewpoints. In stereo matching, similarity measures between pixels are essential. They can be roughly classified into pixel- and window-based measures. The pixel-based measure uses information from a single pixel and thus is very fast and efficient. However, it has low discriminative power and is vulnerable to noise. By contrast, the window-based measure uses information from multiple pixels in the window and offers more discriminative power. In addition, it is more robust against noise than is the pixel-based measure. However, the window-based measures commonly assume fronto-parallel scene surfaces and, therefore, yield a fattening artifact at depth discontinuities and non-fronto-parallel surfaces. For this reason, many methods have been proposed to overcome the limitations of window-based similarity measures. For example, adaptive support weights [20] and over-parameterized disparity estimation [2] truly improve the accuracy of stereo matching by allowing window-based similarity measures to handle depth discontinuities and non-fronto-parallel surfaces more effectively. Actually, some recent methods that employ adaptive support weights and the over-parameterized disparity estimation with a smoothness constraint have produced high quality disparity maps [1, 6, 18].

Nevertheless, when handling stereo image sequences, a disparity map sequence obtained in a frame-by-frame manner remains inconsistent even with the use of state-of-the-art stereo

Pages 152.1-152.12
DOI: https://dx.doi.org/10.5244/C.29.152



(a) Frame 11/12/13 with Gaussian noise ($\sigma = 5$) (b) Ground truth

(c) Results of PMBP [1] (d) Results of PMBP-3D

Figure 1: Stereo matching results for the temple dataset

matching methods. Figure 1(c) shows a disparity map sequence containing a considerable fluctuation artifact in a static background. This fluctuation artifact is also problematic in dynamic regions. Even when scene objects and cameras move smoothly, considerable temporal disparity variation can be generated because of noise and illumination changes between frames.

To solve this inconsistency problem, we propose a window-based spatiotemporal stereo matching method exploiting a 3D disparity profile in a PatchMatch Belief Propagation (PMBP) framework. The 3D disparity profile is a structure that represents disparities and normal vectors of a window over multiple frames. We assume that the elements of a 3D disparity profile vary smoothly with time to produce a consistent disparity map. Our main contribution is a method that produces smooth variations of the 3D profile elements and a means to use the 3D disparity profile for the PMBP framework. Experimental results show the proposed method yields more consistent disparity map sequences than does the original PMBP-based method [1].

1.1 Related works

In general, spatiotemporal stereo matching uses multiple frames to obtain a disparity map of the frame of interest. It usually assumes that disparity maps are similar in consecutive frames and aggregates matching costs from neighboring frames [4, 7, 13] or propagates penalty costs to nearby frames [10]. Davis *et al.* [4] proposed a spatiotemporal framework using a 3D window to aggregate pixel-based costs from neighboring frames. Afterwards, this framework was improved to aggregate pixel-based costs with adaptive support weights. To increase the efficiency for computing adaptive support weights, Richardt *et al.* [13] extended a bilateral grid, and Hosni *et al.* [7] used a guided-filter. Khoshabeh *et al.* [10] proposed a post-processing method based on total variation regularization for consistent disparity maps. However, although these methods prove that the spatiotemporal stereo matching approach improves the consistency of consecutive disparity maps, they do not properly manage disparity variation between frames caused by camera and scene motion.

Two approaches exist for handling disparity variation: blocking undesirable temporal supports [12, 14] and explicitly modeling the variation [8, 15, 19, 22]. Larsen *et al.* [12] proposed a temporal belief propagation based on the spatiotemporal Markov random field having disconnected edges, in which adjacent vertices pass inconsistent messages. Sanchez-Riera *et al.* [14] presented a robust temporal normalized cross correlation (NCC) that only approves temporal supports when corresponding pixels at consecutive frames have similar

NCC values. However, these methods still do not guarantee consistent disparity maps because they have been designed just to avoid problems resulting from disparity variation, not to produce consistent results.

As the second approach, scene flow estimation methods have been proposed. These compute both stereo and motion from two consecutive frame pairs. Huguet and Devernay [8] modeled an energy function for the scene flow and solved it within a variational framework. Vogel *et al.* [19] proposed a super-pixel based scene flow method in which each super-pixel is regarded as a scene plane. However, these methods are not only computationally expensive, but also consistency is induced between only two frames. On the other hand, Zhang *et al.* [22] modeled a spatiotemporal similarity measure using a disparity and the gradient of a disparity function. Shin and Yoon [15] improved this model by applying adaptive support weight that is based on the variant of a guided filter. However, these two methods [15, 22] cannot properly handle image motion between frames. Sizintsev and Wildes [16, 17] proposed spatiotemporal orientation based similarity measures. Because these measures are defined with consecutive frames, they can yield temporally consistent disparity maps. However, similarity measures do not work well under large displacements because of the motion of a camera and objects, as an oriented filter cannot provide meaningful responses in this case.

Recently, Hung *et al.* [9] proposed a method that employs a structure containing disparity values of a point of interest in support frames. They called the structure a disparity profile. Although the proposed method is inspired by this method, remarkable differences exist between the two methods. We extend the disparity profile to utilize window-based similarity measures according to the over-parameterized disparity. In addition, we propose a new PMBP-based framework using the extended profile for computing consistent disparity map sequences. Furthermore, we propose an optical flow transfer method for establishing more reliable disparity profile by using optical flows of both views.

2 Proposed method

2.1 Over-parameterized disparity representation

A disparity is usually defined as a scalar value. However, it can also be defined by an over-parameterized form having additional information besides a disparity. For an image point $\mathbf{p} = [x_p\, y_p\, 1]^\top$, we can define a plane in the disparity space by a disparity value z_p and a normal $\mathbf{n}_p^d = \left[n_{x,p}^d\, n_{y,p}^d\, n_{z,p}^d\right]^\top$ as $n_{x,p}^d x + n_{y,p}^d y + n_{z,p}^d z - n_{x,p}^d x_p - n_{y,p}^d y_p - n_{z,p}^d z_p = 0$. With the plane equation, we can compute a disparity value z_q for any point $\mathbf{q} = [x_q\, y_q\, 1]^\top$ on the disparity plane as,

$$z_q = -n_{xz,p} x_q - n_{yz,p} y_q + \left(n_{xz,p} x_p + n_{yz,p} y_p + z_p\right), \tag{1}$$

where $n_{xz} = n_x^d / n_z^d$ and $n_{yz} = n_y^d / n_z^d$. By Eq.(1), the over-parameterized disparity representation based on z_p and $\mathbf{n}_p^d$ of point $\mathbf{p}$ can define the disparity value of points on the window.

Suppose that the corresponding point $\mathbf{q}'$ in the target image for $\mathbf{q}$ in the reference image is defined as $\mathbf{q}' = \mathbf{q} - [z_q\, 0\, 0]^\top$. We can then express the relationship between two points by a linear transformation as

$$\mathbf{q}' = \begin{pmatrix} 1 + n_{xz,p} & n_{yz,p} & -n_{xz,p} x_p - n_{yz,p} y_p - z_p \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{q}. \tag{2}$$

As a different perspective, if we regard a window as the projection of a scene plane in the world, the linear transformation in Eq.(2) can be explained by a plane-induced homography. For a rectified camera setup, Eq.(2) is rewritten by the plane-induced homography as

$$\mathbf{q}' = \left(I - K \frac{1}{d} \begin{pmatrix} bn_x^s & bn_y^s & bn_z^s \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} K^{-1} \right) \mathbf{q}, \tag{3}$$

where K is an intrinsic matrix composed of focal lengths f_x and f_y, subscripts x and y represent x and y directions. $\mathbf{n}^s = \begin{bmatrix} n_x^s n_y^s n_z^s \end{bmatrix}^\top$ is the normal of a scene plane. b is a baseline between cameras, and d is a distance from the scene plane to the origin that is set to be the location of the reference camera. Specifically, d is defined as $d = -n_x^s x_p^s - n_y^s y_p^s - n_z^s z_p^s$, where $\begin{bmatrix} x_p^s y_p^s z_p^s \end{bmatrix}$ is the point on the scene plane corresponding to the center point of a window.

From Eq.(2) and Eq.(3), we can realize that the location and normal of a scene plane in the world are related to the location and shape of a corresponding window in the target image as [1]

$$z_p = \frac{f_x b}{z_p^s}, \quad n_{xz,p} = \frac{bn_x^s}{n_x^s x_p^s + n_y^s y_p^s + n_z^s z_p^s}, \quad n_{yz,p} = \frac{bn_y^s}{n_x^s x_p^s + n_y^s y_p^s + n_z^s z_p^s} \cdot \frac{f_x}{f_y}. \tag{4}$$

2.2 3D disparity profile

For a point in a frame of interest, if we know the temporal correspondences of the point at support frames, we can collect disparities of the correspondences and normal values of a window over multiple frames. We define a structure concatenating these disparities and normal values in support frames as a 3D disparity profile.

As shown in Eq.(4), a scene plane motion is related to the location and shape variations of a corresponding window. For example, if a fronto-parallel scene plane moves away from the reference camera by increasing z_p^s, then the corresponding window of the target image should be square, and located with decreasing disparity. If a scene plane normal is changed, then the shape of a corresponding window should be changed as well. Hence, we can exploit the characteristics of a scene plane motion to find the proper window location and shape in image frames. In general, objects and cameras in a scene can be assumed to move with smooth variations in direction and velocity. In this case, the scene plane normal and location are smoothly changed, and the window shape and location must also be changed smoothly. Based on this assumption, we generate an improved disparity profile, and use it to obtain consistent disparity maps.

2.3 Energy modeling

To compute consistent disparity maps for a rectified stereo image sequence, we employ an energy minimization scheme for each frame. Our energy function is defined by an over-parameterized disparity $\mathbf{h} = [z \;\; n_{xz} \, n_{yz}]^\top$ and a guidance disparity $\tilde{\mathbf{h}} = [\tilde{z} \;\; \tilde{n}_{xz} \, \tilde{n}_{yz}]^\top$ as

$$E(\mathbf{H}, \tilde{\mathbf{H}}) = \sum_{p \in P} D_p(\mathbf{h}_p) + \sum_{p \in P} \sum_{q \in N_p} V_{p,q}(\mathbf{h}_p, \mathbf{h}_q) + \sum_{p \in P} T_p(\mathbf{h}_p, \tilde{\mathbf{h}}_p), \tag{5}$$

[1] Equations (2) and (3) are only used to show the relationship between the window and scene plane. Hence, the focal length and baseline are not required to perform our method.

where P is a set of pixels in the reference image, and N_p represents four connected neighborhoods of p. D_p is the data cost representing how given disparity and normal values are well fitted to reference and target images, I^{ref} and I^{tar}. It is defined as

$$D_p(\mathbf{h}_p) = \frac{1}{|W_p|} \sum_{q \in W_p} w_{pq} \cdot C(q, \mathbf{h}_p), \tag{6}$$

where W_p is a window centered at pixel p. w_{pq} is an adaptive support weight defined as $exp(-\left|I_p^{ref} - I_q^{ref}\right|_1 / \sigma_c)$, and $|W_p|$ is a normalization factor defined as $\sum w_{pq}$. $C(q, \mathbf{h}_p)$ is a pixel-based cost for the disparity given by Eq.(1). In this paper, we use the combination of census transform [21] and mutual information [11]. It is defined as

$$C(q, \mathbf{h}_p) = (1-\alpha) \cdot \min\left(C_{census}(q, \mathbf{h}_p), \tau_c\right) + \alpha \cdot \min\left(C_{mutual}(q, \mathbf{h}_p), \tau_m\right), \tag{7}$$

where α is a parameter for balancing two costs that have different characteristics, and τ_c and τ_m are truncation parameters for two costs. Census transform is effective at discriminating correct disparity values at textureless regions, while it is susceptible to noise. By contrast, mutual information provides robustness against noise, while it is inappropriate to obtain disparity values in ambiguous regions. Thus, the fused cost can improve accuracy and robustness by complementing weak points of each cost. In addition, because both costs are robust to global illumination variation, the fused cost is also robust against illumination variation.

$V_{p,q}$ is a smoothness cost that encourages a smooth disparity map. It is defined as

$$\begin{aligned} V_{p,q}(\mathbf{h}_p, \mathbf{h}_q) = \beta \cdot \min\Big(&\left| n_{xz,p} \cdot (x_p - x_q) + n_{yz,p} \cdot (y_p - y_q) + (z_p - z_q) \right| + \\ &\left| n_{xz,q} \cdot (x_p - x_q) + n_{yz,q} \cdot (y_p - y_q) + (z_p - z_q) \right|, \tau_v \Big), \end{aligned} \tag{8}$$

where β is a parameter for weighting the smoothness cost. τ_v is a truncation parameter for preserving depth discontinuities.

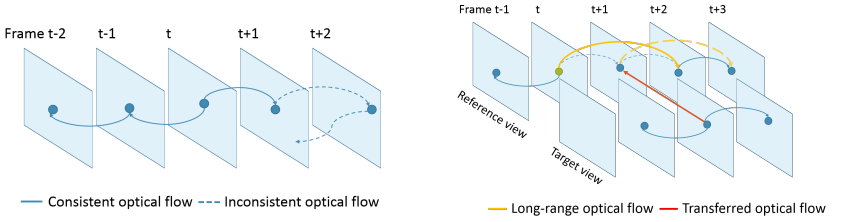
T_p is a temporal cost that encourages consistent disparity values between frames and is defined with guidance disparity and normal values, $\tilde{z}_p$, $\tilde{n}_{xz,p}$ and $\tilde{n}_{yz,p}$, as

$$T_p(\mathbf{h}_p, \tilde{\mathbf{h}}_p) = \delta_p \cdot \gamma \cdot \left[\min\left(\left|z_p - \tilde{z}_p\right|, \tau_z\right) + \min\left(\rho \cdot \left(\left|n_{xz,p} - \tilde{n}_{xz,p}\right| + \left|n_{yz,p} - \tilde{n}_{yz,p}\right|\right), \tau_n\right) \right]. \tag{9}$$

Ideally, $\tilde{z}_p$, $\tilde{n}_{xz,p}$ and $\tilde{n}_{yz,p}$ should be ground truth values. However, because ground truth values are not given, we determine guidance values by exploiting the characteristics of a 3D disparity profile — smoothly varying disparity and normal values between frames are more reliable than largely varying disparity and normal values. Hence, we set values that can make disparity and normal values between consecutive frames smoothly varied as guidance values. To obtain such guidance values, we compose a disparity profile and fit the profile to a polynomial function. This procedure will be described in Sec. 2.5. γ is a parameter for weighting the temporal cost, and ρ is a parameter for adjusting the influence of a normal difference compared to a disparity difference. Lastly, τ_z and τ_n are truncation parameters for preventing incorrect guidance values from overly affecting the energy function.

2.4 Temporal correspondence establishment

To obtain the 3D disparity profile, we must establish the temporal correspondences across frames. To this end, we employ the Large Displacement Optical Flow (LDOF) method [3]



(a) Recursive link of optical flows (b) Optical flow transfer

Figure 2: Establishment of temporal correspondences

to handle large displacements by rapidly moving objects or cameras. After forward and backward optical flows between consecutive frames are obtained, we can find temporal correspondences at distant frames by recursively linking optical flows as in Fig.2(a). By gathering disparity and normal values of temporal correspondences, we can compose a disparity profile. However, if the obtained flows are incorrect at some frames, the disparity profile may contain outliers. Hence, we filter out incorrect flows by a forward-backward optical flow consistency check with the threshold of 1.

However, the consistency check causes the profile to become fragmentary. For example, in Fig.2(b), the optical flow of the reference view is inconsistent between frame t and $t+1$ because of the error at frame $t+1$. Hence, temporal correspondences at frames from $t+1$ cannot support the pixel of interest at frame t. To handle this problem, the early work [9] used long-range optical flows to omit erroneous frames and employed optical flows from frame t to $t\pm1$ / $t\pm2$ / $t\pm3$. Figure 2(b) shows the advantage of this strategy. Because the yellow-colored flow from frame t to $t+2$ is consistent, correspondences at frames after $t+2$ can support the pixel of interest. However, the strategy is not useful when the frame of interest is frame $t+1$. In this case, the pixel of interest at frame $t+1$ cannot use information from other frames due to the absence of reliable flows by the problem of frame $t+1$.

To solve this problem, we propose an optical flow transfer that uses the flows of the other view instead of long-range flows. Given the forward flow from frame t to $t+1$ and disparity maps for frame t and $t+1$ of the target view, we can transfer the flow of pixel p in the target view to the reference view as

$$\overline{u}^{ref}_{t,t+1}(x_p - z^{tar}_{p,t}, y_p) = u^{tar}_{t,t+1}(x_p, y_p) + (z^{tar}_{p,t} - z^{tar}_{p',t+1}), \quad \overline{v}^{ref}_{t,t+1}(x_p - z^{tar}_{p,t}, y_p) = v^{tar}_{t,t+1}(x_p, y_p), \tag{10}$$

where x_p and y_p are x-y coordinates of pixel p in the target view. u and v are horizontal and vertical displacements of initial optical flow o. $\overline{u}$ and $\overline{v}$ are components of transferred flow $\overline{o}$. $z^{tar}_{p,t}$ and $z^{tar}_{p',t+1}$ are disparity values of pixel p at frame t and p' at frame $t+1$, and p' is the temporal correspondence of p. However, the transferred flow could be incorrect because of erroneous flow and disparity values. Thus, we generate the transferred flow using only reliable flow and disparity values. In order to confirm the reliability of a disparity, we use a left-right consistency check as in [6]. Here, we set thresholds for disparity and normal to 1 and 0.3, respectively. In addition, to further remove incorrect transferred flows, we again check the consistency between $o^{ref}_{t,t+1}$ and $\overline{o}^{ref}_{t+1,t}$ as well as that between $\overline{o}^{ref}_{t,t+1}$ and $o^{ref}_{t+1,t}$. Finally, we assign flows of $o^{ref}_{t,t+1}$ or $\overline{o}^{ref}_{t,t+1}$ passed by the consistency check to pixels having inconsistent initial flows. A backward flow is obtained in the same manner.

2.5 Energy minimization

We obtain the optimal solution to Eq.(5) in the PatchMatch Belief Propagation (PMBP) framework. PMBP starts with the initialization of particles using random sampling. Afterwards, in each iteration, particles are propagated from neighborhoods q to pixel p and are then refined. If the disbelief of a candidate state is lower than that of current states of p, the candidate state is selected as a new member of the particle set of p. After a few iterations, the particle with the minimum disbelief is selected as the optimal solution of p.

Our propagation step consists of spatial, view, and temporal propagation. Spatial propagation propagates particles from four spatial neighborhoods of pixel p. The propagation proceeds according to a scan-line order, and already updated particles are only propagated. In [2], authors composed an over-parameterized disparity as $[-n_{xz,q} \;\; -n_{yz,q} \;\; n_{xz,q}x_q + n_{yz,q}y_q + z_q]$. Because this form encapsulates x_q and y_q, it does not require any process for spatial propagation. However, our over-parameterized form uses disparity z_q without x_q and y_q. Thus, if we propagate our over-parameterized disparity, the disparity to propagate, z_q, could be inappropriate to the position of p. Therefore, we update the disparity of q before spatial propagation from pixel q to p as

$$z'_p = z_q + n_{xz,q}(x_q - x_p) + n_{yz,q}(y_q - y_p), \quad n'_{xz,p} = n_{xz,q}, \quad n'_{yz,p} = n_{yz,q}, \tag{11}$$

where z'_p, $n'_{xz,p}$ and $n'_{yz,p}$ are entries of the propagated state to pixel p. Next, we conduct view propagation that propagates particles from the target view. For pixel p in the reference view, if pixel q in the target view is matched by the state of q, the state is propagated to pixel p after transforming the state to be appropriate for the reference view as $z'_p = -z_q$, $n'_{xz,p} = -n_{xz,q}$, $n'_{yz,p} = -n_{yz,q}$. Finally, we conduct temporal propagation. To this end, we propagate particles by using guidance values from the 3D disparity profile obtained in the previous iteration as $z'_p = \tilde{z}_p$, $n'_{xz,p} = \tilde{n}_{xz,p}$, $n'_{yz,p} = \tilde{n}_{yz,p}$. After the propagation step, a random search around the state of p is performed for state refinement as in [2].

Before the next iteration, we update data with an intermediate disparity map sequence having the minimum energy in the current iteration. First, we update optical flows. Because the optical flow transfer uses disparity maps, newly updated disparity maps can improve the quality of transferred optical flows. Next, we update joint histograms for mutual information. Finally, we update guidance values by fitting the disparity profile from the intermediate disparity map sequence. In Eq.(4), the normal of a window is complicatedly changed by the location and normal changes of a corresponding scene plane. Thus, even if the scene plane location and normal are linearly changed, the normal of the window can be changed non-linearly. In addition, because the location of a scene plane can be changed by non-linear motion, disparity values in the disparity profile cannot be properly modeled by a linear function as in [9]. Therefore, we fit disparity and normal values in the profile by a second-order polynomial function. Objective functions for $\tilde{z}_p$, $\tilde{n}_{xz,p}$ and $\tilde{n}_{yz,p}$ of frame t are defined as

$$\begin{aligned}
&\sum_{-S<i<S} w_{z,p}^{t+i}\left(a_p^z \cdot i^2 + b_p^z \cdot i + c_p^z - \frac{1}{z_p^{t+i}}\right)^2, \\
&\sum_{-S<i<S} w_{n,p}^{t+i}\left(a_p^{n_{xz}} \cdot i^2 + b_p^{n_{xz}} \cdot i + c_p^{n_{xz}} - n_{xz,p}^{t+i}\right)^2, \quad \sum_{-S<i<S} w_{n,p}^{t+i}\left(a_p^{n_{yz}} \cdot i^2 + b_p^{n_{yz}} \cdot i + c_p^{n_{yz}} - n_{yz,p}^{t+i}\right)^2,
\end{aligned} \tag{12}$$

where z_p^{t+i}, $n_{xz,p}^{t+i}$ and $n_{yz,p}^{t+i}$ are disparity and normal values in the profile, and superscript $t+i$ represents the frame in which the values are obtained. S is a parameter that defines the number of support frames. w_p^{t+i} is a weight for frame $t+i$, and it is defined as $w_p^{t+i} =$

$exp(-|i|/\sigma_t) \cdot r_p^{t+i}$, where the first term on the right-hand side is a temporal proximity term that yields higher weights for frames nearer to frame t. Note that because normal variation is more complex than disparity variation, σ_t for normal is set to a value smaller than that of σ_t for disparity. The second term on the right-hand side represents the reliability of disparity and normal values, and reliability r_p^{t+i} is determined by using the left-right consistency check. If the consistency check is passed, we set $r_p^{t+i} = 1$. Otherwise, $r_p^{t+i} = 0$. Parameters of the polynomial function, a, b and c, are obtained by weighted least squares regression.

After fitting, we update guidance values as $\tilde{z}_p = 1/c_p^z$, $\tilde{n}_{xz,p} = c_p^{n_{xz}}$, $\tilde{n}_{yz,p} = c_p^{n_{yz}}$. To increase robustness, we use only guidance values for the temporal cost and temporal propagation in the next iteration if the sum of weights for disparity is greater than the threshold η. If we can use guidance values for the temporal cost, we set δ_p of Eq.(9) to 1. Otherwise, $\delta_p = 0$.

Excepting the first iteration, we proceed with the aforementioned process of PMBP on both views by turns, and update data for both views. In the first iteration, because disparity maps do not exist, reliable joint histograms for mutual information cannot be obtained. Thus, before the first iteration, we compute temporary disparity maps using the census transform cost through the single iteration of PMBP. Similarly, because disparity profiles do not exist, we do not use the temporal cost and temporal propagation in the first iteration. Because disparity maps for joint histograms are likely to contain incorrect disparity values if an input image sequence is noisy, we use the disparity maps only for computing the mutual information cost.

3 Experimental results

We evaluated our method using a synthetic dataset presented in [13] and a real dataset provided by KITTI [5]. Parameters of the proposed method were empirically selected as follows. For the data cost, we used a 21×21 window, and set parameters as $\{\sigma_c, \alpha, \tau_c, \tau_m\} = \{30,\ 0.3, 0.4, 0.4\}$. Census transform costs were computed by a 7×7 window, and normalized by the size of the transform window. For smoothness and temporal costs, we set $\{\beta, \tau_v, \gamma, \tau_z, \tau_n, \rho, S, \eta\} = \{0.003, 1, 0.05, 1, 1, 10, 5, 3\}$. In addition, σ_t for disparity and normal were set to 10 and 3, respectively. For the PMBP framework, we used three iterations and a single particle. Optical flows were obtained by the LDOF software provided by authors using basic settings.[2] For post-processing, we conducted a simple scan-line-based hole filling method followed by a left-right consistency check.

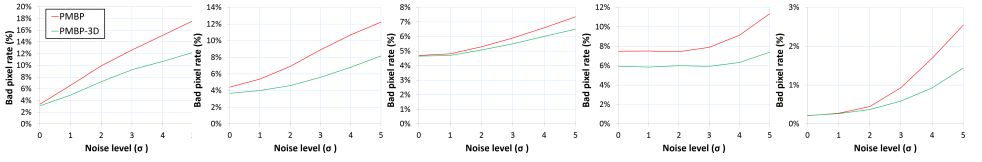
For a comparison, we selected the original PMBP-based method [1] as a baseline. The baseline method used the same settings with the proposed method. In this section, the baseline and proposed methods are referred to as PMBP and PMBP-3D, respectively.

First, we show experimental results for the synthetic dataset. To prove the validity of the proposed method, we analyzed the accuracy of disparity maps for image sequences with noise. We generated an image sequence with additive zero mean Gaussian noise, and measured an average bad-pixel rate of the computed disparity maps for each image sequence. In the experiments, the tolerance for discriminating bad-pixels was set to 1, and bad pixels were only checked in non-occluded regions. Table 1 shows average bad-pixel rates for image sequences with noise levels of 0 and 5. The results for image sequences without noise show that temporal information can help to improve accuracy for sequences containing ambiguous

[2] Basic settings of LDOF are $\{\sigma, \alpha, \beta, \gamma\} = \{0.8, 30, 300, 5\}$ according to the original paper [3].

Table 1: Average bad-pixel rate for synthetic datasets

	Noise level ($\sigma = 0$)					Noise level ($\sigma = 5$)				
	Book	Street	Tanks	Temple	Tunnel	Book	Street	Tanks	Temple	Tunnel
PMBP	3.34	4.41	4.70	7.48	0.21	17.65	12.22	7.35	11.35	2.54
PMBP-3D	3.08	3.66	4.65	5.91	0.21	12.22	8.15	6.50	7.40	1.44



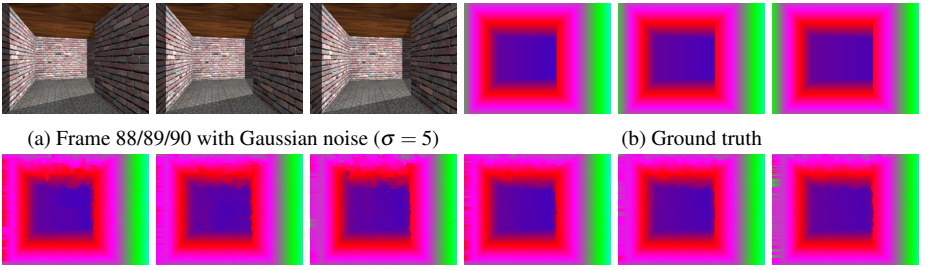
(a) Book (b) Street (c) Tanks (d) Temple (e) Tunnel

Figure 3: Average bad-pixel rates for synthetic data sets according to increasing noise level

regions. The results also show that our method does not yield performance degradation by employing temporal information. We also notice the accuracy improvement in the results for image sequences with noise when temporal information is employed. For all datasets, average bad-pixel rates of PMBP-3D are lower than those of PMBP. To see clearly the consistency improvement, we show average bad-pixel rates according to different noise levels in Fig.3. As the noise level increases, average bad-pixel rates of PMBP increase faster than those of PMBP-3D. This means that PMBP-3D can provide more consistent results despite noise than can PMBP. For a qualitative evaluation, we show the results of *temple* and *tunnel* datasets in Fig.1 and 4, respectively. We can see that PMBP generates inconsistent disparity maps between consecutive frames for noisy sequences. By contrast, our method provides more consistent disparity map sequences.

Next, we analyzed the results for a real dataset. Because the real dataset does not provide ground truth for color image sequences, we performed a qualitative evaluation. Figure 5 illustrates disparity maps for four successive frames. To clearly compare the performance of PMBP and PMBP-3D, we provide magnified results with contrast adjustment for a specific area of the disparity map. The specific area is marked with a yellow square in the disparity map. Because the image captured for a real scene contains ambiguous regions and inevitable noise, PMBP is likely to provide temporally inconsistent disparity maps. By contrast, PMBP-3D generates a consistent disparity map sequence. In addition to the qualitative evaluation, we quantitatively compared two methods using the KITTI multi-view dataset consisting of grayscale image sequences with ground truth. To this end, the first 30 sequences of training image pairs were used for computing the average bad-pixel rates with the error threshold of 3. Similar to the experiment with the synthetic dataset, we analyzed accuracy in according to the existence of additional noise. Without additional noise, PMBP and PMBP-3D had 5.80% and 5.50% average bad-pixel rates, respectively. With the noise level of 5, average bad-pixel rates of PMBP and PMBP-3D were 15.46% and 14.59%, respectively. These results represent that temporal information can help to improve disparity maps for the real dataset.

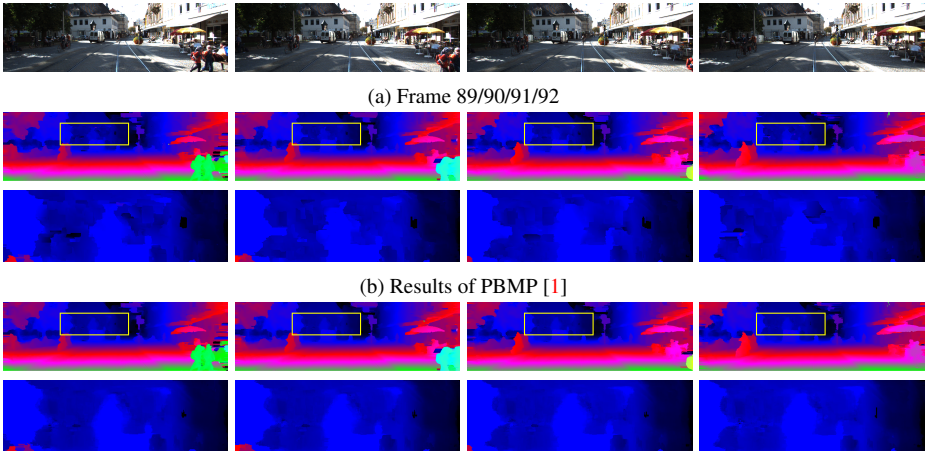
The run time of PMBP-3D depends on parameters. With the described parameters, the run time of PMBP-3D increases by 20% compared to PMBP for computing optical flows and updating data at each iteration.



(a) Frame 88/89/90 with Gaussian noise ($\sigma = 5$) (b) Ground truth

(c) Results of PMBP [1] (d) Results of PMBP-3D

Figure 4: Results for tunnel dataset



(a) Frame 89/90/91/92

(b) Results of PBMP [1]

(c) Results of PMBP-3D

Figure 5: Results for real dataset. A disparity map sequence is provided in the supplementary material.

4 Conclusion

In this paper, we presented a new PMBP-based framework using a 3D disparity profile for computing consistent disparity map sequences. Using the 3D disparity profile, we designed an energy function including temporal costs, and presented a temporal propagation for the PMBP framework. In addition, we introduced an optical flow transfer to obtain more reliable 3D disparity profiles. Our evaluation proves that the proposed method provides more accurate and consistent disparity maps than does the original PMBP-based method.

Acknowledgement

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2015R1A2A1A01005455) and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0101-15-0552, Development of Predictive Visual Intelligence Technology).

References

[1] Frederic Besse, Carsten Rother, Andrew Fitzgibbon, and Jan Kautz. Pmbp: Patchmatch belief propagation for correspondence field estimation. *International Journal of Computer Vision*, 110(1):2–13, 2014.

[2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo - stereo matching with slanted support windows. In *British Machine Vision Conference*, 2011.

[3] Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[4] James Davis, Diego Nehab, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime Stereo: A unifying framework for depth from triangulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2), 2005.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[6] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *International Conference on Computer Vision*, December 2013.

[7] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, and Margrit Gelautz. Temporally consistent disparity and optical flow via efficient spatio-temporal filtering. In *Pacific-Rim Symposium on Image and Video Technology*, 2011.

[8] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *International Conference on Computer Vision*, 2007.

[9] Chun Ho Hung, Li Xu, and Jiaya Jia. Consistent binocular depth and scene flow with chained temporal profiles. *International Journal of Computer Vision*, 102(1-3):271–292, 2013.

[10] Ramsin Khoshabeh, Stanley H. Chan, and Truong Q. Nguyen. Spatio-temporal consistency in video disparity estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[11] Junhwan Kim, Vladimir Kolmogorov, and Ramin Zabih. Visual correspondence using energy minimization and mutual information. In *International Conference on Computer Vision*, pages 1033–1040 vol.2, Oct 2003.

[12] E. Scott Larsen, Philippos Mordohai, Marc Pollefeys, and Henry Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *International Conference on Computer Vision*, 2007.

[13] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, and Neil A. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *European Conference on Computer Vision*, 2010.

[14] Jordi Sanchez-Riera, Jan Cech, and Radu P. Horaud. Robust spatiotemporal stereo for dynamic scenes. In *International Conference on Pattern Recognition*, 2012.

[15] Yongho Shin and Kuk-Jin Yoon. Spatiotemporal stereo matching for dynamic scenes with temporal disparity variation. In *International Conference on Image Processing*, 2013.

[16] Mikhail Sizintsev and Richard P. Wildes. Spatiotemporal stereo and scene flow via stequel matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1206–1219, 2012.

[17] Mikhail Sizintsev and Richard P. Wildes. Spacetime stereo and 3d flow via binocular spatiotemporal orientation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2241–2254, Nov 2014.

[18] Tatsunori Taniai, Yasuyuki Matsushita, and Takeshi Naemura. Graph cut based continuous stereo matching using locally shared labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1613–1620, June 2014.

[19] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *International Conference on Computer Vision*, pages 1377–1384, 2013.

[20] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650 – 656, april 2006.

[21] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision*, 1994.

[22] Li Zhang, Brian Curless, and Steven M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.