VARIATIONAL FUSION OF TIME-OF-FLIGHT AND STEREO DATA USING EDGE SELECTIVE JOINT FILTERING

Baoliang Chen, Cheolkon Jung, and Zhendong Zhang

School of Electronic Engineering, Xidian University, Xian 710071, China
zhengzk@xidian.edu.cn

ABSTRACT

In this paper, we propose variational fusion of time-of-flight (TOF) and stereo data using edge selective joint filtering (ESJF). We utilize ESJF to up-sample low-resolution (LR) depth captured by TOF camera and produce high-resolution (HR) depth maps with accurate edge information. First, we measure confidence of two sensor with different reliability to fuse them. Then, we up-sample TOF depth map using ESJF to generate discontinuity maps and protect edges in depth. Finally, we perform variational fusion of TOF and stereo depth data based on total variation (TV) guided by discontinuity maps. Experimental results show that the proposed method successfully produces HR depth maps and outperforms the-state-of-the-art ones in preserving edges and removing noise.

Index Terms— Edge selective, depth up-sampling, data fusion, stereo vision, time-of-flight, total variation.

1. INTRODUCTION

To produce high-quality depth maps, different approaches have been studied using stereo, TOF, Kinect, and light field cameras. They are classified into two groups: Passive and active. Passive stereo matching provides HR depth estimation while working well on texture scenes. However, it has a limit to estimate depth around occluded or smooth regions. Active devices such as TOF and Kinect cameras perform depth estimation independent of surface textures, but produce LR depth maps compared with HR color images. Thus, data characteristics of TOF and stereo are somewhat complementary, and it is required to fuse them for high-quality depth imaging. To successfully fuse both data, there exist two important issues [1]: 1) Confidence estimation for each device and 2) Data fusion. On the confidence estimation, [2] analyzed various error sources that influenced the performance of TOF cameras. [3] analyzed the effect of scene reflectance influences on the depth estimation. [4] proposed a measurement model of TOF camera errors by considering scene properties, i.e., depth discontinuity and scene reflectance. [5] provided a comprehensive review of stereo matching methods. [6] analyzed total 17 confidence measures for stereo matching based on window-based local stereo matching, and concluded that the attainable maximum likelihood (AML) achieved the best performance on Middlebury benchmark dataset[7]. On the data fusion, [8] provided a detailed review of them. [1] measured a reliable confidence to fuse both TOF and stereo data based on the mixed-pixel effect. [4] proposed a probabilistic framework based on local optimization. [9] used the mixed pixel effect based on global optimization in an MAP-MRF framework. [10] proposed an MAP-MRF framework based on Bayesian formulation with a belief propagation. [11] provided a temporal extension of [10]. Besides, [12] utilized a variational approach for depth fusion, but didn't measure the confidence of TOF depth data. It acquired edges only from an up-sampled gradient map, and thus edges are not preserved well.

In this paper, we propose variational fusion of TOF and stereo data using ESJF. For the confidence estimation, we adopt the adaptive support-weight approach [13] for stereo matching, and compute the confidence measure which combined AML with left-right consistency (LRC). To measure the confidence map for TOF camera, we utilize both a Gaussian noise model and depth discontinuity. For the data fusion, we perform ESJF to up-sample the initial depth map acquired from TOF camera. The up-sampled map has good edge information but is not accurate. Thus, we extract the horizontal and vertical discontinuity maps from the up-sampled depth map, and produce accurate depth maps from them based on TV. Experimental results show that the proposed method effectively preserves edges in HR depth maps while removing noise. Fig. 1 illustrates the flow diagram of the proposed method.

2. PROPOSED METHOD

2.1. Preprocessing

First, we convert the initial depth data D^{ini} from TOF camera to disparity map D^T as follows:

$$D^T = \frac{bf}{D^{ini}} \tag{1}$$

This work was supported by the National Natural Science Foundation of China (No. 61271298) and the International S&T Cooperation Program of China (No. 2014DFG12780).

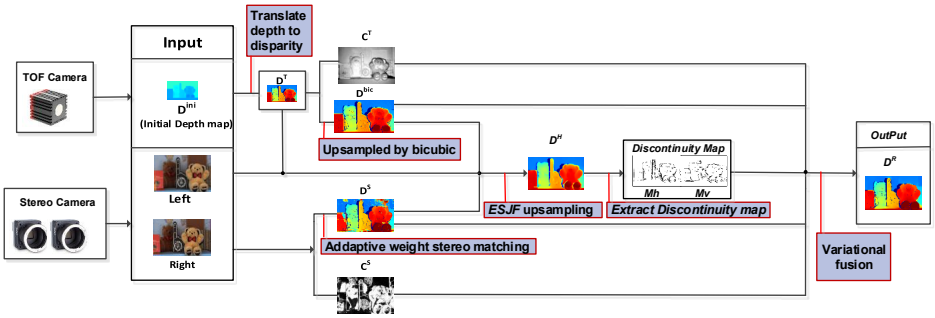

Fig. 1. Block diagram of the proposed method. D^T: Disparity map transformed from initial depth map of TOF camera D^{ini}. D^{bic}: Up-sampled disparity map from TOF camera using bicubic interpolation. C^T: Confidence map of D^{bic}. D^s: Stereo matching result. C^s: Confidence map of D^s. D^H: Up-sampled depth map by ESJF. M_h: Horizontal discontinuity map. M_v: Vertical discontinuity map.

where b is the baseline of the stereo system, and f is the focal length of the rectified stereo camera. When the scene area is flat, the distribution of TOF depth noise is represented by a Gaussian function with standard deviation σ_z as follows:

$$\sigma_z = \frac{c}{4\pi f_{\mathrm{mod}}} \frac{\sqrt{I/2}}{A} \tag{2}$$

The standard deviation σ_d is computed as follows [1]:

$$\sigma_d = bf \frac{\sigma_z}{z^2 - {\sigma_z}^2} \tag{3}$$

where f_{mod} is the infrared frequency, A is the amplitude value, I is the intensity, and c is the speed of light. When the scene area is discontinuous in depth, this model is not effective since its first order Taylor approximation is not correct. The depth values of foreground and background are mixed. Considering the two cases, the confidence for a pixel p is measured as follows:

$$\mathrm{C}^{\mathrm{D}^{\mathrm{ini}}}(p) = \exp[-\frac{\mathrm{var}(p)}{2\sigma_{\mathrm{v}}^2}] * \frac{\sigma_{\max} - \sigma_d}{\sigma_{\max} - \sigma_{\min}} \tag{4}$$

where $var(p)$ is the variance at p with a window size of 3×3; σ_v is a constant parameter; and $\sigma_{max}$is the maximum σ_d. Then, we project this LR depth map $\mathrm{C}_{\mathrm{D}^{\mathrm{ini}}}$ to its HR depth map (the same size as stereo matching result) C^T for depth fusion as follows:

$$C^T(p) = \begin{cases} \mathrm{C}_{\mathrm{D}^{\mathrm{ini}}}(p_\downarrow) & p_\downarrow \; exist \; in \; \mathrm{D}^{\mathrm{ini}} \\ 0 & others \end{cases} \tag{5}$$

where $p_\downarrow$ is the corresponding location of p in the LR depth map. Then, we use an adaptive support-weight approach for stereo matching to get the stereo matching result D^S. We consider both matching cost using attainable maximum likelihood (AML) and the result of left and right consistency (LRC). Even though depth is estimated in smooth regions, confidence of this disparity would be too small to be fused as

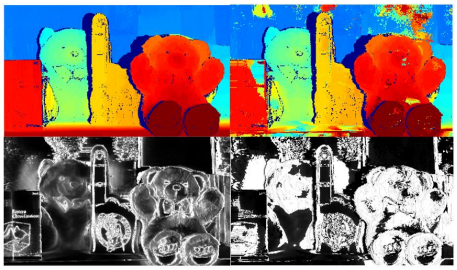

Fig. 2. Top: Ground truth depth map and stereo matching result. Bottom: C_{AML} calculated by AML and C^S calculated by both AML and LRC. Dark blue pixels are ignored in evaluation because occlusions happen or ground truth are not available.

shown in Fig. 2. Thus, we obtain the confidence map C^S by combining AML and LRC as follows:

$$C_{\mathrm{AML}} = \frac{1}{\sum_{\mathrm{d}} \mathrm{e}^{-\frac{(\mathrm{c(d)}-\mathrm{c_1})^2}{2\sigma_{\mathrm{AML}}^2}}} \tag{6}$$

$$C_{\mathrm{LRC}} = \mathrm{e}^{\frac{-|\mathrm{d_l}-\mathrm{d_r}|^2}{2\sigma_{\mathrm{LRC}}^2}} \tag{7}$$

$$C^S(p) = C_{\mathrm{AML}}(p) * \frac{\max\left(C_{\mathrm{LRC}}(p) - \gamma\,,\ 0\right)}{\max\left(C_{\mathrm{LRC}} - \gamma\right)} \tag{8}$$

where $c(d)$ is the matching cost value assigned to a disparity hypothesis d; c_1 is the minimum cost at a pixel; d_l and d_r are left and right stereo matching results, respectively; and γ, σ_{LRC}, and σ_{AML} are three parameters.

2.2. Depth Up-sampling by Edge Selective Joint Filtering

The goal of this step is to get an HR depth map D^H from D^{ini} and extract the horizontal and vertical edge maps. To up-sample D^{ini}, we use I, D^S, and D^{bic}. Here, three types of edges exist in them: 1) Edges exist in the depth map but don't exist in its color image; 2) Edges exist in the color image but its depth map is smooth; 3) Both depth map and color image

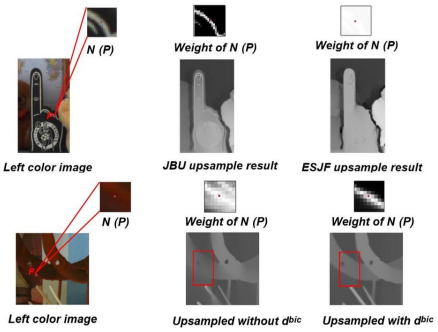

Fig. 3. Depth up-sampling results by ESJF compared with JBU[14] (Top) and without d^{bic} (Bottom)

have the same edges. Depth up-sampling should consider all three cases to successfully preserve edges in depth map. our *ESJF* which meets the requirement:

$$W_I = \mathrm{e}^{-\frac{\|\mathrm{I(p)}-\mathrm{I(q)}\|^2}{\sigma_{\mathrm{I(p)}}^2}} \tag{9}$$

$$W_{d^{bic}} = \mathrm{e}^{-\frac{\left\|\mathrm{d^{bic}(p)}-\mathrm{d^{bic}(q)}\right\|^2}{\sigma_{\mathrm{T}}^2}} \tag{10}$$

$$\mathrm{D^H}(p) = \frac{\sum_{\mathrm{q}\downarrow\in\mathrm{N(p)}} W_{d^{bic}} * W_I * \mathrm{d_T}(\mathrm{q}\downarrow)}{\sum_{\mathrm{q}\downarrow\in\mathrm{N(p)}} W_{d^{bic}} * W_I} \tag{11}$$

and:

$$\sigma_T(\mathrm{p}) = \mathrm{e}^{-\frac{\mathrm{var(d^{bic}(p))}}{2\sigma_{\mathrm{vt}}^2}} \tag{12}$$

$$\sigma_S(\mathrm{p}) = \mathrm{e}^{-\frac{\mathrm{var(d^S(p))}}{2\sigma_{\mathrm{vs}}^2}} \tag{13}$$

$$\sigma_{\mathrm{I}}(\mathrm{p}) = \max(\alpha * \max(\sigma_S(\mathrm{p}), \sigma_T(\mathrm{p})), Th) \tag{14}$$

where $I(p) = [r_p, g_p, b_p]^T$ is the color value at p; $d^{bic}(p)$ and $d^s(p)$ are the disparity values of p in D^{bic} and D^S, respectively; and q is a neighbor pixel of p; the window size is $S \times S$ where S has the same value as the scale on D^T; $var(d^{bic})$ and $var(d^s)$ are the depth variance of p in D^{bic} and D^S, respectively; Th is a threshold of σ_I to prevent too small value; σ_{vs} and σ_{vt} are scale parameters. ESJF has some advantages as follows: As shown in Fig. 3, texture copying artifacts often appear in the first case. This is becasue the texture in color image has different weights on p and its neighbor pixels. If their initial depths are noisy, texture copying artifacts occur. To eliminate the errors, we adjust σ_I with the assistance of D^S and D^{bic} using (13). In the first case, there exists no edge in D^S or D^{bic}, and the values of $var(d^{bic})$ and $var(d^s)$ are close to zero. Eventually, difference between weights of p and its neighbors decreases. We choose the maximum value between σ_S and σ_T because of avoiding large σ_I and false edges may appear by noise in D^{bic} or mismatching in D^S. For the second case, the D^{bic} term is necessary. If we use only color image for guidance, the difference between weights

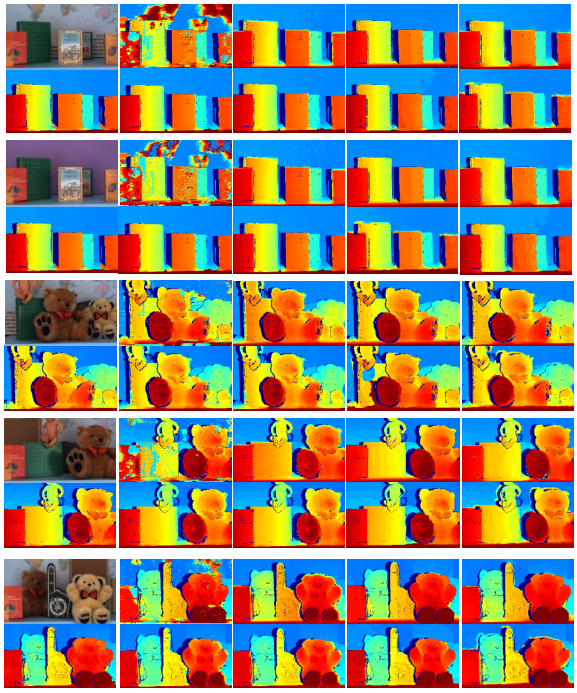

Fig. 4. Experimental results. Top: Color image, D^{bic}, D^S, ground truth, D^R (Proposed method). Bottom: [1], [15], [11], [9], and [16].

of p and its neighbors are reduced due to the color similarity. However, we add the edge information in depth map so that the pixels with different depth have different weights. Thus, the edge is not eliminated by the smooth color region. For the third case, the common edge regions means that p and its neighbor pixels are different in both color and depth, and thus the combination makes the weight in different regions have more diversity and the edge of D^H become more sharp. The up-sampled result is not sufficiently enough because this up-sampling is locally performed on a noisy LR depth map D^T. However, its edges are estimated accurately by the proposed ESJF. Thus, we extract discontinuity maps vertically and horizontally for variational fusion as follows:

$$\mathrm{M}_h(p) = \begin{cases} 0 & \left|\partial \mathrm{x}(\mathrm{D^H(p)})\right| > \mathrm{th\_h} \\ 1 & \text{others} \end{cases}$$
$$\mathrm{M}_v(p) = \begin{cases} 0 & \left|\partial \mathrm{y}(\mathrm{D^H(p)})\right| > \mathrm{th\_v} \\ 1 & \text{others} \end{cases} \tag{15}$$

where $\mathrm{th\_h}$ and $\mathrm{th\_v}$ are horizontal and vertical thresholds, respectively.

2.3. Variational Fusion

Finally, we perform variational fusion of data fidelity and regularization to get the final depth map D^R as follows:

$$\min_{d_R} \mathrm{E}(d_R) = E(d_T) + \lambda_1 E(d_S) + \lambda_2 R_{smooth} \tag{16}$$

Table 1. MSE evaluation results

Scene	1	2	3	4	5	Avg.
TOF	12.25	12.25	15.67	14.86	15.66	14.14
Stereo	46.13	32.30	8.92	16.32	12.61	23.26
Proposed	**6.29**	**6.35**	**4.62**	**4.30**	**3.83**	**5.01**
[1]	6.84	6.84	7.94	7.44	8.24	7.46
[9]	8.36	8.36	7.65	8.32	9.95	8.52
[16]	8.03	8.02	9.41	9.51	9.44	8.89
[15]	7.45	7.44	10.85	10.56	12.36	9.73
[11]	7.76	7.75	11.43	9.65	11.78	9.67

Table 2. SSIM evaluation results

Scene	1	2	3	4	5	Avg.
TOF	0.84	0.84	0.77	0.79	0.76	0.80
Stereo	0.54	0.49	0.80	0.70	0.80	0.67
Proposed	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**
[1]	0.91	0.91	0.87	0.89	0.89	0.89
[9]	0.92	0.91	0.89	0.91	0.91	0.91
[16]	0.90	0.89	0.88	0.87	0.89	0.89
[15]	0.90	0.90	0.89	0.88	0.88	0.89
[11]	0.92	0.92	0.91	0.93	0.92	0.92

$$\begin{aligned} E(d_T) &= \sum_p \left[C_T(p) * \left(\mathrm{d_R}(p) - \mathrm{d_T}(p)\right)^2) \right] \\ E(d_S) &= \sum_p \left[C_S(p) * \left(\mathrm{d_R}(p) - \mathrm{d_S}(p)\right)^2) \right] \end{aligned} \tag{17}$$

The confidence map in the data term makes the result close to TOF and stereo data at different levels. The regularization term based on total variation (TV) is weighted by horizontal and vertical discontinuity maps, which preserves edges and removes noise in depth map:

$$\mathrm{R_{smooth}} = \sum_p \left[\mathrm{M_h}(p) |\partial_{\mathrm{x}}(\mathrm{d_R(p)})|^2 + \mathrm{M_v}(p) |\partial_{\mathrm{y}}(\mathrm{d_R(p)})|^2 \right] \tag{18}$$

where λ_1 and λ_2 are fixed parameters to adjust data fidelity and smoothness, respectively. (16) is converted into a matrix form as follows:

$$\begin{aligned} \underset{\mathrm{D_R}}{Min} \ & \left[(\mathrm{D_R} - \mathrm{D_T})^T \mathrm{C_T} (\mathrm{D_R} - \mathrm{D_T}) \right] \\ & + \lambda_1 \left[(\mathrm{D_R} - \mathrm{D}_S)^T \mathrm{C}_S (\mathrm{D_R} - \mathrm{D}_S) \right] \\ & + \lambda_2 \left[(\partial_{\mathrm{x}} \mathrm{D_R^T} \mathrm{M_h} \partial_{\mathrm{x}} \mathrm{D_R}) + (\partial_{\mathrm{y}} \mathrm{D_R^T} \mathrm{M_v} \partial_{\mathrm{y}} \mathrm{D_R}) \right] \end{aligned} \tag{19}$$

Thus, we obtain $\mathrm{D_R}$ through a pseudo-inverse without iteration as follows:

$$\mathrm{D_R} = (\mathrm{C_T} + \lambda_2 \mathrm{C_S} + \lambda_2 \nabla_{\mathrm{h}} + \lambda_2 \nabla_{\mathrm{v}})^{-1} * (\mathrm{C_T D_T} + \lambda_2 \mathrm{C_S D_S}) \tag{20}$$

where ∇_h,∇_v are two derivative operators.

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, we use the dataset provided in [9]. This dataset contains 5 different scenes acquired by a Mesa SR4000 TOF camera and two Basler video cameras. The TOF sensor has a resolution of 176×144 pixels, while the color cameras have 1032×778 pixels. We set σ_v=15 in (4); γ=0.2 in (8); σ_{AML}=1.6, σ_{LRC}=2.5 in (6) and (7); σ_I=28, σ_T=15 in (10) and (11); and λ_1=5, λ_2=50 in (15) and (16). We set Th=5 in (14), and th_v=3, th_h=3 in (15). We fix them for all tests. In our experiments, the support window size is 30×30 for stereo matching. The proposed method is implemented on a PC with an Intel I7-6700 3.40 GHz CPU and 8 GB RAM using Matlab 2015b. We compare the performance of the proposed method with those of five state-of-the-art approaches: [1], [9], [16], [15], and [11]. All their results are obtained from `http:\\lttm.dei.unipd.it/paper_data/eccv16`. Also, we provide depth estimation results in Fig. 4. From the figure, it can be observed that stereo images have a lot of outliers in textureless regions, while the upsampled disparity maps from TOF camera contain more accurate disparities but are blurred by noise. In particular, the values in edges are mixed, which means a low confidence. In contrast, the proposed method produces more visual pleasing results than the others, and boundaries in depth maps are sharper along edges. In the regions where depth values are gradually changed, the result of the proposed method is more close to ground truth because the proposed method generates better depth values from stereo data. Tables 1 and 2 show quantitative evaluation results in terms of mean square error (MSE) and structural similarity (SSIM). Similar to [1], we only evaluate the measurements on valid pixels: Non-occluded pixels and the pixels with ground truth. It can observed from the tables that our RMSE and SSIM values are comparable with other methods. The average MSE is about 25% lower than [1], which is the best among them. This implies that the proposed method effectively removes noise. Moreover, the proposed methods achieves the best SSIM value because the proposed method produces accurate edge information in depth map.

4. CONCLUSIONS

In this paper, we have proposed variational fusion of TOF and stereo data using ESJF. To effectively fuse TOF and Stereo depth data, we have used confidence and discontinuity maps for data and smoothness terms in a TV framework. First, we have estimated confidence maps from TOF and depth data. Then, we have applied ESJF to the up-sampling of TOF depth. Next, we have obtained discontinuity maps from the up-sampled depth to get weights for TV. Finally, we have utilized TV regularization to fuse TOF and stereo depth data. Experimental results demonstrate that the proposed method effectively generates HR depth maps and outperforms the-stat-of-the-art ones in preserving edges and removing noise.

5. REFERENCES

[1] Giulio Marin, Pietro Zanuttigh, and Stefano Mattoccia, "Reliable fusion of tof and stereo depth driven by confidence measures," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 386–401.

[2] Timo Kahlmann and Hilmar Ingensand, "Calibration and development for increased accuracy of 3d range imaging cameras," *Journal of Applied Geodesy*, vol. 2, no. 1, pp. 1–11, 2008.

[3] Sigurjon Arni Gudmundsson, Henrik Aanaes, and Rasmus Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3d estimation," *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3-4, pp. 425–433, 2008.

[4] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M Cortelazzo, "A probabilistic approach to tof and stereo data fusion," *3DPVT, Paris, France*, vol. 2, 2010.

[5] Beau Tippetts, Dah Jye Lee, Kirt Lillywhite, and James Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *Journal of Real-Time Image Processing*, vol. 11, no. 1, pp. 5–25, 2016.

[6] Xiaoyan Hu and Philippos Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[7] D. Scharstein and R. Szeliski, "Middlebury stereo vision.," in *http://vision.middlebury.edu/stereo/*.

[8] Rahul Nair, Kai Ruhl, Frank Lenzen, Stephan Meister, Henrik Schäfer, Christoph S Garbe, Martin Eisemann, Marcus Magnor, and Daniel Kondermann, "A survey on time-of-flight stereo fusion," in *Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications*, pp. 105–127. Springer, 2013.

[9] Carlo Dal Mutto, Pietro Zanuttigh, and Guido Maria Cortelazzo, "Probabilistic tof and stereo data fusion based on mixed pixels measurement models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2260–2272, 2015.

[10] Jiejie Zhu, Liang Wang, Ruigang Yang, and James Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[11] Jiejie Zhu, Liang Wang, Jizhou Gao, and Ruigang Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 899–909, 2010.

[12] Rahul Nair, Frank Lenzen, Stephan Meister, Henrik Schäfer, Christoph Garbe, and Daniel Kondermann, "High accuracy tof and stereo sensor fusion at interactive rates," in *Proceedings of European Conference on Computer Vision*. Springer, 2012, pp. 1–11.

[13] Kuk-Jin Yoon and In-So Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 2, pp. 924–931.

[14] Kwok-Wai Hung and Wan-Chi Siu, "Improved image interpolation using bilateral filter for weighted least square estimation," in *Proceedings of IEEE International Conference on Image Processing*. IEEE, 2010, pp. 3297–3300.

[15] Qingxiong Yang, Kar-Han Tan, Bruce Culbertson, and John Apostolopoulos, "Fusion of active and passive sensors for fast 3d capture," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2010, pp. 69–74.

[16] CD Mutto, P Zanuttigh, S Mattoccia, and G Cortelazzo, "Locally consistent tof and stereo data fusion," in *Proceedings of ECCV Workshop on Consumer Depth Cameras for Computer Vision*, 2012, vol. 12, pp. 598–607.