DEPTH IMAGE SUPER-RESOLUTION USING INTERNAL AND EXTERNAL INFORMATION

H. Zheng, A. Bouzerdoum, Senior Member, IEEE, and S. L. Phung, Member, IEEE

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong, NSW 2522, Australia

ABSTRACT

The fast development of 3-D imaging techniques has increased demands for high-resolution depth images. Conventional depth super-resolution methods reconstruct the high-resolution image by accessing high frequency information, either internally from a high-resolution intensity image or externally from a high-resolution image database. In this paper, a new depth super-resolution method based on joint regularization is proposed, which exploits both internal and external high frequency information. Specifically, a joint regularization problem with different constraints is formulated, which allows us to solve for the high-resolution image and a sparse code simultaneously. These constraints are constructed by utilizing information from both internal and external high-frequency sources. Experimental evaluation suggests that the proposed method provides improved results over existing approaches, in terms of both visual appearance and objective image quality.

Index Terms— depth super-resolution, sparse representation, non-local constraint, local constraint, joint regularization

1. INTRODUCTION

Image super-resolution techniques have been widely used to enhance the resolution of intensity images. Currently, there exists a great demand for high-resolution (HR) depth images since they are extremely useful for constructing high quality 3D scenes. In practice, capturing high-resolution depth images directly using a camera is either too expensive or time consuming. Therefore, to reduce cost, high-resolution depth images are constructed from a collection of low-resolution (LR) depth versions using super-resolution (SR) methods. Conventional depth image SR methods can be divided into three categories: single-frame, multi-frame and learning-based.

The single-frame SR approach utilizes a registered HR intensity image to help reconstruct the HR depth image. Diebel *et al*. applied Markov Random Fields to estimate the missing pixels in the depth image from the registered HR intensity image [1]. Kopf *et al*. modified the classic bilateral filter [2] into a joint bilateral filter which captures the detail information from the HR intensity image for up-sampling the LR depth image [3]. In [4], a more generalized guided filter was proposed. A linear relationship between the input image and the registered image is established; the coefficients of this relationship are estimated for reconstructing the target image. The guided filter can be used to perform a series of image processing tasks, including image matting, dehazing, denoising and super-resolution.

The multi-frame depth SR approach requires different LR depth images of the same scene taken from slightly displaced vantage points. Ronsenbush *et al*. proposed an SR method that aligns LR depth images on an HR grid and then interpolates the missing pixels [5]. Taking advantage of the robust performance of SR using L1-norm regularization [6], Schuon *et al*. extended the L1-norm regularization to depth image SR [7]. In [8], Schuon *et al*. proposed a new depth SR method called LidarBoost, which combines the data fidelity term and a sum-of-gradient-norm regularization term to form a new regularization problem. In [9], Hu *et al*. proposed a new SR method that utilizes the stereo view of the target depth image to construct a regularization problem.

The learning-based SR approach utilizes high-frequency information from an HR image database to reconstruct the target depth image. This approach is derived from intensity SR methods proposed in [10]. To the best of our knowledge, only a few methods explore the effectiveness of adopting external database to reconstruct HR depth images. Li *et al*. proposed a joint depth SR (J-DSR) method, where different HR patches are reconstructed using a registered HR intensity image and a single dictionary [11]. Zheng *et al*. proposed a multi-dictionary based depth SR method, where each HR patch is reconstructed from its own dictionary [12].

Both single-frame and learning-based SR approaches utilize additional high-frequency information to reconstruct HR images. The difference is that single-frame SR approach accesses the high-frequency information from an HR intensity image, which can be regarded as an internal information source. A learning-based SR approach, on the other hand, relies on the high-frequency information from an external HR depth image database to reconstruct the HR image. In this paper, we propose a new depth SR method by formu-

lating a joint regularization problem to solve for the HR image and the sparse representation simultaneously; constraints in this regularization problem are built from both internal and external information sources. Therefore, the proposed method is hereafter referred to as super-resolution using internal and external information (SRIE). Compared to conventional single-frame and learning-based methods, the proposed method achieves higher reconstruction accuracy.

The remainder of the paper is organized as follows. Section 2 introduces the joint regularization SR problem, followed by the formulation of the local and non-local constraints. Section 3 presents the experimental evaluation and comparisons of SRIE with other SR methods. Section 4 provides concluding remarks.

2. DEPTH SUPER-RESOLUTION BASED ON JOINT REGULARIZATION

This section describes the proposed joint regularization based depth SR method. In Subsection 2.1, the SR problem is formulated as a joint optimization over the target HR image and a sparse code. Then, in Subsection 2.2, the local and non-local regularization constraints are constructed from the registered HR intensity image and the depth LR image, respectively.

2.1. Joint sparse coding and image reconstruction

Given an LR depth image, we aim to reconstruct the HR depth image with an external HR depth database. Let $Y \in \mathbb{R}^{N\times 1}$ and $X \in \mathbb{R}^{M\times 1}$ be lexicographically ordered vectors representing the input LR and the target HR depth images. The degradation model for the super-resolution problem can be formulated as

$$Y = CX + V, \tag{1}$$

where C is a degradation matrix of size $N \times M$, containing blurring and downsampling operations, and V denotes zero-mean additive white Gaussian noise. Similarly to other learning-based SR methods [13, 14], we adopt a patch-based approach, where Y and X are partitioned into overlapping patch sets $\{\mathbf{y}_i \in \mathbb{R}^{n\times 1}; i = 1, \ldots, q\}$ and $\{\mathbf{x}_i \in \mathbb{R}^{m\times 1}; i = 1, \ldots, q\}$, respectively. Assuming $\mathbf{x}_i$ can be sparsely represented over an over-complete dictionary, we can express the HR patches as:

$$\mathbf{x}_i = D^h \boldsymbol{\alpha}_i + \mathbf{v}_i^h, \tag{2}$$

where D^h is a high-resolution dictionary of HR patches extracted from the image database, $\boldsymbol{\alpha}_i$ is a sparse vector, and $\mathbf{v}_i^h$ represents additive noise. Combining all q HR vectors $\mathbf{x}_i$ by stacking them together yields a composite HR vector $\mathbf{x} \in \mathbb{R}^{qm\times 1}$, which can be represented as

$$\mathbf{x} = \mathbf{D}^h \boldsymbol{\alpha} + \mathbf{v}^h \tag{3}$$

where $\boldsymbol{\alpha}$ is a single sparse vector comprising the q sparse codes $\boldsymbol{\alpha}_i$, $\mathbf{v}^h$ is a noise vector, and $\mathbf{D}^h$ is a block diagonal dictionary, containing the HR dictionaries on the main diagonal. Note that although the composite vector $\mathbf{x}$ and the HR vector X are of different sizes, due to the overlap between the patches, they possess the same image content. Define a linear operator $\Gamma \in \mathbb{R}^{M\times qm}$ that stitches all the HR patches together and averages the overlapping areas. The vectors $\mathbf{x}$ and X are related by

$$X = \Gamma\mathbf{x} = \Gamma\mathbf{D}^h\boldsymbol{\alpha} + \Gamma\mathbf{v}^h. \tag{4}$$

Next, a feature constraint is constructed based on the input LR image Y. First, the first and second derivative operators are applied horizontally and vertically to the input image to generate four feature images. Each feature image is partitioned into q patches, and the corresponding patches from the four feature images are concatenated to form a single feature vector $\tilde{\mathbf{y}}_i$ $(i = 1, \ldots, q)$. Similarly, training feature vectors are extracted from the feature images obtained by applying the same derivative operators to the training images in the database. Given the feature dictionary $\tilde{D}^l$ constructed from the training feature vectors, the sparse representation of the feature vector $\tilde{\mathbf{y}}_i$ is given by

$$\tilde{\mathbf{y}}_i = \tilde{D}^l\boldsymbol{\alpha}_i + \tilde{\mathbf{v}}_i. \tag{5}$$

Since the first and second derivatives and the downsampling operation are all linear operations, Eqs. (2) and (5) share the same sparse code. In the image domain, the composite version of the feature vector $\tilde{\mathbf{y}}$ is expressed as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{D}}^l\boldsymbol{\alpha} + \tilde{\mathbf{v}}. \tag{6}$$

where $\tilde{\mathbf{D}}^l$ is a block diagonal dictionary containing the feature dictionaries $\tilde{D}^l$ on the main diagonal. Note that the dictionaries D^h and $\tilde{D}^l$ are trained together by adopting the dictionary training method presented in [15].

The joint regularization problem is formed by jointly minimizing the noise term of Eq. (1) and the sparse code under the two constraints constructed from (4) and (6):

$$\min_{X,\boldsymbol{\alpha}} \left\{ \left\|Y - CX\right\|_2^2 + \lambda\left\|\boldsymbol{\alpha}\right\|_1 + \gamma_1\left\|X - \Gamma\mathbf{D}^h\boldsymbol{\alpha}\right\|_2^2 + \gamma_2\left\|\tilde{\mathbf{y}} - \tilde{\mathbf{D}}^l\boldsymbol{\alpha}\right\|_2^2 \right\}, \tag{7}$$

where λ, γ_1, and γ_2 are regularization parameters.

In the regularization problem formulated in (7), the last two regularization terms are built from the external information source. However, internal information can also be utilized to improve the reconstruction accuracy. To this end, local and non-local constraints are proposed in the next subsection, which exploit information from the LR depth image and the registered HR intensity image, respectively.

2.2. Local and non-local regularization terms

Modern range cameras are able to capture simultaneously the depth image and its corresponding intensity image. The HR intensity image contains high-frequency information that can help enhance the resolution of the corresponding LR depth image. Depth images are often dominated by sharp edges surrounded by plain areas. Therefore, the target HR image X may be modeled as

$$X = UX + \boldsymbol{\epsilon}, \tag{8}$$

where U is a filter that smooths out the plain areas while preserving edges, and $\boldsymbol{\epsilon}$ is the residual error. Similarly to the bilateral filtering approach proposed in [2], the pixels in X are represented by a weighted average of their neighboring pixels. Specifically, let $\mathcal{N}(k)$ denote the set of indices of the neighboring pixels of X_k. Then, the pixel X_k is expressed as

$$X_k = \sum_{l \in \mathcal{N}(k)} U_{k,l} X_l + \epsilon_k \text{ subject to } \sum_{l \in \mathcal{N}(k)} U_{k,l} = 1. \tag{9}$$

In conventional bilateral filtering, the weights $U_{k,l}$ are computed from the input image. In SR problems, however, we have no access to the HR depth image. Instead, we calculate $U_{k,l}$ from the input HR intensity image since the edges in both depth and intensity images often co-exist. Let Z_k denote the intensity pixels corresponding to X_k. The weight $U_{k,l}$ is computed as follows:

$$U_{k,l} = \frac{1}{u_k} \exp\left(- \frac{f(Z_k, Z_l)}{h_f^2} \right) \exp\left(- \frac{g(k,l)}{h_g^2} \right), \tag{10}$$

where the functions $f(Z_k, Z_l)$ and $g(k,l)$ compute the intensity and the geometric distances between Z_k and Z_l, h_f and h_g are the smoothing parameters, and u_k is a normalization constant. The weights obtained from Eq. (10) are used to form the k-th row of the weight matrix U. We should note that the k-th row of U also comprises many zero elements, corresponding to the pixel of X whose indices are not in $\mathcal{N}(k)$.

Natural depth images contain redundant information, which can be utilized to regulate the image reconstruction; we use this redundancy to construct a non-local constraint. Let W denote an $M \times M$ matrix representing a low-pass filter. Similarly to (8), the HR depth image X can be modeled as

$$X = WX + \boldsymbol{\nu}. \tag{11}$$

In other words, every pixel is represented as a weighted linear combination of other pixels in the image plus a noise term. The k-th pixel X_k is represented by its neighboring pixels as follows:

$$X_k = \sum_{l \in \mathcal{N}(k)} W_{k,l} X_l + \nu_k \text{ subject to } \sum_{l \in \mathcal{N}(k)} W_{k,l} = 1, \tag{12}$$

The weights $W_{k,l}$ are obtained by computing the similarity between two patches $\mathbf{x}_k$ and $\mathbf{x}_l$, where k and l represent the centers of the two patches. The weight $W_{k,l}$ is obtained from

$$W_{k,l} = \frac{1}{w_k} \exp\left(- \frac{\|\mathbf{x}_k - \mathbf{x}_l\|_2^2}{h^2} \right), \tag{13}$$

where w_k is the normalization factor and h is the smoothing parameter. Compared to $U_{k,l}$, which is generated from computing the difference between the intensity pixel and its neighboring pixel, $W_{k,l}$ is a function of the distance between the target depth patch and its neighboring patch.

The local and non-local constraints in (8) and (11) can be incorporated into (7) as regularizers. The regularization problem becomes

$$\min_{X,\boldsymbol{\alpha}} \left\{ \left\| Y - CX \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 + \gamma_1 \left\| X - \Gamma \mathbf{D}^h \boldsymbol{\alpha} \right\|_2^2 + \gamma_2 \left\| \tilde{\mathbf{y}} - \tilde{\mathbf{D}}^l \boldsymbol{\alpha} \right\|_2^2 + \gamma_l \left\| (I - U)X \right\|_2^2 + \gamma_{nl} \left\| (I - W)X \right\|_2^2 \right\}. \tag{14}$$

The two unknowns X and $\boldsymbol{\alpha}$ can be computed at the image level simultaneously by taking the partial derivative of (14) with respect to X and $\boldsymbol{\alpha}$, and setting the derivative to 0. To facilitate the differentiation, the L1-norm term is replaced by the Huber norm since the latter is continuously differentiable [16].

3. EXPERIMENTAL RESULTS

This section presents experimental results and comparison with other state-of-the-art depth SR methods. The experimental procedure for testing the proposed SRIE algorithm is introduced first in the next subsection.

3.1. Experimental procedures

The proposed algorithm is evaluated on seven test images: Poster and Tsukuba from Middlebury database; Ballet and Dancer from MSR3DVideo database; Tanks, Pyramid, and Book from the synthetic image database of [17]. These depth images are downsampled by a factor of 3 to obtain the input LR images. For each depth image, a registered HR intensity image is also captured. The training image database contains 33 HR depth images from Middlebury database (excluding the test images). This database is used to train the LR-HR dictionary pair (D^l, D^h).

The parameters for constructing the local and non-local regularization terms are defined as follows. For the local regularization term, the geometric smoothing parameter h_g and the intensity smoothing parameter h_f are set to 3 and 10, respectively; the search window is fixed at 7×7. For the non-local regularization term, the search window for comparing patch similarity is set to 13×13 with a patch size of 5×5,

(a) Ground Truth (b) JBU (c) JDSR (d) SRIE-SP (e) SRIE

Fig. 1. Image Tsukuba reconstructed by different SR methods. Image (a) shows the ground truth HR image. Images (b)-(e) are reconstructed by JBU, J-DSR, SRIE-SP, and SRIE.

and the smoothing parameter h is set to 10. Furthermore, the values of the regularization parameters λ, γ_1, γ_2, γ_l and γ_{nl} in Eq. (14) are set to 1, 0.2, 0.2, 0.1, and 0.1, respectively, throughout the experiment.

Two objective measures are used to evaluate the quality of the reconstructed images: *peak signal-to-noise ratio* (PSNR) and *relative error rate* (RER). Here we briefly introduce the RER measure. Let X be the reference (ideal) image and X^* the reconstructed image. RER maps the relative errors of the reconstructed depth pixels to the range $[0, 1]$, using an exponential function:

$$\mathrm{RER} = \frac{1}{N_0} \sum_{i=1}^{N_0} \exp\Big(- \beta_0 \frac{|X_i - X_i^*|}{X_i} \Big), \qquad (15)$$

where β is a positive parameter that controls the decay rate of the exponential function; in the following experiment, β is fixed at 500.

3.2. Comparison with other depth SR methods

The proposed SRIE method is compared with other depth SR methods, namely bicubic interpolation (BI) [18], joint bilateral upsampling (JBU) [3], and joint learning-based depth SR (J-DSR) [11]. For both JBU and J-DSR, which involve bilateral filtering, the search window is set to 7×7 pixels, and the parameters h_s and h_g are set to 10 and 3, respectively; these setups are the same as those of the proposed SRIE method. In addition, the performance of two versions of the proposed algorithm are also investigated: SRIE with only the sparse reconstruction term (SRIE-SP) (7) and SRIE with all three terms (14). Note that SRIE-SP only exploits high-frequency information from the image database.

Table 1 shows the PSNRs and RERs of the reconstructed images, where the bold values indicate the largest PSNR and the smallest RER. For all test images, the proposed SRIE generates the highest PSNRs and the lowest RERs. Figures 1 shows the HR depth images Tsukuba reconstructed by different SR methods. These figures clearly demonstrate that the proposed SRIE method produces sharper and clearer edges with fewer artifacts.

Table 1. PSNR (dB) and RER (%) performances for different SR methods: JBU, J-DSR, the proposed SRIE-SP and SRIE. For each test image, the top line shows the PSNR, while the bottom line shows the RER.

	JBU	J-DSR	SRIE-SP	SRIE
Poster	42.23	47.64	47.52	**48.06**
	3.45	1.94	1.97	**1.47**
Tsukuba	31.28	34.39	34.30	**34.82**
	12.65	8.66	8.94	**7.43**
Ballet	36.24	39.71	39.35	**40.14**
	5.12	3.81	4.25	**3.39**
Dancer	43.22	45.25	45.04	**45.46**
	2.87	1.88	1.90	**1.86**
Tanks	40.85	42.30	42.14	**42.54**
	8.17	6.11	6.34	**5.76**
Pyramid	38.77	41.96	41.94	**42.68**
	3.79	2.96	3.25	**2.39**
Book	44.53	49.57	49.38	**49.88**
	1.38	0.80	0.79	**0.68**
Average	39.59	42.97	42.81	**43.37**
	5.35	3.74	3.92	**3.28**

4. CONCLUSION

In this paper, we proposed a new depth super-resolution method exploiting internal and external information from the input depth image, the registered HR intensity image, and an image database. A joint regularization based SR problem is formulated with different regularization terms. Specifically, the sparse reconstruction term is formed from a depth image database. The local and non-local regularization terms are built from a high-resolution intensity image and the input depth image. In the experimental stage, the proposed SRIE method was compared to several *state-of-the-art* depth super-resolution methods. The experimental results confirm the superiority of the proposed method.

5. REFERENCES

[1] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 291–298, 5–10 Dec. 2005.

[2] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *International Conference on Computer Vision (ICCV)*, pp. 839–846, 1998.

[3] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transcation on Graphics*, vol. 26, no. 3, pp. 96, 2007.

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

[5] G. Rosenbush, H. Tsai, and R. D. Eastman, "Super-resolution enhancement of flash LADAR range data," *Proceedings of SPIE*, pp. 1–12, 2007.

[6] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[7] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–7, 24–26 Jun. 2008.

[8] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 343–350, 20–25 Jun. 2009.

[9] W. Hu, G. Cheung, X. Li, and O. Au, "Depth map super-resolution using synthesized view matching for depth-image-based rendering," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 605–610, 9–13 Jul. 2012.

[10] W. T. Freeman and E. C. Pasztor, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000.

[11] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 152–157, 9–13 Jul. 2012.

[12] H. Zheng, A. Bouzerdoum, and S. L. Phung, "Depth image super-resolution using multi-dictionary sparse representation," in *IEEE International Conference on Image Processing (ICIP)*, pp. 957–961, 15–18 Sep. 2013.

[13] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 24–26 Jun. 2008.

[14] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[15] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808, 6–9 Dec. 2007.

[16] P. J. Huber, "Robust regression: Asymptotics, conjectures, and monte carlo," *Annals of Statistics*, vol. 1, no. 5, pp. 799–821, 1973.

[17] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *European Conference on Computer Vision Conference on Computer Vision (ECCV)*, pp. 510–523, 5–11 Sep. 2010.

[18] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.