Combining Multi-view Stereo and Super Resolution in a Unified Framework

Haesol Park, Kyoung Mu Lee and Sang Uk Lee
Seoul National University, Seoul, Republic of Korea
E-mail: hspark@spl.snu.ac.kr Tel/Fax: +82-2-880-8428
E-mail: kyoungmu@snu.ac.kr Tel/Fax: +82-2-880-1743
E-mail: sanguk@snu.ac.kr Tel/Fax: +82-2-880-8408

***Abstract*—In multi-view stereo setting, pixel correspondence problem and super resolution problem are inter-related in a sense that the result of each problem could help to solve the other. In this paper, we propose a novel method to solve two problems together by optimizing a unified energy functional. Main difference from the previous works is that the consistency between high resolution images is considered along with consideration to the consistency of high-resolution and low-resolution image pair with the same viewpoint. Experimental results show that our method outperforms the naive combination of single image super resolution and multi-view stereo method.**

I. Introduction

The goal of super resolution is to obtain high-resolution(HR) images from low-resolution(LR) images and it has been widely studied during past decades[1], [2], [3], [4], [5], [6]. One of the main approaches is to estimate one HR image from a number of LR input and it has attracted many researchers because of its theoretical clarity and neat formula. However, its sensitivity to the noise in matching between LR images limits its performance and prevents its use in real application. On the other hand, in stereo problem, another fundamental problem in computer vision, it is often required that the input images be high-quality. The reason is, in conventional binocular stereo, disparity or depth is estimated based on the matching between two images and usually the higher the resolution is, the better the matching become.

In this paper, we propose a method that solves super resolution and stereo problem in one unified framework. Unlike previous work, we consider image consistency between HR images as well as between HR image and LR image pairs. To do that, HR depth maps are estimated instead of LR ones. By considering geometric consistency between depth maps, we can obtain more reliable matchings robust to noise in images. These considerations prevent the error in one problem from propagating to the other problem and effectively cut the positive feedback loop of error accumulation. As a result, our HR image output is free from any undesirable artifact, actually increasing PSNR of reconstruction far higher than the simple bilinear interpolation, and the depth estimates become very accurate compared to the original multi-view stereo method based on LR input. Furthermore, Experimental results show that our method outperforms the naive combination of single image super resolution and multi-view stereo method.

II. Related Work

There are two main approaches to solve super resolution problem. One is learning-based or exemplar-based approach. In the methods that fall into this category, prior knowledge about the HR images such as edge statistics [1], [2] or patch correspondence between HR images and LR images are learned with some training set [3], [4], [5]. These methods produce visually pleasing results. However, there is no guarantee for the visually pleasing estimate obtained by these methods to be actually closer to the real ground truth.

Another main approach for super resolution is to use multiple LR images to reconstruct one HR image, which is well surveyed in [6]. Ideally, by maximizing the consistency between the HR image and the LR images, we can get the result which contain all details that each LR image has. These methods, however, are very sensitive to noise in matching and, thus, hard to apply to non-planar scene where computation of matching is complicated.

There are a few works that combine super resolution and stereo problem. In [7], the authors propose one energy functional which considers two problem simultaneously. However, the accuracy of both the depth map and HR image are not quite impressive, especially for super resolution result which is contaminated by some mosaic artifacts. Another method is to combine texture super resolution and 3D reconstruction problem [8]. While it actually shows improvement on both the reconstructed model and the texture map, it is hard to applicable in real situation because it is targeting 3D reconstruction not general stereo.

In this work, we propose energy functional that combine super resolution problem and multi-view stereo in one framework. We model the combined problem in a novel way so that the HR estimates are forced to be consistent with each other and LR input only have influence to its corresponding HR output. This prevents the complex computation related to warping and downsampling in matching in conventional super resolution formula. In section III, the details of our method will be discussed and the optimization technique will be briefly introduced in IV. Also, qualitative and quantitative evaluation of our method is done with some comparisons to the others in section V.

III. Proposal Method

The goal of our system is to recover HR images, I^H, and relevent HR depth maps, D^H, from given LR images, I^L. The basic assumption about these images is that the target scene should be static to facilitate stereo. Also, for the same reason, we assume that the information about the camera for each view is known up to similiraity transformation.

In the remaining parts of this paper, we use I_i^L to denote the input LR image from the ith viewpoint, while $I_i^L(\mathbf{x})$ represents the color value of this image at pixel position $\mathbf{x}$. Note that both the pixel position and color value are in vector form. Likewise, the notations for the corresponding HR image and HR depth map have been defined as I_i^H and D_i^H, respectively.

The energy functional of our method is defined as follows:

$$E\left(I^H, D^H | I^L\right) = \sum_i E_m\left(D_i^H, I_i^H | \bar{D}_i^H, \bar{I}_i^H\right) + E_c\left(I_i^H | I_i^L\right) + E_r\left(D_i^H\right), \tag{1}$$

where E_m, E_c, and E_r represent matching constraint, consistency constraint, and regularization constraint for depth map and image, respectively. These terms will be described one-by-one in the following subsections. Note that some of the energy terms defined in (1) is conditional, meaning the variables appearing after the bar is given as constants. The variables $\bar{D}_i^H$ and $\bar{I}_i^H$ represent the $D^H - \{D_i^H\}$ and $I^H - \{I_i^H\}$.

A. Matching Constraint

The conventional multi-frame super resolution problem can be modeled by a single equation using matrix-vector multiplication [6], as follows:

$$\mathbf{I}_i^L = \mathbf{S}\mathbf{B}_i\mathbf{W}_{ji}\mathbf{I}_j^H + \mathbf{n}_i. \tag{2}$$

In (2), the images are represented as vectors. This equation states that the HR image is firstly warped to viewpoint i by using $\mathbf{W}_{ji}$ and, then, is captured with low resolution. The capturing process is modeled by multiplication of blur matrix, $\mathbf{B}_i$, and downsampling matrix, $\mathbf{S}$, followed by addition of pixelwise-independent white Gaussian noise $\mathbf{n}_i$.

A.V.Bhavsar and A.N.Rajagopalan [7] use this equation in their method to solve super resolution and stereo simultaneously. However, the naive use of (2) could be problematic. Due to the complex computation related to warping and downsampling, the matching cost in [7] is computed under the uniform depth assumption around the target pixel position and HR images are estimated using Iterated-conditional-mode(ICM) optimization, which has brought mosaic artifacts.

We come up with a different approach. Instead of comparing the HR estimate with LR input images, we generate the HR estimate of all the input images and the matching information is computer based on the estimates instead of LR ones. Our model is formulated as follows:

$$E_m = \sum_{\mathbf{x}} \sum_{j \in N(i)} P\left(I_i^H(\mathbf{x}), I_j^H\left(W_{ji}\left(\mathbf{x}, D_i(\mathbf{x})\right)\right)\right) + G\left(\mathbf{x}, W_{ij}\left(W_{ji}\left(\mathbf{x}, D_i(\mathbf{x})\right), D_j\left(W_{ji}\left(\mathbf{x}, D_i(\mathbf{x})\right)\right)\right)\right). \tag{3}$$

In (3), $W_{ji}(\mathbf{x}, d)$ computes the projection of pixel $\mathbf{x}$ with depth value d in image i onto image j based on camera information of both frames. Function $P(\mathbf{c}, \mathbf{c}')$ and $G(\mathbf{x}, \mathbf{x}')$ are defined as follows:

$$P(\mathbf{c}, \mathbf{c}') = \frac{\|\mathbf{c} - \mathbf{c}'\|^2}{2\sigma_c^2}, \qquad G(\mathbf{x}, \mathbf{x}') = \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_g^2}, \tag{4}$$

where the parameter σ_c and σ_g control the sensitivity to difference in P and G, respectively. The geometric consistency term, G, serves as soft one-to-one correspondence as in [9]. We also employ the temporal selection scheme firstly proposed in [10] to handle occlusion more efficiently.

Using (3), we directly use HR information to compute the depth maps without any downsampling process. Also, because we impose photometric consistency between HR images, the reconstruction result is consistent across the viewpoints.

B. Consistency Constraint

We assume that each HR image should be consistent only with corresponding LR image. The benefit of this assumption is that we can replace complex cross-view blur model, which could possibly include motion blur, with much simpler one. The energy functional for this constraint is as follows:

$$E_i^c = \sum_{\mathbf{y}} \left| I_i^L(\mathbf{y}) - \sum_{\mathbf{x} \in w_{\mathbf{y}}} B(\mathbf{y}, \mathbf{x}) I_i^H(\mathbf{x}) \right|^2. \tag{5}$$

In (5), B is assumed to be a simple averaging weighting function representing box-filtering.

C. Regularization Constraint

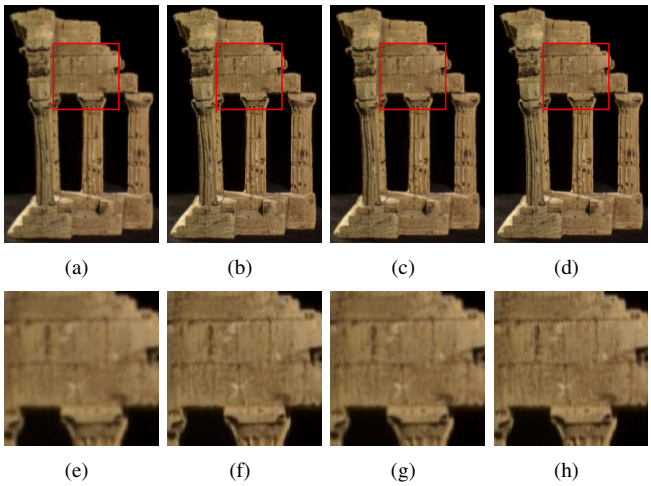
E_r in (1) are regularization function on a depth map. We adopt simple truncated linear smoothness function used in [9] as follows:

$$E_r = \sum_{\mathbf{x}} \sum_{\mathbf{y} \in N(\mathbf{x})} w_s \min\left\{\left|D_i^H(\mathbf{x}) - D_i^H(\mathbf{y})\right|, \eta\right\}. \tag{6}$$

In 6, $N(\mathbf{x})$ is a set of pixel around $\mathbf{x}$ and we assume simple 4-neighborhood system. Also, w_s and η are the user-defined parameters which represent the strength of smoothness and the threshold for truncation, respectively.

IV. Optimization

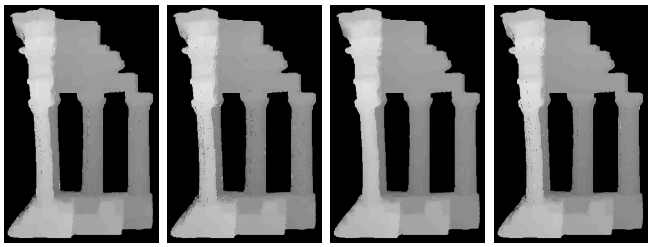
To minimize (1), we introduce iterative optimization approach, in which depth maps and HR images are updated by turns.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1. The comparison of super resolution results. (a) Bilinear interpolation. (b) Result of the SISR method [11]. (c) Result of the proposed method. (d) Ground-truth HR image. (e),(f),(g), and (h) are enlarged view of red rectangles in (a),(b),(c), and (d), respectively.



(a) (b) (c) (d)

Fig. 2. The comparison of multi-view stereo results. For (a), (b), and (d), the input images are each fixed to the result of linear interpolation, result of [11], and ground-truth HR images, respectively. The result of our method is represented in (c). Best viewed in electronic version.

A. Iterative Update

In depth estimation phase, we assume all the other output is fixed except for the target depth map, D_i^H. Then the energy functional becomes a pair-wise MRF for D_i^H. We optimize this MRF function by using tree-reweighted message passing algorithm(TRW-S). The depth maps are updated frame by frame.

For the update of HR images, we fix the depth variables as constants and update images frame by frame. We employ the optimization technique introduced in [12] utilizing the fact that our energy term become the energy of Gaussian distribution. Thanks to the use of this optimization, the super-resolved images has no mosaic artifacts.

B. Initialization

An initialization is necessary for our iterative optimization process to work. We simply upsample the LR images by using biliear interpolation to initialize the HR estimates and the corresponding initialization of depth maps can be obtained via minimizing (1) fixing HR images. During the initialization process, we define another matching constraint without G in (3).

$$E_i^m = \sum_{\mathbf{x}} \sum_{j \in N(i)} P\left(I_i^H(\mathbf{x}), I_j^H\left(W_{ji}\left(\mathbf{x}, D_i(\mathbf{x})\right)\right)\right). \quad (7)$$

In this equation, the geometric consistency term is removed, because we cannot evaluate the consistency between depth maps which is not computed yet.

V. Experimental Results

In order to evaluate the performance of the proposed algorithm, several experiments are carried out on the popular Middlebury data sets[13], [14]. Through all the experiments, the original input multi-view images are first downsampled and the original images are used as ground-truth data. Also. in our experiments, we set the parameters for regularization term, w_s and η, same as in [9]. That is, if we denote the depth range as $[d_{min}, d_{max}]$, then $w_s = 5/(d_{max} - d_{min})$ and $\eta = 0.05(d_{max} - d_{min})$. The other remaining parameters in our method are set as follows: $\sigma_c = 15$, $\sigma_d = 1$.

To show the effect of combined framework, the results are compared to those of the same algorithms as ours except for the HR estimates are fixed to (1) initialization using bilinear interpolation(referred to as *Bilinear+MVS*), (2) ground-truth HR images(referred to as *GT+MVS*), and (3) the result of [11](referred to as *SISR+MVS*). The last one is to compare our results to those of naive combination of single image super resolution and multi-view stereo.

These results are shown in Fig. 1 and Fig. 2 for *temple* dataset. As it is shown, the HR estimate of our method is close to the ground-truth HR image, comparable to that of [11]. It should be mentioned that [11] actually gives overly complex textures in some regions. Note that there are no mosaic artifacts in results of our method, unlike [7]. For quantitative comparison of super resolution performance, the PSNR score for each case is represented in Table I. In Fig. 2, depth estimation result of our method is. Our method outperforms the all the three other cases, including *GT+MVS*, surprisingly. The reason is thought to be the denoising effect of our algorithm obtained by adjusting color values to be consistent across the viewpoints. The result of *SISR+MVS* is more noisy even than *Bilinear+MVS*, due to the view-independent texture synthesis.

The quantitative analysis of our method in terms of multi-view stereo, is done for the *Art, Dolls, Reindeer, Moebius*, and *Books* datasets[14]. The results are illustrated in Fig. 3, and the quantitative evaluation is represented in Table II and

TABLE I
AVERAGE PSNR(DB) OF RECONSTRUCTED HR IMAGES FOR *temple*[13]

Method	*temple*
Proposed	**38.65**
Bilinear Interpolation	36.99
SISR[11]	38.41

Fig. 3. The results of our method with *Middlebury 2005* datasets. From left to right, each column shows ground-truth HR images, bilinear interpolation of LR images, HR estimates of our method, depth estimates of our method, and ground-truth depth map. The resuls are shown only for *Art* and *Dolls* datasets.

TABLE II
PSNR(DB) OF HR IMAGES FOR *Middlebury 2005* DATASETS[14]

Method	*Art*	*Dolls*	*Reindeer*	*Moebius*	*Books*
Proposed	33.43	32.48	33.47	33.25	**29.93**
Bilinear Interpolation	31.30	30.44	31.84	32.17	28.42
SISR[11]	**33.78**	**32.66**	**33.54**	**33.75**	29.72

TABLE III
DISPARITY ERROR RATIOS(%) FOR *Middlebury 2005* DATASETS[14]

Method	*art*	*dolls*	*reindeer*	*moebius*	*books*
Proposed	**3.81**	**3.22**	**0.99**	**2.72**	**6.66**
Bilinear+MVS	3.97	3.60	1.70	3.12	6.76
SISR+MVS	4.30	3.49	2.09	3.24	7.69

Table III. We used all seven views with the same illumination and exposure. As it can be seen from the table, while our method always outperforms the bilinear interpolation in terms of both super resolution and multi-view stereo, the use of HR image obtained by single image super resolution for multi-view stereo turns out to be problematic. Note that, although the PSNR scores are mostly best when using single image super resolution[11], ours are fairly comparable to theirs.

VI. CONCLUSIONS

In this paper, a new approach for the combining super resolution and stereo problem is proposed. Our method solve both problem in one framework by optimizing a unified energy functional. Unlike previous methods, we directly use HR images in matching, while imposing consistency between LR images and HR images only for the same viewpoint. This novel formulation improves accuracy of both multi-view stereo and super resolution results and, especially, eliminates mosaic artifacts from the HR output caused by use of ICM optimization. Experimental results show that the qualities of super resolution and stereo results are better than that of the case where each are handled independently.

REFERENCES

[1] Allebach, J., Wong, P.W.: Edge-directed interpolation. In: Image Processing, 1996. Proceedings., International Conference on. Volume 3. (1996) 707 –710 vol.3

[2] Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y.: Soft edge smoothness prior for alpha channel super resolution. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1 –8

[3] Freeman, W., Jones, T., Pasztor, E.: Example-based super-resolution. Computer Graphics and Applications, IEEE **22** (2002) 56 –65

[4] Sun, J., Zheng, N.N., Tao, H., Shum, H.Y.: Image hallucination with primal sketch priors. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Volume 2. (2003) II – 729–36 vol.2

[5] Kong, D., Han, M., Xu, W., Tao, H., Gong, Y.H.: Video Super-resolution with Scene-specific Priors. In: British Machine Vision Conference. (2006) 549–558

[6] Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine **20** (2003) 21–36

[7] Bhavsar, A., Rajagopalan, A.: Resolution enhancement in multi-image stereo. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010) 1721 –1728

[8] Tung, T., Nobuhara, S., Matsuyama, T.: Simultaneous super-resolution and 3d video using graph-cuts. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (2008) 1 –8

[9] Zhang, G., Jia, J., Wong, T.T., Bao, H.: Consistent depth maps recovery from a video sequence. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 974–988

[10] Kang, S.B., Szeliski, R.: Extracting view-dependent depth maps from a collection of images. Int. J. Comput. Vision **58** (2004) 139–163

[11] Kim, K.I., Kwon, Y.: Example-based learning for single-image super-resolution. In: Proceedings of the 30th DAGM symposium on Pattern Recognition, Berlin, Heidelberg, Springer-Verlag (2008) 456–465

[12] Levi, E.: Using natural image priors-maximizing or sampling? Masters thesis, The Hebrew University of Jerusalem (2009)

[13] Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1. CVPR '06, Washington, DC, USA, IEEE Computer Society (2006) 519–528

[14] Scharstein, D., Pal, C.: Learning conditional random fields for stereo. IEEE Proc. Conf. Computer Vision and Pattern Recognition (2007) 1–8