Simultaneous super-resolution, tracking and mapping

Maxime Meilland, Andrew I. Comport

LABORATOIRE

INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 7271

SIMULTANEOUS SUPER-RESOLUTION, TRACKING AND MAPPING

Maxime Meilland, Andrew Comport

Pôle SIS

Rapport de recherche
ISRN I3S/RR–2012-05–FR

Septembre 2012

Laboratoire d'Informatique de Signaux et Systèmes de Sophia Antipolis - UNSA-CNRS
2000, rte.des Lucioles – Les Algorithmes – Bât Euclide B – B.P. 121 – 06903 Sophia-Antipolis Cedex – France
Tél.: 33 (0)4 92 94 27 01 – Fax: 33 (0)4 92 94 28 98 – www.i3s.unice.fr
UMR7271

RÉSUMÉ :

Cet article propose une technique nouvelle de SLAM visuel qui intègre non seulement le pose en 6ddl et la structure 3D de façon dense, mais il intègre simultanément les informations de couleur contenues dans les images au fil du temps. Il s'agit de développer un modèle inverse pour la création d'une carte de super-résolution à partir de plusieurs images à basse résolution. Contrairement aux techniques classiques de super-résolution, notre approche tient pleinement compte de la translation et rotation 3D dans une formalisme de localisation et cartographie dense. Cela permet non seulement de prendre en compte toute la gamme des déformations de l'image, mais permet également de proposer des critères nouveaux pour combiner les images à faible résolution ainsi que sur la base de la différence de résolution entre les images différentes dans l'espace 6D. Plusieurs résultats sont donnés montrant que cette technique fonctionne en temps réel (30 Hz) et est capable de cartographier les environnements à grande échelle en haute résolution tout en améliorant la précision et la robustesse du suivi.

MOTS CLÉS :

super-resolution, SLAM visuel, localisation, cartographie, suivi 3D

ABSTRACT:

This paper proposes a new visual SLAM technique that not only integrates 6DOF pose and dense structure but also simultaneously integrates the color information contained in the images over time. This involves developing an inverse model for creating a super-resolution map from many low resolution images. Contrary to classic super-resolution techniques, this is achieved here by taking into account full 3D translation and rotation within a dense localisation and mapping framework. This not only allows to take into account the full range of image deformations but also allows to propose a novel criteria for combining the low resolution images together based on the difference in resolution between different images in 6D space. Several results are given showing that this technique runs in real-time (30Hz) and is able to map large scale environments in high-resolution whilst simultaneously improving the accuracy and robustness of the tracking.

KEY WORDS :

super-resolution, visual SLAM, localisation, mapping, 3D tracking

Simultaneous super-resolution, tracking and mapping

Maxime Meilland and Andrew I. Comport I3S CNRS Laboratory
University of Nice Sophia Antipolis
surname@i3s.unice.fr

September 15th 2012

***Abstract*—This paper proposes a new visual SLAM technique that not only integrates 6DOF pose and dense structure but also simultaneously integrates the color information contained in the images over time. This involves developing an inverse model for creating a super-resolution map from many low resolution images. Contrary to classic super-resolution techniques, this is achieved here by taking into account full 3D translation and rotation within a dense localisation and mapping framework. This not only allows to take into account the full range of image deformations but also allows to propose a novel criteria for combining the low resolution images together based on the difference in resolution between different images in 6D space. Several results are given showing that this technique runs in real-time (30Hz) and is able to map large scale environments in high-resolution whilst simultaneously improving the accuracy and robustness of the tracking.**

I. Introduction

The problem of dense real-time localisation and mapping within complex environments is a challenging problem for a wide range of applications ranging from robotics to augmented reality. In this paper the aim is to be able to interact in real-time with the surfaces of the environment so dense approaches are necessary. This work is undertaken as part of a French DGA Rapid project named Fraudo which requires dense localisation and mapping in real-time so as to allow path planning for a mobile robot to traverse uneven ground and surfaces autonomously. Another objective is to allow remote observation of the complex scenes for the operator. The goal is therefore to develop an efficient, accurate and robust *dense visual models* for localisation and mapping. As in all SLAM problems, in order to estimate the unknown maps using a moving sensor, it is necessary to simultaneously estimate the pose of the sensor.

The objectives here require real-time computational efficiency so several bodies of literature are not considered in this short review but are noted to have some overlapping approaches. In particular, the large volume of literature associated with off-line techniques such as Structure From Motion (SFM) and video post-production techniques [1], [2], [3], [4] have similar problems but perform lengthy calculations using all the data simultaneously. 3D volumetric approaches from the computer graphics literature are also very relevant [5]. Equally, we focus on approaches which look at full 6D transformations including rotation and translation since we consider this to be essential. Even so there are many interesting works which have looked at dense approaches in 2D including optic flow [6] or piecewise dense models such as affine [7] or planar geometry [8]. Some real-time stereo algorithms have also been around for quite some time [9], however, only stereo matching is performed and full poses are not estimated.

In the past ten years a lot of work has been carried out to perform robust real-time 6D localisation and mapping. In particular we can note that most visual SLAM approaches have used *feature-based techniques* combined with depth and pose estimation [10], [11], [12], [13]. Unfortunately these approaches are still based on an error prone feature extraction step and are not suited to interact with surfaces since they only provide a sparse set of information and do not provide any information about the dense structure of the surface. Amongst the various RGB-D systems, feature based methods include [14], [15], [16]. All of these methods rely on an intermediary estimation processes based on detection thresholds. This feature extraction process is often badly conditioned, noisy and not robust therefore relying on higher level robust estimation techniques. Furthermore, it is necessary to match these features between images over time which is another source of error (feature mapping is not necessarily one-to-one).

More recently, dense techniques have started to become popular and several groups have demonstrated real-time performance with commodity hardware. In particular, an early work performing dense 6D SLAM in real-time over large distances [17] was based on minimising an intensity in image key-frames. Other photometric approaches include [18] which looks at fully dense omnidirectional spherical RGB-D sensors. Alternatively, other approaches have focused only on geometry [19], [20]. In the later truncated signed distance functions (TSDF) are used to define depth integration in a volumetric space and a classic Iterative Closest Point (ICP) is used to estimate the pose. Recent contributions have included using a moving TSDF with ICP [21]. Uniquely geometric approaches are also common to time-of-flight range sensors [22]. Unfortunately the techniques described here either limit themselves to photometric optimisation in the former case and in the later only geometric information is used. Neglecting one or the other means that important characteristics are overlooked in terms of robustness, efficiency and precision. It can be noted,

however, that in [23], a benchmark test was used to compare both approaches and it was shown that the photometric approach is more precise.

Few techniques have considered optimising an error on both intensity and depth images. In [24] a direct ICP technique was proposed which does this simultaneously using an image-based approach. Alternatively, in [25] both errors were minimised but using a volumetric approach based on Octomap. There are several arguments for and against each approach. In the image based case the resolution of the map is a function of the path taken to acquire it, whereas the volumetric approach is invariant to the path used. In that way the volumetric approach is unable to easily capture the non-linear variation of the image resolution which depends on a particular camera trajectory. More importantly, it should be noted that none of these techniques have tried to "integrate the photometric intensity information", i.e. only pose and depth parameters have been estimated.

To investigate models to integrate the image intensity function we turn to super-resolution (SR) approaches. In this field a great amount of research has been carried out in the past, however, this has mainly been focused on applications such as photography or surveillance so as to obtain better 2D images. More particularly, super-resolution is the art of reconstructing higher resolution images, from a set of lower resolution images. In the most general case, these images are captured from different viewpoints, under different lighting conditions and with sensors of varying resolutions. See Figure 1 (a) for an overview of the image degradation pipeline reconstruction pipeline. Since the paper of [26], super resolution has been extensively studied in the computer vision community, however, most of the research only considers small relative motion between the input images and the major contributions are focused on how to fuse the registered images [27], [28], [6]. Furthermore, the registration techniques are mainly 2D and do not take into account knowledge about the dense depth maps of the scene. Several tutorials of these approaches are available which give basic underlying models and principles [29], [30] and more recent approaches aim at extending them such as [31] who perform spatially adaptive block-based super-resolution.

In this paper we propose an approach to not only simultaneously estimate the 6D pose along with the dense depth map but also the photometric images in a super-resolution format. This is achieved by considering an inverse compositional approach which is efficient for real-time performance since it allows a maximum of pre-computations to be performed. This differs from the classic super-resolution pipeline as is shown in Figure 1. In the model proposed here, dense tracking is used to align the images in 6D while several low resolution images are combined together and integrated to form the high-resolution image (SR). The low resolution (LR) images are combined by minimising their distance to a "virtual image" which is translated and rotated in such a way that it has the same resolution as the high-resolution image. In this way low resolution images are considered better if they are closer to the the same resolution as the target images.

The remainder of the paper is set out as follows. In Section II an overview is first given for the super-resolution process. In Section III the dense SLAM algorithm is defined. In Section IV-B a simulator is used to obtain a ground truth and evaluate the approach. In Section IV-C real-time images are used to perform super-resolution.

II. Observation Model

Consider a RGB-D sensor with a color brightness function $\mathbf{I}(\mathbf{p},t)$ and a depth function $\mathbf{D}(\mathbf{p},t)$, where $\mathbf{p}=(u,v)$ are pixel locations within the image acquired at time t. It is convenient to consider the set of measurements in vector form such that $\mathbf{I}\in\mathbb{R}^{\mathrm{nm}}$ and $\mathbf{D}\in\mathbb{R}^{\mathrm{nm}}$. Consider now a RGB-D image, denoted also an *augmented image* [18], to be the set containing both brightness and depth $\boldsymbol{\mathcal{I}}=\{\mathbf{I},\mathbf{D}\}$. $\mathbf{v}=\{\mathbf{p},\mathbf{D}\}\in\mathbb{R}^{3\times\mathrm{n}}$ are then the 3D vertices of the surface associated with the image points $\mathbf{p}$ and the depth image. $\boldsymbol{\mathcal{I}}$ will be called the *current* image and $\boldsymbol{\mathcal{I}}^*$ the *reference* image. A superscript $*$ will be used throughout to designate the reference view variables.

Now consider a set of low resolution augmented images $\{\boldsymbol{\mathcal{I}}_1,\boldsymbol{\mathcal{I}}_2,\ldots,\boldsymbol{\mathcal{I}}_N\}$, which observe the same scene from different 3D poses, the super-resolution process consists in simultaneously registering and fusing the images onto an augmented super-resolved image $\boldsymbol{\mathcal{I}}^*_{sr}$ such that:

$$\begin{cases}\mathbf{I}^*_{sr}=f\left(\sum_i^N\mathbf{C}_i^{\mathbf{I}}\mathbf{I}_i\left(w\left(\overline{\mathbf{T}}_i,\mathbf{v}_i;\mathbf{K},\mathbf{S}\right)\right)+\eta,\mathbf{B}^{-1}\right)\\ \mathbf{D}^*_{sr}=f\left(\sum_i^N\mathbf{C}_i^{\mathbf{D}}\mathbf{D}_i\left(w\left(\overline{\mathbf{T}}_i,\mathbf{v}_i;\mathbf{K},\mathbf{S}\right)\right)+\eta,\mathbf{B}^{-1}\right)\end{cases}\tag{1}$$

where the matrices $\overline{\mathbf{T}}=(\bar{\mathbf{R}},\bar{\mathbf{t}})\in\mathbb{SE}(3)$ are the true poses of the RGB-D cameras relative to the reference position. Throughout, $\mathbf{R}\in\mathbb{SO}(3)$ is a rotation matrix and $\mathbf{t}\in\mathbb{R}(3)$ the translation vector. The matrix $\mathbf{S}\in\mathbb{R}^{3\times3}$ is the up-sampling matrix, $\mathbf{K}\in\mathbb{R}^{3\times3}$ is the intrisic matrix of the real camera, $\mathbf{C}\in\mathbb{R}nm\times nm$ is the combination matrix and $\mathbf{B}$ is the blur or inverse blur of a given radius. These variables and the warping function will be now explained in detail.

Note that in practice affine illumination parameters are also estimated as in [29]: $\mathbf{I}'=\alpha\mathbf{I}+\beta$ along with vignetting parameters but to improve the clarity of the equations we omit this part of the pipeline.

1) **Geometric warping:** Consider the Figure 1 which shows the processing pipeline. From the first processing block, the motion model $w(\overline{\mathbf{T}}_i,\mathbf{v}_i;\mathbf{K})$ is a 3D warping function, which is related to the 3D pose $\overline{\mathbf{T}}$ of the camera and to the scene vertices $\mathbf{v}$:

$$\mathbf{p}^w=\frac{\mathbf{K}(\overline{\mathbf{R}}\mathbf{v}+\bar{\mathbf{t}})}{\mathbf{e}_3^T\mathbf{K}(\overline{\mathbf{R}}\mathbf{v}+\bar{\mathbf{t}})},\tag{2}$$

where $\mathbf{e}_3$ is a unit vector with the third component equal to 1.

2) **Image up-sampling :** The next block in Figure 1 involves the up-sampling of the LR image to the SR image. Usually this is done by creating intensity values at sub-pixel increments, however, for ease of notation and programming, here we consider the SR image pixels to be smaller than the LR pixels by a scaling factor s. This consists in warping the reference low resolution image by a diagonal homography

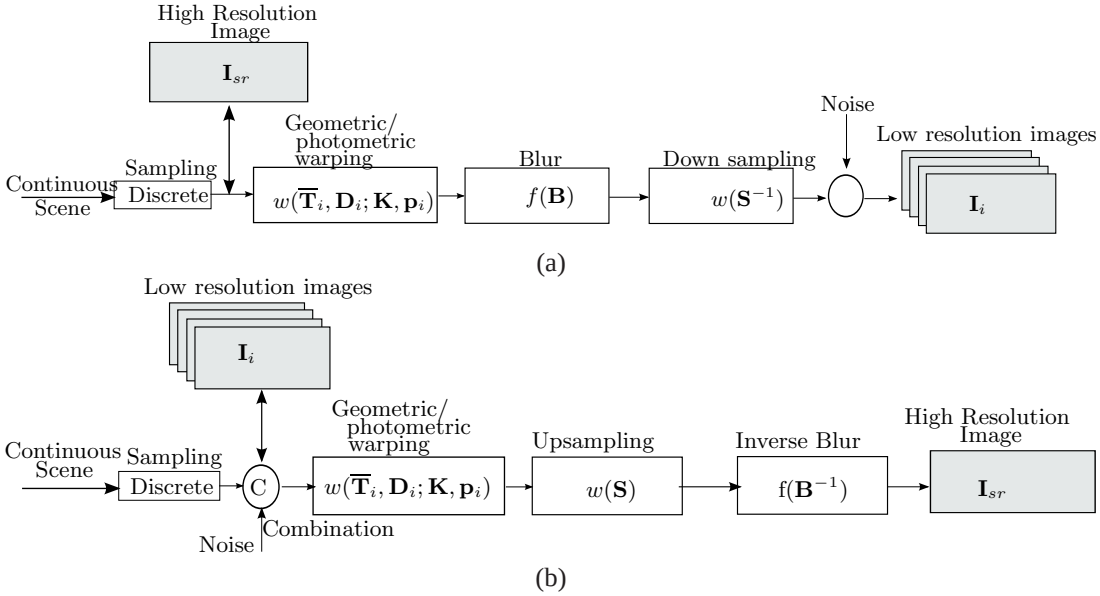


Fig. 1. (a) The image degradation pipeline (forward compositional). On the left an imaging sensor samples the incoming light rays to acquire a SR image. This image is at a particular pose in space and the warping transforms the image. Optical, motion and sensor blur then further degenerate the image before it is down-sampled to produce a low resolution image. (b) The image generation pipeline (inverse compositional). Several low resolution images are sampled from a continuous light field. The images are combined via their weighting wrt. their distance to the ideal image with the same resolution. The low resolution images are transformed to a common reference frame. The images are up-sampled and then inverse blurring is applied.

scaling matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & s^{-1} \end{bmatrix}, \tag{3}$$

where s is the desired scale factor. Note, however, that this means that we must transform between SR pixel units and LR pixel units in the equations.

A pixel in the SR image is then obtained from the LR image by performing a homographic warping as:

$$\mathbf{p}_{sr} = \frac{\mathbf{S}\mathbf{p}_{lr}}{\mathbf{e}_3^T \mathbf{S}\mathbf{p}_{lr}}, \tag{4}$$

where $\mathbf{p}_{lr}$ are the corresponding low-resolution pixels in normalized coordinates. The equivalent intensity and depth up-sampling are done via warping as in (5) given further.

3) **Intensity and depth warping:** The super-resolution image $\mathbf{I}^*_{sr}$ and depth-map $\mathbf{D}^*_{sr}$ of dimensions $sm \times sn$ are finally obtained by warping the corresponding low resolution images such that

$$\begin{cases} \mathbf{I}^*_{sr} = \mathbf{I}^*_{lr}\left(w\left(\mathbf{S}^{-1}, \mathbf{p}^w\right)\right) \\ \mathbf{D}^*_{sr} = \mathbf{D}^*_{lr}\left(w\left(\mathbf{S}^{-1}, \mathbf{p}^w\right)\right) \end{cases} \tag{5}$$

where $\mathbf{p}^w$ are the geometrically warped pixels from (2). The corresponding warped intensities are obtained by interpolation (nearest-neighbour, bi-linear or bi-cubic). In practice the depth warping function is optimised and computed differently as in [24] and bi-linear interpolation is used.

4) **Blur:** The function $f(\mathbf{I}^w, \mathbf{B}^{-1})$, is a filter which performs image deconvolution. This will be assumed to be a post-processing step of the reconstructed SR image, that can be achieved using for example a Wiener filter [32].

5) **Combination matrix:** The matrices $\mathbf{C}_i^{\mathbf{I}}$ and $\mathbf{C}_i^{\mathbf{D}}$ are normalized diagonal "combination" matrices ($\sum_i^N \mathbf{C}_i = \mathbf{I}$) that allow to correctly combine the input depth-map and images into a consistent high resolution one. This will be shown to minimise the difference in image resolution and will be detailed in the next Section.

A. Image resolution distance function

One of the main contributions of this paper is based on how the low resolution images are combined to form a high-resolution image. Classic techniques mainly average the aligned images using a smoothing point-spread function. This naive approach has the effect of simply considering the combination matrices to be $\mathbf{C}^{\mathbf{I}} = \mathbf{I}$. Clearly, this results in a simple average of the input warped images. This often yields a blurred reconstruction since the images taken with a highly different resolution than the SR image and treated the same as those which contain as much detail as those taken by the SR camera. In reality though, the images undergo full 3D transformation and non-linear light field sampling effects are hard to model. To solve this, the aim is to define a distance function with allows to *minimise the effective resolution* of the LR image with the SR image.

The following will show that a LR camera can undergo a 3D transformation with respect to the SR image such that it sees the same effective light rays in space (i.e. the same resolution). This also means that we can compute an "optimal virtual image" with the same resolution as the LR image such that it intersects the same viewing cones as the SR image. This can be seen intuitively as moving the camera toward the scene so that it sees an effective higher resolution (even if it does not cover the same total area as the SR camera).

To better understand, consider the Figure 2. The SR image is defined by the frame $\mathbf{T}$. The current LR image which must be used to generate a part of the SR image is defined in Frame $\mathbf{T}_c$. Both the LR and SR images observe a vertex $\mathbf{v} \in \mathbb{R}^3$ of the scene. The light reflected of the surface at $\mathbf{v}$ forms cones in space that are projected onto the SR and LR images respectively. Now consider moving a virtual camera defined by the frame $\mathbf{T}_o$ and with the same resolution as the LR image. This camera can move in 3D via its homogeneous transformation matrix $\mathbf{T}_o = (\mathbf{R}_o, \mathbf{t}_o)$.

The first goal is to determine the position in 3D space of the virtual camera such that it has the same effective resolution as the SR image. For each viewing cone, this is equivalent to minimising the area between the SR image's pixel size and the intersection between the virtual image plane and the cone. This area is minimised by computing the following equality:

$$\mathbf{S} - \left(\mathbf{R}_o - d^{-1}\mathbf{t_o}\mathbf{n}^T\right) = \mathbf{0}, \tag{6}$$

where $\mathbf{S}$ is the scaling homography defined in Section II-2 and the right hand side is the parametrised homography describing the equivalent transformation in 3D. The viewing cone intersects the 3D surface at vertex $\mathbf{v}$ with a certain radius. This forms a plane with the surface with the normal $\mathbf{n}$, and $\mathbf{n}^T\mathbf{n} = 1$. This normal is known from the dense 3D map, and is obtained by a local cross product on the image grid. $\mathbf{t}_o$ is the translation vector of the virtual camera and d is the distance between the camera centre of projection and the plane:

$$d = |\mathbf{n}^T\mathbf{v}^*|. \tag{7}$$

To reduce the number of solutions we first set $\mathbf{R}_o = \mathbf{I}$ and solve (6) for the translation vector as:

$$\mathbf{t}_o = d(\mathbf{I} - \mathbf{S})\mathbf{n}. \tag{8}$$

Since the scale is invariant to rotations around the Z axis, only the other two axes need to be set. In practice, the rotation is set such that the optical axis of the virtual camera is in the direction of the viewing ray of each pixel. This has the effect of ensuring that we minimise the distance in both translation and rotation between the virtual image and the LR image. This also means that the virtual camera is centred on each pixel in the image which helps avoid optical lens distortion effects and (see Figure 2). This matrix can be computed by

$$\begin{cases} \mathbf{r}_{oz} = \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}, \\ \mathbf{r}_{ox} = \mathbf{r}_y \wedge \mathbf{r}_{oz}, \\ \mathbf{r}_{oy} = \mathbf{r}_{oz} \wedge \mathbf{r}_{ox}. \end{cases} \tag{9}$$

Note that the computation of the virtual camera for each pixel and subsequently its distance, is not computationally expensive.

Following the definition of the optimal pose, it is now possible to define the error metric between the LR pixel and the ideal LR pixel. This transforms directly into a weighting coefficient for each vertex $\mathbf{v}^*$ and each LR pixel intensity, both defined with respect to the current estimated pose $\mathbf{T}_c$:

$$\mathbf{C}^{\mathbf{I}}(\mathbf{v}^*) = \left(\|(\mathbf{T}_c - \mathbf{T}_o)\overline{\mathbf{v}^*}\| + \epsilon\right)^{-1}, \tag{10}$$

where ϵ is a noise constant and $\overline{\mathbf{v}^*}$ is the homogeneous vertex coordinates.

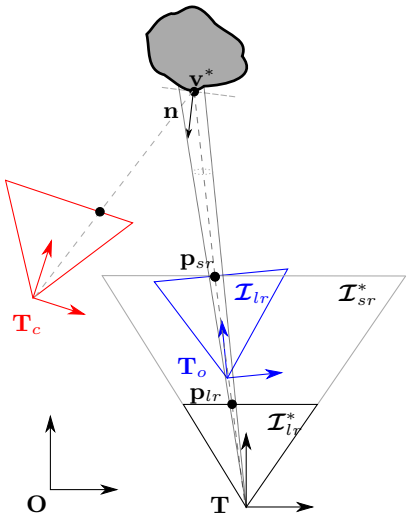


Fig. 2. Super-resolution camera poses. The optimal pose $\mathbf{T}_o$ projects the vertex $\mathbf{v}$ on the low resolution camera with the same resolution as the super-resolved image.

It can be seen from this error that it constrains the 5 degrees of freedom (i.e. not the rotation around Z). Also if the current image (LR) moves towards the optimal resolution then the error is zero whilst as it moves away the error increases. The scale factor which combines the rotational and translation components is determined by the vertex $\overline{\mathbf{v}^*}$ on the surface.

B. Depth weighting coefficients

For the depth weighting, a theoretical random error model proposed by [33] can be used. The depth weighting coefficient is then

$$\mathbf{C}^{\mathbf{D}} = \frac{fb}{m\sigma_d} diag(\mathbf{D})^{-2} \tag{11}$$

where m is a constant, f is the focal length of the camera, b is the baseline and σ_d is the disparity standard deviation.

III. Super-resolution visual SLAM

A. Cost function

The super-resolution visual SLAM problem is defined here to be that which estimates, incrementally, the set of camera poses $\mathbf{T}_i(\mathbf{x}_i)$ whilst simultaneously estimating the super-resolved depth image $\mathbf{D}^*_{sr}$ and the super-resolved intensity measurements $\mathbf{I}^*_{sr}$ from a set of low resolution images. This is achieved by considering the following photometric and depth errors:

$$\mathbf{e_I} = \sum_i^N \mathbf{C}^{\mathbf{I}}_i \left(\mathbf{I}^*_{sr} - \mathbf{I}_i \left(w\left(\widehat{\mathbf{T}}_i\mathbf{T}(\mathbf{x}_i), \mathbf{D}_i, \mathbf{p}, \mathbf{S}\right)\right)\right), \tag{12}$$

$$\mathbf{e_D} = \sum_i^N \mathbf{C}^{\mathbf{D}}_i \left(\mathbf{D}^*_{sr} - \mathbf{D}_i \left(w\left(\widehat{\mathbf{T}}_i\mathbf{T}\mathbf{x}_i, \mathbf{D}_i, \mathbf{p}, \mathbf{S}\right)\right)\right). \tag{13}$$

where it is supposed that for each pose there exists an incremental pose that combines homogeneously with the global pose to give the true transformation: $\exists\tilde{\mathbf{x}}_i : \widehat{\mathbf{T}}_i\mathbf{T}(\tilde{\mathbf{x}}_i) = \overline{\mathbf{T}}_i$. The full state vector representing the variables is then

$$\left[\mathbf{I}^*_{sr}, \mathbf{D}^*_{sr}, \mathbf{x}_1, \ldots, \mathbf{x}_N\right]. \tag{14}$$

Non-linear optimization of this error can then be decomposed via marginalization into three separate non-linear minimization phases which are performed iteratively for each low resolution input image: i.e. pose estimation, depth estimation and intensity estimation. This is the optimal formulation for the joint problem assuming that the initial super-resolution depth and intensities measurements are locally close to the solution.

B. 3D image registration and tracking

For each current image, the unknown motion parameters $\mathbf{x} \in \mathbb{R}^6$ are defined as:

$$\mathbf{x} = \int_0^1 (\boldsymbol{\omega}, \boldsymbol{\upsilon}) dt \in se(3), \tag{15}$$

which is the integral of a constant velocity twist which produces a pose $\mathbf{T}$. The pose and the twist are related via the exponential map as $\mathbf{T} = e^{[\mathbf{x}]_\wedge}$ with the operator $[.]_\wedge$ as:

$$[\mathbf{x}]_\wedge = \begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{\upsilon} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $[.]_\times$ represents the skew symmetric matrix operator.

Thus the pose cost function is then obtained by simultaneously minimising the errors of equation (12) and (13) in a robust least square procedure

$$\mathcal{C}(\mathbf{x}) = \lambda_\mathbf{I}^2 {\mathbf{e_I}}^T \mathbf{W_I} \mathbf{e_I} + \lambda_\mathbf{D}^2 {\mathbf{e_D}}^T \mathbf{W_D} \mathbf{e_D}, \tag{16}$$

where $\lambda_{(.)}$ are weighting scalar gains and where $\mathbf{W}_{(.)}$ are diagonal weighting matrices obtained by M-estimation [34]. The unknown $\mathbf{x}$ is then iteratively estimated using

$$\begin{cases} \mathbf{x} = -(\mathbf{J}^T \mathbf{W} \mathbf{J})^{-1} \mathbf{J}^T \mathbf{W} \begin{bmatrix} \mathbf{e_I} & \mathbf{e_D} \end{bmatrix}^T \\ \widehat{\mathbf{T}} \leftarrow \widehat{\mathbf{T}} \mathbf{T}(\mathbf{x}), \end{cases} \tag{17}$$

where $\mathbf{J}$ contains the stacked Jacobian matrices of the errors and $\mathbf{W}$ contains the stacked weighting matrices. More details on such a minimisation can be found in [24].

Minimising both errors provides a lot of advantages since photometric and depth informations are complementary. First the depth error term usually offers a larger domain of convergence and a fast minimisation. It also allows to track textureless areas, but it is sensitive to noise and not efficient with symmetric objects or with unconstrained scenes. On the other hand, the photometric term allows to track any textured areas with a better precision [23].

C. Super-resolution

Since the matrices ${\mathbf{C^I}}_i$ are diagonal, minimising equation (12) with respect to the photometric parameter $\mathbf{I}^*_{sr}$ can be done incrementally as new low resolution images $\mathbf{I}(t)$ are registered and warped onto the super-resolved frame, leading to the following update rule

$$\begin{cases} {\mathbf{C^I}}_{sr}(t) \leftarrow {\mathbf{C^I}}_{sr}(t-1) + \mathbf{C^I}(t) \\ \mathbf{I}_{sr}(t) \leftarrow \frac{\mathbf{C}^I_{sr}(t-1)\mathbf{I}_{sr}(t-1) + \mathbf{C^I}(t)\mathbf{I}^w(t)}{{\mathbf{C^I}}_{sr}(t)} \end{cases}, \tag{18}$$

where ${\mathbf{C^I}}_{sr}(t-1)$ is the global intensity cost at time $t-1$ and $\mathbf{I}^w(t)$ is the current image, warped from the registration process of Section III-B.

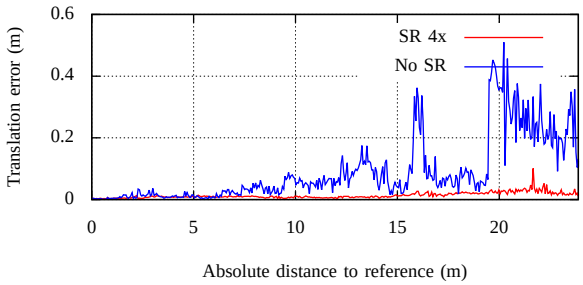


Fig. 3. Absolute translation error with respect to the distance to the reference image. The super-resolution algorithm reduces the localisation error.

The same procedure is applied for the depth parameter $\mathbf{D}^*_{sr}$

$$\begin{cases} {\mathbf{C^D}}_{sr}(t) \leftarrow {\mathbf{C^D}}_{sr}(t-1) + \mathbf{C^D}(t) \\ \mathbf{D}_{sr}(t) \leftarrow \frac{\mathbf{C}^\mathbf{D}_{sr}(t-1)\mathbf{D}_{sr}(t-1) + \mathbf{C^D}(t)\mathbf{D}^w(t)}{{\mathbf{C^D}}_{sr}(t)} \end{cases}. \tag{19}$$

IV. Experiments

A. Real-time implementation

A real-time implementation of the super resolution tracking and mapping algorithm has been made on GPU using the OpenCL library. The algorithm runs at 30 Hz with low resolution input images of size 640×480 pixels and a $4\times$ super-resolution factor on a Nvidia GTX 670 card. For the pose estimation minimisation, a coarse to fine multi-resolution approach is employed. The minimisation begins at the lower resolution, and the result is used to initialize the next level repeatedly until the higher resolution is reached. In this way, larger displacements are minimised at low cost on smaller images. Since the RGB-D sensor usually provides noisy depth measurements, a bilateral filtered is applied to remove noise whilst preserving discontinuities. The filtered depth-map is only used for pose estimation, whilst the raw depth-map $\mathbf{D}$ is used for depth integration, in order to preserve details in the integration process.

B. Simulated results

The algorithm has been tested on a synthetic sequence of images with ground truth poses, generated from the sponza atrium model [35]. The sequence is a 30 meters corridor with textured surfaces. The input images were downscaled to an input resolution of 160×120 pixels. The reference image is taken at the beginning of the sequence and then the camera moves along the corridor. Two experiments were conducted: first a simple tracking is applied to each current image without using super-resolution integration. Then the same sequence is tracked with super-resolution integration, with a scale factor $s = 4$. For both methods, we set the gain which controls the depth error $\lambda_D = 0$, in order to only compare the influence of the photometric integration in the tracking process. The plot reported on Figure 3 shows the absolute translation error with respect to the distance to the reference frame for both approaches. It can be seen that the super-resolution approach clearly improves the localisation error, by integrating new information as the camera moves along the trajectory.

C. Experimental results

The visual SLAM algorithm has been successfully tested on a number of real scenes. The images of Figure 4 reports the results of an experiment performed in an office containing a desk, with different objects and books. The RGB-D camera used for this experiment is an Asus Xtion Pro Live, capturing low resolution images of 640×480 pixels at 30Hz. The super-resolution SLAM is performed in real-time with a scale factor $s = 4$. The original reference image is shown on Figure 4(c). The camera was then moved around the desk with different motions. The image of Figure 4(d) shows the photometric image obtained at the end of the sequence. The Figures 4(e) and 4(f) show a region of interest. We can visually see that the super-resolved image is highly detailed compared to the original one. The Figures 4(a) and 4(b) show the Phong shaded surfaces computed from the depth-maps and the surfaces normals. We can see that compared to the original depth-map, the super-resolved depth-map is much more detailed and less noisy. Other aspects of our visual SLAM method, such as the ability to track during very rapid motion, or its robustness to camera occlusions are illustrated in our submitted video[1].

V. Conclusion

In conclusion, this paper has proposed a new visual SLAM technique which integrates 6DOF pose and dense structure simultaneously with the color information contained in the images over time. A novel inverse model has been provided for creating a super-resolution map from many low resolution images based on a 3D distance criteria which minimises the difference in resolution between the low resolution image in 3D and integrates the super-resolution image. Experimental results are given showing that this technique runs in real-time (30Hz) and is able to map large scale environments in high-resolution whilst simultaneously improving the accuracy and robustness of the tracking.

Future research in this direction will be focused at using the resolution distance criteria proposed here to better choose the position of the key frames in space. It would also be interesting to use this approach to take into account illumination changes on the surface within a dynamic model.

References

[1] S. Seitz, B. C. amd J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 519–526.

[2] J.-P. P. Renaud and R. Keriven, "Modelling dynamic scenes by registering multi-view image sequences," in *In International Conference on Computer Vision and Pattern Recognition*, 2005, pp. 822–827.

[3] P. Merrell, A. Akbarzadeh, L. Wang, J. michael Frahm, and R. Y. D. Nistr, "Real-time visibility-based fusion of depth maps," in *In Int. Conf. on Computer Vision and Pattern Recognition*, 2007.

[4] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1362–1376, 2010.

[5] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 303–312.

[1] http://youtu.be/q51E1NV0Ouw

[6] S. Baker and T. Kanade, "Super resolution optical flow," Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-99-36, October 1999.

[7] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025 –1039, Oct. 1998.

[8] G. Silveira and E. Malis, "Unified direct visual tracking of rigid and deformable surfaces under generic illumination changes in grayscale and color images," *International Journal of Computer Vision*, vol. 89, no. 1, pp. 84–105, 2010.

[9] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," in *In International Conference on Computer Vision and Pattern Recognition*, 2003, pp. 211–218.

[10] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, July 2004, pp. 652–659.

[11] A. J. Davison and D. W. Murray, "Simultaneous localisation and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2002.

[12] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 523–535, 2002.

[13] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real time localization and 3d reconstruction," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 0-0 2006, pp. 363–370.

[14] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments," in *12th International Symposium on Experimental Robotics*, Delhi, India, December 18–21 2010.

[15] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard, "Real-time 3d visual slam with a hand-held rgb-d camera," in *RGB-D Workshop on 3D Perception in Robotics*, Västerås, Sweden, April 8 2011.

[16] J. Sturm, K. Konolige, C. Stachniss, and W. Burgard, "3d pose estimation, tracking and model learning of articulated objects from dense depth video using projected texture stereo," in *RGB-D: Advanced Reasoning with Depth Cameras Workshop, RSS*, Zaragoza, Spain, June 27 2010.

[17] A. I. Comport, E. Malis, and P. Rives, "Accurate quadrifocal tracking for robust 3d visual odometry," in *IEEE International Conference on Robotics and Automation*, April 2007, pp. 40–45.

[18] M. Meilland, A. I. Comport, and P. Rives, "A spherical robot-centered representation for urban navigation," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[19] C. Zach, T. Pock, and H. Bischof, "A globally optimal algorithm for robust tv-l1 range image integration," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[20] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

[21] T. Whelan, J. McDonald, M. Kaess, M. Fallon, H. Johannsson, and J. Leonard, "Kintinuous: Spatially extended KinectFusion," in *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, Jul 2012.

[22] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 343 –350.

[23] F. Steinbruecker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense rgb-d images," in *Workshop on Live Dense Reconstruction with Moving Cameras at the Intl. Conf. on Computer Vision (ICCV)*, 2011.

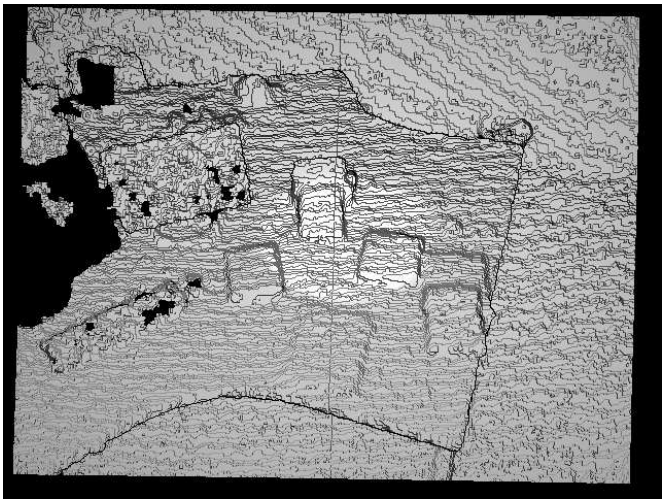
[24] T. Tykkala, C. Audras, and A. I. Comport, "Direct iterative closest point for real-time visual odometry," in *International Workshop on Computer Vision in Vehicle Technology*, 2011, pp. 2050–2056.

[25] D. Damen, A. Gee, W. Mayol-Cuevas, and A. Calway, "Egocentric real-time workspace monitoring using an rgb-d camera," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, August 2012.

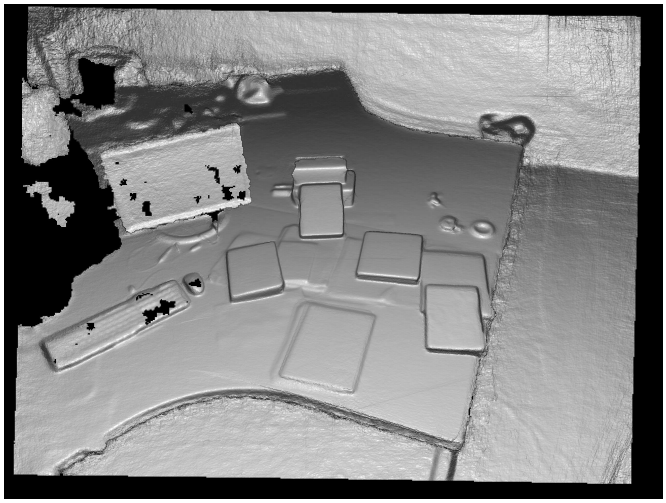
[26] T. S. Huang and R. Y. Tsay, "Multiple frame image restoration and registration," in *Advances in Computer Vision and Image Processing*, vol. 1. Greenwich: JAI, 1984, pp. 317–339.

[27] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[28] B. Bascle, A. Blake, and A. Zisserman, "Motion deblurring and super-resolution from an image sequence," in *In Proceedings of the Fourth*

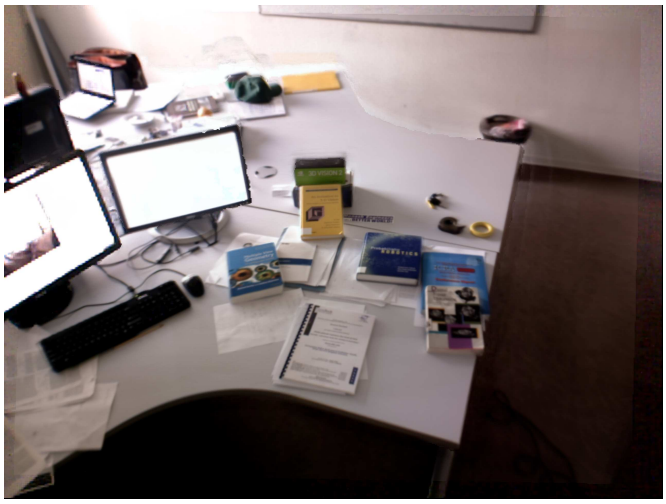


(a) Original 3D surface with phong shading.



(b) Super-resolved 3D surface with phong shading.

(c) Original image.



(d) Super-resolved image.

(e) Original image, area of interest.

(f) Super-resolved image, area of interest.

Fig. 4. On left column, the original LR augmented image of resolution 640×480 pixels. On the right column the SR augmented image with a resolution of 2560×1920 pixels. The first row shows a Phong shaded surface computed from the depth-map of the images. The second row is the photometric map of the images, and the last row is a zoom on an interesting area. It can be seen that the super-resolution SLAM algorithm greatly improves depth measurements as well as intensity measurements.

European Conference on Computer Vision. Springer-Verlag, 1996, pp. 573–582.

[29] D. Capel and A. Zisserman, "Computer vision applied to super resolution," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, may 2003.

[30] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, pp. 21 – 36, may 2003.

[31] H. Su, L. Tang, Y. Wu, D. Tretter, and J. Zhou, "Spatially adaptive block-based super-resolution," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1031 –1045, march 2012.

[32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.

[33] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/2/1437

[34] P. Huber, *Robust Statistics*. New york, Wiley, 1981.

[35] M. Dabrovic and F. Meinl, "Sponza atrium model," 2002, http://www.crytek.com/cryengine/cryengine3/downloads.