SPATIO-TEMPORAL TOF DATA ENHANCEMENT BY FUSION

Frederic Garcia[⋆] *Djamila Aouada*[⋆] *Bruno Mirbach*[†] *Björn Ottersten*[⋆]

[⋆]SnT - Universtity of Luxembourg
{frederic.garcia, djamila.aouada, bjorn.ottersten}@uni.lu

[†]Advanced Engineering - IEE S.A.
bruno.mirbach@iee.lu

ABSTRACT

We propose an extension of our previous work on spatial domain Time-of-Flight (ToF) data enhancement to the temporal domain. Our goal is to generate enhanced depth maps at the same frame rate of the 2-D camera that, coupled with a ToF camera, constitutes a hybrid ToF multi-camera rig. To that end, we first estimate the motion between consecutive 2-D frames, and then use it to predict their corresponding depth maps. The enhanced depth maps result from the fusion between the recorded 2-D frames and the predicted depth maps by using our previous contribution on ToF data enhancement. The experimental results show that the proposed approach overcomes the ToF camera drawbacks; namely, low resolution in space and time and high level of noise within depth measurements, providing enhanced depth maps at video frame rate.

Index Terms— Time of Flight, spatio-temporal data enhancement, sensor fusion, multimodal sensors.

1 Introduction

With the ongoing progress in technology, new depth sensing devices based on the ToF principle are becoming available. In addition to being economic, compact, robust to illumination changes, and of low-weight, ToF cameras are able to provide full-scene depth information at a relatively high frame rate. However, the downside of this promising technology is the low resolution of the provided depth maps and the high contamination by noise in the distance measurements. Besides and despite being much faster than alternative depth sensing systems, their frame rate is still lower than the frame rate of standard 2-D video cameras. As a result, computer vision applications such as the identification of a moving object (or multiple objects) over time, may become intricate or even impossible. Therefore, in computer vision or robotic applications where the ToF camera limitations are critical, a very promising strategy is sensor fusion [1, 2]. In [3], we proposed the so-called Unified Multi-Lateral (UML) filter based upon the bilateral filter. The UML filter enhances the spatial resolution of the ToF data by considering the 2-D data recorded using a hybrid ToF multi-camera rig. In this paper, we propose to extend it from spatial to the spatio-temporal domain.

This work was supported by the National Research Fund, Luxembourg.

We focus on increasing the hybrid ToF multi-camera frame rate up to the same frame rate of the 2-D camera. To that end, we use the flow information between each 2-D frame to predict their corresponding depth maps, then apply the UML filter. The organization of the paper is as follows: In Section 2, we present the literature review on depth enhancement over time, and give the problem statement. Section 3 proposes our concept for spatio-temporal depth enhancement. In Section 4, we quantitatively and qualitatively evaluate our proposed solution. Finally, in Section 5, we give our conclusions and perspectives.

2 Related work and problem statement

Choi et al. [4] proposed to use another bilateral-based fusion filter [1] to tackle the spatial resolution problem for a given low-resolution depth map. Then, they proposed to interpolate depth maps according to their corresponding 2-D frames, as the frame rate of 2-D cameras is usually higher than that of the ToF camera. To that end, they used the motion given by a Full-search Block Matching Algorithm (FBMA) between the previous and the next 2-D frames. The final enhanced depth video is the result of filtering the interpolated depth maps and their corresponding 2-D frames. The same authors proposed in [5, 6] to reduce the temporal fluctuation problem by filtering where they start by simultaneously filtering several depth and 2-D image pairs in order to preserve depth consistency within static regions in the scene. Kim et al. [7] proposed to enhance the spatial resolution of a given depth map by minimizing the unmatched boundary problem between depth and 2-D image pairs using joint bilateral upsampling (JBU), in addition to a boundary refinement to reduce the edge blurring artifact by using linear interpolation with a color segment set. In addition, they minimized temporal depth flickering artifacts on stationary objects, *i.e.,* they preserved depth consistency using the motion between two consecutive frames.

Similarly to the aforementioned approaches, we assume the frame rate of the 2-D camera to be higher than that of the ToF camera. Our aim is to predict the missing low-resolution depth maps by using the motion between consecutive 2-D frames. The resulting enhanced depth video will result from filtering such predicted depth maps and their corresponding 2-D frames using the UML filter. For our purpose and to estimate the dense motion between 2-D video frames, we have

considered the work by Brox et al. [8] who proposed a high accuracy optical flow estimation based on an energy formulation. For two given consecutive frames $\mathbf{I}_i$ and $\mathbf{I}_{i-1}$, taken at times i and $i-1$, respectively, the gray value at a pixel position $\mathbf{p}_i = (u, v)^{\mathrm{T}}$ is assumed invariant to displacement, *i.e.,* $\mathbf{I}_i(\mathbf{p}_i) = \mathbf{I}_{i-1}(\mathbf{p}_i - \mathbf{w}_i)$ with $\mathbf{w}_i = (\tilde{u}, \tilde{v})^{\mathrm{T}}$ being the investigated displacement vector of the pixel $\mathbf{p}_i$ between the frames $\mathbf{I}_i$ and $\mathbf{I}_{i-1}$. In order to overcome the high sensitivity to slight changes in brightness from this first assumption, the gradient of a grey value image is considered to be invariant to displacement, *i.e.,* $\nabla\mathbf{I}_i(\mathbf{p}_i) = \nabla\mathbf{I}_{i-1}(\mathbf{p}_i - \mathbf{w}_i)$, where $\nabla = (\partial_u, \partial_v)^{\mathrm{T}}$ denotes the spatial gradient. The global deviations are minimized due to the gray and gradient constancy assumptions, and are measured by an energy function $E_{data}(\mathbf{w}_i)$. These two assumptions operate locally without considering neighboring pixels. A smoothness flow field $E_{smooth}(\mathbf{w}_i)$ is therefore introduced, and the total energy to minimize becomes

$$E(\mathbf{w}_i) = E_{data}(\mathbf{w}_i) + \alpha E_{smooth}(\mathbf{w}_i), \tag{1}$$

with $\alpha > 0$ being a regularization parameter. Finally, in the case of large pixel displacements between video frames, multi-scale ideas were considered; starting from a coarse, smoothed version of the problem and finishing with a multi-resolution strategy. In [8], Brox et al. proposed the estimation of a high accuracy optical flow by minimizing the non-linear energy function defined in (1). The resulting motion vector $\mathbf{w}_i$ accomplishes that $\mathbf{I}_i(\mathbf{p}_i) = \mathbf{I}_{i-1}(\mathbf{p}_{i-1})$ with

$$\mathbf{p}_{i-1} = \mathbf{p}_i - \mathbf{w}_i = g(\mathbf{p}_i). \tag{2}$$

We note that the function $g(\cdot)$ gives the flow between any pair of two consecutive frames. Then, no subscript is needed as it only depends on its argument. We also note that the subscript i in a pixel position $\mathbf{p}_i$ or motion vector $\mathbf{w}_i$ is to relate the frame $\mathbf{I}_i$ and not their pixel position within the image.
In [9], Sand et al. combined the minimization of Brox et al. with the regularization of the estimated flow proposed by Xiao et al. in [10]. In what follows, we have considered Sand et al's motion estimation algorithm and used its Matlab implementation provided by Chari [1]. We note that the better the motion estimation is, the more accurate will be the enhanced depth map.

We now investigate the problem of depth resolution enhancement over time. That is, we are in the case of a sequence of 2-D frames $\mathbf{I}_i$ taken at a frame rate $1/\tau_{\mathbf{I}}$, where the subscript $i \in \mathbb{N}$, indicates the i^{th} frame taken at time $(i \times \tau_{\mathbf{I}})$. We consider the corresponding sequence of ToF frames $\mathbf{D}_{n\kappa}$, $n \in \mathbb{N}$, taken at a frame rate of $1/\tau_{\mathbf{D}}$, such that the period $\tau_{\mathbf{D}}$ is multiple of $\tau_{\mathbf{I}}$, *i.e.,* $\tau_{\mathbf{D}} = \kappa \cdot \tau_{\mathbf{I}}$. Indeed, during a time period $\tau_{\mathbf{D}}$, the 2-D camera provides κ frames while the ToF camera provides a single one. We refer to the depth maps $\mathbf{D}_{n\kappa}$ as ToF keyframes and to their corresponding frame-synchronised 2-D images $\mathbf{I}_{n\kappa}$ as 2-D keyframes. We recall that our objective is to increase the hybrid ToF multi-camera rig resolution over time. To that end, we first estimate the motion vectors $\mathbf{w}_{n\kappa+i}$ between every consecutive 2-D frames $\mathbf{I}_{n\kappa+i}$ and $\mathbf{I}_{n\kappa+i+1}$, $0 \leq i < \kappa$. Then, we use the estimated motion vectors to predict the missing ToF frames between every consecutive ToF keyframes $\mathbf{D}_{n\kappa}$ and $\mathbf{D}_{(n+1)\kappa}$. For the sake of simplicity, we formulate our concept for the first period $\tau_{\mathbf{D}}$, *i.e.,* $n = 0$.

[1] http://www.mathworks.com/matlabcentral/fileexchange/17500

3 Proposed motion cumulation

The function $g(\cdot)$ introduced in (2) relates the pixel positions between two consecutive frames. However, in general we want to relate pixel positions between non-consecutive frames. Indeed, we want to relate the pixel position of $\mathbf{p}_i$ on the current image frame $\mathbf{I}_i$, $0 < i < \kappa$ with its corresponding pixel position on the keyframe $\mathbf{I}_0$. We therefore propose a *cumulative forward motion estimation* approach and define it as the cumulation of the estimated motion between each pair of 2-D frames starting from the current 2-D frame $\mathbf{I}_i$ until the 2-D keyframe $\mathbf{I}_0$. We prove by induction that

$$\mathbf{p}_0 = g^i(\mathbf{p}_i), \text{ where } g^i = \underbrace{g \circ ... \circ g}_{i \text{ times}}, \tag{3}$$

where $\circ$ is the combination of functions, and $i \in \mathbb{N}^*$ being the number of frames between the current frame $\mathbf{I}_i$ and the keyframe $\mathbf{I}_0$. From (2), we note that the case of $i = 1$ in (3) is true by definition. We then assume that (3) is correct and we show that $\mathbf{p}_0 = g^{i+1}(\mathbf{p}_{i+1})$ as $\mathbf{p}_0 = g^i\big(g(\mathbf{p}_{i+1})\big) = (g^i \circ g)(\mathbf{p}_{i+1}) = g^{i+1}(\mathbf{p}_{i+1})$. The predicted depth map $\acute{\mathbf{D}}_i$, where ‘´’ denotes forward-predicted frame, results from using the estimated cumulative forward motion between the current frame $\mathbf{I}_i$ and the keyframe $\mathbf{I}_0$, on the ToF keyframe $\mathbf{D}_0$, *i.e.,* $\acute{\mathbf{D}}_i(\mathbf{p}_i) = \mathbf{D}_0\big(g^i(\mathbf{p}_i)\big)$ for all pixel positions $\mathbf{p}_i$. The final enhanced depth video results from the fusion between the predicted depth frames $\acute{\mathbf{D}}_i$ and their corresponding 2-D frames $\mathbf{I}_i$ by using the UML filter, or an earlier version referred to as PWAS filter for Pixel Weighted Average Strategy [11], such that,

$$\acute{\mathbf{J}}_i(\mathbf{p}_i) = \frac{\sum_{\mathbf{q}_i \in N(\mathbf{p}_i)} f_{\mathbf{S}}(\mathbf{p}_i, \mathbf{q}_i) f_{\mathbf{I}}\big(\mathbf{I}_i(\mathbf{p}_i), \mathbf{I}_i(\mathbf{q}_i)\big) \mathbf{Q}_i(\mathbf{q}_i) \acute{\mathbf{D}}_i(\mathbf{q}_i)}{\sum_{\mathbf{q}_i \in N(\mathbf{p}_i)} f_{\mathbf{S}}(\mathbf{p}_i, \mathbf{q}_i) f_{\mathbf{I}}\big(\mathbf{I}_i(\mathbf{p}_i), \mathbf{I}_i(\mathbf{q}_i)\big) \mathbf{Q}_i(\mathbf{q}_i)}, \tag{4}$$

with $\mathbf{Q}_i = f_{\mathbf{Q}}(-|\nabla\acute{\mathbf{D}}_i|)$ being a credibility map that weights the reliability of each depth pixel and minimizes the unmatched boundary problem to cope with the edge blurring artifact. The weighting functions $f_{\mathbf{S}}(\cdot)$, $f_{\mathbf{I}}(\cdot)$, and $f_{\mathbf{Q}}(\cdot)$ are taken to be Gaussian functions with standard deviations $\sigma_{\mathbf{S}}$, $\sigma_{\mathbf{I}}$, and $\sigma_{\mathbf{Q}}$, respectively. We note that a stronger edge blurring artifact [11] appears within the enhanced depth maps $\acute{\mathbf{J}}_i$ that are closer in time to their next ToF keyframe $\mathbf{D}_\kappa$ than to their precedent ToF keyframe $\mathbf{D}_0$, from which they have been predicted (compare Fig. 1n and Fig. 1f). The reason is due to the large displacement in both time and space between the

frame $\mathbf{I}_i$ and its preceding keyframe $\mathbf{I}_0$. We therefore propose a *cumulative backward motion estimation* in which, in contrast to the *cumulative forward motion estimation* approach, the predicted depth maps result from the next ToF keyframe $\mathbf{D}_\kappa$. Thus, in this case, the estimated motion vector $\mathbf{w}_i$ verifies that $\mathbf{I}_i(\mathbf{p}_i) = \mathbf{I}_{i+1}(\mathbf{p}_{i+1})$ with $\mathbf{p}_{i+1} = \mathbf{p}_i + \mathbf{w}_i = h(\mathbf{p}_i)$. Similarly to the forward approach, we prove by induction that $\mathbf{p}_\kappa = h^i(\mathbf{p}_{\kappa-i})$. Thus, the predicted depth map $\grave{\mathbf{D}}_i$, where '`' denotes backward-predicted frame, follows as $\grave{\mathbf{D}}_i(\mathbf{p}_i) = \mathbf{D}_\kappa\big(h^i(\mathbf{p}_{\kappa-i})\big)$. Enhanced depth maps that result from considering forward-predicted depth maps $\acute{\mathbf{D}}_i$ are more accurate the closer they are to the precedent ToF keyframe $\mathbf{D}_0$. Instead, enhanced depth maps that result from considering backward-predicted depth maps $\grave{\mathbf{D}}_i$ are more accurate the closer they are to the next ToF keyframe $\mathbf{D}_\kappa$ (compare Fig. 1o with Fig. 1g). We therefore propose to linearly combine the forward-predicted and backward-predicted low-resolution depth frames, and define a *bidirectional motion estimation*, as follows $\hat{\mathbf{D}}_i = \frac{\kappa-i}{\kappa} \cdot \acute{\mathbf{D}}_i + \frac{i}{\kappa} \cdot \grave{\mathbf{D}}_i$, where '^' denotes bidirectionally predicted frame. Enhanced depth maps that result from considering bidirectional motion estimation are expected to present a major advantage of reducing the noise within depth measurements between consecutive ToF frames $\mathbf{D}_0$ and $\mathbf{D}_\kappa$ [6, 7]; hence, preserving depth consistency and reducing the temporal fluctuation problem. In addition, enhanced depth maps from such a combination are more accurate and less noisy than when considering depth maps resulting from a single directional motion estimation. It is, however, important to note that both backward and bidirectional approaches require the next ToF keyframe and thus impose a higher latency.

4 Experimental results

In the following, we present some experimental results computed on a real sequence of a hand moving through the scene. The sequence has been recorded using a hybrid ToF multi-camera rig that comprises a 3D MLI Sensor™ from IEE S.A. [2] and a Flea®2 video camera from Point Grey™ [3]. Both sensors are coupled with a narrow baseline of 36 mm. Also, they are calibrated for a perfect data alignment and frame-synchronised. Whereas the Flea®2 video camera provides (648×488) pixels, the 3D MLI Sensor™ provides a lower resolution of (56×61) pixels. In order to quantify our concept for depth video enhancement, we assume the frame rate of the 2-D camera to be four times higher than the frame rate of the ToF camera, *i.e.,* $\kappa = 4$. That is, three low-resolution depth maps are replaced every four 2-D frames by the predicted low-resolution depth maps. In order to quantify the performance of our proposed method we compute the peak signal-to-noise ratio (PSNR) as well as the structural similarity (SSIM) index between the enhanced depth maps

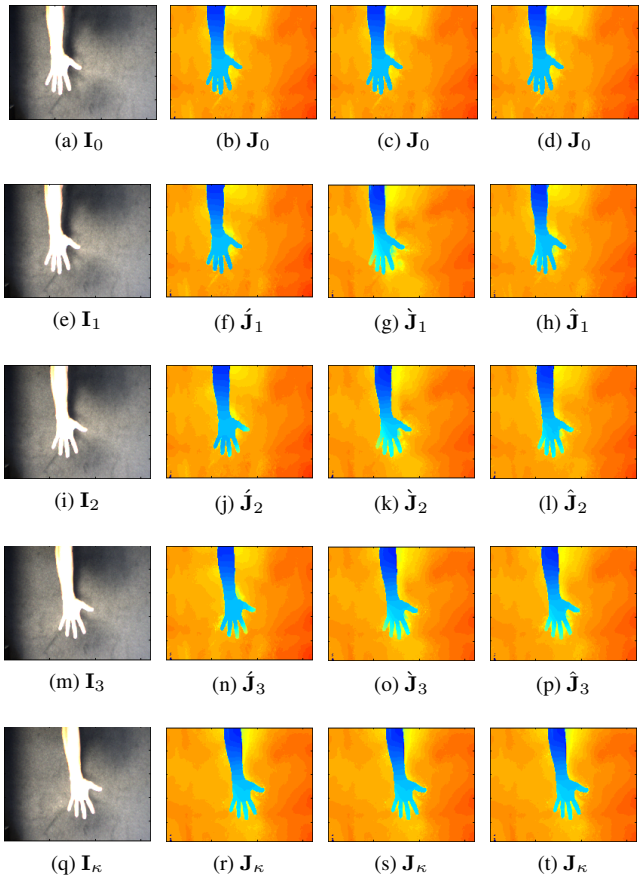


Fig. 1: Predicted and enhanced depth maps using forward, backward, and bidirectional motion estimation.

resulting from filtering using the predicted depth maps and the enhanced depth maps resulting from filtering using the neglected depth maps, *i.e.,* the ground truth. Fig. 1 shows an experiment where enhanced depth maps using the UML filter have been predicted from forward motion estimation (1^{st} column), backward motion estimation (2^{nd} column), and bidirectional motion estimation (3^{rd} column). It can be observed that forward-predicted depth maps are visually better the closer they are to $\mathbf{J}_0$, the enhanced depth map using the precedent ToF keyframe $\mathbf{D}_0$. Instead, the backward-predicted depth maps are better the closer they are to $\mathbf{J}_\kappa$, the enhanced depth map that results form the next ToF keyframe $\mathbf{D}_\kappa$. Thus, the combination of both strategies gives better results as reported in Table 1. Indeed, Table 1 quantifies the predicted enhanced depth maps with their corresponding ground truth, *i.e.,* the enhanced depth maps that result from filtering the pair of $\mathbf{I}_{n\kappa}$ and $\mathbf{D}_{n\kappa}$ given by the camera rig. From the table, we can observe that the predicted depth enhancement frames from bidirectional motion are more similar to the ground truth than considering either forward or backward motion.

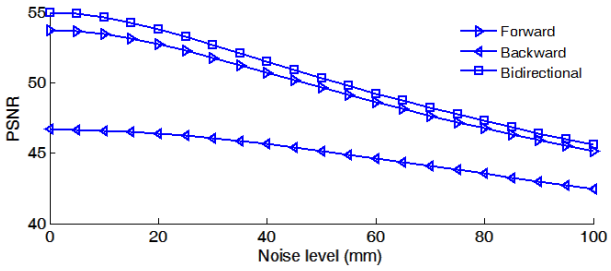
[2] IEE S.A., 3D MLI Sensor™, http://www.iee.lu

[3] Point Grey™, Flea®2, http://www.ptgrey.com/products/flea2/

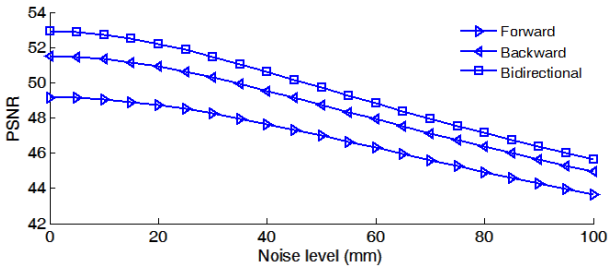
Table 1: Quantification of forward-predicted, backward-predicted and bidirectional-predicted enhanced depth maps.

	Forward		Backward		Bidirectional	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$\hat{\mathbf{J}}_1$	53.72	0.98	46.69	0.97	54.97	0.99
$\hat{\mathbf{J}}_2$	53.41	0.98	48.88	0.97	54.26	0.98
$\hat{\mathbf{J}}_3$	49.17	0.98	51.50	0.98	52.93	0.99

We next quantify the robustness to noise due to the active illumination of ToF cameras [3]. We simulate this behavior by adding Gaussian noise with a standard deviation linearly dependent on the distance measurement. In Fig. 2 we present the response of the proposed depth maps prediction strategies obtained over 50 Monte Carlo runs. The graphs confirm that the forward strategy performs better when predicting frames closer to the precedent keyframe while the backward strategy performs better the closer the predicted frame is to the next keyframe. In contrast, the bidirectional strategy outperforms in all cases. However, the runtime is doubled as both forward and backward motion estimation have to be computed and combined.



(a) $\acute{\mathbf{J}}_1$, $\grave{\mathbf{J}}_1$, and $\hat{\mathbf{J}}_1$.



(b) $\acute{\mathbf{J}}_3$, $\grave{\mathbf{J}}_3$, and $\hat{\mathbf{J}}_3$.

Fig. 2: Responses against to Gaussian noise using cumulative forward ($\triangleright$), backward ($\triangleleft$), and bidirectional ($\square$) motion estimation approaches.

5 Conclusions

We have proposed to extend the sensor fusion concept to the temporal domain. As a result, we enhance both in time and space the low-resolution depth maps delivered by the ToF camera up to the image resolution and frame rate of the coupled 2-D video camera in a hybrid ToF multi-camera rig. Fusion using the UML filter [3] ensures dense depth maps that present more accurate measurements where the depth discontinuities are well defined and adjusted to the 2-D guidance image; hence avoiding the unmatched boundary problem between depth and 2-D image pairs and consequently, reducing the edge blurring artifact within the enhanced depth map. As a consequence of being based upon a bilateral filter, the filtered depth measurements are smoothed. Therefore, the global noise level is significantly reduced. Furthermore, the bidirectional approach reduces the temporal fluctuation problem by preserving depth consistency between ToF keyframes.

6 References

[1] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *ECCVW'08*.

[2] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *SIGGRAPH '07*.

[3] F. Garcia, D. Aouada, B. Mirbach, T. Solignac, and B. Ottersten, "A New Multi-lateral Filter for Real-Time Depth Enhancement," in *AVSS'11*.

[4] J. Choi, D. Min, B. Ham, and K. Sohn, "Spatial and temporal up-conversion technique for depth video," in *ICIP'09*.

[5] J. Choi, D. Min, and K. Sohn, "2D-plus-depth based resolution and frame-rate up-conversion technique for depth video," *IEEE Transactions on Consumer Electronics*, 2010.

[6] J. Choi, D. Min, D. Kim, and K. Sohn, "3D JBU based depth video filtering for temporal fluctuation reduction," in *ICIP'10*.

[7] S. Y. Kim, J. H. Cho, A. Koschan, and M. A. Abidi, "Spatial and Temporal Enhancement of Depth Images Captured by a Time-of-Flight Depth Sensor," in *ICPR'10*.

[8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV'04*.

[9] P. Sand and S. Teller, "Particle Video: Long-Range Motion Estimation using Point Trajectories," in *CVPR'06*.

[10] J. Xiao, H. Cheng, H. Sawhney, C. Rao, M. Isnardi, and S. Corporation, "Bilateral filtering-based optical flow estimation with occlusion detection," in *ECCV'06*.

[11] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel Weighted Average Strategy for Depth Sensor Data Fusion," in *ICIP'10*.