Depth Recovery Using an Adaptive Color-Guided Auto-Regressive Model

Jingyu Yang, Xinchen Ye, Kun Li, and Chunping Hou

Tianjin University, Tianjin, 300072, China
{yjy,yexch,lik,hcp}@tju.edu.cn

Abstract. This paper proposes an adaptive color-guided auto-regressive (AR) model for high quality depth recovery from low quality measurements captured by depth cameras. We formulate the depth recovery task into a minimization of AR prediction errors subject to measurement consistency. The AR predictor for each pixel is constructed according to both the local correlation in the initial depth map and the nonlocal similarity in the accompanied high quality color image. Experimental results show that our method outperforms existing state-of-the-art schemes, and is versatile for both mainstream depth sensors: ToF camera and Kinect.

Keywords: Depth recovery, AR model, nonlocal filtering, depth camera.

1 Introduction

Acquiring depth information of real scenes is an essential task for many applications such as 3DTV, augmented reality, and 3D reconstruction. There are mainly two categories of methods to obtain depth information: passive methods and active methods. Passive methods still suffer from some inherent problems for practical application, e.g., requiring strict image rectification and inefficiency for textureless areas [1]. The alternative is to acquire depth information by active devices, e.g., Time-of-flight (ToF) camera, and Kinect.

Time-of-flight (ToF) based technique is a recent advance in active depth sensing. ToF cameras determine depth information by measuring the phase difference of the emitted light and the reflected light. ToF cameras can capture depth information for dynamic scenes in real time, but are noisy and have low resolutions, e.g., 176×144 and 200 × 200, compared with popular color cameras. Microsoft Kinect is another break through to achieve real-time depth capturing for dynamic scenes. In Kinect, an infrared light source projects some patterns on the scene and an offset infrared camera receives the pattern and estimates the depth information. The generated depth maps contain considerable holes due to the occlusion caused by the relative displacement of projector and camera.

While the new depth sensing techniques are promising, the use of depth cameras is limited by the low quality of produced depth maps, e.g., low resolution, noise, and depth missing in some areas. Some previous work on depth recovery for depth cameras were proposed, and are briefly reviewed as follows.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 158–171, 2012.

Depth Recovery for ToF Cameras: Limited by the sensing mechanism, ToF cameras have a low resolution, which impedes their practical applications. It is difficult to recover high quality depth maps from only the severely undersampled versions due to the loss of salient discontinuities. Fortunately, the depth information and texture information are two descriptions of the same scene from different perspectives, and thus present strong structural correlations. In particular, discontinuities often simultaneously present at the same locations in a depth map and the corresponding (registered) color image, and homogeneous regions in color image tend to have similar depths. Therefore, the common wisdom is to couple a color camera with a ToF camera and recover high quality maps with the help of the accompanied color image [2–6].

Markov Random Field (MRF) that plays an important role in classic stereo matching [1] is also powerful in depth recovery with accompanied color images. Diebel and Thrun [6] proposed a two-layer MRF to model the correlation between range measurements and solve the MRF optimization with the conjugate gradient algorithm. Hu *et al.* [7] further extended this work by designing a data term that fits to the characteristics of depth maps. Zhu *et al.* [8] extended the traditional spatial MRFs to dynamic MRFs so that both the spatial and the temporal relationship can be propagated in local neighbors, improving accuracy and robustness of depth recovery for dynamic scenes. Another category is to use advanced filters such as bilateral filters and non-local means (NLM) filters [9–11]. Joint bilateral filtering and its variations are readily available tools for depth recovery using high quality auxiliary color images for cross filtering [9, 10]. Chan *et al.* designed an adaptive multi-lateral upsampling filter to further address the noise in depth measurements. Min *et al.* [12] proposed a weighted mode filtering method based on a joint histogram of depth video and color video. Huhle *et al.* [13] proposed a fusion scheme based on the non-local principle. Park *et al.* [14] used a NLM term to regularize depth maps and combined with a weighting scheme that involves edge, gradient, and segmentation information extracted from high quality color images.

Depth Recovery for Kinect: The main degradation in depth maps produced by Kinect is random depth missing on the background and structural depth missing for occlusion regions. One possible way is to use inpainting methods. Lai *et al.* [15] filled missing depth values by recursively applying a median filter in their work on RGB-D object dataset construction, but blurring occurs in large occlusion. Matyunin *et al.* [16] proposed a depth restoration method via temporal filtering, which requires consecutive frames and introduces delay. Camplani and Salgado [17] used a joint bilateral filter similar to Ref. [9] to fill missing depth. However, the filtered depth maps present obvious artifacts around discontinuities. The occlusion filling method in Ref. [18] achieves real-time processing, but the recovered depth maps are not always consistent with the accompanied color image, particularly around the boundaries of background and foreground.

These methods achieve good quality for smooth regions, but may introduce artifacts, e.g. jagging, blurring, and ringing, around thin structures or sharp discontinuities. Both taking a low quality depth map and a high quality color image

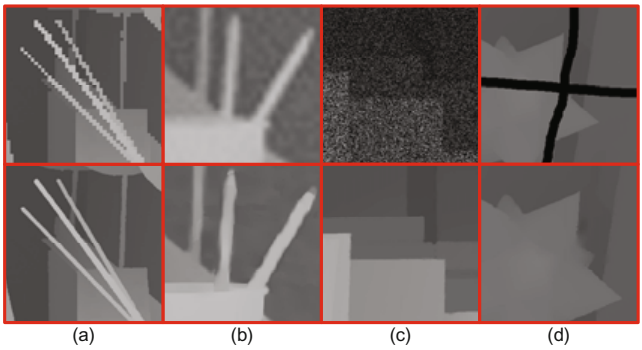

Fig. 1. Depth maps recovered by the proposed method from various types of degradations: (a) undersampling (1/8), (b) undersampling (1/8) with strong signal-dependent additive noise, (c) random missing, and (d) structural missing. The degraded depths and the recovered ones are at the top row and bottom row, respectively.

as input, depth recovery problems for ToF camera and Kinect are the same, but are treated separately in literature. This paper proposes a color-guided AR model to construct a unified depth recovery framework for both ToF and Kinect depth cameras. Observing and verifying the fitness of AR model for depth maps, we design pixel-wise adaptive AR predictors based on the non-local similarity of both the depth map and the accompanied color image. The depth map is recovered by minimizing AR prediction errors subject to observation consistency. Experiments demonstrate that our method can handle all common depth degradation modes, as shown in Fig. 1, and is versatile for various depth capturing systems such as ToF cameras and Kinect. Surprisingly, without resorting to higher level tools such as segmentation used in Ref. [14], our proposed method achieves the best quality among several state-of-the-art depth recovery methods.

2 Degradation Modes and AR Model of Depth Maps

2.1 Degradation Modes of Depth Maps

Current depth capturing systems are far from perfect. A captured depth map is a degraded version of the underlying groundtruth. Let $\boldsymbol{d}$ and $\boldsymbol{d}^0$ denote the vector form of the underlying perfect depth map and the captured one, respectively. The observation model for depth capturing is described as

$$\boldsymbol{d}^0 = \boldsymbol{H}\boldsymbol{d} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{H}$ represents the observation matrix and $\boldsymbol{n}$ is additive noise.

There are mainly four types of degradations: undersampling, random depth missing, structural depth missing, and pollution with additive noise. For the former three ones, the observed depth map $\boldsymbol{d}^0$ has a smaller number of elements

than $\boldsymbol{d}$ and $\boldsymbol{H}$ is a fat matrix, which makes the depth recovery ill-posed to attack. As shown in Fig. 6, the depth maps captured by ToF is undersampled (lower resolution than the accompanied color image), and polluted by noise. After viewpoint registration, the warped depth map contains disoccludded regions around object boundaries, and thus suffers from degradation of structural depth missing. In Fig. 7, the Kinect depth map contains both random and structural missing degradations. Our method is to recovery high quality depth maps from low quality observations, and all the four kinds of degradations are handled in the proposed unified depth recovery framework.

2.2 AR Model of Depth Maps

As shown in Fig. 4-7, depth maps for generic 3D scenes contain mainly smooth regions separated by curves. AR model can well describe such type of 2D signals: The key insight is that a signal can be regenerated by the signal itself. Denote by $\boldsymbol{D}$ a depth map, and $\boldsymbol{D}_{\boldsymbol{x}}$ the depth value at location $\boldsymbol{x}$. The predicted depth map $\tilde{\boldsymbol{D}}$ by the AR model from the depth map $\boldsymbol{D}$ is expressed as

$$\tilde{\boldsymbol{D}}_{\boldsymbol{x}} = \sum_{\boldsymbol{y} \in \mathcal{N}_{\boldsymbol{x}}} a_{\boldsymbol{x},\boldsymbol{y}} \boldsymbol{D}_{\boldsymbol{y}}, \tag{2}$$

where $\mathcal{N}_{\boldsymbol{x}}$ is the neighborhood of pixel $\boldsymbol{x}$ and $a_{\boldsymbol{x},\boldsymbol{y}}$ denotes the AR coefficient for pixel $\boldsymbol{y}$ in the neighborhood $\mathcal{N}_{\boldsymbol{x}}$. The accuracy of the AR model can be measured by the difference between $\boldsymbol{D}$ and $\tilde{\boldsymbol{D}}$, e.g., mean absolute difference (MAD) or root mean squared error (RMSE).

To verify the fitness of the AR model for depth maps, we check the prediction error between the predicted depth maps and the ground truths for a set of test depth maps. Four AR predictors are tested: an average filter, a Gaussian filter, a bilateral filter and the proposed filter, all with a 11×11 neighborhood. As shown in Fig. 2, all the four filters have good prediction for smooth regions. The proposed filter has the smallest prediction error for discontinuities of depth maps. Since the proposed filter adapts the AR model to the nonlocal structures of signals, it almost regenerates the depth map: the average prediction error (MAD) is only 0.051/pixel. These results demonstrate that the AR model is quite effective in modeling the depth maps, which supports the application of this model to the recovery of depth maps.

Depth information and the associated texture information have strong correlation in terms of geometrical structures, and often are acquired and used together [6, 19]. Depth maps are of low resolution and low signal-to-noise ratio while color images are of high resolution and high quality. Exploiting depth-color correlations is quite informative for depth recovery when the accompanied color images are available. As shown in Fig. 2(a) and 2(b), edges in depth maps have their counterparts in color images. This suggests that the locations of edges in depth maps can be inferred from the accompanied color images, and motivates the proposed color-guided AR model for depth recovery from low resolution and incomplete observations.

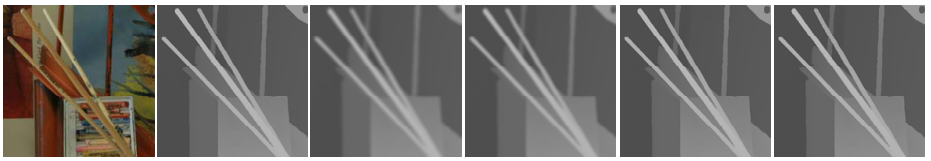

Fig. 2. Prediction efficiency for four AR predictors: (a) the associated color image, (b) the input depth map, and prediction results of AR predictors constructed by (c) average filter, (d) Gaussian filter, (e) bilateral filter, and (f) the propose filter. The prediction error (MAD) between the predictions in (c), (d), (e), and (f) against the original depth maps are 3.992, 3.131, 0.129, 0.051, respectively.

3 Color-Guided AR Model for Depth Recovery

AR model has been successively applied in many image processing applications, such as image interpolation [20] and video frame-rate upconversion [21]. This section describes our AR formulation of the depth map recovery, which takes the benefit of the strong correlation between depth maps and the associated color images.

3.1 Depth Recovery Based on AR Model

Denote by $\boldsymbol{D}^0$ the observed depth map and $\mathcal{O}$ the set of pixels with observed depth values. Given the observed depth map $\boldsymbol{D}^0$, we proposed the following depth recovery model based on AR:

$$\min_{\boldsymbol{D}} E_{\text{data}}\left(\boldsymbol{D}, \boldsymbol{D}^0\right) + \lambda E_{\text{AR}}\left(\boldsymbol{D}\right), \tag{3}$$

where $E_{\text{data}}(\boldsymbol{D}, \boldsymbol{D}^0)$ is the data term to make the recovered depth map consistent with the observation, and $E_{\text{AR}}(\boldsymbol{D})$ is the AR term to impose AR model on the recovered depth map. The data term and the AR term are weighted by λ.

The data term is expressed as

$$E_{\text{data}}\left(\boldsymbol{D}, \boldsymbol{D}^0\right) \triangleq \sum\nolimits_{\boldsymbol{x}\in\mathcal{O}} \left(\boldsymbol{D}_{\boldsymbol{x}} - \boldsymbol{D}^0_{\boldsymbol{x}}\right)^2, \tag{4}$$

and the AR model is incorporated into the depth recovery as the AR term

$$E_{\text{AR}}\left(\boldsymbol{D}\right) \triangleq \sum\nolimits_{\boldsymbol{x}} \left(\boldsymbol{D}_{\boldsymbol{x}} - \sum\nolimits_{\boldsymbol{y}\in\mathcal{N}_{\boldsymbol{x}}} a_{\boldsymbol{x},\boldsymbol{y}} \boldsymbol{D}_{\boldsymbol{y}}\right)^2, \tag{5}$$

where the AR coefficient $a_{\boldsymbol{x},\boldsymbol{y}}$ is defined according to both depth and color information in the following section. The proposed method has a similar form, but is a departure essentially, to related work. For example, the work in Ref. [14] solves the depth upsampling problem with an energy minimization with three terms, i.e., data term, smoothness term, and non-local term. In addition to color information, segmentation and edge saliency are taken into account in confidence

weights. Although such features can be readily incorporated in our recovery model, we found that the elegant AR model can well describe the characteristics of depth maps. Therefore, we insist on the low level processing in depth recovery, and retain the simplicity of the model.

As shown in Sec. 2.2, the AR model is powerful in describing depth maps only when the AR coefficients are properly designed. However, an accurate AR model is difficult to infer from only the degraded depth map $\boldsymbol{D}^0$. Noting that the depth-color pairs have strong structural correlations, the information loss due to depth degradation can be complemented by the accompanied color image. To achieve high quality depth recovery, we design pixel-wise adaptive AR predictors in Sec. 3.2 using both the initial depth map and the auxiliary color image.

3.2 Color-Guided AR Model

In AR-based image processing [20], images are often divided into small units and each unit shares an AR predictor. Preprocessing like segmentation is often used for extraction of homogeneous regions. However, the piece-wise AR model cannot provide sufficient adaptivity when each unit contains considerable variations. Therefore, we design pixel-wise adaptive AR predictors for depth recovery: an AR predictor $\{a_{\boldsymbol{x},\boldsymbol{y}}\}$, $\boldsymbol{y} \in \mathcal{N}_{\boldsymbol{x}}$ is constructed for each pixel $\boldsymbol{x}$ by considering both the depth and color information.

A depth map is reliably recovered with the optimal AR predictors, which can be derived only when the depth map is available. To break this chicken-egg dilemma, we design AR predictors using the available depth map and the accompanied color image. Note that the observed depth map $\boldsymbol{D}^0$ is not directly applicable due to degradations such as the undersampling or depth missing. Denote by $\hat{\boldsymbol{D}}$ the rough estimated depth map obtained by bicubic interpolation from $\boldsymbol{D}^0$. Represent the accompanied color image with $\boldsymbol{I} = \{\boldsymbol{I}^i, i \in \mathcal{C}\}$, where $\boldsymbol{I}^i$ is the intensity of the color channel with index i and $\mathcal{C}$ is the index set of color channels in a certain color space. We had investigated three color spaces (RGB, YUV, and Lab). All three color spaces yield similar results, and we choose the YUV color space due to its slightly better performance, i.e., $\mathcal{C} = \{Y, U, V\}$. The AR coefficient $a_{\boldsymbol{x},\boldsymbol{y}}$ consists of two terms:

$$a_{\boldsymbol{x},\boldsymbol{y}} = \frac{1}{C_{\boldsymbol{x}}} a^{\hat{\boldsymbol{D}}}_{\boldsymbol{x},\boldsymbol{y}} a^{\boldsymbol{I}}_{\boldsymbol{x},\boldsymbol{y}}, \tag{6}$$

where $C_{\boldsymbol{x}}$ is the normalization factor, $a^{\hat{\boldsymbol{D}}}_{\boldsymbol{x},\boldsymbol{y}}$ and $a^{\boldsymbol{I}}_{\boldsymbol{x},\boldsymbol{y}}$ are the depth term and color term, respectively.

The depth term $a^{\hat{\boldsymbol{D}}}_{\boldsymbol{x},\boldsymbol{y}}$ is defined on the initial estimated depth map $\hat{\boldsymbol{D}}$ by a range filter:

$$a^{\hat{\boldsymbol{D}}}_{\boldsymbol{x},\boldsymbol{y}} = \exp\left(- \frac{\left(\hat{\boldsymbol{D}}_{\boldsymbol{x}} - \hat{\boldsymbol{D}}_{\boldsymbol{y}}\right)^2}{2\sigma_1^2} \right), \tag{7}$$

where σ_1 is the decay rate of the range filter. Qualitatively, $a^{\hat{\boldsymbol{D}}}_{\boldsymbol{x},\boldsymbol{y}}$ has a large value if $\hat{\boldsymbol{D}}_{\boldsymbol{x}}$ is close to $\hat{\boldsymbol{D}}_{\boldsymbol{y}}$. This term is also to avoid incorrect depth prediction

due to depth-color inconsistency: pixels of the same depth layers may have very different colors; pixels of similar colors may belong to different depth layers.

The color term $a^{\boldsymbol{I}}_{\boldsymbol{x},\boldsymbol{y}}$ is designed to take benefit of the inter-correlations in the depth-color pair. Edges in a depth map co-occur with their counterparts in the accompanied color image. The color term $a^{\boldsymbol{I}}_{\boldsymbol{x},\boldsymbol{y}}$ should be able to prevent the AR model from predicting across depth discontinuities. Structure-aware filters such as the NLM filter can be used for this purpose. Based on the non-local principle, we propose the following color terms :

$$a^{\boldsymbol{I}}_{\boldsymbol{x},\boldsymbol{y}} = \exp\left(- \frac{\sum_{i\in\mathcal{C}} \| \boldsymbol{B}_{\boldsymbol{x}} \circ \left(\mathcal{P}^i_{\boldsymbol{x}} - \mathcal{P}^i_{\boldsymbol{y}}\right) \|^2_2}{2 \times 3 \times \sigma_2^2} \right), \tag{8}$$

where σ_2 controls the decay rate of the exponential function, $\mathcal{P}^i_{\boldsymbol{x}}$ denotes an operator that extracts a $w \times w$ patch centered at $\boldsymbol{x}$ in color channel i, "$\circ$" represents the element-wise multiplication. The bilateral filter kernel $\boldsymbol{B}_{\boldsymbol{x}}$ is defined in the neighborhood $\mathcal{N}_{\boldsymbol{x}}$ as in the patch operator:

$$\boldsymbol{B}_{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2_2}{2\sigma_3^2}\right) \exp\left(-\frac{\sum_{i\in\mathcal{C}}(\boldsymbol{I}^i_{\boldsymbol{x}} - \boldsymbol{I}^i_{\boldsymbol{y}})^2}{2 \times 3 \times \sigma_4^2}\right), \tag{9}$$

where σ_3 and σ_4 are parameters of the bilateral filter to adjust the importance of the spatial distance and intensity difference.

The difference between the proposed filter in the color term and the standard NLM filter is that the proposed one uses a bilateral kernel to weight the distance of local patches while the standard one uses a Gaussian kernel. The bilateral

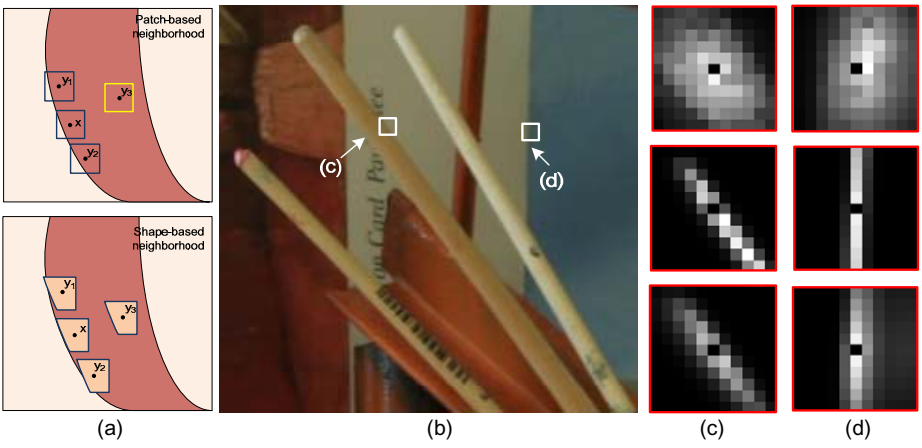

Fig. 3. Illustrations for the color term of AR predictors: (a) patch-based neighborhood and shape-based neighborhood, (b) two pixels centered with neighborhood for similarity search, (c) and (d) presents AR predictors for the two pixels constructed by bilateral filter (top), standard NLM filter (middle), and the proposed filter (bottom)

kernel $\boldsymbol{B}_{\boldsymbol{x}}$ has a strong response for pixels of similar intensities to $\boldsymbol{x}$, and hence carries the shape information of local image structures. This extends the NLM filter from path-based to shape-based in measuring the resemblance of local structures, and has a significant impact on the structure of AR predictors for pixels around edges. As shown in the top part of Fig. 3 (a) , two homogeneous regions are separated by smooth curves and $\boldsymbol{x}$ is nearby a curve. To construct the AR predictor for $\boldsymbol{x}$, the similarity of between $\boldsymbol{x}$ and each pixel $\boldsymbol{y}$ in the neighborhood $\mathcal{N}_{\boldsymbol{x}}$ is evaluated. Constrained by the patch structure, the standard NLM filter produces large coefficients only for pixels that are parallel to the edge, e.g. $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, but small coefficients for other pixels, such as $\boldsymbol{y}_3$, even though they have the same intensity as $\boldsymbol{x}$. On the contrary, our bilateral-weighted NLM filter has a shape-adaptive neighborhood, and increases opportunities to exploit more correlations for pixels around discontinuities. This is illustrated in the bottom part of Fig. 3 (a). With the shape-adaptive neighborhood, the proposed filter produce an equally large coefficient for $\boldsymbol{y}_3$ as for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. This will enlarge the prediction set for AR predictor, and lead to more stable and reliable estimation.

Fig. 3 further shows real examples of AR predictors constructed from the bilateral filter, standard NLM filter, and the proposed filter. The proposed color term of AR predictor can better adapt to the local structures of color images. This merit helps to well determine the inverse problem. Our recovery depth model is equivalent to solve a linear system, say $\boldsymbol{A}\boldsymbol{d} = \boldsymbol{d}_0$, where $\boldsymbol{d}$ is the vector form of the depth map to be recovered, $\boldsymbol{A}$ is the matrix constructed from the assumed model and $\boldsymbol{d}_0$ is the vector form of the observed depth map. The AR predictor for each pixel is associated with each row of $\boldsymbol{A}$. The stability of the linear system tightly depends on the structures of AR predictors. AR predictors estimated by the standard NLM filter tend to have small supports for pixels around edges and contours, which would underdetermine the linear system at least for related pixels; our proposed AR predictors tend to have larger supports and form a more well-determined system, and therefore are more powerful in depth recovery.

4 Experiments and Results

Our method is first evaluated on Middlebury datasets with four kinds of synthetic degradations and compared with several existing methods. Then, our method is applied on two real depth cameras to obtain high quality depth maps.

4.1 Experiments on Datasets with Synthetic Degradations

Three datasets, *Art*, *Book*, and *Moebius*, from the Middlebury's benchmark [22] are used for evaluation. Four kinds of typical degradations are synthesized: subsampling, subsampling with noise, random missing, and structural missing. In the following experiments, the parameters in the AR model are set as: $\lambda = 0.01$, $\sigma_1 = 4$, $\sigma_2 = 6.67$, $\sigma_3 = 3.5$, $\sigma_4 = 0.25$. We investigated the influence of the parameters on the recovery performance, and find that the performance is stable

when the parameters are within certain ranges: $\lambda \in [0.01, 0.10]$, $\sigma_1 \in [3.0, 7.0]$, $\sigma_2 \in [3.8, 19.0]$, $\sigma_3 \in [3, 10]$, $\sigma_4 \in [0.01, 0.45]$. The neighborhood $\mathcal{N}_{\boldsymbol{x}}$ is chosen as a 11×11 patch with $\boldsymbol{x}$ as its center.

For the subsampling degradation, recovery results in terms of MAD against the ground-truth depth maps are reported in Table 1. Beside the bicubic interpolation, the proposed method is compared with four recent methods: MRF [6], bilateral filter [9], guided image filter [23], and edge-weighted NLM-regularization [14] [1]. As shown in Table 1, our method obtains the lowest MAD for all cases, which demonstrates its effectiveness. For visual comparison, $8\times$ upsampled depth maps for *Art* are shown in Fig. 4. The MRF method [6] tends to produce oversmooth results. The edge-weighted NLM-regularized method [14] generates comparable results to ours, but introduces some jaggy artifacts along edges. The computational complexity of our method is similar to the NLM-regularized method [14] since they both solve minimization with quadratic terms, and is higher than methods in Ref. [9, 23] that perform single filtering for each pixel.

Table 1. Quantitative superresolution results from subsampled depth maps on Middlebury datasets at four subsampling rates. Our method consistently achieves the best recovered quality (the lowest MAD) among all the compared methods.

	Art				*Book*				*Mobius*			
	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×
Bicubic	0.48	0.97	1.85	3.59	0.13	0.29	0.59	1.15	0.13	0.30	0.59	1.13
MRFs [13]	0.62	1.01	1.97	3.94	0.22	0.33	0.62	1.21	0.25	0.37	0.67	1.29
Bilateral [16]	0.57	0.70	1.50	3.69	0.30	0.45	0.64	1.45	0.39	0.48	0.69	1.14
Guided [24]	0.66	1.06	1.77	3.63	0.22	0.36	0.60	1.16	0.24	0.38	0.61	1.20
Edge [9]	0.43	0.67	1.08	2.21	0.17	0.31	0.57	1.05	0.18	0.30	0.52	**0.90**
Ours	**0.18**	**0.49**	**0.66**	**2.15**	**0.11**	**0.25**	**0.48**	**0.80**	**0.11**	**0.23**	**0.42**	**0.90**

Recovery results for other three kinds of depth degradations are summarized in Table 2. For subsampling with additive noise, we add strong signal-dependent Gaussian noise whose variance is within the range of [0, 225] and is proportional to depth values. More distant depth layers are polluted with stronger noise. For random-missing degradation, depth values are randomly dropped up to three ratios: 10% , 20% , and 50%. For structural-missing degradation, missing areas are created by drawing black curves on depth maps with a brush. As shown in Table 2, the proposed method provides promising recovery quality for all the three types of depth degradations. Particularly for the random missing and structural missing, the average recovery error is consistently at a low level for all the three datasets. Compared with other two cases, the recovery error for undersampling with additive noise is much higher due to the strength of the noise and the complex nature of signal-dependent noise. Fig. 5 shows the recovered depth maps together with degraded ones. Our method nearly achieves visually perfect recovery for random missing and structural missing. For undersampling

[1] The authors thank J. Park for providing recovered depth maps for comparison.

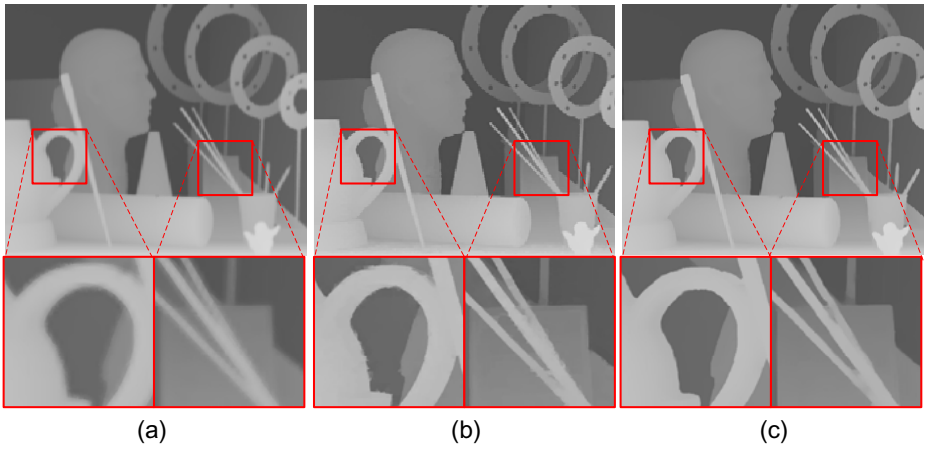

Fig. 4. Depth maps upsampled (8×) by (a) the MRFs method [6], (b) edge-weighted NLM-regularized method [14], and (c) our method

with additive noise, our method obtains good recovery quality for close depth layers (low noise level), but introduces "dirty clouds" for distant depth layers (strong noise level). This suggests that the proposed method owns the capability to fight against moderate-level noise, but would better resort to a denoising filter before depth recovery.

Table 2. Quantitative depth recovery results for other three types of degradations: undersampling with additive noise, random missing, structural missing

	Art			*Book*			*Mobius*		
Additive noise (8×)	4.91			4.50			5.48		
Structural missing	0.12			0.07			0.06		
Random missing	10%	20%	50%	10%	20%	50%	10%	20%	50%
	0.09	0.17	0.45	0.03	0.05	0.14	0.03	0.07	0.17

4.2 Experiments on Real Systems

Recent mainstream depth sensors include ToF depth camera and Microsoft Kinect camera. Both depth sensors have their disadvantages: ToF depth cameras have low resolutions while Kinect camera suffers from the occlusion problem. It is desirable to have higher quality depth maps in practical applications. We apply our method on these two depth sensors to achieve high quality depth recovery from the low quality sensor measurements.

ToF depth camera: For the first depth sensing system, we mount a high resolution Grey Flea2 color camera on a PMD[vision] CamCube3 ToF depth camera to construct a depth-color camera rig. The ToF camera has a resolution of 200×200,

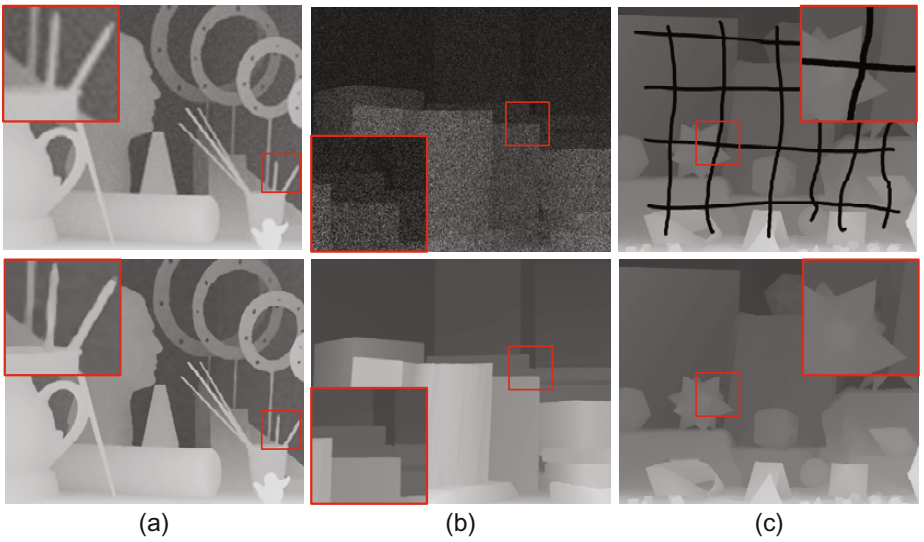

Fig. 5. Recovered depth maps (bottom row) from degraded versions (top row): (a) undersampling (1/8) with additive noise, (b) random missing at the ratio of 50%, and (c) structural missing. The highlighted regions are enlarged for visual inspection.

and the resolution of color camera is set at 640×480 to obtain nearly the same field of view as the ToF camera. To compensate misalignment due to different viewpoints, depth maps are warped to the viewpoint of the color camera using intrinsic parameters and extrinsic parameters for both cameras computed by the camera calibration module in the OpenCV library [24]. Outliers in depth maps are rejected by the associated amplitude images as confidence levels.

As shown in Fig. 6(a), the obtained depth map suffers from the undersampling degradation combined with some random missing on the ceiling. The viewpoint warping between the depth camera and color camera introduces structural missing around depth discontinuities due to disocclusion. To recover depth map from measurements with various degradation modes is more challenging than the previous simulations with single synthetic degradation mode. It is observed in Fig. 6(b) that the recovered depth map quite coincides with the depth relationship as suggested in the accompanied color image. Fig. 6(c) shows the rendered results from the color image and the recovered depth map. We render the image at a different viewpoint from the one of the color camera. The rendered image suggests that our method reliably restores the geometric relationship from a set of low-resolution depth measurements. However, it is observed that there is also slight global blending artifact in the rendering results. We find that the blending artifact is caused by the geometrical distortion of the ToF depth camera. This suggests that geometric distortion compensation can be used a as preprocessing to further improve the recovery accuracy.

Kinect camera: Microsoft Kinect is an integrated sensor array for natural user interaction, consisting of a depth camera and a color camera. The captured

depth maps and color images are of size 640×480. We observe that color images captured by Kinect present jaggy artifacts and fake colors along discontinuities such as contours and edges due to inefficient demosaicing, which severely affects the quality of depth recovery. These artifacts in color images are suppressed by re-demosaicing the color images with a more advanced demosaicing method [25].

Fig. 7 shows the depth-color pair captured by the Kinect camera as well as depth recovery results. It is observed that the depth map contains lots of holes around depth discontinuities due to occlusion. This corresponds to the case of synthetic datasets with structural missing areas in Sec. 4.1. The recovered depth map and the rendered image indicate that the recovered depth map correctly preserves the 3D geometrical relationship of the captured scene. We also observe that there are some artifacts around the depth discontinuities in the rendered view due to the smooth transition of different depth layers (the person, the sofa, and curtain). This is caused by the blurring of the Kinect color image that finally leads to the blurring in the recovered depth map. For applications where the accompanied color images are of low quality, we can improve the recovery quality by applying some pre-processing techniques on the color images such as deblurring and sharpening.

The results in this section demonstrate that the proposed method is versatile for both the two mainstream depth cameras, and is applicable in various

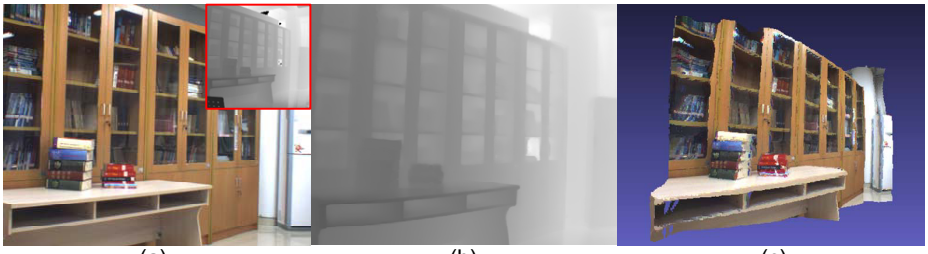

Fig. 6. Depth recovery results for ToF camera: (a) input depth-color pair, (b) recovered depth map, and (c) rendered image at a new viewpoint from the recovered depth and the color image. The depth-color pair is shown at their original ratio of size.

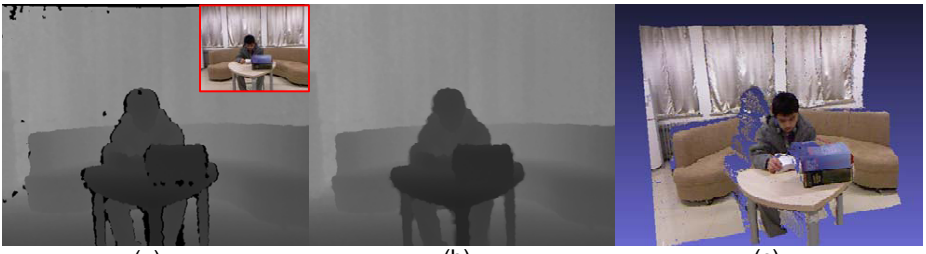

Fig. 7. Depth recovery results for Kinect camera: (a) input depth-color pair, (b) recovered depth map, and (c) rendered image at a new viewpoint. The Kinect color image is shown at a smaller size overlayed on the depth map to save space.

applications involving depth sensing. It is also worth to mention that our depth recovery method should collaborate with other preprocessing techniques such as geometric distortion calibration for depth sensor and color image sharpening to achieve high quality depth acquisition in practical systems.

5 Conclusion

This paper proposes an elegant framework to recover depth maps from low quality measurements with various types of degradations. We show that depth maps are well described by AR models if the AR predictors can adapt to the characteristics of depth maps. Based on this observation, we design pixel-wise adaptive AR predictors using both the depth map and the accompanied color image. The depth map is recovered by minimizing AR prediction errors subject to the observation consistency. Experiments demonstrate that our method achieves high quality depth recovery from low quality versions with various degradation. Experiments on two real systems demonstrate that our method is versatile for various depth capturing systems such as ToF cameras and Kinect.

The proposed framework would be extended and improved in future work: 1) incorporate other regularization to strengthen robustness against noise, 2) optimize model parameters according to depth and image characteristics, 3) derive fast algorithms to achieve real-time implementation for practical applications.

Acknowledgement. This work is supported by the NSF of China (60932007, 61072062), the NCET Program (NCET-11-0376) and Ph.D Program Foundation (20110032110029) from the Ministry of Education of China, and the Tianjin Research Program of Application Foundation and Advanced Technology (12JCYBJC10300).

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 47(1), 7–42 (2002)
2. Prasad, T., Hartmann, K., Weihs, W., Ghobadi, S., Sluiter, A.: First steps in enhancing 3D vision technique using 2D/3D sensors. In: Computer Vision Winter Workshop (2006)
3. Linarth, A., Penne, J., Liu, B., Jesorsky, O., Kompe, R.: Fast fusion of range and video sensor data. In: Advanced Microsystems for Automotive Applications, pp. 119–134. Springer (2007)
4. Lindner, M., Kolb, A., Hartmann, K.: Data-fusion of PMD-based distance-information and high-resolution RGB-images. In: Int. Symp. on Sig., Circ. & Syst., vol. 1, pp. 1–4. IEEE (2007)
5. Guomundsson, S., Larsen, R., Aanaes, H., Pardas, M., Casas, J.: ToF imaging in smart room environments towards improved people tracking. In: CVPR Workshops (2008)
6. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. Advances in Neural Information Processing Systems 18, 291 (2006)

7. Lu, J., Min, D., Pahwa, R., Do, M.: A revisit to MRF-based depth map super-resolution and enhancement. In: ICASSP (2011)
8. Zhu, J., Wang, L., Gao, J., Yang, R.: Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5), 899–909 (2010)
9. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: CVPR (2007)
10. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: CVPR (2010)
11. Huhle, B., Schairer, T., Jenke, P., Straßer, W.: Fusion of range and color images for denoising and resolution enhancement with a non-local filter. CVIU 114(12), 1336–1345 (2010)
12. Min, D., Lu, J., Do, M.: Depth video enhancement based on joint global mode filtering. IEEE TIP 21(3), 1176–1190 (2011)
13. Huhle, B., Schairer, T., Jenke, P., Straßer, W.: Robust non-local denoising of colored depth data. In: CVPR Workshops (2008)
14. Park, J., Kim, H., Tai, Y., Brown, M., Kweon, I.: High quality depth map upsampling for 3D-ToF cameras. In: ICCV (2011)
15. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: ICRA (2011)
16. Matyunin, S., Vatolin, D., Berdnikov, Y., Smirnov, M.: Temporal filtering for depth maps generated by Kinect depth camera. In: 3DTV Conference (2011)
17. Camplani, M., Salgado, L., Grupo de Tratamiento de Imágenes, E.T.S.I de Telecomunicación: Efficient spatio-temporal hole filling strategy for Kinect depth maps. In: SPIE (2012)
18. Berdnikov, Y., Vatolin, D.: Real-time depth map occlusion filling and scene background restoration for projected-pattern based depth cameras. In: Graphic Conf., IETP (2011)
19. Kolb, A., Barth, E., Koch, R., Larsen, R.: Time-of-flight cameras in computer graphics. In: Computer Graphics Forum, vol. 29, pp. 141–159. Wiley Online Library (2010)
20. Zhang, X., Wu, X.: Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation. IEEE TIP 17(6), 887–896 (2008)
21. Zhang, Y., Zhao, D., Ji, X., Wang, R., Gao, W.: A spatio-temporal auto regressive model for frame rate upconversion. IEEE TCSVT 19(9), 1289–1301 (2009)
22. Middlebury datasets, `http://vision.middlebury.edu/stereo/data/`
23. He, K., Sun, J., Tang, X.: Guided Image Filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 1–14. Springer, Heidelberg (2010)
24. Open source computer vision (OpenCV), `http://opencv.org`
25. Paliy, D., Foi, A., Bilcu, R., Katkovnik, V.: Denoising and interpolation of noisy bayer data with adaptive cross-color filters. In: Proc. of SPIE VCIP (2008)