Real-time dense appearance-based SLAM for RGB-D sensors

Cédric Audras and **Andrew I. Comport**
I3S-CNRS, University of Nice Sophia-Antipolis, France.
name.surname@i3s.unice.fr

Maxime Meilland and **Patrick Rives**
INRIA, Sophia-Antipolis, France.
name.surname@sophia.inria.fr

Abstract

In this work a direct dense approach is proposed for real-time RGB-D localisation and tracking. The direct RDB-D localisation approach is demonstrated on a low cost sensor which exploits projective IR light within indoor environments. This type of device has recently been the object of much interest and one advantage is that it provides dense 3D environment maps in real-time via embedded computation. To date all existing tracking approaches using these sensors have been based on a sparse set of geometric features which have to be detected and matched across images in time. These techniques have a downfall in that errors made in matching and extraction cannot be recovered through a purely geometric optimization model. Here a direct 3D tracking approach is proposed such that an error is based directly on the intensity of pixels so that no feature extraction and matching are required. Original work in this direction has been carried out on direct dense tracking for stereo cameras and it is shown here how this approach is ideal for the general class of RGB-D cameras. Some advantages of an IR depth sensor include more accurate maps in untextured environments, however, the sensor is limited to indoor environments and close ranges. The system is made robust via statistical M-estimation along with a saliency selection criterion which favors more interesting pixels. Experimental validation is provided which compares the proposed RGB-D localisation and mapping system with ground truth laser odometry within complex indoor environments.

1 Introduction

The problem of real-time localisation and mapping within complex indoor environments is a challenging problem for a wide range of applications ranging from robotics to augmented reality. It is well known that good localisation requires an accurate and robust map of the environment. In this paper the core issue of 3D visual odometry is considered in the context of a mobile agent bearing a RGB-D camera (colour + depth) which observes a real image stream and who is navigating within large scale buildings which contain moving pedestrians and other types of occluding information.

Simple RGB systems, without depth, have been used extensively for tracking 3D poses in complex indoor environments, however, estimating and extracting the 3D structure and photometric information from the environment is a non-trivial problem. Due to this difficulty, researchers have often resorted to model-based techniques [Drummond and Cipolla, 2002; Comport *et al.*, 2006b] which circumvent the problem by providing a 3D model of the environment *a priori*. 3D Computer Aided Design (CAD) models have shown to be essential for robust, accurate and efficient 3D motion estimation, however, they still need to provide an *a priori* model which is not always available or extremely difficult to obtain as in the case of shapeless objects or large-scale environments. Other approaches have estimated the structure online over time via visual simultaneous localisation and mapping approaches [Davison and Murray, 2002; Chiuso *et al.*, 2002]. In [Mouragnon *et al.*, 2006], a similar monocular technique is proposed but drift is minimised using a local bundle adjustment technique. Recently, dense real-time monocular techniques have begun to appear [Newcombe and Davison, 2010], however, they are quite limited in the quality of depth information and are sensitive to dynamic objects.

One of the biggest interests in RGB-D systems is that they provide a continuous measurement of the 3D structure within the environment. Various types of RGB-D systems have been around for some time and they are often used for performing Self Localisation And Mapping (SLAM) of robots. Renewed interest has, however, been recently ignited with the release of low cost dense RGB-D sensors such as the Microsoft Kinect in November 2010 and the Asus Xtion Pro in February 2011 which are both based on PrimeSense technology [Freedman *et al.*, 2010]. Whilst the objective here is not to give a survey, it can be noted that the various types of RGB-D sensors can be classed into passive and active devices, primarily ac-

Stereo [Comport *et al.*, 2010]

Advantages	Disadvantages
- Passive - Outdoor and indoor - Medium range	- Dependant on texture - Expensive computation

IR structured light [Freedman *et al.*, 2010]

Advantages	Disadvantages
- Known pattern - Semi-continuous (texture-less surfaces)	- Active system - Indoor or nocturne - Close range

Laser [Gallegos *et al.*, 2010; Pitzer *et al.*, 2010; Karg *et al.*, 2010]

Advantages	Disadvantages
- Higher precision - Long range - Outdoor and Indoor	- Active system - High power consumption - Slow acquisition rates

Table 1: A summary of the main differences between **dense** stereo, laser and projective-light systems

cording to the way in which depth is perceived (see Table 1). Stereo vision systems which provide dense matching can be considered as passive systems while active systems include cameras combined with LIDAR, projection of structured light patterns (infra-red (IR) or visible light), sonar, etc.

RGB-D stereo techniques are probably the most developed and wide-spread approaches to performing visual SLAM and it is therefore unavoidable to consider them when developing a model for a structured light type sensor. In [Nistér *et al.*, 2004] stereo and monocular visual odometry approaches are proposed based on a combination of feature extraction, matching, tracking, triangulation, RANSAC pose estimation and iterative refinement. This type of technique has formed the basis for many subsequent approaches. Another stereo approach was proposed originally in [Comport *et al.*, 2007] performs 3D visual odometry based on a direct appearance based model that uses the *dense set* of information provided in a stereo pair. This technique is able to accurately handle large scale scenes efficiently and robustly whilst avoiding error prone feature extraction and matching. Recently this approach has been extended to integrate structure information over time [Tykkala and Comport, 2011a].

Amongst the various techniques used in projective-light RGB-D systems [Freedman *et al.*, 2010], only feature based approaches have been considered. Unfortunately, feature based methods (e.g. [Henry *et al.*, 2010; Engelhard *et al.*, 2011; Sturm *et al.*, 2010]) all 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). Since the global estimation loop is never closed on the image measurements (intensities) these multi-step techniques systematically propagate feature extraction and matching error and accumulate drift. To eliminate drift these approaches resort to techniques such as local bundle adjustment.

On the other hand, appearance and optical flow based techniques are image-based and minimise an error directly based on the image measurements. Unfortunately, they struggle to solve difficulties related to the fact they don't have depth information (i.e. monocular) and therefore make heavy assumptions about the nature of the structure within the scene or the camera model. For example in [Hager and Belhumeur, 1998] an affine camera model is assumed and in [Baker and Matthews, 2001] and [Benhimane and Malis, 2004] planar homography models are assumed. In this way the perspective effects or the effects of non-planar 3D objects are not considered and tracking fails easily under large movements. Of course many papers avoid the problems of monocular algorithms (i.e. scale factor, initialisation, observability, etc.) by using multi-view constraints and a multitude of work exist on multi-view geometry (see [Hartley and Zisserman, 2001] and ref. therein). However, to our knowledge only a limited amount of research has been carried out to derive an efficient dense tracker as in [Comport *et al.*, 2007; Meilland *et al.*, 2010; Tykkala and Comport, 2011a].

The objective of this paper will therefore be to use a recent dense stereo RGB-D SLAM technique [Comport *et al.*, 2010] and adapt it to a low cost Kinect RDB-D sensor. As mentioned previously, all work carried out with the Kinect or similar sensors has used an Iterative Closest Point technique based on extracting features and requiring matching between time instants as in [Henry *et al.*, 2010; Engelhard *et al.*, 2011; Sturm *et al.*, 2010]. The new approach proposed here will allow to avoid feature extraction and matching via an appearance-based approach which is not purely geometric and is well suited to this type of sensor.

In Section 2 an overview of the objective function is given. Section 3.2 the RGB-D warping function is detailed. Section 3.3 outlines the robust non-linear minimisation technique and in 5 the results are presented.

2 Visual Odometry

A framework is described for estimating the trajectory of a RGB-D sensor along a sequence from a designated region within the image. The tracking problem will essentially be considered as a pose estimation problem which will be related directly to the grey-level brightness measurements and depth measurements within the RGB-D sensor via a non-linear model which accounts for the 3D geometric configuration of the scene.

Consider a RGB-D sensor with a colour 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 $\mathbf{I} \in \mathbb{R}^{\mathrm{n}}$ and $\mathbf{D} \in \mathbb{R}^{\mathrm{n}}$. Consider now an RGB-D image, denoted also an *augmented image* [Meilland *et al.*, 2010], to be the set containing both brightness and depth $\boldsymbol{\mathcal{I}} = \{\mathbf{I}, \mathbf{D}\}$. $\boldsymbol{\mathcal{P}}^* = \{\mathbf{p}, \mathbf{D}\} \in \mathbb{R}^{3\times\mathrm{n}}$ are then the 3D points 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. Any set of pixels from the reference image are considered as a reference template, denoted by $\mathcal{R}^* = \{\mathbf{p}_1^*, \mathbf{p}_2^*, \ldots, \mathbf{p}_{\mathrm{n}}^*\}$ where n is the number of points selected in the reference image.

The motion of the camera or objects within the scene induces a deformation of the reference template. The 3D geometric deformation of a structured light RGB-D camera [Freedman *et al.*, 2010] can be fully defined by a motion model $w(\boldsymbol{\mathcal{P}}^*, \mathbf{T}_{IR}, \mathbf{T}_{PR}, \mathbf{K}, \mathbf{K}_{IR}, \boldsymbol{\mathcal{P}}^*_{PR}; \overline{\mathbf{T}}(t))$. The motion model w considered in this paper is the 3D warping function which will be detailed further in Section 3. $\mathbf{K}$ contains the intrinsic calibration parameters for the colour camera, $\mathbf{K}_{IR}$ the intrinsic calibration parameters for the IR camera, $\boldsymbol{\mathcal{P}}^*_{PR}$ are 3D reference points for the projective-light system which have been memorised at a known depth for triangulating a depth map online (see Figure 1 for more detail). $\mathbf{T}_{IR} \in \mathbb{SE}(3)$ is the homogeneous matrix of the extrinsic calibration parameters of the IR camera, $\mathbf{T}_{PR} \in \mathbb{SE}(3)$ are the extrinsic calibration parameters of the projector and $\overline{\mathbf{T}} = (\bar{\mathbf{R}}, \bar{\mathbf{t}}) \in \mathbb{SE}(3)$ is the current pose of the RGB-D camera 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. Since the intrinsic and extrinsic calibration parameters, along with and projective-light reference model, do not vary with time they will be assumed implicit in the following.

It follows that the reference image is obtained by warping the current image as:

$$\boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*) = \boldsymbol{\mathcal{I}}\big(w\big(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}}\big), t\big), \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*. \tag{1}$$

where $\overline{\mathbf{T}}$ is the true pose.

Suppose that at the current image an estimate of the pose $\widehat{\mathbf{T}}$ fully represents the pose of the RGB-D camera with respect to an augmented reference image. The tracking problem then becomes one of estimating the incremental pose $\mathbf{T}(\mathbf{x})$, where it is supposed that $\exists \tilde{\mathbf{x}} : \mathbf{T}(\tilde{\mathbf{x}})\widehat{\mathbf{T}} = \overline{\mathbf{T}}$. The estimate is updated by a homogeneous transformation $\widehat{\mathbf{T}} \leftarrow \mathbf{T}(\mathbf{x})\widehat{\mathbf{T}}$.

The unknown parameters $\mathbf{x} \in \mathbb{R}^6$ are defined as:

$$\mathbf{x} = \int_0^1 (\boldsymbol{\omega}, \boldsymbol{\upsilon}) dt \in se(3), \tag{2}$$

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 and the trajectory of the camera can be estimated by minimising a non-linear least squares cost function:

$$C(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \left(\boldsymbol{\mathcal{I}}\left(w\left(\boldsymbol{\mathcal{P}}^*; \mathbf{T}(\mathbf{x})\widehat{\mathbf{T}} \right) \right) - \boldsymbol{\mathcal{I}}^*\left(\boldsymbol{\mathcal{P}}^* \right) \right)^2. \tag{3}$$

This function is minimised using the robust, efficient and precise non-linear minimisation procedure detailed in Section 3.3. It should be noted that we have also proposed a technique for Direct Iterative Closest Point pose estimation in [Tykkälä and Comport, 2011b]. This approach simultaneously minimises both the intensity and depth error and leads to a better conditioned estimation process. Even so, this approach involves an increase in computational cost which is difficult to justify when the indoor sequences have sufficient precision using only the intensity error and as such we do not present it in this paper. Nevertheless, depth error should definitely be considered in degenerate cases such as when there is no intensity gradient (e.g. the case when the sensor sees a homogeneously textured white wall).

3 Novel view synthesis and warping

The geometric configuration of the RGB-D camera is based on the paradigm that a view of a scene, augmented with depth, satisfies rigid geometric constraints. Thus given an augmented reference image another view can be generated by means of a warping function. This warping function subsequently provides the required relationship between two RGB-D cameras viewing a rigid scene from two different positions (at two time instants).

3.1 Projective Structured Light Geometry

Consider now Figure 1 where the RGB-D projective-light device [Freedman *et al.*, 2010] consists of an IR projector, on the right, an IR camera in the middle, and a RGB camera on the left. The dashed boxes represent a 3D reference plane at a known position onto which the projection pattern has been memorized via an *a priori* learning phase. The device uses an image of this pattern to perform correlation based multi-resolution matching with the image of the pattern projected onto the scene. A quasi periodic pattern is used which has a known spatial frequency spectrum with distinct peaks in order to reduce the effects of ambient light and noise in the correlation phase. The difference between the matched points provides a disparity d (as in a stereo system) which can be used to perform triangulation and determine the depth of the 3D scene point. If the projector and camera are rectified the depth is computed as:

$$Z = \frac{t_x f}{d}, \tag{4}$$

where Z is the depth, t_x is the baseline between the IR camera and the projector, f is the focal length of the camera, and d the disparity.

According to [Freedman *et al.*, 2010], the device also benefits from micro-lenses with varying focal lengths so that the pattern varies with distance from the device.

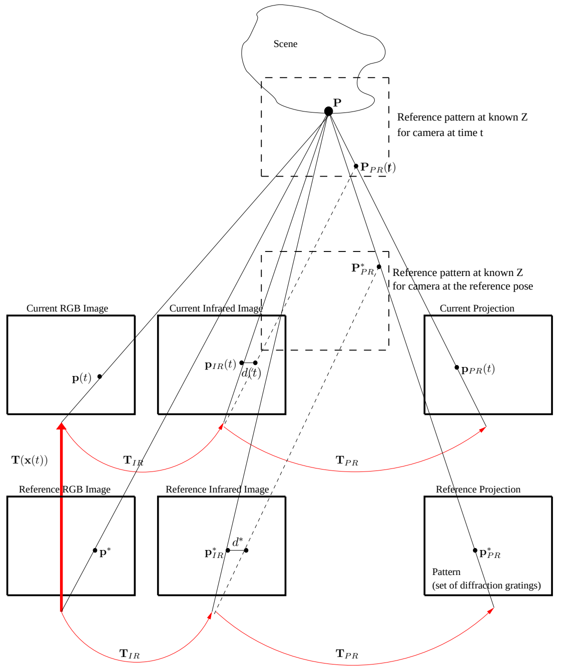


Figure 1: The geometry of a projective-light system at two subsequent time instants. The projection of the point $\mathbf{p}_{PR}$ in the image IR at $\mathbf{p}_{IR}$ is used to perform matching with respect to a known reference pattern which has been memorised at a known depth. Triangulation is then performed between the reference pattern and the IR image to obtain a depth map in the RGB image. The unknown pose $\mathbf{T}(\mathbf{x}(t))$ is estimated via a non-linear warping function which warps all points in the reference RGB-D image $\mathbf{p}^*$ to the current $\mathbf{p}(t)$. The point warping function is chosen to correspond to a spherical projection model. The extrinsic parameters $\mathbf{T}_{IR}$ and $\mathbf{T}_{PR}$ are assumed calibrated and known *a priori*.

In the context of this work the IR and RGB cameras are considered to form a stereo pair which can be calibrated as outlined in Section 5.1. For the purpose of the following developments, it is therefore assumed that the device is calibrated with intrinsic camera parameters $\mathbf{K}$ and $\mathbf{K}_{IR}$ for the RGB and IR cameras respectively and extrinsic parameters $\mathbf{T}_{IR}$ and $\mathbf{T}_{PR}$ denoting the pose from the RGB camera to the IR camera and from the IR camera to the projector respectively. The main objective of this paper is therefore to estimate the unknown motion $\mathbf{T}(\mathbf{x})$ of the sensor with respect to a rigid scene (In Figure 1 only one world point $\mathbf{P}$ is considered, however the scene is modelled by a dense point cloud).

3.2 Augmented image warping

To warp current image intensities to the reference image, reference world points $\boldsymbol{\mathcal{P}}^*$ have to be projected onto the current RGB camera plane with a 3×4 perspective projection matrix $\mathbf{M} = \boldsymbol{K}[\mathbf{R}|\mathbf{t}] \in \mathbb{R}^{3\times3} \times \mathbb{SE}(3)$. The homogeneous vector $\mathbf{p} = (u, v, 1)^T \in \mathbb{RP}(3)$ in pixel coordinates is given by $\mathbf{p} = \mathbf{MP}$. Then current image $\boldsymbol{\mathcal{I}}$ grey levels are interpolated at points $\mathbf{p}$ to obtain intensities corresponding to the reference image. The RGB-D warping function $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}})$ from (3) can now be considered to be composed of a 3D point cloud.

It can be highlighted that the warping operator $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}})$ is a *group action* which allows to exploit the so-called inverse compositional approach. Indeed, the following operations hold:

1. The identity map:

$$w(\boldsymbol{\mathcal{P}}^*; \mathbf{I}) = \boldsymbol{\mathcal{P}}^*, \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4, \tag{5}$$

2. The composition of an action corresponds to the action of a composition $\forall \mathbf{T_1}, \mathbf{T_2} \in \mathbb{SE}(3)$:

$$w(w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}), \mathbf{T_2}) = w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}\mathbf{T_2}) \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4. \tag{6}$$

3.3 Minimization

The aim is to minimize the objective criteria (3) in an accurate, robust and efficient manner. Since this is a non-linear function of the unknown pose parameters, an iterative minimization procedure can be used. The cost function is minimized by $\nabla\mathcal{C}(\mathbf{x})|_{\mathbf{x}=\tilde{\mathbf{x}}} = \mathbf{0}$, where ∇ is the gradient operator with respect to the unknown $\mathbf{x}$ defined in equation (2) assuming a global minimum can be reached in $\mathbf{x} = \tilde{\mathbf{x}}$.

In case of using a first order approximation, the Jacobian can be decomposed into modular parts as:

$$\mathbf{J}(\mathbf{x})|_{\mathbf{x}=\tilde{\mathbf{x}}} = \mathbf{J}_{\boldsymbol{\mathcal{I}}^*}\mathbf{J}_w\mathbf{J}_{\mathbf{T}}. \tag{7}$$

Where $\mathbf{J}_{\boldsymbol{\mathcal{I}}^*}$ is the reference image gradient with respect to pixel coordinates of dimension $n \times 2n$, $\mathbf{J}_w$ is the derivative of perspective projection of dimension $2n \times 3n$, and $\mathbf{J}_{\mathbf{T}}$ depends on the parameterization of $\mathbf{x}$ with a dimension of $3n \times 6$ from Equation (2).

The vector of unknown parameters $\mathbf{x}$ is obtained iteratively from:

$$\mathbf{x} = -\lambda(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{I}}^*), \tag{8}$$

where $(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T$ is the pseudo-inverse $\mathbf{J}^+$ of the matrix $\mathbf{J}$ and λ is the tuning gain which ensures an exponential decrease of the error.

3.4 Robust Estimation

While localising the sensor indoors, the environment can vary between the reference and the current images due to moving objects, pedestrians, lighting changes or self occlusions of corridors perceived from different viewpoints. To deal with these changes, a robust M-estimator is used and included in the objective function (3) which becomes:

$$\mathcal{O}(\mathbf{x}) = \rho\left(\sum_{\boldsymbol{\mathcal{P}}^{*}\in\mathcal{R}^{*}} \boldsymbol{\mathcal{I}}\left(w(\boldsymbol{\mathcal{P}}^{*}, \mathbf{T}(\mathbf{x})\widehat{\mathbf{T}})\right) - \boldsymbol{\mathcal{I}}^{*}(\boldsymbol{\mathcal{P}}^{*})\right). \tag{9}$$

In this case (8) becomes $\mathbf{x} = -\lambda(\mathbf{DJ})^{+}\mathbf{D}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{I}}^{*})$ with $\mathbf{D}$ the diagonal weighting matrix computed via a robust weighting function [Huber, 1996].

3.5 Multi-Resolution Augmented Reference Image

To improve the computational time and the convergence domain, a multi-resolution approach is considered. Each RGB-D image is under-sampled N times (depending on the original image size) by a factor of 2. The minimization begins at the lowest resolution and the result is used to initialize the next level repeatedly until the highest resolution is reached. In this way, larger displacements are minimized at low cost on smaller images. Under-sampling produces smoothed images so strong local edges, which provide more fine tune accuracy in alignment are only used in higher resolutions, when the current estimate is close to the solution. Local minima can also be avoided in this way.

4 Information Selection

The essence of appearance-based methods is to minimize pixel intensities directly between images. RGB images are, however, clearly redundant, i.e. a lot of information is not overly important for localisation, such as completely homogeneous zones. This kind of information can be omitted so as to reduce the dimension, as in feature-based techniques, so as to favour real-time computing. Nevertheless, it will be shown here that pixel selection (instead of feature extraction) means that matching is not required.

A classic approach is to select only the best corners or edges (intensity gradients) in grey-level images, by using a feature detector [Harris and Stephens, 1988]. This naïve approach does not consider the importance of the structure of the scene and can lead to the selection of non-observable measurements. The novelty of the approach proposed in [Meilland *et al.*, 2010] is to quantify the effect of the geometric structure and the image intensity measurements by analysing directly the entire analytical Jacobian which relates scene movement to sensor movement. The aim being to select points which best condition the six degrees of freedom of the vehicle. Indeed, the Jacobian directly combines grey level gradient (image derivatives) and geometric gradient.

More precisely, the reference Jacobian matrix of the perspective projection in equation (7), can be decomposed into six parts corresponding to each degree of freedom of the sensor with: $\mathbf{J} = \{\mathbf{J^1}, \mathbf{J^2}, \mathbf{J^3}, \mathbf{J}^4, \mathbf{J^5}, \mathbf{J^6}\}$. Each column $\mathbf{J^j}$ can be interpreted as a saliency map (see [Meilland *et al.*, 2010]).

A subset of the entire set of pixels :

$$\overline{\mathbf{J}} = \{\overline{\mathbf{J}}^1, \overline{\mathbf{J}}^2, \overline{\mathbf{J}}^3, \overline{\mathbf{J}}^4, \overline{\mathbf{J}}^5, \overline{\mathbf{J}}^6\} \subset \mathbf{J},$$

is sought such that $\overline{\mathbf{J}}$ is the reduced $p \times 6$ version of $\mathbf{J}$ of dimension $n \times 6$ with $p \ll n$, where n is the number of points in the full image, and the rows of $\overline{\mathbf{J}}$ are given by:

$$\overline{\mathbf{J}}_k = \underset{j}{\operatorname{argmax}}(\mathbf{J^j_i} \setminus \tilde{\mathbf{J}}), \tag{10}$$

which this corresponds to selecting an entire line i of $\mathbf{J}$ according to the maximum gradient of one column (direction) j and where k is the next line that is added to $\overline{\mathbf{J}}$. $\tilde{\mathbf{J}} \subset \overline{\mathbf{J}}$ is a intermediary subset of Jacobian that is sought. $\setminus\tilde{\mathbf{J}}$ indicates that it is not possible to reselect the same row. i.e. the lines of $\mathbf{J}_j$ are recursively chosen and inserted into $\tilde{\mathbf{J}}$ until the required number of lines have been selected, in which case it becomes $\overline{\mathbf{J}}$:

$$\overline{\mathbf{J}} = \left[\overline{\mathbf{J}}^{\mathbf{1}}, \ldots, \overline{\mathbf{J}}^{\mathbf{6}}\right]^{\top}. \tag{11}$$

In the recursive selection process, the criteria (10) is applied iteratively in each direction such that an equal number of maximum gradients are chosen for each degree of freedom. In this way the pixels that have been chosen at the end of this algorithm are those that best condition each dof in the pose estimation process.

5 Results

The mobile robot experimental setup is shown in Figure 2 configured with both the Kinect platform and a laser range finder for ground truth. The 3 wheeled robot used for the experiment is the Neobotix mobile platform (MP-S500) equipped with a rotary Sick LD-LRS1000 laser. The acquisition frequency of the laser varies between 5 and 10 Hz. The Kinect sensor has been placed directly above the laser to minimise the extrinsic calibration difference between both systems. The Kinect tracking system is capable of acquiring images at a frame rate of 30Hz and visual odometry is computed at frame-rate on a Dell Latitude e6500 with core 2 duo P6500, 2.5GHz with Fedora 10.

The visual localisation and mapping algorithm can be summarised by the following steps:

1. Take the initial RGB-D image as the world frame.
2. Warp current image using (1), compute the error between the reference and current warped images and the Jacobian as in (7).
3. Estimate the incremental pose with respect to the reference frame according to (9) and integrate the incremental pose into $\widehat{\mathbf{T}} \leftarrow \mathbf{T}(\mathbf{x})\widehat{\mathbf{T}}$.
4. Convergence is checked by comparing the Median Absolute Deviation (MAD) of the error to a predefined threshold. If not iterate estimation to 1.

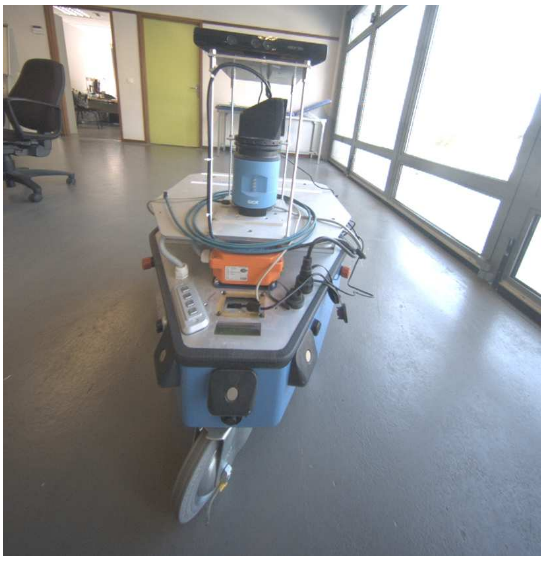
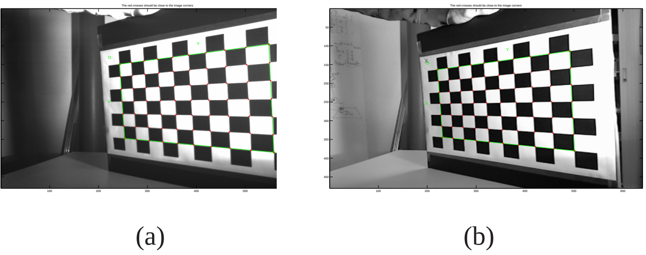


Figure 2: Mobile robot experimental setup with Kinect and laser.



(a) (b)

Figure 3: Extrinsic and intrinsic calibration of the IR (a) and RGB (b) image pair.

5. Take new reference image if the MAD of the error larger than a threshold. Otherwise repeat to 2.
6. Take a new reference RGB-D frame and update the global trajectory using $\widehat{\mathbf{T}}$.

5.1 Calibration

In order to obtain better precision and subsequently avoid error during warping, the Kinect had to be calibrated with the Matlab toolbox [Bouguet, 2010]. First at all, the RGB camera was calibrated using a planar calibration pattern which allows to estimate the intrinsic parameters with distortion parameters. After that, to calibrate the IR camera, the projector had to be hidden so that the pattern from the Kinect would not interfere with the corner detection for the calibration of the IR image. In order to get a better contrast between squares of the pattern, the target was heated with a halogen lamp. Finally, the IR and RGB cameras were calibrated as a stereo pair which allows to estimate the extrinsic parameters.

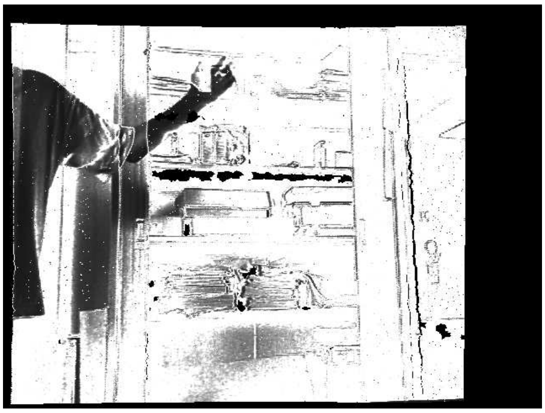


Figure 4: Robust outlier rejection: Image showing the outlier rejection weights. The darker the points, the less influence they have on the estimation process. In this image, it can be seen that a moving person has been rejected. It can also be noted that other outliers are detected in the image. These points generally correspond to illumination change (such as the shadow of the person), sensor noise or the self occlusion of the corners of the corridors.

5.2 Robust Estimation

A robust M-estimation technique (as detailed in [Comport *et al.*, 2006a]) was used to reject outliers not corresponding to the definition of the objective function. The use of robust techniques is very interesting in the case of a highly redundant set of measurement as is the case of a RGB-D sensor. The outliers generally correspond to people, occlusions, illumination changes, sensor noise or the self occlusion of the corners of the corridors.

In Figure 4, a robust weighting image is shown whereby dark pixels have low weight and black pixels are completely rejected. A moving person can be seen, in the left half of the image, to be rejected whilst the rigid parts of the scene, such as the cupboard in the background, were used to estimate the pose. The moving shadow of the person is also rejected. This type of information is useful in applications requiring the trajectory of moving obstacles.

5.3 Trajectory Estimation

In order to test the algorithm, a real sequence has been made in a complex indoor environment with a closed loop trajectory. Two sensors have been used to acquire measurements at the same time in order to provide a relative ground truth:

1. The ground truth system is a SICK laser which provides high precision range measurements especially in indoor environments. The reference is obtained by a 2D scan matching algorithm based on a representation of the scene named *latent map* which consists of a set of piecewise linear functions defined over a spatial grid. The idea of the algorithm is to align all the scan data with the latent map. The algorithm is first applied on several

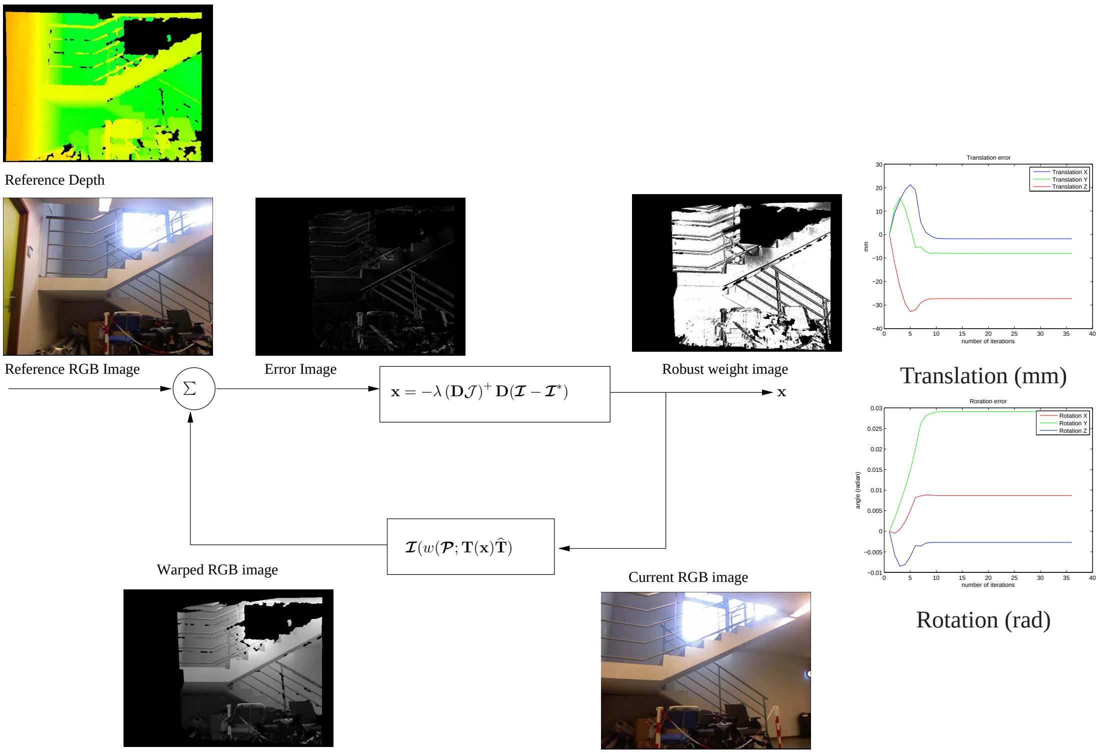


Figure 5: The iterative estimation process as given by equation (8). The augmented reference image is given as input. An error image is then obtained with the warped current image. The robust Jacobian term is inversed to obtain an incremental pose (parameterised as an element of the Lie algebra here). The current image is then re-warped and the process is repeated until convergence. The smooth convergence of the translation and rotation is shown here for a large movement without prediction and at full resolution.

small latent maps which are then combined. Finally, all the scan data are fused in a global map (see [Qi-Xing and Anguelov, 2010]). This provides much better results than classical pair-wise scan matching methods.

2. As already mentioned, the sensor of interest is a RGB-D Kinect sensor which is connected to a mobile computer for performing visual odometry computation. In Figure 5 the different stages involved in 3D tracking and mapping are shown. The augmented reference image $\mathcal{I}^{*}$ including both RGB and depth-map are given to the iterative non-linear estimation loop. The current image is also acquired at time t. The intensity error to be minimised is shown $\mathcal{I}_t^w$-$\mathcal{I}^{*}$ before alignment along with the robust outlier weights. The warped current image $\mathcal{I}_t^w$ is given in the feedback loop.

The computed trajectories, shown in Figure 6, have been estimated in real-time (at 30Hz), however, the bottle neck was due to the bandwidth of the laptop hard-drive which was saving additional data for analysis and presentation of the results. If none of the images are recorded during the experiment it is possible to obtain 30Hz tracking. The plot of this trajectory is particularly illustrative since both laser and RGB-D trajectories are superposed and a return trip has been made so that the drift in the system is visible. It can be noticed that the Kinect trajectory acquires drift on the second bend. This error is due to the fact that the robot took the bend too rapidly and the fact that there were some jumps in image acquisition (since we used crucial processor time to save results to disk). In this case the images were too different from one-another for the algorithm to converge correctly. Considering that the trajectory of the robot was approximately $50m$ long, the absolute deviation in the position was around $1m$ which makes

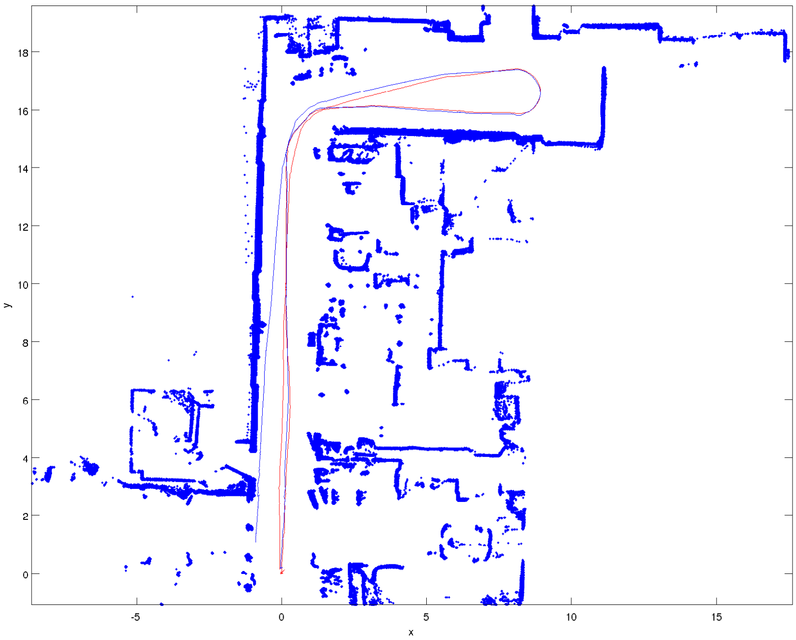


Figure 6: In red, the 2D trajectory computed from laser measurements. In blue the projection of the 3D trajectory computed from Kinect measurements onto the ground plane.

a drift of 2%.

In Figure 7, a reconstructed 3D scene is shown which relies on the precise estimation of the trajectory described above. The depth map aligns very precisely such that the join cannot be detected by visual inspection again confirming the quality of the estimated poses. It can be seen that there is a slight difference in illumination, mainly due to the bad quality camera and difference in exposure from varying position.

6 Conclusions and Future Works

The RGB-D methodology described in this paper has shown to be very efficient, accurate and robust within a complex indoor environment. The approach is interesting because trajectory estimation is integrated into a single global sensor-based process that does not depend of intermediate level features. Furthermore, a compact photometric model of the environment was obtained using the accurate pose measurement obtained by the localisation techniques. The robust non-linear estimator is shown to reject moving objects and outliers whilst the information selection and multi-resolution approach has allowed for real-time performance.

Subsequent work has been devoted to integrating the depth measurements over time in a Direct Iterative Closest Point approach [Tykkälä and Comport, 2011b]. It would be interesting to test loop closing procedures and devise strategies to recognise previously seen places within this framework.

7 Acknowledgements

The authors would like to thank Cyril Joly for his help in providing the laser ground truth data.

References

[Baker and Matthews, 2001] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[Benhimane and Malis, 2004] S. Benhimane and E. Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *IEEE International Conference on Intelligent Robots Systems*, Sendai, Japan, 28 September - 2 October 2004.

[Bouguet, 2010] Jean-Yves Bouguet. Camera calibration toolbox for matlab, July 2010.

[Chiuso *et al.*, 2002] A. Chiuso, P. Favaro, H. Jin, and S. Soatto. Structure from motion causally integrated over time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):523–535, 2002.

[Comport *et al.*, 2006a] A.I. Comport, E. Marchand, and F. Chaumette. Statistically robust 2d visual servoing. *IEEE Transactions on Robotics*, 22(2):415–421, April 2006.

[Comport *et al.*, 2006b] A.I. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless

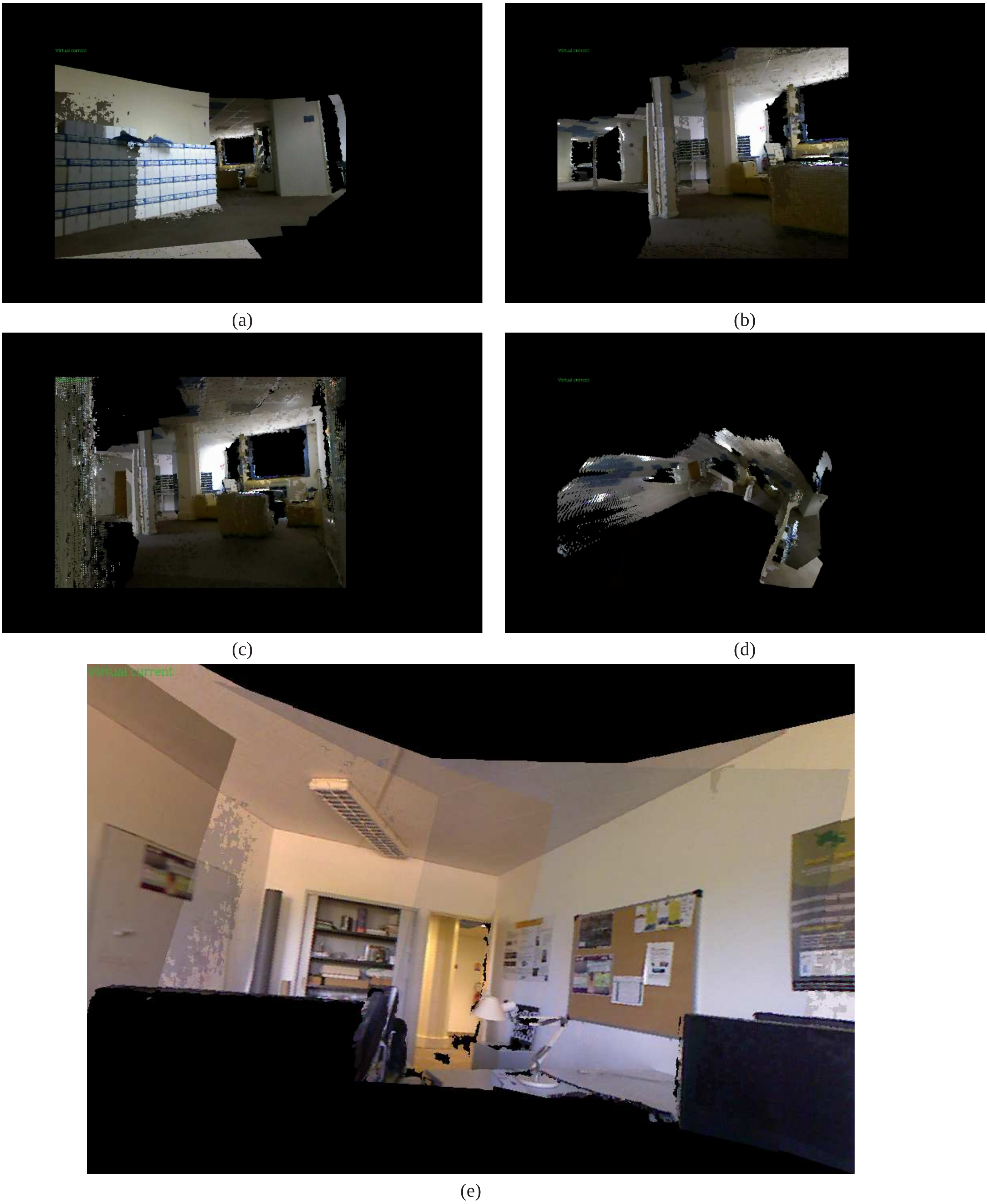


Figure 7: This figure shows a 3D reconstruction using the Kinect depth maps and the estimated trajectory to warp the texture-mapped model into the current reference frame. (a) The automatic shutter on the Kinect can be seen to give different lighting variation across different parts of the model which could be corrected by a more adapted sensor or using blending techniques. (d) Shows an overall view from outside the map. (e) A local map of an office.

tracking for augmented reality: the virtual visual servoing framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):615–628, July 2006.

[Comport *et al.*, 2007] A.I. Comport, E. Malis, and P. Rives. Accurate quadri-focal tracking for robust 3d visual odometry. In *IEEE Int. Conf. on Robotics and Automation, ICRA'07*, Rome, Italy, April 2007.

[Comport *et al.*, 2010] A.I. Comport, E. Malis, and P. Rives. Real-time quadrifocal visual odometry. *International Journal of Robotics Research, Special issue on Robot Vision*, 29(2-3):245–266, 2010.

[Davison and Murray, 2002] 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.

[Drummond and Cipolla, 2002] T. Drummond and R. Cipolla. Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):932–946, July 2002.

[Engelhard *et al.*, 2011] 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.

[Freedman *et al.*, 2010] B. Freedman, A. bd Shpunt, M. Machline, and Y. Arieli. *Depth mapping using projected patterns*. Number 20100118123 in 1. (Binyamina, IL), (Cambridge, MA, US), (Ashdod, IL), (Jerusalem, IL), May 2010.

[Gallegos *et al.*, 2010] G. Gallegos, M. Meilland, A.I. Comport, and P. Rives. Appearance-based slam relying on a hybrid laser/omnidirectional sensor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2010.

[Hager and Belhumeur, 1998] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, October 1998.

[Harris and Stephens, 1988] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 189–192, Manchester University, England, September 1988.

[Hartley and Zisserman, 2001] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge University Press, 2001. Book.

[Henry *et al.*, 2010] 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.

[Huber, 1996] P.J. Huber. Robust statistical procedures. *SIAM Review, second edition*, 1996.

[Karg *et al.*, 2010] M. Karg, K.M. Wurm, C. Stachniss, K. Dietmayer, and W. Burgard. Consistent mapping of multistory buildings by introducing global constraints to graph-based slam. In *IEEE International Conference on Robotics and Automation*, Anchorage, Alaska, May 3–8 2010.

[Meilland *et al.*, 2010] M. Meilland, A.I. Comport, and P. Rives. A spherical robot-centered representation for urban navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, October 2010.

[Mouragnon *et al.*, 2006] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real-time localization and 3d reconstruction. In *IEEE Conference of Vision and Pattern Recognition*, New-York, USA, June 2006.

[Newcombe and Davison, 2010] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1505, June 2010.

[Nistér *et al.*, 2004] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 652–659, CVPR 2004, July 2004.

[Pitzer *et al.*, 2010] B. Pitzer, S. Kammel, C. DuHadway, and J. Becker. Automatic reconstruction of textured 3d models. In *ICRA'10*, pages 3486–3493, 2010.

[Qi-Xing and Anguelov, 2010] H. Qi-Xing and D. Anguelov. High quality pose estimation by aligning multiple scans to a latent map. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 1353 –1360, May 2010.

[Sturm *et al.*, 2010] 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.

[Tykkala and Comport, 2011a] T.M. Tykkala and A.I Comport. A dense structure model for image based stereo slam. In *IEEE International Conference on Robotics and Automation*, Shanghai, China, May 9-13 2011.

[Tykkälä and Comport, 2011b] T.M. Tykkälä and A.I Comport. Direct Iterative Closest Point for Real-time Visual Odometry. In *The Second international Workshop on Computer Vision in Vehicle Technology: From Earth to Mars in conjunction with the International Conference on Computer Vision*, Barcelona, Spain, November 6-13 2011.