Simultaneously Estimation of Super-Resolution Images and Depth Maps from Low Resolution Sensors

Daniel B. Mesquita, Erickson R. Nascimento, Mario F. M. Campos
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
E-mail: {balbino,erickson,mario}@dcc.ufmg.br

***Abstract*—The emergence of low cost sensors capable of providing texture and depth information of a scene is enabling the deployment of several applications such as gesture and object recognition and three-dimensional reconstruction of environments. However, commercially available sensors output low resolution data, which may not be suitable when more detailed information is necessary. With the purpose of increasing data resolution, at the same time reducing noise and filling the holes in the depth maps, in this work we propose a method that combines depth fusion and image reconstruction in a super-resolution framework. By joining low-resolution intensity images and depth maps in an optimization process, our methodology creates new images and depth maps of higher resolution and, at the same time, minimizes issues related with the absence of information (holes) in the depth map. Our experiments show that the proposed approach has increased the resolution of the images and depth maps without significant spawning of artifacts. Considering three different evaluation metrics, our methodology outperformed other three techniques commonly used to increase the resolution of combined images and depth maps acquired with low resolution, commercially available sensors.**

***Keywords*-Super-resolution, convex optimization, RGB-D data, 3D reconstruction, computer vision.**

I. Introduction

The increasing availability of low-cost sensors capable of capturing both image and depth information, also known as RGB-D sensors such as Kinect, opened a wide research horizon for several methodologies involving processing of geometry and visual data. If on one hand these sensors can provide a good initial estimate of the three-dimensional structure of the scene with texture embedded, on the other hand such sensors provide low resolution images and noisy depth maps. Since the results of several algorithms are directly dependent on the quality and amount of detail present in the data, sharpening images and depth maps can increase the accuracies for object recognition and improve precision of 3D alignment and reconstruction [1].

In general, a super-resolution method can be divided into two main steps: registration and interpolation. The first step consists of finding the transformation of the frames to the same coordinate system (*e.g.* the coordinate system defined by camera's pose in the acquisition of the first image in the sequence). The second step estimates the intensity information of pixels for a larger grid.

When the sensor moves freely in the scene, matching between pixels is usually performed by optical flow. However, this approach has high computational cost and cannot be applied when sensor displacements are large. Therefore, several techniques have been proposed to use depth information in order to estimate pixel correspondences making the estimation more efficient. Despite the quality provided by these techniques, it is necessary to provide the depth information and in most cases, image reconstruction methods are used to estimate the depth. Although these strategies are able to determine the correspondence, they use a depth map which is a rough approximation of the real map. This limits the image quality estimation, since such depth maps contain errors that are propagated to the interpolation step.

The main contribution of this work is a new method for 3D reconstruction based on an optimization approach which provides super-resolution intensity and depth images acquired by a low resolution sensor, such as commercially available RGB-D sensors. Our proposed model is defined as a convex optimization problem for which state of the art optimization techniques are employed to find the best solution.

This paper is structured as follows: Section II presents an overview of the works in the area of super-resolution. Initially we discuss the main works in the area of super-resolution for intensity images, followed by a discussion on works that use depth maps. Section III discusses the proposed approach: An optimization method that uses three-dimensional reconstruction techniques for the fusion of images and depth maps to generate high-resolution image results. Results are presented and discussed in Section IV, starting with the validation of our method using synthetic images followed by a quantitative analysis and experiments with RGB-D data. Finally, Section V sums up the conclusions attained with the development and implementation of our methodology and indicates directions for future research.

II. Related Works

In general, the methodologies for super-resolution may be broadly divided into two main categories: Super-resolution by example and reconstruction based super-resolution. While methodologies in the former category use learning-based approaches in order to increase the resolution of images, in the latter it is assumed as a basic premise that the low resolution images (LR) are subsamples with subpixel precision of some high resolution image (HR), and which can be used to reconstruct the original HR image.

Thanks to the increasing access to devices capable of obtaining geometrical data from the scene, several super-resolution techniques have included the use of depth maps in their approach. In [2] the authors proposed a method based on MRF (Markov Random Fields) to integrate the depth information acquired by a laser scanner with images from a high resolution camera. A similar technique is presented in [3]. By using a ToF (Time-of-flight) camera to acquire range images and applying a bilateral filter, they present an iterative algorithm to enlarge the spatial resolution of depth maps.

In [4], the problem of super-resolution is modeled by using a calibrated three-dimensional scene and a probabilistic framework based on MAP-MRF. Although their method efficiently handles occlusions in the depth maps, the algorithm does not estimate depth, which reduces the quality of their results in regions which contain holes in the depth map. In [5], the authors present a method for increasing the precision of the reconstruction from 3D video using multiple static cameras and a formulation based on MRF and graph-cuts.

In [6] the authors used an energy functional that simultaneously estimates the texture and the 3D object surface. The limitation of that work is that the displacement of the camera and the correspondence between images must be known *a priori*.

Unlike [6], which applied an energy functional to simultaneously estimate the texture and geometry in higher resolution, [7] uses a formulation composed of an iterative technique based on graph-cuts and an Expectation Maximization (EM) procedure. The main drawback of their methodologies is the high computational cost required to create the HR images and depth maps.

Several approaches have modeled the reconstruction problem in super-resolution as a convex optimization problem, and the most popular are based on variational methods. In general, these methods use a primal-dual algorithm to solve the optimization problem. In [8], a first order primal-dual algorithm is applied to estimate super-resolution images. Although, the authors show the robustness of their method for different types of noise and different scale factors, the results degenerate in the presence of Gaussian noise.

A recent work related to ours is presented in [9]. Similar to our technique, the authors propose a methodology which uses a first-order primal-dual algorithm to perform three-dimensional reconstruction and obtains super-resolution data simultaneously, using as input a set of images from different views. In spite of the good results achieved for images, the reconstructed depth map does not take into account the depth information, and it is highly sensitive to image noise typically present in low cost RGB-D sensors. In order to overcome such limitations, our approach solves the optimization problem considering both visual and geometrical data.

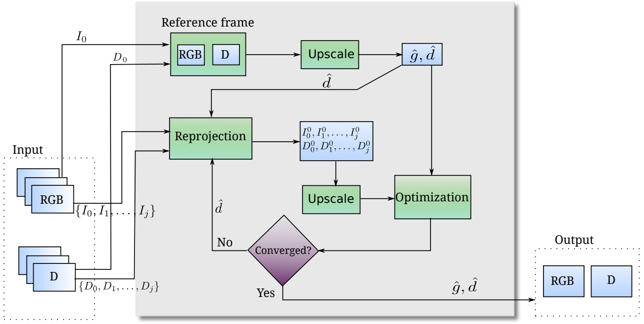


Figure 1. The main steps of the super-resolution methodology. First, all LR RGB-D images are reprojected to the reference frame and then an estimation of the super-resolution image $\boldsymbol{g}$ and depth $\boldsymbol{d}$ with the best reconstruction of the data is obtained in an iterative process.

III. Methodology

In this section we detail the design of our methodology. The input is a set of LR images $\boldsymbol{I}_j$ and depth maps $\boldsymbol{D}_j$ with size $M \times N$ and their corresponding camera poses $\boldsymbol{P}_j \in \mathbb{SE}(3)$ in witch $0 \leq j \leq J$. The output is an image $\boldsymbol{g}$ and a depth map $\boldsymbol{d}$ both with $sM \times sN$ size, $s \in \mathbb{R}$ is the scale factor. Both are estimated w.r.t. the pose of the reference image $\boldsymbol{P}_0$.

Our method is composed of three main steps: i) First, an initial estimation for $\boldsymbol{g}$ and $\boldsymbol{d}$ is computed by upscaling $\boldsymbol{D}_0$ and $\boldsymbol{I}_0$; ii) Then the input images and depth maps are reprojected onto the reference frame by using the depth information from $\boldsymbol{d}$; iii) Finally, the optimization process computes a new estimation for $\boldsymbol{g}$ and $\boldsymbol{d}$. These three steps are repeated until a convergence criteria is satisfied. The diagram in Figure 1 depicts these steps and the data flow from the LR images and depth maps to the final HR images and depth maps.

A. Modeling

Our methodology computes the super-resolution images and depth maps by modeling the problem as a convex optimization problem. The model adopted is based on three main parts:

- Image super-resolution – Establishes the relationship between a super-resolution image $\boldsymbol{g}$ and multiple input LR images $\boldsymbol{I}_i, 0 \leq j \leq J$.
- Depth super-resolution – Defines the relationship between a depth map in super-resolution $\boldsymbol{d}$ and multiple input depths $\boldsymbol{D}_i, 0 \leq j \leq J$.
- Regularization – Maintains consistency of the final solution avoiding degenerate results.

B. Super-Resolution from Images

The relationship between the reference frame and the super-resolution image is given by:

$$\boldsymbol{I}_0 = \boldsymbol{S} * \boldsymbol{B} * \boldsymbol{g}, \tag{1}$$

where $\boldsymbol{B}$ is a blurring Gaussian kernel with standard deviation s and a support size $(s-1)^{(1/2)}$ and $\boldsymbol{S}$ is the downsampling operator.

Assuming photometric consistency, the pixel with coordinates $\boldsymbol{x} = [x, y]^T$ in image $\boldsymbol{I}_0$ and depth $\boldsymbol{D}_0(\boldsymbol{x})$ has its corresponding coordinates in image $\boldsymbol{I}_j$ $(\forall j \in \{0, \ldots, J\})$ given by:

$$\boldsymbol{I}_0(\boldsymbol{x}) = \boldsymbol{I}_j(f(\boldsymbol{x}, \boldsymbol{D}_0(\boldsymbol{x}))), \tag{2}$$

where $f(\boldsymbol{x}, \boldsymbol{D}_0(\boldsymbol{x})) : R^2 \times R \to R^2$ is the function which maps the pixel $\boldsymbol{x}$ with depth $\boldsymbol{D}_0(\boldsymbol{x})$ from the reference image to $\boldsymbol{I}_j$. This mapping function will be presented in more detail in section III-E.

Thus, the problem of computing a super-resolution image consists of estimating an image $\boldsymbol{g}$ that minimizes the error between all low-resolution images:

$$\arg\min_{\boldsymbol{g}} \int_\Omega \sum_{j=0}^{J} \|(\boldsymbol{S} * \boldsymbol{B} * \boldsymbol{g})(\boldsymbol{x}) - \boldsymbol{I}_j(f(\boldsymbol{x}, \boldsymbol{D}_0(\boldsymbol{x}))\| d\boldsymbol{x}, \tag{3}$$

where Ω is the image domain.

C. Super-resolution of the Depth Maps

We also want to find a super-resolution depth map $\boldsymbol{d}$. Similarly to the image enlargement, we minimize the error between all the reprojected LR depth maps:

$$\begin{aligned} E_D &= \arg\min_d \sum_{j=0}^{J} \|\upsilon_j(\boldsymbol{d})\|_1 & (4) \\ &= \arg\min_d \sum_{j=0}^{J} \|\boldsymbol{S} * \boldsymbol{B} * \boldsymbol{d} - \boldsymbol{D}_j^0\|, & (5) \end{aligned}$$

where $\boldsymbol{D}_j^0$ is the depth map $\boldsymbol{D}_j$ reprojected to the reference frame.

D. Regularization

The regularization term is used to keep the consistency of the final solution, and in this work we use the Huber norm [10] for the intensity image $\boldsymbol{g}$ and depth $\boldsymbol{d}$. In addition to preserving discontinuities in the final solution, the Huber norm also prevents degenerate solutions.

The norm for $\boldsymbol{g}$ is defined by the following function:

$$\|\nabla g\|_{\alpha_g}(x, y) = \begin{cases} \frac{|\nabla g|^2}{2\alpha_g} & \text{if} |\nabla g| \leq \alpha_g, \\ |\nabla g| - \frac{\alpha_g}{2} & \text{if} |\nabla g| > \alpha_g, \end{cases} \tag{6}$$

where ∇ is the linear operator corresponding to the image gradient. The Huber norm $\|d\|_{\alpha_d}$ is calculated in the same way.

E. Reprojection Function

The reprojection function that maps a pixel $\boldsymbol{x} = [x, y]^T$ with depth $\boldsymbol{d}(\boldsymbol{x})$ in the reference frame to $\boldsymbol{I}_j$ is defined as:

$$f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x})) = h(\boldsymbol{K}\boldsymbol{P}_{j,0}\boldsymbol{d}(\boldsymbol{x})\boldsymbol{K}^{-1}[x, y, 1]^T), \tag{7}$$

where $\boldsymbol{K}$ is the projection matrix and h is the dehomogenization function.

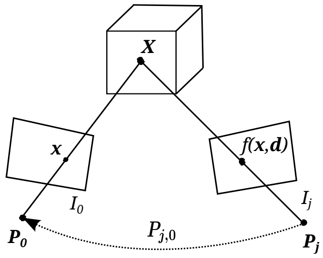


Figure 2. The reprojection fuction $f(\boldsymbol{x}, \boldsymbol{d})$ establishes the correspondence between the pixels in $\boldsymbol{I}_0$ and $\boldsymbol{I}_j$ based on the depth information at the reference frame.

By using the reprojection fuction, the image $\boldsymbol{I}_j(f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x})))$ which has the corresponding pixels of the reference image $\boldsymbol{I}_0(\boldsymbol{x})$ is given by:

$$\boldsymbol{I}_j(f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x}))) = \boldsymbol{I}_j(h(\boldsymbol{K}\boldsymbol{P}_{j,0}\boldsymbol{d}(\boldsymbol{x})\boldsymbol{K}^{-1}[x, y, 1]^T)). \tag{8}$$

To simplify the notation without using pixel reference, we define the reprojection operator $W(\boldsymbol{I}_j, \boldsymbol{d})$ which warps the image $\boldsymbol{I}_j$ to the reference frame in the form:

$$W(\boldsymbol{I}_j, \boldsymbol{d}) = \boldsymbol{I}_j(f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x})))d\boldsymbol{x}, \qquad \forall x \in \Omega. \tag{9}$$

F. Image super-resolution and reconstruction

Although the RGB-D sensor provides an initial estimation of the depth map $\boldsymbol{D}_0$ of the scene, such data is noisy and often there is no depth information in some areas (due to sensor limitations) producing holes in the depth map. To overcome this issue we use a similar approach to the works of [11] and [9] in which we estimate $\boldsymbol{d}$ iteratively and simultaneously to $\boldsymbol{g}$.

Considering the first-order Taylor expansion of $W(\boldsymbol{I}_j, \boldsymbol{d})$ we approximate a change in the image $W(\boldsymbol{I}_j, \boldsymbol{d})$ w.r.t. a small change of depth at an initial value $\boldsymbol{d_0}$ as:

$$W(\boldsymbol{I}_j, \boldsymbol{d}) \simeq W(\boldsymbol{I}_j, \boldsymbol{d}_0) + \frac{\delta}{\delta d} W(\boldsymbol{I}_j, \boldsymbol{d})\bigg|_{\boldsymbol{d}=\boldsymbol{d_0}} \cdot (\boldsymbol{d} - \boldsymbol{d_0}). \tag{10}$$

Therefore, the objective function described by Equation 3 can be linearized as:

$$\arg\min_{g,d} \sum_{j=0}^{J} \|\boldsymbol{S} * \boldsymbol{B} * \boldsymbol{g} - \{W(\boldsymbol{I}_j, \boldsymbol{d}_0) + \boldsymbol{u_j} \cdot (\boldsymbol{d} - \boldsymbol{d_0})\}\|_1, \tag{11}$$

where $\boldsymbol{u_j}$ is the simplified notation for $\frac{\delta}{\delta d} W(\boldsymbol{I}_j, \boldsymbol{d})\big|_{d_0}$, which can be calculated by applying the chain rule:

$$\boldsymbol{u_j} = \frac{\delta}{\delta d} W(\boldsymbol{I}_j, \boldsymbol{d}) = \nabla I_j(f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x}))) \cdot \frac{\delta f(\boldsymbol{x}, \boldsymbol{d}(\boldsymbol{x}))}{\delta d}. \tag{12}$$

In our solution, instead of applying the downsampling operator $\boldsymbol{S}$ to the image $\boldsymbol{g}$ at each iteration, we upscale the input images to the size $sM \times sN$ using bicubic interpolation.

Therefore, the function corresponding to the image term is given by:

$$E_I = \arg\min_{g,d} \sum_{j=0}^{J} \|\rho_j(\boldsymbol{g}, \boldsymbol{d})\| \tag{13}$$

$$= \arg\min_{g,d} \sum_{j=0}^{J} \|\boldsymbol{B} * \boldsymbol{g} - \{W(\hat{\boldsymbol{I}}_j, \boldsymbol{d_0}) + \hat{\boldsymbol{u_d}}.(\boldsymbol{d} - \boldsymbol{d_0})\}\|_1, \tag{14}$$

the operator $\hat{\cdot}$ represents the upscaled version of the image.

G. Final Cost Function

The energy function used in this work takes into account the cost functions E_I and E_D that correspond to merging multiple images and multiple depth maps respectively and the regularization term E_R.

The parameters λ_I and λ_D control the degree of regularization of the energy functional $E(\boldsymbol{g}, \boldsymbol{d})$ as follows:

$$E(\boldsymbol{g}, \boldsymbol{d}) = \|\nabla \boldsymbol{g}\|_{\alpha_g} + \|\nabla \boldsymbol{d}\|_{\alpha_d} + \lambda_I \sum_{j=0}^{J} \|\rho_j(\boldsymbol{g}, \boldsymbol{d})\|_1 + \lambda_D \sum_{j=0}^{J} \|\upsilon_j(\boldsymbol{d})\|_1. \tag{15}$$

H. Solution by Primal-Dual method

In this section we present a solution for Equation 15 based on the first order primal dual algorithm of [12]. Let Equation 15 be a min-max saddle point problem with a primal-dual formulation using Fenchel duality. For simplicity, only one adjacent frame will be used, then Equation 15 can be rewritten as:

$$E(\boldsymbol{g}, \boldsymbol{d}) = \|\nabla \boldsymbol{g}\|_{\alpha_g} + \|\nabla \boldsymbol{d}\|_{\alpha_d} + \lambda_I \|\rho_j(\boldsymbol{g}, \boldsymbol{d})\|_1 + \lambda_D \|\upsilon(\boldsymbol{d})\|_1. \tag{16}$$

Considering the dual variables $\mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{s}$ for each term of Equation 16, the primal dual formulation can be defined by:

$$\begin{aligned}\min_{\boldsymbol{g},\boldsymbol{d}} \max_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},\boldsymbol{s}} & \langle \nabla \boldsymbol{g}, \boldsymbol{p}\rangle + \langle \nabla \boldsymbol{d}, \boldsymbol{q}\rangle + \lambda_I \langle \rho_j(\boldsymbol{g}, \boldsymbol{d}), \boldsymbol{r}\rangle \\ & + \lambda_D \ \langle \upsilon_j(\boldsymbol{d}), \boldsymbol{s}\rangle + \frac{\alpha_g}{2}\|\boldsymbol{p}\|_2^2 + \frac{\alpha_d}{2}\|\boldsymbol{q}\|_2^2 \\ & + F^*(\boldsymbol{p}) + F^*(\boldsymbol{q}) + F^*(\boldsymbol{r}) + F^*(\boldsymbol{s}),\end{aligned} \tag{17}$$

where $F^*(\boldsymbol{p})$ is the dual function of the L^1 norm and is expressed as:

$$F^*(\boldsymbol{p}) = \begin{cases} 0 & \text{if } \|\boldsymbol{p}\| \leq 1 \\ \infty & \text{otherwise.} \end{cases} \tag{18}$$

This formulation expresses the same optimization problem, but with a function which is differentiable at all points.

The method used to solve the optimization problem defined by Equation 17 is an iterative algorithm, which has the steps determined by:

$$\begin{aligned}
\boldsymbol{p}^{n+1} &= \prod\nolimits_{\|\boldsymbol{p}\|_\infty \leq 1} \left\{ \frac{\boldsymbol{p}^n + \sigma \nabla \bar{\boldsymbol{g}}^n}{1 + \sigma \alpha_g} \right\}, \\
\boldsymbol{q}^{n+1} &= \prod\nolimits_{\|\boldsymbol{q}\|_\infty \leq 1} \left\{ \frac{\boldsymbol{q}^n + \sigma \nabla \bar{\boldsymbol{d}}^n}{1 + \sigma \alpha_d} \right\}, \\
\boldsymbol{r}^{n+1} &= \prod\nolimits_{\|\boldsymbol{r}\|_\infty \leq 1} \{\boldsymbol{r}^n + \sigma \rho(\bar{\boldsymbol{g}}^n, \bar{\boldsymbol{d}}^n)\}, \\
\boldsymbol{s}^{n+1} &= \prod\nolimits_{\|\boldsymbol{s}\|_\infty \leq 1} \{\boldsymbol{s}^n + \sigma \upsilon(\bar{\boldsymbol{d}}^n)\}, \\
\boldsymbol{g}^{n+1} &= \boldsymbol{g}^n + \tau \left(div \boldsymbol{p}^{n+1} - \lambda_I \boldsymbol{B}^T \boldsymbol{r}^{n+1} \right), \\
\boldsymbol{d}^{n+1} &= \boldsymbol{d}^n + \tau \left(div \boldsymbol{q}^{n+1} - \lambda_I \hat{\boldsymbol{u}}_{\boldsymbol{d}} \boldsymbol{r}^{n+1} - \lambda_D \boldsymbol{B}^T \boldsymbol{s}^{n+1} \right), \\
\bar{\boldsymbol{g}}^{n+1} &= 2\boldsymbol{g}^{n+1} - \boldsymbol{g}^n, \\
\bar{\boldsymbol{d}}^{n+1} &= 2\boldsymbol{d}^{n+1} - \boldsymbol{d}^n,
\end{aligned} \tag{19}$$

where div is the divergence operator and $\prod_{\|\boldsymbol{p}\|_\infty \leq 1}(\cdot)$ refers to the proximal operator of the dual variables and are subject to the restrictions described in Equation 17 and consists of a simple projection on the unit circle by the following formula:

$$p = \prod_{\|p\|_\infty \leq 1} (\bar{p}) \Leftrightarrow p_{ij} = \frac{\bar{p_{ij}}}{\max\{1, |\bar{p_{ij}}|\}}. \tag{20}$$

The extension for multiple images involves creating dual variables $\boldsymbol{r_i}$ and $\boldsymbol{s_i}$ for each image and update the solution in the form:

$$\boldsymbol{r_j}^{n+1} = \prod\nolimits_{\|\boldsymbol{r_j}\|_\infty \leq 1} \{\boldsymbol{r_j}^n + \sigma \rho_j(\bar{\boldsymbol{g}}^n, \bar{\boldsymbol{d}}^n)\}, \tag{21}$$

$$\boldsymbol{s_j}^{n+1} = \prod\nolimits_{\|\boldsymbol{s_j}\|_\infty \leq 1} \{\boldsymbol{s_j}^n + \sigma \upsilon_j(\bar{\boldsymbol{d}}^n)\}, \tag{22}$$

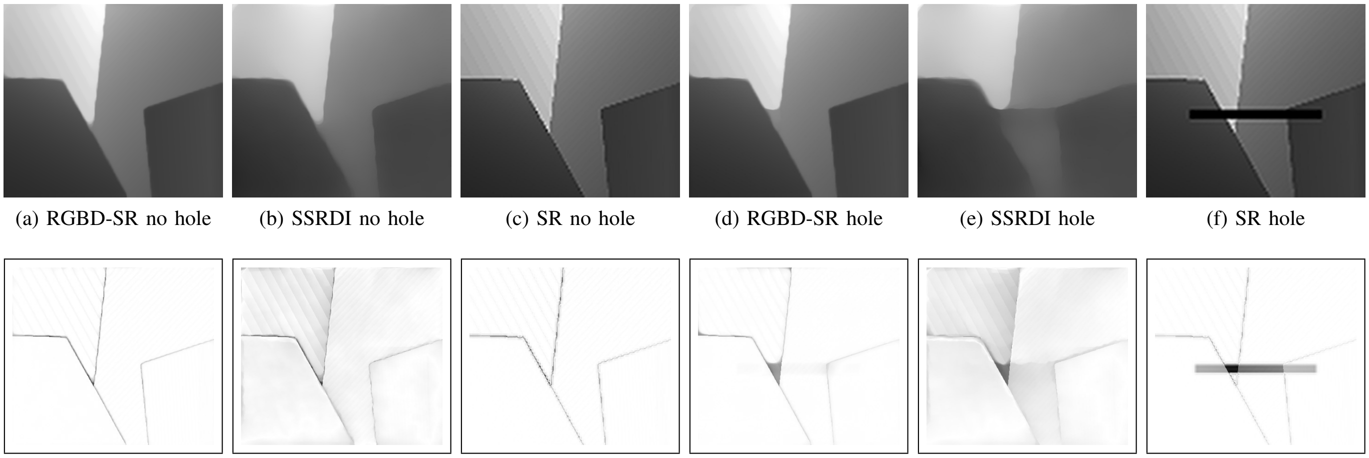
$$\boldsymbol{g}^{n+1} = \boldsymbol{g}^n + \tau div \boldsymbol{p}^{n+1} - \tau \lambda_I \boldsymbol{B}^T \sum_{j=0}^{J} \boldsymbol{r_j}^{n+1}, \tag{23}$$

$$\boldsymbol{d}^{n+1} = \boldsymbol{d}^n + \tau div \boldsymbol{q}^{n+1} - \tau \lambda_I \sum_{j=0}^{J} \hat{\boldsymbol{u_j}} \boldsymbol{r_j}^{n+1} - \tau \lambda_D \boldsymbol{B}^T \sum_{j=0}^{J} \boldsymbol{s_j}^{n+1}. \tag{24}$$

The timesteps σ and τ control the rate of convergence and we choose values according to [12]. Since the reconstruction term in Equations 21 and 24 define a solution space which may not be uniformly convex on all dimensions, we applied a pre-preconditioning process as discussed in [13].

I. Multi-scale Super-resolution approach

The linearization described in Equation 11 is valid only for small displacements. Moreover, the problem of image reconstruction is non-linear and may converge to local minimum. In order tackle with this issues we use a multi-scale coarse-to-fine approach. By using a set of scales $\{s_0, s_1, \ldots, s\}$ in ascending order, we estimate the depth map $\boldsymbol{d}$ and use it as input depth $\boldsymbol{d_0}$ on the next scale in the sequence.



(a) RGBD-SR no hole (b) SSRDI no hole (c) SR no hole (d) RGBD-SR hole (e) SSRDI hole (f) SR hole

(g) RGBD-SR no hole (h) SSRDI no hole (i) SR no hole (j) RGBD-SR hole (k) SSRDI hole (l) SR hole

Figure 3. Resulting depth map with absolute error for the ”Venus” instance for the different methods, in the error map, darker areas indicate larger error. As one can see our method (RGBD-SR) presents the smallest error.

IV. Experiments

We performed several experiments to evaluate our methodology both quantitatively and qualitatively. Synthetic images and real images were used for qualitative analysis. We also compare our results against the works of [9] and [8], which will be denoted as SSRDI and SR respectively. To simplify the nomenclature, we named our method as RGBD-SR.

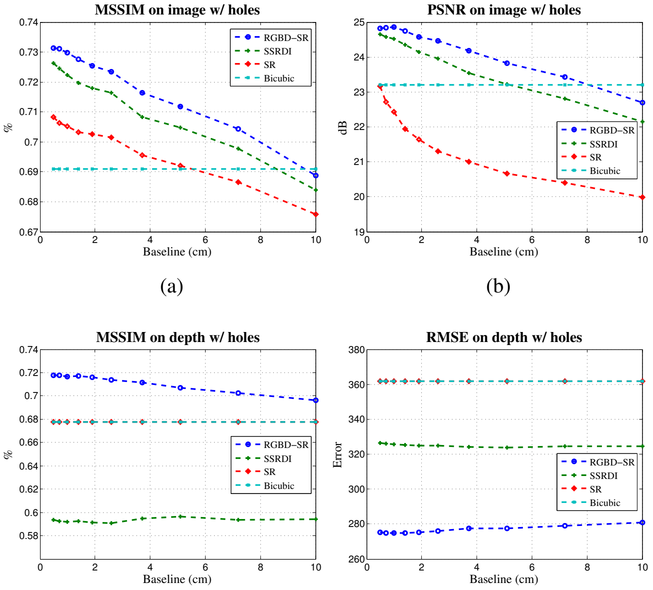
A. Synthetic Images

All synthetic images were created based on the ”Venus” sequence from the Middlebury dataset [14]. The ”Venus” sequence is composed of an intensity image and a depth map of size 434×383 pixels. With the purpose of simulating holes, which is very common in RGB-D low cost sensors (such as Kinect, for example), we included a hole with dimensions 260×20 in the center of the depth map by setting all values to zero (Figure 3f).

In the first experiment, we generated a set with 14 LR images by applying small displacements (rotation and translation transforms). We also included a synthetic hole in 7 images, and downscaled the images to 108×95 pixels. The images were used as input for the super-resolution algorithm in order to restore to its original size, which allow us to evaluate the result of the enlargement. For all tests the virtual camera positions are known and given as input in the optimization process.

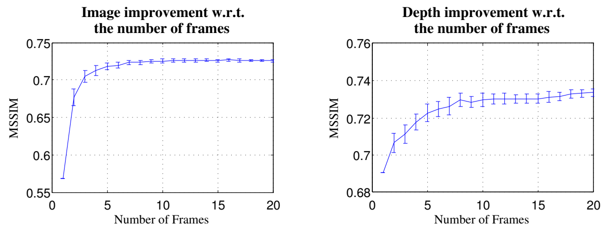
Figure 3(a-f) shows the results for the enlargement of the depth maps. One can clearly see in the error maps in Figures 3 (g-l) (darker pixels represent larger errors) that our methodology has the smallest errors even when there is lack of depth data like in Figure 3j. In addition to the ability of restoring the original size of the LR images, our methodology was also able to reconstruct the depth information without propagating the errors to the whole depth map. As expected a naive enlargement algorithm such as a bicubic interpolation was not capable of reconstructing the depth information.

For the quantitative analysis we generated 11 image sets, each one composed of 7 images and 7 depth maps. We included holes in all depth maps and a virtual camera was placed at arbitrary positions. We used different distances between the virtual camera and the reference frame for each set of images. By increasing this distance we are able to evaluate the robustness of the algorithm for large displacements. We evaluate the resulting HR images and depth maps by comparing then to the original. As quality assessment metric we use MSSIM (Mean Structural Similarity Index) and PSNR (Peak signal-to-noise ratio) for the images and MSSIM and RMSE (Root Mean Squared Error) for the depth maps.



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

Figure 4. Analysis of the robustness of the methodologies for different baselines (distance of the virtual camera w.r.t. to the reference frame).



(a) (b)

Figure 5. Quality improvement w.r.t. the number of input frames. The same experiments were repeated 20 times, allowing to generate a graph of the mean and standard deviation.

Figure 4 shows the results for each set of different distances. One can see that our method presents better results both for the MSSIM and PSNR (considering the results in the images). The blue curve remains consistently above the other curves independent of increasing the distance. Figures 4c and 4d show the results for depth maps. Our method outperforms the others for both metrics (for RMSE metric, the best values are the smallest one).

B. Number of Images

We also evaluate the relation between the number of frames and the errors. As the number of frames used for the super-resolution algorithm increases, one expects to get even better results, however, there is a practical limit for improvements caused by external factors. To verify such limitations, we used the synthetic dataset presented in the section IV-A, where we gradually increased the number of low resolution images and then compare the resulting super-resolution image against the original image using the MSSIM metric. The same experiment was repeated 20 times. The results for intensity and depth images are shown in Figures 5a and 5b respectively. In both cases we found that above 8 images there were no significant improvements in the quality of the results.

C. Noise Robustness

We also analyze the robustness of our method in the presence of noise. We used 7 low resolution synthetic images from the dataset described in Section IV-A. We applied Gaussian noise with increasing σ on the input images, then we compared the estimated super-resolution intensity image against the original high resolution intensity image with the MSSIM metric. The results are shown in the Table IV-C. One can see that, while the Bicubic interpolation quality drops steadily with increased noise, all the super-resolution methods drop slower and at the same rate, in which our method consistently keeps the highest score.

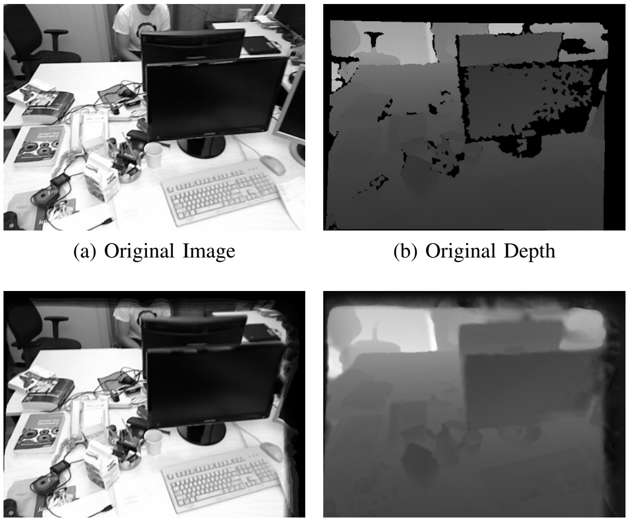
D. Results on a real dataset

For experiments with real images, we used the XYZ1 sequence of the Freiburg dataset [15]. This sequence is composed of frames with small displacements acquired with a low cost RGB-D sensor with its respective pose information. Figures 6a and 6b show the low-resolution image and the depth map used as reference frames. In order to improve the registration already given by the dataset, we also used the method proposed by [16] for all real image experiments.

The super-resolution result obtained after applying our method is shown in Figures 6c and 6d. In this case we used 5 consecutive frames as input. Figure 7a and 7b show the low-resolution and super-resolution point clouds.

	MSSIM			
σ	RGBD-SR	SSRD	SR	Bicubic
0	**0.727**	0.721	0.697	0.690
0.01	**0.709**	0.693	0.680	0.648
0.02	**0.689**	0.673	0.665	0.551
0.04	**0.648**	0.637	0.633	0.374
0.08	**0.511**	0.509	0.521	0.198

Table I
COMPARISON OF THE DIFFERENT METHODS WHEN APPLIED A GAUSSIAN NOISE ON THE INPUT IMAGES.



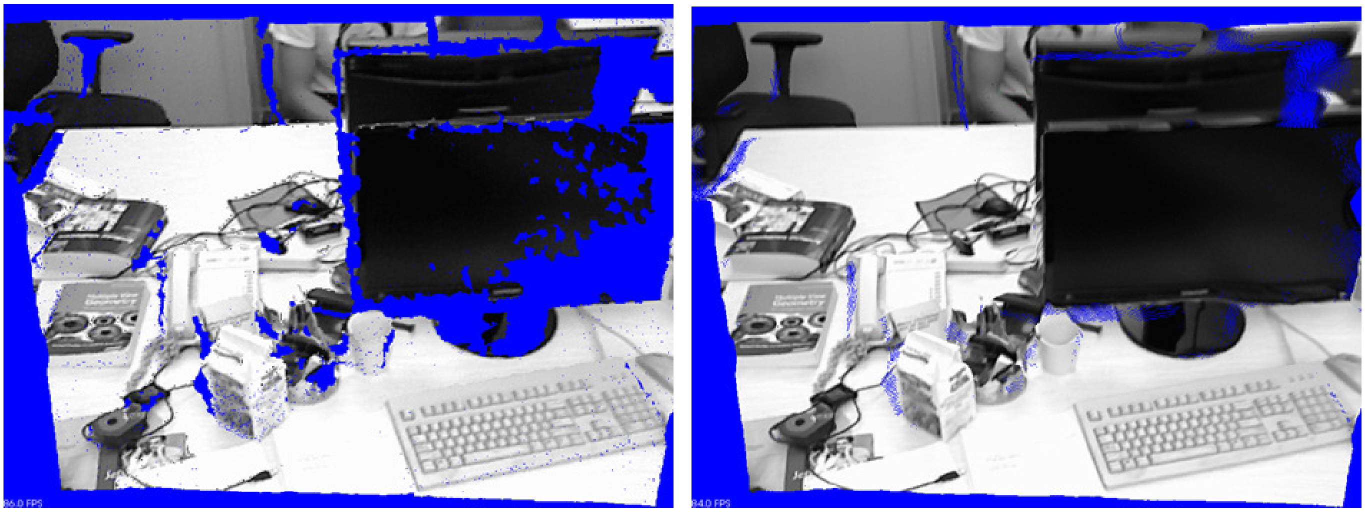
(a) Original Image (b) Original Depth

(c) Super-resolution Image (d) Super-resolution Depth

Figure 6. Experiments with XYZ1 sequence of the Freiburg dataset. We used the images and depth maps of 5 consecutive frames as input. Besides the increase in image and depth map resolution, we can see, by comparing image (b) and (d), that our methodology was able to reconstruct the holes in the depth map.

Aside from the experiments using the Freiburg data set, we also evaluated our approach on outputs from a Kinect sensor. We configured the sensor to capture intensity images with resolution of 640×480 and $1,280 \times 960$. The latter were used as groundtruth. After acquiring 10 images and its respective depth maps with the default 640×480 Kinect resolution (Figure 8), we ran the methodologies to compute such images in resolution of $1,280 \times 960$.

It can be seen in Figure 10 that the depth map produced by the SSRDI method falls into local minimum in some areas and the image reconstruction algorithm is greatly affected by the image noise in areas with low texture information. Since our method uses information from multiple depth maps,



(a) Low Resolution Point Cloud

(b) Super-Resolution Point Cloud

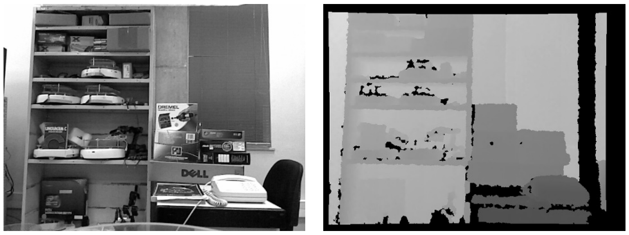
Figure 7. The left image shows a point cloud of the original image from the Freiburg dataset and the right image the corresponding point cloud estimated by our algorithm. It can be seen that our method increase the point cloud resolution (e.g. keyboard and milk box) as well as reconstruct the lost and deteriorated regions of the cloud (e.g. monitor).

the chances that the image reconstruction process converge correctly increases, in addition it is possible to obtain a more robust and accurate result.

The comparison performed by using different metrics against other methods is shown in the Table II. The proposed method in this work (RGBD-SR) leads in performance, followed by SSRDI method which presented a better result than bicubic interpolation. The resulting images for RGBD-SR and SSRDI can be seen in Figure 11. We selected some key regions indicated by red rectangles and compared with the same regions of the groudtruth image (the result of the bicubic interpolation is also shown). It can be readily seen that our method produces the best results in all cases and that the SSRDI method produced some distortions in these areas mainly because the reconstruction optimization process fell into a local minimum.

Metric	RGBD-SR	SSRDI	SR	Bicubic
MSSIM	0.933	0.913	0.887	0.905
PSNR	25.11	24.46	23.97	23.55

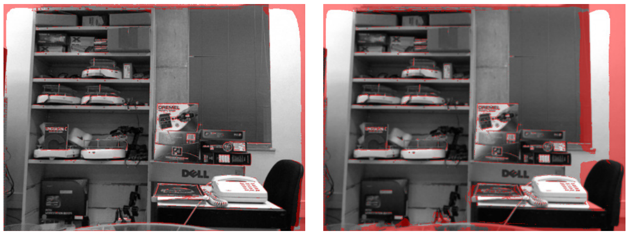
Table II
COMPARISON BETWEEN THE KINECT SVGA IMAGE AND THE HR ESTIMATED FROM A SET OF LR KINECT VGA IMAGES.



(a) Intensity Image

(b) Depth Map

Figure 8. Image and depth maps with resolution 640 × 480 acquired with a Kinect in the laboratory.



(a) RGBD-SR

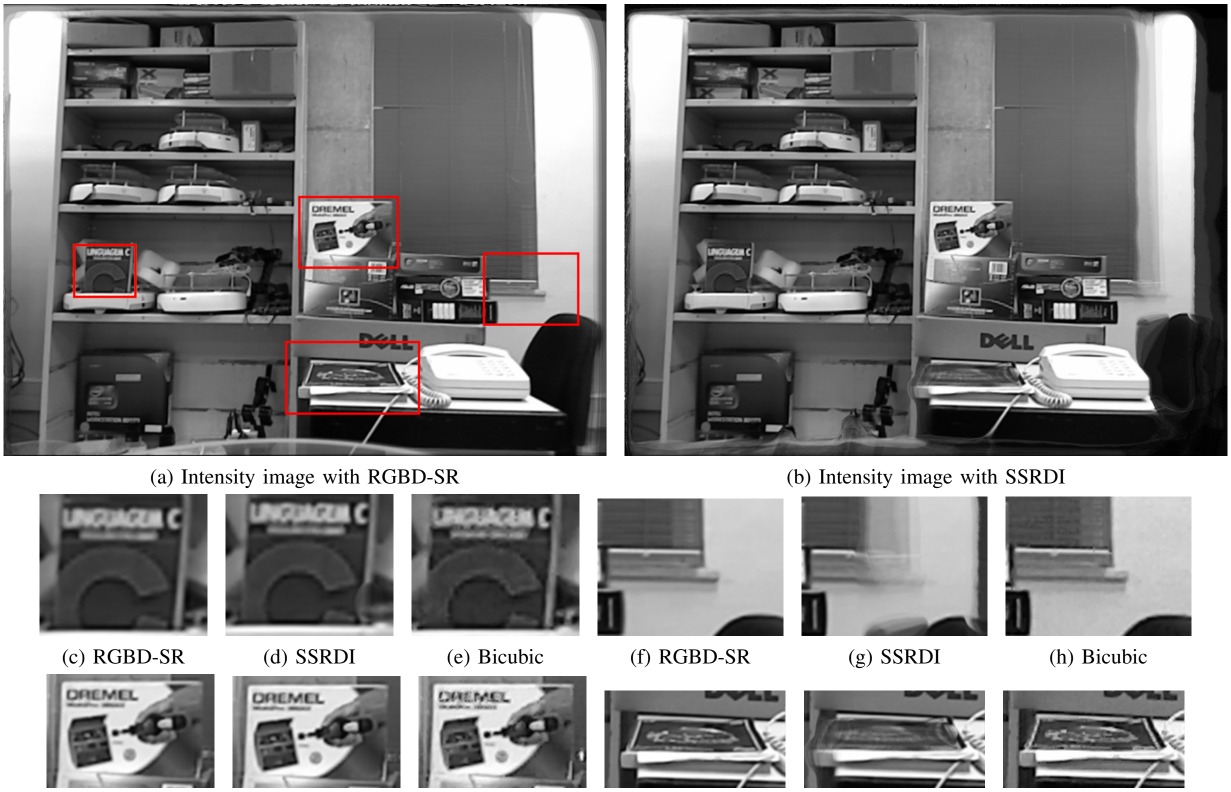
(b) SSRDI

Figure 9. The highlighted red areas show when the absolute difference between the original image and the super-resolution image is above a threshold of 10% from the maximum value. Thanks to the smoothing caused by the regularization, the error concentrates on border areas. Our method presents less areas with errors.

V. Conclusion and Future Work

In spite of the large number of works in super-resolution on imaging and depth maps, there are a few efforts which combine these two kinds of information to improve the final result. To fill this gap, in this work we propose a method capable of estimating simultaneously images and depth maps in higher resolution than that provided by the sensor. Since our method is based on a reconstruction approach, it is also able to estimate the depths not captured by the sensor.

Our experiments showed that, in several cases where there were failures, lost or deteriorated regions (e.g. holes) in the depth map, our methodology correctly estimated the depth information, and thanks to the fusion of image and depth map reconstruction, it was possible to improve the results for both

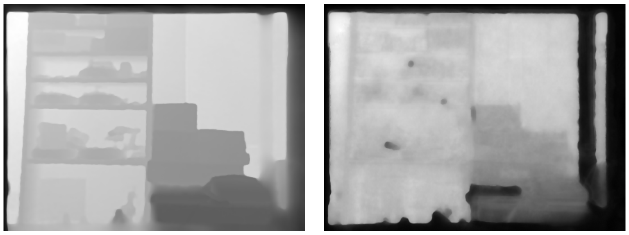


(a) Intensity image with RGBD-SR (b) Intensity image with SSRDI

(c) RGBD-SR (d) SSRDI (e) Bicubic (f) RGBD-SR (g) SSRDI (h) Bicubic

(i) RGBD-SR (j) SSRDI (k) Bicubic (l) RGBD-SR (m) SSRDI (n) Bicubic

Figure 11. Images produced with the our method (RGBD-SR) and the SSRDI method. We can see that both methods produce better visual appearance when compared to the bicubic interpolation. However, the SSRDI method falls into local minimum in some areas as shown in Figures 11g and 11m .



(a) Depth with RGBD-SR (b) Depth with SSRDI

Figure 10. Depth maps produced by the our methodology (RGBD-SR) and the SSRDI method.

images and the depth maps.

As future work we intend to adapt the reconstruction model to different luminance conditions and extend it to handle color information as discussed in [17].

References

[1] M. Meilland and A. I. Comport, "Super-resolution 3D tracking and mapping," *Proc. ICRA*, pp. 5717–5723, May 2013.

[2] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *Advances in neural information processing systems*, 2005, pp. 291–298.

[3] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. CVPR*. IEEE, 2007, pp. 1–8.

[4] U. Mudenagudi, A. Gupta, L. Goel, A. Kushal, P. Kalra, and S. Banerjee, "Super resolution of images of 3d scenecs," pp. 85–95, 2007.

[5] T. Tung, S. Nobuhara, and T. Matsuyama, "Simultaneous super-resolution and 3d video using graph-cuts," in *Proc. CVPR*. IEEE, 2008, pp. 1–8.

[6] B. Goldlücke and D. Cremers, "A superresolution framework for high-accuracy multiview reconstruction," in *Pattern Recognition*. Springer, 2009, pp. 342–351.

[7] A. V. Bhavsar and A. Rajagopalan, "Resolution enhancement in multi-image stereo," *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1721–1728, 2010.

[8] M. Unger, T. Pock, M. Werlberger, and H. Bischof, "A convex approach for variational super-resolution," in *Pattern Recognition*, 2010, pp. 313–322.

[9] H. S. Lee and K. M. Lee, "Simultaneous Super-Resolution of Depth and Images Using a Single Camera," *Proc. CVPR*, pp. 281–288, Jun. 2013.

[10] P. J. Huber, "Wiley series in probability and mathematics statistics," *Robust Statistics*, pp. 309–312, 1981.

[11] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition*, 2010, pp. 11–20.

[12] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[13] T. Pock and A. Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *Proc. ICCV*. IEEE, 2011, pp. 1762–1769.

[14] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[15] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A Benchmark for the Evaluation of RGB-D SLAM Systems," in *Proc. IROS*, Oct. 2012.

[16] F. Steinbrucker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense rgb-d images," in *Proc. ICCV*. IEEE, 2011, pp. 719–722.

[17] B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM Journal on Imaging Sciences*, vol. 5, no. 2, pp. 537–563, 2012.