

A Low-level Fusion of Millimetre-Wave Radar and Nightvision Imaging for Enhanced Characterisation of a Cluttered Environment

Richard Grover, Graham Brooker and Hugh F. Durrant-Whyte

Australian Centre for Field Robotics

The Rose Street Building J04

The University of Sydney, 2006

{r.grover,gbrooker,hugh}@acfr.usyd.edu.au

Abstract

This paper presents a method for the fusion of mm-Wave Radar and Nightvision data. This method differs from traditional approaches in that the information is combined before any judgement is made regarding the number, type or location of any objects within the sensor field. The characteristic information is combined and is shown to yield an environmental representation containing both sensor signatures for each object, significantly simplifying and improving the performance of detection, classification and tracking algorithms. Results of field trials are used to demonstrate the method's effectiveness.

1 Introduction

The characterisation of unstructured environments for the purposes of navigation, threat assessment and object identification, for example, is important in many different robotics applications. These problems are all reducible, at some level, to the extraction of features from sensor data. In this paper, a feature is considered to be an object which is detectable in the operating environment. This includes objects generally considered targets, such as aircraft, ground vehicles, missiles and retro-reflectors, but also encompasses objects generally considered "clutter": trees, land surfaces and foliage, for example. For the purposes of this investigation, the terms target, feature and object will be treated as interchangeable.

Furthermore, data fusion will be considered the act of combining the data from multiple or different sensors to provide a more appropriate representation of the objects within the system environment. In this regard, an abstract state representation is used, similar to that proposed by [Majumder *et al.*, 2001], where the data from each sensor is projected into a common "sensor space" and combined. This combination of information prior to any form of identification or track-

ing process differs greatly from the typical approach to the fusion of data of vastly different dimensionality. [Romine and Kamen, 1996] and [van Huyssteen and Farooq, 1999] provide methods for the combination of visual and radar data in which the data from each sensor is processed to the extent of object recognition and tracking. These processed estimates are then combined to give a 'fused' estimate.

It is well known, however, that feature extraction processes utilising a single sensor are inherently sensitive to disturbances, as demonstrated in [Romine and Kamen, 1996]. The effect of such disturbances can be ameliorated, however, by considering a combined representation, as it is unlikely that the disturbance will affect different sensing modalities in the same way. Indeed [Majumder *et al.*, 2001] notes that these approaches often provide significantly higher reliability.

This paper will first consider two commonly used sensors in the robotics and environmental sensing arenas - mm-Wave radar and light-intensified vision (more commonly known as nightvision). The nature of the data generated and the particular characteristic strengths and weaknesses of each mode will be discussed in the context of real data obtained during field trials. Consideration will then be given to the sensory map representation of Majumder *et al.* and its formulation for these sensors. Finally, the results of the processing for a real environment will be presented.

2 Sensing Characteristics

2.1 mm-Wave Radar

The radar used in this project is a SaabTech automobile collision avoidance radar using a frequency-modulated continuous wave (FMCW) system operating at 76GHz using a frequency sweep with gradient 500kHz/ μ s and length 384 μ s. Data is sampled using a 1.33MHz 12-bit analog to digital converter. This unit has a beamwidth of approximately 1.8° and is mechanically scanned in both azimuth and elevation. During the linear part of the frequency sweep, the return signal

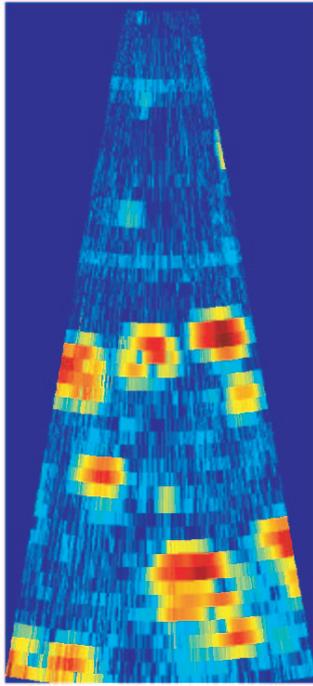


Figure 1: A typical 2D radar scan showing identified targets.

is mixed with the transmitted signal - generating beat frequencies proportional to the propagation delay and hence the distance to the reflecting object. A 256-point fast-fourier transform (FFT) process, therefore, gives the reflected amplitudes over 129 range bins, each with a size of approximately 1.46m. Data is provided at a rate of 235Hz. The unit operates continuously as the system is mechanically scanned with a user selectable rate and field-of-view. Figures 1 and 2 show a single scan at constant elevation and a 3D volume visualisation of the data from a 25° by 12.5° scan at a rate of $10^\circ/s$. Figure 3 shows a photograph of the same region showing the sensor location and features of interest.

These figures clearly show that the radar provides a complete three-dimensional representation of the en-

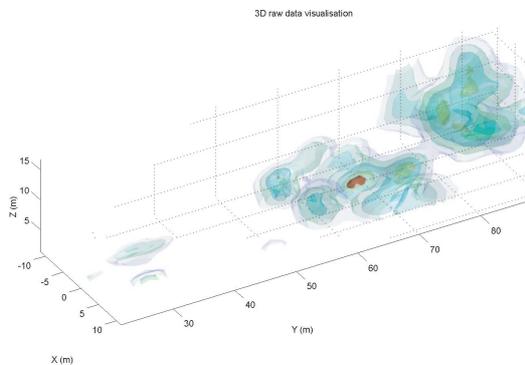


Figure 2: A volume visualisation of the complete data set created using multiple isosurfaces.

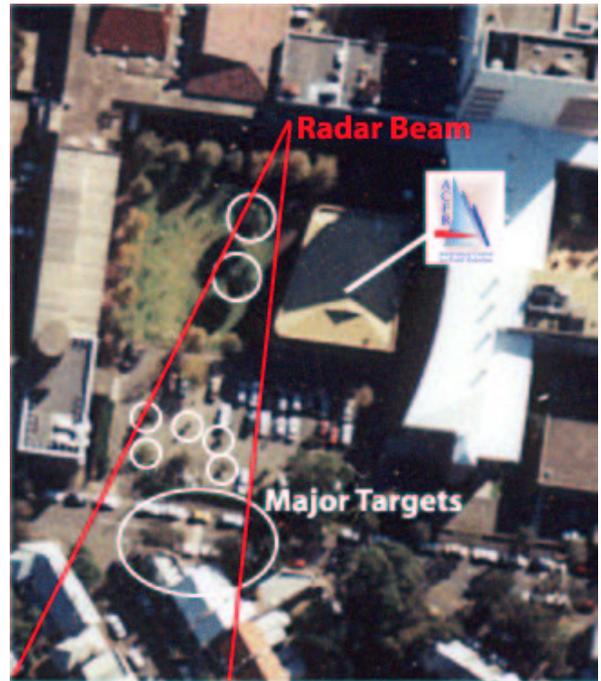


Figure 3: A photograph of the scanned region showing sensor location and identified targets.

vironment, but that the edges of physical features are ‘blurred’ as a result of the large beamwidth and range-bin size; though some sub-beamwidth data is available as the beam is not uniformly illuminated. The data do, however, suffer from multi-path reflections, though this effect is well-known and processing techniques exist to alleviate its effects. Additionally, because of the scanning arrangement of the system, the data is naturally in polar form and can be processed in a fusion system without requiring the back-projection manipulation required to generate the data shown in figures 1 and 2.

It is worth noting, however, that this complicates the representation in the more general case where the sensing platform can move. Under these conditions the



Figure 4: A typical nightvision image taken from the same location as the radar images shown earlier.

transformations required to register the data obtained from fusion at multiple sensing locations involve both translation and rotation phases. Representation of the fused data in a cartesian coordinate system would appear more appropriate as only a translation is required. While this is trivial for the radar, as there is sufficient information to determine the cartesian location directly, the same is certainly not true for the visual data. In the case where only a single image is available, it becomes impossible to determine the cartesian location without *a priori* knowledge of the range to the target. This impacts directly on the uninhibited extraction of the sensor's natural information independently of the other sensors. Thus, while the utilisation of polar coordinates is complicated by rotation and translation operations of a moving sensor platform, it allows the system to remain decoupled prior to the sensor fusion stage. Furthermore, once the data have been fused, they can be readily transformed into a cartesian space to enable further processing and manipulation to be performed easily.

2.2 Nightvision Imaging

The visual data is obtained using an ITT Technologies third-generation nightvision camera, though the techniques utilised in this paper are directly applicable to other 'visual' modes, such as CCD camera, thermal imagers and forward looking infrared (FLIR) devices. The camera provides black and white images of the scene under a variety of different lighting conditions. The camera provides NTSC images at up to 30 frames per second and has a field of view of approximately 13.9° by 8.8° .

The data from the camera differs in three major ways from the radar data. Firstly, while visual cues, such as object occlusion and perspective effects, can provide some relative depth information, no range data is available. Secondly, the camera provides only the intensity of the *first* object along the ray corresponding to any particular pixel. Finally, the visual system provides a significantly improved resolution in azimuth and elevation (commonly known as cross-range resolutions). A typical nightvision image of the same scene is shown in figure 4. Finally, the image data can also be conveniently expressed in polar form without any transformation being required. This means that the data from *both* sensors is used in its 'natural' form and a strong compatibility exists between the data, irrespective of the vastly different dimensionality and other characteristics of the data provided.

3 Data Fusion Processing

3.1 Preprocessing and Information Retention

If this system is to be utilised in practice, however, the data manipulation and association problem must remain tractable. Importantly, note that the sensors

here produce very large quantities of data. Scanning at approximately $25^\circ/\text{s}$ over a 25° pan range and with 25 scan lines gives $235 \times 25 \times 129$ or approximately 750,000 floating-point intensity values for a single data set. Likewise, the visual system provides images at up to 640×480 or 300,000 pixels. Clearly, the association and fusion problem with at least one million values is well beyond the capabilities of any practical system, let alone a system operating under real-time constraints.

This necessitates that some form of pre-processing is required to transform the raw data into a compact and appropriate form. In doing so, however, it is imperative that the maximum quantity of information be retained - in particular that that data which is characteristic to each sensor should be retained. This implies that the data should be processed to the limited extent of extracting detectable entities and their associated properties. This effectively allows the system to generate a table of objects and object properties, significantly reducing the data representation problem. The question remains, however: can the quantity of data be reduced without significantly attenuating the information content of the system?

3.2 Visual Data Pre-processing

As is easily seen from the image shown in figure 4, the nightvision scene is naturally decomposable into unstructured visible entities, commonly known in image processing as 'blobs'. Each of these is identifiable to the extent of grouping pixels of similar intensities together. Importantly, the entities within the image may equally exist with *any* intensity level, a situation differing from many image segmentation problems in which distinct objects are detected against a high-contrast background. For the purposes of this initial investigation, a simple multi-level segmentation was selected.

The image processing method takes advantage of several notable features of the intensified video images. Initially, it is noted that the images are subject to a widely varying level of noise, dependant on the total light intensity of the scene. As this effect is predominately caused by avalanche noise in the photo-multiplier tubes, the relative significance will grow as the intensification increases. Additionally, the camera has an automatic gain control (AGC) system to adjust this level, so that a comparatively well-illuminated scene appears with little 'snow', while a poorly illuminated scene shows significant degradation.

To alleviate the effects of this noise process on the performance of the system several frames are averaged. Figure 5 shows the time varying values of several pixels in a sample image and clearly demonstrates that the noise does, indeed, have zero mean, justifying this approach. Additionally, noting that very few features of interest have a size of the order of several pixels, a low-pass Gaussian filter is applied to smooth the image. The histogram is then expanded to fill the entire

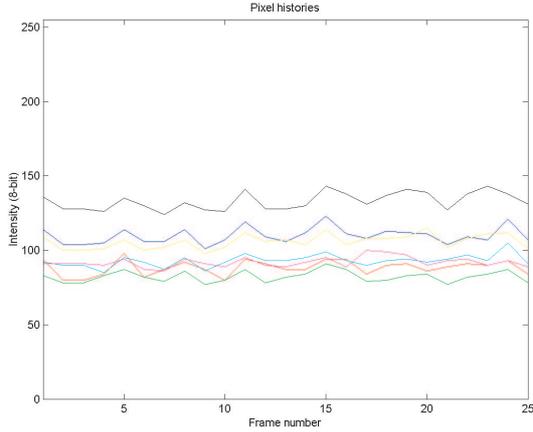


Figure 5: Pixel histories for several points in a nightvision image.

range by using a direct mapping as

$$I_{new} = \frac{I_{old} - I_{min}}{I_{max} - I_{min}}. \quad (1)$$

The intensification of the entire image according to the brightest sources ensures that there is no ‘characteristic’ intensity distribution to be anticipated in this data. Specifically, an image taken without any light sources directly visible will tend to generate a closely packed histogram, while any image with a source visible will tend to be widely separated. This implies that assuming a generalised distribution for the selection of threshold levels is inherently fragile. This problem is overcome through the utilisation of a method of Gaussian mixtures.

In this approach, the data is represented as a superposition of gaussian intensity distributions, each of which is conjectured to correspond to a particular object or group of objects. In this manner, an arbitrary distribution can be compactly represented and analysed. An important point to note, however, is that the Gaussian basis functions cannot be orthogonal, implying that the decomposition is not unique. A simple implementation using a least-squares approximation allows the method to proceed in an efficient and effective manner.

Once the intensity distribution has been constructed, each mean value can be interpreted as corresponding to a particular set of objects with a similar mean intensity. As an initial approach, thresholds are then calculated as the midpoints between these distributions. In a more complete approach, the variance of each distribution could also be taken into account such that the threshold can be selected optimally according to some appropriate measure. The image is then divided into black and white ‘logical maps’ of the regions above and below the threshold levels. These maps are closed and smoothed by the well-known dilation-erosion procedure. The smoothed

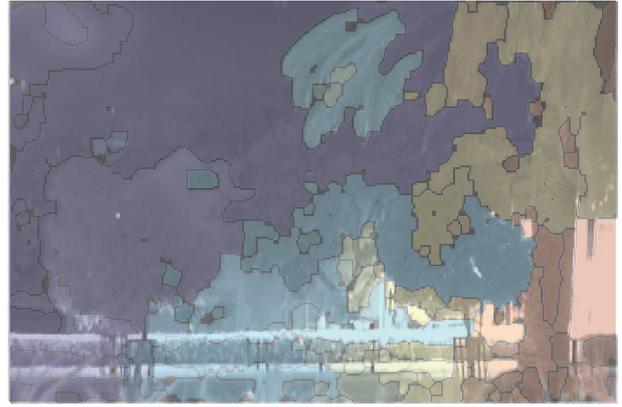


Figure 6: Coloured regions overlaid on original nightvision image.

maps are then dilated further and logically combined. This results in a coarse edge-map of the regions identified earlier. Utilising methods in the image processing toolbox of the MATLAB environment, these edges are reduced to lines and regions are identified. Figure 6 shows the regions identified using this procedure overlaid on the original image. Two notable features are clear in this image, firstly, the regions show a strong correlation with the visual entities within the image; and secondly, the process identifies many very small regions.

Small regions, corresponding to either residual noise in the data, or objects at a great distance can be discarded without a significant loss of information, while regions with similar intensities can be readily combined. Finally, having selected the visual entities from the image data, the properties of each can be recorded in the sensor map for later use. Here the most important properties of the blobs are: the centroid location in the sensor space, in this case the polar co-ordinates; the angular size of the target; the mean pixel intensity across the object; the variance in intensity values - which directly represents the texture of the surface as smooth surfaces will tend to reflect light evenly, while rough surfaces will not; the blob ‘perimeter’; and the angular ‘area’ subtended by the object. Here the perimeter and area are measured in polar co-ordinates as this is a more natural representation of the data from the visual system, and so do *not* represent physical sizes. Of these, the feature most characteristic of the visual sensing processes, as contrasted by the radar, relates to object texture. As this is of great significance in target identification and recognition systems, it is important that differences between objects of differing texture can be utilised effectively.

3.3 Radar Data Pre-processing

The radar data shown in figure 2 can likewise be divided into three-dimensional entities, again unstructured, which represent at least the front surface of ob-

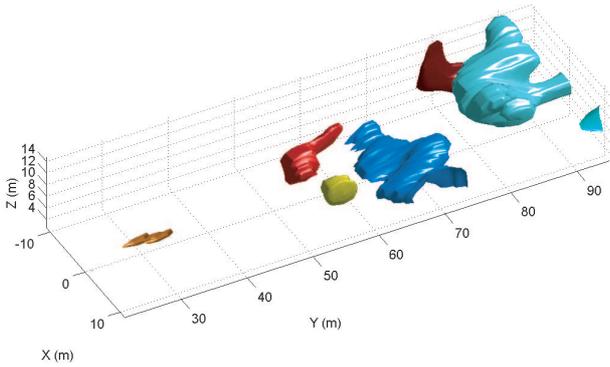


Figure 7: Radar blobs extracted from earlier data and labelled by colour

jects. An example of these are shown in figure 7. Importantly, however, the radar is capable of penetrating many objects of interest, notably trees, and so several blobs can share the same location in orientation space. Like vision, regions of low intensity *can* represent objects of interest - reflective surfaces and radar absorbing targets, amongst others - however, it is significantly more difficult to identify their location in polar space. Indeed, the only method by which these can be identified positively is through the identification of high-intensity objects which encircle the object in orientation space, that is with azimuth and elevation angles surrounding the region. Other effects such as multi-path reflections may further complicate their reliable detection. For the sake of simplicity in this investigation it is assumed that the radar objects of interest are only those of high-intensity.

The radar data should, therefore, be processed to firstly identify the three-dimensional entities with intensities greater than a given threshold. Since the radar returns occur over vast intensity ranges, the detail apparent in the images shown earlier is only visible when consideration is given to logarithmic intensity scaling. The threshold is set as the decimal logarithm of the mean of the data. A single isosurface¹ can be generated from this data and this is subsequently used to identify the individual objects within the scan.

The polar centroid, that is its location in both range and orientation, cannot be considered the true centroid of the target itself, as we can only guarantee that the front surface of the target is illuminated by the beam. We can, however, assign the angular centroid (position in orientation space) to this object. Indeed, this gives a direct association to the apparent angular location of a visual entity. In addition to the angular position, however, we can additionally compute the average range to the object; the range variance; the angular target size; the mean radar intensity; the radar intensity variance;

¹An isosurface is the surface of constant value in a volumetric data set.

and the projected angular area and perimeter. Thus, we find that the radar blobs contain similar information forms as the visual blobs, with the addition of range and range variation data.

3.4 Data Association and Fusion

The data fusion process, that is, the actual combination of the processed data, can then proceed in one of two manners: firstly, an association process can be used to link visual and radar blobs, creating a combined object containing the information from both sensors. Secondly, it is also possible to represent the blobs themselves not as well-defined objects with boundaries, but as statistical distributions describing the probability of object presence. For the radar data this is simply the distribution describing the radar intensity, reduced and represented in an appropriate manner. For the visual information, however, the distributions are not simply the intensity distributions, as dark areas have equal likelihood of being objects. We might therefore assign the distribution to be the absolute intensity variation from the mean image intensity.

Utilising this approach the data are not represented as distinct objects, though the segmentation into blobs assists with the later characterisation of the combined representation and remain an important part of the system operation. The distributions are, in general, arbitrary, and the compact representation of the distribution is of great importance in the application of this method. A ‘sums-of-Gaussians’ (SoG) approach, where the data is represented as the summation of a series of gaussian distributions, can be shown to be appropriate for the consideration of such data. Once the data have been represented, several methods exist for the statistically appropriate combination of the distributions.

In this initial study, however, the first approach is used to demonstrate the feasibility of the fusion of these disparate modalities. Figures 6 and 7 show the successful segmentation of the data into individual entities, each of which contains the most important information available from each sensor. We note several very important points here before considering the fused representation. Firstly, the segmentation of the radar gives results which appear in good agreement with the physical reality. Secondly, while the visual image appears well-segmented, there are several important problems with the segmentation. Most notably, the tree in the left hand half of the frame merges with the wall of the building in the distance, leading to an incorrectly merged object. Note however, that the intensities of these two objects are almost identical and that consideration of other visual cues such as texture allows the viewer to distinguish the individual items despite this intensity match. Finally and most importantly, there exist objects which are visible to one sensor but which are not detectable using the second. This implies that there will exist objects which must remain



Figure 8: Three-dimensional visualisation of fused scene

within the environmental model but which cannot be used in the fusion.

As an initial visualisation of a representation combining both the radar and visual information, the nightvision image is mapped onto the surface describing the radar blobs. Once the radar and nightvision information has been registered to allow direct association in spherical coordinates, each pixel in the image is examined in turn. The projection which this pixel makes through the radar data identifies which, if any, of the blobs the viewing ray impinges. Note, however, that the radar objects extend beyond the physical limits of the items of interest. This effect is well known, and accounts for the observance of blurring in radar imagery. To overcome this, the radar blobs have a layer of thickness equal to one quarter of the beamwidth removed from their extremities prior to performing the projection. This improves the performance of the visualisation and enables a more realistic image to be created. Figure 8 shows the scene described earlier in this representation. It is clear that the intensity data contained in the visual representation is not used when performing the association and depth augmentation, this is left to the user's own faculties.

3.5 Further Work

While this method provides an interpretable image, significant scope exists for extending the processes described in this paper; much of which relates to the association, representation and visualisation stages. In particular, there are two major approaches which warrant further attention: utilising the visual intensity data to improve the texture-mapping process and implementing a probabilistic association and fusion strategy. The first of these will enable the system to quickly characterise those parts of the visual image which do not appear in the radar representation; say, for example they could be displayed with a different colouring scheme or transparency to distinguish them from the matched surfaces. This allows the visualisation to show both those items for which a correlation exists, but addition-

ally, those for which no depth information is available.

The second area relates to extending the approach described here to automatic identification, tracking and post-processing situations. By applying a statistical fusion approach, such as Bayesian Filtering, it is possible to combine the information from the two representations into a single multi-sensor form, as described earlier in this paper.

4 Conclusions

This paper has outlined the development of an approach to the fusion of information from disparate sensing modes in a significant and appropriate manner. The characteristics of a combined representation have been shown to provide an improved information source for applications such as target identification and tracking. Practical methods for the extraction of important features and their properties from the visual and radar fields have been described and the results of practical implementation have shown a high degree of success. Finally the information from the two sensors has been fused in a visualisation to provide a verification of the conceptual fusion methods proposed.

References

- [Majumder *et al.*, 2001] Somajyoti Majumder, Steve Scheding, and Hugh Durrant-Whyte. Multisensor data fusion for underwater navigation. *Robotics and Autonomous Systems*, 35:97–108, 2001.
- [Romine and Kamen, 1996] Jay Brent Romine and Edward W. Kamen. Modelling and fusion of radar and imaging sensor data for target tracking. *Opt. Eng.*, 35(3):659–673, 1996.
- [van Huyssteen and Farooq, 1999] David van Huyssteen and Mohamad Farooq. A partially decentralised architecture for fusing active (radar) and passive (infrared) measurement data. *SPIE*, 3719:196–208, 1999.