Simultaneous Super-Resolution of Depth and Images using a Single Camera

Hee Seok Lee Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University, 151-742, Seoul, Korea
ultra21@snu.ac.kr kyoungmu@snu.ac.kr
http://cv.snu.ac.kr

Abstract

In this paper, we propose a convex optimization framework for simultaneous estimation of super-resolved depth map and images from a single moving camera. The pixel measurement error in 3D reconstruction is directly related to the resolution of the images at hand. In turn, even a small measurement error can cause significant errors in reconstructing 3D scene structure or camera pose. Therefore, enhancing image resolution can be an effective solution for securing the accuracy as well as the resolution of 3D reconstruction. In the proposed method, depth map estimation and image super-resolution are formulated in a single energy minimization framework with a convex function and solved efficiently by a first-order primal-dual algorithm. Explicit inter-frame pixel correspondences are not required for our super-resolution procedure, thus we can avoid a huge computation time and obtain improved depth map in the accuracy and resolution as well as high-resolution images with reasonable time. The superiority of our algorithm is demonstrated by presenting the improved depth map accuracy, image super-resolution results, and camera pose estimation.

1. Introduction

In 3D reconstruction with a single camera, the accuracy of camera pose and scene structure estimation is highly affected by the conditions of input images such as noise, contrast, blur, and resolution. In particular, image resolution is an important factor for achieving sufficient accuracy of various geometry-related computer vision algorithms including 3D reconstruction, since it influence the feature detection, localization and matching. Even in an image of a scene, the resolutions of objects vary according to their sizes and depths. Note that small measurement error does not bring large errors in object position and camera pose when an object is close to the camera, while, it does significantly when the object is far from the camera. Therefore, it is necessary to enhance the image resolution to reduce the sensitivity to the image measurement error and achieve reliable and accurate 3D reconstruction.

Image super-resolution, the method for enhancing image resolution, has two different approaches: reconstruction-based approach and learning-based approach. The reconstruction-based approach, which is related to our approach, infers the high-resolution pixel by merging multiple observations of a target pixel. Multiple observations are obtained by finding corresponding pixels through an image sequence. Therefore, finding accurate pixel-wise correspondences is the key for the success of the reconstruction-based super-resolution. For general scenes, these correspondences can be obtained up to sub-pixel accuracy using optical flow algorithms. However, optical flow in low-resolution images usually do not provide enough accuracy in correspondences, producing unsatisfactory results. Some iterative methods [4, 7] alternately estimate a high-resolution image and pixel correspondences, and show better results. However, these methods usually take a very large amount of computation time, and thus they are not appropriate for real-time applications such as visual odometry and SLAM.

Note that if we employ the information about the 3D scene geometry, the super-resolution problem can be solved more efficiently since we can directly use it for enhancing the accuracy of the correspondences. That is, with estimated camera poses, the problem of finding pairwise pixel correspondences through an image sequence can be converted into estimating the depth value of corresponding pixels. Although this converted problem has an error source related to the camera pose error, because it is casted in a much lesser dimensional solution space than the original pairwise correspondence problem, it can be solved much easily and faster. Therefore, depth reconstruction and super-resolution problems are interrelated and boost each other's accuracy. So, in this work, we combine the depth estimation and the high-resolution image estimation in a unified framework, and propose a simultaneous solution to both problems.

In the proposed method, the depth estimation and image super-resolution are formulated with a single convex energy function, which consists of data term and regularization term. The solution is estimated by convex optimization of the energy function. Although both pixel correspondences (re-parameterized by depth) and high-resolution image are estimated, the computational cost is not so expensive compared to the conventional high-resolution image estimation only because we do not use alternating methods like EM. Additionally, due to the simultaneous estimation of depth and high-resolution image, the results of the two problems are greatly enhanced.

2. Related works

In this section, we review some works that are similar to our work in combine 3D reconstruction and super-resolution. Then, we discuss the works on the primal-dual algorithm for 3D reconstruction or super-resolution.

2.1. 3D reconstruction and image super resolution

In [1, 9, 14, 5], the close relationship between super-resolution and 3D scene structure is pointed out and their cooperative solution is studied. In [9], the super-resolution is formulated with the calibrated 3D geometry and solved using the MAP-MRF framework. Occlusions are effectively handled in their super-resolution method using depth information, but super-resolution does not contribute to depth map estimation in this method. In [14], a method for increasing the accuracy of 3D video reconstruction using multiple static cameras is presented. The 3D video is composed of texture images and 3D shapes, and increasing their accuracy is achieved by simultaneous super-resolution using MRF formulation and graph-cuts. High-quality texture and 3D reconstruction is presented in [5] where texture and shape of a 3D model are alternately estimated with joint energy functional. Compared to [5] our work has more challenging settings in which neither accurate camera motions nor initial pixel correspondences are available.

The work most closely related to ours with respect to its objective is [1]. The authors formulate a full frame super-resolution problem combined with a depth map estimation problem, and attempt to enhance the results of both problems. However, their solution is not fully simultaneous but follows an EM-style alternating method instead. They fix the current high-resolution image for the estimation of the depth map, and vice versa. Graph-cut and iterated conditional modes (ICM) are used for the depth and high-resolution image estimation, respectively, for each iteration, which result in an inevitably large computation cost. In contrast, we search the globally optimum solution directly with a single convex energy function and achieve very fast optimization speed for dense real-time 3D reconstruction.

2.2. Primal-dual algorithm for 3D reconstruction and super-resolution

The formulation of our algorithm is based on the variational approach, especially the primal-dual algorithm [2, 3, 6]. The first-order primal-dual algorithm is a very effective tool for convex variational problems due to its parallelizable characteristics. The algorithm has been used in various computer vision problems, with the wide use of parallel computing acceleration such as general-purpose computing on graphics processing units (GPGPU).

The first-order primal-dual algorithm has been applied recently for the 3D reconstruction and super-resolution problems. In [10] and [13], a dense 3D reconstruction is studied and its real-time implementations are demonstrated. They used conventional energy functions consisting of photometric consistency-based data term and L^1 or Huber norm-based smoothness term, but achieved a breakthrough performance in computation time using the primal-dual algorithm combined with the GPGPU implementation.

In [15], the first-order primal-dual algorithm is applied to the super-resolution problem. The reconstruction-based super-resolution is formulated by image downsampling, blurring, and warping, and then the latent high-resolution image is estimated with the Huber norm regularization. This method achieves a fast computation of high-quality super-resolution comparable to other methods, but has certain limitations such that highly accurate initial image warping is required and no updating procedure is involved in estimating the super-resolution.

Our novel combined 3D reconstruction and super-resolution algorithm is also formulated in the first-order primal-dual framework. However, unlike [10] and [13], the proposed super-resolution combined framework enables more accurate depth map estimation with respect to its resolution. Our image super-resolution is also accelerated by finding pixel correspondences in a depth domain instead of optical flows between images with the help of camera geometry obtained from the 3D reconstruction.

3. Model

In this work, we propose a new energy function for a simultaneous estimation of depth map and high-resolution image. The inputs are $M \times N$ size low-resolution image sequence $\mathbf{I}_j \in \mathbb{R}^{MN}$ and their corresponding camera poses $\mathbf{P}_j \in \mathbb{SE}(3)$ with $j \in \{0, ..., J\}$. Let $\mathbf{g} \in \mathbb{R}^{s^2MN}$ be the latent super-resolution image with the gray scale, and $\mathbf{d} \in \mathbb{R}^{s^2MN}$ be the latent inverse depth map, where s is the predefined upscale factor. The solution of $\mathbf{g}$ and $\mathbf{d}$ is estimated with respect to the reference view $\mathbf{P}_0$. The energy function to solve this problem is composed of the data cost E_{data} based on the photometric constancy and the regularization cost E_{reg} for smoothing undesirable artifacts. With

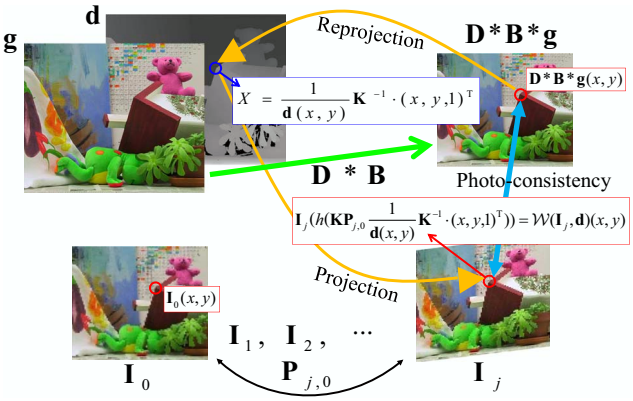


Figure 1. The relationship between the low-resolution input sequence $\mathbf{I}_j$ and the super-resolution image $\mathbf{g}$, induced by the depth map $\mathbf{d}$: The photometric consistency should hold for $\mathbf{I}_j$ and the simulated low-resolution image $\mathbf{D} * \mathbf{B} * \mathbf{g}$.

the parameter λ which controls the degree of regularization, the energy function has the form $E(\mathbf{g}, \mathbf{d}) = E_{reg} + \lambda E_{data}$. The super-resolution image $\mathbf{g}$ can also be the color, but we use the gray scale notation here for simplicity and show the color image results in the experiment section.

3.1. Data cost

We start with the relationship between the high-resolution image $\mathbf{g}$ for the reference image $\mathbf{I}_0$ and the low-resolution image $\mathbf{I}_j$ from an adjacent view. With the camera internal parameter $\mathbf{K}$ including the focal length and the principal point, the reprojected 3D position X of pixel (x, y) in $\mathbf{I}_0$ with the inverse depth $\mathbf{d}(x, y)$ by the reference camera $\mathbf{P}_0$ is given by $X = \frac{1}{\mathbf{d}(x,y)}\mathbf{K}^{-1} \cdot (x, y, 1)^\top$, and its projection to the adjacent view with $\mathbf{P}_j$ is calculated as $h(\mathbf{K}\mathbf{P}_{j,0}\frac{1}{\mathbf{d}(x,y)}\mathbf{K}^{-1} \cdot (x, y, 1)^\top)$, where $\mathbf{P}_{j,0} = \mathbf{P}_j\mathbf{P}_0^{-1}$ and h is the dehomogenization function such that $h((x, y, z)^\top) = (x/z, y/z)$. Fig. 1 illustrates these relationships.

For notational simplicity, the non-bold characters g and d are used for the pixel-wise values $\mathbf{g}(x, y)$ and $\mathbf{d}(x, y)$, respectively, and their corresponding dual variables later. We define the image warping $\mathcal{W}(\mathbf{I}_j, \mathbf{d})$, which transforms the image $\mathbf{I}_j$ to the reference image $\mathbf{I}_0$, using the pixel projection and reprojection discussed above,

$$\mathcal{W}(\mathbf{I}_j, \mathbf{d})(x, y) = \mathbf{I}_j(h(\mathbf{K}\mathbf{P}_{j,0}\frac{1}{d}\mathbf{K}^{-1} \cdot (x, y, 1)^\top)). \quad (1)$$

Then, by the photometric consistency between the reference image and the adjacent image, the equation

$$\mathbf{I}_0(x, y) = \mathbf{I}_j(h(\mathbf{K}\mathbf{P}_{j,0}\frac{1}{d}\mathbf{K}^{-1} \cdot (x, y, 1)^\top)) = \mathcal{W}(\mathbf{I}_j, \mathbf{d})(x, y) \quad (2)$$

holds for all $j \in \{0, ..., J\}$ if the inverse depth d has the exact value. By incorporating the image resolution degradation model, the equation

$$(\mathbf{D} * \mathbf{B} * \mathbf{g})(x, y) = \mathbf{I}_0(x, y) = \mathcal{W}(\mathbf{I}_j, \mathbf{d})(x, y) \quad (3)$$

also holds for all $j \in \{0, ..., J\}$. Here, $\mathbf{D}$ and $\mathbf{B}$ are the downsampling and the blurring operator, respectively. From the equality in Eq. (3), we can set our objective which finds an optimum value of $\mathbf{g}$ and $\mathbf{d}$, such that

$$\arg\min_{\mathbf{g},\mathbf{d}} \sum_{j=0}^{J} \|\mathbf{D} * \mathbf{B} * \mathbf{g} - \{\mathcal{W}(\mathbf{I}_j, \mathbf{d})\}\|_1. \quad (4)$$

To find the optimized value of $\mathbf{d}$ through an iterative update, we apply the first-order Taylor expansion to $\mathcal{W}(\mathbf{I}_j, \mathbf{d})$ to approximate a change in image $\mathcal{W}(\mathbf{I}_j, \mathbf{d})$ with respect to a small change of depth at the initial value $\mathbf{d}_0$,

$$\mathcal{W}(\mathbf{I}_j, \mathbf{d}) \simeq \mathcal{W}(\mathbf{I}_j, \mathbf{d}_0) + \frac{\partial}{\partial \mathbf{d}}\mathcal{W}(\mathbf{I}_j, \mathbf{d})\Big|_{\mathbf{d}=\mathbf{d}_0} \cdot (\mathbf{d} - \mathbf{d}_0). \quad (5)$$

Then, our objective (4) can be rewritten as a linearized form,

$$\arg\min_{\mathbf{g},\mathbf{d}} \sum_{j=0}^{J} \|\mathbf{D} * \mathbf{B} * \mathbf{g} - \{\mathcal{W}(\mathbf{I}_j, \mathbf{d}_0) + \mathbf{I}_{j\mathbf{d}} \cdot (\mathbf{d} - \mathbf{d}_0)\}\|_1, \quad (6)$$

where $\mathbf{I}_{j\mathbf{d}}$ is the simplified notation of the image derivative $\frac{\partial}{\partial \mathbf{d}}\mathcal{W}(\mathbf{I}_j, \mathbf{d})$, which can be calculated pixel-wise using the chain-rule,

$$\mathbf{I}_{j\mathbf{d}} = \frac{\partial \mathcal{W}(\mathbf{I}_j, \mathbf{d}_0)}{\partial d} = \frac{\partial \mathcal{W}(\mathbf{I}_j, \mathbf{d}_0)}{\partial x}\frac{\partial x}{\partial d} + \frac{\partial \mathcal{W}(\mathbf{I}_j, \mathbf{d}_0)}{\partial y}\frac{\partial y}{\partial d}. \quad (7)$$

The blur kernel $\mathbf{B}$ is predefined with the simple Gaussian blur model, with the standard deviation s and the kernel size of $(s-1)^{1/2}$. To handle the downsampling operator $\mathbf{D}$ efficiently, we upscale the low-resolution input images to the high-resolution size $sM \times sN$ as $\mathbf{I}_j \in \mathbb{R}^{MN} \rightarrow \hat{\mathbf{I}}_j \in \mathbb{R}^{s^2MN}$ using bicubic interpolation and perform the optimization process with the resized image space $\mathbb{R}^{s^2MN}$. The resulting data cost then has the form,

$$\begin{aligned} E_{data} &= \int_{X,Y} \rho(\mathbf{g}, \mathbf{d}) \\ &= \int_{X,Y} \sum_{j=0}^{J} \|\mathbf{B} * \mathbf{g} - \{\mathcal{W}(\hat{\mathbf{I}}_j, \mathbf{d}_0) + \hat{\mathbf{I}}_{j\mathbf{d}}(\mathbf{d} - \mathbf{d}_0)\}\|_1. \end{aligned} \quad (8)$$

Fig. 2 shows an example of the convexity of data cost $\rho(\mathbf{g}, \mathbf{d})$ for different image points. The shape of the cost function is obviously convex, but the shape of the function varies from image point to point according to the image gradient. In a low texture region, the data cost is dominated by the high-resolution intensity $\mathbf{g}$ than the depth $\mathbf{d}$. Therefore, regularization is required to get a more plausible solution for depth $\mathbf{d}$.

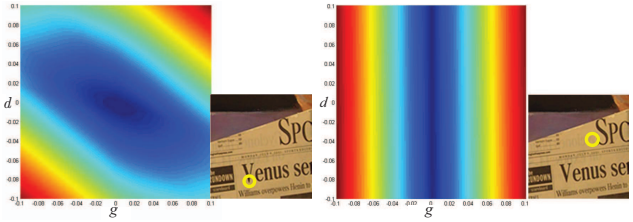


Figure 2. The shape of data cost $\rho(\mathbf{g}, \mathbf{d})$ for textured (left) and untextured (right) region.

3.2. Regularization

For image intensity $\mathbf{g}$ and inverse depth $\mathbf{d}$, we use a Huber norm based regularization to get a smoothed and discontinuity-preserved result. The Huber norm for $\mathbf{g}$ is defined by following pixel-wise function:

$$\|\nabla \mathbf{g}\|_{\alpha_g}(x,y) = \begin{cases} \frac{|\nabla g|^2}{2\alpha_g}, & \text{if} |\nabla g| \leq \alpha_g \\ |\nabla g| - \frac{\alpha_g}{2}, & \text{if} |\nabla g| > \alpha_g \end{cases}, \quad (9)$$

where ∇ is the linear operator that computes derivatives of x and y direction. The Huber norm for $\|d\|_{\alpha_d}$ is defined in the same way. In our implementation, we set $\alpha_g = \alpha_d = 0.001$.

By combining the data cost (8) and the regularization (9), we get our objective energy function $E(\mathbf{g}, \mathbf{d})$,

$$E(\mathbf{g}, \mathbf{d}) = \int_{X,Y} \|\nabla \mathbf{g}\|_{\alpha_g} + \|\nabla \mathbf{d}\|_{\alpha_d} + \lambda \rho(\mathbf{g}, \mathbf{d}). \quad (10)$$

In the next section, we describe the solution of this energy function.

4. Solution

4.1. Initial depth estimation

In the data cost (8), the first-order Taylor expansion, which can only handle a small update for $\mathbf{g}$, and $\mathbf{d}$ is applied. This step requires the starting point of optimization to be close to the global optimum. The initial value of $\mathbf{g}$ can be easily obtained by upscaling the input image at reference view using simple bicubic interpolation. However, the initial value of $\mathbf{d}$ should be estimated using the low-resolution input sequence.

The cost function for initial depth estimation is easily obtained from Eq. (8) and (10) by replacing $\mathbf{B} * \mathbf{g}$ and $\hat{\mathbf{I}}_j$ with the low-resolution images $\mathbf{I}_0$ and $\mathbf{I}_j$, respectively, and removing the regularization on $\mathbf{g}$. The resulting energy function for low-resolution depth map $\check{\mathbf{d}}$ is

$$E(\check{\mathbf{d}}) = \int_{X,Y} \|\check{\mathbf{d}}\|_{\alpha_d} + \lambda \sum_{j=1}^{J} \|\mathbf{I}_0 - \{\mathcal{W}(\mathbf{I}_j, \check{\mathbf{d}}_0) + \mathbf{I}_{j\check{\mathbf{d}}} \cdot (\check{\mathbf{d}} - \check{\mathbf{d}}_0)\}\|_1. \quad (11)$$

The equation (11) is actually a conventional formulation for depth map estimation. The optimization of this energy function is almost similar to the optimization of Eq. (10), which will be explained below, so the optimization of (11) is skipped here. The limitation of a small update also holds for Eq. (11). Thus, a coarse-to-fine approach is used to approach the global optimum of $\mathbf{d}$ gradually by starting from an arbitrary initial solution, *e.g.*, filled with 1.0. The depth result obtained at the finest level is upscaled using bicubic interpolation and is fed to the optimization of (10) as an initial value.

4.2. High-resolution image and depth estimation

Now we will describe a solution of Eq. (10) based on the first-order primal-dual optimization algorithm. By interpreting our objective function (10) as the primal-dual formulation, we can rewrite it as a generic saddle point problem with the dual variables $\mathbf{p}$ and $\mathbf{q}$, which corresponds to $\mathbf{g}$ and $\mathbf{d}$, respectively:

$$\min_{\mathbf{g},\mathbf{d}} \max_{\mathbf{p},\mathbf{q}} \langle \nabla \mathbf{g}, \mathbf{p} \rangle + \langle \nabla \mathbf{d}, \mathbf{q} \rangle + \lambda \|\rho(\mathbf{g}, \mathbf{d})\|_1 - \delta_{\mathbf{P}}(\mathbf{p}) - \frac{\alpha_g}{2}\|\mathbf{p}\|_2^2 - \delta_{\mathbf{Q}}(\mathbf{q}) - \frac{\alpha_d}{2}\|\mathbf{q}\|_2^2, \quad (12)$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, and the functions $\delta_{\mathbf{P}}$ and $\delta_{\mathbf{Q}}$ are the indicator functions given as $\delta_{\mathbf{P}}(\mathbf{p}) = \begin{cases} 0, & \text{if } \|\mathbf{p}\|_\infty \leq 1 \\ \infty, & \text{if else.} \end{cases}$ and $\delta_{\mathbf{Q}}(\mathbf{q}) = \begin{cases} 0, & \text{if } \|\mathbf{q}\|_\infty \leq 1 \\ \infty, & \text{if else.} \end{cases}$, respectively.

This problem can be optimized through the iteration,

$$\begin{aligned} (\mathbf{p}, \mathbf{q})^{n+1} &= \mathcal{R}_{\mathbf{p},\mathbf{q}} \left((\mathbf{p}, \mathbf{q})^n + \sigma \nabla (\bar{\mathbf{g}}, \bar{\mathbf{d}})^n \right) \\ (\mathbf{g}, \mathbf{d})^{n+1} &= \mathcal{R}_{\mathbf{g},\mathbf{d}} \left((\mathbf{g}, \mathbf{d})^n - \tau \nabla^* (\bar{\mathbf{p}}, \bar{\mathbf{q}})^n \right) \\ (\bar{\mathbf{g}}, \bar{\mathbf{d}})^{n+1} &= 2(\mathbf{g}, \mathbf{d})^{n+1} - (\bar{\mathbf{g}}, \bar{\mathbf{d}})^n. \end{aligned} \quad (13)$$

where the operator ∇^*, the conjugate of ∇ as $\nabla^* = -\operatorname{div}$, computes the divergence [2], and $\bar{\mathbf{g}}$ and $\bar{\mathbf{d}}$ are the intermediate variables for the convergence of algorithm. The initial value $(\mathbf{g}, \mathbf{d})^0$ is obtained from Section 4.1, and $(\mathbf{p}, \mathbf{q})^0$ is set to zero. The operators $\mathcal{R}_{\mathbf{p},\mathbf{q}}$ and $\mathcal{R}_{\mathbf{p},\mathbf{q}}$ are the resolvent operators that search lower energy values using subgradients. τ and σ are constants that control the convergence of primal and dual variable, respectively. The resolvent operators will be discussed in more detail.

Our regularization term (10) is a typical form used in [2]. Thus, the resolvent operator of the dual variables is a pixel-wise projection

$$\mathcal{R}_{p,q}(p,q) = \left(\frac{p}{max(1,|p|)}, \frac{q}{max(1,|q|)} \right). \quad (14)$$

On the other hand, the data cost has a difference with the standard form in previous primal-dual algorithm applications. This difference comes from the summation of absolute value in the data cost for image sequence. Since we use

a L^1 norm for the difference between two images, there are some critical (non-differentiable) points in their summation. Therefore, these non-differentiability should be handled in the optimization procedure. The minimization of similar cost function is introduced in [13], but the solution space of [13] is for the depth map only, so the minimization can be efficiently achieved by evaluating and sorting all critical points. On the other hand, the solution space of our problem is composed of depth map and image intensity, so there are J^2 critical points. Searching them is not straightforward, and thus optimization by evaluating and sorting critical points is inefficient. Instead, the general gradient descent and critical point searching are combined to accelerate the minimization procedure.

Let per-image data cost $\|\rho_j(\mathbf{g},\mathbf{d})\|_1 = \|\mathbf{B} * \mathbf{g} - \{\mathcal{W}(\hat{\mathbf{I}}_j, \mathbf{d}_0) + \hat{\mathbf{I}}_{j\,\mathbf{d}} \cdot (\mathbf{d}-\mathbf{d}_0)\}\|_1$, then we can write $\rho(\mathbf{g},\mathbf{d})$ as

$$\rho(\mathbf{g},\mathbf{d}) = \sum_{j=0}^{J} \|\rho_j(\mathbf{g},\mathbf{d})\|_1 = \sum_{j=0}^{J} sgn(\rho_j(\mathbf{g},\mathbf{d})) \cdot \rho_j(\mathbf{g},\mathbf{d}), \tag{15}$$

where $sgn(\cdot)$ is a signum function. Then the derivatives of (15) are calculated as

$$\partial\rho(\mathbf{g},\mathbf{d}) = \sum_{j=0}^{J} sgn(\rho_j(\mathbf{g},\mathbf{d})) \cdot \left(1, -\hat{\mathbf{I}}_{j\,\mathbf{d}}^{\top}\right). \tag{16}$$

We divide the domain of resolvent operator based on the cost ρ and the magnitude of gradient $\|\partial\rho\|_2^2$, and apply the gradient descent search and critical point search,

$$\mathcal{R}_{g,d}(g,d) = \begin{cases} (g,d) - \tau\lambda\left(\partial\rho(g,d)\right), \\ \quad \text{if } \rho(g,d) > \tau\lambda\|\partial\rho(g,d)\|_2^2 \\ (g,d) - \frac{\rho_j^*(g,d)\cdot\partial\rho_j^*(g,d)}{\|\partial\rho_{j^*}(g,d)\|_2^2}, \\ \quad \text{if } \rho(g,d) < \tau\lambda\|\partial\rho(g,d)\|_2^2 \end{cases}, \tag{17}$$

where

$$j^* = \underset{\{j|\rho_j(g,d)\cdot sgn(\nabla\rho(g,d))>0\}}{\arg\min} \|\rho_j(g,d)\|_1. \tag{18}$$

The operation of the second case in (17) is searching the closest critical point with a lower cost value by (18), and moving the variable to this critical point. By iterating Eq. (13) and checking the amount of changes in total cost (10), we can terminate the iteration and can get the final results of $\mathbf{g}$ and $\mathbf{d}$.

5. Implementation of 3D Reconstruction

5.1. Camera localization

To use the proposed depth map estimation and super-resolution algorithm in the single camera 3D reconstruction system, the camera localization algorithm needs to be incorporated. Before the depth map is estimated for an initial few frames, we rely on the sparse point feature-based SLAM for camera localization. After the initial depth map is created, the image registration method similar to the 2.5D image registration in [10] is used between the input frame and the pre-warped image from the estimated high-resolution image and depth map to estimate a new camera pose $\mathbf{P}_{J+1}$ as:

$$\begin{aligned}\mathbf{P}_{J+1} = &\arg\max_{\mathbf{P}} \int_{X,Y} \|\mathbf{g}(x,y) - \\ &\mathbf{I}_{J+1}(h(\mathbf{K}\mathbf{P}\mathbf{P}_0^{-1}\frac{1}{d}\mathbf{K}^{-1}\cdot(x,y,1)^{\top}))\|.\end{aligned} \tag{19}$$

The optimization of this function can be achieved by predicting $\mathbf{P}_{J+1}$ using the motion dynamics and iteratively approaching to optimum value using the gradient-based method.

There are advantages to estimating a camera pose using high-resolution image $\mathbf{g}$. The image registration can be robust to image degradation such as image noise, downsampling, and blurring. Since the input images are the degraded version of a scene by those effects, the recorded images are different from the real appearance of the scene. The estimated image $\mathbf{g}$ can be regarded as the most probable appearance of a real scene, because it is estimated from a number of instance images.

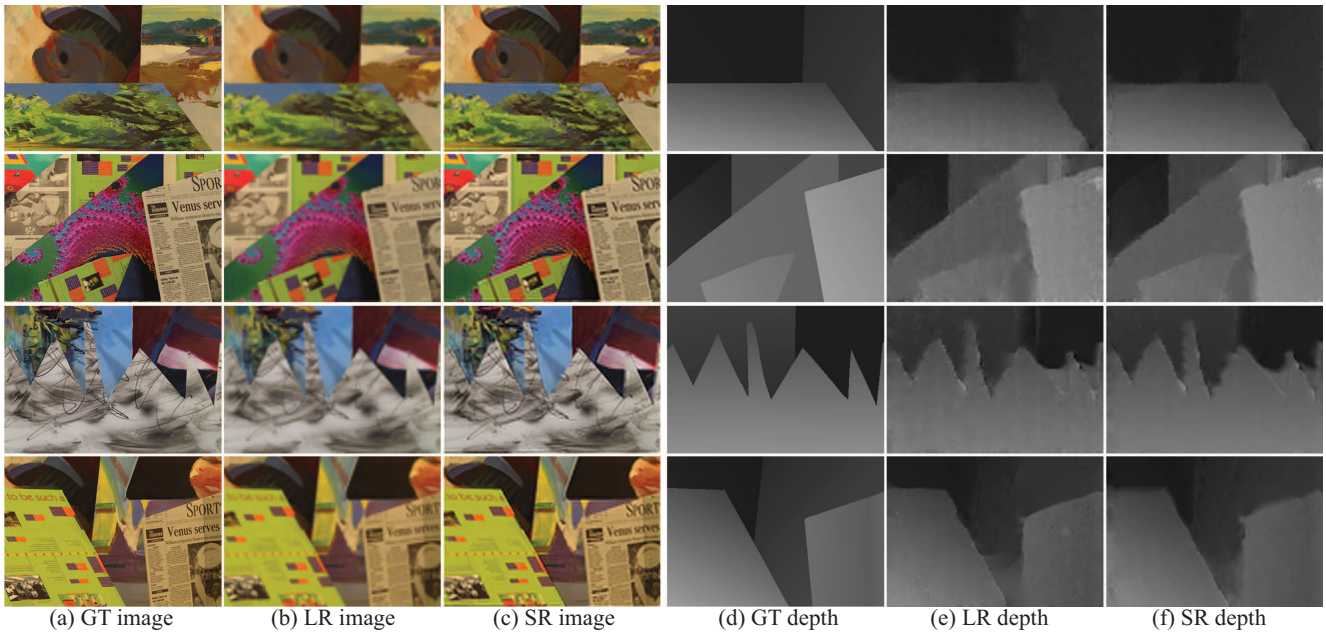
5.2. Map management

Our method estimates an inverse depth map instead of 3D points of sparse features or full 3D surface; hence, the map does not increase continuously. The depth map is reconstructed for some selected keyframes, and the relationship between depth maps is calculated and stored as a relative representation [8]. Although the depth map-based representation does not provide a visually attractive 3D surface, it has the advantage that the depth map merging step which takes large amount of computation is not required in this representation.

When the overlap between the reconstructed depth map and the current input image goes below threshold, then we perform a new depth map and high-resolution image estimation. The overlapped depth map is propagated to new depth estimation and used as an initial value. The relative pose between the previous keyframe and the new keyframe is stored, and the current camera pose is set to identity. The camera poses for subsequent frames are estimated with respect to the current keyframe's pose.

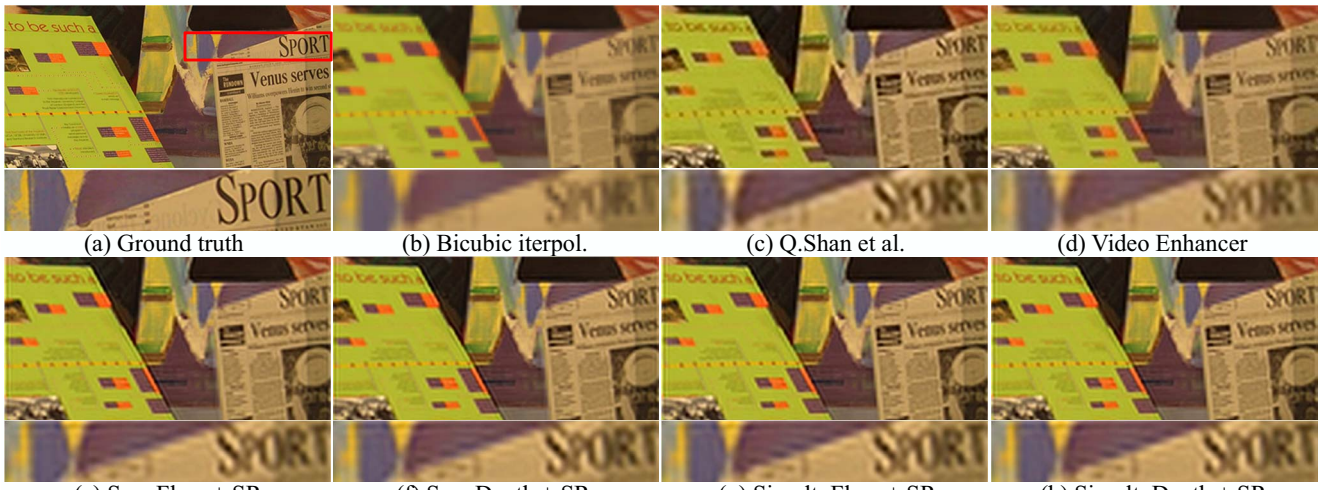
6. Experiments

We implement the proposed algorithm using NVIDIA's CUDA for GPGPU parallelization, and test the implementation using 3.3GHz quad core processor and GeForce GTX



(a) GT image (b) LR image (c) SR image (d) GT depth (e) LR depth (f) SR depth

Figure 3. Depth map estimation and super-resolution results on the synthesized low-resolution image sequences *Bull*, *Poster*, *Sawtooth*, and *Venus* in [11]. (a) Original images. (b) Synthesized low-resolution images. (c) Super resolution images. (d) Ground truth depth. (e) Depth map without super-resolution. (f) Depth map with super-resolution.



(a) Ground truth (b) Bicubic iterpol. (c) Q.Shan et al. (d) Video Enhancer

(e) Seq. Flow + SR (f) Seq. Depth + SR (g) Simult. Flow + SR (h) Simult. Depth + SR

Figure 4. Comparison of super-resolution results ($\times$ 4) on the synthesized *Venus* sequence with other super-resolution methods.

570 which has 480 stream processors. The algorithm performance is evaluated by three factors; super-resolution result, depth map estimation result, and registration error for camera localization. We evaluate our algorithm by performing a quantitative analysis using synthetic data and a feasibility test using real image sequence.

6.1. Results on simulated data

We use images and depth maps from [11], which have no occlusion information. For a given high-resolution image and its ground truth depth map from a reference view, the low-resolution image set is synthesized by warping and downsampling the high-resolution image. The virtual camera motion is simulated with a combination of arbitrary translation and rotation, and 20 low-resolution images of one-fourth scale (for example, 109 $\times$ 96 size for *Venus* image data) are obtained. The super-resolution image and depth map are estimated with their original scale, and their errors with respect to ground truth are calculated.

Fig. 3 shows our results on the synthetic data. The low-resolution images and depth maps Fig. 3-(b, e) are obtained by bicubic interpolation of the input images and initial depth maps. From the results of the proposed algorithm shown in Fig. 3-(c, f), we can see the improved depth map re-

Table 1. PSNR (in dB), SSIM (Structural similarity, closer to 1 is better), and computation time (in second) of various super-resolution algorithm.

Image		Bicubic	[12]	[16]	Seq. Flow+SR	Seq. Depth+SR	Sim. Flow+SR	Sim. Depth+SR
Bull	PSNR	15.69	16.08	16.59	16.78	16.76	16.82	**16.83**
	SSIM	0.77	0.7762	**0.79**	0.78	0.78	0.79	**0.79**
Poster	PSNR	13.71	12.69	**13.98**	13.65	13.67	13.85	13.87
	SSIM	0.54	0.57	0.57	0.56	0.56	**0.57**	0.57
Sawtooth	PSNR	12.67	12.63	13.17	12.91	12.86	**13.20**	13.19
	SSIM	0.66	0.67	0.69	0.67	0.67	**0.69**	0.69
Venus	PSNR	15.14	14.75	15.66	15.74	15.74	**15.87**	15.86
	SSIM	0.71	0.71	0.72	0.72	0.72	**0.73**	0.73
Avg. comp. time		-	22.93	1.21	19.05	1.625	18.26	0.97

sult as well as super-resolution image. In the closed-up region, the low-resolution input image has a degraded texture which makes depth estimation difficult. By recovering high-resolution texture using super-resolution, we can also recover the correct depth map.

Various methods for the super-resolution are tested to analyze the accuracy and efficiency of our algorithm. To test the contribution of simultaneous estimation, we replace our simultaneous formulation with the sequential method. In the sequential algorithm, the energy function (10) is minimized with a fixed $\mathbf{g}$ obtained from the bicubic interpolation of reference view, and then $\mathbf{g}$ is estimated with the obtained $\mathbf{d}$ fixed. The result of sequential method is shown in Fig. 4-(f), where we can see the limitation of sequential methods in the quantitative analysis in Table 1.

The efficiency of depth based formulation for super-resolution is also verified by comparing the results and computation time with the pairwise correspondence (optical flow) based formulation in which the optical flow vectors between the reference view and the other view are estimated simultaneously. The objective has a form similar to Eq. (6) as follow:

$$\underset{\mathbf{g},\mathbf{v}_1,\ldots,\mathbf{v}_J}{\arg\min} \sum_{j=0}^{J} \|\mathbf{D}*\mathbf{B}*\mathbf{g}-\{\mathcal{W}(\mathbf{I}_j,\mathbf{v}_j)+\mathbf{I}_{j_{\mathbf{v}_j}}^{\top}\cdot(\mathbf{v}_j-\mathbf{v}_{j0})\}\|_1, \tag{20}$$

where $\mathcal{W}(\mathbf{I}_j,\mathbf{v}_j)$ is the image warping by flow $\mathbf{v}_j$, and $\mathbf{I}_{j_{\mathbf{v}_j}}$ is the image derivative in the x and y direction, respectively. The results are shown in Fig. 4-(g), together with its sequential estimation version in Fig. 4-(e). Fig. 4-(g) shows very similar accuracy with the proposed algorithm shown in Fig. 4-(h), but it and its sequential version take much more computation time due to their high dimensional $(2\times J+1)$ solution space. Table 1 summarizes the PSNR, SSIM, and computation time for each algorithm, together with the results from other high-performance super-resolution algorithms [12] and [16] whose executables are available for public.

6.2. Results on real sequence

Different from the synthesized data, our real data have camera pose errors because it is estimated from the real image sequence. Therefore, the effect of camera pose error in our algorithm can be analyzed using a real data set. A wide FOV camera is used for the effective 3D reconstruction, and the radial distortion is removed in advance. Fig. 5 shows the reconstructed depth map and super-resolution images, and Fig. 6 shows the comparison of various super-resolution algorithms previously discussed in the simulated data experiments. The results indicate that the camera pose error is not an important error factor for super-resolution.

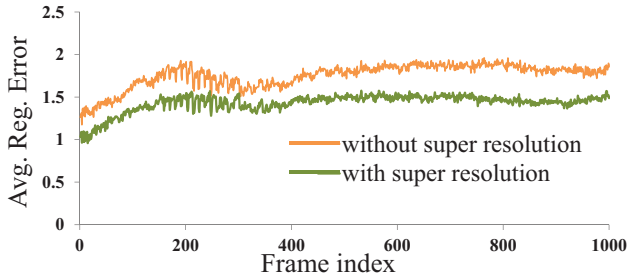


Figure 7. Plot of registration error for camera localization with high-resolution and low-resolution image and depth map for *outdoor* sequence.

6.3. Camera localization performance

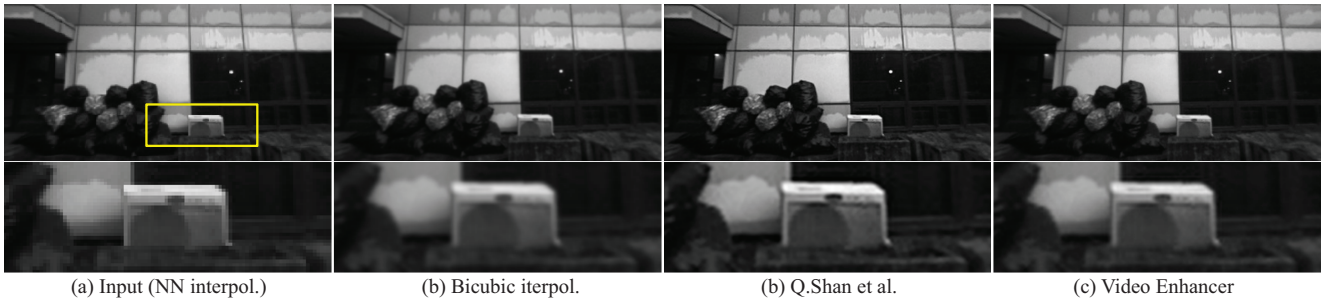
We test an improvement of the camera localization performance, measured by registration error from Eq. (19) through the image sequence. For a fair comparison, the original input images are used in the registration error calculation, because super-resolution images can reduce the photometric errors by themselves. Thus, only the depth map and the camera pose can affect the registration error, and the system which has a consistent depth map and camera trajectory through the whole sequence will have a lower average registration error. The plot of registration error for *indoor* sequence is shown in Fig. 7. The average per-pixel registration error (with intensity interval [0, 255]) with the high-resolution estimation is 1.430, whereas it is 1.752 for the camera localization with low-resolution images and depth map.

7. Conclusions

A novel optimization framework for simultaneous super-resolution and depth map estimation is proposed. Two closely related problems are formulated by a single convex problem using the camera geometry and solved effi-

(a) LR image (Bicubic) (b) SR image (c) Depth from LR image (d) Depth with SR

Figure 5. Depth map estimation and super-resolution results on the real image sequences. (a) Input images. (b) Super resolution images. (c) Depth map without super-resolution. (d) Depth map with super-resolution.



(a) Input (NN interpol.) (b) Bicubic iterpol. (b) Q.Shan et al. (c) Video Enhancer

(e) Seq. Flow + SR (f) Seq. Depth + SR (g) Simult. Flow + SR (h) Simult. Depth + SR

Figure 6. Comparison of super-resolution results on the real image sequences.

ciently by the first-order primal-dual algorithm. Our simultaneous solution gives results comparable to other high-performance algorithms for each problem, but takes much less computation time. Thus, the proposed framework can be applied to real-time 3D reconstruction systems for improving their accuracy. In our future work, more sophisticated super-resolution models, including occlusion, and geometry-aware downsampling, and their optimization will be discussed. We are also interested in investigating the use of depth sensors to facilitate better depth solution.

References

[1] A. V. Bhavsar and A. Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010. 2

[2] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 2011. 2, 4

[3] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *Journal on Imaging Sciences*, 3, 2010. 2

[4] R. Fransens, C. Strecha, and L. V. Gool. Optical flow based super-resolution: A probabilistic approach. *Computer Vision and Image Understanding*, 106, 2007. 1

[5] B. Goldlucke and D. Cremers. A super-resolution framework for high-accuracy multiview reconstruction. In *Proc. DAGM conference on Pattern recognition*, 2009. 2

[6] J. Lellmann, D. Breitenreicher, and C. Schnörr. Fast and exact primal-dual iterations for variational problems in computer vision. In *Proc. ECCV*, 2010. 2

[7] C. Liu and D. Sun. A bayesian approach to adaptive video super resolution. In *Proc. CVPR*, 2011. 1

[8] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid. Rslam: A system for large-scale mapping in constant-time using stereo. *International Journal of Computer Vision*, 94, 2010. 5

[9] U. Mudenagudi, A. Gupta, L. Goel, A. Kushal, P. Kalra, and S. Banerjee. Super resolution of images of 3d scenecs. In *Proc. ACCV*, 2007. 2

[10] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *Proc. ICCV*, 2011. 2, 5

[11] D. Scharstein and R. Szelisk. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47, 2002. 6

[12] Q. Shan, Z. Li, J. Jia, and C.-K. Tang. Fast image/video upsampling. *ACM Transactions on Graphics (Siggraph Asia)*, 27(5), 2008. 7

[13] J. Stühmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Proc. DAGM conference on Pattern recognition*, 2010. 2, 5

[14] T. Tung, S. Nobuhara, and T. Matsuyama. Simultaneous super-resolution and 3d video using graph-cuts. In *Proc. CVPR*, 2008. 2

[15] M. Unger, T. Pock, M. Werlberger, and H. Bischof. A convex approach for variational super-resolution. In *Proc. DAGM conference on Pattern recognition*, 2010. 2

[16] Video Enhancer. http://www.infognition.com/videoenhancer/, 2012. Version 1.9.7. 7