INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS
VOL.16,NO.2,JUNE 2011,142-159

Visual Metrics for the Evaluation of Sensor Data Quality in Outdoor Perception

Christopher BRUNNER, Thierry PEYNOT and Teresa VIDAL-CALLEJA

Abstract— **This paper proposes an experimental study of quality metrics that can be applied to visual and infrared images acquired from cameras onboard an unmanned ground vehicle (UGV). The relevance of existing metrics in this context is discussed and a novel metric is introduced. Selected metrics are evaluated on data collected by a UGV in clear and challenging environmental conditions, represented in this paper by the presence of airborne dust or smoke. An example of application is given with monocular SLAM estimating the pose of the UGV while smoke is present in the environment. It is shown that the proposed novel quality metric can be used to anticipate situations where the quality of the pose estimate will be significantly degraded due to the input image data. This leads to decisions of advantageously switching between data sources (e.g. using infrared images instead of visual images).**

Index Terms— **Perception, Computer Vision, Visual/Infrared Camera, Unmanned Ground Vehicle, Quality Metrics.**

1. INTRODUCTION

1.1. Problem Statement

The purpose of this work is to promote integrity and reliability in perceptual systems, with a focus on perception for unmanned ground vehicles (UGVs). Perception can be defined as the interpretation of sensor data to produce a representation of the environment that is used to achieve a task. It is arguably one of the most critical components of an autonomous vehicle as this is the first element in contact with the environment and its output is fundamental for all other components needed to ensure autonomy. Considerable progress has been achieved over the last decades to obtain perception algorithms that can handle the uncertainty in sensor data. By rigorously modelling these uncertainties, accurate solutions can be obtained in most regular cases [1]. Nevertheless, the main difficulty remains the interpretation of sensor data. The most significant perception errors are often caused by aspects that cannot be modelled systematically (e.g. interpretation errors due to the presence of a dust cloud obscuring the environment).

This paper proposes an experimental study of quality metrics that can be applied to visual and infrared (IR) images acquired from cameras onboard a UGV. The relevance, in the context of UGV applications, of various visual metrics that can be found in the literature of the television and video industries is discussed. This leads to a selection of potentially appropriate metrics depending on the application to be implemented. Following previous work from the authors [2], which considered a few information-based metrics, namely: Shannon Information (ShI), Spatial Information (SI) and Temporal Information (TI), the selected metrics were evaluated on data collected by a UGV in clear and challenging environmental conditions, represented in this paper by the presence of airborne dust or smoke. The metrics are evaluated for their effectiveness in detecting challenging conditions and identifying their impact on sensing data quality for UGVs. Additionally, a novel metric is introduced to overcome some identified limitations of SI.

Finally, an example of application to a UGV mission is provided. Monocular simultaneous localisation and mapping (SLAM) is used to estimate the pose of the UGV while smoke is present in the environment. It is shown that the novel metric proposed in this paper can be used to anticipate situations where the quality of the pose estimate will be significantly degraded due to the impact smoke has on the quality of the input visual data. This leads to decisions of advantageously switching between data sources (e.g. using IR images instead of visual images).

The paper is organised as follows. The following Section 1.2 discusses existing metrics in the literature and Section 1.3 discusses the concept of image quality in the context of UGV perception. Section 2 describes the experiments used to analyse the various metrics in Section 3. Section 4 further discusses the interpretation of the metrics. An example of application to monocular SLAM is then proposed in Section 5. Finally, Section 6 draws some conclusions.

1.2. Related Work

The television and video industries have been developing "quality metrics" to attempt to quantify objectively how a human viewer would evaluate the quality of a video stream or image [3, 4]. While the metrics are generally developed to capture the errors caused by compression and transmission and are frequently tailored to the human vision system (HVS), there are many metrics that can still be relevant to the evaluation of UGV perception quality. For example, pictures that are colourful, well-lit, sharp with high contrasts are considered attractive to humans given the choice of dark, low contrast, blurry pictures. Most of these characteristics are also relevant for perception applications on a UGV.

The TV transmission infrastructure allows for metrics that compare the output to a known reference input (Full-Reference and Reduced-Reference metrics). However, fidelity of a transmitted image rarely correlates to the perceived quality of the

Manuscript received December 10, 2010, accepted March 31, 2011. This work was supported in part by the Centre for Intelligent Mobile Systems (CIMS), funded by BAE Systems as part of an ongoing partnership with the University of Sydney, and the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government.

C. Brunner, T. Peynot and T. Vidal-Calleja are with the Australian Centre for Field Robotics (ACFR), The University of Sydney, NSW, Australia.

output and, in robotic systems, a reference ground truth is rarely known. Therefore, the subset of metrics known as No-Reference are usually more appropriate to the context of this work. These metrics deal strictly with the quality of an image without relying on any knowledge of what the image should look like.

Video Quality metrics can be further classified into three groups [3, 4]. Data Metrics or "the Fidelity Approach" are purely Full-Reference metrics as they compare images directly. Feature Extraction Based (FEB) metrics or "the Engineering Approach" evaluate specific distortions in an image that are already known to occur and to degrade quality. Vision Model Based metrics (VMB) or "the Psychophysical Approach" model the HVS and evaluate human physical and/or psychological responses to aspects in an image.

FEB and VMB metrics provide the richest area of potential quality metrics for perception systems. Ironically, recent developments in the field of video quality metrics are less likely to be suitable for robotic perception as their metrics are strongly tailored to the HVS or specific artefacts due to transmission or compression. For example, metrics designed to react to phenomena such as blocking [5, 6, 7] and ringing [8] are not particularly relevant to robotic perception. Similarly, high-level metrics modelled on physical aspects of the HVS [3, 4] were excluded. In the context of UGV perception systems, the proposed selection of metrics include FEB and VMB metrics evaluating Brightness, Contrast, Blur, Sharpness and Spatial Information.

1.3. On Image Quality

Colour and infrared cameras are common sensors on autonomous outdoor robots. Acquired images are used in various crucial high level applications such as localisation, terrain modelling, motion detection, tracking or recognition/classification. Many of the fundamental techniques employed in these applications rely on low-level operations that are often quite similar and can be broken into two families: feature-based methods (FBM) and area-based methods (ABM). FBMs actively identify features such as edges, corners, ridges, blobs or shapes/segments. They are typically used in applications such as recognition (e.g. path extraction), sparse stereovision, visual SLAM and visual odometry. ABMs directly analyse the intensity in the images without exploiting the saliency of objects. They use criteria similar to a correlation, a Fourier transform or Mutual Information. Examples of ABM applications are dense stereovision and motion estimation using optical flow.

A good quality image for a UGV perception system is one that captures sufficient required information about the environment and allows the application to perform the task without failure. In this context, quality is degraded when the image data do not match the environmental ground truth and/or the environment itself does not contain enough information to perform the task. As quality is application-dependent, metrics should be analysed considering their relation with the performance of the two categories of applications: FBM and ABM. The experimental setup used for this analysis is detailed in the next section.

Fig. 1. The Argo UGV sensing the *static* trial area. Representative images from the visual and infrared sensors can be found in Figs. 2 and 3

2. Experimental Setup

In previous work [9], synchronised multi-sensor data were collected from a stationary vehicle observing a "reference" scene (see Fig. 1) in controlled and variable environmental conditions. These included challenging environmental conditions, represented by the presence of airborne dust, smoke or rain. The list of sensors included a Prosilica mono-CCD RGB camera acquiring images of resolution 1360×1024 at 15 frames per second (fps) and a Raytheon infrared camera with a spectral response range of $7 - 14\mu m$, using a frame grabber to acquire an average of $12.5fps$. The same static scene was observed with these cameras in clear conditions and (separately) in the presence of airborne dust, smoke, and rain, all at different times of the day. Other data sets also figure a moving UGV, experiencing the same type of conditions in an open and unknown environment. Accurate time-stamping of all these visual and infrared images allowed synchronisation of the data a posteriori for this experimental study.

Although the metrics were evaluated on various data sets, two particular sequences of images were chosen to illustrate the utility of the metrics in this paper. The first sequence ($70s$ long) features the presence of variable amounts of airborne dust. The second one ($90s$) features variable presence of smoke. Figs. 2 and 3 show four representative images from both the colour and infrared cameras for the *Dust* and *Smoke* sequences respectively, to demonstrate the characteristic changes in the environment over the course of data collection.

Although the actual correlation of the signals is not tackled in this paper, using common areas of images from different cameras allows us to illustrate the reaction of metrics when applied to data from different sources of information about the same environment. Therefore, a smaller section of the visual camera images was obtained by trimming the images to manually register with the field of view of the infrared camera. The resolution of the visual image was adjusted to match the resolution of the infrared image. A demonstration of the resulting pair of images is shown in Fig. 4. Because the sensors are mounted in different physical positions on the vehicle, the

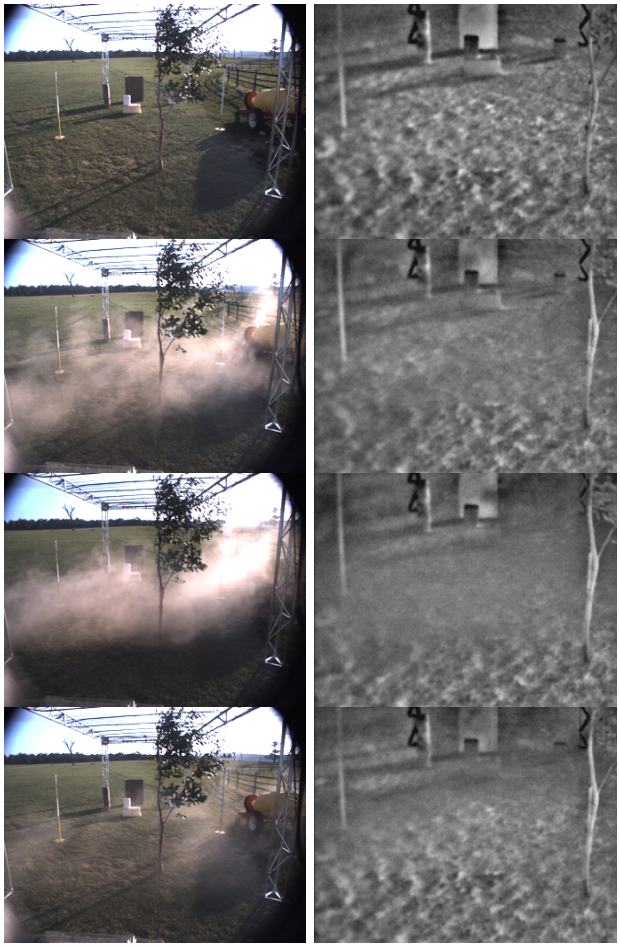


Fig. 2. Representative pairs of colour (left) and infrared (right) images for the *Dust* data set. From top to bottom; Clear conditions at $t = 6s$; Very light dust covering most of image at $t = 11s$; Thick dust cloud at $t = 24s$; Thin dust cloud at $t = 36.2s$.

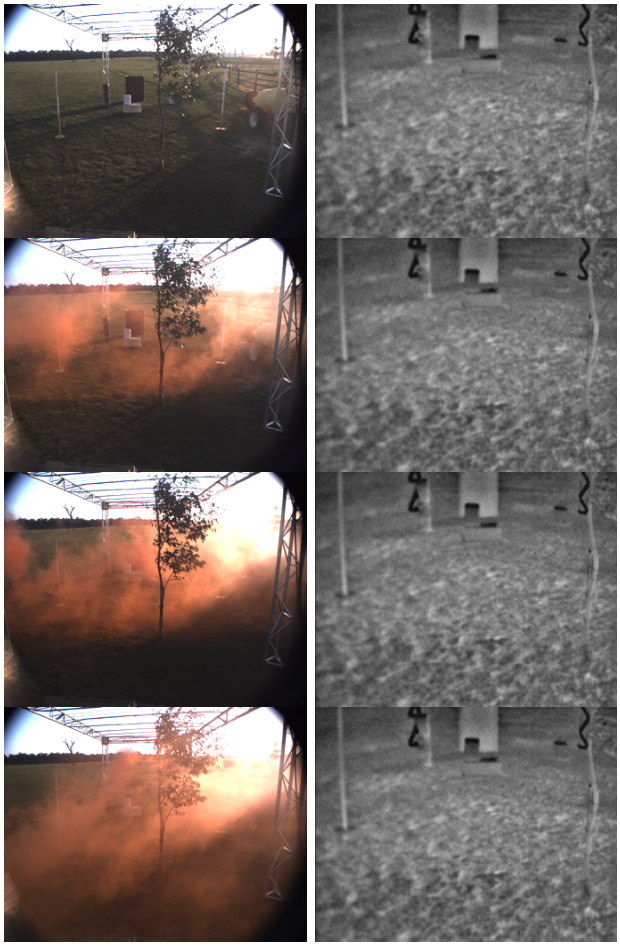


Fig. 3. Representative pairs of colour (left) and infrared (right) images for the *Smoke* data set. From top to bottom; Clear conditions at $t = 1s$; Smoke covering most of image at $t = 50.7s$; Thick smoke cloud at $t = 33.9s$; Thin dust cloud at $t = 28.6s$. Note that smoke is not visible in the infrared images.

positions of objects in the field of view do not necessarily match very accurately. However, the information content of the two images is comparable. Thus, metrics computed on the trimmed images from both cameras will be comparable. Extrinsic calibration between the sensors may be used to further compensate for this perspective difference.

The quantity of dust or smoke in each image of the data sets is illustrated in Fig. 5, showing a direct comparison of individual R,G,B pixel values at any time with those from "reference images" taken from early in the corresponding data set when there are no known environmental impacts. In practice, a pixel is considered as matching the reference if the distance between corresponding points in the RGB space is lower than a pre-defined threshold, chosen to account for noise in the data. These illustrations can be used as a reference when evaluating the quality metrics in the following section. Note that these graphs indicate the proportion of the image that is affected by an obscurant; they do not reflect the relative density of dust or smoke in the environment.

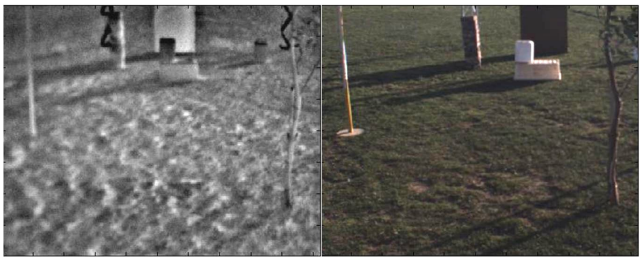


Fig. 4. Representative IR (left) and visual (right) images after the visual image has been trimmed and resized.

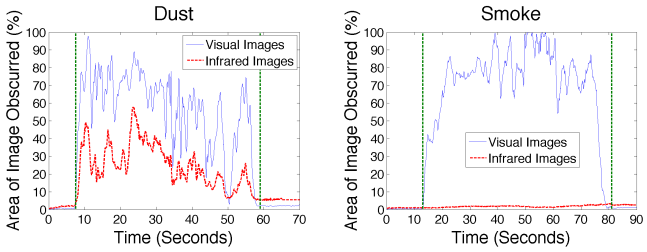


Fig. 5. Percentage of the visual (blue) and infrared (dashed red) image affected by dust (left) and smoke (right) in the course of the representative *Dust* and *Smoke* data sets.

3. Visual Quality Metrics

This section proposes to evaluate a selection of metrics that have been considered potentially relevant for evaluating the appropriateness of data for robotic perception. For each metric considered, the structure of the analysis is the following. Firstly, the metric is introduced and defined. Secondly, following the discussion on image quality in Section 1.3, the contribution of the metric to quality is considered for perception applications divided in two main categories: FBM and ABM. Note that the Monocular SLAM used as an example of perception application in Section 5 is a feature-based method (FBM). Therefore, conclusions drawn for FBM methods below are applicable to our Monocular SLAM application. Finally, results are shown for the metric applied to our datasets of known challenging conditions for perception and a discussion of these results is given.

One of the motivations of this work is to compare the responses of both the IR camera data and the visual camera data in the presence of challenging conditions. Therefore, colour-based metrics were not considered for this study since the IR camera provides gray-scale images only.

3.1. Brightness

Brightness is a measure of the average luminosity of all the pixels in an image.

3.1.1) Contribution to Quality: It is often considered that a bright environment is preferable to a dark environment as objects can be observed more clearly. However, Brightness also needs to be limited to avoid saturation. In general, extremely dark or bright conditions are not desirable for perception applications. For robotic perception, the brightness of an image is less important than for human vision since image processing algorithms rely more on the dynamic range of pixel values than where they lie on the brightness scale. The contrast of an image is discussed in Section 3.2. Typically, changes in Brightness strongly affect all area-based methods. The effect on feature-based methods is much more limited but not necessarily absent if it causes some weak features to be lost. The value of overall Brightness may be relevant to FBM methods only in extreme cases (e.g. at night time or in very low visibility). The main problem with metrics measuring Brightness in the context of outdoor robotics is that they are directly affected by the lighting conditions of the environment. Challenging conditions such as dust and smoke can also influence the Brightness of parts of the image depending on the background environment and the refraction of light.

3.1.2) Discussion: A minimum and a maximum threshold can be set on Brightness to identify extreme situations when an image is not useful for the considered application. Out of these extreme cases this metric is usually not a relevant indicator of image quality in its own, but it could be used in combination with other metrics for the discrimination of situations where apparent variations of quality are in fact only due to a change in lighting conditions.

3.2. Contrast

Contrast is a measure of the relative luminance in an image or region. It can be defined as the difference in brightness of objects within the same field of view. Higher contrast of an image is often associated with better quality as it makes features in the image easier to extract.

3.2.1) Definition: Although various contrast methods can be found in the literature, this experimental study focuses on the *Root Mean Square (RMS) Contrast* [10]. Both a global (i.e. on the whole image) and a local (on patches in the images) method are considered.

By assuming that the histogram of intensities of the pixels in an image can be modelled by a Gaussian distribution, the first standard deviation of this distribution provides a measure of the Contrast of the whole image:

$$C^{RMS} = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{i=0}^{M-1}(L_{ij}-\overline{L})^2} \qquad (1)$$

where L_{ij} is the i^{th} and j^{th} element of the two dimensional image of size $M \times N$ and $\overline{L}$ is the average luminance in the image.

In a more local analysis of Contrast using the same RMS method, for each pixel of the image a local Contrast is computed using a patch of 10×10 neighbouring pixels. The Contrast value for the image is then calculated by averaging all the local Contrast values across the whole image.

3.2.2) Contribution to Quality: Good Contrast is crucial for many feature-based methods, as corners, ridges or edges are identified using the relative intensity of neighbouring pixels. A higher Contrast is also preferable for area-based methods, to have a better signal-to-noise ratio. A minimum of Contrast is usually needed in both cases. Therefore, a corresponding threshold (for a minimum quality) can be defined. However, apart from this extreme case, the variation of Contrast will be relevant to ABM methods only.

Challenging conditions such as dust and smoke partially obscure areas in the image, and therefore affect its global Contrast. Whether it increases or decreases as a result depends on the relative intensity of the dust and smoke compared to the background environment. However, locally, within the smoke/dust cloud, the Contrast is consistently diminished, reducing the quality of corresponding portions of the image.

3.2.3) Experimental Results: Fig. 6 shows the evolution of RMS Contrast for the whole image for the *Dust* and *Smoke* data sets. In these data sets, there is a clear sudden increase in the global RMS Contrast with the appearance of dust and smoke. However, this effect is very strongly dependent on the *relative* intensity of the background and the smoke or dust, and the effect of sunlight scattering from the dust and smoke clouds.

Fig. 7 shows the evolution of Contrast using the local method for the *Dust* and *Smoke* data sets. The average Contrast of the image drops in both *Dust* and *Smoke* data sets as dust or smoke begin to obscure the background of the scene.

3.2.4) Discussion: Without utilising further information, the RMS Contrast of the whole image is a poor method of evaluating the quality of an image, due to the dependence on

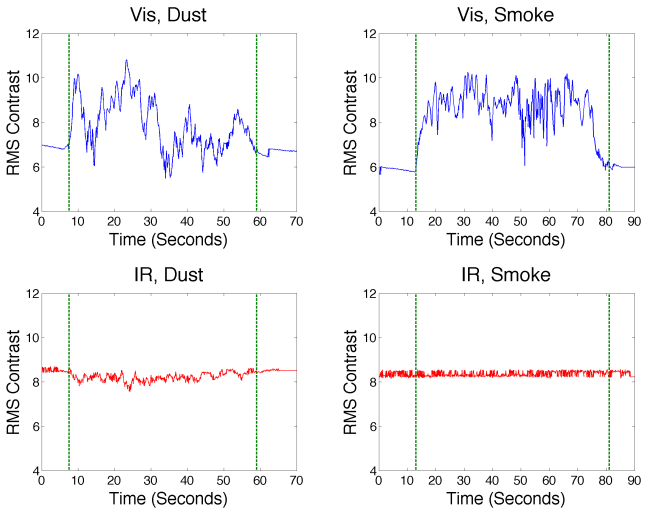


Fig. 6. RMS Contrast measurement of the whole image, for *Dust* (left) and *Smoke* (right). Top line: visual (Vis) camera, bottom line: infrared (IR) camera. Note that hereafter the times of appearance and then disappearance of dust/smoke will be indicated by dashed vertical lines.

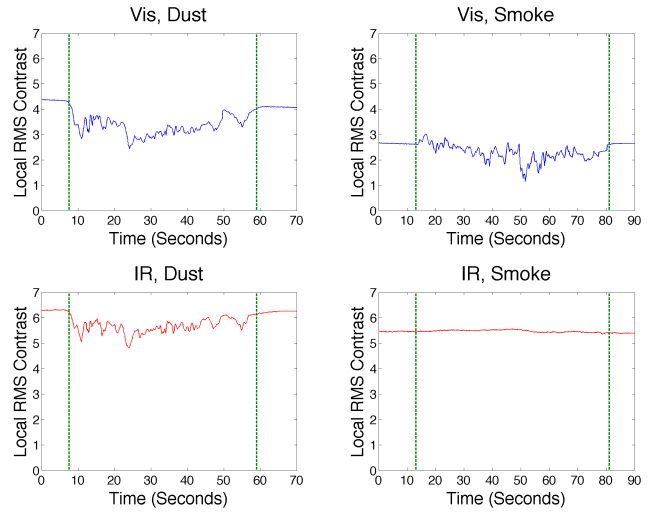


Fig. 7. *Local* RMS Contrast measurement for *Dust* (left) and *Smoke* (right). Top line: visual camera, bottom line: IR camera.

the background characteristics and the Gaussian assumption. The Gaussian approximation for the distribution of intensity values of the pixels means that a few very bright or very dark pixels can cause a large standard deviation when Contrast in much of the image is in fact very low. This metric can only be used by itself if the background is known or the Contrast drops to a critically low value. In the latter case, the image can be judged as poor quality and unlikely to provide any useful data for the perception algorithms considered in our context.

Calculating the *local* RMS Contrast at a pixel-by-pixel level using the method described above has a very high computational cost. Thus, a more appropriate method for real-time applications would use regions of interests (ROI) defined by specifying an appropriate size for a sub-image or by focussing on areas where challenging conditions are expected to appear (if such information is available) and checking the evolution of Contrast in them.

3.3. Shannon Information

3.3.1) Definition: The Shannon Information (ShI) of an image is defined as the measure of entropy of the distribution of intensities in the luminosity image. The image I is composed of pixels with a discrete set of possible intensity values ($i \in A_I$). If the probability of observing any particular intensity value, i, in the image is given by $P(i)$, ShI is defined as [11]:

$$ShI(I) = \sum_{i \in A_I} P(i) log_2 \frac{1}{P(i)} \tag{2}$$

and is expressed in average bits of information per observation.

The more variety in the intensities in the image observation, the more information content is considered to be contained within that image. A uniform distribution for all possible intensities corresponds to the maximum entropy (i.e. maximum amount of information) that is possible for a whole image.

3.3.2) Contribution to Quality: A broad spectrum of intensity values is important for many feature-based methods as they rely on differentiation in an image using intensity values. A decrease in ShI indicates that the pixel intensities in the image are becoming more homogeneous. Images that are too homogeneous are less likely to be useful for perception.

Challenging conditions such as dust and smoke partially or totally obscure background features in the environment. Therefore, in those regions that are obscured, the amount of Shannon Information tends to decrease as the luminosity values become more homogeneous. However, unless the obscurant is similarly covering the entire background, it may add to the amount of information within the image as it contrasts with the background. Similarly to Brightness and Contrast, as ShI is directly linked to the intensity levels in the image, this metric is relevant to ABM methods but its utility is more limited for FBM methods.

3.3.3) Experimental Results: Fig. 8 shows the evolution of Shannon Information for the *Dust* and *Smoke* data sets. In the visual images, the appearance of both dust and smoke causes an increase in Shannon Information for the visual images. However, the dust causes a decrease in the ShI of the infrared data. The cause of this difference is that, in this case, the dust and smoke are relatively bright in the visual images in comparison to the background environment and therefore add a broader range of pixel intensities to the distribution than those that are obscured. Alternatively, for the case of dust in the infrared images, the dust clouds are approximately at the same average temperature as the background environment and are relatively homogeneous compared to the background. Therefore, they reduce the overall distribution of intensities in the image.

3.3.4) Discussion: Without context or further information from another metric, Shannon Information on its own is a poor method of evaluating the quality of an image except in extreme cases when most of the image is close to the same luminosity. The ShI measure is too dependent on the background environment to be useful to discriminate challenging

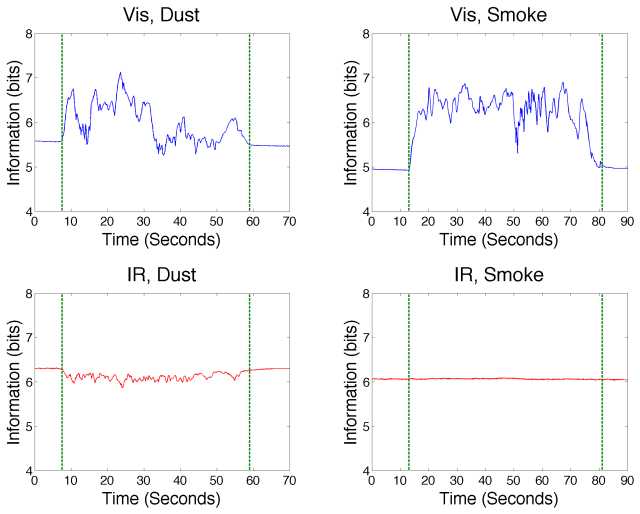


Fig. 8. Shannon Information measurement of the whole image, for *Dust* (left) and *Smoke* (right). Top line: visual camera, bottom line: IR camera.

conditions, which have been shown to increase the level of ShI despite obscuring the background. Further discussion of Shannon Information in relation to challenging conditions can be found in [2].

3.4. Blur

Blurred features are harder to differentiate as the boundaries become smeared. This may lead to difficulties in image analysis and scene interpretation.

3.4.1) Definition: Among the different techniques in the literature, *Marziliano Blur* [8] was identified as a method of measuring Blur that is quick to compute and intuitive. The method is as follows: first, strong vertical edges are identified by applying a vertical Sobel filter to the image and then thresholding to eliminate noise and low intensity edges. In this case, the Sobel-filtered image is scaled so the maximum possible intensity is the same as in the original image (255 for an 8-bit image). A threshold of 50 was found to eliminate most of the noise and insignificant edges in the visual image while retaining relevant edges from objects in the environment. The same threshold was used for infrared images. Second, each row of the processed image is scanned for pixels corresponding to an edge location. The start and end positions of the edge in the horizontal row provide a spatial reference position of the edge in the original image. Third, for each location of an edge, the local maximum and minimum of luminosity values are found along the horizontal rows of the original image. The distance between these local extrema, expressed in number of pixels, is considered the local Blur value. The global Blur value (measured in pixels) is then found by averaging the local Blur measurements over all suitable edge locations:

$$Blur = \frac{1}{N}\sum_{i=1}^{N}(d_i) \tag{3}$$

where N is the number of edge positions used to calculate the Blur and d_i is the distance in pixels between the two local extrema of luminosity in the original image around edge i. A more detailed discussion of the process can be found in [8].

3.4.2) Contribution to Quality: The blurriness of the image can strongly affect feature-based methods as it reduces the saliency of objects. On the contrary, the effect of blurriness on area-based methods is limited. In normal conditions, the Blur from a sensor remains relatively constant regardless of the background environment and changes to lighting conditions.

3.4.3) Experimental Results: Fig. 9 shows the evolution of Blur for the images in the *Dust* and *Smoke* data sets. In both

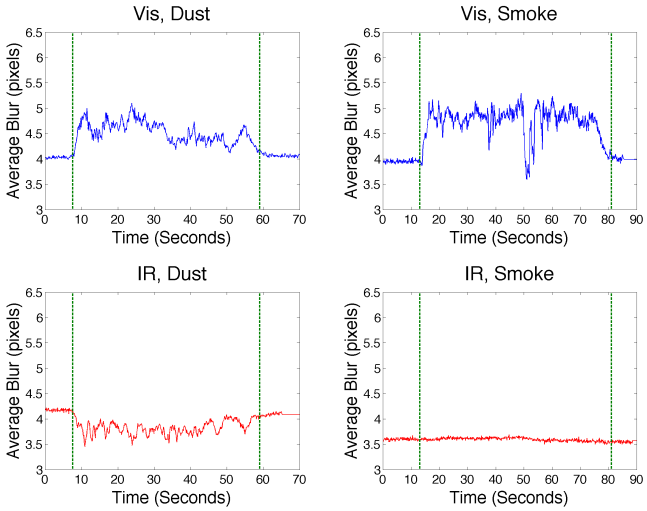


Fig. 9. Blur measurement (in pixels) for *Dust* (left) and *Smoke* (right). Top line: visual camera, bottom line: IR camera.

Dust and *Smoke* data sets there is a characteristic increase in Blur for the visual images in the presence of dust and smoke. The most significant troughs in the Blur signal that can be observed in the visual images in the presence of smoke can be attributed to a significant drop in the number of edges on which Blur is calculated (e.g. see at time $t = 51s$ in Fig. 9). Note that there is a decrease in Blur in the infrared images in the presence of dust. The difference between the Blur in visual and infrared sensors is likely to be related to the intensity threshold that is used to choose edges for the Blur calculation.

3.4.4) Discussion: In challenging conditions such as dust and smoke, the overall Blur of the visual images is seen to increase significantly. By setting an upper threshold on the Blur, challenging conditions could be identified. However, as mentioned above, the value of Blur highly depends on the threshold applied to the edge image, and on the number of edges considered in the calculation of the metric, as a result. In the case of the IR images, the signal-to-noise ratio is much lower. When the background is obscured by challenging conditions, strong edges are dimmed or lost, and small edges due to noise become dominant in the calculation of Blur, resulting in an increase of the metric. This makes this Blur metric difficult to use for the IR camera in its current form.

3.5. Sharpness

Sharpness (or acutance) [12] describes the rate of change of luminosity with respect to spatial position.

3.5.1) Definition: The Sharpness of an image is found by averaging the gradient between neighbouring cells [12].

$$Gx^2 = \sum(\frac{\Delta I^2}{n}) \tag{4}$$

$$Acutance = (Gx^2/I_0) \times C \tag{5}$$

where ΔI is the difference in the grey scale value between a pixel and each of the 8 surrounding pixels; n is the total number of contributing values, that is, the number of pixels multiplied by 8; I_0 is the mean luminosity value of the image; and C is a scaling factor.

3.5.2) Blur and Sharpness: Blur (Section 3.4) and Sharpness are metrics designed to measure a similar aspect of images: the rate of change of luminosity. However, the method of calculating the metrics are subtly different and, while often correlated, Blur and Sharpness can provide different results. Sharpness measures the rate of change of luminosity between *all* neighbouring cells in an image. Often an image with noise will be perceived as being sharper than one without and the Sharpness metric tries to capture this. Alternatively, the Blur metric process extracts specific features (i.e. strong edges) and uses only these to calculate the blur in the image.

3.5.3) Contribution to Quality: Sharpness strongly affects feature-based methods, in particular those using features such as edge or corner detectors. Area-based methods do not rely on a sharp image. However, both categories of methods (FBM and ABM) experience difficulties with image sequences when Sharpness changes rapidly. Note that Sharpness is dependent on the focus of the sensor being used. The appearance of challenging conditions are shown to decrease the Sharpness of images as the background edges are dulled.

3.5.4) Experimental Results: Fig. 10 shows the evolution of Sharpness for the images in the *Dust* and *Smoke* data sets.

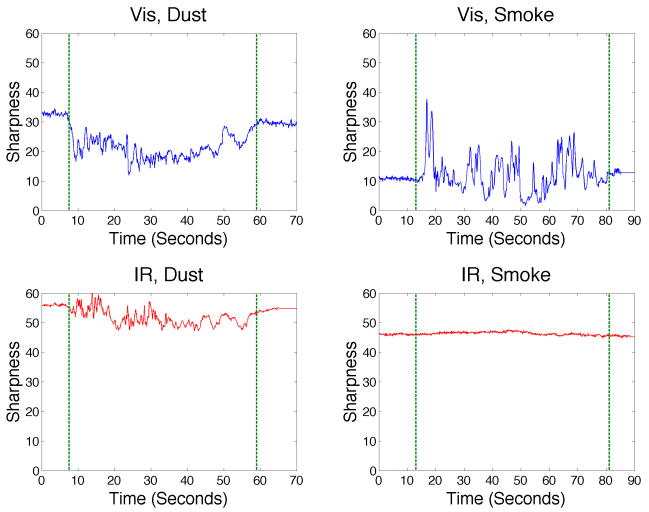


Fig. 10. Sharpness for *Dust* (left) and *Smoke* (right). Top line: visual camera, bottom line: IR camera.

3.5.5) Discussion: The Sharpness method uses a simple edge detector and averages the intensities of the edges over the whole image. Indeed, the response of the Sharpness metric (Fig. 10) to both the *Dust* and *Smoke* data sets is very similar to the SI metric (see below). Averaging the intensities of an edge-filtered image provides no more information than can be found using Spatial Information and Spatial Entropy, which have been preferred, as explained below. Therefore, this metric was not selected for our applications.

3.6. Spatial Information

In previous work [2], among the existing information-theory based metrics, Spatial Information (SI) was found to be the most promising one in the context of perception in challenging environmental conditions. SI is evaluating the amount of structure in an image, and so it can be applied in the same way to heterogeneous sensors such as a visual camera and an infrared camera. However, in this section we show the limitations of SI, mainly due to the Gaussian distribution assumption.

3.6.1) Definition: To compute SI, an edge detector such as a Sobel filter is first used on the input image. SI is then defined as the first standard deviation of the resulting distribution of intensities in the Sobel image, i.e. the intensities of edges in the original image. More specifically, the image I is composed of N pixels with a range of intensity values. Similarly, the Sobel-filtered image is composed of N pixels with a discrete set of possible intensity values ($i \in Sob_I$). If the average intensity value of all the pixels in Sob_I is μ, then the standard deviation is:

$$SI(I) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(i_n - \mu)^2} \tag{6}$$

where i_n is the intensity value of the n^{th} pixel. Spatial Information is expressed as an intensity.

3.6.2) Contribution to Quality: SI measures the amount of structure in an image. As such, it is particularly relevant to feature-based methods of perception. Setting a minimum threshold of SI can allow identification of when an image is unlikely to be useful for applications using feature-based methods. On the other hand, SI has little relevance to area-based methods. Challenging conditions such as dust and smoke are shown to reduce the value of SI when they are in the foreground and obscure the background environment.

3.6.3) Limitations of SI: SI was developed by the television and video industries to measure the amount of structure in an image. It is one of the very few international standard metrics to measure the quality of an image [13]. However, the validity of SI relies on the assumption that the distribution of intensities in the Sobel-filtered image can be modelled as a Gaussian, with an average close to zero. In that case, the distribution can be characterised by its standard deviation. In most edge-filtered-images captured in normal outdoor environments, this assumption is found to be reasonable. Since dust or smoke clouds tend to obscure features in the background and contain very little structure, SI was shown to be a useful tool to monitor the appearance of such environmental conditions [2], at least in cases where dust or smoke was shown to dim or obscure most background features, i.e. when the Gaussian assumption was acceptable.

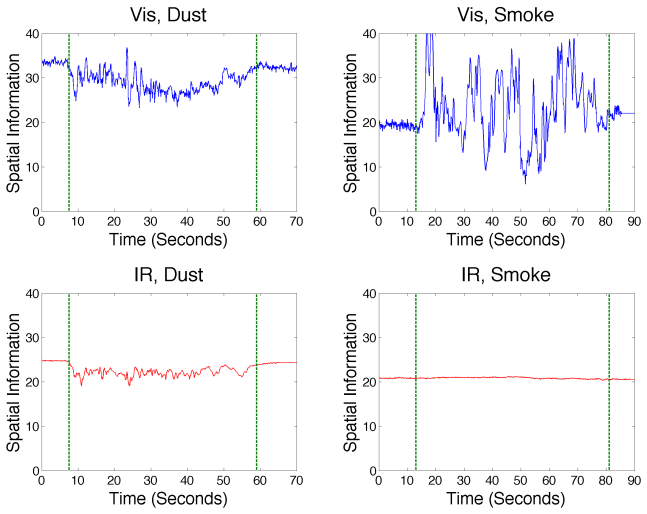


Fig. 11. Evolution of SI for *Dust* (left) and *Smoke* (right), Visual (top row) and IR Camera (bottom)

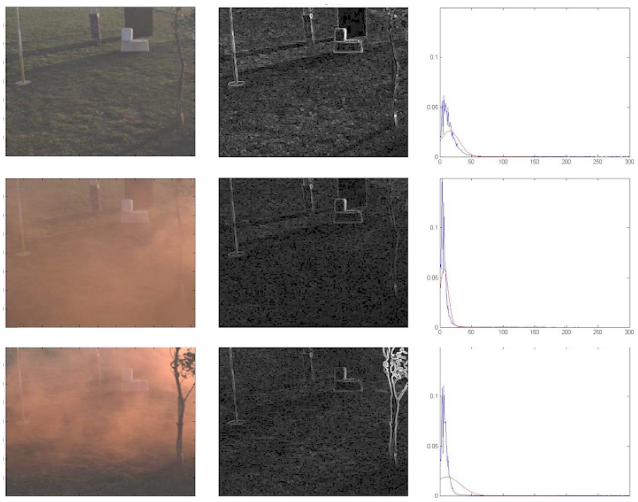


Fig. 12. Representative images (left column) for *Smoke* with Sobel-filtered image (middle column). Right column: corresponding distribution of edge intensities (in blue) and Gaussian approximation of this distribution (in red). Clear Conditions at $t = 1s$; Smoke covering most of image at $t = 50.7s$; Thick smoke cloud at $t = 33.9s$.

However, in some situations, e.g. in the presence of foreground objects, the known challenging conditions can actually highlight the edges of these foreground objects while still obscuring background features. An example is shown in the *Smoke* data set where the smoke obscures much of the background but mostly stays behind the tree that is in the foreground on the right side of the image (see Fig. 12). In this situation the edges of the tree are highlighted as the tree contrasts more with the smoke behind it than with the original background. These high intensity edges have a strong impact on SI (see the high peaks in Fig. 11, right column), making it increase significantly, despite the reduction of structure everywhere else in the image. This same condition is also observed for short moments in the *Dust* data set (most notably at the spike at $t = 24s$). Fig. 12 shows that in these situations the Gaussian assumption is clearly not valid, which means the first standard deviation (and therefore SI) does not characterise the distribution, i.e. the actual amount of structure in the image.

3.7. Spatial Entropy

To overcome the identified limitations of SI brought about due to the Gaussian assumption, we introduce a new metric that we call *Spatial Entropy* (SE).

3.7.1) Definition: SE models the intensity distribution of an edge-filtered image using entropy. The entropy of an edge-filtered-image measures the variety of edge intensity values without giving weight to the magnitude of the intensity values, as happens with SI. An image I is composed of pixels with a range of intensity values. Similarly, the Sobel-filtered image is composed of pixels with a discrete set of possible intensity values ($i \in Sob_I$). If the probability of observing any particular intensity value, i, in the Sobel-filtered image is given by $P(i)$, SE is defined as [11]:

$$SE(I) = \sum_{i \in Sob_I} P(i) log_2 \frac{1}{P(i)} \tag{7}$$

and is expressed in average bits of information per observation. By using entropy, no assumption on the shape of the distribution is made for this metric.

3.7.2) Contribution to Quality: SE measures the amount of structure in an image, which is a relevant metric for feature-based methods of perception but has little relevance to area-based methods. Setting a minimum threshold of SE can allow identification of when an image is unlikely to be useful for applications using feature-based methods. Challenging conditions such as dust and smoke are shown to reduce the value of SE.

SE can be used to help discriminate the challenging conditions that SI failed to identify. Furthermore, we show that combining SI and SE can contribute to discriminating more situations. In situations where most features are behind the obscurant, SI and SE both decrease and jointly confirm that features are being dimmed or lost (e.g. see the *Dust* dataset). However, SI and SE disagreeing usually means the Gaussian approximation of SI is not appropriate. For example if SI is increasing but SE is decreasing, then some features in the environment are being obscured while others are becoming more intense (as in Fig. 12).

3.7.3) Experimental Results: Examples of the evolution of Spatial Entropy for both *Dust* and *Smoke* data sets are shown in Fig. 13. The value of SE consistently decreases when dust appears and spreads for visual and infrared images and for smoke in visual images. Note that the highest peaks in the amount of dust and smoke, as seen in Fig. 5, correspond to the lowest points in the SE evolution, as observed in Fig. 13. Once dust and smoke have cleared, SE returns to a nominal value corresponding to clear conditions. For clear and unchanging conditions, such as seen in the first 8 seconds of both data sets, the value of SE is stable.

3.7.4) Discussion: Very low values for Spatial Entropy usually indicate that there is very little information in the image. However, a black and white image of edges (e.g. a

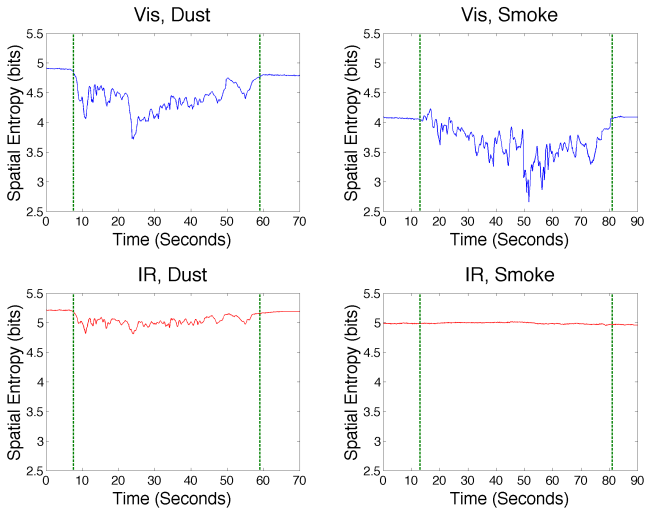


Fig. 13. Evolution of SE for *Dust* (left) and *Smoke* (right), Visual (top row) and IR Camera (bottom)

checker board) may contain a great deal of useful information for perception but because it uses only two values for intensity, its entropy will be very low. In natural environments such a situation is extremely unlikely, making SE relevant on its own. However, it should be noted that in this case the value of SI will be high. Therefore, both SI and SE can be used in conjunction to differentiate such situations.

When conditions are clear, both SI and SE are stable. When smoke is covering most of the image, including the tree in the foreground (e.g. at $t = 50.7s$ in Fig. 13), this corresponds to a clear drop in both SI and SE. However, when a thick cloud of smoke is covering much of the background but passes behind the tree in the foreground ($t = 33.9s$), SI increases rapidly while SE decreases. Table I summarises the conditions that can be discriminated by monitoring the relationship between SI and SE metrics.

4. METRICS INTERPRETATION

4.1. Suitable Metrics for Robotic Applications

In the previous section, existing visual quality metrics were considered for appropriateness in evaluating image data quality for robotic perception particularly in challenging conditions such as dust and smoke.

The metrics Brightness, Contrast, Sharpness and Shannon Information are based on the relative intensity and distribution of pixels in an image. They were identified as important for area-based methods of perception. However, while potentially useful in extreme situations when the environment is entirely obscured, these metrics are heavily dependent on the original background and the lighting conditions, as well as how an obscurant interacts with light in the environment. Therefore, in outdoor robotics it is difficult to interpret much based on these metrics alone as their reactions to challenging conditions can be variable. The utility of these metrics for feature-based methods was found to be limited to extreme cases.

The Blur metric actively identifies edges and then measures how wide these edges are. A high level of blur will negatively affect feature based methods of perception but not be as much of a hindrance to area-based methods providing the blur is consistent. While this seemed a promising metric to identify when images are being degraded for robotic perception, significant tuning of parameters is required depending on the specific sensor to account for noise and filter out weak edges.

Spatial Information (SI) detects the amount of structure in an image. The amount of structure in an image can be a strong indicator for how well a feature-based method of perception, such as monocular SLAM, would perform, since FBMs actively search for certain structured features in an image. In most cases SI was shown to be a very useful indicator of degraded perception conditions as challenging conditions such as dust and smoke tend to obscure the background detail in the environment. However, in some observed cases, SI had the opposite effect as background features were lost because foreground features were simultaneously emphasised. This was due to the invalidity of the Gaussian distribution assumption.

SE measures the amount of structure in an image without making a Gaussian assumption (i.e. it does not give extra weight to small but intense edge features). SE was consistently shown to decrease when challenging conditions such as smoke and dust appear and/or get stronger. Thus, SE appears as the most likely metric to aid in detecting when image data is degraded, particularly for feature-based methods, for which the structure aspect is extremely relevant.

Following the conclusions of this metric analysis, for the example application of vision-based SLAM discussed in Section 5, we have chosen to use SE to evaluate the quality of sensing data, as it has the most potential to capture the effect of challenging conditions on images and correlate to a degradation in the performance of monocular SLAM, an example of a feature-based method.

4.2. Metrics Evaluation

We consider three main ways to use these metrics to evaluate image quality, in particular in the presence of challenging conditions. The first is to check a single metric value, the second is to monitor its evolution and the third is to compare the relative evolution of multiple metrics.

As discussed for each metric above, due to the generality of metrics and the diversity of environments UGVs operate in, it is rarely possible to discriminate poor quality images from any single metric value, except in extreme cases. Only when the value for the metric is approaching maximum or minimum values, for example if Contrast drops below a critical level, can it definitively indicate that the sensor data is unlikely to be useful for further perception.

The evolution of an individual metric can be used to monitor changes in the environment. Since unmodelled changes in environments will effect perception applications, unexpected changes in a metric may indicate that challenging conditions are occurring. For example, a sudden reduction in SI occurs when features are obscured or lost from the perceived environment (if other features in the foreground are not simultaneously emphasised by the obscurant, as discussed earlier). Similarly, most metrics were unstable under challenging conditions

TABLE I

THE RELATIONSHIP IN THE EVOLUTION OF SI AND SE CAN BE USED TO DISCRIMINATE SITUATIONS.

NORMAL TEXT: WHAT HAPPENS IN THE IMAGE. *Emphasised text: meaning in the case of challenging conditions.* ↗ AND ↘ STAND FOR "INCREASES" AND "DECREASES", RESPECTIVELY.

	SI ↗	**SI ↘**
SE ↗	Amount of structure is increasing *Less Dust/Smoke in front of most objects*	General amount of structure is increasing but some strong edges are getting weaker *Less Dust/Smoke in background. There are objects in front of the cloud*
SE ↘	General amount of structure is decreasing but some edges are getting stronger *More Dust/Smoke in background. There are objects in front of the cloud*	Amount of structure is decreasing *More Dust/Smoke in front of most objects*

compared to normal conditions. However, due to the motion of the UGV, the metric will evolve according to the change in the area of the environment which is currently perceived. Consequently, a method to compensate for this influence of motion on the metrics is needed. This is demonstrated in Section 4.3.

4.3. Compensation for Motion

To allow for the use of the selected quality metrics when the UGV is moving, a simple method is used to compensate for the motion of the vehicle. The technique approximates this motion between successive frame acquisitions by an affine transformation, which is calculated using sensing data that must be fully independent from the cameras. In this work we used the Inertial Measurement Unit (IMU) which is available on the platform. Once this transformation has been applied, the images can be cropped so that consecutive images approximately register.

Taking into account the field of view and frame rate of the sensors, as well as the operating speeds of our UGV, it was found that five consecutive images guarantee sufficient overlap in images. The evolution of metrics on these overlapping sub-images can then be evaluated for actual changes in quality (e.g. due to challenging conditions) without significant influence from the motion of the vehicle (see Fig. 14).

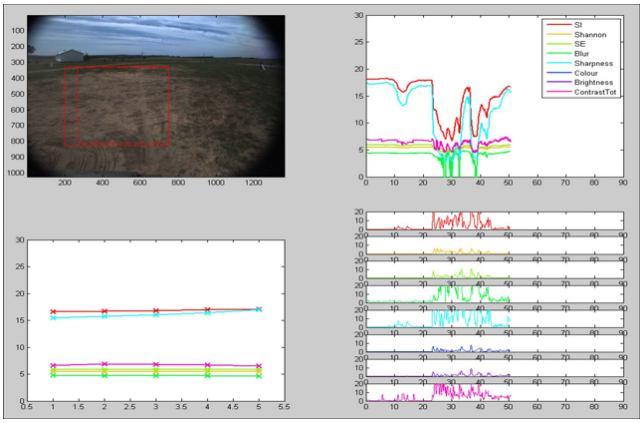


Fig. 14. Evolution of multiple metrics for a moving UGV, in a data set with variable presence of airborne dust. Top left: current frame and evaluated ROI in red. Top right: direct evolution of metrics in the ROI. Bottom left: variation of metrics for the last five overlapping regions. Bottom right: variation of the metrics after compensation for motion.

5. APPLICATION TO MONOCULAR SLAM

In Section 3.6, SE was identified as one of the most promising quality metrics to evaluate image data in the context of UGV missions. Thus, in this section the utility and usage of image quality metrics are illustrated with SE in the presence of smoke for a selected application which is SLAM using monocular vision (hereafter *monocular SLAM*).

A state-of-the-art monocular SLAM technique is applied to visual images and to infrared images, separately. The performance of SLAM using the visual camera (hereafter named *Visual-SLAM*) is compared to the same SLAM algorithm using the infrared camera (hereafter named *IR-SLAM*). It is shown that in clear conditions Visual-SLAM outperforms IR-SLAM, illustrating that in this case the available visual information is of "better quality" than the infrared data. However, in the case of a significant presence of smoke the opposite is observed.

The following shows that by using SE as a quality metric, situations where the quality of IR data clearly becomes higher than Visual data (in this case because the latter is degraded due to the presence smoke) can be detected and the performance of the application can be *anticipated*. As a result, the appropriate sensing modality can be selected *before* the operations of feature extraction and matching have to be performed. This can make the perception application robust to challenging conditions such as the presence of smoke (i.e. conditions that affect some of the sensing modalities but not all of them) while keeping comparable computational performance.

5.1. Monocular SLAM

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. Monocular SLAM gained popularity back in 2003 thanks to the first full real-time implementation by [14], based on [15], the original solution to SLAM which employs an extended Kalman filter (EKF) as the central estimator. Davison's technique elegantly solved a great number of problems, but the remaining one was the problem of landmark

initialisation. The problem was successfully solved with the inverse-depth landmark parameterisation (IDP) by [16].

The core algorithm of this application is a landmark-based EKF-SLAM with IDP parametrisation based on [17]. The only sensor used to produce the vehicle trajectories is the camera, for this reason, a 6-DOF constant velocity model is used to predict the motion. The model based on quaternions for the orientation is

$$\mathbf{r}^+ = \mathbf{r} + \mathbf{v}\Delta t \quad (8)$$

$$\mathbf{q}^+ = \mathbf{q} \times v2q(\omega\Delta t) \quad (9)$$

$$\mathbf{v}^+ = \mathbf{v} + a \quad (10)$$

$$\omega^+ = \omega + \alpha \quad (11)$$

where $()^+$ means the forward predicted value, $\times$ is the quaternions product, and $v2q(\omega\Delta t)$ transforms the local incremental rotation vector $\omega\Delta t$ into a quaternion (quaternions are systematically linearised). At each time step, perturbations $a, \alpha \sim \mathcal{N}(0; \sigma_v^2, \sigma_\omega^2)$ add variances to the linear and angular velocities proportionally to the elapsed time Δt. Note that the solution obtained using only a single camera, with no aid of other sensors, is subject to scale.

a) Inverse-Depth Parametrisation: The IDP is encoded by the direction vector from the current camera position $\mathbf{r}_0$ to the observed point $\mathbf{p}$, with just elevation and azimuth angles (ε, α) of the observed optical ray joining $\mathbf{r}_0$ to $\mathbf{p}$. 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 $\mathbf{r}_0$ as an anchor to improve the linearity leads to the 6D-vector,

$$L = \begin{bmatrix} \mathbf{r}_0 \\ (\varepsilon, \alpha) \\ \rho \end{bmatrix} = \begin{bmatrix} x_0 & y_0 & z_0 & \varepsilon & \alpha & \rho \end{bmatrix}^\top \quad (12)$$

b) Feature Extraction and Map Management: Sparse interest points are extracted using SIFT detectors and matched using SIFT descriptors [18]. The same type of interest points are used for both cameras, Visual and IR.

As proposed in [19], we used the Gaussian expectation of the visible mapped points to reject outliers in the image space. 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 used 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, because the Gaussian expectation already account 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. In this implementation there is no strategy to explicitly enforce loop closures.

5.2. Experimental Setup

For this experiment two datasets from [9] are used. In the first one, the Argo UGV (see Section 2) is driven in a natural environment in *clear conditions*, i.e. in the absence of any of the aforementioned challenging environmental conditions. In the second dataset, the UGV is driven in the same natural scene but in the presence of smoke, generated using a smoke bomb (see details in [9]). The positions of the vehicle given by the cm-accuracy dGPS/INS unit are only used as a reference to compare the performance of the SLAM estimates.

5.3. Visual vs. Infrared

The visual and IR cameras used in the following experiments are the same as described in Section 5.2. In these datasets the framerates are: 10 frames per second (fps) for the visual camera and $12.5fps$ for the IR camera. The horizontal field of view (FOV) is about 68° for the Visual camera and 36° for the IR camera. Consequently, in clear conditions the monocular SLAM application is favourable to the Visual camera, as more information about the motion can be captured with a larger FOV. Furthermore, the signal-to-noise ratio is significantly stronger for the visual images than the IR images.

5.4. Performance of Visual-SLAM and IR-SLAM in clear conditions

Fig. 15 shows the trajectories estimated by the monocular SLAM algorithm using the visual and IR streams of images, separately. They are compared to the dGPS/INS trajectory, used as a reference. To make this comparison, the Visual-SLAM and IR-SLAM trajectories are scaled using the differences of velocities with the reference to recover the scale factor. At the end of the $58.6m$-long trajectory, the Visual-SLAM solution finishes with a 3D position estimate 3 metres different from the dGPS/INS trajectory, while the IR-SLAM ends approximately 8 metres away from the dGPS/INS end point. In terms of the orientation, the Visual-SLAM solution finishes with a 0.14 radians difference from the dGPS/INS measured yaw, compared to 0.2 radians for the IR-SLAM solution.

The evolution of the uncertainties in the position estimates are shown in Fig. 16. Note that the uncertainties grow more rapidly in the case of IR images than Visual. This is mainly due to the difference of the FOV, as well as the quality of the sensors. Also, in this test, a better localisation is obtained using Visual data rather than IR data. Interestingly, along this trajectory, the average number of SIFT point matches (35 for visual images and 32 for IR images) and the average ratio of matches over total number of points (69% for visual images and 72% for IR images) are comparable for both sources of data.

These results show that in clear conditions Visual-SLAM outperforms IR-SLAM (at least for the available sensors), although IR-SLAM performs reasonably well.

5.5. Performance of Visual-SLAM and IR-SLAM in challenging conditions (presence of smoke)

Fig. 17 shows the trajectories estimated by monocular SLAM using the visual and IR streams of images, separately,

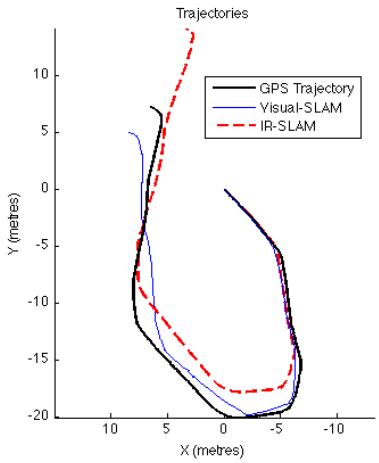


Fig. 15. Estimated trajectories using Visual-SLAM (blue), IR-SLAM (dashed red) and the measured $58.6m$ long dGPS/INS path (thick black), in clear conditions.

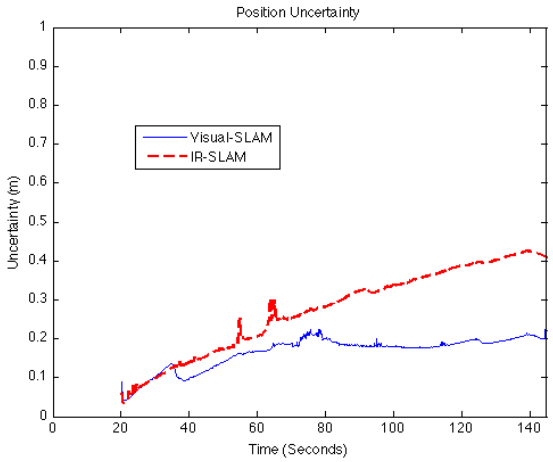


Fig. 16. Evolution of the uncertainties for Visual-SLAM (blue) and IR-SLAM (dashed red), in clear conditions. For both Visual and IR, the value shown over time is the square root of the sum of eigenvalues of the position covariance matrix.

for the dataset with smoke present[1]. They are also compared to the dGPS/INS trajectory. As in clear conditions, the SLAM trajectories are scaled using the differences in velocity with respect to the reference.

Fig. 18 illustrates the variable amount of smoke that can be seen in the visual images during the sequence. It was obtained by identifying the number of pixels in an image that had a colour corresponding to the reference colour for this particular smoke[2]. This amount of smoke is then reported on the estimated trajectory in Fig. 19 for Visual-SLAM and Fig. 20 for IR-SLAM.

In this dataset, the pose estimation actually starts at $t = 22s$. The original position at this time was set to $(0, 0)$ for illustration. Initially, the UGV is going forward with a slight and progressive right turn, slowly getting closer to the smoke cloud, i.e. smoke is progressively covering a larger area of the images. From event A ($t = 25s$, see Figs. 19 and 21) the visible amount of smoke becomes quite significant ($> 15\%$ of the image) and remains so until about $t = 54s$ (event D), after which it goes back to a low level. At $t = 43s$ (event C), the UGV experiences a sharper (and short) turn to the right. Between $t = 48s$ and $t = 60s$ the UGV is not facing the smoke cloud any more, only small light residuals of smoke are visible (e.g. see event D in Figs. 19 and 21). At $t = 63s$ (event E), immediately following a strong turn to the left, the UGV suddenly faces the smoke cloud again. When this happens, smoke covers a large majority of the image (see Fig. 21, E, and Fig. 18), before quickly dissipating.

From Fig. 17, we can already see that the presence of smoke in this experiment has strongly affected Visual-SLAM compared to IR-SLAM, which is not perturbed by smoke. In particular, when the visual camera is facing a very large amount of smoke (e.g. Event E at $t = 63s$ has more than 80% of the image covered), Visual-SLAM completely loses track of the trajectory (see Fig. 22), while the IR-SLAM solution

[1]see adjoint videos `Argo-Visual-smoke.avi` for Visual-SLAM and `Argo-IR-smoke.avi` for IR-SLAM at: http://www-personal.acfr.usyd.edu.au/tpeynot/IJICS/Videos/

[2]Note this "smoke detector" is very specific to this type of smoke and this dataset. It is only used for the purpose of illustration.

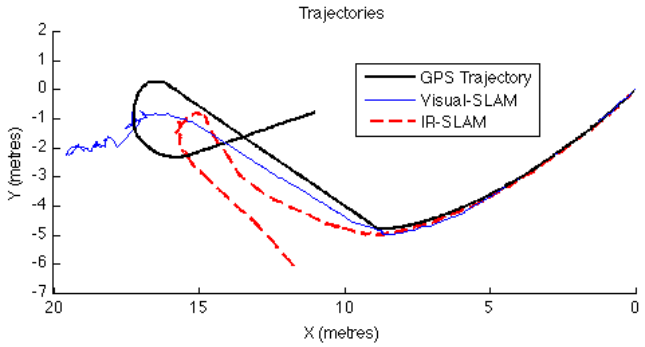


Fig. 17. Estimated trajectories using Visual-SLAM (blue), IR-SLAM (dashed red) and the measured $28.4m$ long dGPS/INS path (thick black), in the presence of smoke.

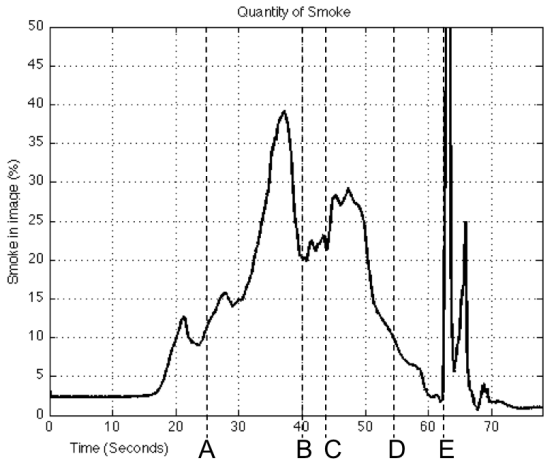


Fig. 18. Quantity of smoke in the images (% of the image area that is covered by smoke).

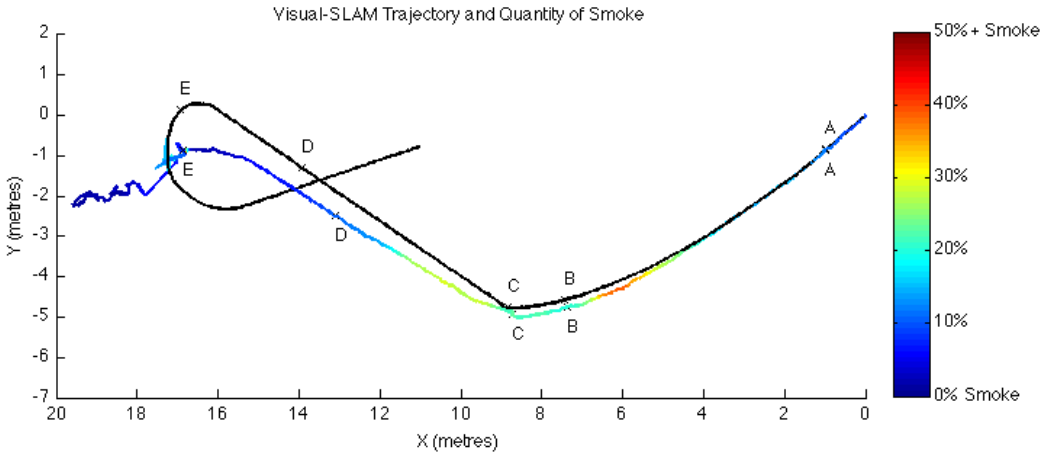


Fig. 19. Trajectory estimated by Visual-SLAM vs. GPS reference. The colour on the trajectory represents the amount of smoke visible in the (visual) images. Sample images for events A, B, C, D, E are shown in Fig. 21.

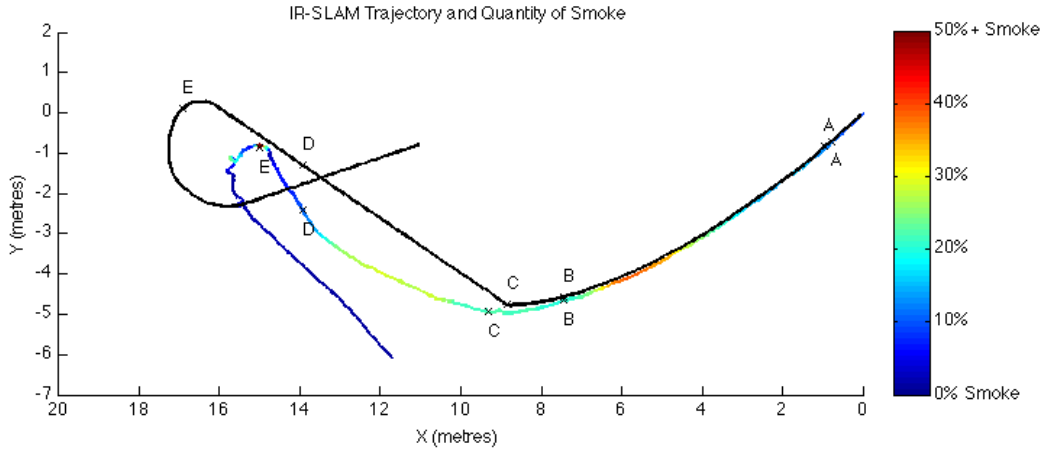
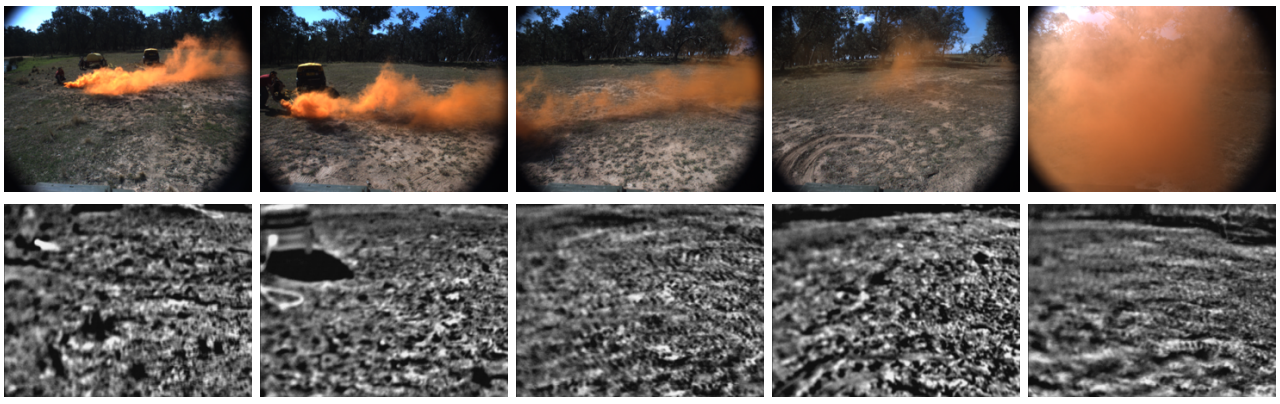


Fig. 20. Trajectory estimated by IR-SLAM vs. GPS reference. The colour on the trajectory represents the amount of smoke visible in the (visual) images. Sample images for events A, B, C, D, E are shown in Fig. 21.



(a) Event A (b) Event B (c) Event C (d) Event D (e) Event E

Fig. 21. Sample visual images (top) and IR images (bottom) for events A, B, C, D, E as shown on the trajectories in Figs. 19 and 20.

remains reasonable and consistent.

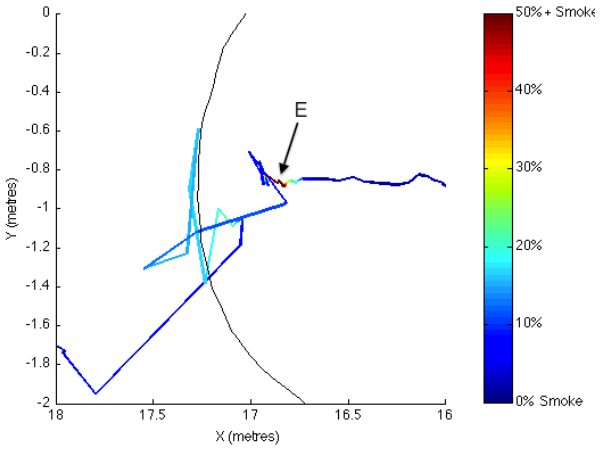


Fig. 22. Zoom on the Visual-SLAM trajectory shown in Fig. 19, around event E. Following the sudden appearance of large amounts of smoke (section in red on the trajectory), the Visual-SLAM solutions becomes erratic.

The difference in absolute position in three dimensions between the estimated trajectories and the dGPS/INS trajectory is shown in Fig. 28. The difference in absolute yaw between the estimated trajectories and the dGPS/INS trajectory is shown in Fig. 29. It can be observed that in this experiment, as long as the quantity of smoke remains limited, the Visual-SLAM solution retains superior yaw estimation compared to the IR-SLAM while there is arguably little difference in the overall position difference. However, when there is a large quantity of smoke (e.g. Event E at $t = 63s$), the yaw estimate and the position estimate of the Visual-SLAM are significantly affected and IR-SLAM clearly outperforms Visual-SLAM. Note that the significant increase in the error around event E is partly due to the strong turn that can be seen on Figs. 19 and 20 on the reference trajectory. Inevitably, this turn affects both SLAM solutions, but the presence of smoke makes the error on Visual-SLAM much worse, as will be discussed below. Recall that the horizontal FOV of the visual camera is much larger than the one of the IR camera, which means that *in clear conditions* Visual-SLAM would typically cope with strong turns better than IR-SLAM.

Another observation is that in this dataset the position uncertainties grow more rapidly for Visual-SLAM than IR-SLAM (see Fig. 23), in particular when the smoke cloud covers a significant part of the visual images (recall that the contrary has been observed in clear conditions). This is due to the fact that in the Visual images, almost no features can be matched within the smoke cloud, and the few matches that might occur are not enough to reduce the uncertainty in the 6-DOF of the camera pose.

Fig. 24 shows the number of SIFT matches that could be used in the SLAM algorithm. It can be seen that this number is comparable for Visual and IR when smoke quantity is low (typically below 10%), which happens between $t = 54s$ (event D) and $t = 62s$ (event E) and for $t > 68s$, for example. When more smoke is present, e.g. between $t = 25s$ (event A) and

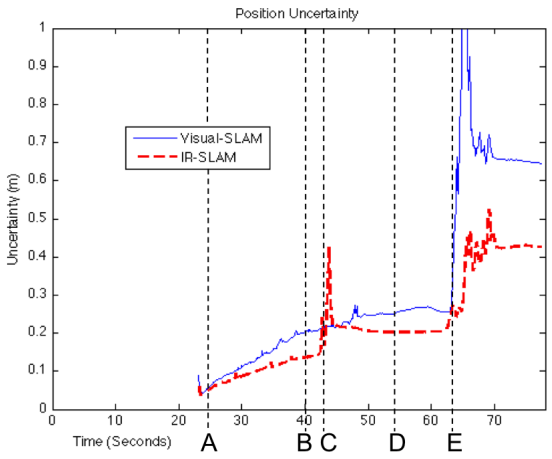


Fig. 23. Position uncertainties of Visual-SLAM (blue) and IR-SLAM (dashed red), in the presence of smoke.

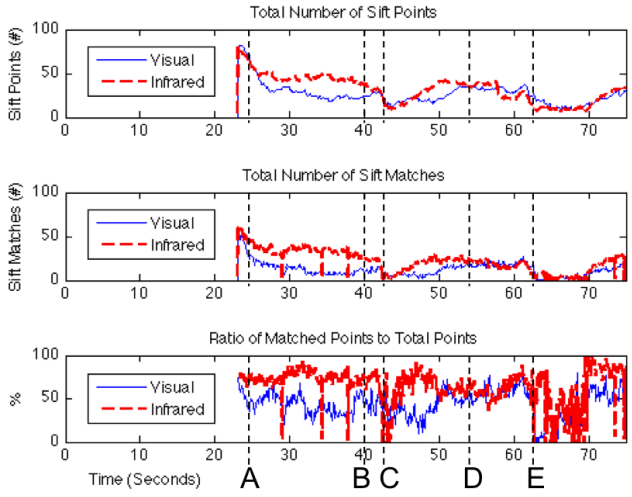


Fig. 24. Number of SIFT points (top) and number of SIFT matches (middle) for each Visual (blue) and IR image (red). The bottom line shows the ratio (%) of matched points over the total number of SIFT points found in the images.

$t = 54s$ (event D), it is clear that more SIFT matches are found in the IR data. Note that the drop in the number of SIFT matches for IR at about $t = 43s$ (event C) is due to the aforementioned sharp turn, which affects IR more than Visual due to the smaller FOV of the camera. The turn just before $t = 63s$ (event E) affects both SLAM solutions, but for a few seconds the visual data contains almost no matches at all due to the large presence of visible smoke following this event, which causes a very large error in the Visual-SLAM pose estimation. Once the smoke is gone, Visual-SLAM can finally re-initialise some landmarks and progressively recover an acceptable relative pose estimation. The global position cannot be corrected without a loop closure, and even in this case the local pose estimation around event E would still have large errors.

Along the trajectory, the average number of SIFT point matches is 13 for Visual and 25 for IR images, and the average

ratio of matches over total number of SIFT points is 46% for Visual and 67% for IR images.

Fig. 25 shows the evolution of SE for Visual and IR during this test. It can be seen that SE in Visual images decreases significantly when the smoke cloud covers a large section of the image, while there is no significant change in SE for IR. In particular, a major drop of quality of Visual data is visible at $t = 63s$ (event E).

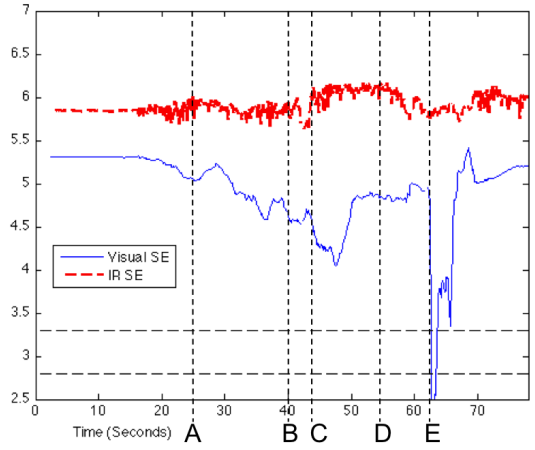


Fig. 25. SE for Visual (blue) and IR (red).

The decision on the quality of the two types of data, based on SE, was made using the technique proposed in [20], to choose what source of data should be used to compute the pose estimation with the monocular SLAM algorithm. Fig. 26 illustrates that decision. The alarms indicate when it is believed that the quality of the data is inappropriate. In practice, every time an alarm is triggered for visual data, this means IR data will be preferred, i.e. IR-SLAM will be chosen as the local pose estimator rather than Visual-SLAM. Using these

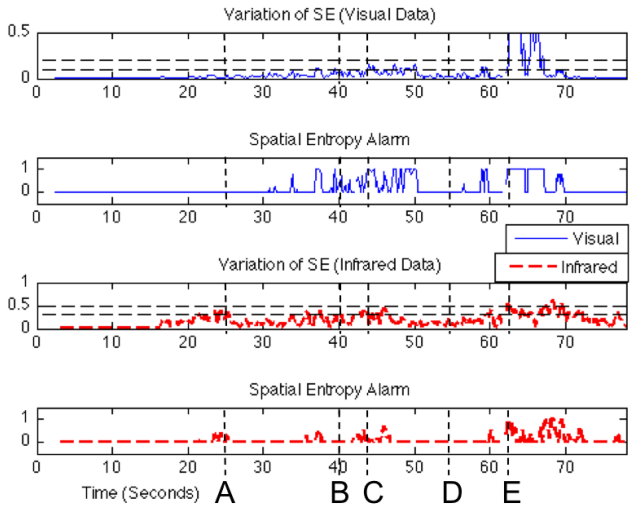


Fig. 26. Alarms for Visual (top, blue) and IR (bottom, red).

alarms, based on the analysis of the variation of SE after motion compensation, situations where the quality of visual data will cause a degradation of the performance of Visual-SLAM, while the performance of IR-SLAM is maintained, can be anticipated. This is achieved *before* the SLAM computation is made and any evaluation of the quality of the solution can be done (e.g. checking the uncertainties).

5.6. *Switching Modalities based on Quality Metrics*

In this section, the decision on the quality of Visual and IR data is used to switch between sensing modality to obtain a better and more reliable pose estimation with the monocular SLAM algorithm. Initially, Visual-SLAM is preferred, as its estimation is more accurate in (mostly) clear conditions. SE is used to measure the quality of visual and IR data. When a drop in the quality of visual data is detected at time t_d, the preferred pose estimation becomes the one from IR-SLAM. In practice, an initial pose estimate is set as the last estimate from Visual-SLAM (at t_d) and for $t > t_d$ the relative motion from this position is computed using the IR-SLAM solution, until the quality of visual data becomes higher again, or the UGV mission is stopped.

Fig. 27 shows the trajectory obtained using that process off-line. The estimations were combined after the full computation of both separate trajectories, using the two different sources of information.

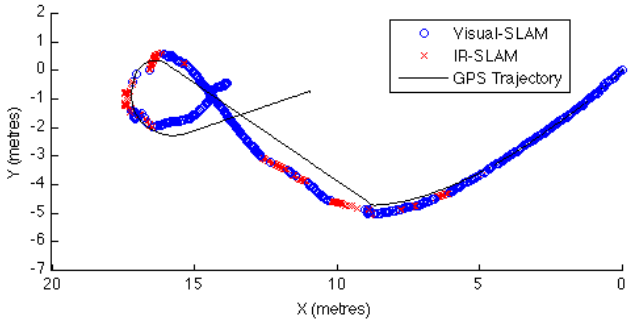


Fig. 27. Combined trajectory using the pose estimates from Visual-SLAM (blue points) *or* from IR-SLAM (red points). The dGPS/INS trajectory is shown in black, for reference.

The difference in absolute position between the estimated trajectories and the dGPS/INS trajectory is shown in Fig. 28. The difference in absolute yaw between the estimated trajectories and the dGPS/INS trajectory is shown in Fig. 29. It can be seen that in most cases the solution obtained is of far better quality than what was obtained with Visual-SLAM or even IR-SLAM only. The overall error in position and yaw estimation at the end of the dataset is improved substantially from either of the Visual-SLAM or IR-SLAM alone. Between $t = 45s$ and $t = 65s$, the Visual-SLAM estimate of the yaw is better than the switching solution as the IR-SLAM solution was chosen to be used around the first sharp corner (Event C) and, despite high levels of smoke in the visual data, IR-SLAM does not estimate the turn quite as well as Visual-SLAM. However, despite this momentary offset, the overall yaw estimate is greatly improved when significant challenging conditions are encountered, such as

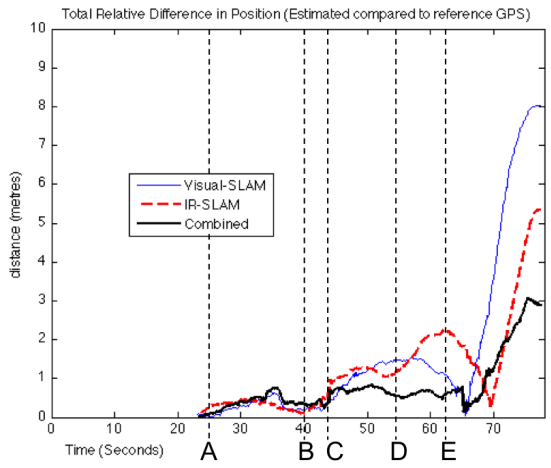


Fig. 28. Total difference in metres (in 3 dimensions) between the estimated combined trajectory calculated by switching between Visual-SLAM (blue), IR-SLAM (dashed red) and Combined (thick black), and the measured trajectory from the dGPS/INS in smoke conditions over time.

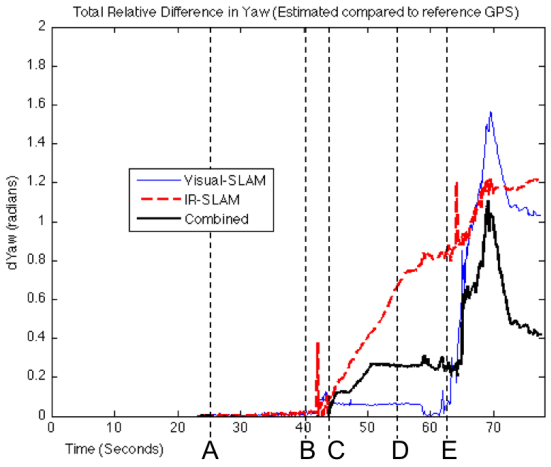


Fig. 29. Total difference between the estimated yaw calculated by switching between Visual-SLAM (blue), IR-SLAM (red) and Combined (black), and the measured yaw from the dGPS/INS in smoke conditions over time.

Event E at $t = 63s$. The switching meant that the perceptual failure, occurring when using a visual camera, was mitigated. Additionally, the position estimate (Fig. 28) of the combined solution is significantly better than either IR-SLAM or Visual-SLAM throughout the trajectory estimation.

Using a process as described for switching between modalities, an overall better estimate can be obtained both in clear (where visual is preferred) and challenging conditions (where IR substitutes visual when necessary), which means that an application robust to conditions such as presence of smoke can be obtained, as Figs. 27, 28 and 29 have shown.

In summary, the proposed method provides a perception application robust to challenging environmental conditions by mitigating perceptual failures, although in some cases this can come to the cost of some temporary degradations of the performance of the application (accuracy in the case of SLAM).

5.7. Discussion

The main reason for the degradation of the accuracy of Visual-SLAM in the presence of smoke is the limited number of remaining feature matches, which causes limited observability of the 6-DOF of the camera pose. Note that inherently there is no integration of corrupted observations to the filter, as the feature management process of the monocular SLAM (see Section 5.1) effectively discards features that do not match according to the prediction. Detecting the limited observability would still require to compute all the features and the matching, which is computationally expensive, even if features faster to compute than SIFT, such as SURF [21], were to be used. Then, following the result of this test, the process would have to switch to the other source, the IR camera, and perform the feature extraction and matching process again.

By using a quality metric (which is much simpler and faster to compute and evaluate), our system can *anticipate* these situations and request the modality switching before any feature computation is made. This results in a significant reduction of the computation time, as we only need to do these costful operations for one type of image at any time. Thanks to this process, the perception application (monocular SLAM) becomes robust to challenging conditions such as the presence of smoke, while keeping computational performances comparable to the original Visual-SLAM.

Note that this example of automatic switching between sensing modalities is based on the current evaluation of the quality of data only. It does not take into account requirements of the application such as, in this case of SLAM, the necessity to keep tracking features between several images from the same source (*i.e.* type of camera). Accounting for such requirements will be needed before any operation of switching between different sensing sources can be applied on-line.

6. Conclusion

6.1. Summary

This paper has presented image quality metrics and how they relate to evaluating the quality of image data for outdoor robotic perception systems. The existing metrics analysed in this context included in particular Brightness, Contrast, Blur, Sharpness and Spatial Information. While all metrics are affected differently by challenging conditions, the metrics Local Contrast, Blur and SI were found to be the most promising in our context. However, more work would be required with the Blur metric to tune it for specific sensors. Also, Local Contrast in the current form cannot realistically be performed on a real-time system.

Spatial Entropy (SE) was introduced as a novel metric evaluating structure in an image that is more robust in detecting challenging conditions than Spatial Information (SI). Additionally, SI and SE were shown to provide even better discrimination of situations when interpreted together.

We have discussed how the metrics may be interpreted specifically to discriminate challenging conditions for perception including thresholding individual values, monitoring the evolution of the metrics and comparing the relationship of metrics.

An example of exploitation of such metrics in a realistic robotics application has been proposed, where the application is monocular SLAM. This example showed that the quality metric SE could be used to anticipate situations where SLAM using IR images is likely to perform better than SLAM using visual images. By switching between sensing modalities accordingly, it is possible to obtain an application robust to challenging conditions such as the presence of smoke. Note that the proposed technique of modality switching based on data quality evaluation is not SLAM specific and could be applied to other localisation techniques relying on similar types of features, such as visual odometry.

6.2. Future Work

Future work with the metrics presented in this paper will include analysis of how they react in a broader range of conditions such as at night, on the road or in urban areas. While the metrics discussed in Section 3 were considered to be the most appropriate for robotic perception, there are many more metrics and even different methods of measuring the same aspect in an image that are available in the literature. For example, although they are not mentioned in this paper because the SLAM application extracted only gray-scale features, colour quality metrics (e.g. see [22, 23, 7, 24]) could be used to augment the visual image quality evaluation for applications that use colour information.

The overall quality of an image could be better expressed by considering combinations of metrics that have some relationship. For example, Brightness, Contrast and Shannon Information all measure different aspects of the intensity in an image. There are relationships between these metrics that can better discriminate what is occurring overall in the image (such as demonstrated in the SI-SE relationship in Table I).

Additional work will include evaluation of how metrics may be utilised more effectively to better reflect what is occurring in the image. For example, most of these metrics are designed to provide one global value representing the quality of the image. However, they could be adapted to evaluate specific sub-images or regions of interest in the original image. This could be particularly relevant when challenging conditions are present but not covering the whole image.

Besides checking the quality of the data provided by a single sensor, one of the future objectives of this work is to be able to check the consistency of data between heterogeneous sensors such as a visual camera and an infrared camera. Exploiting methods to make direct comparisons of data provided by multimodal sources will help to discriminate the most appropriate sensor data to use. SI and SE were clear candidates for such multimodal comparisons but other methods and metrics will be investigated.

In the current monocular SLAM application, the maps from the different sources (e.g. visual and IR cameras) are kept separately, i.e. the visual and the IR landmarks are completely uncorrelated. Future work will consider combining the landmarks from the different cameras in the same map. This should make the online application straightforward and will prevent from initialising landmarks every time the sensor modality switches on and off, making the SLAM more robust than an online switching algorithm with a non-shared map.

ACKNOWLEDGMENTS

The authors would like to thank J. Solà for software collaboration.

REFERENCES

[1] J. P. Underwood, A. Hill, T. Peynot, and S. J. Scheding, "Error modeling and calibration of exteroceptive sensors for accurate mapping applications," *Journal of Field Robotics, Special Issue: Three-Dimensional Mapping, Part 3*, vol. 27, no. 1, pp. 2–20, 2010.

[2] C. Brunner, T. Peynot, and J. Underwood, "Towards discrimination of challenging conditions for ugvs with visual and infrared sensors," in *ARAA Australasian Conf. on Robotics and Automation*, 2009.

[3] Z. Wang and A. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[4] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. John Wiley & Sons Ltd, 2005.

[5] T. Vlachos, "Detection of blocking artifacts in compressed video," *Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, 2000.

[6] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *IEEE Int. Conf. on Image Processing*, 2000.

[7] S. Susstrunk and S. Winkler, "Color image quality on the internet," in *IS&T/SPIE Electronic Imaging: Internet Imaging V*, 2004.

[8] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to jpeg2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.

[9] T. Peynot, S. Scheding, and S. Terho, "The Marulan Data Sets: Multi-Sensor Perception in Natural Environment with Challenging Conditions," *Int. Journal of Robotics Research*, vol. 29, no. 13, pp. 1602–1607, 2010.

[10] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A*, vol. 7, pp. 2032–2040, 1990.

[11] D. Mackay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2007.

[12] Y. Choong, F. Rakebrandt, R. North, and J. Morgan, "Acutance, an objective measure of retinal nerve fibre image clarity," *British Journal of Ophthalmology*, vol. 87, no. 3, pp. 322–326, 2003.

[13] ITU-T, *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union Recommendation P.910, 1999.

[14] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Int. Conf. on Computer Vision*, vol. 2, 2003.

[15] R. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *Int. Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1987.

[16] J. M. M. Montiel, J. Civera, and A. J. Davison, "Unified inverse depth parametrization for monocular SLAM," in *Robotics: Science and Systems*, 2006.

[17] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. on Robotics*, vol. 24, no. 5, pp. 932–945, 2008.

[18] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[19] A. J. Davison, "Active search for real-time vision," *Int. Conf. on Computer Vision*, 2005.

[20] C. Brunner and T. Peynot, "Perception quality evaluation with visual and infrared cameras in challenging environmental conditions," in *Int. Symposium on Experimental Robotics*, 2010.

[21] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[22] C. C. Koh, J. M. Foley, and S. K. Mitra, "Color preference and perceived color naturalness of digital videos," in *SPIE Proceedings, Human Vision and Electronic Imaging XI*, 2006.

[23] D. Hasler and S. Sasstrunk, "Measuring colourfulness in natural images," in *Human vision and electronic imaging. Conf. No. 8*, 2003.

[24] R. Hunt, *Measuring Colour*, 3rd ed. Fountain Pr Ltd, January 2001.

Christopher Brunner is a postgraduate student at the Australian Centre for Field Robotics, The University of Sydney, Australia. He received a B.E. degree in mechatronics (space) engineering and B.Sc. degree in advanced science (physics) from the University of Sydney in 2006. After a brief two year stint working in a biomedical engineering company, he returned to university in to start Ph.D. studies in 2009. In 2010 he visited LAAS-CNRS in Toulouse, France. His current research interests include high integrity and robust perception systems for outdoor autonomous ground vehicles.

Thierry Peynot received M.E. and Ph.D. degrees from the University of Toulouse, France, in 2002 and 2006, respectively. He prepared both thesis at LAAS-CNRS, in Toulouse. In 2005 he visited NASA Ames Research Center, Moffett Field, California. From 2005 to 2007 he was Associate Lecturer at the University of Toulouse, pursuing his research at LAAS-CNRS. Since the end of 2007 he is a Research Fellow at the Australian Centre for Field Robotics (ACFR), The University of Sydney, Australia. His current research interests include unmanned ground vehicles, persistent autonomy, multi-sensor perception, perception integrity, diagnosis for robotics and automatic reconfiguration of robotic systems.

Teresa Vidal-Calleja received the M.E. degree from the Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico, the M.S.E.E. degree from Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV-IPN), Mexico City, and the Ph.D. degree from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2007. During her Ph.D. studies, she was Visiting Scholar with the Active Vision Lab, University of Oxford, Oxford, U.K., and the Australian Centre for Field Robotics (ACFR), The University of Sydney, Sydney, NSW, Australia. In 2008, she was a Postdoctoral Fellow with LAAS-CNRS, Toulouse, France. She was on leave from the Institut de Robòtica i Informatica Industrial, CSIC-UPC, Barcelona to ACFR in 2009. She currently is research fellow at ACFR. Her current research interests include perception, visual SLAM, place recognition, cooperative aerial and ground vehicles, and autonomous navigation.