PIXEL WEIGHTED AVERAGE STRATEGY FOR DEPTH SENSOR DATA FUSION

Frederic Garcia[*†] *Bruno Mirbach*[*] *Bjorn Ottersten*[†] *Frédéric Grandidier*[*] *Ángel Cuesta*[*]

[*]Advanced Engineering - IEE S.A., Luxembourg
[fga, bmi, fgr, acu]@iee.lu

[†]SnT - Universtity of Luxembourg
bjorn.ottersten@uni.lu

ABSTRACT

This paper introduces a new multi-lateral filter to fuse low-resolution depth maps with high-resolution images. The goal is to enhance the resolution of Time-of-Flight sensors and, at the same time, reduce the noise level in depth measurements. Our approach is based on the joint bilateral upsampling, extended by a new factor that considers the low reliability of depth measurements along the low-resolution depth map edges. Our experimental results show better performances than alternative depth enhancing data fusion techniques.

Index Terms— Machine vision, active vision, multisensor systems, image resolution, nonlinear filters.

1. INTRODUCTION

Time-of-Flight (ToF) sensors are a novel technology based on the ToF principle. A modulated near-infrared light is emitted by the sensor and simultaneously detected and demodulated by the entire sensor. The phase shift difference between the emitted and the received modulated light allows the sensor to compute the distance to the target. Moreover, ToF sensors avoid common artifacts in stereo vision setups such as occlusions or shadows [1]. Nevertheless, ToF sensors still present two main disadvantages: First, the resolution of their depth maps is far lower than the resolution of depth maps acquired with stereo techniques. Second, the depth measurements are strongly affected by noise.

Some attempts have been made to overcome these drawbacks by fusing ToF data with high-resolution 2D data. The application of Markov Random Fields (MRFs) to the problem of generating high-resolution depth maps from a low-resolution depth map and a high-resolution image was first presented by Diebel et al. [2], and extended by Gloud et al. [3]. Both methods are not suitable for real-time applications due to the computational requirements needed to solve the problem using MRF. In contrast, the problem of generating high-resolution depth maps may be tackled in real time using the bilateral filter [4]. Indeed, Kopf et al. [5] proposed a modified bilateral filter, called Joint Bilateral Upsampling (JBU), to upsample the low resolution depth maps by considering a high resolution guidance image taken from the same scene. Crabb et al. [6] implemented this alternative sensor fusion strategy in a real-time method for foreground/background segmentation of a colour video sequence. Yang et al. [7] presented another method that uses an iterative refinement module with bilateral filtering of the cost volume.

The referenced works share the same assumption when considering the information coming from the guidance image. They assume that depth discontinuities in a scene co-occur with colour or brightness changes within the associated high-resolution image, which is typically the case but not always justified. As a result, two main artifacts will appear on the fused data. The first one is *texture copying*. The textures from the guidance image are considered as edges that must be preserved according to the bilateral filter principle and, hence, they appear in the depth-enhanced map. The second artifact is *edge blurring*. It occurs when real depth discontinuities are not visible in the guidance image, that is, when targets at different depths share similar colours. To deal with these challenges, Chan et al. [8] proposed an adaptive multi-lateral upsampling filter (NAFDU) which is an extension of the JBU filter. It behaves in a different way depending on the pre-filtered data. However, in spite of the NAFDU promising results, two parameters remain to be tuned manually. Herein, we develop and analyse a novel extension of the JBU filter that addresses the two challenges commonly encountered in ToF depth acquisitions. Our contribution relies on a new factor that favours depth discontinuities over those in the guidance image. Thus, our multi-lateral filter is able to prevent texture copying and reduce edge blurring.

2. BILATERAL FILTER

The basis of our approach is the bilateral filter, whose output at each pixel is a weighted average of its neighbours; smoothing the image while preserving edges [9].

It analyses both the spatial domain S and the range domain R of an image. We denote by $I(\mathbf{x})$ and $I(\mathbf{y})$ the range image values at pixel positions $\mathbf{x}$ and $\mathbf{y}$, respectively. The filtered image J at $\mathbf{x}$ is:

$$J(\mathbf{x}) = \frac{\sum_{\mathbf{y}\in N(\mathbf{x})} f_S(\mathbf{x},\mathbf{y})\cdot f_R(I(\mathbf{x}),I(\mathbf{y}))\cdot I(\mathbf{y})}{\sum_{\mathbf{y}\in N(\mathbf{x})} f_S(\mathbf{x},\mathbf{y})\cdot f_R(I(\mathbf{x}),I(\mathbf{y}))} \qquad (1)$$

where $N(\mathbf{x})$ is the neighbourhood of $\mathbf{x}$. f_S and f_R are the spatial and range filter kernels, respectively. In [5], Kopf et al. suggested the JBU technique or cross/joint bilateral filtering that computes the range function based on another image D. The resulting filtered image J_D is defined at the position $\mathbf{x}$ as:

$$J_D(\mathbf{x}) = \frac{\sum_{\mathbf{y}\in N(\mathbf{x})} f_S(\mathbf{x},\mathbf{y})\cdot f_R(D(\mathbf{x}),D(\mathbf{y}))\cdot I(\mathbf{y})}{\sum_{\mathbf{y}\in N(\mathbf{x})} f_S(\mathbf{x},\mathbf{y})\cdot f_R(D(\mathbf{x}),D(\mathbf{y}))} \quad (2)$$

The JBU filter enforces the texture of the final image J_D to be similar to the texture of D. A possible application of the JBU technique is depth map enhancement by smoothing the low-resolution depth map while considering the edge information from a high resolution 2D image. The output is an enhanced depth map with much less discontinuities in their edges and a significantly reduced noise level. However, these enhanced depth maps present texture and blurring artifacts, as confirmed by our experiments (Fig. 2(d)). We therefore propose in the next section an extension of the JBU filter that strongly reduces such undesired behaviour.

3. PWAS: PIXEL WEIGHTED AVERAGE STRATEGY FOR DEPTH SENSOR DATA UPSAMPLING

The starting point of our method is, as in [8], the JBU filter [5]. We propose a new strategy for fusing low-resolution depth maps with high-resolution images in order to tackle the common artifacts encountered in data fusion. Our strategy is based on an additional factor to the kernels in (2), henceforth referred to as credibility map (CM).

A requirement for any low-level data fusion is that the filter input data must be perfectly aligned. In our case, we deal with the data matching through a mapping procedure that maps the data related to each sensor to a common reference frame [1].

Fig. 1 presents an overview of the framework of our method. The first step consists in mapping the low-resolution depth maps I', the high-resolution images D' and the image gradient of I' into a common reference frame where the entire data is pixel aligned, $I' \mapsto I$, $D' \mapsto D$ and $|\nabla I'| \mapsto C$. The low resolution of ToF sensors implies that one depth map pixel can represent several centimetres in the scene. As the depth measurement is inaccurate on edge pixels, the mapping procedure expands these pixels to stripes along edges where the depth measurement is inaccurate. Our concept is to define a credibility map that assigns to these pixels a lower weight in the filter kernel. Given the strength of the depth edge in terms of the absolute gradient of the low-resolution depth map $|\nabla I'|$, the application of the mapping procedure yields the upsampled depth edge strength C. The credibility map is then defined as a Gaussian kernel G_{σ_c} with variance σ_c^2, such that $CM(\mathbf{x}) = G_{\sigma_c}(C(\mathbf{x}))$. Similarly, we use Gaussian kernels for f_S and f_R with variances σ_s^2 and σ_r^2, respectively.

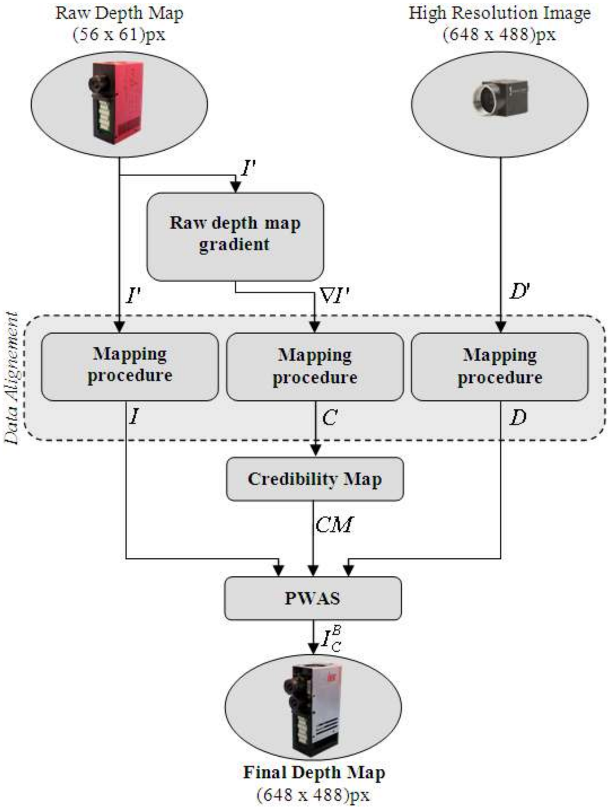


Fig. 1. Framework of our multi-lateral filter. The low-resolution depth map and the high-resolution image are mapped to a unified reference frame where the mapped images are pixel aligned. Both together with the already generated and mapped credibility map serve as the inputs for our multi-lateral filter.

By using our Pixel Weighted Average Strategy for depth sensor data fusion (PWAS), (2) takes the following form:

$$J_C(\mathbf{x}) = \frac{\sum\limits_{\mathbf{y}\in N(\mathbf{x})} G_{\sigma_s}(\|\mathbf{x}-\mathbf{y}\|)G_{\sigma_r}(|D(\mathbf{x})-D(\mathbf{y})|)\,CM(\mathbf{y})I(\mathbf{y})}{\sum\limits_{\mathbf{y}\in N(\mathbf{x})} G_{\sigma_s}(\|\mathbf{x}-\mathbf{y}\|)G_{\sigma_r}(|D(\mathbf{x})-D(\mathbf{y})|)\,CM(\mathbf{y})} \quad (3)$$

The standard deviations of the Gaussian kernels are related to the application and to the input raw data. The σ_s should be chosen greater than the upsampling rate used during the mapping procedure. The σ_r operates on the high-resolution guidance image D', being related to the edge amplitude, *i.e.*, the mean of the gradient along the edge. The σ_c behaves as the σ_r, operating on the low-resolution depth map I'.

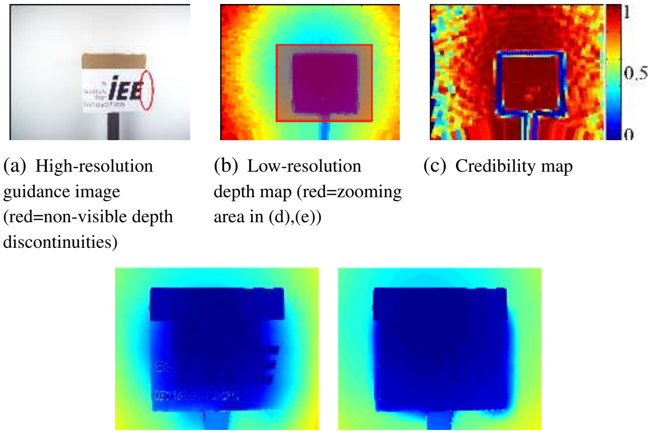
Note that the higher the credibility value, the greater the reliability over the depth measurement. A credibility value close to zero indicates that the corresponding range value is not reliable and thus not taken into account in the filter. As a result, the range value of the pixels in the image region with

[1] Work to be extensively reported in a different paper.

low credibility are instead replaced by an average over the neighbouring pixels. Thereby the weight is determined by the guidance image D', such that the depth edge will be sharpened by stretching the depth measurements until the guided position. Edge blurring only occurs in the case where a true depth discontinuity is not visible in the guidance image D', Fig. 2(a). Nevertheless, this drawback is restricted to the credibility map boundaries, performing better than previous sensor fusion approaches, Fig. 2(d).

4. EXPERIMENTAL SETUP AND RESULTS

The experimental setup used for raw data acquisition is a *ToF-based pair-sensor system*[2] (shown at the bottom of Fig. 1) that integrates a 3D MLI Sensor$^{\text{TM}}$from IEE S.A. [3] and a Flea®2 video camera from Point Grey$^{\text{TM}}$[4]. Both sensors are coupled for narrow baseline stereo vision. Also, they are frame-synchronised. Whereas the Flea®2 video camera provides (648×488) pixels, the 3D MLI Sensor$^{\text{TM}}$provides a lower resolution of (56×61) pixels.



(a) High-resolution guidance image (red=non-visible depth discontinuities)

(b) Low-resolution depth map (red=zooming area in (d),(e))

(c) Credibility map

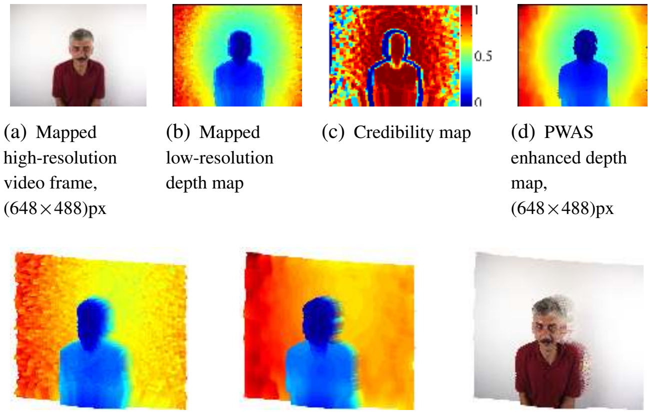
(d) JBU enhanced depth map (zoomed)

(e) PWAS enhanced depth map (zoomed)

Fig. 2. Visual comparison between our approach and JBU. (a), (b), (c) are the inputs for our approach while (e) is the depth-enhanced map. Note in (e) that we reduce texture copying and edge blurring artifacts (d).

Our approach reduces common output artifacts left by other fusion-based upsampling methods, Fig. 2. The credibility map utterly defines the depth map pixels with low reliability in their range values, Fig. 2(c) and Fig. 3(c). Thus, such pixels are not taken into account by the multi-lateral filter. As can be observed from Fig. 3(d), our filter presents a good performance width well defined edges, adjusted to the guidance image. In addition, when there is no contrast between foreground and background in the guidance image (depicted in red in Fig. 2(a)), the filter restricts the edges within the credibility map boundaries, Fig. 2(e).

As shown in Fig. 3, our method successfully increases the low-resolution depth maps, (56×61) pixels (Fig. 3(b)) to the (648×488) pixels of the guidance image resolution (Fig. 3(a)). Besides this considerable upsampling, the geometric detail provided by the guidance image is also preserved in the output depth maps.



(a) Mapped high-resolution video frame, (648×488)px

(b) Mapped low-resolution depth map

(c) Credibility map

(d) PWAS enhanced depth map, (648×488)px

(e) Low-resolution depth map

(f) PWAS enhanced depth map

(g) Textured enhanced depth map

Fig. 3. (a), (b), and (c) are the PWAS inputs whose result is shown in (d). Depth maps can be represented as a 3D geometry (e) and (f), that can also be textured by simply assigning the intensity value located in the same indices, as shown in (g).

We use the *Venus* scene provided by the Middlebury stereo dataset [5], Fig. 4(a) to compare our multi-lateral filter with alternative fusion-based upsampling methods. The Middlebury datasets provide intensity images together with its ground truth depth maps. We downsampled with a factor rate of 2, 4, and 8 the provided ground truth to be used as the low-resolution depth map input. The intensity images are directly used as high-resolution guidance images. After filtering, the root-mean-squared error (RMSE) between the processed depth maps and the provided ground truths is computed and presented in Table 1.

Table 1 shows that under global error measure our method performs better than the alternative selected fusion-based methods and more so when increasing the resolution of depth maps. Note that NAFDU only starts outperforming JBU at a high downsampling rate of 8, while our filter performs better in all cases. In addition, texture copying and edge

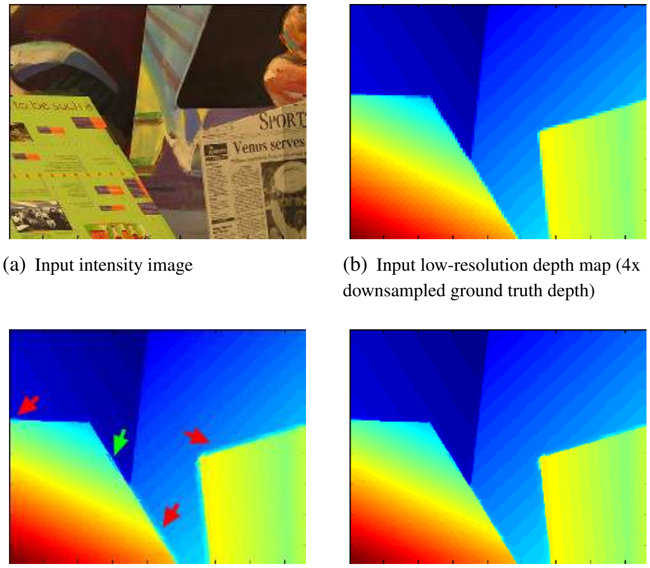
[2] We refer to such a ToF-based pair-sensor system to any multi-modal system that integrates a ToF camera and a 2-D camera.

[3] IEE S.A., 3D MLI Sensor$^{\text{TM}}$, http://www.iee.lu

[4] Point Grey$^{\text{TM}}$, Flea®2, http://www.ptgrey.com/products/flea2/

[5] Middlebury Stereo Dataset, http://vision.middlebury.edu/stereo/data

blurring (green and red arrows, respectively, in Fig. 4(c)) are significantly reduced, as depicted in Fig. 4(d).



(a) Input intensity image

(b) Input low-resolution depth map (4x downsampled ground truth depth)

(c) JBU 4x enhanced depth map

(d) PWAS 4x enhanced depth

Fig. 4. Although JBU and our method increase the low-resolution depth map to the input guidance image, in contrast to our result, the JBU performance exhibits slight texture copy and edge blurring in the areas marked with green and red arrows, respectively.

Table 1. RMSE quantitative comparison on depth-enhanced maps against MRF, NAFDU and JBU methods using the Venus scene from the Middlebury dataset. Note that values marked with '*' have been reproduced from [8].

Downsampled	Raw*	MRF*	NAFDU*	JBU	PWAS
2x	2	2.1	1.73	1.29	1.16
4x	2.97	2.28	2.18	2.10	1.61
8x	4.86	3.13	2.95	3.38	2.82

5. CONCLUSION

In this paper we have presented a new multi-lateral filter technique to fuse low-resolution depth maps with high-resolution colour images. We have extended the joint bilateral technique with an additional factor, the credibility map. As a result, we generated high resolution depth maps with more accurate depth measurements where the depth discontinuities are well defined and adjusted to the guidance image. Our experiments showed that our technique prevents texture copying and reduces edge blurring in the final depth-enhanced maps. Moreover, the results of an experimental comparison with the recent fusion-based approaches clearly denoted a better performance for our method.

Acknowledgements

Frederic Garcia would like to thank Dr. Djamila Aouada for fruitful discussions and critical review of the manuscript.

6. REFERENCES

[1] C. Matabosch Geronès, *Hand-held 3D-scanner for large surface registration*, Ph.D. thesis, University of Girona, 2007.

[2] J. Diebel and S. Thrun, "An Application of Markov Random Fields to Range Sensing," in *NIPS*. 2005, pp. 291–298, MIT Press.

[3] S. Gloud, P. Baumstarck, M. Quigley, Y. Ng Andrew, and K. Daphne, "Integrating Visual and Range Data for Robotic Object Detection," in *Workshop on M2SFA2, ECCV*, 2008.

[4] S. Paris and F. Durand, "A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach," in *International Journal of Computer Vision*. 2009, vol. 81, pp. 24–52, Kluwer Academic Publishers.

[5] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," in *SIGGRAPH*, New York, NY, USA, 2007, p. 96, ACM.

[6] R. Crabb, C. Tracey, A. Puranik, and J. Davis, "Real-time Foreground Segmentation via Range and Color Imaging," in *CVPRW*, 2008, pp. 1–5.

[7] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-Depth Super Resolution for Range Images," in *CVPR*, 2007, pp. 1–8.

[8] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A Noise-Aware Filter for Real-Time Depth Upsampling," in *Workshop on M2SFA2, ECCV*, 2008.

[9] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," in *ICCV*, 1998, pp. 839–846.