Robust direct vision-based pose tracking using normalized mutual information

Luo, Hang, Pape, Christian, Reithmeier, Eduard

Hang Luo, Christian Pape, Eduard Reithmeier, "Robust direct vision-based pose tracking using normalized mutual information," Proc. SPIE 10819, Optical Metrology and Inspection for Industrial Applications V, 108190T (2 November 2018); doi: 10.1117/12.2500857

SPIE.

Event: SPIE/COS Photonics Asia, 2018, Beijing, China

Robust Direct Vision-based Pose Tracking using Normalized Mutual Information

Hang Luo*[a], Christian Pape [a], Eduard Reithmeier [a]
[a] Institute of Measurement and Automatic Control, Leibniz University Hannover, Nienburger Str.17, D-30167 Hannover

ABSTRACT

This paper presents a novel visual tracking approach that combines the NMI metric and the traditional SSD metric within a gradient-based optimization frame, which can be used for direct visual odometry and SLAM. We firstly derivate the closed form expression for first- and second-order analytical NMI derivatives under the assumption of rigid-body transformations, which then can be used by subsequent Newton-like optimization methods. Then we develop a robust tracking scheme that utilizes the robustness of NMI metric while keeping the optimization characteristics of SSD-based Lucas-Kanade (LK) tracking methods. To validate the robustness and accuracy of the proposed approach, several experiments are performed on synthetic datasets as well as real image datasets. The experimental results demonstrate that our approach can provide fast, accurate pose estimation and obtain better tracking performance over standard SSD-based methods in most cases.

Keywords: direct visual tracking, normalized mutual information, sum of squared differences, nonlinear optimization

1. INTRODUCTION

Real-time vision-based pose tracking is a critical technique for a wide range of robotic applications. Most of current available visual tracking approaches can be divided into two primary classes: feature-based and direct tracking methods. Compared to traditional feature-based methods normally limited to certain feature type, direct tracking is able to take full advantage of the available image information and thereby save the costly computation spent on feature extraction and matching. Therefore, direct methods have become increasingly popular recently in scenarios requiring fast tracking speed and robustness towards feature-less scenes (e.g. visual odometry and SLAM). This paper mainly deals with direct tracking approaches applied in visual odometry.

When performing such direct tracking, the tracking problem is usually transformed into a non-linear optimization problem, which aims to maximize or minimize a registration function that represents the similarity between a template image and the current image. Many recent works have used the traditional SSD metric that directly compares the luminance differences between two images, which is real-time capable due to its standard least squares optimization structure. However, SSD-based approaches are based on the brightness constancy assumption that can be violated quite easily in real-world applications, for example by partial occlusions, shadows and variations in lighting conditions. Several solutions have been proposed to deal with these variations. [1, 2] use robust M-estimators and [3] introduce an elegant re-weighting scheme from the perspective of probability, which increases the robustness to occlusions by assigning relatively small weights to those occluded regions. An affine brightness transfer function [2, 4] has been employed to handle more fine-grained illumination changes in a joint optimization frame. However, in spite of taking such measures above, the tracking performance of SSD-based approaches would still inevitably suffer certain deteriorations in case of scene variations.

In contrast to the SSD metric, mutual information (MI) that measures the information quantity shared by two images (in this context) has proved to be robust towards partial occlusion and lighting changes, and has been commonly used in medical image registration [5, 6]. [7] subsequently presented a normalized measure of mutual information (i.e. NMI) to increase the robustness to the changes in overlapping regions, which obtained a distinct improvement in the behavior of rigid registration of MR-CT and MR-PET images. Since then, various MI-based implementations have been proposed, which can be divided into two main categories: non-gradient and gradient based methods. In earlier works, non-gradient based methods (such as hill climbing [6], Powell's method [8]) were quite popular, since MI is originally calculated from the discrete joint intensity histogram (i.e. a two-dimensional histogram of combinatorial intensity levels in two images) that makes an explicit solution to the derivatives of MI function unavailable. Normally, these non-gradient optimization methods are computationally inefficient, thereby limiting their applications to real-time tracking. With the introduction of

Optical Metrology and Inspection for Industrial Applications V, edited by Sen Han, Toru Yoshizawa, Song Zhang, Proc. of SPIE Vol. 10819, 108190T · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2500857

partial volume interpolation technique [9, 10], an analytic computation of MI-related derivatives becomes available. In [10], the authors formulate the MI metric as a continuous and differentiable function of the estimated parameters using B-spline Parzen windows. Based on this, a Levenberg-Marquardt (LM) optimizer is then employed for parameters estimation. [11] presents a novel Newton-like optimization approach within an inverse compositional optimization frame, which enables the real-time template-based tracking. In their work, they consider second order terms in the computation of Hessian matrix for the first time, and demonstrate that this is of significance to the optimization results. Although [11] claims that their proposed modified Newton-like optimization approach has a wider convergence domain than the traditional Newton approach, we find that it still easily gets stuck in incorrect local optima or even goes divergent in our experiments, which has also been pointed out in [12].

However, most efficient optimization methods mentioned above only focus on the optimization of MI criterion, and quite few works feature normalize mutual information (NMI) within a gradient-based optimization strategy and apply it to visual odometry, which might be because the derivation of analytical gradients of NMI is more complicated than its counterpart of standard MI. In [13], the authors only present the first-order analytical derivatives of NMI function without further considering second-order terms. Only using the first-order gradients, [14] use a Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimizer for NMI optimization in visual localization, rather not visual odometry. The BFGS optimizer is essentially a quasi-Newton method in combination with the use of line search strategies, which only requires a Jocabian matrix to be supplied and then iteratively construct a Hessian during optimization. In contrast, our approach evaluates a Hessian matrix directly and employ a Levenberg-Marquardt algorithm, a real Newton-type method, for subsequent optimization. A direct comparison of these two methods is beyond the scope of this work.

In this paper, we propose a hybrid tracking scheme that utilizes the inherent robustness of NMI metric while maintaining the relatively wide convergence basin of SSD metric. A series of experiments are conducted on both synthetic and real-sensor benchmark datasets [12, 15, 16]. The experimental results show that our proposed method has a distinct advantage over standard SSD-based methods, in terms of accuracy and robustness. To the best of our knowledge, this is the first work to employ NMI criterion within a Newton-type optimization frame for the task of direct visual odometry without any prior data.

2. PROBLEM FORMULATION

In this section, we firstly present the general mathematical formulation of direct visual tracking problems as well as the classical LK tracking methods with SSD metric. Then, a brief introduction of NMI metric will be given.

2.1 Formulation of direct visual tracking

Direct visual tracking essentially refers to a class of approaches based on the optimization of an image registration function, which usually involves a template (reference) image I_r, a current (target) image I_t, a geometric warp model $\omega(\theta)$ parametrized by the displacement parameters θ, and a similarity function $f(\cdot)$. Generally, this problem can be written as

$$\theta^* = \arg\max_{\theta \in \mathrm{V}(\theta)} f(I_r, \omega(I_t, \theta)) \tag{1}$$

where V represents the available parameter space under given constraints, and θ^* is the displacement parameters that maximize the similarity between the template I_r and the warped current image I_t. Note that although the formulation (1) is a maximization problem, it can be easily rewritten as a minimization form in case of the use of dissimilarity metrics, such as the standard SSD metric. The warp model $\omega(\theta)$ depends on the specific transformation models, e.g. homograph transformation and affine transformation. In this paper, we assumes a rigid body transformation model, and the displacement parameters θ are defined in $\mathfrak{se}(3)$ lie algebra associated to the SE(3) group with six dimensions.

As can be seen from (1), different similarity functions lead to different registration (cost) functions and based on the individual forms of functions, different optimization methods would be employed, which results in different optimization characteristics (e.g. convergence domain and rate). Considering the requirement of real-time tracking in visual odometry, the SSD-based LK tracking method is the mostly commonly used approach in recent works [3, 17, 18].

2.2 Lucas-Kanade Framework with SSD

[19] gives detailed description of the LK method as well as its several variants, which has been applied for various parameter estimation tasks. In direct image registration, SSD metric is often used within the LK framework. In this case, the formulation (1) can be rewritten as

$$\theta^* = \arg\min_{\theta \in \mathrm{V}(\theta)} \sum_{x \in I_r} \left\| I_r(x) - I_t(\omega(x,\theta)) \right\|^2 \tag{2}$$

where x denotes the pixel coordinates in image plane, and the cost function here is aimed to minimize the photometric error based on the SSD of pixel intensities. Note that this optimization problem in (2) is in the least-squares form, for which there has existed several efficient nonlinear optimizer, such as Gauss-Newton and LM methods. In each iteration, the algorithms search the update that minimize the function in (2) as

$$\Delta\theta^k = \arg\min_{\Delta\theta \in \mathrm{V}(\theta)} \sum_{x \in I_r} \left\| I_r(x) - I_t(\omega(\omega(x,\Delta\theta),\theta^k)) \right\|^2 \tag{3}$$

The current parameters can then be updated as follows:

$$\omega(\omega(x,\Delta\theta^k),\theta^k) \rightarrow \omega(x,\theta^{k+1}) \tag{4}$$

In [19], the authors also propose an inverse compositional (IC) variant to classical LK, which can save large amounts of calculations. In the IC structure, the optimization function and the corresponding update are adapted as

$$\Delta\theta^k = \arg\min_{\Delta\theta \in \mathrm{V}(\theta)} \sum_{x \in I_r} \left\| I_r(\omega(x,\Delta\theta)) - I_t(\omega(x,\theta^k)) \right\|^2 \tag{5}$$

$$\omega(\omega^{-1}(x,\Delta\theta^k),\theta^k) \rightarrow \omega(x,\theta^{k+1}) \tag{6}$$

Since the update parameters are applied to the reference image instead of the current image, the derivatives of the cost function can be partially precomputed only once using the gradients of the reference image. In our approach, we also employ this IC optimization structure.

2.3 Normalized mutual information

Although the SSD based LK formulation has superior optimization characteristics with the use of some least-square optimizers, it has proved to be vulnerable to scene variations (such as occlusions and illumination changes) that can easily violate the assumption of constant brightness. In contrast to directly comparing the differences of pixel intensities, distributions of pixel intensities can provide a much more robust criterion to measure the similarity (or difference) between two images. Mutual information, an information-theoretical concept proposed by Shannon [20], can indicate the underlying relations between the probability distributions of two random variables (images), which now is applied quite successfully in medical and remote-sensing image registration [6, 11]. For convenience, we will describe the mutual information for two images, as used in image registration, not in a general sense. The most intuitive definition of MI is given as

$$\mathrm{MI}(I_r, I_t) = \mathrm{H}(I_r) + \mathrm{H}(I_t) - \mathrm{H}(I_r, I_t) \tag{7}$$

where I_r, I_t still denote the reference and current images as before. $\mathrm{H}(I_r)$, $\mathrm{H}(I_t)$ represent the entropy values of corresponding images, and $\mathrm{H}(I_r, I_t)$ is the joint entropy of two images. If we combine this formulation with the previously defined direct tracking problem in (1), the current image I_t can be considered to depend on the displacement parameter θ, and then MI can be rewritten as a function with respect to θ

$$\mathrm{MI}(\theta) = \mathrm{H}(I_r) + \mathrm{H}(\omega(I_t,\theta)) - \mathrm{H}(I_r, \omega(I_t,\theta)) \tag{8}$$

Substituting the definition of entropy [20], this expression can be adapted as the following expression

$$\mathrm{MI}(\theta) = \sum_{r,t} p_{I_r I_t}(r,t;\theta) \log\left(\frac{p_{I_r I_t}(r,t;\theta)}{p_{I_r}(r)\, p_{I_t}(t;\theta)} \right) \tag{9}$$

In (9), r and t are possible pixel values of I_r and I_t respectively, and $p_{I_r}(r) = P(I_r = r)$ is the probability distribution function of image (or variable) I_r, and $p_{I_t}(t;\theta) = P(\omega(I_t,\theta) = t)$ the probability function of I_t. And $p_{I_r I_t}(r,t;\theta) = P(I_r = r, \omega(I_t,\theta) = t)$ represents the joint probability distribution of these two images. From the perspective of probability, this formulation can be interpreted as the Kullback-Leibler divergence between the joint probability distribution of two images $p_{I_r I_t}(r,t;\theta)$, and the joint distribution in case of complete independence of the images $p_{I_r}(r)\, p_{I_t}(t;\theta)$. The assumption is that the divergence between these two distributions of images would reach the maximum when they are totally identical, namely the displacement parameters are correctly estimated, and it becomes zero when the distributions are independent that normally means severely misaligned.

In [7], the authors point out that the traditional MI metric could be influenced adversely by the size of overlapping part of images, and propose a normalized measure of mutual information, i.e. NMI, which is proved to be more robust to changes in overlap

$$\mathrm{NMI}(\theta) = \frac{\mathrm{H}(I_r) + H(\omega(I_t,\theta))}{\mathrm{H}(I_r, \omega(I_t,\theta))} = \frac{\sum_{r,t} p_{I_r I_t}(r,t;\theta) \log(p_{I_r}(r)\, p_{I_t}(t;\theta))}{\sum_{r,t} p_{I_r I_t}(r,t;\theta) \log p_{I_r I_t}(r,t;\theta)} \tag{10}$$

In order to make this function differentiable, the Parzen windowing method with a third-order B-spline kernel are normally employed [9, 11, 13]. Firstly the joint distribution can be calculated from a normalized bi-dimensional histogram of the two images using the following expression

$$p_{I_r I_t}(r,t;\theta) = \frac{1}{N_x} \sum_x \phi(r - \overline{I}_r(x))\, \phi(t - \overline{I}_t(\omega(x,\theta))) \tag{11}$$

where N_x is the number of pixels selected in reference image, and ϕ is the B-spline kernel function. $\overline{I}_r$ and $\overline{I}_t$ are the respective scaled version of the images, and r and t are the possible values of the scaled images, normally $\{r,t\} \in [0, N_c]^2$ with N_c the number of histogram bins. Then the respective marginal probability distribution of the two images can be integrated from the joint probability distribution, e.g. for the reference image I_r

$$p_{I_r}(r) = \sum_t p_{I_r I_t}(r,t;\theta) = \frac{1}{N_x} \sum_x \phi(r - \overline{I}_r(x)) \sum_t \phi(t - \overline{I}_t(\omega(x,\theta))) \tag{12}$$

Due to the partition of unity constraint of B-spline kernel function [9], this term $\sum_t \phi(t - \overline{I}_t(\omega(x,\theta))) = 1$. The formulation (12) is thus written as

$$p_{I_r}(r) = \frac{1}{N_x} \sum_x \phi(r - \overline{I}_r(x)) \tag{13}$$

Similarly, the marginal distribution of the current image I_t is as follows

$$p_{I_r}(t;\theta) = \sum_r p_{I_r I_t}(r,t;\theta) = \frac{1}{N_x} \sum_x \phi(t - \overline{I}_t(\omega(x,\theta))) \tag{14}$$

Although several efficient optimization methods have been propose for standard MI optimization, few works investigate the optimization strategies of NMI metric, especially in those real-time applications. The next section presents a full derivation of analytical derivatives of NMI, which enables the possibility of applying Newton-type optimization methods.

2.4 Derivatives of NMI metric

As mentioned in section 2.2, an inverse compositional formulation is used in optimization, and the corresponding Jocabian and Hessian matrix can be given as

$$\mathrm{J} = \frac{\partial\, \mathrm{NMI}(\omega(I_r, \Delta\theta), \omega(I_t, \theta))}{\partial \Delta\theta} \tag{15}$$

$$\mathrm{H} = \frac{\partial^2\, \mathrm{NMI}(\omega(I_r, \Delta\theta), \omega(I_t, \theta))}{\partial \Delta\theta^2} \tag{16}$$

In this work, the Jacobian matrix is firstly considered and for simplifying the expressions, we define two notations $\mathrm{A} \triangleq \mathrm{H}(I_r) + H(\omega(I_t, \theta))$ and $\mathrm{B} \triangleq \mathrm{H}(I_r, \omega(I_t, \theta))$ to represent the joint entropy of the two images and the sum of their respective entropies, with respect to the update parameter $\Delta\theta$. Then substitute (10) into (15) and use the derivative chain rules

$$\begin{aligned} \mathrm{J} &= \frac{\partial\, \mathrm{NMI}(\Delta\theta)}{\partial \Delta\theta} = \frac{\partial\, \mathrm{B}(\Delta\theta)/\mathrm{A}(\Delta\theta)}{\partial \Delta\theta} \\ &= \frac{1}{\mathrm{A}^2(\Delta\theta)} \left(\mathrm{A}(\Delta\theta) \cdot \frac{\partial\, \mathrm{B}(\Delta\theta)}{\partial \Delta\theta} - \mathrm{B}(\Delta\theta) \cdot \frac{\partial\, \mathrm{A}(\Delta\theta)}{\partial \Delta\theta} \right) \end{aligned} \tag{17}$$

Applying (11)-(14) to (17) to yield the following expressions

$$\frac{\partial\, \mathrm{A}(\Delta\theta)}{\partial \Delta\theta} = \sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \left(1 + \log p_{I_r I_t}\right), \quad \frac{\partial\, \mathrm{B}(\Delta\theta)}{\partial \Delta\theta} = \sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \log p_{I_r} \tag{18}$$

$$\mathrm{J} = \sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \left(\frac{1}{\mathrm{A}} \log p_{I_r} - \frac{\mathrm{B}}{\mathrm{A}^2} \left(1 + \log p_{I_r I_t}\right) \right) \tag{19}$$

Then further derivate the Hessian matrix using the first-order gradient in (19)

$$\begin{aligned} \mathrm{H} &= \frac{\partial\, \mathrm{J}(\Delta\theta)}{\partial \Delta\theta} = \frac{\partial}{\partial \Delta\theta} \left(\sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \cdot \mathrm{M}(\Delta\theta) \right) \\ &= \sum_{r,t} \frac{\partial^2\, p_{I_r I_t}}{\partial \Delta\theta^2} \cdot \mathrm{M}(\Delta\theta) \; + \; \sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \cdot \frac{\partial\, \mathrm{M}(\Delta\theta)}{\partial \Delta\theta} \end{aligned} \tag{20}$$

where the first data term on the right side of the equation is the second-order part of Hessian, and the second data term represents the first-order part. As pointed out in [11, 21], the influence of the second-order part on MI optimization cannot be neglected, and we thus keep this term in Hessian expression of the NMI metric as well. As for the first-order part, using the chain rule on the derivative of M and then apply (18) to the derivative, can yield an accurate expression as follows

$$\begin{aligned} \sum_{r,t} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \cdot \frac{\partial\, \mathrm{M}(\Delta\theta)}{\partial \Delta\theta} &= 2 \cdot \left(\frac{\mathrm{B}}{\mathrm{A}^3} \cdot \left(\frac{\partial\, \mathrm{A}}{\partial \Delta\theta} \right)^2 - \frac{1}{\mathrm{A}^2} \cdot \frac{\partial\, \mathrm{A}}{\partial \Delta\theta} \cdot \frac{\partial\, \mathrm{B}}{\partial \Delta\theta} \right) \\ &- \sum_{r,t} \frac{\mathrm{B}}{\mathrm{A}^2} \frac{1}{p_{I_r I_t}} \cdot \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta}^{\mathrm{T}} \frac{\partial\, p_{I_r I_t}}{\partial \Delta\theta} \; + \; \sum_{r} \frac{1}{\mathrm{A}} \frac{1}{p_{I_r}} \cdot \frac{\partial\, p_{I_r}}{\partial \Delta\theta}^{\mathrm{T}} \frac{\partial\, p_{I_r}}{\partial \Delta\theta} \end{aligned} \tag{21}$$

As we can see from (19), (20) and (21), the derivatives of NMI are related to the derivatives of the joint distribution of these two images and the marginal distribution of the reference image.

Using the (11) and (13), we can get the following expressions

$$\begin{aligned}
\frac{\partial p_{I_r}}{\partial \Delta\theta} &= \frac{1}{N_x}\sum_x \frac{\partial \phi(r-\overline{I}_r(\omega(x,\Delta\theta)))}{\partial \Delta\theta} \\
\frac{\partial p_{I_r I_t}}{\partial \Delta\theta} &= \frac{1}{N_x}\sum_x \phi(t-\overline{I}_t(\omega(x,\theta)))\frac{\partial \phi(r-\overline{I}_r(\omega(x,\Delta\theta)))}{\partial \Delta\theta} \\
\frac{\partial^2 p_{I_r I_t}}{\partial \Delta\theta^2} &= \frac{1}{N_x}\sum_x \phi(t-\overline{I}_t(\omega(x,\theta)))\frac{\partial^2 \phi(r-\overline{I}_r(\omega(x,\Delta\theta)))}{\partial \Delta\theta^2}
\end{aligned} \tag{22}$$

In (22), the derivatives of B-spline function ϕ with respect to the update can be obtained as follows:

$$\begin{aligned}
\frac{\partial \phi(r-\overline{I}_r(\omega(x,\Delta\theta)))}{\partial \Delta\theta} &= -\frac{\partial \phi}{\partial r}\frac{\partial \overline{I}_r}{\partial \Delta\theta} \\
\frac{\partial^2 \phi(r-\overline{I}_r(\omega(x,\Delta\theta)))}{\partial \Delta\theta^2} &\simeq \frac{\partial^2 \phi}{\partial r^2}\frac{\partial \overline{I}_r}{\partial \Delta\theta}^{\mathrm{T}}\frac{\partial \overline{I}_r}{\partial \Delta\theta}
\end{aligned} \tag{23}$$

with

$$\frac{\partial \overline{I}_r}{\partial \Delta\theta} = \frac{\partial \overline{I}_r(x)}{\partial x}\frac{\partial x}{\partial \Delta\theta} \tag{24}$$

Since using the inverse compositional formulation, the derivatives of image intensities with respect to update parameters represented by (24), are all computed in the reference image at the point of $\Delta\theta = 0$. This means, once the reference image is chosen, the expressions of (23) and (24) would keep constant during the whole optimization and thus allow to be precomputed only once. Besides, although the expression of Hessian from (20) seems complicated and computation-costly, the computation in practice can be optimized by reusing the components utilized to calculate the function value as well as its Jacobian values (e.g. A, B, ∂A and ∂B), thereby saving large amounts of computations.

3. ROBUST DIRECT TRACKING SCHEME

The main idea of this hybrid scheme is to firstly employ a SSD based optimization approach to yield a coarse estimation of the displacement, and this coarse result is then refined using a NMI based LM method. The whole scheme is efficiently performed on a Gaussian pyramid, where the SSD optimization is performed on higher levels and NMI optimization on lower levels.

As illustrated above, a SSD optimizer is more capable of tracking relatively larger wraps (motions) than NMI optimizer, due to its larger convergence basin. It is thus natural to consider using SSD optimizer to provide an initial estimation that more likely falls into the narrow convergence domain of subsequent NMI optimizer. However, the occurring scene variations can easily bias the solutions of SSD optimization, which might result in much worse initial guesses. In order to cope with this situation, we employ the reweighting optimization scheme proposed by [3] based on the Students' t-distribution [22] noise model. Using this scheme, the original optimization problem in (2) is adapted into a reweighted least squares form as follows

$$\theta^* = \arg\min_{\theta\in \mathrm{V}(\theta)} \sum_{x\in I_r} w(x;\theta)\left\| I_r(x) - I_t(\omega(x,\theta))\right\|^2 \tag{25}$$

with

$$w(x;\theta) = \frac{\nu+1}{\nu + \left\| I_r(x) - I_t(\omega(x,\theta))\right\|^2 / \sigma^2} \tag{26}$$

where ν denotes the degrees of freedom and σ is the scale factor of t-distribution. The only difference from our work to [3] is that we use a generalized expectation-maximization (GEM) method to update the scale factor and the weights

alternately, instead of solving an equation iteratively in [3]. The experimental results presented in the next section show that this reweighting mechanism enhance the algorithms' robustness to partial occlusion and slight illumination changes to a certain extent, thereby more likely to provide a better coarse pose estimation.

From (19)-(24), we can see that one pixel with the higher intensity gradient magnitude contributes more to the final Jacobian and Hessian than its lower-magnitude counterparts. From the subset of pixels preselected by SSD optimizer, we further choose those points whose gradient magnitudes are beyond a preset threshold as the final candidates for NMI metric calculation. Then a classical LM method is chosen for the optimization of NMI metric and the optimization is only performed on the lower pyramid levels (higher resolution) with the initial guess from SSD optimization, which can both reduce the number of iterations and increase the convergence rate effectively.

4. EXPERIMENTS

Several experiments are conducted to evaluate the proposed tracking approach. The whole evaluation process can be divided into two steps: On the first step, we only run the experiments on single image pairs and test how the algorithm responds to perturbations under different scene variations; on the second step, the evaluation is conducted on several benchmark datasets in order to have a quantitative measure of its accuracy and robustness. The two steps above are both be conducted on synthetic data and real sensor data respectively [12, 15, 16]. The used Gaussian pyramid comprises of five levels, where the top three levels are used for SSD optimization and the rest 2 levels for NMI optimization.

4.1 Experiments on single image pairs

The experiments in this section are used to analyze the convergence domain of the algorithms (i.e. SSD, NMI and hybrid methods) through a series of tests on single image pairs with different initial positioning error in different scene variations. As proposed in [11], we use the root mean-square (rms) distance between the reference pixel coordinates in current image and the projected pixel coordinates with the estimated transformation as the evaluation criterion, which is given as

$$e(\theta)=\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left\|x_i^c-\omega(x_i,\theta)\right\|^2} \tag{27}$$

where x_i is the coordinate of the pixel selected for evaluation in the reference image, $\omega(x_i,\theta)$ denotes the projected coordinate of x_i in the current image, and x_i^c is the accurate (reference) coordinate of the correspondence of x_i in the current image. Unlike the work in [11] that only uses four corners for evaluation, we use SIFT descriptors and geometry consistency tests to select sufficient correspondence (no more than 50) for evaluation in order to increase the statistical power in convergence rate estimation.

A. Selection of correspondences for evaluation

In order to find corresponding pixels reliably, the selection process mainly comprises of two steps based on appearance similarity and spatial consistency. Firstly, extracting and matching SIFT features in the two images to generate an initial set of pixel pairs. Secondly, for every pair of the initial set, project the pixel in the reference image onto the current image using the ground-truth transformation offered by the dataset, and if the distance between the projected coordinates and the corresponding pixel coordinates is less than a preset value (in our experiments, 1px for synthetic images and 3px for real images), this pair would be used for evaluation.

B. Experiments on synthetic data

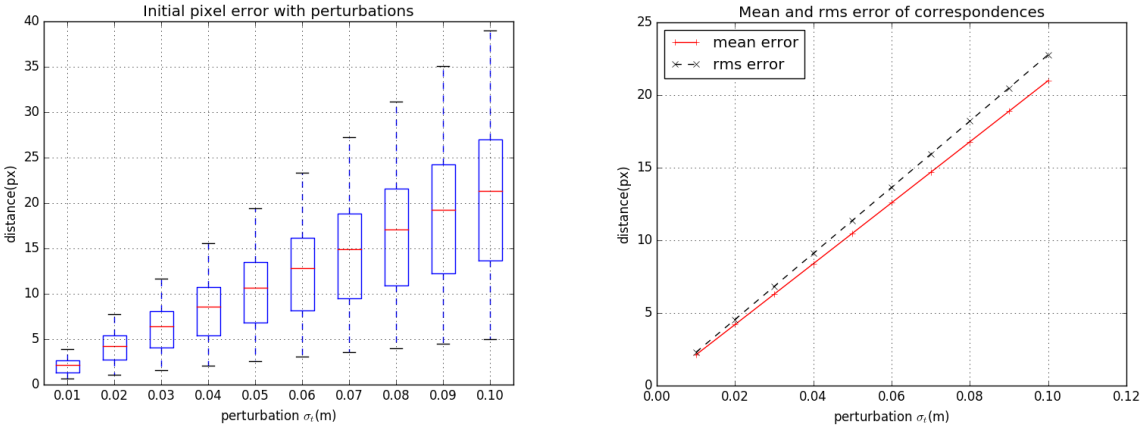
We extract two images from the datasets provided by [15], and firstly select corresponding pixel pairs in a nominal condition (without any scene variations). Fig.1 shows the two images used for subsequent experiments and the distribution of the distance error of selected correspondences under the ground-truth transformation.

(a) (b)

Figure 1. (a) Reference image and current image from left to right. (b) Distribution of the distance error of correspondences, with the mean error = 0.24px, rms error = 0.26px

Then we conduct the image alignment experiments with respect to a set of different initial poses under four different conditions: nominal condition, illumination changes, slight occlusion and severe occlusion. The initial pose parameters are generated automatically by adding a white Gaussian noise with the chosen σ to the ground-truth pose. In our experiments, the deviation σ_t for the translations varies from 0.001 to 0.1m and a fixed σ_r of 0.01 rad is chosen for rotations. However, only the variance values of perturbations on the pose cannot reflect the impact imposed on the images quite well, since this impact is also concerned with the scene depth (A smaller scene depth obviously leads to stronger modifications of the image, with respect to a same perturbation). Similar to [21], we present the distribution of initial pixel errors in Fig. 2, so as to discover thoroughly how strong the perturbations' impact on the scene are.

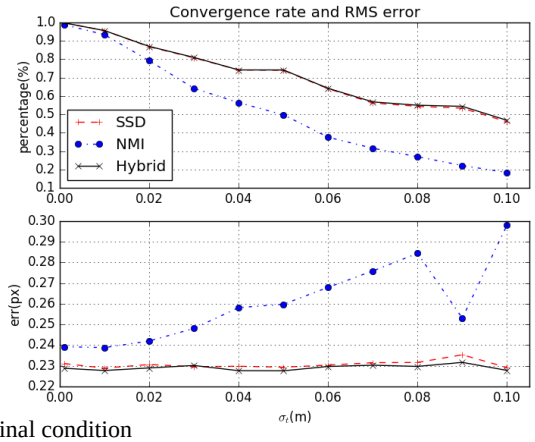
Note that for each sigma, we calculate the distance of selected correspondences 500 times with randomly generated parameters and measure the mean outcome of each correspondent, so that Fig. 2(a) actually shows the statistical distribution of the results of 500 trials. And further using the statistical data from (a) to generate the mean- and rms distance of all correspondences. Then with the initial positioning errors in Fig. 2, we perform the alignment experiments under different conditions, and employ (27) to analyze the convergence rate (the optimization is considered to be convergent if the criterion is blow 0.5px) and estimation accuracy. For each method in each condition with each perturbation, the alignment experiment is repeated 500 times and the experimental results are presented in the following Fig. 3.



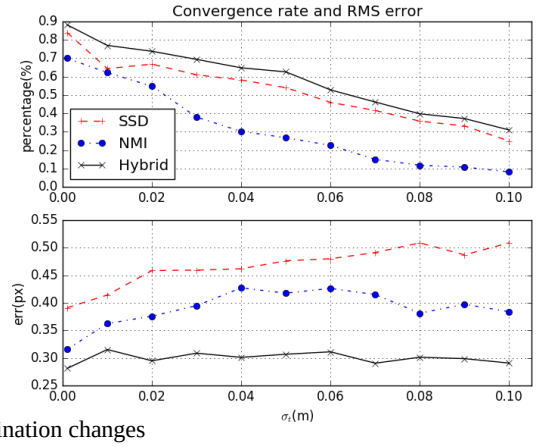
(a) (b)

Figure 2. (a) Boxplot for the initial pixel error with respect to perturbations. The boxes delimit the 25% and 75% quartiles, and the whiskers indicate the minimum and maximum. The red line is the median error of all correspondences. (b) Mean and rms pixel error of corresponding pixels. It shows that there is an approximated linear relation between the perturbation sigma and the corresponding mean- and rms distance.

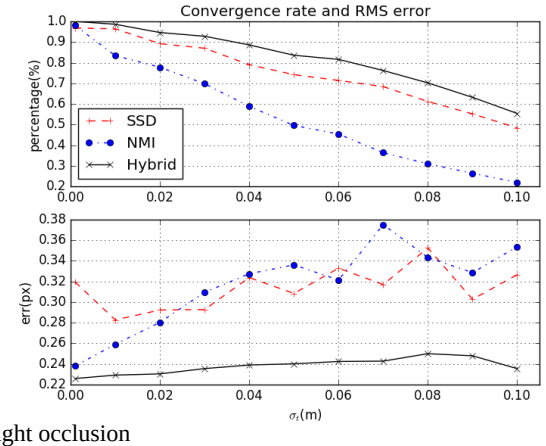
Then with the initial positioning errors in Fig. 2, we perform the alignment experiments under different conditions, and employ (27) to analyze the convergence rate (the optimization is considered to be convergent if the criterion is blow 0.5px) and estimation accuracy. For each method in each condition with each perturbation, the alignment experiment is repeated 500 times and the experimental results are presented in the following Fig. 3.



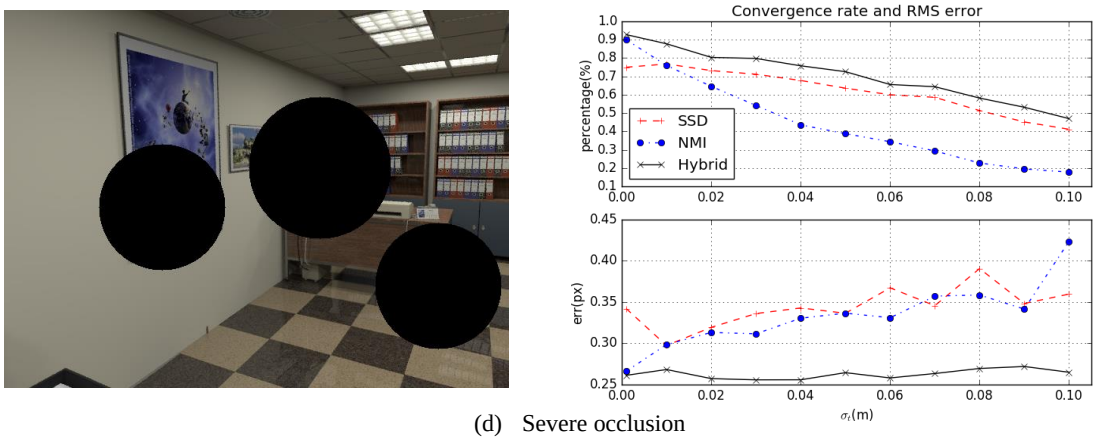
(a) Nominal condition



(b) Illumination changes



(c) Slight occlusion



(d) Severe occlusion

Figure 3. Different current images, and corresponding convergence rates and rms errors (a) Nominal condition. (b) Illumination changes, with strong local changes and relatively slight global changes. (c) Slight occlusion, with only one artificial patch added to the image. (d) Severe occlusion, with multiple patches added to the image.

From the results, we can see that in the nominal condition, the re-weighted SSD and our hybrid approach perform almost the same in terms of the convergence rate and the final residual error, both better than the NMI approach. When encountering the illumination changes (specifically refers to strong local changes and slight global changes in our experiments. Strong global changes are not considered, since it is quite rare in visual odometry), the SSD method can still maintain a relatively high convergence percentage in spite of suffering an apparent deterioration in performance, and the NMI method outperforms the SSD method in the alignment accuracy, but with a much lower convergence rate. The hybrid strategy is capable of achieving a better precision while keeping a higher convergence rate. This is pretty the same for the occlusions, where it is worth nothing that the NMI approach performs better than SSD approach when starting with smaller perturbations, and the hybrid approach always the best.

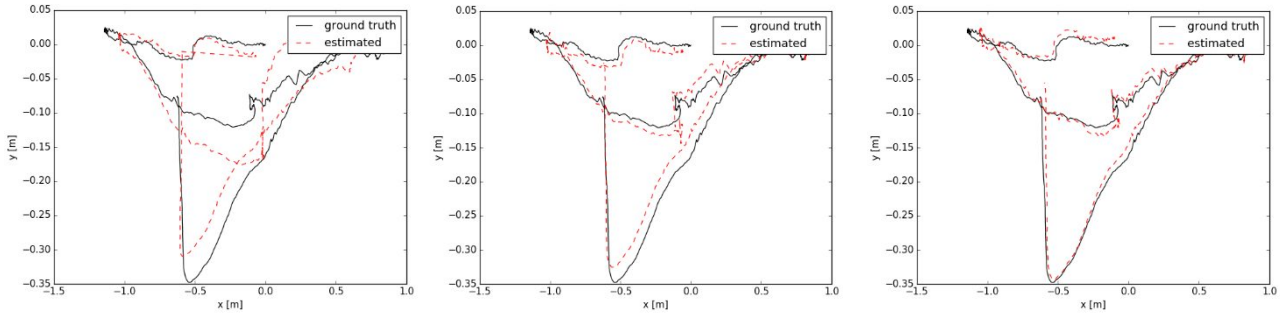
4.2 Experiments on benchmark datasets

This section will test our hybrid approach on some popular reference datasets, and we also provide the results of the re-weighted SSD approach, and the TUM DVO project [23] as a reference (TUM-DVO is essentially a complete SLAM project with loop detection and optimization modules, and in this paper we only present its tracking result before the final optimization for comparison). In our experiments, the tracking process only using NMI method gets stuck in incorrect local extreme quite easily, and further suffers unrecoverable tracking failures. Therefore, the tracking results of using NMI metric alone are not presented in the following part.

A. Synthetic datasets

Firstly, the algorithms are tested on a synthetic office room scene in ETH-ICL datasets [12]. In this dataset, the authors add random noise to the original depth data and simulate several illumination changes (we focus on the local changes) in the images. It also provides the camera ground-truth trajectory that can be used for evaluation. Fig. 4 presents some images of this scene and the estimated trajectories are shown in Fig. 5.

Figure 4. Images of the office room scene with illumination changes



(a) TUM-DVO (b) Re-weighted SSD (c) SSD + NMI

Figure 5. Estimated trajectories on the noisy dataset from TUM-DVO, SSD and SSD+NMI.

Similar to [15], we use the absolute trajectory error (ATE) to quantify the accuracy of an estimated trajectory. Three standard statistics are computed for evaluating the ATE: RMSE, Mean and Median. The ATE statistics of these four trajectories are listed in Table 1.

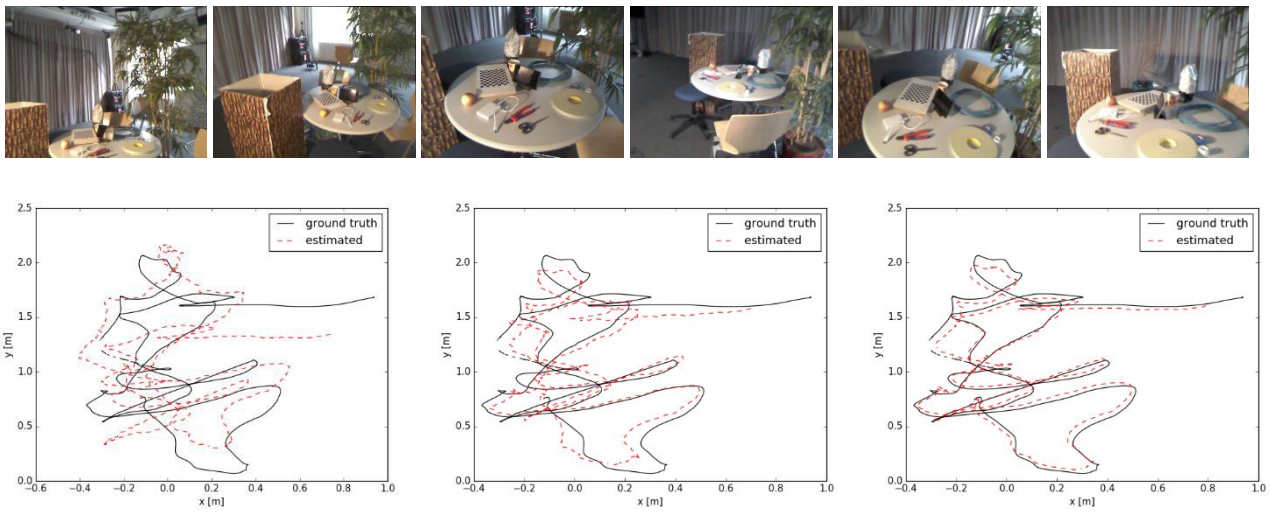
Table 1. Absolute trajectory error (ATE) for trajectories on the office room scene.

Statistics (m)	TUM-DVO	SSD	SSD+NMI
RMSE	0.119961	0.093862	**0.082909**
Mean	0.101134	0.089770	**0.075346**
Median	0.109709	0.078844	**0.065110**

Obviously, the proposed method achieves the highest tracking accuracy when encountering noise-contaminated depth data and illumination changes in scenes. It is worth nothing that the TUM-DVO method used here actually cannot be regarded as a pure vision-based tracking method, because except photometric residuals it also takes depth residuals directly into its optimization frame. Since photometric residuals and depth residuals are both affected severely in this scene, it thus suffers the largest degradation in tracking performance.

B. Real sensor datasets

In practice, the actual noise, occlusion and illumination changes from the real sensor are usually much more complex than that in synthetic data. Therefore, we continually evaluate our method on some real world datasets.



(a) TUM-DVO (b) Re-weighted SSD (c) SSD + NMI

Figure 6. Estimated trajectories on the real dataset with illumination variations from TUM-DVO, SSD and SSD+NMI.

The authors of ETH-ICL dataset also provides real RGB-D sequences with illumination variations acquired by a 1st-generation Kinect, which is used for our evaluation firstly. We still consider the sequence with local illumination changes that more likely occur in the context of visual odometry. Similar to experiments on synthetic data, several images of this scene and estimated trajectories are shown in Fig. 6. The ATE statistics of each method are listed in Table 2. The results are basically consistent with that on synthetic data, showing that our hybrid tracking strategy is more capable of coping with gradual lighting changes in the scenes.

Table 2. Absolute trajectory error (ATE) for trajectories on the office room scene.

Statistics (m)	TUM-DVO	SSD	SSD+NMI
RMSE	0.244890	0.094815	**0.053491**
Mean	0.233010	0.089577	**0.049016**
Median	0.243662	0.082388	**0.044572**

Then we also evaluate on several sequences from the commonly-used TUM RGB-D dataset [16] to allow comparisons between our results and other published works, although this dataset do not exhibit much illumination variation. From this dataset we choose four sequences: fr1_xyz, fr1_desk, fr2_desk and fr3_long_office, to test our method.

Table 3. Absolute trajectory error (ATE) for sequences in TUM RGB-D dataset.

Approach	Statistics (m)	fr1_xyz	fr1_desk	fr2_desk	fr3_long_office
TUM-DVO	RMSE	**0.030732**	0.168639	0.124842	0.132312
	Mean	**0.027709**	0.140536	0.115781	0.111144
	Median	**0.025783**	0.118550	0.095214	0.098825
SSD	RMSE	0.046604	0.076606	0.097196	0.097152
	Mean	0.042353	0.070361	0.096168	0.079650
	Median	0.039532	0.071350	0.095036	0.055902
SSD+NMI	RMSE	0.034698	**0.072053**	**0.092882**	**0.042874**
	Mean	0.032768	**0.066530**	**0.092015**	**0.037013**
	Median	0.032279	**0.068276**	**0.092327**	**0.031498**

We show the respective ATE statistics in Table 3, and note that there exists obvious partial occlusion (caused by a wooden board placed between two desks) in the fr3_long_office. From Table 1-3, we can conclude that the proposed method performs only slightly better than SSD based method on nominal scenes (such as fr1_xyz and fr1_desk, with little variations), and demonstrates a distinct superiority when encountering the appearance variations from occlusion and illumination changes.

5. CONCLUSION

This paper presented a robust and accurate visual tracking scheme for direct visual odometry, which utilizes the robustness of NMI similarity metric and the inherent wide convergence basin of SSD metric. A full derivation of first- and second-order analytical NMI derivatives are presented so that it can be used with the LM optimization method. Then a novel hybrid tracking strategy has been proposed that preserves the strengths of NMI with respect to occlusions and illumination variations, and simultaneously increase the convergence rate by taking advantage of the superior optimization characteristics of SSD. The proposed method has been evaluated on both synthetic and real sensor datasets. The experimental results have verified its robustness and accuracy, which shows an apparent advantage compared with classical approaches.

REFERENCES

[1] Odobez, J.-M. and Bouthemy, P., "Robust multiresolution estimation of parametric motion models," Journal of visual communication and image representation 6(4), 348-365 (1995).

[2] Hager, G. D. and Belhumeur, P. N., "Efficient region tracking with parametric models of geometry and illumination," IEEE transactions on pattern analysis and machine intelligence 20(10), 1025-1039 (1998).

[3] Kerl, C., Sturm, J. and Cremers, D., "Robust odometry estimation for rgb-d cameras," Proc. Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, 3748-3754 (2013).

[4] Silveira, G. and Malis, E., "Real-time visual tracking under arbitrary illumination changes," Proc. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 1-6 (2007).

[5] Pluim, J. P., Maintz, J. A. and Viergever, M. A., "Mutual-information-based registration of medical images: A survey," IEEE transactions on medical imaging 22(8), 986-1004 (2003).

[6] Studholme, C., Hill, D. L. and Hawkes, D. J., "Automated 3d registration of truncated mr and ct images of the head," Proc. BMVC, Citeseer, 95, 27-36 (1995).

[7] Studholme, C., Hill, D. L. and Hawkes, D. J., "An overlap invariant entropy measure of 3d medical image alignment," Pattern recognition 32(1), 71-86 (1999).

[8] Okker, B., Yan, C., Zhang, J., Ong, S. and Teoh, S., "Accurate and fully automatic 3d registration of spinal images using normalized mutual information," Proc. Biomedical Circuits and Systems, 2004 IEEE International Workshop on, IEEE, S3/1-S5 (2004).

[9] Dowson, N. and Bowden, R., "A unifying framework for mutual information methods for use in non-linear optimisation," Proc. European Conference on Computer Vision, Springer, 365-378 (2006).

[10] Thévenaz, P. and Unser, M., "Optimization of mutual information for multiresolution image registration," IEEE transactions on image processing 9(12), 2083-2099 (2000).

[11] Dame, A. and Marchand, E., "Second-order optimization of mutual information for real-time image registration," IEEE Transactions on Image Processing 21(9), 4190-4203 (2012).

[12] Park, S., Schöps, T. and Pollefeys, M., "Illumination change robustness in direct visual slam," Proc. Robotics and Automation (ICRA), 2017 IEEE International Conference on, IEEE, 4523-4530 (2017).

[13] Xu, R., Chen, Y.-W., Tang, S.-Y., Morikawa, S. and Kurumi, Y., "Parzen-window based normalized mutual information for medical image registration," IEICE transactions on information and systems 91(1), 132-144 (2008).

[14] Pascoe, G., Maddern, W. P. and Newman, P., "Robust direct visual localisation using normalised information distance," Proc. BMVC, 70.71-70.13 (2015).

[15] Handa, A., Whelan, T., McDonald, J. and Davison, A. J., "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," Proc. Robotics and automation (ICRA), 2014 IEEE international conference on, IEEE, 1524-1531 (2014).

[16] Sturm, J., Engelhard, N., Endres, F., Burgard, W. and Cremers, D., "A benchmark for the evaluation of rgb-d slam systems," Proc. Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, IEEE, 573-580 (2012).

[17] Engel, J., Schöps, T. and Cremers, D., "Lsd-slam: Large-scale direct monocular slam," Proc. European Conference on Computer Vision, Springer, 834-849 (2014).

[18] Engel, J., Koltun, V. and Cremers, D., "Direct sparse odometry," IEEE transactions on pattern analysis and machine intelligence 40(3), 611-625 (2018).

[19] Baker, S. and Matthews, I., "Lucas-kanade 20 years on: A unifying framework," International journal of computer vision 56(3), 221-255 (2004).

[20] Shannon, C. E., "A mathematical theory of communication," ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 3-55 (2001).

[21] Fraissinet-Tachet, M., Schmitt, M., Wen, Z. and Kuijper, A., "Multi-camera piecewise planar object tracking with mutual information," Journal of Mathematical Imaging and Vision 56(3), 591-602 (2016).

[22] Lange, K. L., Little, R. J. and Taylor, J. M., "Robust statistical modeling using the t distribution," Journal of the American Statistical Association 84(408), 881-896 (1989).

[23] Kerl, C., Sturm, J. and Cremers, D., "Dense visual slam for rgb-d cameras," Proc. Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, IEEE, 2100-2106 (2013).