Resolution Enhancement in Single Depth Map and Aligned Image

Yang Xian
The Graduate Center
The City University of New York
yxian@gradcenter.cuny.edu

Yingli Tian
The Graduate Center and City College
The City University of New York
ytian@ccny.cuny.edu

Abstract

Depth resolution enhancement aims to recover a high quality depth map from one or multiple low-resolution depth input(s) with missing pixels. While a registered high-resolution intensity image is often utilized to assist, little attention has been paid to the circumstances when there is only one pair of low-resolution depth map and aligned intensity image available.

In this paper, we propose a novel resolution enhancement approach that targets at improving the quality of both the input depth map and the low-resolution RGB image. By exploiting the statistical dependency between the input pairs, a label matrix is generated utilizing the support vector machine classifier. Guided by the constructed label matrix and the aligned intensity image, the missing values in the depth map are well predicted in a manner consistent with the embedded structure. After that, the completed depth map and the intensity image are super-resolved through a set of regression models trained via external exemplars. Extensive experiments demonstrate that our framework is effective with satisfying performance.

Figure 1. Enhancement results of 'cone' (partial) ($\times 4$). Left column present the low-resolution depth map and the intensity image while the right column are the corresponding high-resolution results generated by the proposed enhancement framework. The LR instances are upsampled by nearest-neighbor interpolation for a better illustration. The figure is better viewed on screen with HR display.

1. Introduction

Depth maps are convenient in representing and storing the distance information of the objects' surfaces given a viewpoint. They can be easily obtained through 3D imaging hardware such as time-of-flight (TOF) cameras and cost-effective consumer RGB-D cameras (e.g., Microsoft Kinect camera). Quality of the captured depth maps are crucial in their relevant applications, e.g., reliable 3D reconstruction, accurate human pose recognition, proper semantic scene analysis, and other geometry-related computer vision systems. However, due to the limitations of the depth sensors, depth maps suffer from low spatial resolution especially when the objects are far from the camera. Moreover, missing depth values exist due to the short distance between the object and the depth camera, disparity between the projector and the sensor, or poor reflection of the light patterns [28]. Under these circumstances, we reply on computer vision algorithms to enhance the quality of the depth maps.

Image super-resolution (SR) aims at estimating a fine-resolution image from one or multiple coarse-resolution images. Depending on the number of input images, image SR can be broadly divided into multi-image SR and single image SR. Single image SR methods can be further classified as interpolation-based, reconstruction-based, and learning-based. Commonly used interpolation-based approaches, e.g., nearest-neighbor, bilinear, and bicubic, are simple and efficient. But visual artifacts such as jaggies and blurring exist in the generated high-resolution (HR) results. To suppress the visual artifacts and generate results with sharper edges, more sophisticated interpolation-based methods [30, 40, 39] were proposed. Reconstruction-based image upscaling methods [8, 27, 31] aim to enforce certain statistical priors during the estimation of the target image.

This group of SR approaches emphasize on restoring sharp edges but tend to be less effective in hallucinating rich texture regions. Recently popularized learning-based SR explores the relationship between HR and their corresponding low-resolution (LR) exemplars either via an external dataset [36, 32, 35, 6] or within the input image [10, 9, 13].

Depth maps can be viewed as gray scale images where each pixel stores the depth information. Different from research in intensity image SR, single depth SR is not that commonly seen. From the aspect of input, research in depth image resolution enhancement mainly falls into two classes: multi-frame SR and depth image SR with the assistance of an aligned HR RGB image. Multi-frame SR methods make use of the presence of aliasing in multiple depth inputs of the same scene to produce one fine-resolution depth map. Schuon *et al.* [25] verified that multi-frame SR designed for intensity images also function in the 3D domain. They applied image SR scheme proposed in [7] to depth images taken by a 3DV^{TM} TOF Camera. Campbell *et al.* [3] adopted a discrete label Markov Random Fields (MRF) optimization to pose a spatial consistency constraint in extracting the fine depth map based on stored depth hypotheses. In [26], Schuon *et al.* incorporated a data fidelity term and a geometry prior term into an optimization framework. The former constraint ensures the fidelity between the HR depth map and the LR measurements and the later term guides the energy minimization to a plausible solution. Later, Cui *et al.* [5] proposed a probabilistic scan alignment approach to fuse noisy scans into high quality 3D shapes. Real-time GPU-based algorithm was designed in [14] to merge LR images captured by Kinect and accomplish 3D reconstruction.

With the presence of an aligned HR RGB image, the second category of depth image SR tends to jointly use both depth and color information of the same scene. Statistical dependency exists between the registered intensity and depth images based on the observation that depth discontinuities often co-occur with intensity changes. In [37], Yang *et al.* utilized the HR intensity image to build the cost volume and iteratively refined the input LR range image. Park *et al.* [21] introduced nonlocal means filtering to regularize depth maps during the reconstruction and the HR intensity input provides additional features to better preserve the structure. Li *et al.* [20] utilized piece-wise planar assumption to regulate global geometry of the scene and proposed a Bayesian approach by taking the uncertainty of depth measurements into consideration. Kiechle *et al.* [16] presented a bimodal co-sparse analysis model to capture the interdependency of registered intensity and depth information. In [19], a unified framework is proposed to combine multiple constrains where the aligned HR intensity image could be incorporated as an additional term if available.

To obtain high-quality HR depth maps, missing values in the input depth map need to be filled using image inpainting techniques. Image inpainting aims at predicting the missing pixel values with the known regions and to successfully replicate visually plausible background textures. In the depth domain, other than generating visually plausible results, the predicted depth values should be accurate in a manner consistent with the registered intensity images if available. Shen *et al.* [28] proposed a probabilistic model to capture various types of uncertainties in the depth measurement. Depth layers are utilized to achieve a depth correction and completion process where the layer labels are obtained through solving a maximum-a-posteriori estimation problem. In [16], with the assistance of an aligned HR RGB image, an algorithm for simultaneous performing depth map SR and inpainting is presented.

Among numerous work in depth image quality enhancement, little attention has been paid to the situation where only one pair of registered LR RGB image and depth map is available. Lee and Lee [18] employed a convex optimization framework for simultaneous estimation of super-resolved depth map and intensity image but required LR depth map sequences as inputs. In this paper, we propose a novel sequential resolution enhancement framework which takes only one pair of aligned LR depth and intensity images as input to obtain both HR intensity image and depth map. By exploiting the statistical dependency between the input pair, a label matrix is learnt through a support vector machine (SVM) classifier to differentiate the foreground objects and the background scene. Guided by the aligned LR RGB image and the constructed label matrix, the missing values in the LR depth map are accurately predicted. Afterwards, the HR depth map and intensity image are recovered through a set of pre-built regression models learnt from external exemplars. Fig. 1 shows the HR depth map and intensity image generated by the proposed resolution enhancement framework over 'cone' in dataset [23] under the magnification factor of 4. Due to the space limit, in order to illustrate details more clearly, only part of the image is presented. As observed, missing pixels in the input depth map are correctly predicted consistent with the structure revealed by the registered intensity image. After resolution enhancement, the HR depth map and intensity image have finer details including sharper edges and richer textures.

Contributions of the proposed framework are fourfold:

- A novel image quality enhancement system is proposed to handle circumstances where only one pair of registered LR intensity and depth images are available. Moreover, the proposed framework can be flexibly disassembled to solve different tasks, i.e., depth completion, single depth/intensity SR.
- We present a novel depth inpainting scheme by exploiting the statistical dependency between the aligned intensity image and the depth map.
- An external example-based learning pipeline is adopt-

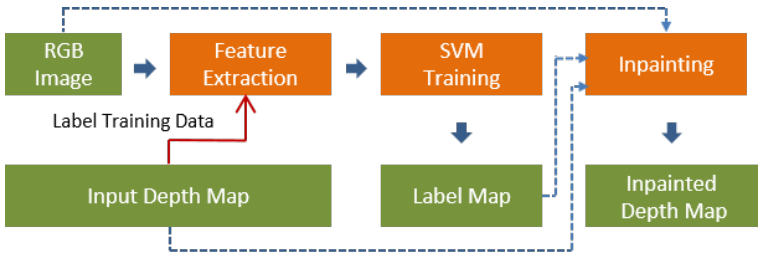

Figure 2. Schematic flowchart of the proposed depth completion process. Pixel-level features are extracted based on the intensity image. The extracted features serve as the inputs to a SVM classifier which is learnt via the training data labeled by the available depth values in the input depth map. A guided inpainting is then performed over the input depth map with the assistance of the label map generated by the SVM model and the aligned RGB image.

ed to train a set of regression models that could be applied to super-resolve LR instances.

- The proposed framework generates visually-pleasing results and outperforms the representative peer methods both quantitatively and qualitatively.

The rest of this paper is organized as follows. Section 2 provides a detailed description of the proposed resolution enhancement system. Section 3 presents the experimental results and discussions. The conclusions including the future work are listed in Section 4.

2. Resolution Enhancement

Provided with a pair of registered LR depth map and RGB image, we aim at recovering their fine-resolution correspondences with missing depth values predicted. The depth completion and SR processes are performed in a sequential manner. A completed LR depth map is first recovered where all missing depth pixels are predicted guided by the aligned intensity image. Traditional inpainting algorithms focus on reconstructing visually-plausible images and in depth domain, a precise structure consistent with the intensity image also matters. Therefore, a label map which differentiates the foreground objects from the background scene is constructed through a SVM classifier. The input of the SVM model is pixel-level feature which encodes the local color and texture characteristics extracted from the intensity image. The training samples of the classifier are labeled by the available depth information. Afterwards, with the assistance of the generated label map, a targeted inpainting is performed to predict the missing depth values.

The HR depth map and RGB image are then recovered by feeding the completed depth map and LR intensity image into a group of pre-built regression models. To ensure an effective regression learning, the Gaussian Mixture Models (GMM) are adopted to model the feature space and the training is performed in a divide-and-conquer way. Details are presented in the two subsections below.

2.1. Depth Completion

Given an input depth map D_l and its aligned LR RGB image I_l, we first recover a completed depth map D_{lc} where the missing pixel values in D_l are predicted with the guidance of I_l. Fig. 2 presents the flowchart of the depth completion process. In order to preserve the correct structure during inpainting, a label map is first generated utilizing a SVM classifier.

The pixel-level features are extracted through Gabor filters and local homogeneity model to encode both color and texture information. Gabor filters have been successful in a variety of image processing related applications including image segmentation [15, 33] and texture classification [11, 2]. In the spatial domain, a 2D Gabor filter is a 2D Fourier basis function multiplied by an origin-centered Gaussian function, defined as:

$$G(x,y) = \exp(-\frac{x^2+y^2}{\sigma^2})\exp(2\pi\theta i(x\cos\phi + y\sin\phi)) \quad (1)$$

where θ represents the spatial frequency, ϕ stands for the corresponding orientation, σ indicates the standard deviation of the Gaussian kernel.

Gabor features are constructed from responses of Gabor filters by utilizing multiple filters on different frequencies and orientations. In our implementation, we apply the Gabor filters to the luminance channel of the color image. Filters of 5 scales at 8 orientations are adopted and the complex responses are expanded into real and imaginary parts respectively.

The second portion of the pixel-wise feature is extracted according to the local homogeneity model. The color image is first transformed from RGB to CIE Lab color space. For each component, we extract pixel-level feature which encodes the local intensity information. Based on component $i(i \in L, a, b)$, matrices g^i and d^i are computed where g^i stands for the gradient magnitude and d^i represents the standard deviation for each pixel within a $a \times a$ neighborhood centered at it. g^i and d^i are then normalized to range

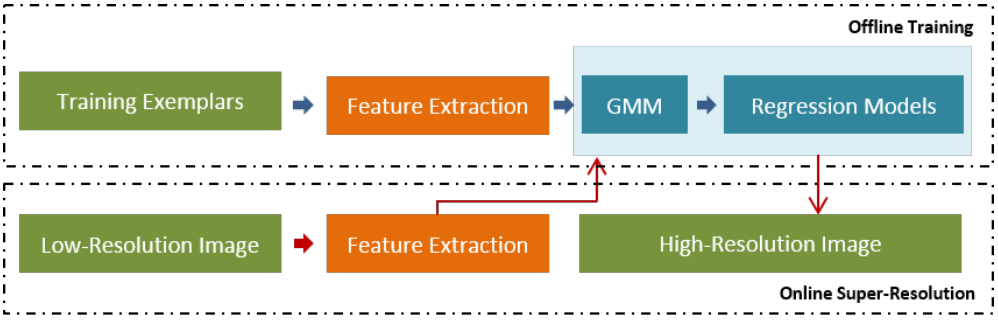

Figure 3. Schematic pipeline of the proposed SR framework. To ensure a targeted learning, the input feature space extracted from an external dataset is modeled with Gaussian Mixture Models. Within each Gaussian component, a regression model is generated. Given a LR instance, the corresponding HR instance is constructed through the regression models.

between 0 and 1 respectively. The normalized matrices are denoted as $\overline{g^i}$ and $\overline{d^i}$. For pixel j under component i, the extracted feature $f_j^i = 1 - \overline{d_j^i} \cdot \overline{g_j^i}$ in which $\cdot$ represents the dot product of two matrices. The final feature is represented as the concatenation of the features extracted from the three components along with the Gabor responses.

The extracted features are employed as inputs of a SVM classifier. To train the SVM model, we first label a portion of the pixels in the intensity image utilizing the available depth information. Generally speaking, the foreground objects are closer to the camera and the background scene has larger absolute depth values. Therefore, after eliminating the pixels whose depth values are unavailable, we annotate the N_1 pixels with the largest depth values with label l_1 and the N_2 pixels with the smallest depth values with label l_2. The annotated data serves as the training samples for the SVM model. We then apply the trained SVM classifier to predict the labels of all the remaining pixels including those whose depth values are missing. The labeling results including the previous training labels are saved as a label map L_d utilized to assist the depth completion process.

We denote the unknown region(s) in D_l as mask M. The completion order of the masked depth pixels is calculated similar to [34] illustrated as follows:

$$Pri(i) = C(i) \cdot T(i) = \frac{\sum_{j \subseteq \{P(i) \cap \overline{M}\}} C(j)}{b^2} \cdot \sqrt{\frac{\nabla D_{l_i}^{\perp} \cdot u_i}{A}}, \tag{2}$$

where $P(i)$ represents the instant patch centered at pixel i in D_l, $\overline{M}$ is the unmasked region in which all the depth values are available, A is a normalization factor, u_i stands for a unit vector orthogonal to the front at pixel i. The initialization for $C(j)$ is set to $C(j) = 1$ if pixel value j is known and $C(j) = 0$ otherwise. D_{l_i} represents the depth value at pixel i in D_l. As shown in Eq. (2), priority at a given pixel is measured as the product of two terms: the confidence term $C(\cdot)$ and the data term $T(\cdot)$. Both terms are normalized to range between 0 and 1.

After computing the priority for every pixel along the boundary of the masked region, the depth value of pixel p with the highest priority is predicted first. We calculate the similarity between pixel p with the qualified pixel(s) within a neighborhood of $b \times b$ centered at p. Only pixels with valid depth values and share the same label as p measured in the label map are considered qualified. The similarity between two pixels is measured in the mean square error (MSE) of the corresponding pixel-level features extracted in the intensity image introduced above. Then the depth value for pixel p is filled with the corresponding depth value of the most similar one. After updating the confidence term and the data term, the above process is repeated until all the depth values within mask M are filled.

2.2. Super Resolution

Single image SR is a numerically ill-posed problem and therefore relies on additional assumptions or priors to finalize the output. External example learning-based SR hinges on learning statistical priors or models from a large image dataset and leads to a stable SR performance. Compared with internal example learning-based SR, learning externally allows introducing new information other than employing only the input image and thus is more robust when the upscaling factor is relatively large. Moreover, the learning process is normally performed off-line and is therefore computationally tractable during the testing phase.

In the proposed SR framework, the completed depth map and the input intensity image are fed into a group of externally-trained regression models to recover the final HR outputs. The regression models are trained separately for depth maps and intensity images utilizing the same learning pipeline over different datasets consisted of corresponding exemplars. Fig. 3 illustrates the schematic pipeline of the proposed SR process. A large set of HR/LR exemplar patch pairs with magnification factor s are collected from a dataset consisted of training images. Original images in the dataset are considered HR images and the corresponding L-

Table 1. Comparison of the proposed approach with other depth SR schemes over 14 depth maps in dataset [23] measured in terms of FSIM [38] under the magnification factor of 4. The best performance in FSIM for each image is marked in bold.

Middlebury2001 [23]	barn1-disp2	barn1-disp6	barn2-disp2	barn2-disp6	bull-disp2	bull-disp6	map-disp0
nearest-neighbor	0.9527	0.9530	0.9532	0.9560	0.9598	0.9614	0.9325
Aodha *et al.* [1]	0.9626	0.9570	0.9479	0.9464	0.9481	0.9467	**0.9669**
Ours	**0.9777**	**0.9782**	**0.9815**	**0.9824**	**0.9902**	**0.9911**	0.9566
Middlebury2001 [23]	map-disp1	poster-disp2	poster-disp6	sawtooth-disp2	sawtooth-disp6	venus-disp2	venus-disp6
nearest-neighbor	0.9322	0.9582	0.9624	0.9505	0.9459	0.9573	0.9578
Aodha *et al.* [1]	0.9373	0.9701	0.9655	0.9553	0.9529	0.9633	0.9629
Ours	**0.9561**	**0.9842**	**0.9843**	**0.9742**	**0.9752**	**0.9818**	**0.9819**

R images are generated through a blur and downsampling process. For an instance patch pair, both patches are normalized by extracting the mean value of the LR patch. After the normalization and vectorization, the LR and HR features are represented as $\boldsymbol{L} \in \mathbb{R}^{l \times S}$ and $\boldsymbol{H} \in \mathbb{R}^{h \times S}$ where l and h denote the feature dimensions and S indicates the number of samples.

Due to the diversity of patch patterns, before the regression model training, we first model the input LR feature space with GMM to ensure a more targeted learning. GMM is a generative model with the capacity to model any given probability distribution function when the number of Gaussian components is large enough. The probability of a feature $\boldsymbol{x}_i$ given a GMM with K components is

$$p(\boldsymbol{x}_i|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\right), \tag{3}$$

where w_k represents the prior mode probability that satisfies the constraint $\sum_{k=1}^{K} w_k = 1$, and $\mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\right)$ indicates the kth normal distribution with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\sigma}_k$:

$$\mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\right) = \frac{\exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{\sigma}_k)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)}{2}\right)}{(2\pi)^{m/2} \left|\boldsymbol{\sigma}_k\right|^{1/2}}, \tag{4}$$

where $\boldsymbol{x}_i \in \mathbb{R}^l$, $\boldsymbol{\mu}_k \in \mathbb{R}^l$, and $\boldsymbol{\sigma}_k \in \mathbb{R}^{l \times l}$. The GMM parameters $\theta = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, k = 1, \ldots, K\}$ are estimated using the Expectation-Maximization (EM) algorithm to optimize the Maximum Likelihood (ML) from a large number of features.

We then assign each LR feature $\boldsymbol{x}_i \in \boldsymbol{L}$ to corresponding Gaussian component according to the posterior. Assume there are N_k patches associated with the kth Gaussian component and $\boldsymbol{L}_k \in \mathbb{R}^{l \times N_k}$, $\boldsymbol{H}_k \in \mathbb{R}^{h \times N_k}$ stand for the corresponding LR/HR features, a linear regression model is then trained with the coefficient C_k learnt through:

$$C_k^* = \underset{C_k}{\operatorname{argmin}} \{|\boldsymbol{H}_k - C_k \hat{\boldsymbol{L}}_k|^2\}, \tag{5}$$

where $\hat{\boldsymbol{L}}_k{}^T = [\boldsymbol{L}_k{}^T \ \mathbf{1}]$. After performing the training process over a dataset of natural images, K regression models are learnt for upscaling the intensity images. Another set of K regression models is obtained through learning over a depth-map dataset in a similar manner. Given a testing LR instance, patch-based features are extracted by performing normalization and vectorization for every patch. After that, according to the posterior, each feature is assigned to a Gaussian component where the corresponding regression model is applied to obtain the HR patch. Weighted average is adopted to blend overlapping pixels to generate the final HR image. Back-projection is utilized as a post processing.

3. Experimental Results

In this section, the proposed resolution enhancement system is evaluated on the Middlebury Stereo Datasets [23, 24, 22, 12] and RGBD Scenes dataset v2 [17]. We present multiple generated HR depth maps and intensity images compared with the recent state-of-the-art methods both quantitatively and qualitatively.

3.1. Implementation Details

During the depth completion, the pixel-level feature vector has a dimension of 83 where 80 of them come from the Gabor responses of multiple filters at 5 scales and 8 orientations. The rest are extracted according to the local homogeneity model. The neighborhood size a defined to calculate the standard deviation is set to 5. The number of training samples N_1 and N_2 are the same and equal to 10% of the total number of pixels in the LR RGB image. We employ SVM with RBF kernel utilizing LIBSVM [4]. In Eq. (2), b is set to 7 and the normalization factor A is 255.

In the SR phase, the training dataset used for intensity regression model learning is the same as in [35] with $6,152$ natural images. $1,449$ completed depth maps in NYU Depth Dataset V2 [29] are employed to train the depth regression models. The original examples in the datasets are treated as the HR instances and the corresponding LR exemplars are generated through a blurring and downsampling process. We extract all patches with size 7×7 from the LR instances. Only the central $3s \times 3s$ pixels in the corresponding HR patch are captured to formulate the HR features where s is the scaling factor. Total of $200,000$ LR/HR fea-

Table 2. Comparison of the proposed approach with peer methods for intensity images in datasets [23, 24, 22, 12] measured in terms of average FSIM [38] under the scaling factor of 4. The best performance in FSIM for each dataset is marked in bold.

Datasets	nearest	bicubic	ScSR [36]	ANR [32]	Yang [35]	Ours
M2001 [23]	0.7034	0.7705	0.7716	0.8052	**0.8169**	0.8129
M2003 [24]	0.7128	0.7894	0.7932	0.8200	0.8213	**0.8215**
M2005 [22]	0.9826	0.9863	0.9702	0.9930	0.9902	**0.9940**
M2006 [12]	0.9855	0.9892	0.9728	0.9944	0.9918	**0.9949**

tures are randomly selected to train the GMM model with 512 components. After that, the remaining features are assigned to corresponding Gaussian component with the highest probability. Finally, a linear regression model is learnt for each Gaussian component. Same as many existing image SR methods, for color images, the proposed SR algorithm is performed on the luminance channel in YUV color space while the other two color channels are upscaled by bicubic interpolation. The number of iterations for back-projection is set to 10. The time cost of the proposed system varies depending on the size of the input image, the scaling factor, and the number of missing pixels in the depth map. Generally speaking, under a scaling factor of 4, for an input depth map (120×100) with missing pixels less than 10% of the overall number of pixels, it takes couple of seconds to finish the depth completion and the SR.

3.2. Quantitative Comparison

We evaluate the proposed enhancement framework on a variety of depth maps and RGB images in Middlebury Stereo Datasets [23, 24, 22, 12] under the scaling factor of 4. Middlebury DB 2001 [23] consists of depth maps without missing values and depth maps provided in the rest datasets listed, i.e., Middlebury DBs 2003, 2005, 2006, all contain unknown regions. In our experiments, the LR inputs are generated from the depth maps and intensity images in the datasets by performing the nearest-neighbor downsampling. Since the completed ground-truth depth maps for Middlebury DBs 2003, 2005, 2006 are unavailable, it is infeasible to provide numerical statistics to measure the depth completion performance. Visual results are presented instead in Section 3.3.

Quantitative evaluations for SR performance are provided over depth maps in Middlebury DB 2001 and intensity images for all listed datasets. The recent proposed image quality assessment criterion Feature Similarity (FSIM) [38] is adopted for measurement since the human visual system (HVS) understands an image mainly according to its low-level features and therefore FSIM achieves higher consistency with the subjective evaluation compared with other metrics. In the proposed depth SR, statistics from the aligned RGB image is not utilized. Therefore, we compare our results with the state-of-the-art single depth SR method [1]. As illustrated in Table 1, our approach outperforms [1] in 13 out of 14 depth maps measured in terms of FSIM. [36, 32, 35] are the state-of-the-art external example-based single image SR approaches. Table 2 presents the comparison of the proposed approach with these methods listed for intensity images in 4 different datasets. Our framework adopts GMM to model the input feature space and therefore a more targeted learning is ensured for the regression model training. Statistics from Table 2 reveals that the proposed SR framework is the most effective in 3 out of 4 evaluated datasets for intensity outputs measured in average FSIM.

3.3. Qualitative Comparison

The enhancement performance is further evaluated qualitatively in this subsection. Fig. 4 presents our enhancement results under the scaling factor 4 of 'Baby3', 'Midd1', 'Moebius', and 'Cone' from the Middlebury datasets. In 'Baby3', it is challenging to correctly predict the missing depth values due to the cluttered background (i.e., map with irregular contours and patterns) and the color similarity between certain foreground objects and the background scene. As observed, after the proposed enhancement framework, missing depth values are well filled in a manner consistent with the embedded structure. Effectiveness of the proposed depth completion algorithm can be further demonstrated in 'Cone' where complex texture patterns exist along with large missing depth areas. Rich details and clear edges are preserved with minimal artifacts as illustrated from both the HR depth maps and the HR intensity outputs.

Fig. 5 presents more enhancement results in RGBD Scenes dataset v2 [17] under the magnification factor of 4. Different from the Middlebury datasets, depth maps in [17] suffer from large missing areas and the registered intensity images are of low quality with blurry visual artifacts. The proposed enhancement framework manages to reconstruct structurally correct completed depth maps. Clear contours are recovered in the HR results.

We further compare our results with representative state-of-the-art peer algorithms [1, 36, 32] in depth and intensity SR. In Fig. 6, a set of SR results is provided on depth maps 'bull' and 'sawtooth' compared with nearest-neighbor interpolation and the outputs generated by [1]. For a better illustration, only part of each depth map is presented. The nearest-neighbor interpolated results suffer from blurry visual artifacts. Irregular zigzag patterns occur in results constructed by [1]. Compared with the ground-truths, our results best recover the contours with minimal visual artifacts. Corresponding HR intensity outputs are presented in Fig. 7 compared with nearest-neighbor interpolation, bicubic interpolation, ScSR [36], and GR [32]. While interpolation-based methods produce over-smoothed results, ringing artifact exists in results generated by [32]. [36] reconstructs results with gridded patterns that do not exist in the orig-

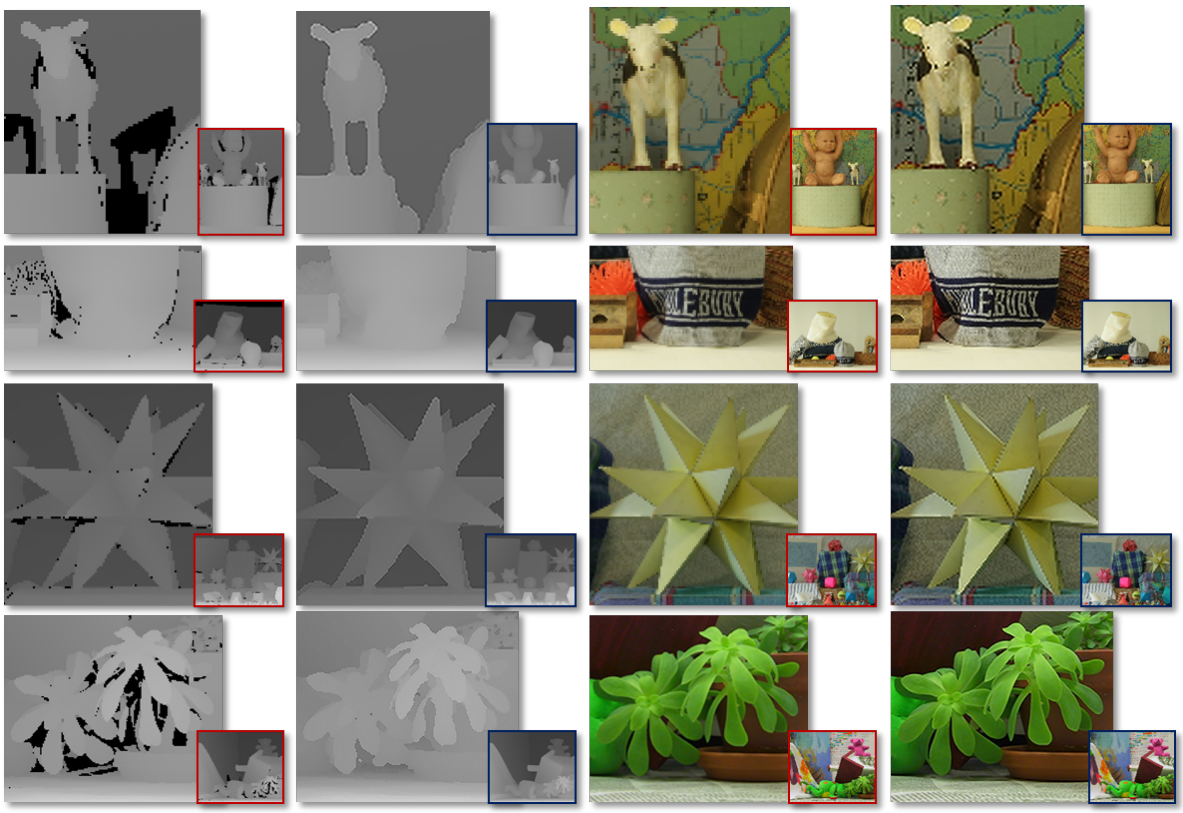

Figure 4. Resolution enhancement results of 'Baby3', 'Midd1', 'Moebius', and 'Cone' ($\times 4$). From left to right, the columns represent LR input depth maps, generated HR depth maps, LR intensity images, and HR intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation. This figure is better viewed on screen with HR display.

Figure 5. Resolution enhancement results in RGBD Scenes dataset v2 [17] ($\times 4$). From left to right, the columns represent LR input depth maps, generated HR depth maps, LR intensity images, and HR intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation. This figure is better viewed on screen with HR display.

inal image. As observed from the edges and textures, our recovered HR images reveal more natural patterns and finer details.

4. Conclusions and Future Work

In this paper, we have proposed a resolution enhancement system to improve the quality of a registered pair of low-resolution depth map and intensity image. Missing values in the input depth map are predicted utilizing the

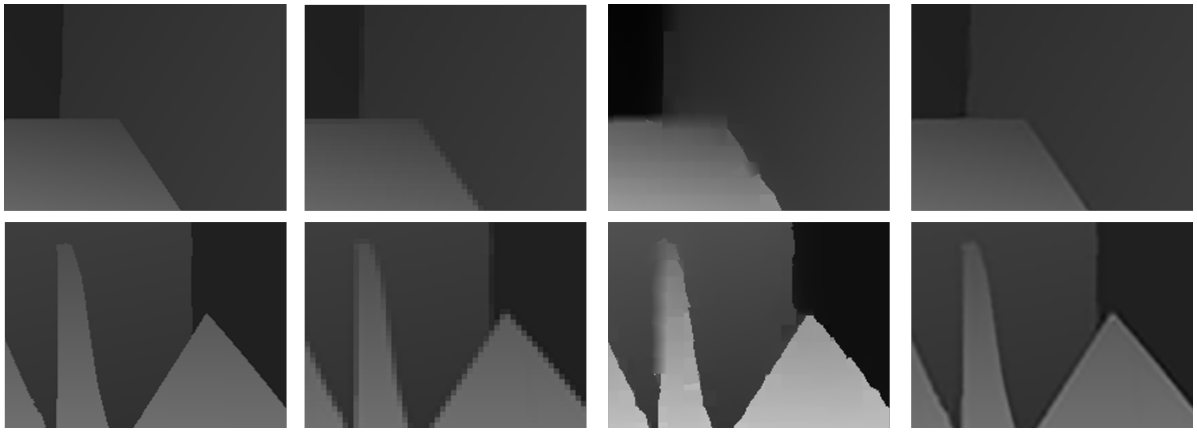

Figure 6. SR results of depth maps 'bull' and 'sawtooth' ($\times 4$). From left to right, the columns represent the ground-truth, results generated utilizing nearest-neighbor interpolation, patch-based [1], and the proposed framework. Only partial of the original depth maps are shown for a clearer presentation. The figure is better viewed on screen with HR display.

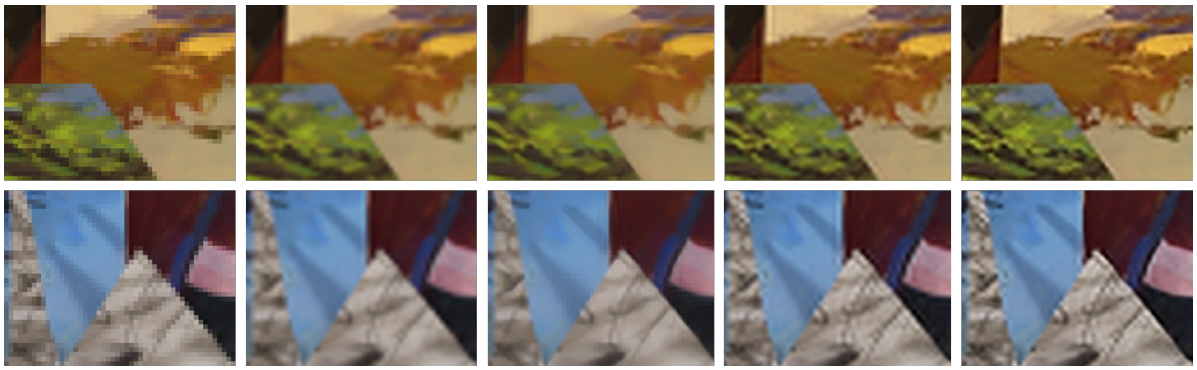

Figure 7. SR results of RGB images 'bull' and 'sawtooth' ($\times 4$). From left to right, the columns represent the results generated through nearest-neighbor interpolation, bicubic interpolation, ScSR [36], GR [32], and the proposed system. Only partial of the images are shown for a clearer presentation. The figure is better viewed on screen with HR display.

statistical structure dependency between the aligned low-resolution RGB image and the depth map. Fine details are further recovered by feeding the completed depth map and the low-resolution intensity image to a set of regression models. The regression models for upscaling depth maps and RGB images are trained separately utilizing a uniform learning pipeline in a divide-and-conquer manner. As demonstrated by the extensive experimental results, the proposed enhancement framework is effective with satisfying performance measured quantitatively in FSIM. Compared with representative peer approaches, the generated high-resolution depth maps and images have clearer contours and more natural textures close to the ground-truths.

Currently, structural dependency between the aligned low-resolution RGB image and the depth map is well exploited to accomplish depth completion. In the future, we will investigate more in the statistical relations between the input pairs in the joint simultaneous super-resolution phase.

Acknowledgment

This work was supported in part by ONR Grant N000141310450, FHWA Grant DTFH61-12-H-00002, and NSF Grants EFRI-1137172, IIP-1343402.

References

[1] O. M. Aodha, N. D. Campbell, A. Nair, and G. J. Brostow. Patch based synthesis for single depth image super-resolution. In *ECCV*, 2012.

[2] F. Bianconi and A. Fernandez. Evaluation of the effects of gabor filter parameters on texture classification. *Pattern Recognition*, 40(12):3325–3335, 2007.

[3] N. D. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. Using multiple hypothese to improve depth-maps for multi-view stereo. In *ECCV*, 2008.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[5] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *CVPR*, 2010.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014.

[7] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004.

[8] R. Fattal. Image upsampling via imposed edge statistics. In *ACM SIGGRAPH*, 2007.

[9] G. Freedman and R. Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics*, 28(3):1–10, 2010.

[10] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009.

[11] S. E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *IEEE Transactions on Image Processing*, 11(10):1160–1167, 2002.

[12] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007.

[13] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

[14] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, 2011.

[15] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. In *IEEE SMC*, 1990.

[16] M. Kiechle, S. Hawe, and M. Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *ICCV*, 2013.

[17] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, 2014.

[18] H. S. Lee and K. M. Lee. Simultaneous super-resolution of depth and images using a single camera. In *CVPR*, 2013.

[19] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha. Similarity-aware patchwork assembly for depth image super-resolution. In *CVPR*, 2014.

[20] J. Li, G. Zeng, R. Gan, H. Zha, and L. Wang. A bayesian approach to uncertainty-based depth map super resolution. In *ACCV*, 2012.

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

[22] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *CVPR*, 2007.

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1).

[24] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003.

[25] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. High-quality scanning using time-of-flight depth superresolution. In *CVPR TOF Workshop*, 2008.

[26] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In *CVPR*, 2009.

[27] Q. Shan, Z. Li, J. Jia, and C. K. Tang. Fast image/video upsampling. In *ACM SIGGRAPH Asia*, 2008.

[28] J. Shen and S. ching S. Cheung. Layer depth denoising and completion for structured-light rgb-d cameras. In *CVPR*, 2013.

[29] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[30] D. Su and P. Willis. Image interpolation by pixel-level data-dependent triangulation. *Computer Graphics Forum*, 23:189–201, 2004.

[31] J. Sun, J. Sun, Z. Xu, and H. Y. Shum. Image super-resolution using gradient profile prior. In *CVPR*, 2008.

[32] R. Timofte, V. D. Smet, and L. V. Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013.

[33] X.-Y. Wang, T. Wang, and J. Bu. Color image segmentation using pixel wise support vector machine classification. *Pattern Recognition*, 44(4):777–787, 2011.

[34] Y. Xian and Y. Tian. Robust internal exemplar-based image enhancement. In *ICIP*, 2015.

[35] C.-Y. Yang and M.-H. Yang. Fast direct super-resolution by simple functions. In *ICCV*, 2013.

[36] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008.

[37] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *CVPR*, 2007.

[38] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20:2378–2386, 2011.

[39] Q. Zhang and J. Wu. Image super-resolution using windowed ordinary kriging interpolation. *Optics Communications*, 336:140–145, 2015.

[40] F. Zhou, W. Yang, and Q. Liao. Interpolation-based image super-resolution using multisurface fitting. *IEEE Transactions on Image Processing*, 21:3312–3318, 2012.