Depth Map Restoration and Upsampling for Kinect v2 Based on IR-Depth Consistency and Joint Adaptive Kernel Regression

C. Wang, Z. C. Lin and S. C. Chan

Department of Electronic and Electrical Engineering
The University of Hong Kong, Hong Kong
{cwang, zclin, scchan}@eee.hku.hk

***Abstract*—This paper presents a depth map restoration scheme for both the raw and projected depth map from Kinect v2 sensor. Based on IR-depth consistency, erroneous depth readings around foreground objects are removed by an edge aware consistency correction method. Moreover, a joint adaptive kernel regression algorithm is designed to upsample the sparse depth map after the projection from Kinect v2 sensor's depth camera to its full HD video camera. The structural information in the high resolution color image is implicitly utilized to guide the upsampling of depth map. The effectiveness of the proposed upsampling algorithm is illustrated by experimental results and comparisons on both real Kinect v2 data and Middlebury dataset.**

Keywords—Kinect v2 sensor; ToF; Kernel Regression

I. INTRODUCTION

Depth maps provide a totally new dimension on how machines sense the world besides texture and make computers one step closer to human beings. It has become an important ingredient in many real world applications under active research. Thus, increasing focus has been placed on acquiring depth maps using computer vision techniques [1] or infra-red (IR) based depth sensing devices [2, 3]. Computer vision techniques can provide high resolution depth maps, but their performances are limited in texture-less and occlusion areas. While on the other hand, conventional depth sensing devices are limited in terms of their resolution, range, and noise performance. With the introduction of Microsoft Kinect for Xbox360 [2] which is based on coded lighting using IR illumination, the resolution of depth map has significantly improved to 320x240 resolution with 11 bits/pixel compared to more expensive conventional sensors, say the SR4000 from Mesa imaging has a resolution of only 176x144. This has motivated a rapidly expanding research in using Kinect to assist various computer vision problems [4] as well as the restoration of its depth map [5].

An even more encouraging news is that Microsoft has launched a new generation of Kinect for Windows v2 sensor recently in 2014, generally knows as Kinect v2, with a full HD video camera and a depth sensor having a resolution of 521x424. Since ToF depth sensing is used, the occlusion area is significantly reduced and the depth values are more stable than the first generation Kinect. Kinect v2 also provides a wider field of view (70° horizontally and 60° vertically), a similar working range (0.5-8 meters) and a same frame rate (30 FPS) as its predecessor.

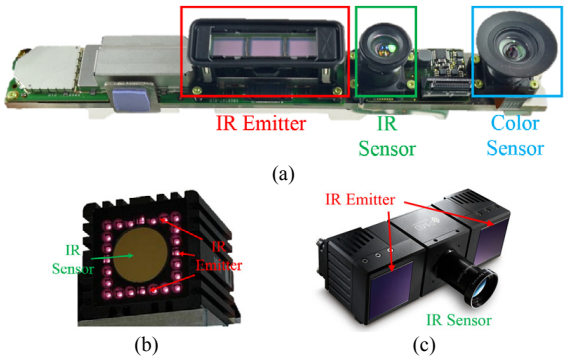
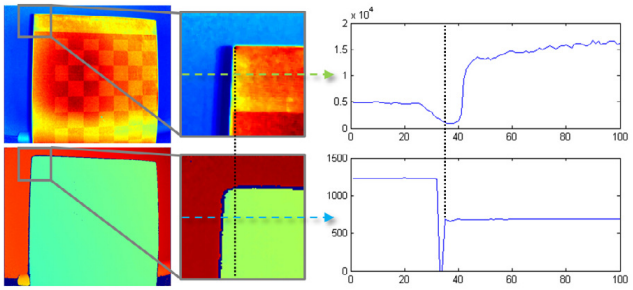


Fig. 1. Difference between Kinect v2 and other ToF cameras. (a) The inside structure of Kinect v2 (extracted from www.ifixit.com) (b) Mesa Imaging AG SwissRanger 4000. (c) PMD[Vision] CamCube 2.0.

However, Kinect v2, which is based on time-of-fight (ToF) depth sensing, shares the same imperfections as other ToF cameras which include depth distortion, IR amplitude related error, lighting scattering errors, etc. [3]. Hence, depth map restoration is still a crucial step for most depth based applications. In this paper, two key restoration problems, namely inconsistent depth reading removal and depth map upsampling, are considered for Kinect v2. To be more specific, an edge aware IR-depth consistency correction method is presented to remove the light scattering errors in raster scan order, while a joint adaptive kernel regression algorithm is proposed to upsample the projected sparse depth map by using both the depth map and high resolution color image.

The rest of the paper is organized as follows. We first introduce a raster-scan-based edge aware correction method to address Kinect v2 sensor's IR-depth inconsistency issue in Section II. Then a joint adaptive kernel regression algorithm is proposed to upsample the projected depth map. Experimental results on both real Kinect v2 data and Middlebury dataset will be given in Section IV. Finally, we conclude the paper in Section V.

This work was supported in part by the General Research Fund (GRF) of Hong Kong Research Grant Council (RGC).



(a) (b)

Fig. 2. IR-depth inconsistency in Kinect v2 due to occlusion. (a) IR image (upper) and corresponded depth map (lower). (b) 1D amplitude curves (green and blue dashed line in (a)) of the IR image (upper) and depth map (lower) across the boundary of the foreground board.

II. IR-DEPTH CONSISTENCY

Although the new Kinect v2 uses the same principle as the traditional ToF cameras, there exists a notable difference on the IR illumination scheme. Comparing to SwissRanger 4000 (Fig. 1(b)) and CamCube 2.0 (Fig. 1(c)), the IR emitter of Kinect v2 is located only on one side of the IR sensor as shown in Fig. 1(a). Such setting may introduce undesirable problems due to the occlusion caused by the displacement between IR emitter and IR sensor.

A. Inconsistent Depth Values

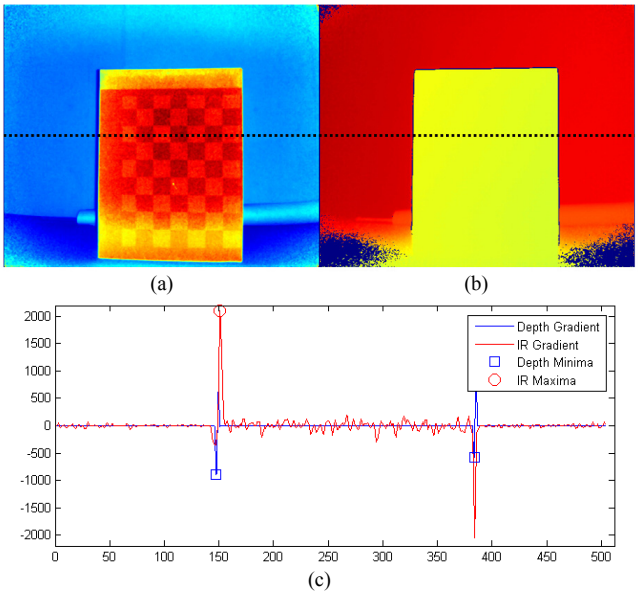
One major issue caused by this occlusion is the IR-depth inconsistency problem. Although the depth map is accurately aligned with the IR image, depth errors still occur around the foreground boundary as shown in Fig. 2(a). It can be seen that, the IR amplitudes are extremely low (dark blue) in the occlusion area, while it leads to erroneous depth value in this region. The 1D amplitude curves of the IR image and depth map given in Fig. 2(b) clearly show the problem of IR-depth inconsistency, i.e. foreground depth readings appear in the background region.

Since Kinect v2 is new on the market, there is little research mentioning this issue. This may be categorized as a kind of light scattering errors [3] occurred in the area with low IR amplitude. Multiple light reflections between the camera lens and its sensor also produces interference between nearby objects. In occlusion area where the IR amplitude is low, depth readings may be highly affected by the neighboring pixel reflections [3]. It thus explains why the occlusion region is mistakenly assigned the depth values of its nearby foreground objects.

B. Edge Aware IR-Depth Consistency Correction

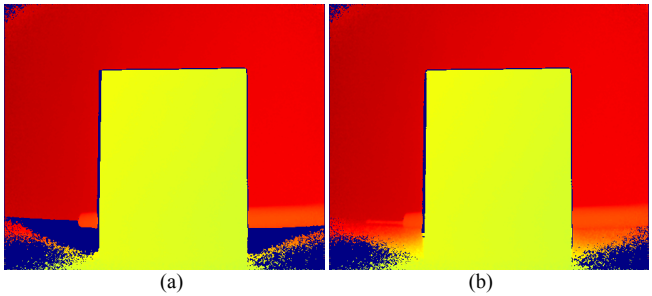
This kind of erroneous depth readings must be removed carefully, as it will result severe artifacts in various depth map based applications. A straightforward method is using low-amplitude filtering to remove the corrupted readings [3] directly. However, it also removes a lot of correct depth readings, such as the floor region shown in Fig. 4(a).

Considering the fact that the occlusion only occurs in horizontal direction, we propose an edge aware IR-depth consistency correction method to remove the IR-depth inconsistent error caused by scattering lights. From Fig. 2(a)



(a) (b) (c)

Fig. 3. Edge-aware IR-depth consistency correction. (a) and (b) are a pair of IR image and corresponded depth map, respectively. (c) 1D IR and Depth gradient curvs along the black dotted line in (a) and (b). Red circle and blue squares denote the maxima in IR and depth gradients, respectively.



(a) (b)

Fig. 4. Light scattering errors in Fig. 3(b) removed by (a) IR amplitude based thresholding and (b) the proposed edge aware IR-depth consistency correction method.

and (b), it can be seen that the errors start from the depth edge and end at the IR edge. Therefore, we try to locate these jumping points on the edges according to the horizontal gradients $g_{IR}(x,y)$ and $g_d(x,y)$ of the IR image and depth map respectively along the y-th scan line, where x is the horizontal index. Fig. 3(c) shows a typical example of IR gradient (red line) and depth gradient (blue line). Then, we search for distinct depth local minima and IR local maxima located at $(x_{d,\min}, y_{d,\min})$ and $(x_{IR,\max}, y_{IR,\max})$ as shown in the blue squares and red circles in Fig. 3(c), respectively. For each IR local maximum, if there is a depth local minimum satisfying $x_{d,\min} < x_{IR,\max}$, the interval $[x_{d,\min}, x_{IR,\max}]$ is denoted as the occlusion region. Hence the depth readings within this interval are considered as corrupted and set as 0 to remove the errors. The whole depth map $d(x,y)$ is processed line by line in raster scan order as follows

$$\hat{d}(x,y) = \begin{cases} 0, & x \in [x_{d,\min}, x_{IR,\max}] \\ d(x,y), & otherwise \end{cases}, y = 1,\ldots,M, \quad (1)$$

where $d(x,y)$ is the observed raw depth map and M is the height of depth map.

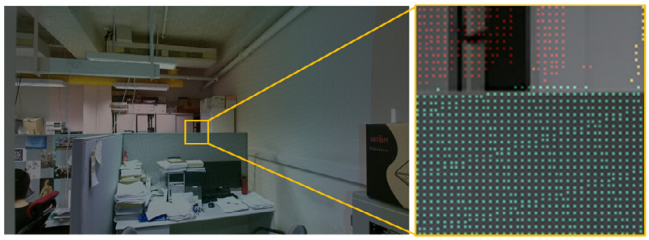


Fig. 5. Sparse depth map projected to the HD color camera in Kinect v2.

An example of the depth map after error removal by the proposed method is shown in Fig. 4(b). Comparing with low-amplitude filtering [3] (Fig. 4(a)), our method only remove those erroneous depth readings in the occlusion area, while the correct ones in low IR amplitude region remain unchanged.

III. Joint Adaptive Kernel Regression for Depth Map Upsampling

To utilize depth map in various applications, projection of the corrected raw depth map to different image space, such as external high resolution cameras, is usually performed. The Kinect SDK has provided a built-in function to map the depth map to its HD color camera. However, the projected depth map is quite sparse as shown in Fig. 5. It is mainly due to the difference in resolution between the depth map and color image. Therefore, the sparse depth map needs to be upsampled with the help of external high resolution cameras.

In previous works, steering kernel regression (SKR) [6] and local polynomial regression (LPR) [7, 8] have shown promising performance on both single image upsampling and video/multiview super-resolution. In this paper, we would like to extend the kernel regression (KR) framework to jointly utilize the depth map and high resolution.

A. Adpative Kernel Regression

In SKR and LPR, the color information at location $\boldsymbol{x}$ is modeled locally as a polynomial. Let y_j , $j=1,...,p$, be the observation within a small neighborhood $N(\mathbf{x}_i)$ of the location $\mathbf{x}_i$ of interest, i.e. $\mathbf{x}_j \in N(\mathbf{x}_i)$. Ideally, the neighborhood $N(\boldsymbol{x}_i)$ should be chosen so that the image can be approximated well by the given polynomial. It was shown in [6] that the local polynomial coefficients can be obtained by weighted least squares (WLS) fit of the given polynomial to the observations, which gives:

$$\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \| \mathbf{W}_{K_i}(\mathbf{y}_i - \mathbf{A}_i\mathbf{b}_i) \|_2^2 , \tag{2}$$

where $\boldsymbol{b}_i$ is the vector of polynomial coefficients to be determined, $\boldsymbol{y}_i = [y_1,...,y_p]^T$ is the vector of observations at $\mathbf{x}_i$,

$$\mathbf{A}_i = \begin{bmatrix} 1 & (\mathbf{x}_1-\mathbf{x}_i)^{\mathrm{T}} & vech^{\mathrm{T}}\{(\mathbf{x}_1-\mathbf{x}_i)(\mathbf{x}_1-\mathbf{x}_i)^{\mathrm{T}}\} & \cdots \\ 1 & (\mathbf{x}_2-\mathbf{x}_i)^{\mathrm{T}} & vech^{\mathrm{T}}\{(\mathbf{x}_2-\mathbf{x}_i)(\mathbf{x}_2-\mathbf{x}_i)^{\mathrm{T}}\} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (\mathbf{x}_p-\mathbf{x}_i)^{\mathrm{T}} & vech^{\mathrm{T}}\{(\mathbf{x}_p-\mathbf{x}_i)(\mathbf{x}_p-\mathbf{x}_i)^{\mathrm{T}}\} & \cdots \end{bmatrix}, \tag{3}$$

$\mathbf{W}_{K_i} = diag[K_h(\mathbf{x}_1-\mathbf{x}_i), K_h(\mathbf{x}_2-\mathbf{x}_i), \cdots, K_h(\mathbf{x}_p-\mathbf{x}_i)]$ is the weight matrix defined by a kernel function $K_h(u) = \frac{1}{h}K(\frac{u}{h})$ which puts more emphasis to observations near $\mathbf{x}_i$, $K(u)$ is the kernel function, h is the bandwidth of the kernel which controls the size of the neighborhood, and $vech\{\cdot\}$ is the half-vectorization operator. The first element of $\hat{\boldsymbol{b}}_i$ is the smoothed pixel value of at $\boldsymbol{x}_i$ which is given by

$$\hat{z}(\mathbf{x}_i) = \mathbf{e}_1^T(\mathbf{A}_i^T\mathbf{W}_{K_i}\mathbf{A}_i)^{-1}\mathbf{A}_i^T\mathbf{W}_{K_i}\mathbf{y}_i , \tag{4}$$

where $\mathbf{e}_1 = [1,0,\cdots,0]^T$. Note that $\boldsymbol{x}_i$ can be any location in the image and hence the image can be interpolated to a higher resolution. Since the local structure is usually anisotropic, the following locally adaptive kernel has been proposed in [6], which can better adapt to the local gradient of the image:

$$K_h(\mathbf{x}_n - \mathbf{x}_i) = \frac{\sqrt{\det(\mathbf{C}_n)}}{2\pi h^2}\exp(-\frac{1}{2h^2}\| \mathbf{C}_n^{1/2}(\mathbf{x}_n - \mathbf{x}_i) \|_2^2) , \tag{5}$$

where $\boldsymbol{C}_n$ is the inverse of the covariance matrix estimated from the gradients in $N(\mathbf{x}_n)$ and $\mathbf{x}_n \in N(\mathbf{x}_i)$.

One of the key parameters in KR is the bandwidth h of the adaptive kernel in (5), which is estimated from the local image model. In [8], the intersection of confidence interval (ICI) rule was proposed to select the appropriate bandwidth h. The estimated pixel by this bandwidth adaptive KR is therefore derived from (3) as

$$\hat{z}(\mathbf{x}_i) = \mathbf{e}_1^{\mathrm{T}}(\mathbf{A}_i^T\mathbf{W}_{K_i}\mathbf{A}_i)^{-1}\mathbf{A}_i^T\mathbf{W}_{K_i}\mathbf{y}_i \,|_{h_i} , \tag{6}$$

where the suboptimal bandwidth h_i is selected for each $\mathbf{x}_i$.

B. Joint Depth Map Upsampling

In our case, depth map upsampling aims to estimate a dense depth map from the sparse samples. The sparse depth map can be modeled as images blurred and down-sampled from a dense image, i.e.

$$Y = DHI + N = DZ + N , \tag{7}$$

where I denotes a high-resolution (HR) image, Y denotes the projected sparse observations, D and H are the corresponding down-sampling and blurring operators respectively, N is the noise term, Z is the blurred HR image which is the target of the estimation. Combining (2) and (7), we can get the following formulation for the depth map upsampling problem

$$\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \| \mathbf{W}_{K_i}D^T(\mathbf{y}_i - D\mathbf{A}_i\mathbf{b}_i) \|_2^2 , \tag{8}$$

where D^T denotes the upsampling operator with zero padding. Therefore, (6) can be reformulated as

$$\hat{z}(\mathbf{x}_i) = \mathbf{e}_1^{\mathrm{T}}(\mathbf{A}_i^TD^TD\mathbf{W}_{K_i}D^TD\mathbf{A}_i)^{-1} \mathbf{A}_i^TD^TD\mathbf{W}_{K_i}D^T\mathbf{y}_i \,|_{h_i} . \tag{9}$$

It should be noted that the depth map is viewed as a single image in (6), which means the information of the HD color camera will not be used. To utilize both the depth map and the high resolution color image, we select the weight matrix in (9) according to the color image to perform a joint adaptive kernel regression. In other words, two adaptive kernels $K_h^d(\mathbf{x}_n - \mathbf{x}_i)$ and $K_h^c(\mathbf{x}_n - \mathbf{x}_i)$ are estimated on depth map and color image, respectively. Then the color information is incorporated in the

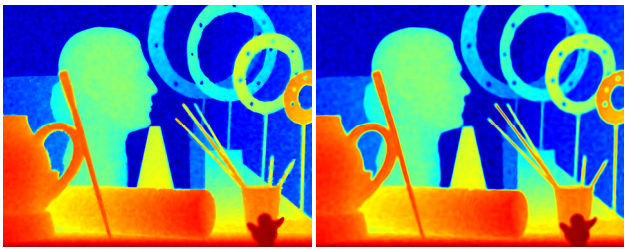


Fig. 7. Visual comparison of 4x up-sampling on the Middlebury Art dataset. Left and right are recovered by JBU [12] and proposed joint adaptive kernel regression, respectively.

TABLE. I **RMSE** (PIXELS) RESULTS ON THE MIDDLEBURY DATASET FOR TWO DIFFERENT MAGNIFICATION FACTORS (2X, 4X).

Algorithms	**Moebius**		**Books**	
	2x	***4x***	***2x***	***4x***
Guided [10]	2.4806	2.8315	2.3748	2.7369
MRFs [11]	2.1271	3.1054	2.0642	3.0017
JBU [12]	1.9970	2.5773	2.1093	2.6380
Proposed	**1.9903**	**2.5591**	**1.9444**	**2.5828**

depth upsampling process as

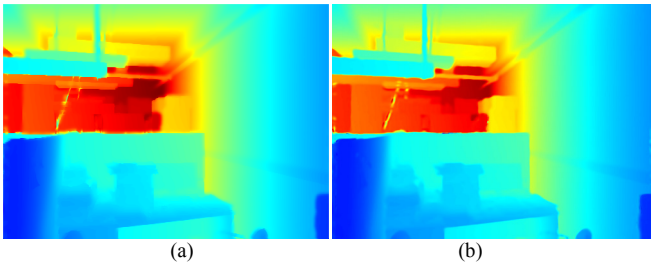
$$\hat{z}(\mathbf{x}_i) = \mathbf{e}_1^{\mathrm{T}} (\mathbf{A}_i^T D^T D \mathbf{W}_{K_i^d} \mathbf{W}_{K_i^c} D^T D \mathbf{A}_i)^{-1} \mathbf{A}_i^T D^T D \mathbf{W}_{K_i^d} \mathbf{W}_{K_i^c} D^T \mathbf{y}_i |_{h_i}, \quad (10)$$

where $\mathbf{W}_{K_i^c} = diag[K_h^c(\mathbf{x}_1 - \mathbf{x}_i), K_h^c(\mathbf{x}_2 - \mathbf{x}_i), \cdots, K_h^c(\mathbf{x}_p - \mathbf{x}_i)]$ and $\mathbf{W}_{K_i^d} = diag[K_h^d(\mathbf{x}_1 - \mathbf{x}_i), K_h^d(\mathbf{x}_2 - \mathbf{x}_i), \cdots, K_h^d(\mathbf{x}_p - \mathbf{x}_i)]$ are the weight matrices on color and depth samples, respectively. Since color weight matrix $\mathbf{W}_{K_i^c}$ is estimated on the high resolution color image, the color structure information is implicitly utilized in the process of depth map upsampling. The effectiveness of the additional weight matrix will be demonstrated in the next section.

IV. EXPERIMENTAL RESULTS

We first evaluate the proposed upsampling algorithm using the images from the Middlebury dataset [9]. The disparity images are used as the ground truth and the original RGB images are utilized to calculate the color weight matrix in (10). To simulate real acquisition process, we use the test dataset modified by [13], which contain disparity dependent noise. Using these datasets, we compare our results with the Markov Random Field (MRF) based approach [10], the guided image filtering approach [11] and joint bilateral upsampling (JBU) [12]. The numerical comparison in terms of the root mean squared error (RMSE) is shown in Table I. The proposed algorithm outperforms the compared methods.

A visual comparison on the Middlebury Art dataset is shown in Fig. 7, while another visual comparison on the real Kinect v2 data is presented in Fig. 8. It can be seen that the proposed algorithm produces more smooth results due to the nature of kernel regression. Moreover, it can also preserve the detailed structures, such as the cables in Fig. 8, with the help of the corresponded high resolution color image.



(a) (b)

Fig. 8. Visual evaluation on real Kinect v2 data. (a) JBU [12]. (b) Proposed joint adaptive kernel regression.

V. CONCLUSIONS

Solutions to two key restoration problems related to the recently launched Kinect v2 sensor have been presented. The erroneous depth readings caused by low IR amplitude and light scattering are removed by the proposed edge aware IR-depth consistency correction method based on IR-depth consistency. Corrected depth map is further upsampled to HD resolution. A new joint adaptive kernel regression algorithm is proposed to utilize the implicit structural information in the HD color image to assist the process of depth map upsampling. The effectiveness of the proposed algorithm is illustrated by experimental results on both the Middlebury dataset and real Kinect v2 data.

REFERENCES

[1] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Second Edition, Cambridge University Press, March 2004.

[2] Microsoft Kinect for windows sensor [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/.

[3] S. Foix, G. Alenya and C. Torras, "Lock-in Time-of-Flight (ToF) Cameras: A Survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917-1926, Sep. 2011.

[4] J. Han, L. Shao, D. Cu and J. Shotton, "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review," *IEEE Trans. Cybernetics,* vol. 43, no. 5, pp. 1318-1334, Oct. 2013.

[5] C. Wang, Z. Y. Zhu, S. C. Chan and H. Y. Shum, "Real-time Depth Image Acquisition and Restoration for Image Based Rendering and Processing Systems," *Journal of Signal Processing Systems*, pp. 1-18, 2013, 10.1007/s11265-013-0819-2.

[6] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel Regression for Image Processing and Reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 349-366, Feb. 2007.

[7] Z. G. Zhang, S. C. Chan, and C. Wang, "A New Regularized Adaptive Windowed Lomb Periodogram for Time-Frequency Analysis of Nonstationary Signals With Impulsive Components," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 8, pp. 2283-2304, Aug. 2012.

[8] C. Wang and S. C. Chan, "A New Bandwidth Adaptive Non-Local Kernel Regression Algorithm For Image/Video Restoration and Its GPU Realization," in Proc. ISCAS, 2013.

[9] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," In Proc. CVPR, 2007.

[10] J. Diebel and S. Thrun. "An application of markov random fields to range sensing," In Proc. NIPS, 2006.

[11] K. He, J. Sun, and X. Tang. "Guided image filtering," In Proc. ECCV, 2010.

[12] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, 2007.

[13] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," In Proc. ICCV, 2011.