Selective Combination of Visual and Thermal Imaging for Resilient Localisation in Adverse Conditions: Day and Night, Smoke and Fire

Christopher Brunner[1], Thierry Peynot[1], Teresa Vidal-Calleja[2] and James Underwood[1]

[1] Australian Centre for Field Robotics

The University of Sydney

NSW 2006, Australia

{c.brunner,tpeynot,j.underwood}@acfr.usyd.edu.au

[2] Centre for Autonomous Systems

Faculty of Engineering and IT

University of Technology Sydney

NSW 2007, Australia

Teresa.VidalCalleja@uts.edu.au

Abstract

Long-term autonomy in robotics requires perception systems that are resilient to unusual but realistic conditions that *will* eventually occur during extended missions. For example, unmanned ground vehicles (UGVs) need to be capable of operating safely in adverse and low-visibility conditions, such as at night or in the presence of smoke. The key to a resilient UGV perception system lies in the use of multiple sensor modalities, e.g. operating at different frequencies of the electromagnetic spectrum, to compensate for the limitations of a single sensor type. In this paper, visual and infrared imaging are combined in a Visual-SLAM algorithm to achieve localisation. We propose to evaluate the quality of data provided by each sensor modality prior to data combination. This evaluation is used to discard low-quality data, i.e. data most likely to induce large localisation errors. In this way,

perceptual failures are anticipated and mitigated. An extensive experimental evaluation is conducted on data sets collected with a UGV in a range of environments and adverse conditions, including the presence of smoke (obstructing the visual camera), fire, extreme heat (saturating the infrared camera), low-light conditions (dusk), and at night with sudden variations of artificial light. A total of 240 trajectory estimates are obtained using 5 different variations of data sources and data combination strategies in the localisation method. In particular, the proposed approach for selective data combination is compared to methods using a single sensor type or combining both modalities without pre-selection. We show that the proposed framework allows for camera-based localisation resilient to a large range of low-visibility conditions.

1 Introduction

In the near future, unmanned ground vehicles (UGVs) are expected to operate for long periods of time in unknown and unstructured environments with minimum supervision. Long-term autonomy requires perception systems that are resilient to unusual but realistic conditions that *will* eventually occur during extended missions. For example, robots need to be able to operate safely at night or in the presence of thick fog or smoke.

Perception is arguably the most critical component of a UGV system. It can be defined as the interpretation of sensor data to provide a representation of the environment that is appropriate for a particular application, e.g. path planning or localisation, see Fig. 1. Perception is widely recognised as a bottleneck in this search for long-term operation of UGVs. This is because the realm of situations the vehicle may encounter is unbounded, thus it is not possible to define a model for each one, and it is difficult to generalise with a single perception model. In particular, state-of-the-art robotic perception models relying on a single type of sensor have shown their limitations in adverse environmental conditions, causing data misinterpretation and subsequent failures. Examples include laser-based perception in the presence of airborne dust (Urmson et al., 2008) or smoke (Castro and Peynot, 2012) and visual camera in smoke (Brunner and Peynot, 2010). This paper proposes to extend the range of situations that UGV perception systems are resilient to, with a focus on conditions of low visibility.

Conditions of low visibility for a particular sensor are found when the sensor is employed outside of its nominal operating environments. For example, common causes include insufficient lighting, saturation, or the effects

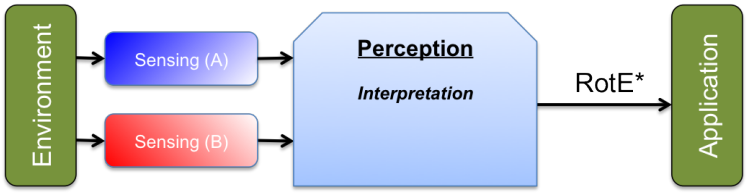
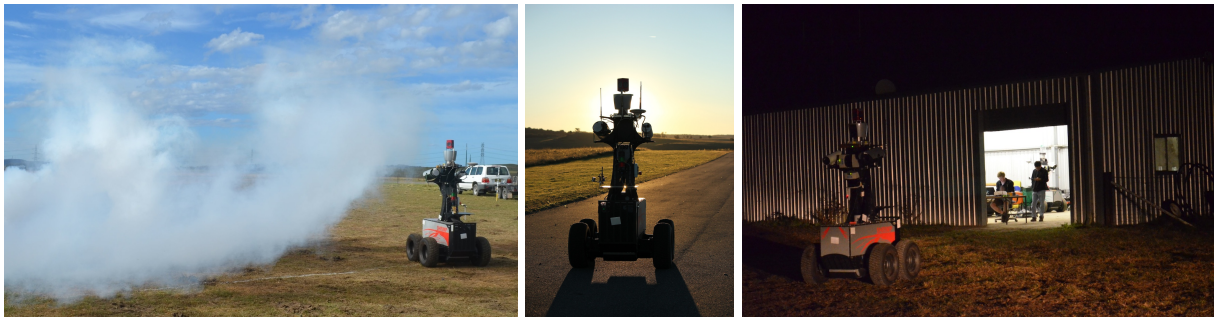


Figure 1: Example of perception system with two sensing modalities. Sensors perceive the environment and the perception system interprets the sensed data to produce a Representation of the Environment (*RotE*) that is suitable for an application.



(a) Smoke (b) Sunset (c) Night

Figure 2: Challenging low-visibility conditions for perception.

of obscurants such as fog or smoke. Fig. 2 illustrates examples of low-visibility conditions for a visual camera (i.e. operating in the visible spectrum). These situations are particularly challenging for perception, whether it is because of lack of information or misleading information that cause data misinterpretation and subsequent perceptual failures.

The key to a resilient UGV perception system lies in the use of multiple *sensor modalities*, i.e. sensors that perceive the environment using distinct physical processes, e.g. operating at different frequencies of the electromagnetic spectrum. Using different sensor modalities enables to compensate for the limitations of a single sensor type, which means *common-mode failures*[1] can be mitigated. In this paper, we investigate the detection and mitigation of perceptual failure due to low-visibility conditions by using visual and infrared (IR) sensors onboard a UGV for a *Visual-SLAM*[2] application to estimate the pose of the platform. Visual cameras and infrared cameras can be complementary in a range of low-visibility conditions. For instance, Figs. 3 and 4 show the effect of smoke and darkness respectively on visual and IR data. In these cases, the IR camera is providing much better information of the background environment than the visual camera, which is greatly affected. Alternatively, as shown in Fig. 5, infrared images can be affected in environments with high heat, where visual images are not. In this paper, we show that by using this complementarity of visual and IR cameras, we can implement a camera-based localisation system that is resilient to a range of

[1]Multiple sensors all using the same physical process will similarly fail in the same type of low-visibility conditions. For example, multiple visual cameras do not provide better perception at night than a single visual camera.

[2]In this work, Visual-SLAM refers to *camera-based* Simultaneous Localisation and Mapping.

Figure 3: Representative images in clear conditions (left column) and in smoke (right) from the visual (top row) and IR (bottom) cameras.

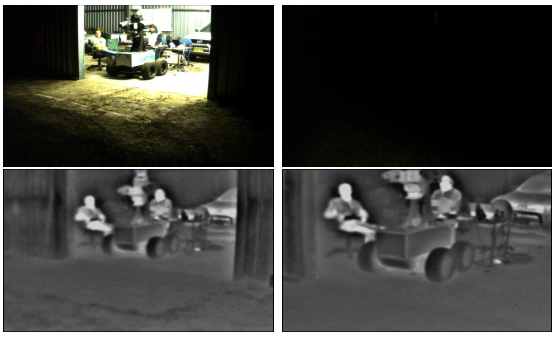


Figure 4: Representative images in lit conditions (left column) and in complete darkness (right) from the visual (top row) and IR (bottom) cameras.

Figure 5: Perception of a flame by visual (left) and IR (right) cameras.

low-visibility conditions.

When combining all available sensor data into the perception algorithm, inappropriate sensor data (e.g. affected by challenging conditions) may cause significant errors in the localisation algorithm, leading to perceptual failures. Additionally, error detection, diagnosis and recovery after data combination can be extremely difficult, and often impossible. Therefore, error mitigation requires the anticipation of potential errors as early as possible in the system. In this work, we propose to evaluate the quality of data provided by each sensor modality prior to data combination. This evaluation is used to discard low-quality data, i.e. data most likely to induce large localisation errors. In this way, perceptual failures are anticipated and mitigated. We show that our proposed framework for *selective* data combination allows for resilient localisation in various low-visibility conditions.

Through an extensive experimental evaluation, we compare different versions of the camera-based localisation algorithm, using data sets collected with a UGV in a range of environments and adverse conditions. These include the presence of smoke (obstructing the visual camera), extreme heat (saturating the IR camera), low-light conditions (dusk) and darkness (night) with sudden variations of artificial light. The study first shows that we can obtain a localisation that is robust to smoke and darkness using an IR camera within the Visual-SLAM algorithm. Second, we show that better results can be obtained by exploiting the combination of both modalities: visual and IR cameras. Third, the results show that the proposed method for selective

data combination based on prior quality evaluation further mitigates errors in low-visibility conditions. Fourth, we improve over the previous technique by achieving a *local* evaluation of data quality, prior to data selection and combination of selected data for localisation. We analyse experimental results from a total of 240 trajectory estimates obtained using 5 different variations of data sources and data combination strategies in the localisation method, in 6 different situations and environments.

The paper is organised as follows. Section 2 discusses related work. Section 3 introduces the proposed method to evaluate the quality of sensor data prior to using them in a perception algorithm. Section 4 presents the core algorithm used for camera-based localisation in this paper. Section 5 specifies the proposed process of automatic data pre-selection for localisation. Section 6 presents the experimental setup (robotic platform and data sets). Section 7 illustrates and discusses the experimental results. Finally, Section 8 summarises the conclusions of this study and discusses future work.

2 Related Work: Resilient Perception in Low-Visibility Conditions

A resilient UGV perception system requires that the operational environment be visible to the available sensor(s) at all times. Visibility is defined as the distance one can perceive a high-contrast object (Pearsall, 1999). The visibility of a scene in sensor data (e.g. in an image) depends on: 1) the intensity[3] of the electromagnetic (EM) spectrum in the scene and the attenuation of the spectrum due to the atmospheric conditions (Brooker, 2009), 2) the capability of the sensor to capture EM energy (e.g. spectral range and sensitivity), and 3) the conversion of the EM input signal into output data by physical and electronic processes (e.g. amplification, discretisation, quantification, compression). Low visibility due to the environment can be *global*, with similar visibility conditions in every direction (e.g. night or dense fog as viewed by a visual camera). In other common cases (e.g. presence of smoke), visibility conditions can be *locally variable*, in position, density, and over time, which is more difficult to model. Variable visibility conditions are recognised as having a significant effect on the performance of state-of-the-art UGVs (Kelly et al., 2006; Thrun, 2006), and general perception problems caused by sensor data interpretation errors are still largely unsolved for robotic perception systems (Underwood, 2009; Urmson et al., 2008; Leonard et al., 2008).

This section discusses related work concerned with improving perception for robots in low-visibility conditions with an emphasis on camera-based applications. First, developments in sensor hardware and resilient

[3] as a combination of radiation and reflection

perception algorithms using a single sensor modality are presented. We then focus on methods to evaluate the quality of image data and their role in detecting low-visibility conditions. Finally, we discuss resilient perception systems that make use of multiple sensor modalities to avoid common-mode failures.

2.1 Sensor Hardware Development

A variety of hardware options can aid perception in low-visibility conditions. Hardware design is often engineered into sensing systems to provide maximum visibility depending on the prevalent conditions. Most off-the-shelf cameras will adapt to low illumination with automated shutter time and gain control to improve the brightness and dynamic range of the output image, respectively. Increasing the shutter time and gain of a camera may improve what is visible in darker regions of an image but can also lead to blurring, saturation and higher noise. A solution to improve visibility, particularly at night, is to employ onboard lights to actively illuminate the environment (Dubbelman et al., 2007). Alternatively, night vision cameras work by enhancing the acquired images by *photon boosting*, which amplifies available light to an observable level. Another solution is to operate in the infrared spectrum thereby sensing the heat of objects, which is less dependent on the illumination of the environment. (Owens and Matthies, 1999) have tested a range of different thermal based and image intensifier vision systems to aid stereo navigation of UGVs at night with some success. Similarly, hyperspectral cameras have been considered to provide improved visibility at night and in smoke (Fay et al., 2000; Toet, 2003). These advances in sensor hardware increase the operating range of available sensors, which contributes to making perception algorithms more resilient to low-visibility conditions. However, limits to the operating range of any sensor remain, while the realm of environmental situations still is unbounded. This means inappropriate data (acquired while outside of the sensor operating range) *will* be acquired eventually. Therefore, resilient perception systems need to be able to handle the inappropriate data to mitigate failures.

2.2 Resilient Perception Algorithms

Many techniques have been developed to improve image-based perception algorithms in low-visibility conditions. These methods can be divided into two families; 1) the first category applies advanced image processing techniques to directly remove or compensate for the condition, 2) the second category focusses on endowing the perception algorithm with an invariance to changes in visibility.

The quality of image data can be enhanced for various low-visibility conditions, although this quality is

rarely explicitly evaluated. For example, (Ferwerda et al., 1996; Tumblin and Rushmeier, 1993; Ward, 1994) enhance the visibility of poorly-illuminated objects by increasing the contrast in images. (Narasimhan and Nayar, 2003; Nayar and Narasimhan, 1999; Tan, 2008) have developed physical models of how visible light is attenuated in fog in order to filter and reverse the effect. The compensation of shadow areas in images has been studied for many years (Finlayson et al., 2006; Gu et al., 2005; Scanlan et al., 1990). These compensation methods require computationally expensive and complicated models, can often introduce other artefacts, and typically require a-priori knowledge that the specific condition is present.

To the best of our knowledge, such models are not available for a range of other low-visibility conditions, in particular for localised and variable visibility phenomena such as smoke clouds. Besides, it would be difficult to specifically anticipate and exhaustively model all environmental phenomena that an outdoor vehicle might experience. Additionally, while the methods discussed above are designed to enhance sensing data, they do not consider whether the output data are appropriate for use in a subsequent perception algorithm.

A generic approach that avoids this problem is to create perception algorithms that are capable of adapting to variations in visibility. Scale Invariant Feature Transform (SIFT) descriptors are robust to moderate illumination variations (Lowe, 2004). Other feature descriptors have been developed specifically for robustness to changes in intensity and apparent colour caused by the illumination of the scene (Zabih and Woodfill, 1996; Moreno-Noguer, 2011; Kobayashi and Kameyama, 2010). Typically, illumination changes are modelled by a transformation of pixel intensities (region normalisation and invariant steerable filters (Mikolajczyk and Schmid, 2005), isotropic (Toth et al., 2000), monotonic (Zabih and Woodfill, 1996)) to normalise the variations of gradient magnitude over time. More abrupt changes in lighting have been modelled with hue and chromaticity (Scandaliaris and Sanfeliu, 2010). (Yu et al., 2012) discuss the maximum illumination change before feature-based matching will fail. To mitigate the issues with feature extraction robustness in some challenging outdoor conditions, (Nuske et al., 2009) propose to use prior knowledge of the environment in the form of an edge map, and show this can allow for robust visual localisation in extreme lighting and rain conditions.

Some of the most difficult situations for any perception technique occur when a sensor is employed outside its expected operating conditions, making it unable to perceive the environment in a useful way in the first place. Although methods discussed in this section can improve the robustness of algorithms in low-visibility conditions, they will still fail when events occur that are not expected by the model used to interpret the sensing data. The next section discusses methods to evaluate the quality of the data (specifically image data) prior to interpretation to identify if it is appropriate for the model.

2.3 Image Data Quality Evaluation

In the telecommunication literature, methods have been proposed to explicitly evaluate the *quality* of image data for human consumption (Winkler, 2005; Wang and Bovik, 2006; (ITU-T), 1999). Previous work by the authors (Brunner et al., 2009; Brunner et al., 2011b) proposed a study of many of these image quality metrics and their relevance to outdoor robotic perception specifically for visual and IR imagery. Many quality metrics are designed to capture the errors caused by compression and transmission and are typically tailored to the human vision system. However, pictures that are colourful, well-lit, sharp and have high contrasts are considered more attractive to humans, and these characteristics are also positive for a number of UGV perception applications. The study found that many of the characteristics captured by existing quality metrics are similarly caused by sensing in low-visibility conditions. A number of limitations of these metrics were also identified. For example, many were highly dependent on the type of background perceived, or were inappropriately relying on Gaussian distributions assumptions. To address these issues, a novel quality metric named Spatial Entropy (SE) was introduced (Brunner and Peynot, 2010). SE is defined as the entropy of distribution of intensities in a Sobel-filtered image. This was shown to be a good indicator of image quality, especially in low-visibility conditions (Brunner et al., 2011b) and to be appropriate to evaluate both visual and infrared data.

2.4 Multiple Modalities of Sensing for Resilient Perception

Combining multiple modalities of sensing for perception is beneficial for two reasons; 1) the additional information enables enhanced discrimination, and, 2) the redundant information provides robustness, as the perception system is less prone to common-mode failure (Underwood, 2009). Employing multiple modalities of sensing has long been identified as crucial for many systems that are affected by low-visibility conditions (Foyle et al., 1993; Perbet et al., 1993). Different modalities provide redundancy in variable environmental conditions. In a military visualisation context, (Martinsen et al., 2008) concludes that broad spectrum sensing (from visual to long-wave IR) is required for long-term mission success in varying weather. Enhanced Vision Systems (EVS) (Hines et al., 2005) combine vision and IR sensing (with manual selection) to assist pilots in fog. Typically these systems rely on the operator to choose the best sensor to use at any time.

Visual and IR images have long been fused to aid in detection and tracking (Nandhakumar and Aggarwal, 1988; Toet et al., 1989) because relevant objects that are indistinguishable from the background in one

sensing modality can have higher contrast in another (Ardeshir Goshtasby and Nikolov, 2007). (Lanir et al., 2006) proposes a review of multispectral image fusion methods, including averaging, edge enhancement, false colour and principal component analysis. Fusion has been used to enhance the available information in poor visibility, such as in the presence of smoke (Lewis et al., 2007; Waxman et al., 1998; Sadjadi, 2005), heat (Lewis et al., 2007), or at night (Lewis et al., 2007; Waxman et al., 1998; Waxman et al., 1997). In all cases, image quality was subjectively considered to have been improved by fusion, since specific features were more visible in the final image than in either of the original images separately. However, in fusing the images, information was lost, the low-visibility condition (such as a smoke cloud) remained apparent, and inconsistent data from different sources was merged. The quality of these fused images with regards to perception applications was not evaluated.

Modern UGV robotic platforms are often equipped with multiple sensing modalities such as cameras, lasers and radars (Urmson et al., 2008; Leonard et al., 2008; Thrun et al., 2007). However, these are primarily adopted for their discriminative capabilities as opposed to redundancy. In (Thrun et al., 2007) radar is used to identify obstacles, the visual camera is used to detect roads in the distance and the laser provides a near field map. (Roberts et al., 2008) demonstrate the benefits of using redundant information in laser and vision for robust localisation. Laser-based and visual-based localisation filters are run in parallel and an arbitration step decides which localisation estimate the vehicle should use. Multiple-modality Visual-SLAM was recently used for night and day localisation (Maddern and Vidas, 2012), showing that the combination of thermal IR and vision was required for robust recognition.

A growing field of work exploits the redundancy of multiple modality sensor data to filter appropriate data. In (Carlson and Murphy, 2005), a *conflict metric* is used for sonar and laser data to detect when the use of either sensor would lead to an incorrect interpretation of the environment. The system adapts by switching to the more appropriate sensor. (Soleimanpour et al., 2008) discuss using the *compatibility* of sensor data prior to ensuring that only compatible information is combined but do not discuss specifically how to measure compatibility between different modalities of data. A multisensor indoor robot for operation in flame and smoke conditions is implemented in (Luo and Su, 2003), where four different modes of proximity sensor are used to confirm each other and to treat any incompatible sensor as having failed. (Peynot et al., 2009; Castro and Peynot, 2012) consider the range disparity between overlapping radar and laser data to identify inconsistencies and filter laser data that are affected by dust for an improved map of the environment. (Peynot and Kassir, 2010; Borges et al., 2010) compare edges in visual image data and overlapping 2D and 3D laser data respectively to determine the likelihood that the sensor data between the two modalities are

consistent.

In (Brunner et al., 2011b), the authors showed a preliminary example of the use of visual and infrared cameras for visual localisation in the presence of smoke. First, the relative motions of the platform were estimated using visual images and infrared images separately. Because of the presence of smoke, localisation using only the visual camera failed, while the localisation using only IR was successful but relatively inaccurate. A better trajectory was then obtained offline by using relative motion estimates from one single source of information at any time (only IR when smoke was present, and only visual data otherwise). The results obtained on this single trajectory suggested that pre-selection of appropriate image data might improve the performance of a camera-based localisation system.

In this paper, we make the most of the two types of available sensors by combining both types of data in the localisation algorithm, and extend over the work presented in (Brunner et al., 2011a). We propose to use data quality evaluation on both sensor modalities, both globally (on the entire image) and locally (on sections of each image), to decide which data should be combined and used in the localisation algorithm. Localisation techniques are compared in multiple experiments using 1) a single sensing modality, 2) a full combination of both sensing modalities and 3) a selective combination of both sensing modalities based on the image quality evaluation. The proposed approach is shown to be resilient to a large variety of low-visibility conditions, including at dusk, at night, and in the presence of smoke, high heat and fire.

3 Image Quality Evaluation and Automatic Data Selection

Evaluating the quality of raw sensor data prior to interpretation may enable anticipation of the performance of a perception system. As suggested in (Brunner and Peynot, 2010), errors can be mitigated by discarding inappropriate sensor data. To achieve this, the link between the data quality and the performance of the perception application first needs to be established. This section discusses this link for a perception system using camera image data. Section 3.1 discusses image data requirements for perception systems. Section 3.2 presents the proposed framework to evaluate the quality of sensor data and make a selection of the appropriate data for a perception system.

3.1 Data Requirements

Previous work by the authors (Brunner et al., 2011b), discussed the link between data quality and the performance of a perception system based on cameras. This work showed that *poor quality* or inappropriate data will provide little useful information for the perception application or, in the worst case, may corrupt the solution by providing misleading information. The study considered two main families of image-based perception techniques; feature-based methods (FBMs) and area-based methods (ABMs) (Zitova and Flusser, 2003). A range of quality metrics (see Section 2) were evaluated to identify appropriate and inappropriate data in the context of these two families of perception techniques.

In this paper, Visual-SLAM is the proposed perception application. Our Visual-SLAM approach is based on a FBM, as it first detects SIFT features and then matches these features between consecutive images to estimate changes in the robot pose. Good quality images for FBMs contain lots of unique structure that is easily distinguishable spatially and temporally. On the other hand, inappropriate or poor-quality image data contain very few unique and invariant features. Besides, in these situations, the SIFT features that are detected are likely to be incorrectly matched (Brunner and Peynot, 2010), leading to an incorrect pose estimation. Low-visibility conditions often cause poor quality data for FBMs because the background environment is often lost or significantly attenuated, thereby reducing the discrimination between features. An example is a visual image in dense fog or at night.

Spatial Entropy of a gray-scale image, is defined as the entropy of the distribution of intensities in the Sobel-filtered image, $Sob(I)$ (Brunner et al., 2011b). $Sob(I)$ is composed of pixels with intensities from a discrete set of possible values $i \in A_i$. The probability of observing any particular intensity value i in $Sob(I)$ is given by $P(i)$. SE of an image is defined as the entropy of the ensemble of intensities in the image (Mackay, 2007):

$$SE(I) = \sum_i P(i) log_2 \frac{1}{P(i)} \tag{1}$$

and is expressed in average bits of information per observation (i.e. per image or sub-image).

SE quantifies the amount of structural information contained in an image and is therefore a relevant measure of the quality of an image for a large family of perception applications. Images with high SE values contain a large variety of edges, which occur naturally as a separator between distinct objects in a scene. Hence the ability to detect edges in perceptual data correlates generally to discriminative potential. Such images are considered as good quality input for FBMs. Low-visibility conditions are shown to obscure and reduce the

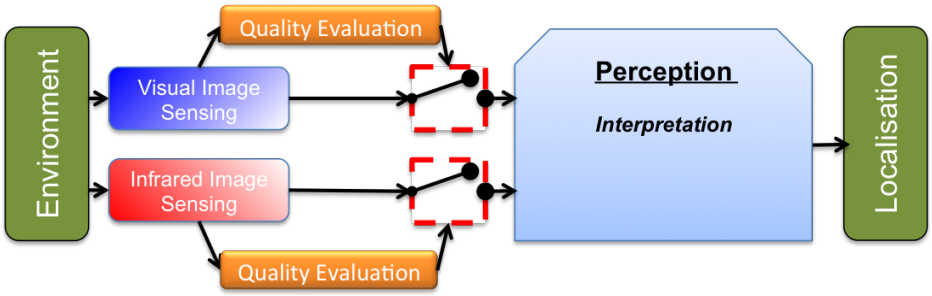


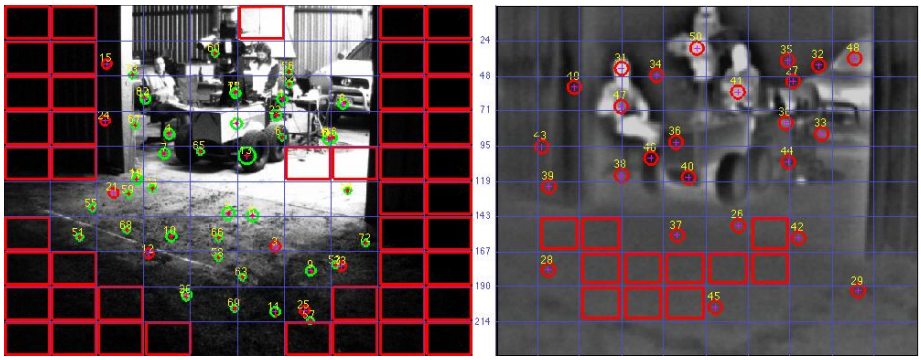
Figure 6: Proposed framework for multimodal camera perception with pre-selection of image data (IR and visual), based on quality evaluation.

contrast of structure in the environment, leading to low values of SE. The absolute value of Spatial Entropy is dependent on the nature of the background of the observed scene. Therefore, to reduce this dependency, we also use the derivative of SE (dSE) (Brunner et al., 2011b). High values of dSE indicate that the amount and uniqueness of structure in the environment is suddenly changing, which may indicate a quality loss for FBMs. For example, this typically happens when a smoke cloud appears between a visual camera and the background of the scene (Fig. 3, top). SE and dSE are useful metrics to evaluate the quality of images used for FBM-based field robotic perception applications as they respond to obscurants in the environment and lack of features in the scene. In this paper they are used to anticipate failures in camera-based localisation using both visual and IR images.

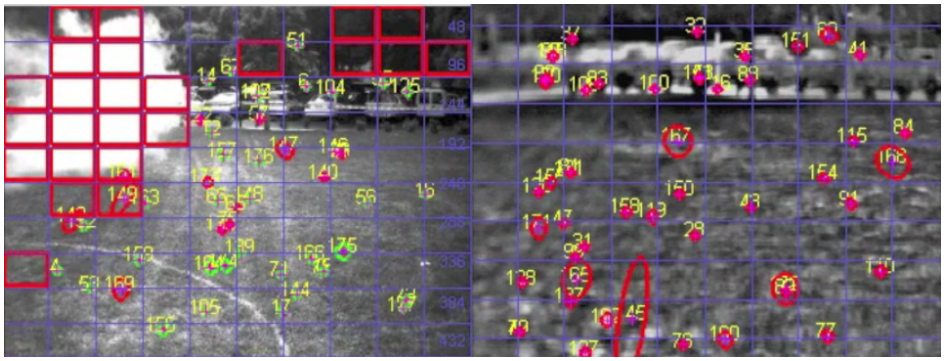
3.2 Strategy for Data Selection

In (Brunner et al., 2011b), visual image quality was evaluated, and only high quality images were used for UGV localisation in smoke. The IR camera was assumed to be providing good quality data at all times since it is not affected by smoke. In this paper, we propose to adopt a similar strategy, as shown in Fig. 6. However, the quality evaluation and consequent selection of the image data are performed for both sensors to enable resilient perception in a larger set of environmental conditions. Additionally, we enhance the previous approach by using a *local* analysis of data quality: the images are partitioned and then each sub-image is evaluated independently. Fig. 7 shows examples of this local quality evaluation on pairs of visual and IR images, taken at night, and in the presence of smoke.

In the process of data pre-selection, only local regions of data (i.e. sub-images) are discarded. The rest of the images will still be used by the localisation algorithm. This is particularly useful when low-visibility conditions only affect a portion of an image (see the left image of Fig. 7(b), for example). The availability of larger amounts of discriminative data will usually lead to higher performance of perception algorithms, as will be shown in Section 7. This also highly reduces the chance that no data are available to the perception



(a) Nighttime with artificial lighting



(b) Daytime with smoke cloud

Figure 7: Visual (left) and IR (right) images, at night with some artificial lighting (top), and during the day in the presence of a smoke cloud (bottom). The bold red squares indicate regions of poor quality data as identified using the SE metric. The pink crosses and corresponding numbers show SIFT features that are matched within the Visual-SLAM algorithm. The red and green ellipses illustrate the uncertainties of the landmark positions.

algorithm at a given time, an event that may be more likely if entire images are discarded. In addition, local analysis allows for better spatial and temporal responsiveness to low quality data within images. Inappropriate data can be identified earlier than what is possible when evaluating entire images.

4 Multiple-Modality Camera Localisation

A block diagram of the complete perception system proposed in this paper, including data quality evaluation and selection of visual and IR images prior to the vehicle localisation estimation, is shown in Fig. 8. This section describes the Visual-SLAM algorithm based on visual and infrared cameras. Section 4.1 discusses required background to visual localisation. Sections 4.2 and 4.3 introduce the adopted localisation algorithm and its implementation. Section 4.4 describes how the performance of the localisation estimation is evaluated in this paper. Finally, Section 4.5 presents the different strategies of image data combination for localisation that are implemented and evaluated in this paper.

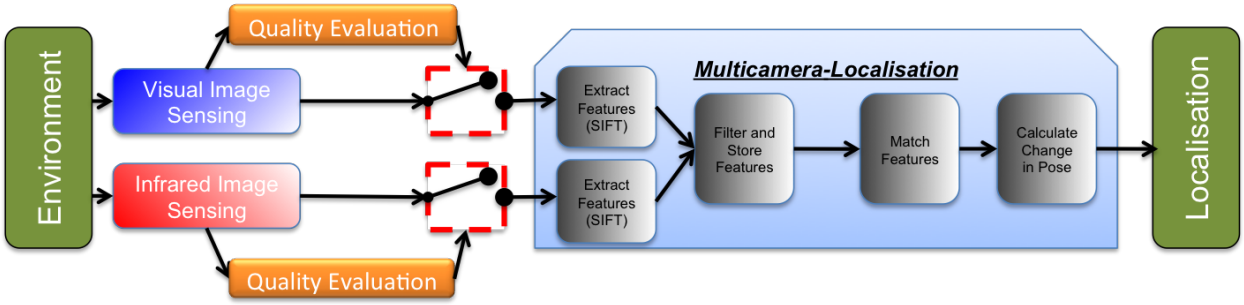


Figure 8: Proposed framework for resilient camera localisation using visual and IR cameras. A step of evaluation and selection of the image data occurs prior to localisation estimation (blue box). The heterogenous sensor data are not fused directly but independent features are stored in a common map.

4.1 Background

Monocular-SLAM is the problem of concurrently estimating the structure of the surrounding world (the *map*) while getting *localised* in it, using a single projective camera as the only exteroceptive sensor. This problem was successfully solved, using a filtered solution, with the work of Davison (Davison, 2003) and the inverse-depth landmark parametrisation (IDP) in (Montiel et al., 2006).

Multicamera-SLAM involves fusing information from different cameras mounted on the same vehicle. When the cameras have similar properties such as spectral range, field of view or distortion, fusing their information requires the data-association problem to be solved. The common way to solve this problem is to match visual features in the image space. From the SLAM point of view, matching features can be done as a preprocessing step to initialise $3D$ points in the map, e.g. Stereo-SLAM (Jung and Lacroix, 2003), or as a data-association step to update landmarks already contained in the map, e.g. Bicam-SLAM (Solà et al., 2008), or even a combination of both methods (Paz et al., 2008). On the other hand, matching features between corresponding images is not always possible for multiple modalities of sensing (i.e. with very different properties such as for visual and IR cameras). Nevertheless, these cameras may still contribute independently to the vehicle's localisation and share the same map.

4.2 Algorithm

The core algorithm of this application is a landmark-based EKF-SLAM with inverse-depth parametrisation (IDP) based on (Solà et al., 2008). Let us consider a visual and an IR camera. The state-space vector is given by:

$$X^\top = \left[\mathcal{R}^\top \quad {\mathcal{L}_{vis_1}}^\top \quad \dots \quad {\mathcal{L}_{vis_N}}^\top \quad {\mathcal{L}_{ir_1}}^\top \quad \dots \quad {\mathcal{L}_{ir_M}}^\top\right], \tag{2}$$

where $\mathcal{R}$ represents the current robot pose, $\mathcal{L}_{vis_i}$ is the i^{th} landmark of the visual camera and $\mathcal{L}_{ir_j}$ is the j^{th} landmark of the IR camera, both parametrised as IDP. The cameras are intrinsically and extrinsically calibrated, therefore both independently update the robot pose. Note that the same algorithm is applied when using a single camera, but IDP landmarks are extracted from only one image modality.

The cameras are the only sensors used to estimate the robot trajectories, for this reason, a 6-DOF[4] constant velocity model is adopted to predict the motion (as defined in (Solà et al., 2008)). The predicted robot pose is given by $\mathcal{R}^+ = \mathbf{f}(\mathbf{p}, \mathbf{q}, \mathbf{v}, \omega, a, \alpha, \Delta t)$, where $\mathbf{p}$ is the robot position, $\mathbf{q}$ the orientation quaternion (systematically linearised), and $\mathbf{v}$ and ω are the linear and angular velocities respectively. At each time step, perturbations $a, \alpha \sim \mathcal{N}(0; \sigma_{\mathbf{v}}^2, \sigma_{\omega}^2)$ add variances to the linear and angular velocities proportionally to the elapsed time Δt.

A common issue with Monocular-SLAM is that motion estimates and map structure can only be recovered up to scale, due to the projective nature of a single camera (Davison, 2003). The solution obtained using two different cameras, with no landmarks in common and no aid of other sensors, is similarly subject to scale since there is no direct data association between the features of the two cameras.

The IDP is encoded by the direction vector from the current camera position $\mathbf{c}_0$ to the observed point ℓ, with just elevation and azimuth angles (ε, α) of the observed optical ray joining $\mathbf{c}_0$ to ℓ. When these angles are appended with the inverse of the distance $\rho = 1/d$, the result is a $3D$ point in modified polar coordinates, $(\varepsilon, \alpha, 1/d)$. Adding the current camera position $\mathbf{c}_0$ as an anchor to improve the linearity leads to the $6D$-vector $\mathcal{L}_{cam} = \begin{bmatrix} \mathbf{c}_0 & \varepsilon & \alpha & \rho \end{bmatrix}^\top$. Note that during the initialisation ρ must be provided as a prior.

4.3 Feature Extraction and Map Management

In the proposed framework, different sensing modalities (i.e. IR and visual cameras) contribute independently to the overall vehicle localisation because data are not being associated between them. This is done by sharing the same map in a Bicam-SLAM algorithm. As images become available from each sensor, sparse interest points (*features*) are extracted using SIFT detectors and matched using SIFT descriptors (Lowe, 2004) using the 8-bit gray-scale images from each sensor. SIFT features between visual and IR images are not matched in the process, as their appearance descriptor is very different. These features parametrised as IDPs are stored in the EKF map. Although the input processes are independent, the features from each camera become correlated, meaning that an update from an IR image will correct the mapped locations of visual features.

[4]Degrees of freedom

To ensure approximately uniform sampling, each image is divided in a regular grid and features are randomly selected within this grid. In order to keep the computational complexity bounded, the algorithm keeps a maximum number of landmarks. The feature set of each sensor modality has a unique identifier and is maintained separately from other modalities, although all features are still stored in the same map. Therefore, each sensor always maintains a fixed number of features. As a given modality image becomes available, the mapped features are used to correct the position of all the landmarks in the map together with the current 6D robot pose. The scale is recovered using the velocity information from an onboard IMU outside the filter.

As proposed in (Davison, 2005), the Gaussian expectation of the visible mapped points is used to reject outliers in the sensor frame. The Gaussian expectation is defined as the ellipse $\mathcal{E} = \mathcal{N}(u - e; E)$, with u being the measured pixel position, and with mean e and covariance matrix E of expected point position in the image. $\mathcal{E}$ is usually gated at 3σ, giving place to an elliptic region in the image where the landmark must project with 99% probability. Note that there is no need to apply expensive outlier rejection algorithms, such as RANSAC (Fischler and Bolles, 1981), because the Gaussian expectation already accounts for most of the wrong SIFT matches.

Unstable and inconsistent landmarks are deleted from the map to avoid map overpopulation and corruption. Unstable refers to landmarks that are expected but not observed, and inconsistent refers to those landmarks that are observed but lie outside the 3σ bound defined by $\mathcal{E}$. Based on the ratio of unstable and inconsistent landmarks, the decision of a landmark being deleted is taken. To make the algorithm faster and because the interest is in the localisation and not in the mapping, landmarks that have not been observed for a certain time are also deleted. In consequence, loop closures are unlikely to happen automatically and there is no strategy to explicitly enforce them.

4.4 Evaluation of Localisation Performance

The performance of a localisation application can be evaluated by comparing the estimated trajectory to a *reference* trajectory. The accuracy of an estimated trajectory is given by the difference in pose δP_{est} with the reference δP_{ref}, $\delta P = ||\delta P_{est} - \delta P_{ref}||$. However, an error at the start of the trajectory will cause a large δP at the end, even if the estimation was locally accurate for the remainder of the experiment. Therefore, in this work, the evaluation is performed by comparing the reference and estimated trajectories locally as suggested in (Burgard et al., 2009). The relative local difference in pose, $\delta P_{dt} = ||\delta P_{est(t+dt)} - \delta P_{ref(t+dt)}||$, between the two trajectories is calculated for each dt along the trajectory (Fig. 9). The mean of local differences

provides a measurement (γ) of the accuracy of that trajectory estimation:

$$\gamma = \frac{1}{N} \sum_{t_1}^{t_N} \delta P_{dt} \,, \tag{3}$$

where N is the number of local differences, δP_{dt}, for the duration of an experiment. By considering relative changes in pose between the reference and estimated trajectories, we mitigate the problem of accumulative errors and can also quantify the performance at specific times of the run. The variance and distribution of δP_{dt} for a run can also provide insight into the overall performance of the localisation application.

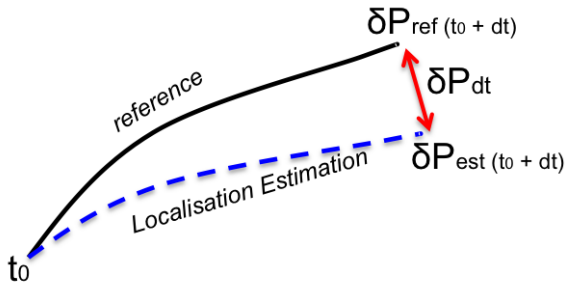


Figure 9: Local errors in pose (δP_{dt}) are evaluated by computing the relative changes in pose between the reference and the localisation estimation over a short period of time dt (Burgard et al., 2009).

It should be noted that it is not always suitable to compare the performance of localisation between different experiments. Different environments (e.g. indoor, urban, unstructured), the duration of the experiment, the velocity and the pattern of motion of the robot (e.g. travelling in straight line, turning, rolling) will have an effect on the overall performance of the localisation. However, it is appropriate to compare the performance of different localisation methods during the same experiment.

4.5 Localisation Methods using Different Sources of Data

In this work, both IR and visual cameras are sources of image data for the localisation application. The experiments in Section 7 evaluate and compare a range of localisation methods considered in this paper. All methods are based on the Visual-SLAM algorithm described in Sections 4.2 and 4.3 but use different sources of data, or different strategies for combining those sources. These methods are defined and named as follows:

- *Visual Camera Localisation* (*Vis.Loc*) estimates the vehicle pose using one visual camera only.
- *Infrared Camera Localisation* (*IR.Loc*) estimates the vehicle pose using one IR camera only.

- *Multiple-Modality Camera Localisation* (*MM.Loc*) combines the landmarks extracted from both sources, without any data quality evaluation or pre-selection.

- *Selective-Multiple-Modality Camera Localisation* (*SMM.Loc*) also combines data from both sensing modalities, but only after evaluating image quality and rejecting poor-quality images accordingly (see Fig. 8). Data quality evaluation and pre-selection are made on entire images (i.e. *globally* w.r.t. the image).

- *Locally-Selective-Multiple-Modality Camera Localisation* (*LSMM.Loc*) combines data from both sensing modalities after data quality evaluation and pre-selection as well, but evaluation and selection are achieved on sub-images (i.e. *local* regions within the image). In this paper, each original input image was partitioned into 10×10 sub-images (see Section 3.2).

Table 1 summarises the differences between the methods.

Table 1: Localisation using different strategies of image data combination.

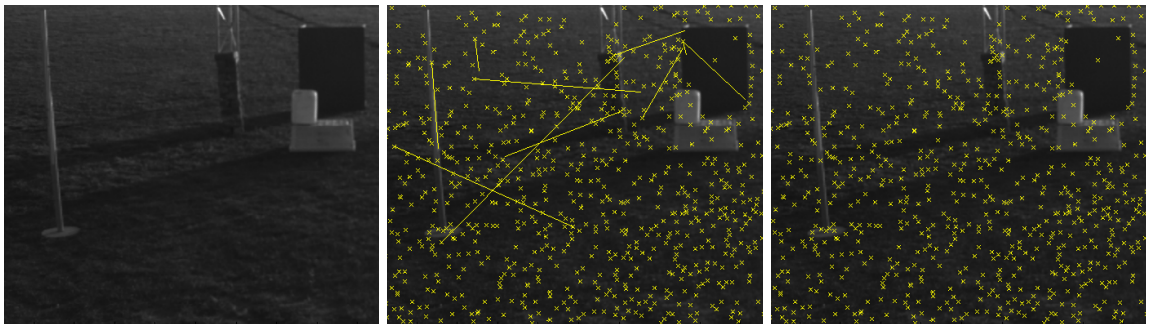
Method	Visual	Infrared	Data Selection Method
Vis.Loc	✔	X	X
IR.Loc	X	✔	X
MM.Loc	✔	✔	X
SMM.Loc	✔	✔	✔(Global)
LSMM.Loc	✔	✔	✔(Local)

The next section further specifies the implementation of image data quality evaluation for multi-camera localisation estimation.

5 Automatic Data Selection to Mitigate Localisation Errors

The main source of localisation error in the Visual-SLAM algorithm used in this work lies in the extraction, selection and matching of features in the input images. Any errors in SIFT matching will propagate through the system (see Fig. 8) and are most likely to have an impact on the accuracy of the estimated trajectory. Therefore, anticipating and mitigating these errors is a fundamental requirement for resilience.

While internal mechanisms are used to find coherent matches, eliminate outliers and thereby reduce the error (as described in Section 4), these methods are not always sufficient, especially in low-visibility conditions (Brunner et al., 2011a). Previous work (Brunner and Peynot, 2010) suggested that the SE of an image could be used to anticipate errors in SIFT matching, particularly in low-visibility conditions.



(a) Original Visual Image (b) SIFT matches in consecutive images. (c) MAPSAC-filtered matches.

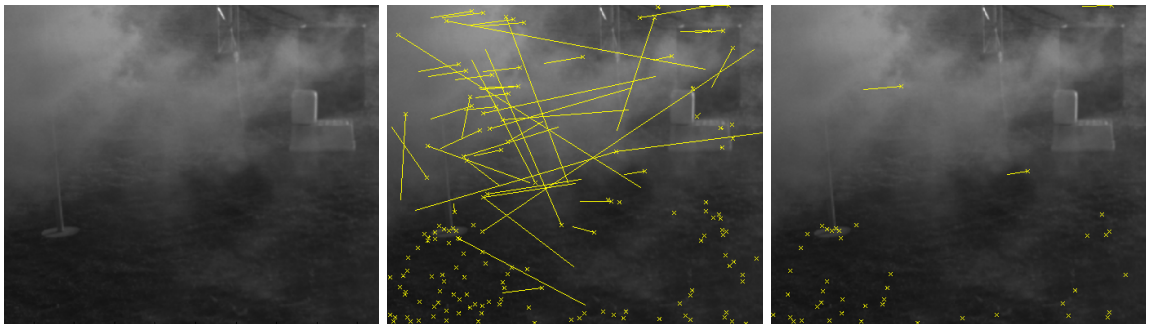
Figure 10: SIFT matching between visual images in clear conditions. All SIFT features are represented by yellow crosses. The SIFT matches between two consecutive images are represented by a yellow line that connects the position of the match with the position of the original feature in the previous image (i.e. long yellow lines indicate large matching errors, since the camera is static).

In Section 5.1, we evaluate the performance of feature matching and state-of-the-art outlier rejection techniques in low-visibility conditions by using stationary cameras in a controlled static test environment. Section 5.2 discusses the direct link between image quality and SIFT matching errors. Subsequently, we show that discarding low-quality data prior to feature matching mitigates matching errors.

5.1 SIFT-Matching Errors in Low-Visibility Conditions

To evaluate experimentally the performance of SIFT matching in low-visibility conditions, we used data from (Peynot et al., 2010). Stationary cameras were set to view a static test environment. Environmental conditions were initially clear, meaning that visibility was considered high for both the visual and infrared cameras. This corresponds to daytime conditions with no environmental phenomena such as fog, smoke or airborne dust. An obscurant was then introduced that reduced the visibility of one of the sensors. Such obscurants included smoke clouds (obscuring the visual camera sensing), and hot air and flames (affecting the IR camera sensing). For clarity, a representative data set using a visual camera in the presence of smoke conditions is used for the remainder of this section. Fig 10(a) and Fig. 11(a) are representative images of clear and smoke conditions, respectively.

We used the implementation of VLFeat library (Vedaldi and Fulkerson, 2008) for SIFT extraction and matching. In the static test environment, ground truth is available since correctly matched features are in the same pixel location in consecutive images. Therefore, matching errors were calculated by measuring the absolute distance between matched features measured in pixels. The average of all the individual matching errors in an image provides a single value of *matching error* per image.



(a) Original Visual Image (b) SIFT matches in consecutive images. (c) MAPSAC-filtered matches.

Figure 11: SIFT matching between visual images in smoke conditions. All SIFT features are represented by yellow crosses. The SIFT matches between two consecutive images are represented by a yellow line that connects the position of the match with the position of the original feature in the previous image (i.e. long yellow lines indicate large matching errors, since the camera is static).

Images in smoke conditions (see Fig. 11(b)) illustrate that the presence of low-visibility conditions cause many wrong associations in the SIFT-matching algorithm. The top row of Table 2 shows the average matching error for the whole data set and the average matching error for the data set divided into periods of clear and for periods of smoke. Smoke causes a major increase in the mean matching error (46.3 pixels compared to 1.71 in clear conditions).

Table 2: Mean Matching Error (in pixels) over time for the representative smoke data set after applying different techniques of outlier rejection to the output of the SIFT matching algorithm. The error is also calculated separately for periods of clear and smoke conditions.

Outlier Rejection Method	Mean Error in pixels (variance)		
	Total Data Set	Clear Conditions Only	Smoke Conditions Only
None	33.3 (1044)	1.71 (0.3)	46.3 (900)
Gaussian Expectation	0.38 (0.03)	0.36 (0.004)	0.41 (0.06)
RANSAC	1.93 (9.20)	0.42 (0.0008)	2.56 (11.7)
MSAC	1.81 (8.82)	0.42 (0.0007)	2.38 (11.4)
MAPSAC	1.68 (11.76)	0.32 (0.0015)	2.25 (12.6)

In the camera localisation technique implemented in this paper, outliers are rejected by finding inconsistent matches using the Gaussian expectation of visible mapped points (see Section 4.3 for details). This method requires the 3D position of the features and a prediction of the camera motion. In this test, since the camera is static, the velocity in the motion model is set to zero and the features are initialised at $\rho = 0.25$ (see Section 4.2); therefore, the features are projected into the image at the same position. Table 2 shows that in clear conditions, the Gaussian expectation eliminated most of the outliers and improved on SIFT matching alone with a mean matching error of 0.36 pixel. The maximum matching error during clear conditions was 0.66 pixel. During times of smoke, this consistency check also improved on SIFT matching alone with a mean matching error of 0.41 pixel. However, in those conditions, the maximum matching error was 5.1 pixels,

and there were 24 images where the average matching error was greater than 2 pixels, indicating that the Gaussian expectation can fail if a significant amount of smoke is present in the environment. Note that in some extreme cases of smoke the Gaussian expectation eliminates all matches. These cases were not included in the table since no error can be computed.

In the literature, other outlier rejection techniques such as random sample consensus, i.e. RANSAC (Fischler and Bolles, 1981) or more recent variants, are often used to eliminate wrong associations such as those produced by SIFT matching. We evaluated the performance of RANSAC, MSAC (Torr and Murray, 1997) and MAPSAC (Torr, 2002) with the fundamental matrix model, calculated with the normalised 8-point algorithm (Hartley and Zisserman, 2004), in low-visibility conditions. For conciseness, in the remainder of the section we only show the best results obtained when varying the parameters of each technique.

A thorough search ($10,000$ iterations) was run for each set of matched SIFT features to obtain a coherent subset of matches from the stationary data sets. Figs 10(b) and 10(c) show a representative example where MAPSAC successfully removed the incorrect matches in clear conditions. Fig. 11(c) shows matched features kept after applying MAPSAC in smoke conditions. In this representative image, although MAPSAC eliminated many outliers, some incorrectly associated features were retained as inliers. Table 2 shows the average matching error obtained with RANSAC, MSAC and MAPSAC in clear and smoke conditions. While these performed very well in clear conditions, at times of smoke significant errors can be observed. Note that the best performance of the RANSAC variants were obtained with MAPSAC.

These outlier rejection techniques rely on the availability of a sufficient proportion of correct matches to obtain coherence in the data. In low-visibility conditions, coherent groups of matches can be found from biased subsections out of the large number of incorrect matches. Therefore, state-of-the-art outlier rejection techniques are unable to sufficiently mitigate the effect of low-visibility conditions on feature matching. Consequently, anticipating situations with large SIFT-matching errors is beneficial to obtain a SIFT-based localisation resilient to low-visibility conditions. This can be done using SE to evaluate the quality of the image. The next section specifically establishes the link between SE and SIFT-matching error.

5.2 Spatial Entropy Evaluation to Anticipate SIFT-Matching Errors

In this section, we identify thresholds for the quality metrics that are appropriate to discard low-quality data prior to the localisation algorithm, allowing for the data pre-selection mechanism proposed in this paper. The top plot in Fig. 12 shows the evolution of Spatial Entropy over time in the smoke data set used in the

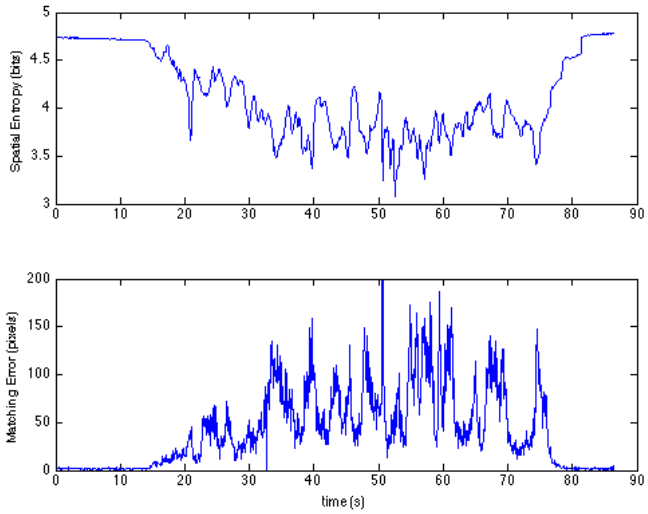


Figure 12: Static visual camera data in the presence of smoke. Smoke is present between $t = 15s$ and $t = 77s$. Top: the evolution of Spatial Entropy over time. Bottom: the corresponding average matching error (pixels) of all the SIFT matches found between consecutive images.

previous section, while the middle plot shows the SIFT matching errors. The data set is initially clear for about $15s$ (220 total images), after which smoke is introduced and is highly variable for the next $77s$ (930 total images). Note that at $15s$, the value of SE drops considerably as the presence of smoke begins to be significant and there is a corresponding increase in the SIFT matching error.

Supervised learning was used to quantify the link between the SE of images and the matching error, using the controlled static data sets for training. Images were labelled as providing poor quality data when the average matching error was found to be higher than the error in clear conditions plus two standard deviations. On the contrary, when the matching error was lower than this, images were labelled as providing good quality data.

Receiver operating characteristic (ROC) curves (Green and Swets, 1989) were generated by varying the value of SE and dSE for the stationary data sets to obtain thresholds that provided the best ratio of true positive rate (TPR) to false positive rate (FPR). This method was performed for a number of stationary data sets to determine the values that would be used to discard data in operation. The ROC curves obtained by varying SE for the two data sets used in Section 5.1 are shown in Fig. 13. The curves show that evaluating SE allows us to anticipate SIFT-matching errors. The thresholds obtained are: $T^{SE}_{Vis} = 4.13bits$ on SE and $T^{dSE}_{Vis} = 0.41bits$ on dSE for the visual images, $T^{SE}_{IR} = 4.60bits$ on SE and $T^{dSE}_{IR} = 0.35bits$ on dSE for the IR images. We used these thresholds to decide whether images should be used in the localisation algorithm. Images with values of SE below T^{SE} and dSE above T^{dSE} were considered to be low quality data, likely to generate errors. Therefore, they were discarded before SIFT extraction and matching. This decision rule

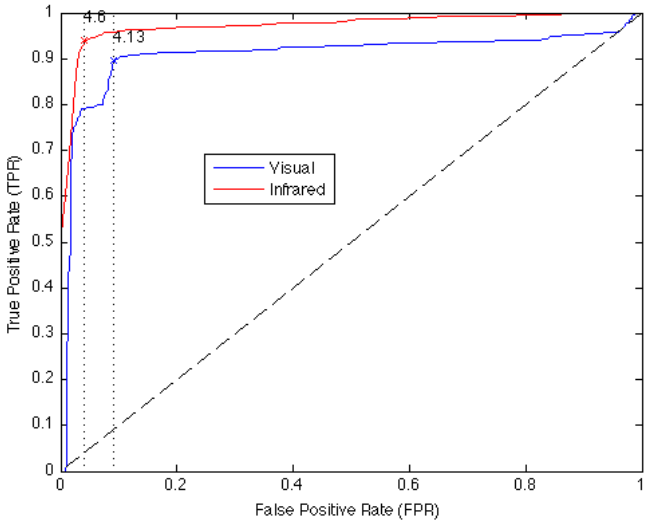


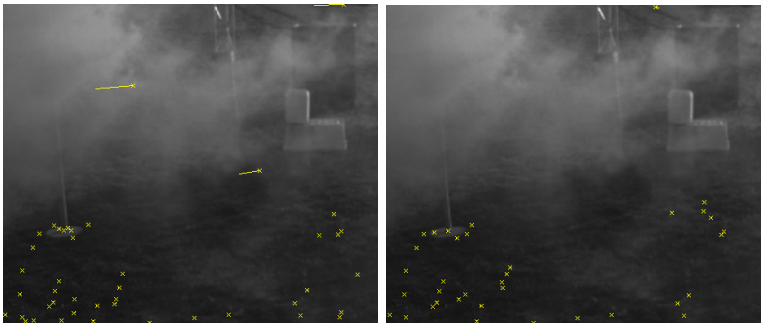
Figure 13: ROC illustrating the predictive power of SE for SIFT-matching error. The ROC curves were generated by varying the value of the SE threshold on visual images (blue) and IR images (red) for the stationary data sets used in Section 5.1. The selected values, providing the best ratio of true positive to false positive, are shown at the top-left corner: $T_{Vis}^{SE} = 4.13bits$ and $T_{IR}^{SE} = 4.60bits$.

was also used when evaluating quality locally (i.e. on sub-images).

5.3 Data Pre-Selection and SIFT-Matching Errors in Low-Visibility Conditions

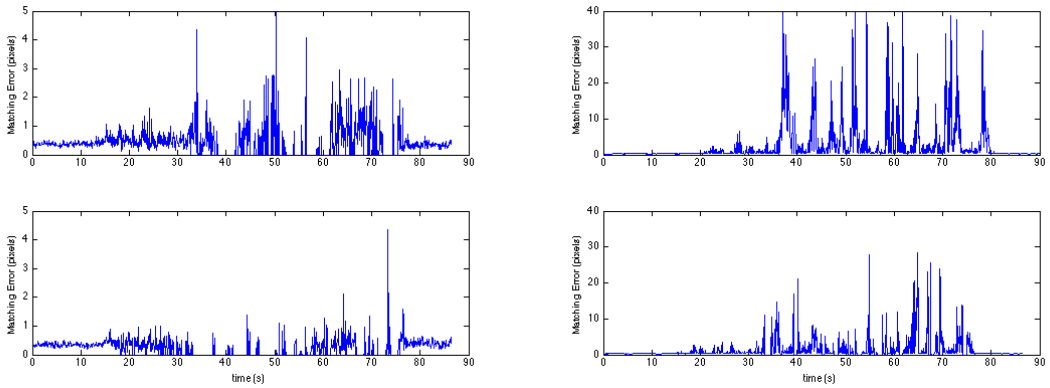
We evaluated the matching errors obtained with the outlier rejection techniques in the smoke data set *after* applying quality-based data pre-selection using the threshold values above. Images were divided into 10×10 regions and SE was evaluated locally. Features were discarded if they were located within regions of the image that were considered low quality. As in Section 5.1, the remaining features were then matched between consecutive images and outliers were rejected using the Gaussian expectation, RANSAC, MSAC and MAPSAC, respectively. We illustrate these results with MAPSAC, since it provided the lowest matching errors of the RANSAC variants, as indicated in Section 5.1. Fig. 14(b) shows an example of result, obtained for the same sequence as in Fig. 11. In this representative image in smoke, thanks to the pre-selection of data MAPSAC was actually able to remove all the incorrect matches. This is to compare with Fig. 14(a) (reproduced for convenience), where MAPSAC was applied without data pre-selection.

The use of the SE quality metric to pre-select image data is found to improve the final matching results between many smoke images. The bottom plot of Fig. 12 shows the evolution of the SIFT matching error over time after features have been eliminated using the quality evaluation. The top plots in Fig. 15 show the evolution of matching error after outliers are rejected using the Gaussian expectation and MAPSAC. The bottom plots show the corresponding evolution of matching error when SE has been used to eliminate



(a) MAPSAC-filtered matches. (b) MAPSAC-filtered SIFT matches after SE quality evaluation.

Figure 14: SIFT matching (in yellow) between visual images in smoke conditions. (a) corresponds to Fig 11(c) (reproduced for convenience).



(a) Gaussian expectation (b) MAPSAC

Figure 15: Static visual camera in the presence of smoke. Smoke is present during $15 - 77s$. Top: the average matching error (pixels) between Gaussian expectation (left) and MAPSAC (right) filtered SIFT matches between consecutive images. Bottom: the average matching error (pixels) between Gaussian expectation (left) and MAPSAC (right) filtered SIFT matches between consecutive images after features have been filtered using SE.

features prior to matching and then rejecting outliers. In all cases, there is a general reduction in matching error during times of smoke. Note that when the graphs indicate zero error, this is because there were no inliers, or no matches left after the quality evaluation. Table 3 shows the mean matching errors after quality-based data pre-selection for the smoke data set compared to the matching errors obtained in Section 5.1. Errors are shown separately for the total data set, during times of clear and smoke conditions. In these columns, the average error is only calculated for images where both methods find inliers. The table shows that generally the matching error in smoke is improved using the quality evaluation and at the very least performs as well as the other methods. The cases where no inliers were found are discussed below.

The final column of Table 3 shows the matching errors in the smoke data set when images were entirely rejected by the data pre-selection. This means that no inliers could be found by the outlier rejection

Table 3: Mean Matching Error (in pixels) over time for the representative smoke data set after applying different techniques of outlier rejection to the output of the SIFT matching algorithm.

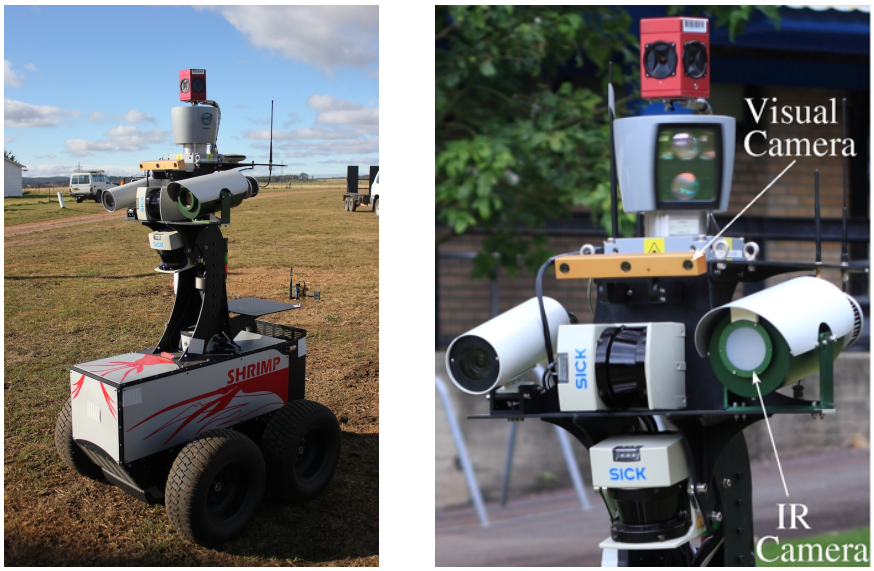
Outlier Rejection Method	SE	Mean Error in pixels (variance)			
		Total Data Set	Clear Conditions Only	Smoke Conditions Only	Whole Image Rejected by SE
None	-	33.3 (1043)	1.71 (0.3)	46.3 (900)	105.8 (1280)
	Yes	**27.1 (771)**	**1.63 (0.3)**	**37.7 (714)**	N/A
Gaussian expectation	-	0.38 (0.03)	0.36 (0.004)	0.41 (0.06)	1.64 (25.3)
	Yes	**0.39 (0.03)**	**0.35 (0.003)**	**0.42 (0.06)**	N/A
RANSAC	-	1.93 (9.10)	0.42 (0.0008)	2.55 (11.6)	9.9 (92.2)
	Yes	**1.60 (7.68)**	**0.41 (0.0006)**	**2.10 (10.1)**	N/A
MSAC	-	1.79 (8.71)	0.42 (0.0007)	2.36 (11.2)	9.8 (88.8)
	Yes	**1.54 (7.86)**	**0.41 (0.0008)**	**2.01 (10.4)**	N/A
MAPSAC	-	1.68 (11.76)	0.32 (0.0015)	2.25 (15.6)	11.7 (130.4)
	Yes	**1.51 (9.74)**	**0.31 (0.0013)**	**2.01 13.0**	N/A

techniques during these times because no features were available to them. In these cases no matching error can be calculated. On the other hand, without quality evaluation, the outlier rejection methods generally were able to find some inliers, however, large errors can be observed at these times. The quality evaluation has anticipated this and stopped the errors from propagating. This can be observed in Figs. 12 and 15 where the major spikes of matching errors have been reduced or eliminated.

5.4 Conclusion

In this section we have shown that low-visibility conditions can highly affect feature matching, and that state-of-the-art outlier rejection techniques can be insufficient in these situations. We also demonstrated that, by pre-evaluating the quality of image data, matching errors can be anticipated, which will result in the mitigation of camera localisation errors.

We note that in this process of data selection, some good quality data can be discarded. Further, at times of extreme low-visibility entire images might be rejected, resulting in the absence of data from the corresponding camera. For example, if smoke covers the full field of view of the visual camera, no visual feature will be available to the localisation algorithm. In these situations, the use of complementary sensing modalities allows us to maintain a flow of input data to feed the camera localisation algorithm, which is essential for resilience. In our previous example, features will still be available for localisation thanks to the IR camera. In the remaining sections, we implement the proposed approach on a mobile robot equipped with a visual camera and an IR camera.



(a) Shrimp (b) Sensor frame

Figure 16: The Shrimp robot and its sensors.

6 Experimental Platforms and Data Sets

This section describes the platform and data sets used for the experiments in this paper.

6.1 Shrimp UGV Platform

The ACFR Shrimp platform (Fig. 16(a)) is based on the Segway RMP 400 module. It is equipped with multiple sensor modalities, including a Novatel RTK[5] dGPS/INS SPAN[6] unit, composed of a Novatel ProPak-G2plus GPS receiver and a Honeywell HG1700 AG17 IMU. It provided the 6-DOF *reference* localisation (with an average $2cm$ global accuracy) used in the experimental validation.

We used a Raytheon Thermal-Eye 2000B IR camera and the left camera of a Point Grey Bumblebee XB3 camera set (see Fig. 16(b) for sensors and Figs. 3, 4 and 5 for examples of captured data). The sensor specification parameters are shown in Table 4. Note that the visual images were converted to gray-scale and down-sampled to 640×480 for the experiments, a resolution comparable to that of the IR images. The two cameras were not synchronised in hardware, but images were logged on the same computer and accurately timestamped.

[5]Real-Time Kinematic
[6]Synchronised Position Attitude and Navigation

Table 4: Sensor Parameters

Camera	Spectral Range	Field of View	Raw Resolution	Framerate
Visual	$390 - 750n$m	43°	1280×960	15 fps
IR	$7 - 14\mu$m	35.8°	480×576	12.5 fps

6.2 Data Sets

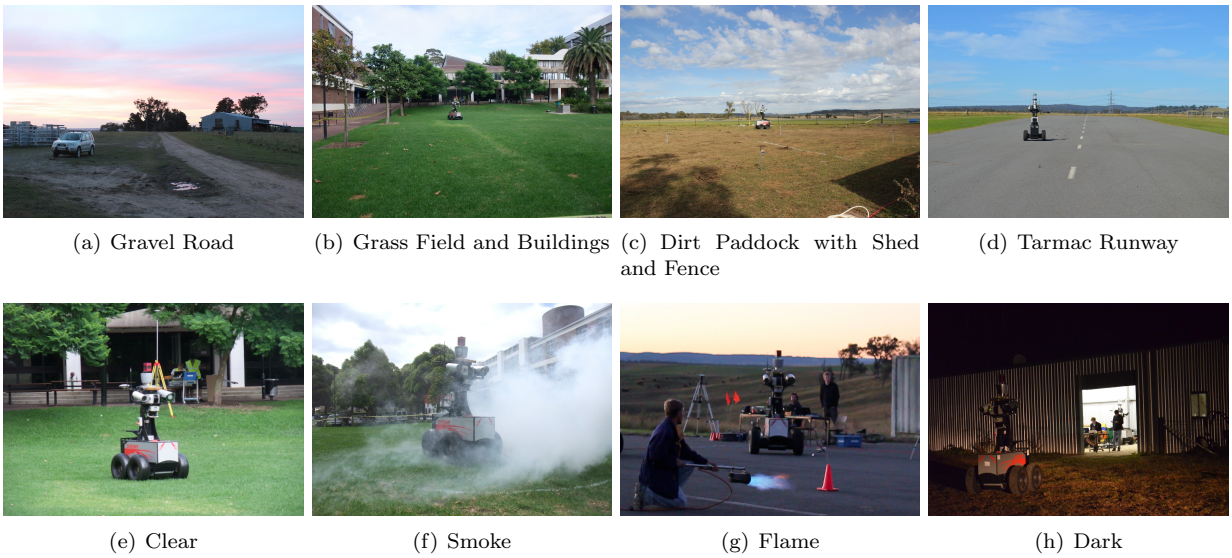
Data sets were acquired by remotely driving the *Shrimp* robot along a range of controlled trajectories through semi-urban environments in varying visibility conditions, while logging sensor data. Table 5 summarises the main characteristics of each data set, including environmental conditions, trajectory type and statistics. Environments ranged from a gravel road, a grass field surrounded by buildings, a dirt paddock with a shed and fence and a tarmac runway (see Fig. 17(a)-(d)) and were recorded at different times of the day and night. Depending on the visibility conditions, data sets are labelled as *Clear*, *Smoke*, *Dark* or *Flame*. Examples of Shrimp operating in these conditions are shown in Fig. 17(e)-(h) respectively.

Table 5: Data Set Summary

Data Set	**Trajectory**	**Duration**	**Distance**	**Average Velocity**	**Average Yaw Rate**	**Visibility Changes (times*)**
Clear	Circle	$70s$	$20.0m$	$0.29\,m/s$	$2.4\,°/s$	Clear
Smoke A	Straight Line	$45s$	$22.9m$	$0.51\,m/s$	$0.7\,°/s$	Smoke $(18 - 44.5s)$
Smoke B	Circle	$60s$	$34.2m$	$0.58\,m/s$	$6.2\,°/s$	Smoke $(18 - 30s$ & $34 - 53s)$
Smoke C	Circle	$50s$	$28.3m$	$0.57\,m/s$	$5.5\,°/s$	Smoke $(0 - 13s)$
Flame	Turn	$52s$	$9.7m$	$0.19\,m/s$	$2.17\,°/s$	Flame $(12 - 33s)$
Dark	Turn	$86s$	$7.4m$	$0.09\,m/s$	$1.0\,°/s$	Dark $(10 - 30s$ & $56 - 86s)$

* The rest of the time, conditions are clear (Smoke and Flame data sets) and the scene is lit (Dark data set).

In these experiments, Shrimp is moving and acquiring images from both IR and visual cameras. Data sets labelled as *Clear* refer to normal daytime operating conditions with good visibility for both sensors. In the other data sets, a variation was introduced into the environment (e.g. smoke) that provoked a change in the visibility conditions for one of the sensors. Unless specified otherwise, initially the conditions are clear.



(a) Gravel Road (b) Grass Field and Buildings (c) Dirt Paddock with Shed and Fence (d) Tarmac Runway

(e) Clear (f) Smoke (g) Flame (h) Dark

Figure 17: Environments in which the data sets were captured (a)-(d) and the Shrimp Robot operating in variable visibility conditions (e)-(h).

For *Smoke* data sets, a smoke machine using a water-based poly-glycol[7] introduced varying levels of white smoke into the environment, provoking poor-visibility conditions for the visual camera (see Fig. 3 for a sensor view). In the Smoke A and Smoke B data sets, Shrimp started in clear conditions and smoke was introduced after some time. In the Smoke C data set, thick smoke was present initially.

For the *Flame* data set, a gas-fired flame gun was used to heat the air. The robot was driven on a tarmac runway on a cold afternoon as the sun was setting, meaning that there were fewer features available for both sensors than in other data sets. While the flame itself was observed by the visual camera, the fire generated limited heat hazing and no visible smoke, therefore the visual images were not significantly affected in this data set. On the other hand, the IR data were affected by the flame and the surrounding heat, meaning background objects were obscured (refer to Fig. 5 for a sensor view).

For the *Dark* data sets, the robot was driven at night in various artificial lighting conditions. In the example used in this paper, at the start of the data set, the robot drove towards a large shed opening. At first, all lights inside the shed were on (see Fig. 4, left, for a sensor view). All lights were then switched off for $20s$, resulting in near-complete darkness (Fig. 4 right), then turned back on. The robot then drove past the shed and turned away into an unlit area, also creating poor visibility for the visual camera.

[7] JEM-ZR22 machine with Jem Pro-Smoke Super Fluid.

7 Experimental Results

In this section we present the results of trajectory estimation using the localisation algorithm described in Section 4, with different strategies of combination of IR and visual image data. The mean local error (γ with $dt = 2s$, see Eq. (3)) between estimated trajectories and the reference, provided by the dGPS/INS unit onboard the robot, was used to compare and evaluate these results. Firstly, overall results are presented for each data set. Secondly, we specifically consider the effect of low-visibility conditions on the localisation accuracy and compare the performance of the different data-selection techniques.

For each data set, the robot trajectory was estimated using each of the 5 different variations of data sources and data combination methods described in Section 4.5 and Table 1, namely: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc and LSSM.Loc. Additionally, we performed a statistical analysis over 8 runs of the estimation algorithm, due to the random selection of the features in the images. This resulted in 5 groups of trajectories shown for each data set, each group containing 8 trajectory estimates. For each data set, we compare the performance of the 5 methods considered. However, results are not compared between different data sets because the specific trajectory, velocity and background environment all contribute to the error recorded.

The estimated trajectories were computed off-line but the selection of image data was performed during the execution of the localisation algorithm. The parameters of the localisation algorithm were initialised identically for each run. Features were randomly selected from a 10×10 grid in each image. A maximum of 50 features were maintained in a common map and these were split evenly when two sensors were being used (i.e. 25 features each). The scale of the trajectories was recovered *a posteriori* using the velocity information from an onboard IMU.

Section 7.1 provides an overview of the results obtained for all the methods considered and all data sets. Sections 7.2 - 7.5 provide specific analyses for each type of environmental condition. Finally, Section 7.6 presents a summary of all these results.

7.1 Results Overview

Table 6 shows the root-mean-square (RMS) of the local error γ over the 8 runs executed for each method. The worst and best performing methods for each data set are highlighted in red and green respectively.

Global trajectories are shown in the left column of Figs. 19-24 for Clear, Smoke, Flame and Dark conditions. The reference is plotted in black, while the estimated trajectories are in colour. The robot was initially

Table 6: RMS of γ over 8 runs for each localisation method (standard deviation in parenthesis). All units are mm.

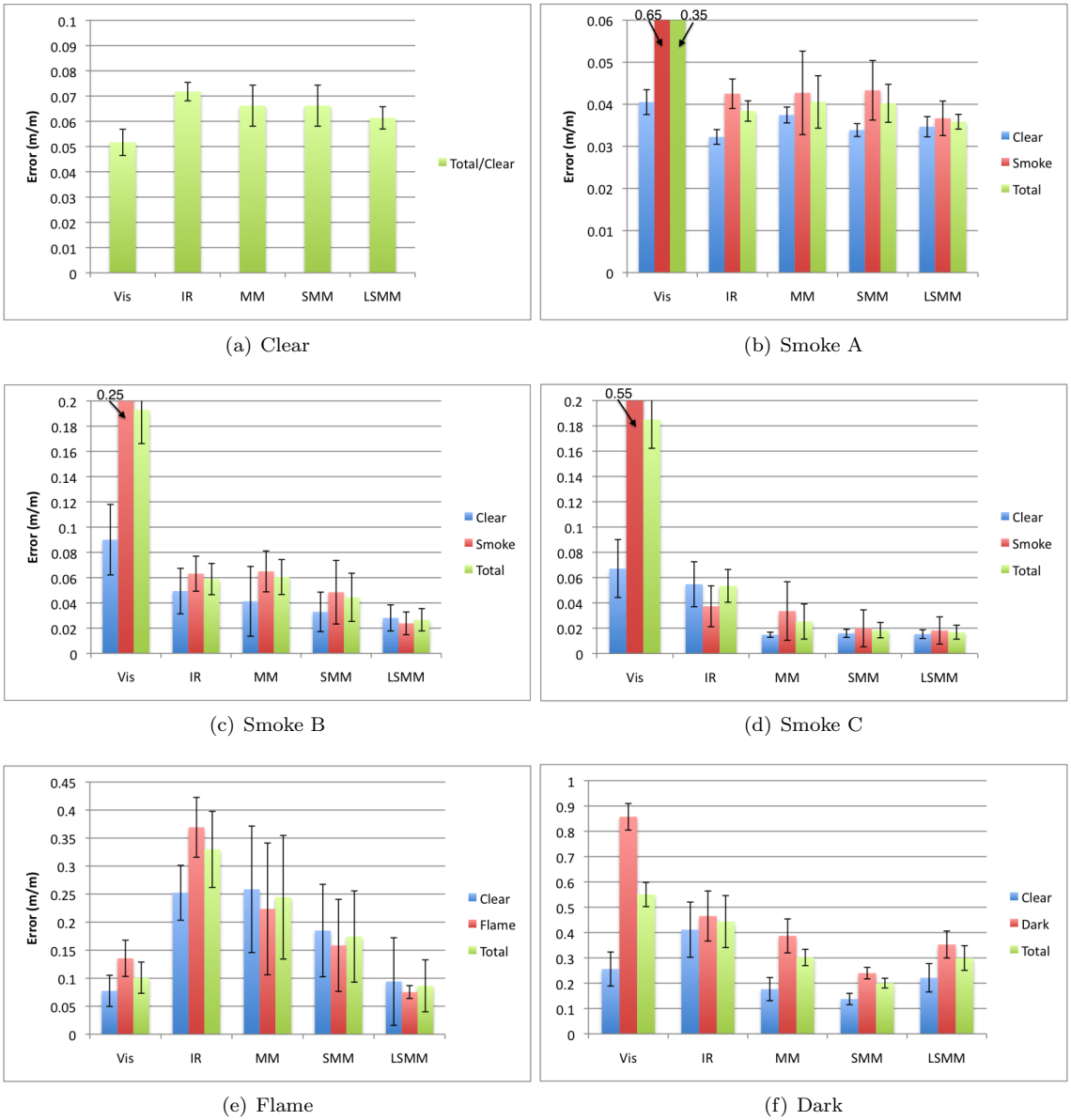
Data Set	Average Distance Travelled (measured by the reference)	RMS Error				
		VisCam-Loc	IRCam-Loc	MMCam-Loc	SMMCam-Loc	LSMMCam-Loc
Clear	582	**30.0 (3.0)**	41.7 (2.1)	38.5 (4.7)	38.5 (4.7)	35.7 (2.6)
Smoke A	957	337 (51)	36.8 (2.3)	38.8 (5.9)	38.5 (4.3)	**34.3 (1.7)**
Smoke B	1093	210 (29)	64.3 (13.5)	66.1 (15.1)	48.5 (20.9)	**29.1 (9.6)**
Smoke C	1142	211 (26.0)	61.0 (14.8)	28.9 (15.9)	21.1 (6.9)	**19.3 (6.3)**
Flame	360	36.4 (10.0)	118 (24.4)	88.1 (39.7)	62.7 (29.3)	**31.1 (16.7)**
Dark	160	88.1 (7.6)	71.0 (16.4)	48.4 (5.2)	**32.1 (3.1)**	48.0 (7.8)

positioned at (0,0), facing along the x-axis. The right column of these figures show the instantaneous local pose difference (δP) vs. time for the corresponding group of 8 runs. The blue line shows the mean δP over the 8 runs and the red dashed lines show the maximum and minimum δP observed. The presence of low-visibility conditions (i.e. image data containing significant smoke, flame or dark) was identified manually and is displayed as shaded areas in the figures.

The metre-per-metre (m/m) error was obtained by dividing γ by the average distance travelled and was evaluated exclusively for times of clear and low-visibility conditions. These results are shown in Fig. 18 by the blue and red columns respectively, with the green column showing the total m/m error. The variance in the m/m error over the 8 runs in each set is shown by the black error bars. Low overall error and small variance indicate good localisation performance as all 8 runs match up to the reference trajectory closely.

7.2 Clear Conditions

Clear conditions are found throughout the Clear data set, at the start of data sets Smoke A, Smoke B, Flame and Dark, and at the end of data set Smoke C. Clear conditions are shown by unshaded regions of the right column of Figs. 19-24 and the m/m error is displayed for these conditions in the blue columns in Fig. 18. The results show that all methods provide a comparably accurate estimate of the localisation during clear conditions.



(a) Clear

(b) Smoke A

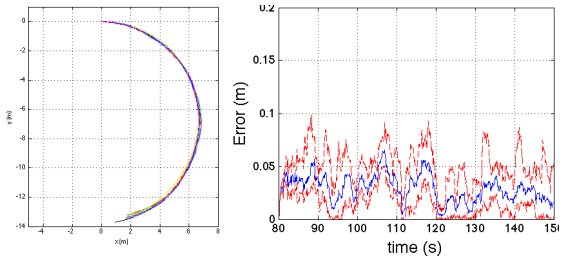
(c) Smoke B

(d) Smoke C

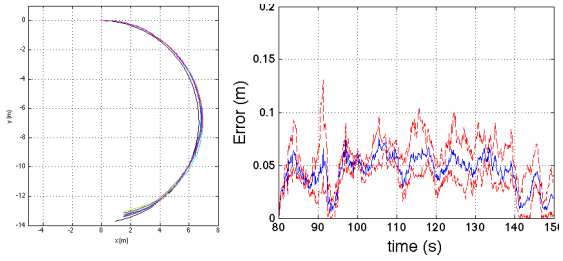
(e) Flame

(f) Dark

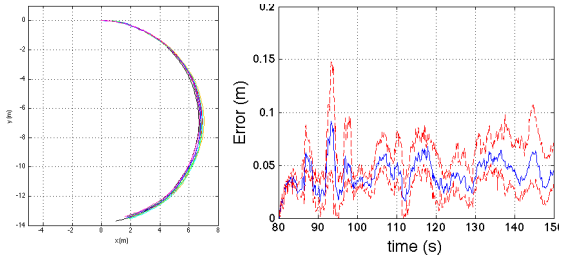
Figure 18: The RMS m/m error for 8 localisation runs using the different methods considered. The blue column shows the error observed during clear conditions, the red column indicates the error observed during low-visibility conditions, and the green column shows the total error for the entire run. The subtitle refers to the data set.



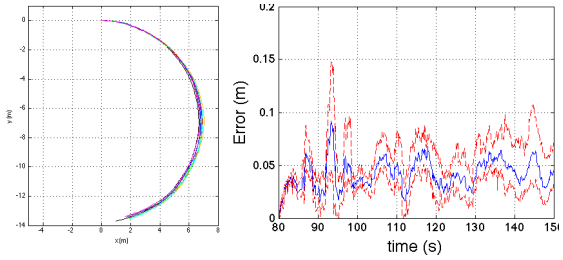
(a) Vis.Loc (b) Vis.Loc



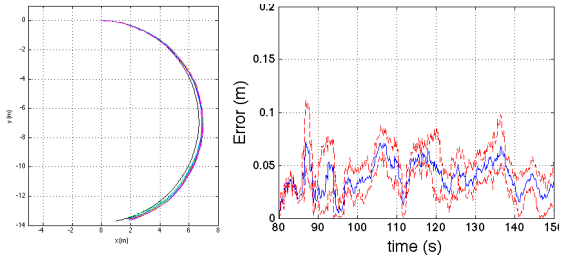
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc

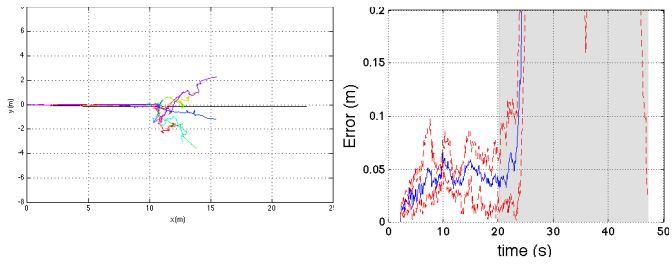


(g) SMM.Loc (h) SMM.Loc

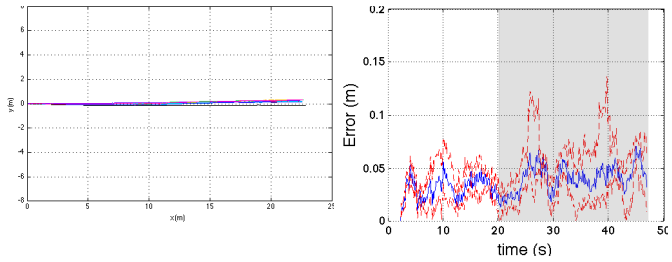


(i) LSMM.Loc (j) LSMM.Loc

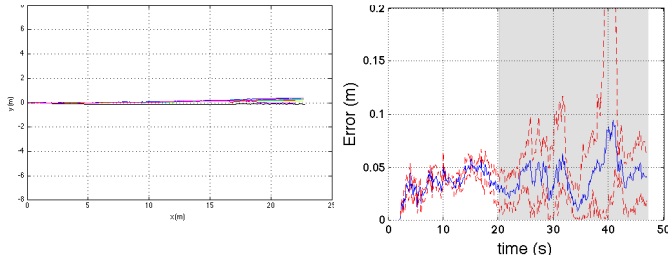
Figure 19: Results for the **Clear** data set using the different data combination and selection methods for 8 runs. Top to bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line.



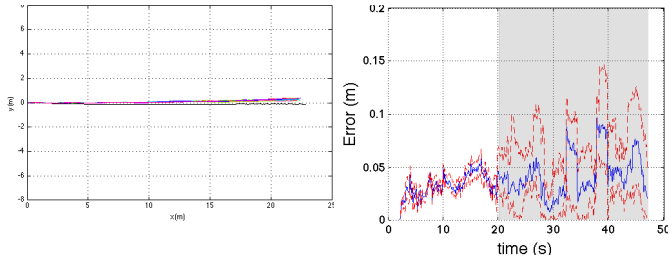
(a) Vis.Loc (b) Vis.Loc



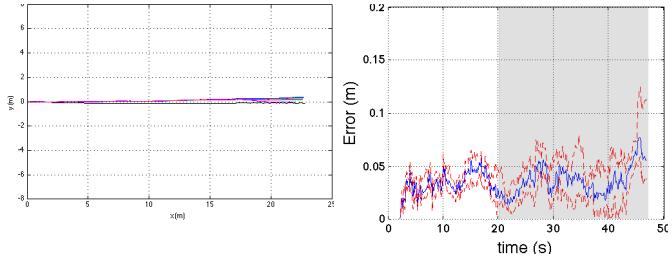
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc

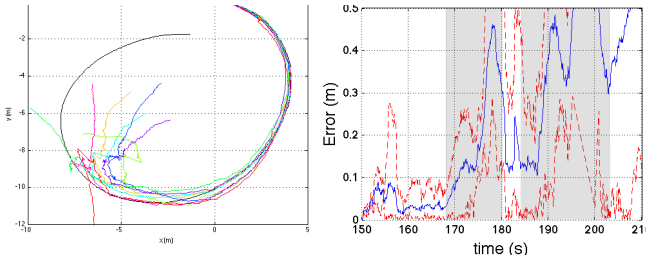


(g) SMM.Loc (h) SMM.Loc

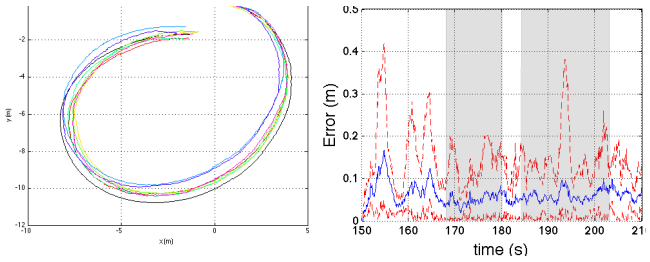


(i) LSMM.Loc (j) LSMM.Loc

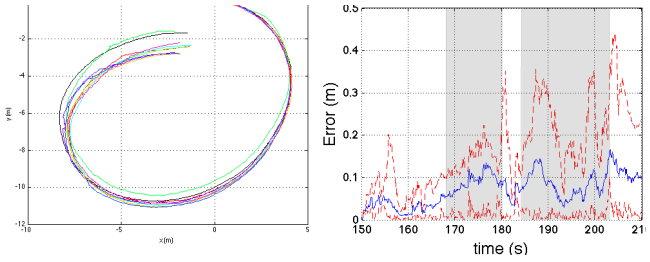
Figure 20: Results for the **Smoke A** data set using different data combination and selection methods for 8 runs. Top to Bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line. The shaded areas indicate the presence of smoke.



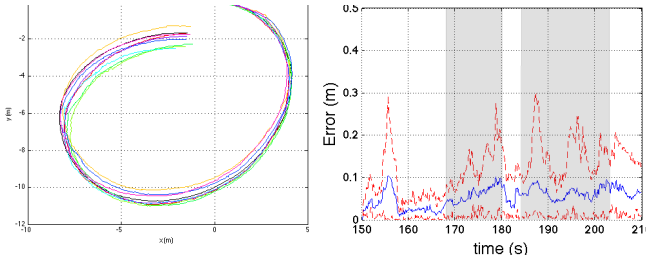
(a) Vis.Loc (b) Vis.Loc



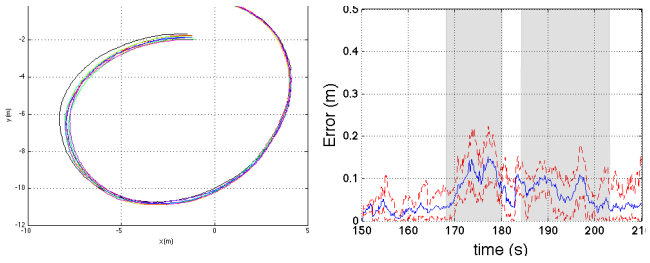
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc

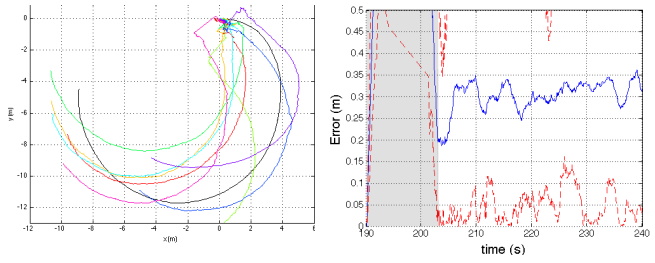


(g) SMM.Loc (h) SMM.Loc

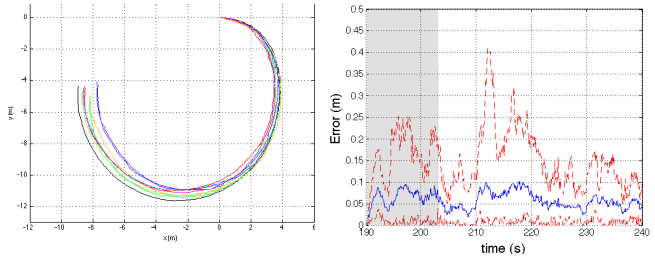


(i) LSMM.Loc (j) LSMM.Loc

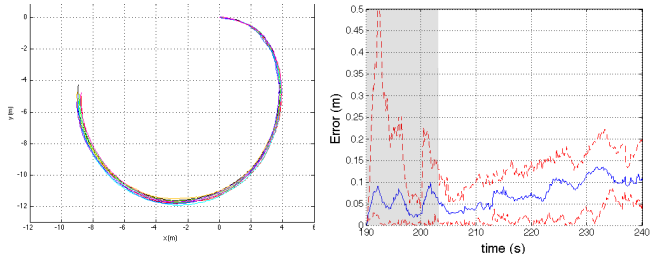
Figure 21: Results for the **Smoke B** data set using different data combination and selection methods for 8 runs. Top to bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line. The shaded areas indicate the presence of smoke.



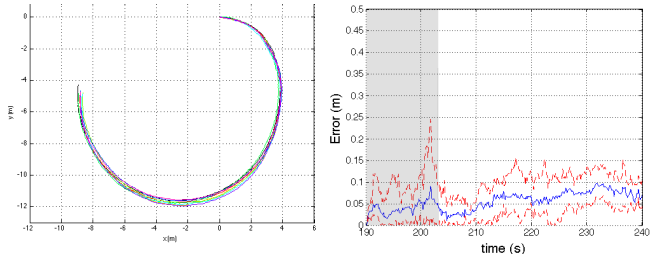
(a) Vis.Loc (b) Vis.Loc



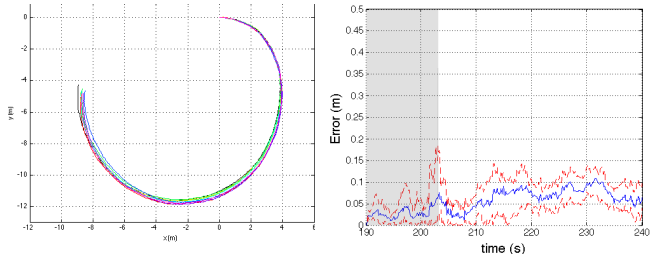
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc

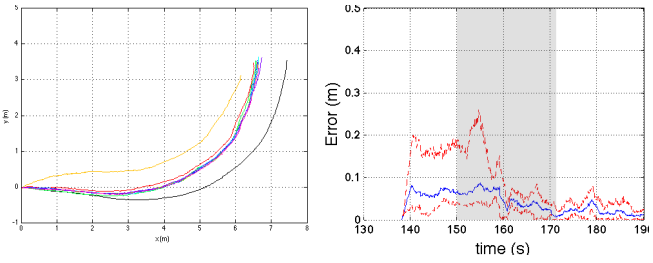


(g) SMM.Loc (h) SMM.Loc

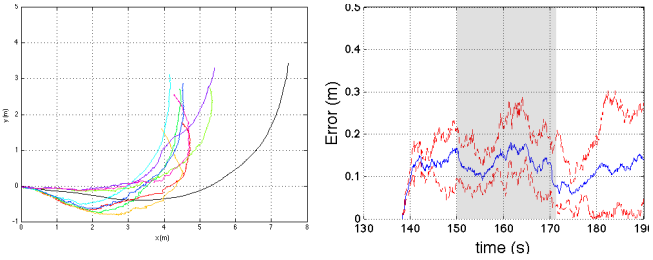


(i) LSMM.Loc (j) LSMM.Loc

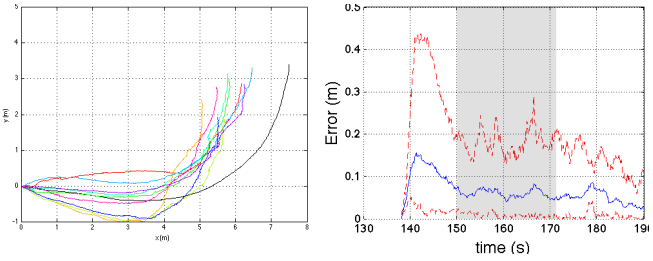
Figure 22: Results for the **Smoke C** data set using different data combination and selection methods for 8 runs. Top to bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line. The shaded areas indicate the presence of smoke.



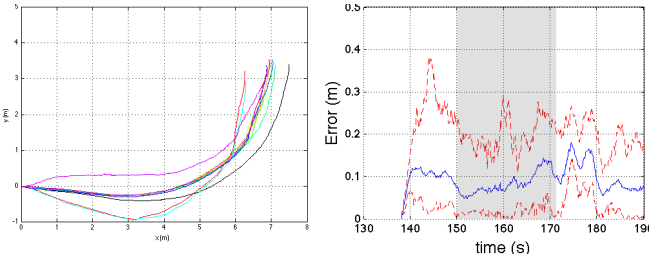
(a) Vis.Loc (b) Vis.Loc



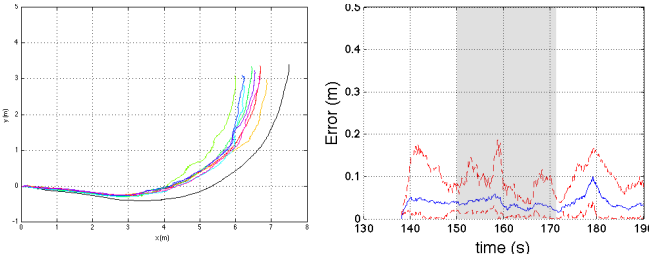
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc

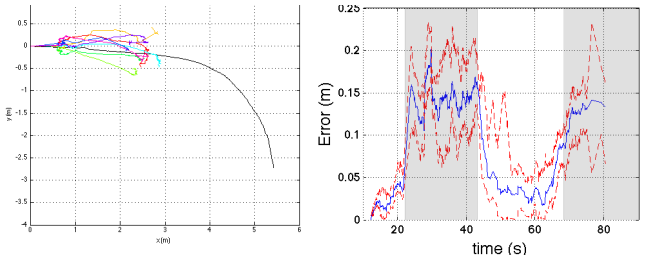


(g) SMM.Loc (h) SMM.Loc

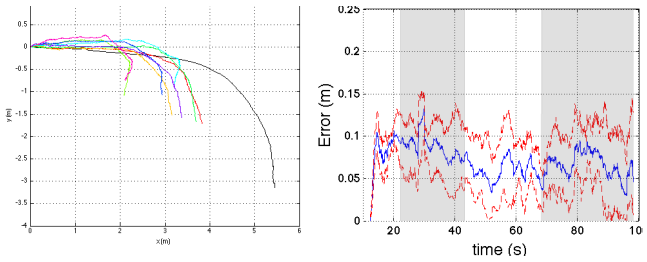


(i) LSMM.Loc (j) LSMM.Loc

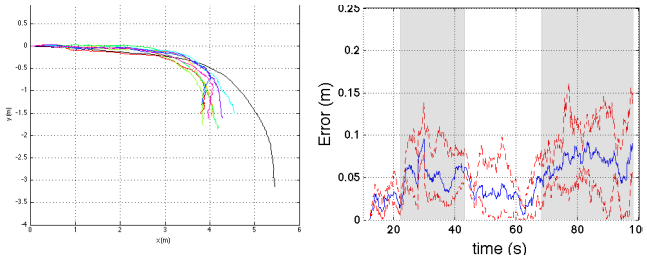
Figure 23: Results for the **Flame** data set using different data combination and selection methods for 8 runs. Top to bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line. The shaded areas indicate the presence of the flame/heat.



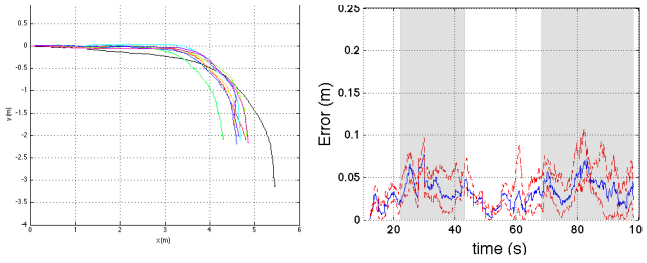
(a) Vis.Loc (b) Vis.Loc



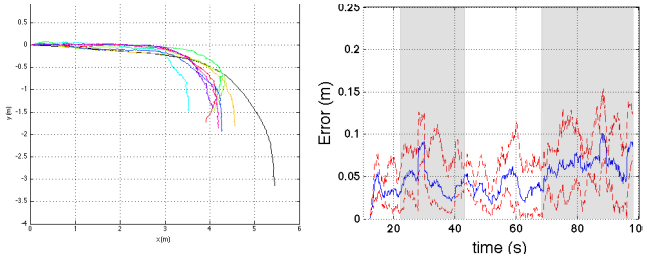
(c) IR.Loc (d) IR.Loc



(e) MM.Loc (f) MM.Loc



(g) SMM.Loc (h) SMM.Loc



(i) LSMM.Loc (j) LSMM.Loc

Figure 24: Results for the **Dark** data set using different data combination and selection methods for 8 runs. Top to bottom: Vis.Loc, IR.Loc, MM.Loc, SMM.Loc, LSMM.Loc. The left column shows the estimated trajectories next to the reference in black, projected on (x, y) plane. The right column shows the corresponding average local error over time (δP with $dt = 2s$) in blue with the minimum and maximum local error shown with a red dashed line. The shaded areas indicate periods of (almost) complete darkness, i.e. absence of artificial light.

The use of an individual sensor, as shown for Vis.Loc or IR.Loc, is sufficient to produce a good trajectory in clear conditions. In data sets Smoke A and Smoke B, Vis.Loc is less accurate than IR.loc, but the opposite is true in the clear conditions of Flame and Dark. The difference in performance can be explained by the background environment. Flame and Dark data sets were captured in colder conditions (in the evening) and were less favourable to IR sensing than the warmer conditions of the smoke data sets.

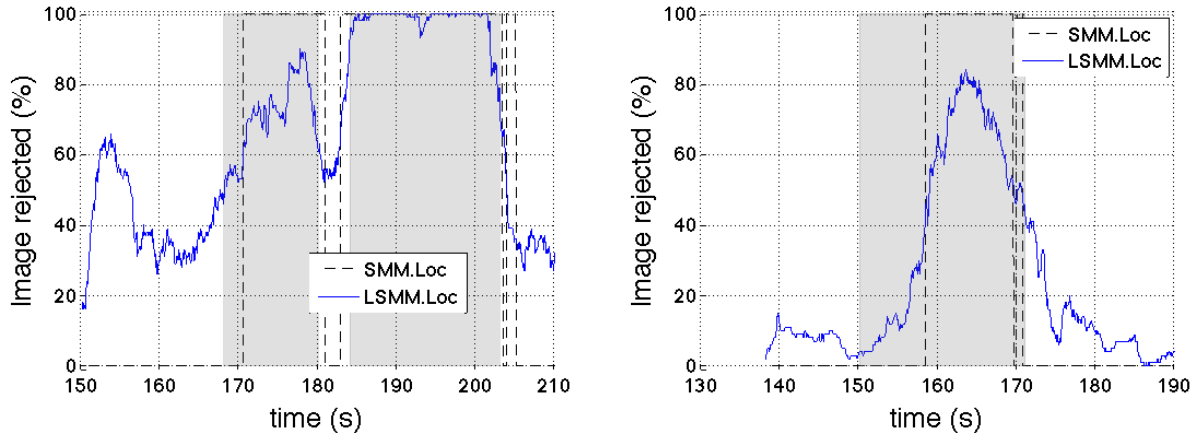
By combining the two sensing modalities, MM.Loc typically performed at least as well as either Vis.Loc or IR.Loc in clear conditions. For the Dark data set, the MM.Loc error is significantly lower than either Vis.Loc or IR.Loc. This is because the information provided by images from each sensor is complementary (see Fig. 7(a)), which results in a greater spatial distribution of features in the image and, subsequently, a better estimate of the motion of the robot. In the Smoke A and Flame data sets, the variance of MM.Loc is high, see Fig. 18. In both cases (see (f) of Figs. 20 and 23), the presence of the smoke or flame had an effect on the error even after the low-visibility condition was no longer present. This is because the features are still present in the map for some time afterwards and continue to cause errors.

In clear conditions, the estimates computed by multiple-modality methods (MM.Loc, SMM.Loc, LSMM.Loc) have comparable errors because very little data were rejected by either SMM.Loc or LSMM.Loc and, therefore, any variability in results is mainly due to the feature selection within the localisation algorithm. In some cases, the accuracy of LSMM.Loc is similar or higher than that of MM.Loc because LSMM.Loc removed low-quality image data that were captured during clear conditions. For example, in the Smoke B data set, the error of MM.Loc in clear conditions at $155s < t < 158s$ (Fig. 21(h)) is high compared to other times of clear conditions, due to the saturation of part of the image caused by the reflection of sunlight on a cement surface. LSMM.Loc (Fig. 21(j)) significantly reduced this error by rejecting up to 60% of the visual image (see Fig. 25(a)) during this period.

7.3 Smoke Conditions

This section refers to the Smoke A, Smoke B and Smoke C data sets only, where visual data were affected by low-visibility conditions. The estimated trajectories can be seen in the left column of Figs. 20-22 and the local error over time in smoke is shown within the shaded regions of the right column. The m/m errors in the presence of smoke for the different selection methods are given by the red columns in Figs. 18(b)-18(d).

In data sets Smoke A and Smoke B, Vis.Loc initially produces an accurate pose estimation in clear conditions. However, as smoke appears and corrupts visual data, Vis.Loc fails dramatically (see Figs. 20(a) and 21(a)).



(a) Smoke B - Visual Data

(b) Flame - IR Data

Figure 25: Proportion of image rejected during by SMM.Loc or LSMM.Loc during the Smoke B (left) and Flame (right) data sets. Periods of smoke or high heat and flame are shown in grey. SMM.Loc (blue) rejects whole images and LSMM.Loc (black dashes) rejects smaller regions of an image. Note that even during clear times, LSMM.Loc rejects portions of the image as inappropriate for the localisation application.

Data set Smoke C starts in the presence of smoke and, therefore, Vis.Loc immediately fails to produce an accurate trajectory. IR.Loc significantly outperforms Vis.Loc because the operational spectrum of the IR camera is not affected by smoke and continues to provide good quality data.

In smoke, MM.Loc is as accurate as IR.Loc, which shows that it is able to mitigate many of the poor features caused by inappropriate visual data and maintain a reasonable estimate of the trajectory. However, the variance of MM.Loc is higher than IR.Loc, which shows that despite the careful selection of features in MM.Loc (as detailed in Section 4), smoke can still have an effect on the overall performance of the localisation.

The rejection of low-quality images in SMM.Loc using the global approach contributed to the reduction of the localisation errors in the presence of smoke compared to MM.Loc. However, the variance of SMM.Loc remained higher than that of IR.Loc because images that contained a relatively small amount of smoke were still considered to be good-quality data overall, see Fig. 25(a). Subsequently, low-quality data corrupted the localisation.

During smoke conditions, LSMM.Loc used all the IR data, as did SMM.Loc, but combined them with *only* the proportion of the visual data that was good quality, see Fig. 25(a). As a result, the LSMM.Loc estimates have a lower overall error and lower variance in smoke than any of the other techniques evaluated.

7.4 Dark Conditions

This section refers to the Dark data set only, where the visual camera was often unable to perceive the environment. The trajectories can be found in Fig. 24 and the average local error in dark conditions for different selection methods is given by the red columns in Fig. 18(f). Many similar results can be observed for Dark as for Smoke. For example, when visual data are corrupted by a low-visibility condition, Vis.Loc failed to provide a reasonable estimate of the trajectory since there were no suitable features to use in the visual images. In extreme cases, when the entire image was affected, the localisation estimated the robot to have stopped completely. IR.Loc performs equally well in light and dark as the artificial visual illumination does not have a significant impact on the thermal properties of the environment. However, note that at nighttime the contrast of IR images was naturally lower than at daytime due to a more uniform distribution of temperatures in the environment, particularly in the last part of the dataset. This caused more error in the IR matches, with a stronger impact on the localisation accuracy when IR was the main source of information (e.g. IR.Loc). However, the methods based on selective data combination were able to mitigate some of the effect of these errors.

Fig. 18(f) shows that the accuracy of MM.Loc and LSMM.Loc is comparable during periods of darkness. This is because the regions that were found to be low quality by LSMM.Loc were black and, therefore, no SIFT features could be extracted by MM.Loc either. The figure also shows that during periods of darkness, SMM.Loc is more accurate than both MM.Loc and LSMM.Loc. Fig. 24(h) shows that the error reduction mainly occured during the second dark phase of the data set. As described in Section 6, this phase involved the robot turning away from a lit doorway, causing the region of good-quality visual data to be progressively constricted to one side of the image. The result was a limited observability of the 6-DOF camera motion in the visual images, leading to increased errors for the methods using these visual data, i.e. MM.Loc and LSMM.Loc. SMM.Loc was not affected by this problem because it rejected these images and used only the IR camera during these times. This indicates that the local version of the data quality evaluation used in LSMM.Loc could be improved by accounting for the requirements on the spatial distribution of the contributing data in the image. This will be left to future work.

7.5 Extreme Heat Conditions (Flame)

The Flame dataset provides the counter-example to smoke and dark conditions in that the IR data are corrupted by low-visibility conditions instead of the visual data. As shown in Fig. 18(e), SMM.Loc and

LSMM.Loc improve on MM.Loc error by selecting appropriate data to use for the localisation. The estimates of LSMM.Loc have the lowest overall error and lowest variance. Fig. 25(b) shows the amount of IR data that was rejected by LSMM.Loc prior to its use in the localisation algorithm.

7.6 Results Summary

Significant errors in the localisation estimates were observed during low-visibility conditions for methods that rely on a single sensor type (Vis.Loc and IR.Loc). For example, the performance of Vis.Loc is poor in smoke or darkness compared to clear conditions. The same is true for IR.Loc in high heat and flame.

Combining data from both sensor modalities (MM.Loc) allowed for a clear reduction of the overall error and improved consistency in these adverse conditions, thanks to the additional contribution of the unaffected sensor. However, MM.Loc was still largely affected by low-visibility conditions, as indicated by increased error and higher variance in the estimations during these times, compared to the performance in clear conditions. This is because data from the affected sensor still contributed to the localisation estimation, and so corrupted the solution.

By eliminating most of the affected data, methods using data evaluation and pre-selection (SMM.Loc and LSMM.Loc) were shown to further mitigate the errors caused by low-visibility conditions and improve the consistency of results. The error was also reduced in some cases in clear conditions. In particular, by evaluating and selecting data *locally*, LSMM.Loc was shown to be superior to the *global* method SMM.Loc in most cases, as it anticipated the impact of low-visibility conditions earlier and kept a larger quantity of good-quality data in general.

8 Conclusion

Resilient perception is a fundamental requirement for long-term autonomy in robotics. In this paper, we addressed the problem of maintaining reliable perception for UGVs in low-visibility conditions, where low-quality, potentially inappropriate sensor data are likely to be introduced to the perception system, leading to unbounded errors.

Using a suite of sensors that operate in different spectral ranges (i.e. different sensing modalities), perception systems can be resilient to a larger variety of environmental conditions, since it is more likely that at least one sensor will have sufficient visibility at any point in time. In this paper, we used visual and IR cameras

onboard a UGV to localise the platform with a state-of-the-art Visual-SLAM algorithm based on SIFT matches. Combining visual and IR imaging was shown to improve the resilience of camera-based localisation in many environments, however, we demonstrated that adverse environmental conditions can cause large errors in SIFT matching. Importantly, we showed that despite careful selection of matches using robust outlier rejection and data-association techniques, a significant proportion of these errors remain. The paper demonstrated experimentally that SIFT-matching errors could be anticipated by evaluating image data quality. Therefore, to mitigate these errors, we proposed to discard low-quality data prior to localisation.

An extensive experimental evaluation was conducted to validate the proposed approach. The robot was driven in a range of low-visibility conditions, such as smoke, high heat and darkness. First, only visual or IR images were used to estimate the trajectory. Second, trajectories were estimated using a combination of all the image data provided by both visual and IR cameras. Third, the proposed quality evaluation and data pre-selection framework was used prior to data combination. The quality evaluation was performed in two ways: using a global approach (i.e. over the whole image) and a local approach (i.e. over sub-images).

Combining all data from both sensor modalities allowed for a clear reduction of the errors in adverse conditions compared to methods using a single sensor type, thanks to the additional contribution of the unaffected sensor. However, the effects of low-visibility conditions on this method were still significant. By eliminating most of the affected data, methods using data evaluation and pre-selection were shown to further mitigate the errors caused by these conditions. In particular, evaluating and selecting data locally was shown to be superior to the global method. In conclusion, the proposed selective data combination approach allowed for resilient localisation in a large range of low-visibility conditions.

In future work, we will consider applying a local-map strategy as in (Vidal-Calleja et al., 2011) to localise the vehicle in longer datasets with a variety of challenging conditions. The proposed methods of selective data combination will be applied to a larger range of perception applications to demonstrate the generality of the process of quality evaluation and the associated data pre-selection. We will also investigate techniques of data quality evaluation for other sensing modalities, including lasers and radars. A greater range of sensors would provide the opportunity for a more nuanced process of selecting appropriate data. In addition, checking the consistency between heterogeneous sensor data will further contribute to discriminating the most appropriate input data for a resilient perception system.

Acknowledgments

This work was supported in part by the Australian Centre for Field Robotics (ACFR), by the Centre for Intelligent Mobile Systems (CIMS), funded by BAE Systems as part of an ongoing partnership with the University of Sydney, and by the New South Wales State Government. The authors would like to thank Joan Solà for software collaboration and Matthieu Simmoneau for his contribution to code development.

References

Ardeshir Goshtasby, A. and Nikolov, S. (2007). Guest editorial: Image fusion: Advances in the state of the art. *Information Fusion*, 8(2):114–118.

Borges, P., Zlot, R., Bosse, M., Nuske, S., and Tews, A. (2010). Vision-based localization using an edge map extracted from 3d laser range data. In *IEEE International Conference on Robotics and Automation*, Anchorage, AK.

Brooker, G. (2009). *Introduction to Sensors for Ranging and Imaging*. SciTech Pub.

Brunner, C. and Peynot, T. (2010). Perception quality evaluation with visual and infrared cameras in challenging environmental conditions. In *International Symposium on Experimental Robotics*, Delhi, India.

Brunner, C., Peynot, T., and Underwood, J. (2009). Towards discrimination of challenging conditions for ugvs with visual and infrared sensors. In *ARAA Australasian Conference on Robotics and Automation*, Sydney, Australia.

Brunner, C., Peynot, T., and Vidal-Calleja, T. (2011a). Combining multiple sensor modalities for a localisation robust to smoke. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, CA.

Brunner, C., Peynot, T., and Vidal-Calleja, T. (2011b). Visual metrics for the evaluation of sensor data quality in outdoor perception. *International Journal of Intelligent Control and Systems*, 16(2):142–159.

Burgard et al., W. (2009). A comparison of slam algorithms based on a graph of relations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Saint Louis, MO.

Carlson, J. and Murphy, R. (2005). Use of dempster-shafer conflict metric to adapt sensor allocation to unknown environments. In *American Association for Artificial Intelligence*, Pittsburgh, PA.

Castro, M. and Peynot, T. (2012). Laser-to-radar sensing redundancy for resilient perception in adverse environmental conditions. In *ARAA Australasian Conference on Robotics and Automation*, Sydney, Australia.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *International Conference on Computer Vision*, Nice, France.

Davison, A. J. (2005). Active search for real-time vision. In *International Conference on Computer Vision*, Los Alamitos, CA.

Dubbelman, G., van der Mark, W., van den Heuvel, J., and Groen, F. (2007). Obstacle detection during day and night conditions using stereo vision. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Diego, CA.

Fay et al., D. (2000). Fusion of multi-sensor imagery for night vision: color visualization, target learning and search. In *International Conference on Information Fusion*, Paris, France.

Ferwerda, J., Pattanaik, S., Shirley, P., and Greenberg, D. (1996). A model of visual adaptation for realistic image synthesis. In *Annual Conference on Computer Graphics and Interactive Techniques*, New Orleans, LA.

Finlayson, G., Hordley, S., and Drew, M. (2006). Removing shadows from images. *European Conference on Computer Vision*.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commununications of the ACM*, 24(6):381–395.

Foyle, D., Ahumada, A., Larimer, J., and Townsend Sweet, B. (1993). Enhanced/synthetic vision systems: Human factors research and implications for future systems. *SAE Transactions*, 101.

Green, D. M. and Swets, J. A. (1989). *Signal Detection Theory and Psychophysics*. Peninsula Pub.

Gu, X., Yu, D., and Zhang, L. (2005). Image shadow removal using pulse coupled neural network. *IEEE Transactions on Neural Networks*, 16(3):692–698.

Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.

Hines, G., Rahman, Z., Jobson, D., Woodell, G., and Harrah, S. (2005). Real-time enhanced vision system. In *SPIE Enhanced and Synthetic Vision*, Orlando, FL.

(ITU-T), I. T. U. (1999). Subjective video quality assessment methods for multimedia applications.

Jung, I. and Lacroix, S. (2003). High resolution terrain mapping using low altitude aerial stereo imagery. In *International Conference on Computer Vision*, Nice, France.

Kelly et al., A. (2006). Toward reliable off road autonomous vehicles operating in challenging environments. *The International Journal of Robotics Research*, 25(5-6):449–483.

Kobayashi, M. and Kameyama, K. (2010). Partial image retrieval using sift based on illumination invariant features. In *International Conference on Multimedia Computing and Information Technology*, Sharjah, UAE.

Lanir, J., Maltz, M., Yatskaer, I., and Rotman, S. (2006). Comparing multispectral image fusion methods for a target detection task. In *IEEE International Conference on Information Fusion*.

Leonard et al., J. (2008). A perception driven autonomous urban vehicle. *Journal of Field Robotics*, 25(10):727–774.

Lewis, J., OCallaghan, R., Nikolov, S., Bull, D., and Canagarajah, N. (2007). Pixel-and region-based image fusion with complex wavelets. *Information fusion*, 8(2):119–130.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Luo, R. and Su, K. (2003). A multiagent multisensor based real-time sensory control system for intelligent security robot. In *IEEE International Conference on Robotics and Automation*, Taipei, Taiwan.

Mackay, D. (2007). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press.

Maddern, W. and Vidas, S. (2012). Towards robust night and day place recognition using visible and thermal imaging. In *Beyond Laser and Vision: Alternative Sensing Techniques for Robotics Perception, Workshop, Robotics: Science and Systems*, Sydney, Australia.

Martinsen, G., Hosket, J., and Pinkus, A. (2008). Correlating military operators visual demands with multi-spectral image fusion. In *Proceedings of SPIE*.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630.

Montiel, J. M. M., Civera, J., and Davison, A. J. (2006). Unified inverse depth parametrization for monocular SLAM. In *Robotics: Science and Systems*, Philadelphia, PA.

Moreno-Noguer, F. (2011). Deformation and illumination invariant feature point descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Nandhakumar, N. and Aggarwal, J. (1988). Integrated analysis of thermal and visual images for scene interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):469–481.

Narasimhan, S. and Nayar, S. (2003). Shedding light on the weather. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI.

Nayar, S. and Narasimhan, S. (1999). Vision in bad weather. In *International Conference on Computer Vision*, Corfu, Greece.

Nuske, S., Roberts, J., and Wyeth, G. (2009). Robust outdoor visual localization using a three-dimensional-edge map. *Journal of Field Robotics*, 26(9):728–756.

Owens, K. and Matthies, L. (1999). Passive night vision sensor comparison for unmanned ground vehicle stereo vision navigation. In *IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, Fort Collins, CO.

Paz, L. M., Piniés, P., Tardós, J. D., and Neira, J. (2008). Large scale 6DOF slam with stereo-in-hand. *IEEE Transactions on Robotics*, 24(5):946–957.

Pearsall, J., editor (1999). *The concise Oxford dictionary.* Oxford Univ. Press, Oxford, UK, 10th edition.

Perbet, J., Baron, L., Parus, R., and Quancard, S. (1993). Enhanced vision systems. In *International Symposium on Head Up Display, Enhanced Vision and Virtual Reality*, Amsterdam, Netherlands.

Peynot, T. and Kassir, A. (2010). Laser-camera data discrepancies and reliable perception in outdoor robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan.

Peynot, T., Scheding, S., and Terho, S. (2010). The Marulan Data Sets: Multi-Sensor Perception in Natural Environment with Challenging Conditions. *International Journal of Robotics Research*, 29(13):1602–1607.

Peynot, T., Underwood, J., and Scheding, S. (2009). Towards reliable perception for unmanned ground vehicles in challenging conditions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Saint Louis, MO.

Roberts, J., Tews, A., and Nuske, S. (2008). Redundant sensing for localisation in outdoor industrial environments. In *6th IARP/IEEE-RAS/EURON Workshop on Technical Challenges for Dependable Robots in Human Environments.*

Sadjadi, F. (2005). Comparative image fusion analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Scandaliaris, J. and Sanfeliu, A. (2010). Discriminant and invariant color model for tracking under abrupt illumination changes. In *International Conference on Pattern Recognition*, Istanbul, Turkey.

Scanlan, J., Chabries, D., and Christiansen, R. (1990). A shadow detection and removal algorithm for 2-d images. In *International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM.

Solà, J., Monin, A., Devy, M., and Vidal-Calleja, T. (2008). Fusing monocular information in multi-camera SLAM. *IEEE Transactions on Robotics*, 24(5):958–968.

Soleimanpour, S., Ghidary, S., and Meshgi, K. (2008). Sensor fusion in robot localization using ds-evidence theory with conflict detection using mahalanobis distance. In *IEEE International Conference on Cybernetic Intelligent Systems*, Chengdu, China.

Tan, R. (2008). Visibility in bad weather from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK.

Thrun, S. (2006). Winning the darpa grand challenge: A robot race through the mojave desert. In *IEEE/ACM International Conference on Automated Software Engineering*, Tokyo, Japan.

Thrun et al., S. (2007). *Stanley: The robot that won the DARPA Grand Challenge*. Springer.

Toet, A. (2003). Natural colour mapping for multiband nightvision imagery. *Information Fusion*, 4(3):155–166.

Toet, A., Valeton, J., and van Ruyven, L. (1989). Merging thermal and visual images by a contrast pyramid. *Optical Engineering*, 28(7):789–792.

Torr, P. H. and Murray, D. W. (1997). The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300.

Torr, P. H. S. (2002). Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61.

Toth, D., Aach, T., and Metzler, V. (2000). Illumination-invariant change detection. In *IEEE Southwest Symposium on Image Analysis and Interpretation*.

Tumblin, J. and Rushmeier, H. (1993). Tone reproduction for realistic images. *IEEE Computer Graphics and Applications*, 13(6):42–48.

Underwood, J. (2009). *Reliable and safe autonomy for ground vehicles in unstructured environments*. PhD thesis, University of Sydney, Sydney, Australia.

Urmson et al., C. (2008). Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466.

Vedaldi, A. and Fulkerson, B. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. http://www.vlfeat.org/.

Vidal-Calleja, T. A., Berger, C., Sola, J., and Lacroix, S. (2011). Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain. *Robotics and Autonomous Systems*, 59(9):654–674.

Wang, Z. and Bovik, A. (2006). *Modern Image Quality Assessment*. Morgan & Claypool.

Ward, G. (1994). *A contrast-based scalefactor for luminance display*. Academic Press Professional, Inc., San Diego, CA.

Waxman, A., Aguilar, M., Fay, D., Ireland, D., and Racamato, J. (1998). Solid-state color night vision: fusion of low-light visible and thermal infrared imagery. *Lincoln Laboratory Journal*, 11(1):41–60.

Waxman, A., Gove, A., Fay, D., Racamato, J., Carrick, J., Seibert, M., and Savoye, E. (1997). Color night vision: opponent processing in the fusion of visible and ir imagery. *Neural Networks*, 10(1):1–6.

Winkler, S. (2005). *Digital Video Quality: Vision Models and Metrics*. John Wiley & Sons Ltd, New Delhi.

Yu, Y., Huang, K., Chen, W., and Tan, T. (2012). A novel algorithm for view and illumination invariant image matching. *IEEE Transactions on Image Processing*, 21(1):229–240.

Zabih, R. and Woodfill, J. (1996). A non-parametric approach to visual correspondence. In *IEEE Transactions on Pattern Analysis and Machine intelligence*.

Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000.