Super-Resolution 3D Reconstruction from Multiple Cameras

Tomoaki Nonome, Fumihiko Sakaue and Jun Sato
Department of Computer Science and Engineering, Nagoya Institute of Technology, Gokiso, Showa, Nagoya 466-8555, Japan

Keywords: Super-Resolution, Multiple Cameras, 3D Reconstruction, High Resolution.

Abstract: In this paper, we propose a novel method for reconstructing high resolution 3D structure and texture of the scene. In the image processing, it is known that image super-resolution is possible from multiple low resolution images. In this paper, we extend the image super-resolution into 3D space, and show that it is possible to recover high resolution 3D structure and high resolution texture of the scene from low resolution images taken at different viewpoints. The experimental results from real and synthetic images show the efficiency of the proposed method.

1 INTRODUCTION

Recovering structure of the scene is one of the very important objectives in computer vision, and many efficient reconstruction methods have been proposed in the past research.

The early studies in this field revealed what kind of constraints exist in multiple images, and what kind of information can be obtained from these images (Hartley and Zisserman, 2000; Faugeras et al., 2004). For this objective, two-view, three-view and multi-view geometry have been studied extensively (Longuet-Higgins, 1981; Shashua and Werman, 1995; Hartley and Zisserman, 2000; Faugeras et al., 2004). The bundle adjustment (Triggs et al., 1999) has been combined with these theoretical advances, and the sparse 3D reconstruction has been achieved.

More recently, multiple images are used for recovering large scale structures of the scene efficiently. One of the mile stone research in this field was presented by Agarwal et al. (Agarwal et al., 2011), who showed that whole buildings and cities can be reconstructed automatically from vast amount of images. Furthermore, the accuracy of 3D reconstruction of large scale scenes has been improved drastically in recent years (Galliani et al., 2015; Schonberger et al., 2016). However, these existing methods use multiple images mainly for reducing outliers and noises in reconstructed 3D structures. That is, the multiple images have been used for improving the stability of 3D reconstruction. On the contrary, we in this paper propose a method which uses multiple images for recovering finer 3D structures of the scene.

In the image processing research field, the super-resolution of 2D images has been studies extensively. The existing methods in this field can be classified into two groups. The first group of methods are based on statistical priors which are obtained from advanced learning (Glasner et al., 2009; Kim et al., 2016; Ledig et al., 2016). These methods can obtain a high resolution image just from a single low resolution image, since the statistical priors can compensate the lack of high frequency term in the image. However, these methods are heavily depend on the priors, and if the priors do not agree with the input images, they output wrong high resolution images. The second group of methods are based on multiple observations (Hardie et al., 1997; Tom et al., 1994). Although these methods needs multiple images, they can recover high resolution images accurately without any wrong inference. In this paper, we propose a new method for recovering fine 3D structures of the scene by extending the image super-resolution based on multiple images.

In our method, we recover fine 3D structures of the scene, whose resolutions are much higher than the input image resolutions. For this objective, we recover the high resolution structures of the scene directly from the image intensity of low resolution images. Thus point correspondences among multiple images are not required in our method. Instead, we recover the high resolution texture of the scene as well as the high resolution 3D structure of the scene. By estimating the high resolution textures and the high resolution 3D structures simultaneously, we can recover

Nonome, T., Sakaue, F. and Sato, J.
Super-Resolution 3D Reconstruction from Multiple Cameras.
DOI: 10.5220/0006718304810486
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 481-486
ISBN: 978-989-758-290-5

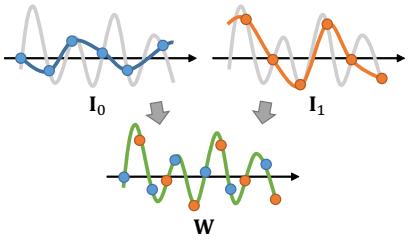


Figure 1: Image super-resolution from multiple images. By obtaining multiple image, $\mathbf{I}_0$ and $\mathbf{I}_1$, with different sampling phases, a high resolution image $\mathbf{W}$ can be recovered by combining these images.

fine structure of the scene from low resolution images. As as result, we can recover fine 3D structures, which could not be recovered by the existing multiple view reconstruction methods.

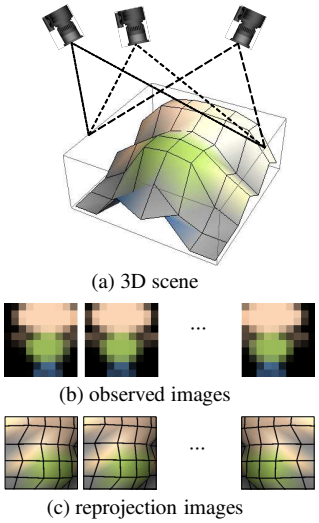
2 IMAGE SUPER-RESOLUTION

Before considering the super-resolution 3D reconstruction, we revise the standard image super-resolution from multiple images. In the image super-resolution, multiple images observed at different viewpoints are combined together, so that these multiple observations compensate the lack of observation in single image as shown in Fig. 1. The maximum a posteriori probability (MAP) estimation is often used for obtaining a high resolution image $\mathbf{W}$ from multiple low resolution images $\mathbf{I}_i$ $(i = 1, \cdots, N)$ under the existence of image noise as follows:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \|\mathbf{I}_i - \mathbf{A}_i\mathbf{W}\|^2 + \alpha\|\mathcal{L}\mathbf{W}\|_2 \quad (1)$$

The first term in Eq.(1) is a data term, and $\mathbf{A}_i$ denotes a matrix which represents down sampling in ith image, i.e. a down sampling at ith viewpoint. The second term is a regularization term, and $\mathcal{L}$ denotes the Laplacian filter for smoothness constraints. $\|\cdot\|_2$ denotes the L_2 norm, and α denotes the magnitude of the regularization term.

The image super-resolution assumes that the objective surface is planar, and the difference of sampling phase in each image is constant. Thus, if the objective surface is not planar, the standard image super-resolution fails. On the contrary, we in this paper consider non-planar objects, and propose a method for reconstructing high resolution 3D surfaces as well as their high resolution textures. We call it super-resolution 3D reconstruction.



(a) 3D scene

(b) observed images

(c) reprojection images

Figure 2: observed images and reprojection images.

3 SUPER-RESOLUTION 3D RECONSTRUCTION

Our super-resolution 3D reconstruction is achieved by estimating high resolution 3D structures and high resolution textures simultaneously by minimizing a cost functions defined by low resolution images observed at multiple viewpoints.

Let us consider a 3D surface whose high resolution structure and texture are $\mathbf{D}$ and $\mathbf{W}$ respectively. Suppose the 3D surface is projected into N cameras $\mathbf{C}_i$ $(i = 1, \cdots, N)$, and N low resolution images $\mathbf{I}_i$ $(i = 1, \cdots, N)$ are observed as shown in Fig. 2. Then, these projections can be described by projection functions $\mathcal{P}_i$ $(i = 1, \cdots, N)$ as follows:

$$\mathbf{I}_i = \mathcal{P}_i(\mathbf{D}, \mathbf{W}) \quad (2)$$

The projection functions $\mathcal{P}_i$ represent not only the relative position and orientation among N cameras, but also the down sampling in these cameras. In this research, we assume that the cameras are calibrated, and the projection functions $\mathcal{P}_i$ are known. Also, we assume the ambient light is constant in all the orientations around the 3D surface, and the local orientation of the surface does not affect the intensity in images.

Then, the objective of our method is to estimate the high resolution structure $\mathbf{D}$ and the high resolution texture $\mathbf{W}$ simultaneously, which best fit low resolution images $\mathbf{I}_i$ $(i = 1, \cdots, N)$ observed by N cameras as follows:

$$\{\hat{\mathbf{D}}, \hat{\mathbf{W}}\} = \arg\min_{\mathbf{D},\mathbf{W}} \sum_{i=1}^{N} \|\mathbf{I}_i - \mathcal{P}_i(\mathbf{D}, \mathbf{W})\|^2 \quad (3)$$

In the real scenes, we can assume that the structure and the texture of the 3D surface does not change drastically except the boundary of objects and the boundary of textures.Thus, we can make the estimation more stable by adding the smoothness constraints $\mathcal{S}(\cdot)$ on structure and texture as follows:

$$\{\hat{\mathbf{D}},\hat{\mathbf{W}}\} = \arg\min_{\mathbf{D},\mathbf{W}} \sum_{i=1}^{N} \|\mathbf{I}_i - \mathcal{P}_i(\mathbf{D},\mathbf{W})\|^2 + \mathcal{S}(\mathbf{D}) + \mathcal{S}(\mathbf{W}) \quad (4)$$

Unfortunately, the simultaneous estimation of high resolution structures and textures described in Eq.(4) is very difficult and unstable, since we have to minimize the cost function in very high dimensional space. In the next section, we describe a practical method for estimating high resolution structures and high resolution textures based on Eq.(4).

4 PRACTICAL SUPER-RESOLUTION 3D RECONSTRUCTION

Suppose we have N cameras, and each of which obtains a low resolution image $\mathbf{I}_i = [I_1,\cdots,I_P]^\top$ $(i = 1,\cdots,N)$, where P denotes the number of pixels in a low resolution image. Then, we consider one of these cameras as a basis camera, and its camera coordinates are considered as the basis 3D coordinates of the scene. Thus, the high resolution 3D structure of the scene is represented by a high resolution depth image $\mathbf{D} = [D_1,\cdots,D_Q]^\top$ observed at the basis camera, where Q denotes the number of pixels in a high resolution image. Also, the high resolution texture of the scene is represented by a high resolution intensity image $\mathbf{W} = [W_1,\cdots,W_Q]^\top$ observed at the basis camera. Naturally, we assume $P \leq Q$. Then, our objective is to estimate $\mathbf{D}$ and $\mathbf{W}$ from $\mathbf{I}_i$.

Since the simultaneous estimation of high resolution structures and textures shown in Eq.(4) is difficult and unstable, we in this paper estimate high resolution structures and high resolution textures alternately by iterating the following two steps.

4.1 Estimation of High Resolution Textures

We fist estimate a high resolution texture $\mathbf{W}$ given an estimated high resolution structure $\mathbf{D}$.

Suppose we have a high resolution structure $\mathbf{D} = [D_1,\cdots,D_Q]^\top$. Then, the low resolution images $\mathbf{I}_i$ $(i = 1,\cdots,N)$ observed by N cameras can be described by using the high resolution texture $\mathbf{W}$ as follows:

$$\mathbf{I}_i = \mathbf{A}_i(\mathbf{D})\mathbf{W} \quad (5)$$

where, $\mathbf{A}_i(\mathbf{D})$ denotes a $P \times Q$ matrix, which represents a projection from the high resolution texture at the basis camera to the low resolution image at the ith camera given a high resolution structure $\mathbf{D}$. Thus, the high resolution texture $\mathbf{W}$ can be estimated from low resolution images $\mathbf{I}_i$ observed at N cameras by solving the following minimization problem:

$$\hat{\mathbf{W}}(\mathbf{D}) = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \|\mathbf{I}_i - \mathbf{A}_i(\mathbf{D})\mathbf{W}\|^2 + \alpha\|\mathbf{L}\mathbf{W}\|_2 \quad (6)$$

where, $\mathbf{L}$ denotes a matrix for computing the Laplacian of $\mathbf{W}$, and $\|\cdot\|_2$ denotes the L_2 norm. Thus the second term represents the smoothness constraints $\mathcal{S}(\mathbf{W})$ on high resolution textures, and α is its weight. From Eq.(6), the high resolution texture $\mathbf{W}$ can be estimated given a high resolution structure $\mathbf{D}$.

Note, the estimation of $\mathbf{W}$ in Eq.(6) is a linear problem, and thus $\mathbf{W}$ can be estimated linearly.

4.2 Estimation of High Resolution Structures

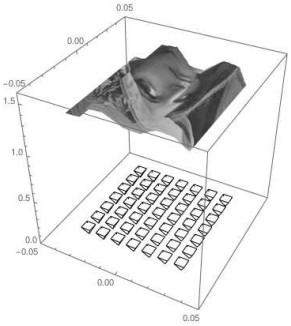
We next estimate a high resolution structure $\mathbf{D}$ given an estimated high resolution texture $\mathbf{W}$.

Given a high resolution texture $\mathbf{W} = [W_1,\cdots,W_Q]^\top$, the low resolution camera images $\mathbf{I}_i$ can be described by Eq.(5) as before. Then, the high resolution structure $\mathbf{D}$ can be estimated from low resolution images $\mathbf{I}_i$ observed at N cameras as follows:

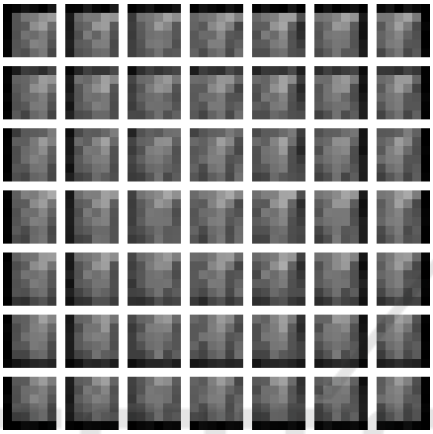
$$\hat{\mathbf{D}}(\mathbf{W}) = \arg\min_{\mathbf{D}} \sum_{i=1}^{N} \|\mathbf{I}_i - \mathbf{A}_i(\mathbf{D})\mathbf{W}\|^2 + \beta\|\mathbf{L}\mathbf{D}\|_2 \quad (7)$$

The second term represents the smoothness constraints $\mathcal{S}(\mathbf{D})$ on high resolution structures, and β is its weight. α and β are chosen empirically in our experiments.

By iterating Eq.(6) and Eq.(7) alternately, we can estimate the high resolution structure $\mathbf{D}$ and texture $\mathbf{W}$ of 3D surfaces. In this estimation, we also use coarse to fine technique to stabilize the super resolution estimation. That is, we gradually increase the scale of estimated texture and structure during the iteration. Since we need initial values of 3D structure and 2D texture in our estimation, we used a flat surface as the initial structure and used the low resolution image of the basis camera as the initial texture. By using the proposed method, the high resolution structures and textures can be estimated efficiently.

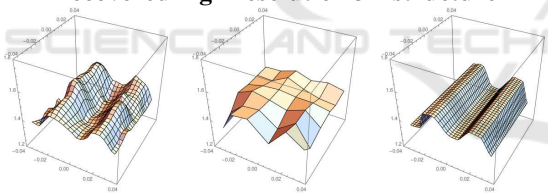


(a) 3D object and 7×7 cameras



(b) low resolution camera images

Recovered high resolution 3D structure

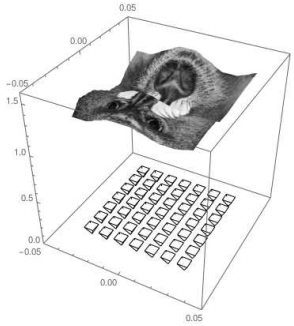


(c) proposed (d) stereo (e) ground truth

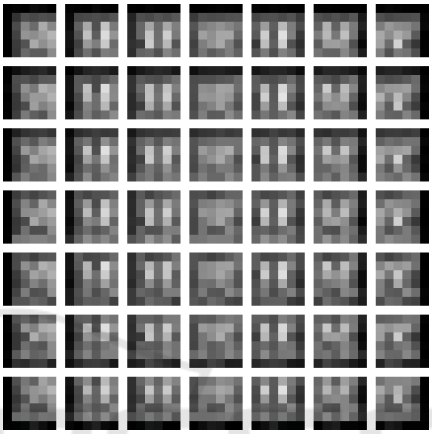
Recovered high resolution texture

(f) proposed (g) bi-cubic (h) ground truth

Figure 3: Results of super-resolution reconstruction (lenna).

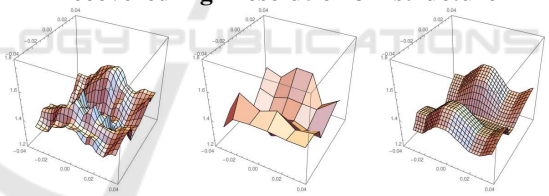


(a) 3D object and 7×7 cameras



(b) low resolution camera images

Recovered high resolution 3D structure



(c) proposed (d) stereo (e) ground truth

Recovered high resolution texture

(f) proposed (g) bi-cubic (h) ground truth

Figure 4: Results of super-resolution reconstruction (mandrill).

5 EXPERIMENTS

We next show the efficiency of the proposed method by using synthetic images as well as real images.

5.1 Synthetic Image Experiments

We first show that high resolution 3D structures and textures can be obtained from the proposed method by using synthetic images.

Fig. 3 (a) shows a 3D object used in our synthetic image experiment. The 7×7 quadrangular py-

ramids in this figure show the position and the orientation of 49 cameras used in our super resolution reconstruction. These cameras were assumed to be calibrated in this experiment. The image resolution of each camera is 6 pix $\times$ 6 pix, and 49 images obtained from Fig. 3 (a) are shown in Fig. 3 (c). The intensity range of these images is 0 to 1, and the random Gaussian noise with the standard deviation of 0.01 was added to the image intensity for simulating the image noise in observation. These low resolution images were used for recovering high resolution 3D structures and textures with the resolution of 24 pix $\times$ 24 pix.

Fig. 3 (c) shows a high resolution 3D structure obtained from the proposed method. The ground truth structure is shown in Fig. 3 (e). For comparison, we also reconstructed the 3D structure from the low resolution images by using the standard stereo method. The obtained 3D structures is shown in Fig. 3 (d). As shown in these figures, the proposed method provides us fine structure of the original shape, while the standard stereo method suffers from the aliasing problem, and cannot recover correct shape of object.

We next show a high resolution texture, i.e. high resolution image at the basis camera, recovered from the proposed method. Fig. 3 (f) shows the result from the proposed method, and Fig. 3 (h) shows the ground truth texture. For comparison, the result from the standard bi-cubic interpolation is also shown in Fig. 3 (g). As shown in these figures, the proposed method provides us the high resolution texture of the object accurately, even if the input images are very low resolution. On the contrary, the result from the standard bi-cubic interpolation is very bad.

Fig. 4 shows the results from another synthetic 3D object. Again, the proposed method provides us very accurate high resolution structure and texture, while the standard stereo method and bi-cubic interpolation cannot recover high resolution structure and texture. The numerical accuracy of recovered structures and textures shown in table 1 and table 2 also show the efficiency of the proposed method.

Table 1: Accuracy of recovered high resolution 3D structure.

	proposed method	existing method
lenna	0.0185	0.0405
mandrill	0.0195	0.0406

Table 2: Accuracy of recovered high resolution texture.

	proposed method	existing method
lenna	0.0636	0.0910
mandrill	0.1035	0.1734

5.2 Real Image Experiments

We next show the results from real image experiment. In this experiment, we recovered the high resolution structure and texture of a plaster face shown in Fig. 5. The plaster face was observed by a camera which was translated in 2 directions by using a moving stage shown in Fig. 5, and 5×5 images were obtained with every 2cm translation. For obtaining the ground truth image of high resolution texture, we generated the low resolution images by taking the average of 4×4 pixels, and used these low resolution images for super resolution 3D reconstruction. The ground truth shape of the object was measured by using structured lights projected from the projector in Fig. 5, and the camera internal parameters were calibrated in advance by using a calibration board. Fig. 6 shows 5×5 low resolution images obtained from the camera. The resolution of these images is 8×8. We used these low resolution images for recovering the high resolution structure and texture whose resolution is 32×32.

Fig. 7 (a) shows the high resolution 3D structure recovered from the proposed method, and (c) shows the ground truth of the structure. For comparison the result from the standard stereo method is shown in Fig. 7 (b). As shown in this figure, the result from the proposed method is very fine and accurate, while the result from the standard stereo is very rough and inaccurate. The high resolution textures recovered from the proposed method is also compared with that of the

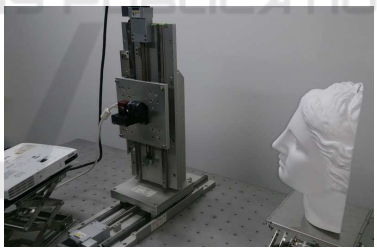
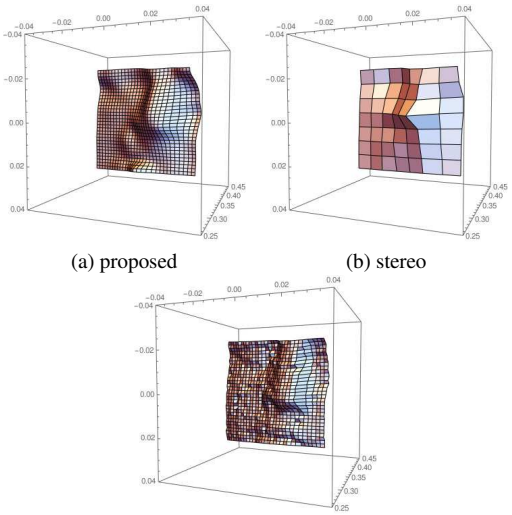


Figure 5: The experimental setup of our real image experiment.

Figure 6: Low resolution images obtained from a moving camera.



(a) proposed (b) stereo

(c) ground truth

Figure 7: The high resolution 3D structures recovered from low resolution images.

(a) proposed (b) bi-cubic (c) ground truth

Figure 8: The high resolution textures recovered from low resolution images.

bi-cubic interpolation in Fig. 8. Again, the proposed method is superior to the standard bi-cubic method.

From these results, we find that the proposed method is very efficient to recover accurate high resolution 3D structures and textures.

6 CONCLUSION

In this paper, we proposed a novel method for reconstructing high resolution 3D structure and texture of the scene. For this objective, we extended the 2D image super-resolution into 3D space, and showed that it is possible to recover high resolution 3D structure and high resolution texture of the scene from low resolution images taken at different viewpoints.

We showed the efficiency of the proposed method by using real and synthetic image experiments comparing with the existing methods.

REFERENCES

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. (2011). Building rome in a day. *Commun. ACM*, 54(10):105–112.

Faugeras, O., Luong, Q., and Papadopoulo, T. (2004). *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications*. MIT press.

Galliani, A., K.Lasinger, and Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. ICCV*.

Glasner, D., Bagon, S., and Irani, M. (2009). Super-resolution from a single image. In *Proc. ICCV*.

Hardie, R., Barnard, K., and Armstrong, E. (1997). Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633.

Hartley, R. and Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Kim, J., Lee, J., and Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2016). Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414?–2423.

Longuet-Higgins, H. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135.

Schonberger, J., Zheng, E., Pollefeys, M., and Frahm, J.-M. (2016). Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV*.

Shashua, A. and Werman, M. (1995). Trilinearity. In *Proc. ICCV*, pages 920–925.

Tom, B., Katsaggelos, A., and Galatsanos, N. (1994). Reconstruction of a high resolution image from registration and restoration of low resolution images. In *Proc. IEEE International Conference on Image Processing*, pages 553–557.

Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (1999). Bundle adjustment - a modern synthesis. In *Proc. International Workshop on Vision Algorithms*.