Robotics and Autonomous Systems 61 (2013) 195–208

Contents lists available at SciVerse ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

# Cross-spectral visual simultaneous localization and mapping (SLAM) with sensor handover

Marina Magnabosco, Toby P. Breckon *

*School of Engineering, Cranfield University, United Kingdom*

ARTICLE INFO

*Article history:*
Received 17 April 2012
Received in revised form
14 September 2012
Accepted 21 September 2012
Available online 23 October 2012

*Keywords:*
Thermal imaging SLAM
Sensor handover
Cross-spectral SLAM

ABSTRACT

In this work, we examine the classic problem of robot navigation via visual simultaneous localization and mapping (SLAM), but introducing the concept of dual optical and thermal (cross-spectral) sensing with the addition of sensor handover from one to the other. In our approach we use a novel combination of two primary sensors: co-registered optical and thermal cameras. Mobile robot navigation is driven by two simultaneous camera images from the environment over which feature points are extracted and matched between successive frames. A bearing-only visual SLAM approach is then implemented using successive feature point observations to identify and track environment landmarks using an extended Kalman filter (EKF). Six-degree-of-freedom mobile robot and environment landmark positions are managed by the EKF approach illustrated using optical, thermal and combined optical/thermal features in addition to handover from one sensor to another. Sensor handover is primarily targeted at a continuous SLAM operation during varying illumination conditions (e.g., changing from night to day). The final methodology is tested in outdoor environments with variation in the light conditions and robot trajectories producing results that illustrate that the additional use of a thermal sensor improves the accuracy of landmark detection and that the sensor handover is viable for solving the SLAM problem using this sensor combination.



## 1. Introduction

Autonomous robotics is an increasingly growing area in research and development within both academia and industry [1]. A key part of autonomous navigation for robotics used in a wide range of applications is the ability to localize within a given environment and additionally map that same environment. This is the classical simultaneous localisation and mapping (SLAM) problem of mobile robotics [2].

In this work, we uniquely use co-registered optical and thermal camera sensors and develop a SLAM system on a mobile robot platform capable of exploring and mapping a given environment. Furthermore, we introduce the concept of sensor handover, specifically to address the issues of extreme changes in illumination over a long-timescale SLAM mission (i.e. multi-day missions) where the advantages of thermal sensing are key under certain twilight/night illumination conditions whilst optical sensing remains for brighter illumination periods.

The novel aspects of this work are the cross-spectral SLAM (i.e., combined thermal and optical sensing) for mobile robot and additionally the concept of sensor handover between different sensors on a given SLAM mission. An example can be seen in Fig. 1, where thermal features introduce additional information to the scene not present from the optical sensor alone.

A simultaneous localization and mapping (SLAM) approach based on purely optical imagery features is susceptible to changing environmental lighting conditions for long-duration SLAM missions and additionally limits feature availability for nocturnal SLAM operations [3,1,2,4,5]. For long-duration SLAM missions within a military/security operational setting (e.g., transport or sentry), an autonomous system would require continuous daylight and nocturnal operation. A cross-spectral SLAM approach offers a robust navigation able to operate during day and night illumination transition without interruption of operations. Furthermore, a requirement exists to perform handover from one sensing modality to the other in order to carry out such continuous operations over changing illumination conditions. This is achieved using a passive sensing approach (i.e., visual SLAM). In addition, the cross-spectral capability of dual optical/thermal sensing facilitates includes scene detail and detects object presence (Fig. 2) not readily available in prior optical-only SLAM approaches [3].

In this work, we combine the use of speed-up robust feature points (SURF [6]), the proposed visual SLAM approach of [4] and additional robust RANSAC-based feature matching [7] as a solution to the visual SLAM problem over cross-spectral optical and thermal imagery. Additionally we consider the novel concept of sensor handover within this multi-modal sensing context.

* Corresponding author. Tel.: +44 7855549973.
*E-mail address:* toby.breckon@cranfield.ac.uk (T.P. Breckon).
*URL:* http://www.cranfield.ac.uk/~toby.breckon/ (T.P. Breckon).

0921-8890/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2012.09.023

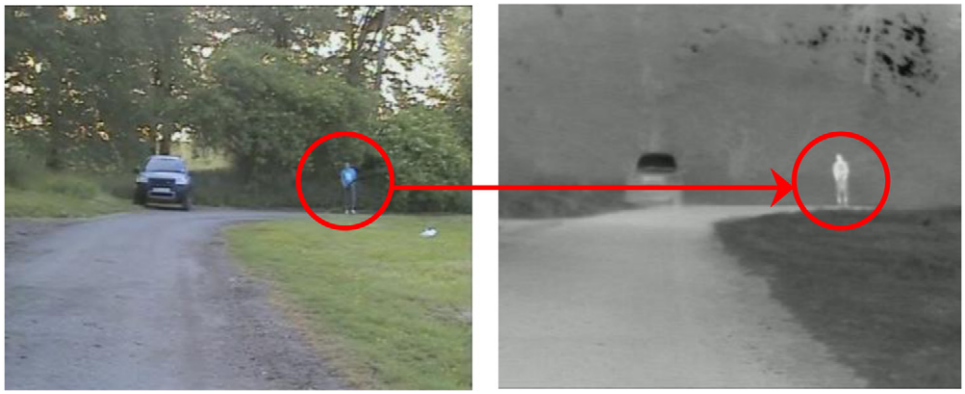

**Fig. 1.** Detection of human presence using optical and thermal cameras.
*Source:* SATURN project, Salisbury Plain, Wiltshire, UK, June 2009.

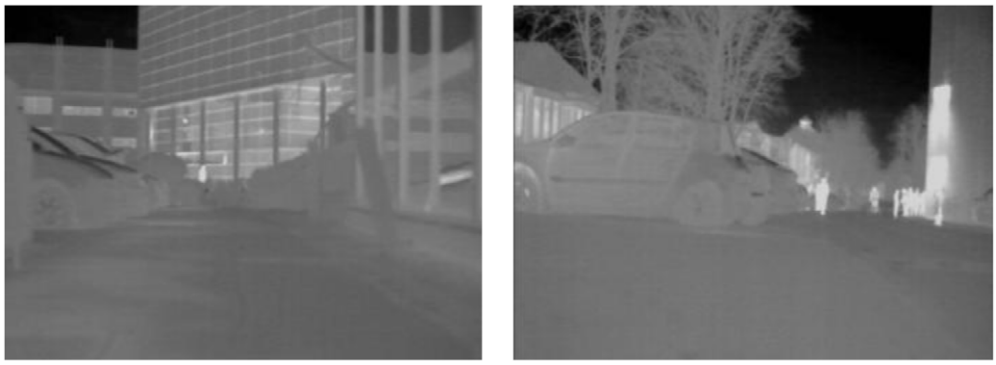

**Fig. 2.** Urban environment—thermal sensing examples.

## 2. Prior work

From the survey of [1], mobile robot navigation can be summarized into three principal subgenres: (a) *map-based navigation* (the mobile robot has an *a priori* map of the environment), (b) *map-building-based navigation* (the robot constructs a map of the environment) and (c) *mapless navigation* (the mobile platform does not use an explicit model and navigates based on object identification).

The SLAM problem is essentially a *map-building-based navigation* in which the system has a known kinematic model of the mobile robot performing the mapping task. Starting from an unknown position inside the environment, the aim of SLAM is to localize the vehicle within the environment and concurrently build an incremental navigation map using the observed landmarks from sensor information.

A useful method to decrease the positioning error of the robot inside the environmental map is the *loop closing method*. This method is key within SLAM; it consists of re-estimating the map when the robot returns to a previously visited location and, in this case, the robot is able to recognize the SLAM landmarks and increase the accuracy of the overall map. This method permits the construction of map with greater consistency through reducing the localization and the position errors of the landmarks [8], increasing the robustness of the overall pose estimation within the environment.

A wide range of devices can be used for robot navigation and they allow the detection of the necessary SLAM landmark information. Commonly SLAM is achieved using laser range sensors [8,10], which facilitates the recovery of explicit scene depth information or an alternative active sensing approach (e.g., LIDAR [11], millimetre wave radar [2]).

By contrast, visual SLAM uses a passive sensing approach based commonly on multiple views from a single camera (monocular SLAM) [11–15,4] or stereo-based approaches [14,16] to recover scene depth information. Visual SLAM [9] is commonplace within autonomous robotics and it is frequently augmented with odometry, inertial navigation and satellite-based GPS navigation sensors [11,2,16,17].

In general, one or more sensors are utilized and combined using a sensor fusion approach [18] to increase the overall SLAM robustness. In this work we consider the fusion of cross-spectral video sources (optical/thermal) to target long-term illumination-independent visual SLAM using a passive sensing approach. These are augmented with inertial and GPS sensors on our robot platform.

### 2.1. Visual SLAM

From [9], within the visual SLAM problem we can identify two main classes related to the method used to extract the information from the sensing devices. The first method is called *feature-based methods* [9], and it consists of extracting a sufficient number of features (e.g., points, lines, edges) and sequentially matching them between successive images. The matching stage is the key of all SLAM approaches: erroneous matching means erroneous pose estimation and hence erroneous map construction. The second class of methods based on [9] is the *direct methods*. The required information or parameters are directly extracted from the pixel intensity values (such as image brightness and illumination-based cross-correlation) [19].

Several techniques within the literature have been used for SLAM, including corner detection [13,20], scene flow [9,21], generalized feature points [6,22] and generalized segmentation [3,15].

In this work we uniquely extend earlier visual SLAM approaches to consider the use of such feature-based methods over a cross-spectral imagery with the target of being able to perform a sensor handover from optical to thermal sensing enabling long-term

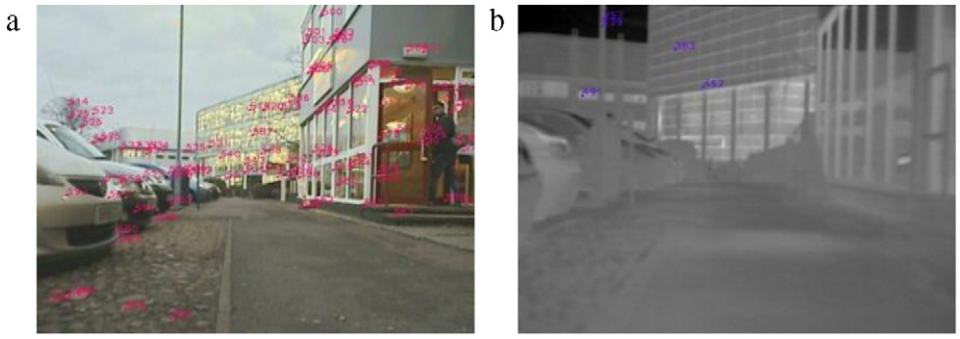


**Fig. 3.** SURF features extracted for (a) optical camera and (b) thermal camera.

visual SLAM, from passive image-based sensing, under extreme illumination variance.

## 3. Optical and thermal features

In our visual SLAM we require a feature-based method to extract information from our cross-spectral imagery. To meet both the scale and rotational invariance requirements of SLAM together with our desire for computational efficiency (enabling real-time performance) we select the SURF-based approach of [6].

### 3.1. SURF features

The speeded-up robust feature (SURF) method [6] is a robust image feature point detector and descriptor partially based on the seminal scale-invariant feature transform (SIFT) method [22]. As with the SIFT method, the SURF method is both scale and rotation invariant, but it is computationally faster, and thus makes it a good candidate feature detection algorithm for our application.

The first step of the SURF method is the detection of characteristic feature points in the image whilst the second step is to assign a unique descriptor to each detected point. The descriptor is a parameter vector of the feature point that aims to be unique because it is then used to subsequently match feature points between images. It has also to be noise robust, and invariant to scale and rotation in order to match the same points in different images.

The detector used in the SURF method [6] is based on the Hessian matrix of the image. Given a point $\boldsymbol{x} = \{x, y\}$ in an image $I$, the Hessian matrix $H(\boldsymbol{x}, \sigma)$ in $\boldsymbol{x}$ at scale $\sigma$ is defined as follows:

$$H(\boldsymbol{x}, \sigma) = \begin{bmatrix} L_{xx}(\boldsymbol{x}, \sigma) & L_{xy}(\boldsymbol{x}, \sigma) \\ L_{xy}(\boldsymbol{x}, \sigma) & L_{yy}(\boldsymbol{x}, \sigma) \end{bmatrix}, \tag{1}$$

where $L_{xx}(\boldsymbol{x}, \sigma)$ is the convolution of the Gaussian second-order derivative $\delta^2/\delta x^2 g(\sigma)$ with the image $I$ at the point $\boldsymbol{x}$, and similarly for $L_{xy}(\boldsymbol{x}, \sigma)$ and $L_{yy}(\boldsymbol{x}, \sigma)$, as described in [6]. To obtain a scale-invariant feature, the images are successively smoothed with a Gaussian filter and spatially subsampled [23].

In general, the SURF method [6] has being shown to perform better than other methods relative to its feature extraction performance against computational cost [24]. In our implementation we extract a 128-dimension SURF descriptor. SURF features are initially detected independently in each of the optical and thermal images (Fig. 3). In order to facilitate the relative correspondence of features occurring in either of these images, thermal to optical sensor calibration is required to calculate the relative image plane transformation between the two sensors.

### 3.2. Thermal to optical sensor calibration

An initial calibration of each optical and thermal camera is carried out to calculate the intrinsic and extrinsic camera parameters [25,26] denoted by optical and thermal camera calibration matrices, $K_O$ and $K_T$, respectively. These are required in order to recover the transformation between the two image planes permitting the use of common reference frames for the two camera sensors (local and global; see Section 4.1). This transformation allows the mapping of feature points detected in the thermal camera image to the spatial reference frame of the optical camera image, thus allowing the features detected in each image to be used in unison. This transformation is denoted as the *planar homography* that projects one image plane (thermal) to the other camera (optical). This type of mapping can be expressed in terms of matrix multiplication [25]. Given a point $\boldsymbol{x}_T$ in the thermal image and a correspondent point $\boldsymbol{x}_O$ in the optical image (taken from the same scene) we can express the operation of the homography mapping as follows:

$$\boldsymbol{x}_O = sH\boldsymbol{x}_T, \tag{2}$$

where $\boldsymbol{x}_O = \{x_O, y_O, 1\}^T$ and $\boldsymbol{x}_T = \{x_T, y_T, 1\}^T$ are in homogeneous coordinates, the parameter $s$ is an arbitrary scale factor and $H$ is a $3 \times 3$ transformation matrix.

The homography can be calculated using four correspondent planar points identified in each of corresponding optical and thermal scene images [25]. In Fig. 5, we see the thermal image transformed using the calculated homography $H$ overlain on the corresponding optical image based on the two original optical and thermal input images shown in Fig. 4.

In order to compute the homography matrix $H$, four points in both views (i.e. four correspondent points in the thermal image and in the optical image) are selected and successively calibrated using the correspondent calibration matrix (i.e., $K_T$ and $K_O$) to correct for camera lens and barrel distortions as follows:

$$\widetilde{\boldsymbol{x}}_T = K_T\boldsymbol{x}'_T \quad \text{and} \quad \widetilde{\boldsymbol{x}}_O = K_O\boldsymbol{x}'_O, \tag{3}$$

where $\boldsymbol{x}'_T = \left\{x'_T, y'_T, 1\right\}^T$ is a point in the thermal image after the distortion correction and $\widetilde{\boldsymbol{x}}_T = \left\{\tilde{x}_T, \tilde{y}_T, 1\right\}^T$ is the corresponding calibrated point ($\boldsymbol{x}'_O = \left\{x'_O, y'_O, 1\right\}^T$ and $\widetilde{\boldsymbol{x}}_O = \left\{\tilde{x}_O, \tilde{y}_O, 1\right\}^T$ are the optical points after distortion correction).

Successively, the points $\boldsymbol{x}'_T$ and $\boldsymbol{x}'_O$ are used to compute the homography matrix $H_c$ between the calibrated points. Given a calibrated thermal point $\widetilde{\boldsymbol{x}}_T$, the corresponding optical point is now estimated as follows:

$$\widetilde{\boldsymbol{x}}_O = H_c\widetilde{\boldsymbol{x}}_T, \tag{4}$$

where $H_c$ is now the homography based on the corrected thermal and optical images with respect to the distortion characteristics identified in the imagery. This calibration procedure to recover the homography $H_c$ is required for each individual camera set up.

Looking at the detail of Fig. 5(b), we can see that the optical and thermal images do not completely match due to parallax between the images, but this homography is empirically a good result for the feature matching required in this work.

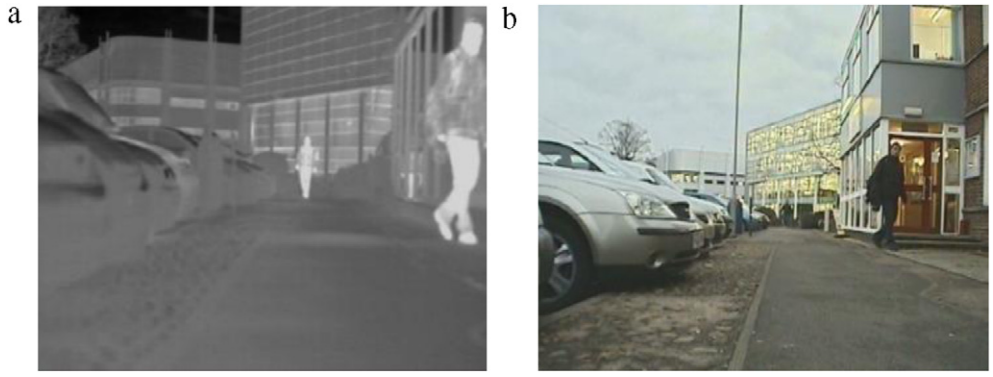


**Fig. 4.** Images used to compute the homography that maps the (a) thermal images to the (b) optical images.


**Fig. 5.** (a) Untransformed and (b) transformed images (overlay of the optical and thermal images using calibrated homography $H_c$).

As introduced earlier, the homography matrix $H$ is used to express the thermal image or thermal detected features within the reference frame of the optical camera (i.e., the reference frame used for the SLAM system). To compute the transformation from a thermal feature point in the thermal image plane to the optical image plane, we use the homography matrix $H$ and the intrinsic camera parameters $K_T$ and $K_O$ resulting from the single calibration of the thermal and optical camera, respectively. Following this approach the transformed thermal feature points are now expressed in the optical camera reference frame and they are ready for use in the initialization of the SLAM approach described subsequently.

## 4. From a video to a map

Following the mono-SLAM approach of [4], we construct a system based on a single camera sensor capable of building a 3D environmental map and self-localizing itself within that map over a given period of time. From [4] we develop an initial solution for monocular SLAM using optical sensing, inertial wheel encoder inputs and GPS—upon which thermally sensed features are latterly overlain.

In summary, the general approach is to extract SURF feature points and subsequently either initialize new features or match detected features against those within an existing feature database (i.e., known features from previous images). New features are initialized to select a set of possible 3D positions (i.e., 3D possible coordinates). Each new feature is initialized as a sum of Gaussians that is then updated with each subsequent observation [4]. From several successive observations, the depth of the feature and its 3D coordinates can be recovered, and the feature is selected as a candidate landmark to be added to the 3D environment map. The resulting map thus contains information about the position estimation of each identified landmark and its associated positional error [4].

In addition a dead reckoning approach using inertial wheel encoders and a secondary GPS receiver are used as additional sensors to resolve the global robot motion and position. As the robot continues in motion through the environment, management and merging of this information is required to maintain the environmental map. To facilitate the combination of this multi-sensor data for the estimation of the robot current position, an implementation of an extended Kalman filter [27] is used.

### 4.1. Feature extraction and initialization

From each captured optical and thermal image frame, SURF features [6] are extracted. Subsequently the extracted features are matched between successive frames on a per-sensor basis. This initial set of matches is used to build an initial feature database for each sensor (separately). These databases are then used to track the image to image features as the robot transits through the environment (optical and thermal separately). New feature points are added to these databases every $n$ frames based on a parallel matching technique. This technique consists in matching features among $m$ image frames, allowing the system to add features which are more frequently present in the field of view of the camera sensors. With this method we then are able to increase the probability of landmark detection as the initial feature points used are more stable over time.

After a feature is detected we follow the approach of [4], and each SURF feature point is initialized with a set of candidate positions and updated using successive observations as described

in [4]. For each update of the feature positions, the 3D coordinates of the initial set is pruned until a last update selects one candidate as the initial 3D coordinates (see Section 4.2).

For a single feature point, once extracted from the image, the only information about its position in space that can be immediately recovered is the 2D pixel coordinates in the image with respect to the camera reference frame. Given a 3D point $P$ with coordinates $\boldsymbol{X} = \{X, Y, Z\}$ in the scene, we can express its corresponding 2D image pixel coordinates using the calibration matrix $K$ as follows:

$$\lambda \boldsymbol{x} = \lambda \begin{Bmatrix} u \\ v \\ 1 \end{Bmatrix} = \begin{bmatrix} fs_x & 0 & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} X \\ Y \\ Z \end{Bmatrix} = K\boldsymbol{X}, \tag{5}$$

where $\lambda$ is the distance (depth) of point $X$ along the $z$ axis of the camera, $\boldsymbol{x} = \{u, v\}$ the 2D pixel coordinates in the image camera plane of $P$ and $f$ the camera focal length, whilst $s_x$ and $s_y$ are the lengths of the pixel along the horizontal and vertical directions respectively and $o_x$ and $o_y$ are the $x$ and $y$ coordinates of the optical centre of the image plane. The matrix $K$ and the relative elements are recovered from prior calibration [26, see Section 3.2].

From Eq. (5), we can see that the depth $\lambda$ is the $Z$ coordinate of a feature, which is in general unknown at this stage of SLAM when using a monocular camera model [4]. Multiplying the 2D pixel coordinates $\boldsymbol{x} = (u, v)$ by the inverse of the calibration matrix, we obtain the expression to recover the 3D coordinates $\boldsymbol{X}$ of $P$ up to a scalar factor $\lambda$:

$$\boldsymbol{X} = \begin{Bmatrix} X \\ Y \\ Z \end{Bmatrix} = \lambda K^{-1} \begin{Bmatrix} u \\ v \end{Bmatrix}. \tag{6}$$

The scalar factor $\lambda$ represents the unknown depth of the point $P$. Division of Eq. (6) by $\lambda$ gives:

$$\boldsymbol{X}/\lambda = \begin{Bmatrix} X/\lambda \\ Y/\lambda \\ Z/\lambda \end{Bmatrix} = \begin{Bmatrix} X/Z \\ Y/Z \\ 1 \end{Bmatrix} = \begin{Bmatrix} x \\ y \\ 1 \end{Bmatrix} = K^{-1} \begin{Bmatrix} u \\ v \\ 1 \end{Bmatrix}. \tag{7}$$

The next stage is the recovery of the direction of the point with respect to the camera reference frame. The direction of a feature is represented by the angles $\theta$ and $\phi$:

$$\begin{Bmatrix} \theta \\ \phi \end{Bmatrix} = \begin{Bmatrix} \arctan\ (y/x) \\ -\arctan\ (1/\sqrt{x^2 + y^2}) \end{Bmatrix}, \tag{8}$$

and these angles refer to a polar reference frame, as shown in Fig. 6.

As we use images from a single camera, the current depth of a feature point $P$, translated as radius $\rho$ in a polar coordinates system (see Fig. 6), is unknown, as previously outlined. Based on [4], we estimate the spatial position (and thus the depth) of a feature initializing the correspondent 3D coordinates using a sum of Gaussians. This sum of Gaussians is successively updated using observations of the same feature over time based on our feature matching methodology (to achieve consistent match features we use a nearest-neighbour technique [28] in combination with the RANSAC algorithm [7]). The main hypothesis that allows us to compute this initial set of Gaussians for a feature point is the established specific depth range $[\rho_{\min}, \rho_{\max}]$ for any given sensor. However, the initial values of $\rho_{\min}$ and $\rho_{\max}$ are somewhat arbitrary, and can be set empirically based on the environment and known sensor capabilities.

Following the approach of [4], the 3D coordinates of a point $P$ can thus be expressed as:

$$P(\theta, \phi, \rho) = \Gamma(\theta, \sigma_\theta) \cdot \Gamma(\phi, \sigma_\phi) \cdot \sum_i \omega_i \Gamma_i(\rho_i, \sigma_{\rho_i}), \tag{9}$$

where $P(\theta, \phi, \rho)$ represents the spatial position of $P$ with its relative error in polar coordinates and $\sum_i \omega_i \Gamma_i(\rho_i, \sigma_{\rho_i})$ represents

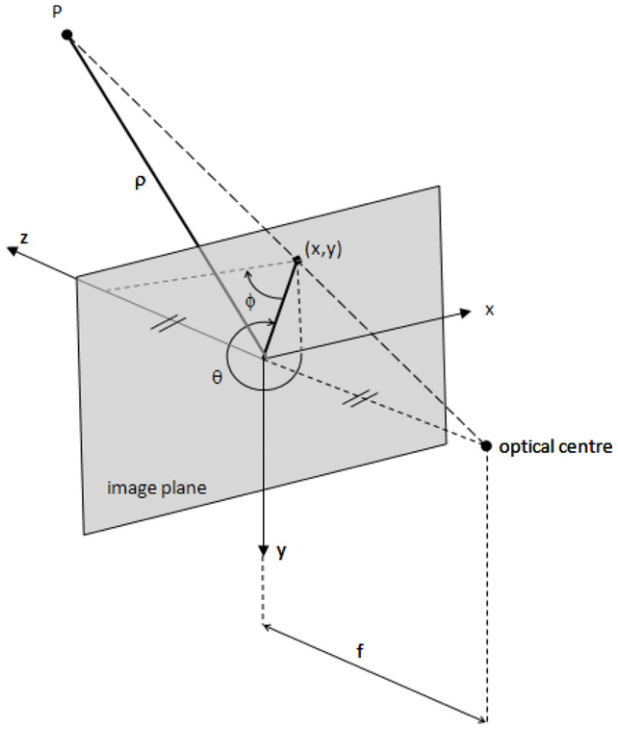


**Fig. 6.** Representation of a feature in polar coordinates $(\theta, \phi)$.

the sum of Gaussians that approximates the *a priori* knowledge of the depth. According to [4] the depth of the feature can be computed using the following geometric series:

$$\rho_0 = \rho_{\min}/(1 - \alpha), \tag{10}$$

$$\rho_i = \beta^i \cdot \rho_0, \qquad \sigma_{\rho_i} = \alpha \cdot \rho_i, \qquad \omega_i \propto \rho_i, \tag{11}$$

$$\rho_{n-2} < \rho_{\max}/(1 - \alpha), \qquad \rho_{n-1} \geq \rho_{\max}/(1 - \alpha). \tag{12}$$

Once again the values of $\alpha$ (=0.25) and $\beta$ (=2.5) are chosen empirically but following constraints related to the distribution of a Gaussian that we want to obtain [4]. After this initialization, each Gaussian $\left\{\mu_i^p = \{\rho_i, \theta, \phi, \}, \sum_i^p = \left\{\sigma_{\rho_i}^2, \sigma_\theta^2, \sigma_\phi^2\right\}\right\}$ is converted from polar coordinates to Cartesian coordinates:

$$\mu_i^c = \boldsymbol{g}\begin{pmatrix} \theta \\ \phi \end{pmatrix} = \begin{Bmatrix} \rho_i \cos\phi \cos\theta \\ \rho_i \cos\phi \sin\theta \\ -\rho_i \sin\phi \end{Bmatrix} = \begin{Bmatrix} x_i \\ y_i \\ z_i \end{Bmatrix} \tag{13}$$

$$\sum_i^c = G \sum_i^p G^T = \left\{\sigma_{x_i}^2, \sigma_{y_i}^2, \sigma_{z_i}^2\right\}, \tag{14}$$

where $G = \partial \boldsymbol{g}/\partial \boldsymbol{X}|_{(\rho_i, \theta, \phi)}$.

The reference frame used to express the initial 3D coordinates of a feature point is the robot reference frame at the instant when the feature is seen the first time.

After this initialization, we obtain $n$ 3D coordinates of the same feature, and an update stage is necessary to select which Gaussian best approximates the feature pose. An example of the output of the initialization process is shown in Fig. 7, where each Gaussian in the set is represented as an ellipsoid of error and the orientation is related to the first relative direction in which the feature was initially seen.

### 4.2. Initial position update and landmark initialization

Using successive observations of the same feature point, a selection procedure can be performed computing an estimation of

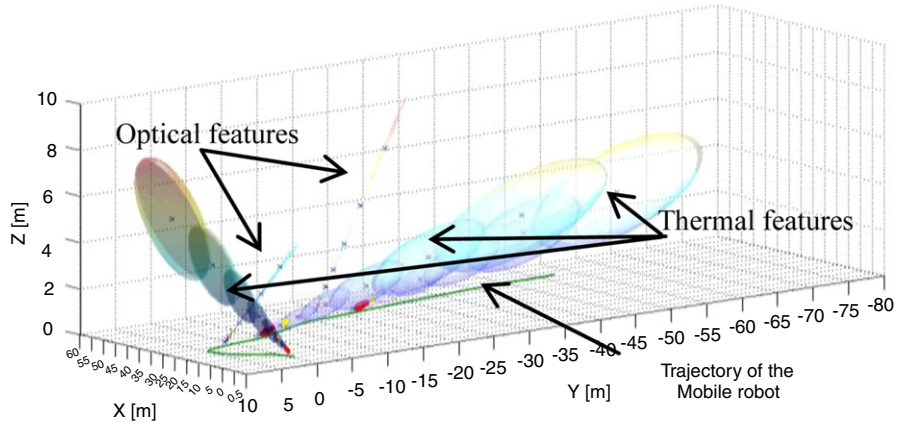


**Fig. 7.** Feature initialization—landmark selection and estimation related to the robot position.

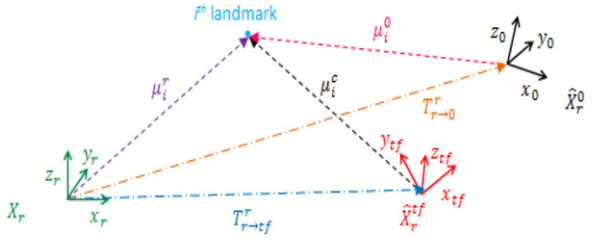


**Fig. 8.** Feature/landmark position vector with respect to different robot reference frames.

**Table 1**
Description of the variables used in our formulation.

| Symbols | Description |
|---|---|
| $X_r$ | *global* (or *world*) *reference frame*, taken as the first position of the mobile robot |
| $\hat{X}_r^{tf}$ | *robot reference frame* where the landmark is initially seen, coordinates are expressed with respect to the global reference frame |
| $\hat{X}_r^0$ | *current robot reference frame*, coordinates are expressed with respect to the global reference frame |
| ${}^r_{tf}R$, $T^r_{r\to tf}$ | rotation matrix, $R$, and translation vector, $T$, that express the rotation from the robot reference frame at time $t$ to the global reference frame |
| ${}^0_rR$, $T^r_{r\to 0}$ | rotation matrix, $R$, and translation vector, $T$, that expresses the rotation from the global reference frame to the current robot frame |
| $\mu_i^r$ | 3D coordinates of the $i$th landmark with respect to the *global* reference frame |
| $\mu_i^c$ | 3D coordinates of the $i$th landmark with respect to the *robot* reference frame |
| $\mu_i^0$ | 3D coordinates of the $i$th landmark with respect to the *current* robot reference frame |

the normalized likelihood for each Gaussian, $\Gamma_i$. The likelihood of $\Gamma_i$ to be an estimation of the observed feature is computed as follows:

$$L_i^t = \frac{1}{2\pi\sqrt{|S_i|}} \exp\left(-\frac{1}{2}\left(z_t - \hat{z}_i\right)^T S_i^{-1}\left(z_t - \hat{z}_i\right)\right), \tag{15}$$

where $S_i$ is the covariance of the innovation $z_t - \hat{z}_i$ [4]. The prediction of the observation $\hat{z}_i = (\theta_i, \phi_i)$ is estimated considering each Gaussian in the current robot frame (i.e., at time $t$) and $z_t$ is the observation at the corresponding time $t$.

In Fig. 8 and Table 1 we present three different robot reference frames generally used and the notation of the landmark vector with respect to these reference frames (i.e., $\mu_i^r$, $\mu_i^c$ and $\mu_i^0$). All the information from the images is acquired in the image reference frame, while the information stored in the EKF and in the subsequent 3D environmental map refers to the robot reference frame. The robot reference frame differs from the image reference frame by two simple rotations along the $z$ and $x$ axes of $\pi/2$.

From the specification of the variables, we can thus formulate a computation for $\hat{z}_i$, the prediction of the $i$th observation in the *current* robot reference frame, as follows:

$$\hat{z}_i = \boldsymbol{h}\left(\text{to}\left(\hat{X}_r^0, \text{from}\left(\hat{X}_r^{tf}, \mu_i^c\right)\right)\right) = H\left(\hat{X}_r^0, \hat{X}_r^{tf}, \mu_i^c\right), \tag{16}$$

where

$$\mu_i^r = \text{from}\left(\hat{X}_r^{tf}, \mu_i^c\right) = {}^r_{tf}R \cdot \mu_i^c + T^r_{r\to tf}, \tag{17}$$

$$\mu_i^0 = \text{to}\left(\hat{X}_r^0, \mu_i^r\right) = {}^0_rR \cdot \mu_i^r - T^r_{r\to 0}, \tag{18}$$

$$\hat{z}_i = \begin{Bmatrix}\theta_i \\ \phi_i\end{Bmatrix} = \begin{Bmatrix}\arctan\left(-z_i^0/x_i^0\right) \\ -\arctan\left(\dfrac{y_i^0}{\sqrt{\left(x_i^0\right)^2 + \left(-z_i^0\right)^2}}\right)\end{Bmatrix}, \tag{19}$$

based on the formulation of [4]. From Eqs. (18) and (19) we can see how the variable $\hat{z}_i$ is the Cartesian coordinates $\mu_i^0$ transformed into the polar reference frame, from where we can indicate $\boldsymbol{h}()$ to be the transformation from Cartesian coordinates to polar coordinates [4].

Furthermore, in Eq. (15) we use the matrix $S_i$, the covariance of the innovation $z_t - \hat{z}_i$, which is computed as follows:

$$S_i = H_1 P_{X_r^0} H_1^T + H_2 P_{X_r^{tf}} H_2^T + H_1 P_{X_r^0, X_r^{tf}} H_2^T + H_2 P^T_{X_r^0, X_r^{tf}} H_1^T + H_3 \sum_i^c H_3^T + R_t, \tag{20}$$

where

$$H_1 = \frac{\partial H}{\partial X_r^0}, \qquad H_2 = \frac{\partial H}{\partial X_r^{tf}} \quad \text{and} \quad H_3 = \frac{\partial H}{\partial \mu_i^c} \tag{21}$$

where $R_t$ is the covariance associated with the observation $z_t$ and $P_{(k)}$ is the co-variance matrix associated with a given reference frame $k$ from Table 1 (where reference frame $k$ in Table 1 is relative to the origin with associated transformation done by $T$ and $R$).

Following [4], to compare the likelihood of each Gaussian, we use the normalized likelihood; that is, for hypothesis $i$, the product of likelihoods obtained for $\Gamma_i$ is

$$\Lambda_i = \frac{\prod_t L_i^t}{\sum_j \prod_t L_j^t}. \tag{22}$$

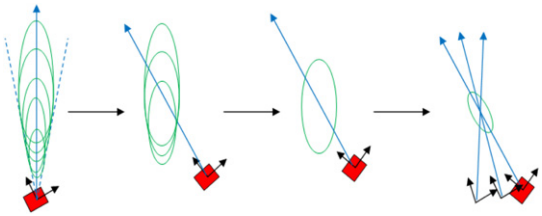

**Fig. 9.** From an observed feature point in the image frame to a landmark in the 3D environmental map.

The normalized likelihood is computed every time we have a new observation $z_t$ of the feature point for each Gaussian, and the Gaussian associated with the worst hypotheses is pruned if $\Lambda_i < \tau$ ($\tau = 0.5/n, n =$ number of Gaussians remaining). Following several observations we end up with a single Gaussian, and the associated 3D coordinates of the feature are compared with the last observation using the $\chi^2$ test [4,5]. If the coordinates pass the $\chi^2$ test, the associated feature is declared as a landmark and is inserted in the landmark database and deleted from the correspondent features database (i.e., optical or thermal). If this is not the case, this means that the feature is not in the pre-specified depth range $[\rho_{\min}, \rho_{\max}]$ that we have identified in the initialization stage or alternatively the observations were not consistent enough so the feature has been rejected by the $\chi^2$ test.

The process from initial feature to a new landmark is shown in Fig. 9 with the reduction in spatial Gaussian (illustrated in two dimensions). The leftmost image (Fig. 9) represents the initialization stage of a feature as a sum of Gaussians. The successive two images in Fig. 9 show that, thanks to successive observations of the same feature, some Gaussians are subsequently pruned,leaving a single estimate of the 3D position with a Gaussian error.

When only one Gaussian remains and passes the $\chi^2$ test, the feature is declared as a landmark and it is projected into the global reference frame. At the final stage, the past observations of the feature point are used to update the estimation of the landmark position and reduce the overall error related to the Gaussian (see right image in Fig. 9) prior to be added to the 3D environment map. As the number of landmarks detected from the environment increases, we slowly build up a 3D landmark map of the environment as the mobile robot transits through the environment.

### 4.3. Robot position and 3D map update

Once we start constructing the environment map, we use observations of the identified landmarks as additional information input to the extended Kalman filter (EKF) [27,29]. This is used in combination with the inertial wheel encoders and additional GPS receiver data to better estimate the 3D map of the environment and the position of the robot within the environment map.

The state of the EKF is composed by the current robot position, the past $n$ robot positions and the $m$ landmark estimates. The prior $n$ robot position are not subjected to update or prediction within the EKF framework. These robot positions are kept to allow the position update process to identify the robot position where a feature is initially detected.

The overall EKF state can be represented as follows:

$$X = \begin{Bmatrix} X_r^t \\ \vdots \\ X_r^0 \\ X_{l,1}^t \\ \vdots \\ X_{l,m}^t \end{Bmatrix}, \qquad P = \begin{bmatrix} P_{X_r^t} & \cdots & P_{X_r^t,X_r^0} & P_{X_r^t,X_{l,1}^t} & \cdots & P_{X_r^t,X_{l,m}^t} \\ & \ddots & \vdots & \vdots & \cdots & \vdots \\ & & P_{X_r^0} & P_{X_r^0,X_{l,1}^t} & \cdots & P_{X_r^0,X_{l,m}^t} \\ & & & P_{X_{l,1}^t} & \cdots & P_{X_{l,1}^t,X_{l,m}^t} \\ & & & & \ddots & \vdots \\ & & & & & P_{X_{l,m}^t} \end{bmatrix}, \tag{23}$$

where $X$ is the state vector and $X_r^t$ represents the current position of the robot, $X_r^0$ is the first saved past position of the robot and $X_{l,i=1,\ldots,m}^t$ is the position of the $i$th landmark at time $t$. The matrix $P$ is the covariance matrix of the entries of the EKF state vector, and it represents the error of the estimations.

After a given amount of iteration, the overall size of the EKF state increases due to the addition of new landmarks within the environment map and the constant recording of prior robot positions. In order to reduce both the memory footprint and the computational cost of the overall EKF process, the size of the EKF state vector is managed by removing prior robot positions that are not used for subsequent feature tracking from the vector used in the initial position update process (see Section 4.2).

## 5. Results

This work is realized on Pioneer 3-AT mobile robot platform equipped with both optical and thermal cameras (Visionhitech VC57WD-24 CCTV—spectral range: 400–700 nm/Thermoteknix

**Fig. 10.** Pioneer 3-AT equipped with optical and thermal sensors.


**Fig. 11.** Outdoor test environment: (a) optical camera, (b) thermal camera.

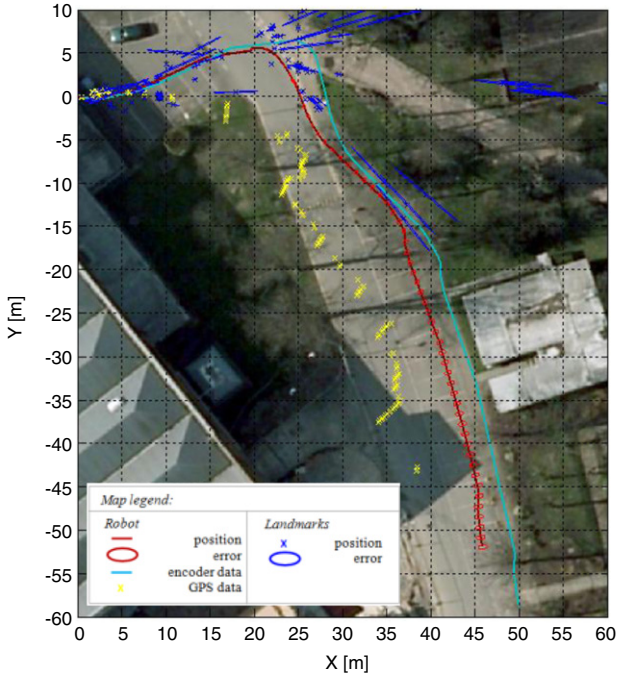


**Fig. 12.** 2D environmental map for the optical sensor SLAM analysis (case 2).

MIRICLE 110K—spectral range: 8–12 μm). Furthermore, the mobile robot is equipped with a GPS GlobalSat BU–353 receiver and it is manually controlled at a variable speed within the test scenario used. The system configuration is shown in Fig. 10.

The system is tested over a range of outdoor environments around the Cranfield University campus (Fig. 11(a)–(b)).

In general, within these test environments we see a diverse range of features at a varying depth detected from the optical camera due to the large level of detail within the image and our choice of the feature detector [6, Refer to Fig. 3 and Fig. 11 (Fig. 3 for the features and Fig. 11 for the difference in the details of the image)]. By contrast, from the thermal camera a variety of features are still detected with varying signature features, such as people and aspects of building within the environment (Fig. 11(a)). In general the feature density within the thermal imagery is lesser than the corresponding optical imagery (Fig. 3). However, it has to be noted that the density of the available optical imagery features is dependent on the illumination condition, whereas the thermal features are largely constant and dependent on the thermal dynamics of the environment rather than varying illumination conditions.

### 5.1. Optical SLAM

As first reference implementation we use a single camera system replicating the proposed solution of [4] using only the optical camera sensor. For all the reported examples, satellite imagery [30] is overlaid onto the 2D environmental map (see Fig. 13(a)). For this case 135 optical landmarks are detected with average error ±2.10 m, ±0.40 m and ±0.21 m along the $x$, $y$ and $z$ axes, respectively.

A second example is presented in a different environment setting for comparison with the resulting 2D map/satellite imagery shown in Fig. 12. For this example case 129 optical landmarks are detected with average error ±1.46 m, ±0.60 m and ±0.43 m along the $x$, $y$ and $z$ axes, respectively.

### 5.2. Thermal SLAM

The experiments from Section 5.1 are repeated using the thermal sensor (features detected on the same physical 'run' of the robot). Both cameras are initialized identically, with the only exception being the initial value of the depth range limits $[\rho_{\min}, \rho_{\max}]$. In general, the minimum ray $\rho_{\min}$ used to initialize new thermal feature points is larger than the one used for optical features (see Section 4.1). The motivation for this is related to the different field of view of the thermal sensor, which is significantly narrower than that of the optical sensor. As a consequence, following the initialization process (see Section 4), the ellipses of error for the thermal landmark results are larger than for optical

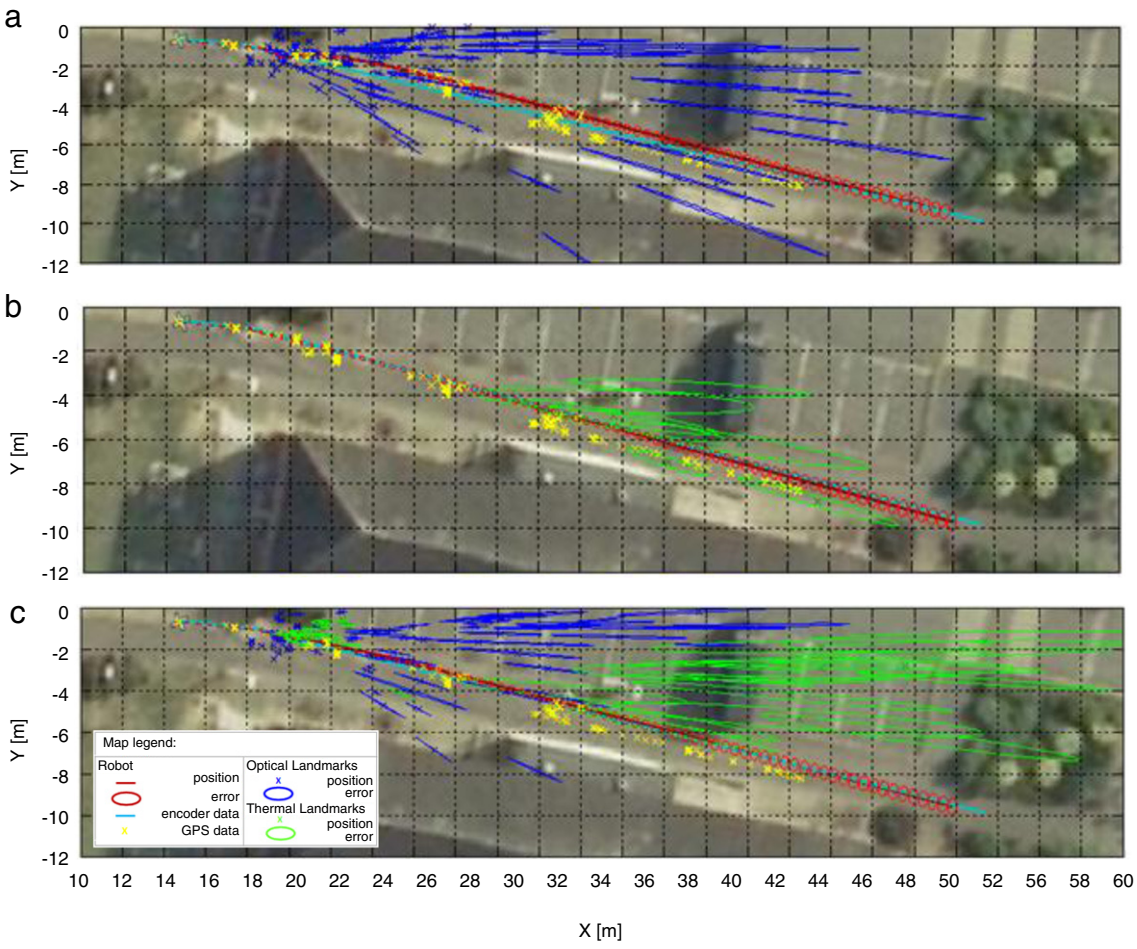


**Fig. 13.** 2D environmental map for the (a) optical sensor SLAM analysis, (b) thermal sensor SLAM analysis and (c) cross-spectral SLAM analysis (case 1).

landmarks (compare Fig. 13(a) and (b)). For the thermal case shown in Fig. 13(b), the analysis using the thermal sensor detects seven thermal landmarks with average error ±8.31 m, ±1.30 m and ±0.97 m along the $x$, $y$ and $z$ axes, respectively.

Comparing the optical and thermal analyses for this case (compare Fig. 13(a) and (b)), we see a significant reduction in the landmarks detected during the exploration of the area. This can be attributed to a low level of thermal detail within this particular portion of the environment and additionally the narrower field of view of the thermal camera sensor.

For the second thermal SLAM test case, the 2D environmental map is shown in Fig. 14; it detects 122 thermal landmarks with average error ±0.94 m, ±0.68 m and ±0.31 m along the $x$, $y$ and $z$ axes, respectively. For this case, we notice a similar level of landmark detection as in the corresponding optical case (see Fig. 12), and additionally note the comparable average errors over the set of detected landmarks. Overall we can see that thermal sensor-based SLAM can perform at a comparable level to that obtained using the optical sensor.

### 5.3. Cross-spectral SLAM

The experimental analysis undertaken in Sections 3.1 and 3.2 is repeated using both the optical and thermal camera sensors for independent feature detection as a conduit to landmark detection. For the first analysis case, using the cross-spectral SLAM approach (Fig. 13(c)), the level of optical landmark detection is comparable to the prior optical sensor only case (Fig. 13(a)), but we see a significant increase in the number of thermal landmarks detected

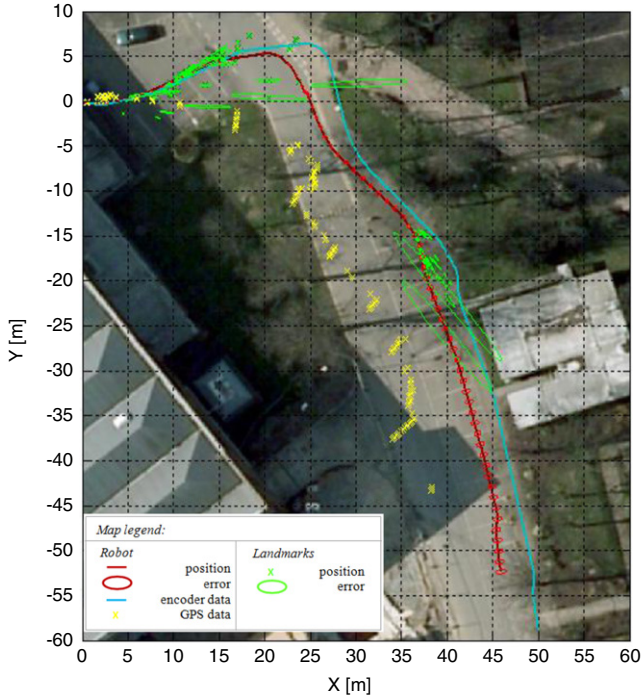


**Fig. 14.** 2D environmental map for the thermal sensor SLAM analysis (case 2).

with respect to the prior thermal sensor only analysis (Fig. 13(b)). Furthermore, the environmental map is represented in three

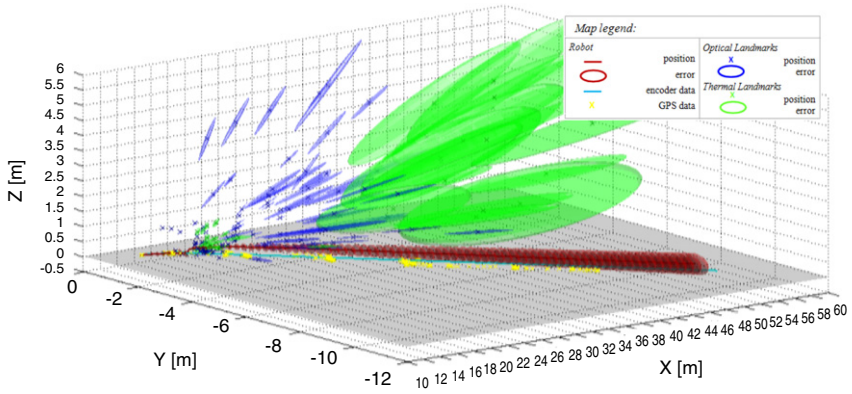


**Fig. 15.** 3D environmental map for the cross-spectral SLAM analysis (case 1).

**Table 2**
Final value of the EKF state vector $\{X, Y, Z, \gamma, \beta, \alpha\}$ for the cross-spectral SLAM, the optical sensor only SLAM and thermal sensor only SLAM (case 1).

| | Cross-spectral SLAM | Optical SLAM | Thermal SLAM |
|---|---|---|---|
| $X$ (m) | $51.82 \pm 0.39$ | $51.61 \pm 0.39$ | $51.91 \pm 0.40$ |
| $Y$ (m) | $-9.59 \pm 0.66$ | $-9.41 \pm 0.65$ | $-9.73 \pm 0.65$ |
| $Z$ (m) | $0.22 \pm 0.45$ | $0.37 \pm 0.40$ | $0.06 \pm 0.47$ |
| $\gamma$ (°) | $-0.55 \pm 8.96$ | $-3.85 \pm 8.04$ | $-1.60 \pm 9.06$ |
| $\beta$ (°) | $-2.73 \pm 8.56$ | $-4.25 \pm 7.81$ | $-5.32 \pm 7.84$ |
| $\alpha$ (°) | $-14.42 \pm 4.59$ | $-14.67 \pm 4.59$ | $-14.40 \pm 4.59$ |

dimensions in Fig. 15 to give an alternative illustration of the overall estimated map following the SLAM result presentation style of the original base work [4].

For this case, the combination of the optical and thermal sensors allows the system to detect 96 optical landmarks and 36 thermal landmarks with average error $\pm 3.06$ m, $\pm 0.34$ m and $\pm 0.31$ m along the $x$, $y$ and $z$ axes, respectively. In comparison to the optical sensor only implementation of [4] (Section 4.1) we show that the cross-spectral SLAM approach gives an increase of 18.2% of the landmarks detected. The is attributable to the fact that the cross-spectral SLAM using two sensors is able to extract more information from the explored environment.

During this experiment a number of dynamic objects (people) with significant thermal and optical features also enter the environment within the field of view of the sensors. It is important to outline that the presence of these dynamic objects in the scene did not affect the overall operation of the approach. This can be attributed to the threshold added before the use of any given observation vector in the update stage of the EKF that allows the system to have a robust feature/landmark matching process (Sections 4.2 and 4.3).

Table 2 shows the final estimated positions of the mobile robot for the three sensor variation approaches (i.e., cross-spectral SLAM, optical SLAM and thermal SLAM). The final error associated with the robot position estimation using combined sensing is almost identical to those estimated from the optical sensor only case, but notably the use of the thermal camera provides additional information during the mapping phase of the SLAM approach with the advantage of providing further thermal features/landmarks. This is achieved without introducing further significant errors into the system.

From Table 2, if we consider that the mobile robot is moving on a surface that can be considered planar, we can observe that the results obtained for the cross-spectral SLAM system give a better estimate of the robot position with particular regards to the robot orientation along the $x$ and $y$ axes.

Fig. 16 shows the 2D environmental map of the second analysis case for combined cross-spectral sensing. In comparison with the optical sensor only case (Fig. 12) and the thermal sensor only case (Fig. 14) this cross-spectral SLAM analysis obtained 104 optical landmarks and 67 thermal landmarks with average error of $\pm 1.06$ m, $\pm 0.43$ m and $\pm 0.26$ m along the $x$, $y$ and $z$ axes, respectively.

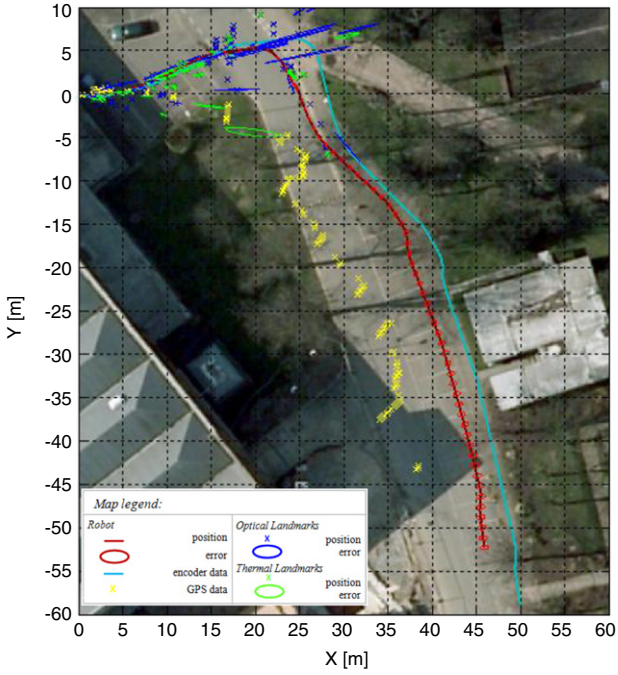


**Fig. 16.** 2D environmental map for the cross-spectral SLAM analysis (case 1).

Despite the decrease of the number of thermal landmarks detected in the combined system with respect to the thermal sensor only analysis (Section 5.2), the positional errors (ellipses in Figs. 13(b)–14 and Figs. 13(c)–16) for the thermal landmarks in the cross-spectral SLAM analysis are actually smaller than those recovered in the thermal sensor only case. Comparing the cross-spectral SLAM system with respect to the single optical camera implementation of [4] we have an increase of 32% in the number of landmarks detected and a decrease of 27.4%, 28.3% and 39.5% in the average landmark error along the $x$, $y$ and $z$ axes, respectively. It is important to notice that the increase in the overall number of landmarks results in a higher number of landmark observations. These combined optical and thermal landmark observations are dependent on the current robot position (the landmark position observations are indirect), and when used as input to the EKF this dependency on the current status of the mobile robot positioning allows us to reduce the overall EKF state vector error. The corresponding 3D map of the combined system for this second case is shown in Fig. 17.

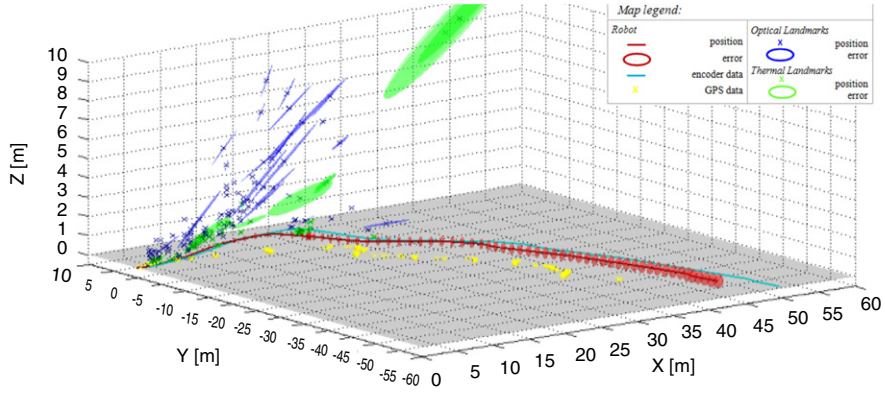


**Fig. 17.** 3D environmental map for the cross-spectral SLAM analysis (case 2).

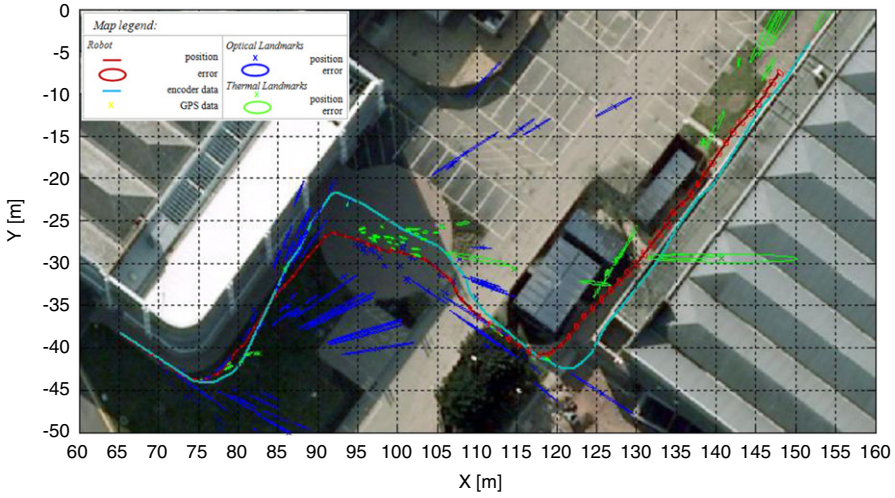


**Fig. 18.** 2D environmental map for the cross-spectral SLAM analysis (case 3).

**Table 3**
Final value of the EKF state vector $\{X, Y, Z, \gamma, \beta, \alpha\}$ for the cross-spectral SLAM, the optical sensor only SLAM and thermal sensor only SLAM (case 2).

| | Cross-spectral SLAM | Optical SLAM | Thermal SLAM |
|---|---|---|---|
| $X$ (m) | $45.96 \pm 0.96$ | $45.90 \pm 0.96$ | $45.92 \pm 0.96$ |
| $Y$ (m) | $-52.28 \pm 0.43$ | $-51.95 \pm 0.43$ | $-52.34 \pm 0.43$ |
| $Z$ (m) | $0.05 \pm 0.47$ | $0.07 \pm 0.47$ | $0.01 \pm 0.48$ |
| $\gamma$ (°) | $-3.65 \pm 9.48$ | $-20.18 \pm 8.94$ | $-2.91 \pm 9.49$ |
| $\beta$ (°) | $-12.82 \pm 9.46$ | $-18.59 \pm 8.73$ | $-11.14 \pm 9.49$ |
| $\alpha$ (°) | $-81.49 \pm 4.20$ | $-81.41 \pm 4.20$ | $-81.76 \pm 4.20$ |

From Table 3 we can observe that for this case the cross-spectral SLAM performs better in terms of estimated final robot position than in the optical sensor only case, suggesting that the cross-spectral SLAM approach benefits from the addition of thermal landmark observations improving the mobile robot localization within the environment.

This experiment is carried out for a third case and the 2D environmental map is shown in Fig. 18. In this example we can see a prevalence of optical landmarks at the beginning of the navigation whilst a detection of thermal landmarks seems to dominate the last part of the mobile robot route. This is related to a higher presence of optical features in the first part of the mobile robot route in comparison to the end of the route, whilst for thermal features the opposite trend is apparent. In this example 213 landmarks are detected: 141 are optical landmarks and 72 are thermal landmarks, with an average error $\pm 1.82$ m, $\pm 1.49$ m and $\pm 0.58$ m along the $x$, $y$ and $z$ axes, respectively.

For this third case, different optical and thermal landmarks are detected during the navigation, and a wide range of error ellipses is present in the resulting environment map (Figs. 18 and 19). This extent of landmark positioning errors is related to the initial coordinates used for a new landmark; landmarks that are detected further away from the camera sensors are going to be initialized with a larger correspondent error as the initial error is proportional to the distance of the landmark from the camera (see Section 4.1). Fig. 19 represents the 3D environmental map for the combined system of this third case.

### 5.4. Sensor handover within SLAM

As shown in Section 5.3, the combined use of optical and thermal sensors allows the mobile robot to perform the SLAM task independent of illumination conditions and improves estimation of the robot position compared to a single sensor SLAM approach [4]. Another advantage of this approach is that if features from one of the two sensors become unavailable the system is still able to explore the environment, performing SLAM and detecting features/landmarks, as the data from the remaining sensor is used to complete this portion of the SLAM task. To test the performance of the system, we simulate a sensor handover situation where we essentially remove the optical sensor as an input to the system within the SLAM mission and rely on thermal features only for a mid-portion of the mission. This is performed over a short period of a longer mission to simulate a significant illumination change (darkness) affecting the availability of optical features for visual SLAM.

In Fig. 20, we see a second analysis case (previous case 2 from Section 5.3) where optical camera features are not used from frame 220 until frame 1560 of the overall data sequence (the optical

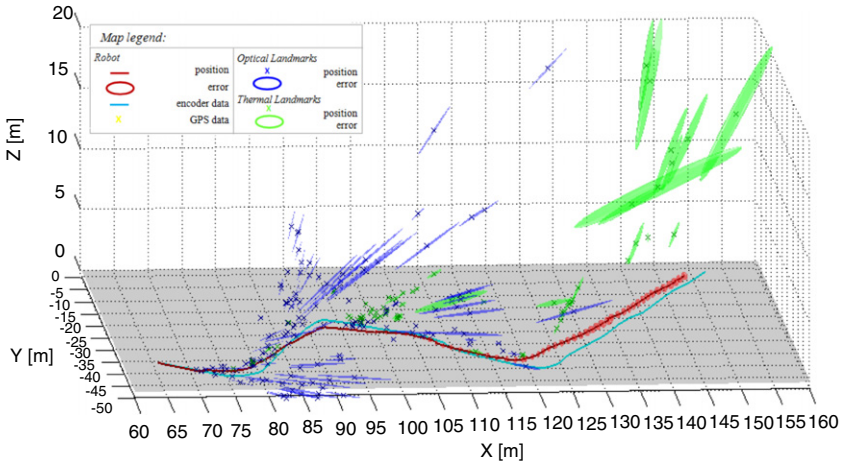


**Fig. 19.** 3D environmental map for the cross-spectral SLAM analysis (case 3).

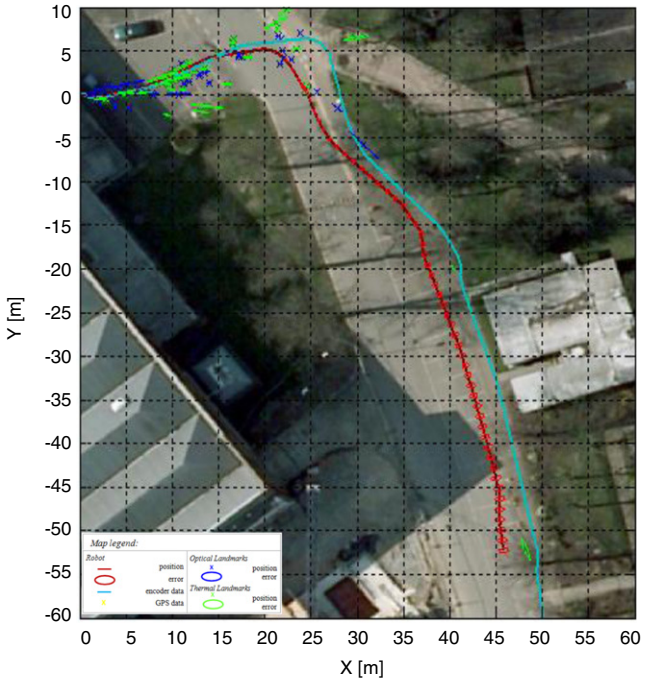


**Fig. 20.** 2D environmental map for the handover analysis (case 2).

features are absent over 30% of the mission). In this first handover case, 118 landmarks are detected, of which 62 are thermal and 56 are optical detected landmarks, with an average error of ±0.59 m, ±0.22 m and ±0.15 m along the $x$, $y$ and $z$ axes, respectively.

Comparing the sensor handover condition with the cross-spectral SLAM case, it is notable that fewer landmarks are detected with particular reference to optical landmarks. This is somewhat to be expected because it is the data from the optical camera in which we are simulating an outage, requiring handover, for approximately a third of the overall mobile robot navigation route. In Table 4, we report the final position of the robot for both the cross-spectral SLAM and the sensor handover conditions. The estimated position for the mobile robot under the sensor handover condition is comparable in terms of the error intervals with the cross-spectral SLAM case, and from this we can see that the combined sensing approach is able to cope with the missing sensor data that could be caused by extreme changes in illumination conditions (i.e., day/night operation) or sensor malfunction.

The same handover analysis is carried out on the third case (see Fig. 21). Similarly, for this analysis case, the optical camera

**Table 4**
Final value of the EKF state vector $\{X, Y, Z, \gamma, \beta, \alpha\}$ for regular cross-spectral SLAM and sensor handover case (case 2).

| | Cross-spectral SLAM | Handover |
|---|---|---|
| $X$ (m) | $45.96 \pm 0.96$ | $45.84 \pm 0.97$ |
| $Y$ (m) | $-52.28 \pm 0.43$ | $-52.22 \pm 0.44$ |
| $Z$ (m) | $0.05 \pm 0.47$ | $0.05 \pm 0.47$ |
| $\gamma$ (°) | $-3.65 \pm 9.48$ | $-2.47 \pm 9.47$ |
| $\beta$ (°) | $-12.82 \pm 9.46$ | $-12.85 \pm 9.39$ |
| $\alpha$ (°) | $-81.49 \pm 4.20$ | $-81.42 \pm 4.24$ |

is not being used for approximately 30% of the navigation route to simulate a significant change in the lighting condition which limits the usefulness of the optical camera features. This result is illustrated graphically in Fig. 21 and additionally with comparison to the cross-spectral SLAM case in Table 5.

In this case the approach is able to detect 167 landmarks; 62 are thermal and 105 are optical landmarks, with an average error of ±1.29 m, ±0.54 m and ±0.09 m along the $x$, $y$ and $z$ axes, respectively. The number of landmarks detected is again fewer than in the cross-spectral SLAM case but this is related to the absence of optical data information for part of the robot navigation route.

The final robot position estimation is reported in Table 5 with the relative errors and compared to the cross-spectral SLAM case. From Table 5 we can see the difference in the robot position estimation of 4 m for the $X$ position and 2 m for the $Y$ position. In this final case the thermal camera is notably positioned to an offset to the optical sensor field of view (see Fig. 22) in an attempt to maximize the number of near field features detected by the thermal sensor. This configuration in addition with the narrow field of view of the thermal sensor (in comparison with the optical camera) shows that purely thermal camera navigation is not ideal and the overall system suffers to a greater degree with a lack of optical data during the handover condition. However this does not affect the ability of the system to estimate its position and detect landmarks, but does appear to affect the localization of its overall final performance. Future work will look at using multiple thermal sensors or a wider field of view to give greater coincidence coverage of both the optical and thermal field of views within this task (Fig. 5).

In general, we can see that the SLAM approach can cope well under optical to thermal sensor handover conditions, simulating daylight to nocturnal illumination changes, and we see that the impact on the overall localization and mapping components of the SLAM task is minimal.

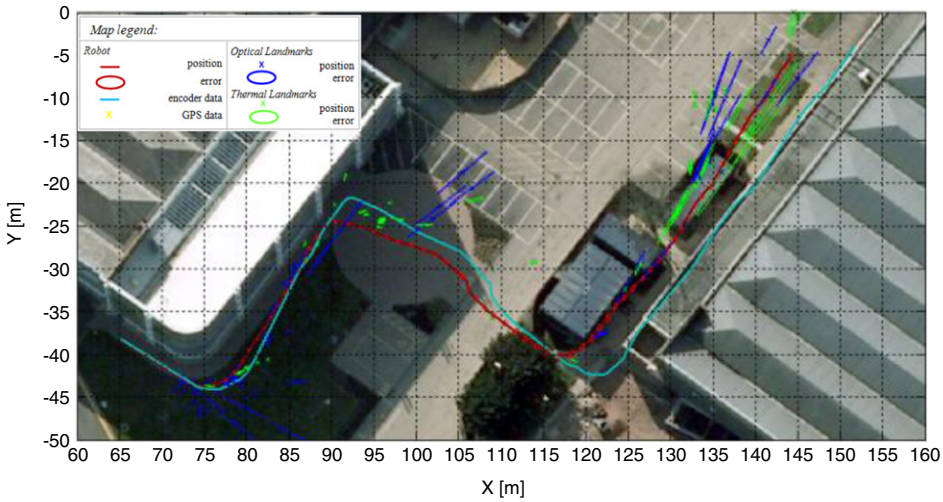


**Fig. 21.** 2D environmental map for the sensor handover case analysis (case 3).

**Fig. 22.** Example of the output during the navigation (case 3).

**Table 5**
Final value of the EKF state vector $\{X, Y, Z, \gamma, \beta, \alpha\}$ for regular cross-spectral SLAM and sensor handover case (case 3).

| | Cross-spectral SLAM | Handover |
|---|---|---|
| $X$ (m) | $148.00 \pm 0.60$ | $144.21 \pm 0.34$ |
| $Y$ (m) | $-7.52 \pm 0.63$ | $-5.16 \pm 0.47$ |
| $Z$ (m) | $0.35 \pm 0.44$ | $0.27 \pm 0.34$ |
| $\gamma$ (°) | $10.99 \pm 4.28$ | $3.72 \pm 2.33$ |
| $\beta$ (°) | $-11.13 \pm 6.68$ | $-11.22 \pm 5.27$ |
| $\alpha$ (°) | $51.09 \pm 2.98$ | $58.30 \pm 1.85$ |

## 6. Conclusion

We present a solution for the SLAM problem using combined optical and thermal sensing. An implementation of the technique of [4] is successfully achieved, and it permits the introduction of two novel aspects of this work: (1) the additional use of a secondary thermal sensor to complement the existing optical sensor in the SLAM navigation task (cross-spectral SLAM navigation) and (2) sensor handover between cameras operating in different parts of the spectrum over a single SLAM mission. The evaluation of the approach presented confirms that the information added by the thermal camera improves the performance of the monocular SLAM approach (despite being used as independent, non-stereo sensors) as it increases the number of detected landmarks and decreases the average error related to the landmark positions. Furthermore, it improves the estimation of the mobile robot position throughout the time integration steps (i.e., improved localization). This performance of the system is confirmed for different types of environment and in varying lighting conditions including the presence of moving objects within the scene.

In addition, we illustrate the first use of cross-spectral sensor handover where for a single SLAM mission we perform part of the route with combined optical and thermal (cross-spectral) sensing and perform handover to a single sensor (thermal camera) during the mission. This work extends the state of the art with respect to the monocular single sensor SLAM approach of [12,13,4] by illustrating its extension to cross-spectral sensing and additionally with the ability of handover between multi-sensor and single sensor sensing with minimum effect on the overall localization and mapping.

Further work could develop the integration of an optical flow technique [21] as a supplementary tool for the measurement of the robot motion, cross-spectral stereo [31] and additionally consideration of SLAM loop closing in a multi-sensing/cross-sensing environment.

## References

[1] G.N. Desouza, A.C. Kak, Vision for mobile robot navigation: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2) (2002) 237–267.
[2] M.W.M.G. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, M. Csorba, A solution to the simultaneous localization and map building (SLAM) problem, IEEE Transactions on Robotics and Automation 17 (3) (2001) 229–241.
[3] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image and Vision Computing 22 (10) (2004) 761–767.
[4] T. Lemaire, C. Berger, I. Jung, S. Lacroix, Vision-based SLAM: stereo and monocular approaches, International Journal of Computer Vision 74 (3) (2007) 343–364.
[5] Y. Lee, T. Kwin, J. Song, SLAM of a mobile robot using thinning-based topological information, International Journal of Control, Automation and Systems 5 (5) (2007) 577–583.
[6] H. Bay, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, in: European Conference of Computer Vision, 2006, p. 404.
[7] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the Association for Computing Machinery 24 (6) (1981) 381–395.
[8] A. Nüchter (Ed.), 3D Robotic Mapping the Simultaneous Localization and Mapping Problem with Six Degrees of Freedom, in: Springer Tracts in Advanced Robotics, 2009.
[9] C.C. Wang, C. Thorpe, Simultaneous localization and mapping with detection and tracking, in: IEEE International Conference on Robotics and Automation, vol. 3, 2002, pp. 2918–2924.

[10] J. Choi, S. Ahn, W.K. Chung, Robust sonar feature detection for the SLAM of mobile robot, in: International Conference on Intelligent Robots and Systems, 2005, pp. 3415–3420.
[11] J.C. Last, R. Main, Techniques used in autonomous vehicle systems: a survey, University of Northern Iowa, 2009.
[12] B. Williams, G. Klein, I. Reid, Real-time SLAM relocalisation, in: IEEE International Conference on Computer Vision, vol. 0, 2007, pp. 1–8.
[13] Z. Zhang, Y. Huang, C. Li, Y. Kang, Monocular vision simultaneous localization and mapping using SURF, in: World Congress on Intelligent Control and Automation, 2008, pp. 1651–1656.
[14] L.M. Paz, P. Pinies, J.D. Tardos, J. Neira, Large-scale 6-DOF SLAM with stereo-in-hand, IEEE Transactions on Robotics 24 (5) (2008) 946–957.
[15] P. Sturgess, K. Alahari, L. Ladickỳ, P.H.S. Torr, Combining appearance and structure from motion features for road scene understanding, in: British Machine Vision Conference, London, September 2009.
[16] N. Muhammad, D. Fofi, S. Ainouz, Current state of the art of vision based SLAM, in: Society of Photo-Optical Instrumentation Engineers, SPIE, Conference Series, vol. 7251, 2009.
[17] H. Choset, K. Nagatani, Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization, IEEE Transactions on Robotics and Automation 17 (2) (2001) 125–137.
[18] J. Sola, A. Monin, M. Devy, T. Vidal-Calleja, Fusing monocular information in multicamera SLAM, IEEE Transactions on Robotics 24 (5) (2008) 958–968.
[19] M. Irani, P. Anandan, About Direct Methods, in: B. Triggs, A. Zisserman, R. Szeliski (Eds.), Vision Algorithms: Theory and Practice, Springer, Berlin, Heidelberg, 2000, pp. 267–277.
[20] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Vision Conference, Manchester, 1988, pp. 147–151.
[21] F. Liu, V. Philomin, Disparity estimation in stereo sequences using scene flow, in: British Machine Vision Conference, London, September 2009.
[22] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision, 2, 1999, pp. 1150–1157.
[23] P. Pinggera, T.P. Breckon, H. Bischof, On cross-spectral stereomatching using dense gradient features, in: Proc. British MachineVision Conference, 2012. pp. 526.1–526.12.
[24] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. VanGool, A comparison of affine region detectors, International Journal of Computer Vision 65 (1–2) (2005) 43–72.
[25] Y. Ma, S. Soatto, J. Kosecka, S.S. Sastry, An Invitation to 3-D Vision: From Images to Geometric Models, Springer-Verlag, 2003.
[26] Z. Zhang, A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (11) (2000) 1330–1334.
[27] S. Schmidt, Applications of state–space methods to navigation problems, in: Advances in Control Systems, Vol. 3, 1966, pp. 293–340.
[28] M. Mount, S. Arya, ANN: a library for approximate nearest neighbor searching, Version 1.1.2. Available at: http://www.cs.umd.edu/~mount/ANN/ (accessed: 04/29).
[29] E. Nebot, Navigation system design, Unpublished Centre of Excellence for Autonomous Thesis, University of Sydney, Australia, May 2005.
[30] Google Maps, Explore the world using interactive maps. Available at: http://www.google.co.uk/help/maps/tour/ (accessed: 12.02.11).
[31] C.J. Solomon, T.P. Breckon, Fundamentals of Digital ImageProcessing: A Practical Approach with Examples in Matlab, Wiley-Blackwell, 2010.

**Marina Magnabosco** is Automation Controls System Developer at Ocado Technology (UK). Her work is focused on the delivery of efficient, reliable and robust software solutions for automatic distribution systems core to the company activities.

Marina Magnabosco holds an M.Sc. by Research in Image Processing and Robotics (2011) from Cranfield University (UK), where she applied an optical–thermal combined sensing for autonomous navigation. She also holds an M.Eng. (2009) and a B.Eng. (2007) in Aerospace Engineering from the University of Padova (Italy). Her key research interests are related to autonomous systems, image processing and computer vision.

**Toby P. Breckon** is currently a senior lecturer within the School of Engineering, Cranfield University (UK). His key research interests lie in the domain of computer vision and image processing, and he leads a range of research activity in this area.

Dr. Breckon holds a Ph.D. in informatics from the University of Edinburgh (UK). He is a visiting member of faculty at the Ecole Suprieure des Technologies Industrielles Avances (France) and has additionally held visiting positions at Northwestern Polytechnical University (China), Shanghai Jiao Tong University (China) and Waseda University (Japan).

Dr. Breckon is a Chartered Engineer, Chartered Scientist and professional member of the IET and BCS. In addition, he is an Accredited Imaging Scientist and an Associate of the Royal Photographic Society. He led the development of image-based automatic threat detection for the 2008 UK MoD Grand Challenge winners (R.J. Mitchell Trophy, (2008), IET Innovation Award (2009)). His work is recognised via the Royal Photographic Society Selwyn Award for early-career contribution to imaging science (2011).