Image Assisted Upsampling of Depth Map via Nonlocal Similarity

Wentian Zhou
West Virginia University
Morgantown, WV 26506
Email: wzhou2@mix.wvu.edu

Xin Li
West Virginia University
Morgantown, WV 26506
Email: xin.li@ieee.org

Daryl Reynolds
West Virginia University
Morgantown, WV 26506
Email: daryl.reynolds@mail.wvu.edu

***Abstract*—The depth resolution of TOF cameras is poor and the resulting depth maps are noisy. Therefore, it is highly desirable to develop powerful image processing tools to enhance the resolution and suppress the noise of depth maps. In this paper, we propose a new image-assisted upsampling method for depth map. A spatially adaptive iterative singular-value thresholding (SAIST) with image-guided patch clustering strategy is developed and compared with previous image-guided depth map upsampling techniques. Overall, the proposed scheme is capable of better preserving salient global structure information. Extensive experimental results are reported to justify the superiority of the proposed method.**

I. Introduction

Capturing a high resolution depth map has been a challenging task. Before the introduction of time-of-flight (TOF) cameras, people attempted to use expensive laser scanners to obtain the depth information. This pixel-per-pixel scanning procedure often requires a large amount of time and computational resources which limits its use in practice. Recently, Microsoft announced their second-generation Kinect, which was an inexpensive and portable depth map capturing device. Unlike the previous generation, it was equipped with a TOF camera providing faster depth map acquisition. However, the depth resolution of even this new-generation Kinect is still low and the acquired depth map often suffers from the impact of acquisition noise. Consequently, developing powerful depth map processing tools such as upsampling and denoising has become desirable for the purpose of obtaining high quality depth maps.

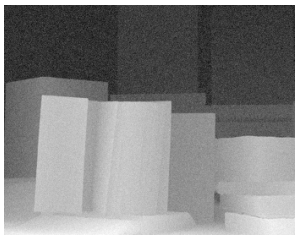
Early upsampling methods include two major classes, one is filter-based [1], [2], and the other is Markov Random Field (MRF) based [3]. Joint Bilateral Upsampling (JBU) [1] and MRF [3] both promote the idea of extracting additional information from a high resolution image to assist the upsampling process and produce good results. Thus, various information can be obtained from a high resolution image and used for upsampling. More recent works include guided image filtering [4] that can be used as an edge-preserving smoothing operator; Anisotropic Total Generalized variation approach [5] formulate the upsampling as a convex optimization problem using primal-dual algorithm, and high resolution upsampling [6] construct this problem using constrained optimization. Most recently, depth map denoising has also been studied in the literature (e.g., [7]).

In this paper, we view depth map denoising and upsampling as a joint problem and propose to exploit the patch dependency/similarity of color images to guide the upsampling process of low-resolution depth maps (please refer to Fig.1). Our method relies on the fact that depth discontinuity information of a targeted scene is often embedded within the corresponding high resolution color image (with the only exception of textured regions) [8]. Meantime, even though homogeneous texture regions of natural images tend to appear on the same depth layer, such seemingly inconsistency does not affect the patch similarity result. Therefore, it is plausible to leverage the patch clustering result of high resolution color images into the nonlocal similarity enforcement during the upsampling of low-resolution depth maps. Such new insight distinguishes our approach from the previous ones.

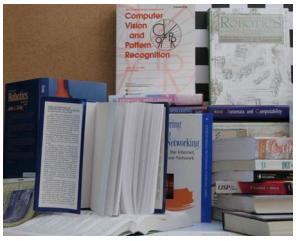
More specifically, we propose to exploit the nonlocal similarity enforcement for depth map upsampling via low-rank methods. Low-rank methods have attracted increasingly more attention in recent years and been successfully applied to various applications including matrix completion [9], [10] and image restoration [11]. Recently, the connection between nonlocal sparse representations and low rank methods was explored for natural images in [11] and further extended by the tool of Graph-based Transform [7] for depth maps. By utilizing this connection, excellent image and depth map denoising results have been obtained. When compared with those works, this paper can be viewed as another extension of the proposed Spatially Adaptive Iterative Singular-Value Thresholding (SAIST) [11] algorithm into image-guided depth map upsampling where the patch dependency among high-resolution depth maps is "learned" from the accompanying color images. Experimental results with the widely used Middlebury database have shown that our method is comparable with previous developed image-guided depth map upsampling methods.

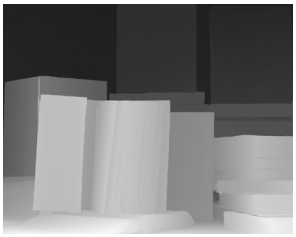
The outline of this paper is as follows. First, we describe the proposed image-guided SAIST method in section II; In section III, we report our experimental results to demonstrate the performance of our method; in the end, we discuss and conclude our work in the last section IV.

Acknowledgment: This work is partially supported by the NSF Award ECCS-1305661.



(a) Low resolution depth map with noise





(b) High resolution image (c) High resolution depth result

Fig. 1: Example of upsample a low resolution depth map: (a) Low resolution depth with noise (273×344) (b) Corresponding high resolution RGB image (1088×1376) (c) Depth map after image-guided SAIST (1088×1376).

II. Algorithm Development

In this section, we develop a depth map upsampling algorithem based on image-guided SAIST. The methodology of this approach can be divided into three steps: (1) Registering the low resolution depth map into the high resolution image coordinate (section II-A), (2) learning nonlocal dictionaries of high resolution image (Section II-B), and (3) importing the dictionaries into SAIST (section II-C).

A. Depth Map Registration

In our setup, we define $X_d = (X, Y, Z, 1)^T$ as the global coordinate of the captured 3D scene, and the TOF camera as the coordinate origin. Each measurement $d_{i,j}$ at low resolution pixel position $x_{ij} = (i, j, 1)^T$ represents the distance between the origin and a single point $X_{i,j,k}$ at the global coordinate X_d. To be noticed, this measurement is an averaged depth value of multiple points around $X_{i,j,k}$, and the entire depth map is treated as a projection of the 3D scene. We then re-project these measurements into corresponding coordinate on the high resolution image space Ω_H by calculating

$$\tilde{x}_{i,j} = P_H d_{i,j} P_L^{\dagger}, \tag{1}$$

where $P_L^{\dagger}$ is the pseudo-inverse of the low resolution projection matrix, and P_H is the projection matrix for the high resolution depth map.

Therefore, a new depth map has been created with sparse set of depth values at position $\tilde{x}_{i,j}$. Since the global minimum of upsampling result remains unknown in terms of MSE, we interpolate the missing pixels between known positions to initiate the algorithm to find a better local minimum. We name the initialized depth map as y.

B. Dictionary Learning

We re-evaluate the strength of patch-based image processing by investigating nonlocal similarity properties under the context of upsampling. Previous studies on nonlocal upsampling for magnetic resonance image [12] and depth map [6] confirm the advantages of relying on nonlocal properties. Therefore, It's plausible that exploitation of nonlocal similarity by patch clustering could lead to fine details and structures.

Instead of collaborating patch clustering directly on the depth map, we apply it under the guidance of high resolution color image by leveraging the fact that depth map and the corresponding color image are related. We can view this relationship subjectively by comparing a pair of depth map and RGB color image to observe their structural similarity. Discard the fact that depth map intensity values represent the physical distance while intensity values of color images stand for different textures, the remaining content are similar edges/structure information. In Fig.2, we use k-means clustering with patch size of five to show the structural similarity between depth map and color image. To put it another way, structure perimeter information "borrowed" from high resolution image will gradually increase the sharpness and accuracy of upsampled result.

(a) k-means clustering of color (b) k-means clustering of depth

Fig. 2: Comparison of k-means clustering between depth map and color image of middlebury *Books*.

Our upsampling method increases the low resolution of captured depth map by making the best use of the strategies described above. For a given $N \times M$ high resolution image I and initialized depth map y, we extract all patches separately and formulate two matrices by converting every patch as a single column, named as M_I and M_y. Both matrices are highly correlated where the corresponding columns are given the same labels by their position. Then, for each column in M_I (exemplar patch), we search for its KNNs by comparing its Euclidean distances with all other patches inside searching window. Next, we construct our nonlocal dictionary D by stacking the label of exemplar and the labels of its KNN as columns.

C. Spatially Adaptive Iterative Singular-Value Thresholding

We first overview simultaneous sparse coding as a standard low-rank approximation problem [11]. Sparse coding means

that a patch can be represented by a learned dictionary $\mathbf{U}_i$ and a collection of sparse weight vectors α_i, as $y_i \approx \mathbf{U}_i\alpha_i$. Therefore, a minimization problem for different image restoration tasks can be written as

$$(\mathbf{U}_i, \alpha_i) = \underset{\mathbf{U}_i,\alpha_i}{\operatorname{argmin}} \|y_i - \mathbf{U}_i\alpha_i\|_2^2 + \tau\|\alpha_i\|_0, \tag{2}$$

where τ is a Lagrange multiplier serving as a regularization parameter. However, regardless the assumption of independence between patches, the idea of group sparse coding is to exploited dependence among all patches by encoding the exemplar and its KNN jointly using the same learned dictionary $\mathbf{U}$:

$$(\mathbf{U}, \mathbf{A}) = \underset{\mathbf{A}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}\mathbf{A}\|_F^2 + \tau \sum_{i=1}^{N} \|\alpha_i\|_1^2. \tag{3}$$

We use the predefined dictionary D to find all KNNs and further write equation 3 using low-rank approximation as

$$(\mathbf{U}, \Sigma, \mathbf{V}) = \underset{\mathbf{U},\Sigma,\mathbf{V}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{U}\Sigma\mathbf{V}^T\|_F^2 + \tau \sum_{i=1}^{K} \lambda_i, \tag{4}$$

where,

$$(\mathbf{U}, \Sigma, \mathbf{V}) = \mathsf{svd}(\mathbf{Y}). \tag{5}$$

In addition, iterative regularization has been studied to update the result. The original idea of iterative regularization is to add filtered noise back to the image as

$$\hat{y}^{(k+1)} = \hat{y}^k + \delta(y - \hat{y}^k), \tag{6}$$

where y is the input noisy low resolution depth map, $\hat{y}^k$ is the output of the k-th iteration and δ is a relaxation parameter. However, instead of using this technique on the upsampled depth map, signal variance and noise variance will be estimated and updated alternatively through

$$\hat{\sigma}_w^k = \gamma\sqrt{\sigma_w^2 - \|y - \hat{y}^{(k+1)}\|_{l2}}, \tag{7}$$

$$\hat{\sigma}_i^{(k+1)} = \sqrt{\max((\hat{\lambda}_i^k)^2/m - (\hat{\sigma}_w^k)^2, 0)}. \tag{8}$$

As a result, threshold/regularization parameter can be iteratively determined by $\tau = 2\sqrt{2}\sigma_w^2/\sigma_i$ [13] and applied to soft thresholding $\hat{\Sigma} = \mathbf{S}_\tau(\Sigma)$. The new data matrix can be reconstructed by $\hat{\mathbf{Y}} = \mathbf{U}\hat{\Sigma}\mathbf{V}^T$.

With the help of iterative regularization, we usually observe the Mean-Square-Error (MSE) of upsampled depth map monotonically decreases and disclose more details. A step-by-step description of image-guided depth map upsampling algorithm is given in Algorithm 1.

III. Experiment Result

In this section, we present our experimental results with image-guided SAIST upsampling algorithm developed in Section II. We compare against several competing upsampling methods, and further demonstrate its performance by showing its ability in depth map upsampling.

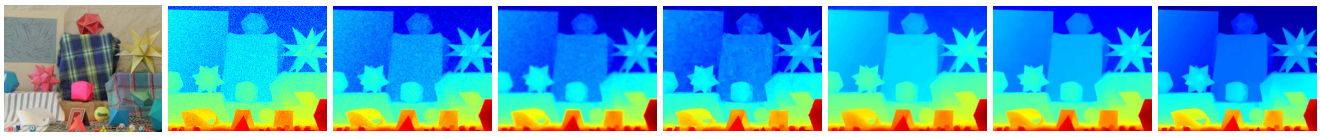
Algorithm 1 Depth Map Upsampling via Image-Guided SAIST

1: **Input**: One low resolution depth map and one high resolution image I;
2: **Input Initialization**: Register low resolution depth map on high resolution image space and create y;
3: **Dictionary Learning**: Find the KNN for each exemplar patch in M_I and extract their labels to create the dictionary D;
4: **Initialization**: $\hat{y}^{(1)} = y$;
5: **for** $k = 1$ to $iter$ **do**
6: Patch clustering: Search D for the KNN labels of the i-th exemplar patch of y, and create data matrix $\mathbf{Y}_i$ by stacking corresponding columns from M_d;
7: SVD for each data matrix $\mathbf{Y}_i$: ($\mathbf{U}_i$,Σ_i,$\mathbf{V}_i$) = svd($\mathbf{Y}_i$);
8: Signal variance and noise variance update: update both variance use Eqs. (7) and (8);
9: Thresholds update: $\tau = 2\sqrt{2}\sigma_w^2/\sigma_i$;
10: Singular value thresholding: $\hat{\Sigma} = \mathbf{S}_\tau(\Sigma)$;
11: Depth map update: Obtain improved depth map by combining all reconstructed data patches;
12: **end for**
13: **Output**: The upsampled depth map $\hat{y}^k$.

A. Evaluations using the Middlebury stereo dataset

We evaluate our image-guided SAIST algorithm with three Middlebury stereo depth maps *Art*, *Moebius* and *Books*. We downsample these depth maps by different upsampling factors (*e.g.* $\times 2, \times 4$) to create low resolution depth maps. More specifically, we have a set of low resolution depth maps, three high resolution RGB color images and three high resolution depth maps as groundtruth. We first demonstrate the strength of iterative regularization. In Fig.3, we calculate the PSNR of the input y (*Art*) to indicate the starting point (red baseline). Two user adjustable parameters, patch size and number of similar patches, are varied to demonstrate the impact of nonlocal properties. For depth map *Art*, large patch size and increased number of similar patches will cause misleading on obtaining accurate structure information and patch locations, which leads to sub-optimal results. We can easily observe that the depth map is greatly intensified after the first iteration and improves continuously as iteration progresses. Notice, in Fig.3 upsampled results tend to converge after several iterations.

Base on the iteratively updated results, we compare with bilinear interpolation, MRF by Diebel and Thrun [3], guided image filtered approach by He *et al.* [4], non-local means filtering by Park *et al.* [6] and anisotropic total generalized variation approach by Ferstl *et al.* [5]. The numerical results of these experiments in terms of Peak Signal-to-Noise Ratio (PSNR) are presented in Table I. These experiments provide an objective comparison on the effectiveness and accuracy of different algorithms. It is easy to observe that our algorithm produces the leading results, which achieved more than one dB gain for upsampling ratio of two. For upsampling ratio



(a) RGB (b) Input Depth (c) Diebel (d) Chan (e) Park (f) Ferstl (g) OURS (h) Groundtruth

Fig. 4: Subjective quality comparison of $\times 4$ upsampling of middlebury *Moebius* dataset with added AWGN noise. (a) RGB high resolution color image. (b) Low resoltuion noisy input image. (c) Upsampling using MRF approach [3]. (d) Adaptive bilateral upsampling approach [2]. (e) Nonlocal means upsampling approach [6]. (f) Anisotropic total generalized variation [5]. (g) Our Upsampling result using image-guided SAIST. The results in (c), (d) and (e) are still suffer from noise. (f) removes majority of the noise and sharp edges around small structure. Our algorithm removes noise and retains sharp edges.

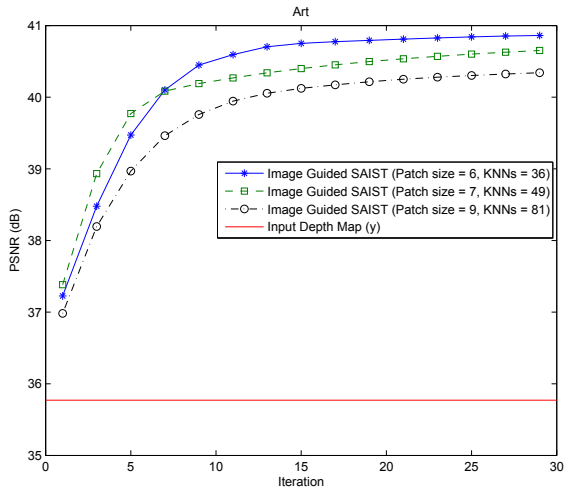


Fig. 3: The numerical (PSNR) results of iteratively upsampled depth map *Art*. Patch size and number of similar patches are varied to indicate the effectiveness of regular nonlocal parameters.

	No Noise					
	Art		Books		Moebius	
	x2	x4	x2	x4	x2	x4
Bilinear	37.36	32.86	45.19	40.48	46.19	41.11
Diebel *et al.* [3]	38.25	36.55	46.51	44.34	46.64	44.97
He *et al.* [4]	38.78	36.56	46.82	44.20	47.34	45.00
Park *et al.* [6]	39.08	37.25	46.58	44.63	47.59	45.53
Ferstl *et al.* [5]	38.50	36.57	45.92	44.03	47.07	44.85
OURS	40.86	37.30	47.38	41.94	48.82	45.98

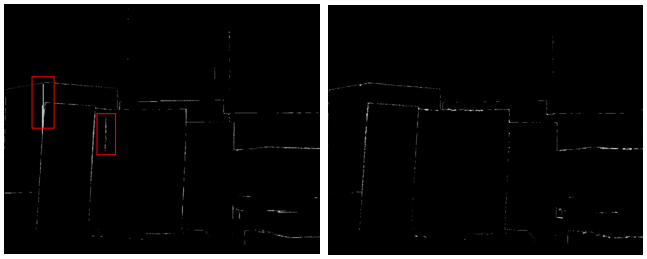
TABLE I: Objective quality comparison results without noise. The performance is measured as PSNR(dB) for two different factors ($\times 2, \times 4$).

of four, our algorithm is still capable of producing better results on *Art* and *Moebius*. For *Books*, we compare the error maps in Fig.5. We observe overall a less error density, but there are two relatively large errors (highlighted in Fig.5a). The possible explanation is that our algorithm misjudged the structural information at these particular locations and kept reinforcing these nonexistent edges. Next, we compare these

	Noise					
	Art		Books		Moebius	
	x2	x4	x2	x4	x2	x4
Bilinear	34.93	33.14	36.22	35.46	35.69	34.95
Diebel *et al.* [3]	37.27	35.04	41.83	38.58	41.57	38.28
He *et al.* [4]	37.13	35.23	40.61	39.38	40.23	39.09
Park *et al.* [6]	36.63	34.94	42.34	39.80	42.30	40.14
Ferstl *et al.* [5]	38.06	35.95	44.48	41.23	44.75	41.98
Chan *et al.* [2]	37.40	35.13	41.72	39.27	41.78	39.31
OURS	37.90	33.41	43.29	38.76	45.12	40.88

TABLE II: Objective quality comparison results with noise. The performance is measured as PSNR(dB) for two different factors ($\times 2, \times 4$).

results with the outcome of noisy depth map upsampling in Section III-B to demonstrate the ability of our algorithm over noisy depth maps.



(a) OURS (b) Park

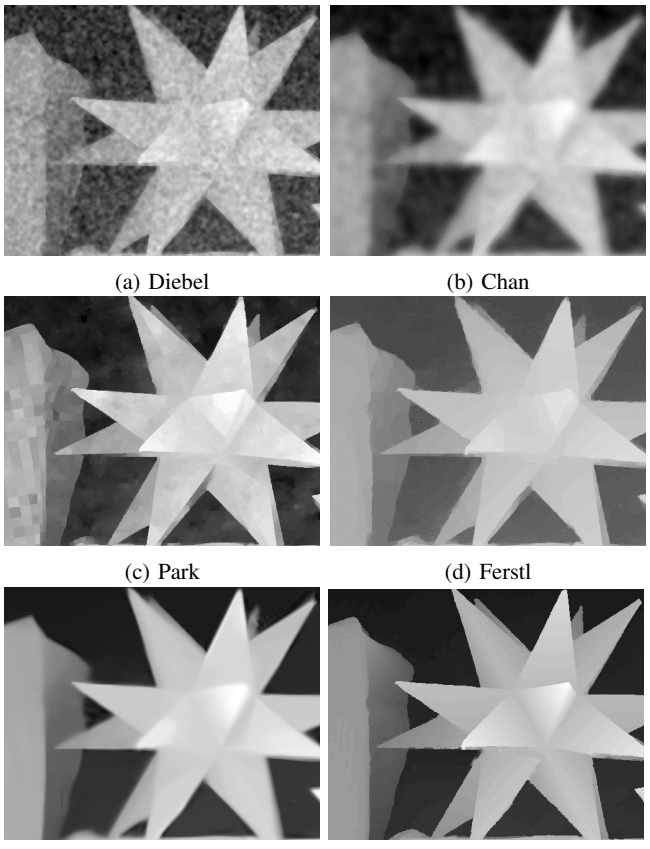
Fig. 5: Error map comparison of *Books* with upsampling factor of four between our algorithm and nonlocal means upsampling approach by Park *et al.* [6].

B. Evaluation with synthetic noise

To simulate the acquisition process, additive white Gaussian noise is added to all low resolution depth maps. Although the real noise distribution is more sophisticated than Gaussian model, we implement it only for a fair comparison with previously developed algorithms. In order to further demonstrate the ability of our algorithm against noise, we compare all results to the noise-aware bilateral filter approach by Chan

Please refer to end of Section III-A for more details.

et al. [2]. The qualitative comparison results are summarized in Table II. We can see that our result is comparable with the competing algorithms, and outperform them for *Moebius* with ratio of two. A subjective quality comparison of different methods is presented in Fig.4 for noisy depth map *Moebius*, and we enlarge the polygon on the right side in Fig.6 for further comparison. It can be observed that the subjective quality of our result exhibit clear structure, smooth surface and suppressed noise level, while others remain noisy. The fact that our numerical and subjective results are not consistent is due to the anti-aliasing edges created by our algorithm which dramatically reduced our numerical results compared with groundtruth. However, the visual results confirm that similar texture in high resolution color image provide desirable references for our algorithm to locate similar patches in depth map. Hence, accurate patch clustering leads to exploit the capacity of group sparsity, which is why our algorithm successfully suppress noise level and retain more structure, compare with other depth map upsampling methods.



(a) Diebel (b) Chan

(c) Park (d) Ferstl

(e) OURS (f) Groundtruth

Fig. 6: Enlarged comparison results of *polygon* inside *Moebius* from Fig. 4.

IV. Conclusion

In this paper, we present a novel depth map upsampling algorithm using a low resolution depth map captured by TOF camera under the guidance of a high resolution color image. Our algorithm is based on the similar structure relation between color and depth map. We formulate a dictionary using similar patch location in color image and pass it to form group sparse coding via low rank approximation. The upsampled depth map also can be iteratively updated with the help of iterative regularization. When testing our algorithm on the well-known dataset, we show our algorithm typically outperform several competing approaches in non-noisy environment. We further demonstrate the ability of our algorithm against noise. The visual results are consistent with the fact that nonlocal patch clustering is effective for reveling details in image domain, and we believe that we have open a new door to acquire high resolution image information for assisting low resolution depth map upsampling. In the future, we plan to use more sophisticated noise models to simulate acquisition process, and test our algorithm on real-world data. We also anticipate to develop a stopping criteria for applications where groundtruth is unknown for upsampled result.

References

[1] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)*, vol. 26, no. 3, p. to appear, 2007.

[2] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noiseaware filter for real-time depth upsampling," in *In Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[3] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.

[4] K. He, J. Sun, and X. Tang, "Guided image filtering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1397–1409, June 2013.

[5] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 993–1000.

[6] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 1623–1630.

[7] W. Hu, X. Li, G. Cheung, and O. Au, "Depth map denoising using graph-based transform and group sparsity," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept 2013, pp. 001–006.

[8] A. Torralba and W. Freeman, "Properties and applications of shape recipes," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, June 2003, pp. II–383–90 vol.2.

[9] E. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[10] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[11] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: A low-rank approach," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 700–711, Feb 2013.

[12] J. V. Manjón, P. Coupé, A. Buades, V. Fonov, D. Louis Collins, and M. Robles, "Non-local mri upsampling," *Medical image analysis*, vol. 14, no. 6, pp. 784–792, 2010.

[13] S. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *Image Processing, IEEE Transactions on*, vol. 9, no. 9, pp. 1532–1546, Sep 2000.