# DISPARITY ESTIMATION IN STEREO VIDEOS USING SPATIO-TEMPORAL DISPARITY HYPERPLANE MODELS

*Hiroki Nakano, Daisuke Sugimura, and Takayuki Hamamoto*

Graduate School of Engineering, Tokyo University of Science, 125-8585, Tokyo, Japan

## ABSTRACT

We propose a method for disparity estimation in stereo video. We address the problems associated with spatially-temporally-correlated disparity variations (STCDV). STCDV problems are caused by complex motions, e.g., yaw-rotation, pan-tilt-zoom camera movements, etc. The key novelty of this study is to introduce a spatio-temporal disparity hyperplane (STDH) model. The proposed STDH model represents a hyperplane defined in four-dimensional space spanned by disparity, image plane, and time coordinates. Our STDH model is represented by surface normals varying with the spatially-temporally-correlated changes in disparity. Thus, our STDH model is effective in estimating disparity in a stereo video including STCDVs. We estimate video disparity by incorporating our STDH model into the PatchMatch brief propagation framework. Our experiments demonstrate that the proposed method outperforms other methods.

***Index Terms*—** disparity estimation, stereo video, spatio-temporal hyperplane model

## 1. INTRODUCTION

Methods for stereo matching have been well studied [1, 2]. They were used to search on the correspondences based on similarities that are measured using a local support window. Methods using the slanted support window have shown their effectiveness in estimating the disparity [3–8]. Because the slanted support window allows to capture the spatial changes in the disparity in a local window, it enables disparity estimation to be performed with sub-pixel accuracy.

In recent years, techniques of disparity estimation using a stereo video have attracted considerable attention among researchers. However, there exist crucial problems to perform an accurate disparity estimation in a stereo video. Since the stereo video observes the movement of objects in a scene while the camera undergoes movement itself, the corresponding disparity varies temporally. Thus, techniques for compensating for the temporal disparity variation are required.

Researchers have proposed methods for estimating the video disparity [3, 9–19]. Methods [3, 9, 10] have achieved the performance enhancement of a video disparity estimation in a frame-by-frame manner by incorporating a temporal coherency between consecutive frames. However, these methods used motion vectors to obtain the temporal coherency; thus, an accurate optical flow estimation is essentially required for video disparity estimation.


**Fig. 1**. Spatially-temporally-correlated disparity variations (STCDV) problem. This stereo video simulates movements where the moving cameras turn right. This camera movement causes STCDVs. In this case, the local window (the red rectangular in the target image) is required to be deformed spatially temporally, to estimate video disparity accurately.

In contrast, the authors of [12–14, 16, 17, 19] have proposed spatio-temporal window matching to estimate the video disparity. Compared to methods employing a frame-by-frame manner, their approach does not need optical flow computation to estimate video disparity. Furthermore, noise tolerance in disparity estimation can be improved, as reported in [15]. Methods [12–14] have utilized an undeformable spatio-temporal window (i.e., the shape of local window is a cuboid in the spatio-temporal space) for searching on the correspondences between the left and right videos. However, these methods are hard to capture the temporal changes in disparity because the window used is difficult to be deformed spatially temporally. On the other hand, methods [16–19] have introduced spatio-temporal similarity measures to address the temporal disparity variations.

In general, however, complex movements, such as yaw rotations, simultaneous pan-tilt-zoom movements of cameras, etc., are likely to be observed in stereo videos. Figure 1 shows an example of stereo videos simulating movements that the moving stereo camera turns right. Such scenes will be observed in a real world (e.g., robots or drones move

around to search on objects or to reconstruct 3D scene). This camera movement contains the forward movements and the yaw-rotations. We consider that such movements cause spatially-temporally-correlated disparity variations (STCDV) in a stereo video. When STCDVs are observed, the local window requires to be deformed spatially temporally as shown in the red rectangular in the target image in Figure 1. In such scenes, previous methods [3, 9–19] would be unable to process effectively, because they have implicitly assumed that STCDVs would not be observed (e.g., see Eq. (2) in [17]).

To address STCDV problems that previous methods are hard to deal with, we propose a spatio-temporal disparity hyperplane (STDH) model for accurately estimating the video disparity. The STDH model is a hyperplane defined in a four-dimensional space that is spanned by disparity, image plane, and time coordinates. Specifically, our STDH model is characterized with surface normals varying with the spatially-temporally-correlated changes in disparity. Thus, it enables to deform the shape of the local window spatially temporally. With our STDH model, we perform disparity estimation by utilizing a PatchMatch berief propagation (PMBP) [8] extended with a spatio-temporal pairwise term.

## 2. PROPOSED STDH MODEL

The proposed STDH model is defined in the four-dimensional space spanned by disparity, an image plane, and time coordinates (we refer to this as the spatio-temporal disparity space). We derive our STDH model by first extending the original slanted window model [3] to a spatio-temporal window.

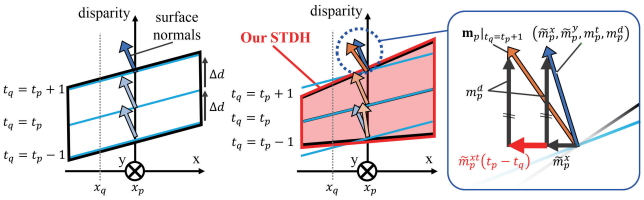
Let the position of the $p$-th point in this space be $\mathbf{x}_p = (x_p, y_p, t_p, d_p)$ where $x_p$ and $y_p$ are the pixel position on the image plane, $t_p$ is the time, and $d_p$ denotes the corresponding disparity. We define the four-dimensional hyperplane $h_p$ at $\mathbf{x}_p$ in the spatio-temporal disparity space with the surface normal $\mathbf{m}_p = (m_p^x, m_p^y, m_p^t, m_p^d)$. Assuming that the $q$-th point $\mathbf{x}_q = (x_q, y_q, t_q, d_q)$ is on the hyperplane $h_p$, the disparity value $d_q$ can be expressed as

$$d_q = \frac{m_p^x}{m_p^d}(x_p - x_q) + \frac{m_p^y}{m_p^d}(y_p - y_q) + \frac{m_p^t}{m_p^d}(t_p - t_q) + d_p \,. \quad (1)$$

When STCDVs are observed, we can consider that the spatial components of $\mathbf{m}_p$ ($m_p^x$, $m_p^y$, and $m_p^d$) would also vary temporally. Assuming that the temporal changes in them are small in the consecutive frames, the temporal changes in $m_p^d$ would be negligible, while those in $m_p^x$ and $m_p^y$ are approximated to a linear model. With these assumptions, we apply a Taylor series expansion to $m_p^x$ and $m_p^y$. We then obtain the approximated linear models as,

$$m_p^x = \tilde{m}_p^x + \tilde{m}_p^{xt}(t_p - t_q)\,, m_p^y = \tilde{m}_p^y + \tilde{m}_p^{yt}(t_p - t_q)\,, \quad (2)$$

where $\tilde{m}_p^x$ and $\tilde{m}_p^y$ are the components of the normal vector representing the spatial variations, respectively. Further, the second terms on the right-hand side in Eq. (2) represent the temporally-varying component of $m_p^x$ and $m_p^y$. According to theorem of Taylor series expansion, they are derived as the temporal derivatives at $t_p$, $\tilde{m}_p^{xt} = (\mathrm{d}m_p^x/\mathrm{d}t)|_{t=t_p}$ and $\tilde{m}_p^{yt} = (\mathrm{d}m_p^y/\mathrm{d}t)|_{t=t_p}$, respectively.




(a) a model in Eq. (1) (b) STDH model in the $x$-$d$ space

**Fig. 2**. Proposed STDH model. (a): An extension of [3] (Eq. (1)). This allows us to capture the temporal variation in disparity. (b): Our STDH model in the $x$-$d$ space. Based on the assumption that the temporal changes in the disparity between the consecutive frames would be small, $m_p^x$ can be represented as a linear model ($m_p^x = \tilde{m}_p^x + \tilde{m}_p^{xt}(t_p - t_q)$), while $m_p^d$ is approximately the same between the consecutive frames. Our STDH model can capture STCDVs in a stereo video.

We finally obtain our STDH model by plugging Eq. (2) in Eq. (1). Consequently, $d_q$ is given by

$$d_q = \frac{\tilde{m}_p^x}{m_p^d}(x_p - x_q) + \frac{\tilde{m}_p^y}{m_p^d}(y_p - y_q) + \frac{m_p^t}{m_p^d}(t_p - t_q) + \left(\frac{\tilde{m}_p^{xt}}{m_p^d}(x_p - x_q) + \frac{\tilde{m}_p^{yt}}{m_p^d}(y_p - y_q)\right)(t_p - t_q) + d_p \,. \quad (3)$$

In Eq. (3), the first and second terms correspond to the slanted window model [3], and the third term represents the temporal disparity variations. The fourth term (spatio-temporal cross-term) enables to capture the STCDVs. Figure 2 shows the above derivation for our STDH model.

## 3. DISPARITY ESTIMATION

We estimate the disparity sequence $\mathbf{h}$ parameterized by our STDH model. We define this as $\mathbf{h} = (\mathbf{h}^{(1)}, ..., \mathbf{h}^{(T)})$, where $\mathbf{h}^{(t)} = (\mathbf{h}_1^{(t)}, ..., \mathbf{h}_N^{(t)})$ ($N$ denotes the number of pixels in a frame, and $T$ is the number of frames). Each component $\mathbf{h}_u^{(t)}$ can be represented with our STDH model as $\mathbf{h}_u^{(t)} = (d_u, \tilde{m}_u^x, \tilde{m}_u^y, m_u^t, m_u^d, \tilde{m}_u^{xt}, \tilde{m}_u^{yt})$. We minimize the following energy function to estimate the latent disparity sequence,

$$E(\mathbf{h}) = \sum_{u \in V} \phi_u(\mathbf{h}_u) + \beta \sum_{u \in V} \left( \sum_{v \in N_{\mathrm{st}}(u)} \psi_{uv}(\mathbf{h}_u, \mathbf{h}_v) \right), \quad (4)$$

where $V$ represents a set of all the pixels in the disparity sequence to be estimated, and $N_{\mathrm{st}}(u)$ represents the

spatio-temporal neighboring pixels of $u$. In proposed method, we consider fourteen adjacent pixels in the spatio-temporal space. The weight $\beta$ is employed to balance the influences of the unary and pairwise terms.

The unary term $\phi_u(\mathbf{h}_u)$ is defined as

$$\phi_u(\mathbf{h}_u) = \sum_{k \in W(u)} w(u,k)\rho(k,k') \,, \tag{5}$$

where $W(u)$ is a spatio-temporal window centered at the $u$-th pixel, and $k$ is the pixel index in $W(u)$. Similar to [20], we utilize a spatio-temporal adaptive support weight $w(u,k)$ that is defined as $w(u,k) = \exp(-||\mathbf{I}_u - \mathbf{I}_k||_1/\gamma)$, where $\mathbf{I}_k$ is an RGB pixel value at the $k$-th pixel in $W(u)$, and $\gamma$ is a parameter. Furthermore, $\rho(k,k')$ is a cost function that is defined using the following two costs with a balancing parameter $\alpha$ as $\rho(k,k') = (1-\alpha)\min(||\mathbf{I}_k - \mathbf{I}_{k'}||_1, \tau_{\rm col}) + \alpha\min(||\nabla\mathbf{G}_k - \nabla\mathbf{G}_{k'}||_1, \tau_{\rm grad})$ , where $k'$ is the pixel index in the target image corresponding to $k$ in the reference image. It is determined by warping with the disparity $d_k$ given by Eq. (3). In computing $\rho(k,k')$, $\nabla$ is an operator for computing the spatial gradient along the horizontal direction, and $\mathbf{G}_k$ denotes an intensity value at the $k$-th pixel. In addition, $\tau_{\rm col}$ and $\tau_{\rm grad}$ denote truncation parameters.

The spatio-temporal pairwise term $\psi_{uv}(\mathbf{h}_u, \mathbf{h}_v)$ calculates the deviation between two STDHs at the $u$- and the $v$-th node. Because $\tilde{m}_p^{xt}$ and $\tilde{m}_p^{yt}$ are the temporal derivatives of $m_p^x$ and $m_p^y$, the deviation of them is measured separately from that of $\tilde{m}_p^x$, $\tilde{m}_p^y$, $m_p^t$ and $m_p^d$. Thus, we compute the deviation of STDHs with truncation parameters $\tau_{\rm c}$ and $\tau_{\rm td}$ as

$$\psi_{uv}(\mathbf{h}_u, \mathbf{h}_v) = [\min(dev_{\rm c}, \tau_{\rm c}) + \min(dev_{\rm td}, \tau_{\rm td})] \,, \tag{6}$$

with

$$dev_{\rm c} = |\mathbf{m}_u^{\rm sp} \cdot (\mathbf{x}_u^{\rm sp} - \mathbf{x}_v^{\rm sp})| + |\mathbf{m}_v^{\rm sp} \cdot (\mathbf{x}_v^{\rm sp} - \mathbf{x}_u^{\rm sp})| \,, \tag{7}$$

$$\begin{aligned} dev_{\rm td} &= |\left\{\mathbf{m}_u^{\rm td} \cdot (\mathbf{x}_u^{\rm td} - \mathbf{x}_v^{\rm td})\right\}(t_u - t_v)| \\ &\quad + |\left\{\mathbf{m}_v^{\rm td} \cdot (\mathbf{x}_v^{\rm td} - \mathbf{x}_u^{\rm td})\right\}(t_v - t_u)| \,, \end{aligned} \tag{8}$$

where we define $\mathbf{m}_u^{\rm sp} = (\tilde{m}_u^x, \tilde{m}_u^y, m_u^t, m_u^d)$ , $\mathbf{m}_u^{\rm td} = (\tilde{m}_u^{xt}, \tilde{m}_u^{yt})$, $\mathbf{x}_u^{\rm sp} = (x_u, y_u, t_u, d_u)$, and $\mathbf{x}_u^{\rm td} = (x_u, y_u)$.

We minimize Eq. (4) by utilizing a PMBP [8]. We then estimate $\mathbf{h}^*$.

Lastly, we perform post-processing to improve the estimated disparity sequence. To find outliers in the estimated disparity sequence, we first perform a left-right consistency checking as was done in [3, 8]. For the detected outlier regions, we interpolate the corresponding disparity values by using those that are consistently matched at the neighboring pixels. This interpolation is performed according to our STDH model (Eq. (3)).

Because of high computational cost in PMBP optimization, we consider a temporal subset in a stereo video as the target to be optimized. We compute the disparity for the entire stereo video by sliding the target subset frame by frame.

## 4. EXPERIMENTAL RESULTS

We present the experimental results for the video disparity estimation. We evaluated the performance of the proposed method quantitatively by analyzing the average bad pixel rate (0.5 sub-pixel accuracy) of the disparity estimation results for non-occluded regions. We empirically set the parameters as as follows. The number of iterations for PMBP optimization was set to 3, and the number of particles used for PMBP was set to 1. For the other parameters, we set $\{\gamma, \alpha, \tau_{\rm col}, \tau_{\rm grad}, \beta, \tau_{\rm c}, \tau_{\rm td}\} = \{10, 0.6, 10, 8, 0.57, 1.0, 1.0\}$. We evaluated the proposed method by comparing differences in the disparity models and the energy functions. Specifically, methods used in this comparison experiments are: (1) STDH + Eq. (4) (proposed method), (2) slanted window [3] + Eq. (4) (Baseline (1)), and (3) slanted window [3] + Eq. (1) in [8] (Baseline (2)). The proposed method enables to deform the shape of local window spatially temporally, while Baseline (1) (spatio-temporal matching) and (2) (frame-by-frame matching) are able to do spatially only. For each method, PMBP optimizer was used to estimate the latent disparity sequence.

### 4.1. Results for STCDV Sequences

In this experiment, we analyzed the accuracy of video disparity estimation in a stereo video where STCDVs are observed. We originally synthesized a stereo video (we named it "turnRight"). In this sequence, the moving stereo cameras turned right in an indoor scene. The size of spatio-temporal window $W(u)$ was set to $31 \times 31 \times 3$ for the proposed method and Baseline (1) ("3" means the number of frames), while $31 \times 31 \times 1$ was used for Baseline (2) (single frame).

Figure 3 shows comparison results using the sequence "turnRight". Further, the average bad pixel rate for each method was obtained as 5.03 (proposed method), 15.42 (Baseline (1)), and 5.74 (Baseline (2)), respectively. We observed that proposed method was able to estimate the video disparity more accurately.

### 4.2. Noise Tolerance

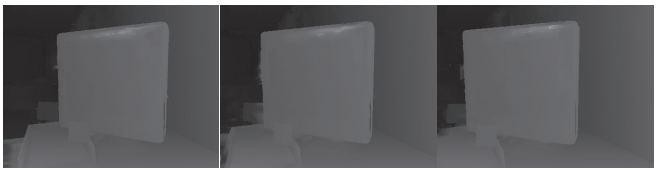
We also evaluated the ability of the proposed method to tolerate noise using the "turnRight", and the other sequences ("tunnel" and "temple") provided by [13]. For each image sequence, we added zero-mean Gaussian noise with a noise level $\sigma = 0, 3, 5, 7$. The size of spatio-temporal window $W(u)$ was set to $21 \times 21 \times 3$ for the proposed method and Baseline (1). For Baseline (2), we set the size of spatial window to $21 \times 21 \times 1$.

Table 1 lists the average bad-pixel rates. We can see that the results obtained using the proposed method showed the noise robustness against the other comparison methods.

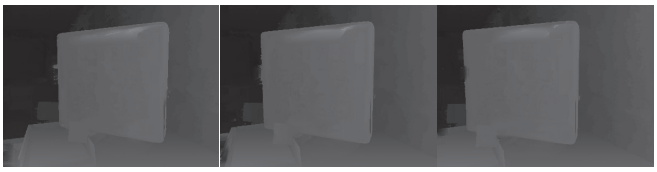
**Table 1**. Average bad pixel rate for image sequences ($\sigma = 0, 3, 5, 7$). The best scores are represented in bold.

| Seq. id | turnRight | | | | tunnel [13] | | | | temple [13] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0$ | $\sigma = 3$ | $\sigma = 5$ | $\sigma = 7$ | $\sigma = 0$ | $\sigma = 3$ | $\sigma = 5$ | $\sigma = 7$ | $\sigma = 0$ | $\sigma = 3$ | $\sigma = 5$ | $\sigma = 7$ |
| Proposed | **5.03** | **7.10** | **10.47** | **16.28** | 1.29 | **2.49** | **5.16** | **9.03** | 14.00 | 17.77 | **22.62** | 34.56 |
| Baseline (1) | 15.42 | 18.97 | 22.55 | 27.93 | 4.88 | 6.83 | 10.37 | 17.07 | **13.29** | **17.49** | 24.08 | **32.85** |
| Baseline (2) | 5.74 | 7.30 | 11.16 | 18.05 | **0.55** | 2.94 | 5.93 | 10.24 | 14.00 | 20.94 | 28.88 | 37.61 |

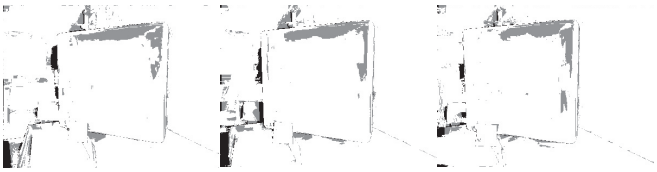
(a) Input sequence (frame no: 17,18,19)



(b) Estimated disparity sequence using proposed method



(c) Estimated disparity sequence using Baseline(1)



(d) Bad pixel sequence obtained using proposed method

(e) Bad pixel sequence obtained using Baseline(1)

**Fig. 3**. Disparity estimation results using sequence "turnRight". In Figs. 3-(d) and (e), the gray and the white pixels were classified as the wrong and the correct estimates with 0.5 sub-pixel accuracy, respectively. The black pixels represent occluded regions.

### 4.3. Results of 3D Scene Reconstruction for Real Scene

We tested proposed method using a real scene provided by KITTI dataset [21]. In this experiment, we compared 3D scene reconstruction results. They were obtained via triangulation using the disparity sequences that are estimated with proposed method and Baseline (1). Figure 4 shows 3D scene reconstruction results. We can qualitatively see that proposed method enabled to reconstruct 3D scene more accurately.

(a) Input sequence (frame no: 890, 891, 892)

(b) Estimated disparity at the frame 891
(left: proposed, right: Baseline(1))

(c) Close-up regions (yellow rectangular in (a) and (b))
of the reconstructed 3D scene
(left: proposed, right: Baseline(1))

**Fig. 4**. A comparison in 3D scene reconstruction in a sequence provided by KITTI dataset [21]. They were obtained using disparity estimated by proposed method (left image) and Baseline (1) method (right image).

## 5. CONCLUSION

We proposed a disparity estimation method for stereo video. We tackled a problem of spatially-temporally-correlated disparity variation (STCDV) that are caused by spatially-temporally-correlated motions. To address STCDV problems, we proposed a spatio-temporal disparity hyperplane (STDH) model. The STDH model is a hyperplane with the surface normals varying with the changes in the spatially-temporally-correlated disparity. We incorporated our STDH model into a framework of disparity estimation based on a PMBP. Through the experiments, we demonstrated the effectiveness of the proposed method by comparing its estimation accuracy with that of the other methods.

## 6. REFERENCES


[1] D. Marr and T. A. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.

[2] M. Bleyer and C. Breitender, *Stereo Matching - State-of-the-Art and Research Challenges*. Springer, 2013.

[3] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *Proc. of British Machine Vision Conference*, 2011, pp. 14.1–14.11.

[4] J. Lu, H. Yang, D. Min, and M. N. Do, "Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 1854–1861.

[5] S. Xu, F. Zhang, X. He, X. Shen, and X. Zhang, "Pm-pm: Patchmatch with potts model for object segmentation and stereo matching," *IEEE Trans. Image Processing*, vol. 24, no. 7, pp. 2182–2196, 2015.

[6] P. Heise, S. Klose, B. Jensen, and A. Knoll, "Pm-huber: Patchmatch with huber regularization for stereo matching," in *Proc. of IEEE Int. Conf. Computer Vision*, 2013, pp. 2360–2367.

[7] T. Taniai, Y. Matsushita, and T. Naemura, "Graph cut based continuous stereo matching using locally shared labels," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1613–1620.

[8] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "Pmbp: Patchmatch belief propagation for correspondence field estimation," *Int. Journal of Computer Vision*, vol. 110, no. 1, pp. 2–13, 2013.

[9] Z. Lee, J. Juang, and T. Q. Nguyen, "Local disparity estimation with three-moded cross census and advanced support weight," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1855–1864, 2013.

[10] Y. Shin and K.-J. Yoon, "Spatiotemporal stereo matching with 3d disparity profiles," in *Proc. of British Machine Vision Conference*, 2015, pp. 152.1–152.12.

[11] M. Bleyer and M. Gelautz, "Temporally consistent disparity maps from uncalibrated stereo videos," in *Proc. of Int. Symposium on Image and Signal Processing and Analysis*, 2009, pp. 383–387.

[12] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusimkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 296–302, 2005.

[13] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Proc. of European Conf. Computer Vision*, 2010, pp. 510–523.

[14] A. Hosni, C. Rhemann, M. Bleyer, and M. Gelautz, "Temporally consistent disparity and optical flow via efficient spatio-temporal filtering," in *Proc. of Pacific Rim Symposium on Image and Video Technology*, 2011, pp. 165–177.

[15] A. K. Jain and T. Q. Nguyen, "Discriminability limits in spatio-temporal stereo block matching," *IEEE Trans. Image Processing*, vol. 23, no. 5, pp. 2328–2342, 2014.

[16] J. Sanchez-Riera, J. Cech, and R. Horaud, "Robust spatiotemporal stereo for dynamic scenes," in *Proc. of Int. Conf. Pattern Recognition*, 2012, pp. 360–363.

[17] Y. Shin and K.-J. Yoon, "Spatiotemporal stereo matching for dynamic scenes with temporal disparity variation," in *Proc. of Int. Conf. Image Processing*, 2013, pp. 2242–2246.

[18] M. Sizintsev and R. P. Wildes, "Spatiotemporal stereo and scene flow via stequel matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1206–1219, 2012.

[19] L. Zhang, B. Curless, and S. M. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 367–374.

[20] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, 2006.

[21] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.