Proceedings of the 1996 IEEE
International Conference on Robotics and Automation
Minneapolis, Minnesota - April 1996

3D Robot Sensing from Sonar and Vision

Huzefa AKBARALLY and Lindsay KLEEMAN

Intelligent Robotics Research Centre
Department of Electrical & Computer Systems Engineering
Monash University, Australia

ABSTRACT

We describe a sensor that fuses sonar and visual data to create a three dimensional (3D) model of the environment with application to robot navigation. The environment is characterized by a set of connected horizontal and vertical lines. 3D sonar data is augmented by making deductions concerning the connection and definition of lines in 2D visual data. Any errors that may result from incorrect interpretation of the 2D camera data, such as false connections between lines, can be detected by moving the robot. Experimental results from the sensor are presented.

1. Introduction

The sensing and modeling of the environment are important in realizing an autonomous robot vehicle. Due to inadequacies of single sensor systems, the use of multiple sensor data fusion is increasingly popular. The choice of combining sonar and visual data in this paper is motivated by the complementary nature of the two information streams. We present a novel method of combining sonar and visual data to create a 3D sensing combination that models structured indoor environments. The sensor combination is intended for autonomous mobile robots operating indoors on problems such as localisation and map building. Few systems based on the fusion of sonar and visual data have been reported in the literature [2, 9, 11], due mainly to poor sonar sensors. This paper reports high resolution and accuracy, due to the state-of-the-art 3D sonar sensor [3, 7, 8] and the novel vision fusing strategy.

The structure of the paper is as follows: Section 1 introduces sensor data fusion and the motivation for combining sonar and visual data. In Section 2, an overview of the 3D environmental modeling technique is presented and the assumptions employed are stated. Section 3 details the algorithms and their limitations. In Section 4, the performance is illustrated with experimental results. Finally, conclusions and extensions are presented in Section 5.

1.1. Fusion of Vision and Sonar

Sensor data fusion is the combination of two or more information streams. By exploiting redundancy and complementarity of information, sensor data can be improved by reducing uncertainty and processing time, and by increasing accuracy and diversity [10].

Data from a CCD camera represents a 2D projection of 3D features in the environment. Stereoscopy, laser strip lighting and other techniques rely on triangulation in 2D to extract the third dimension of range. Conversely, a sonar sensor is intrinsically a detector of the range and employs triangulation and data association on three receivers to resolve the two bearing angles to targets. By combining the two information streams, we hope to exploit the accurate range sensing capability of sonar sensor while capitalizing on the detailed bearing perception afforded by a camera.

Processing of visual data can be time consuming as algorithms tend to maintain and exploit the 2D relationships that exists in raw visual data. In contrast, sonar sensing relies on specular reflections and diffraction from targets. For the sonar sensor to detect a target, a reflected pulse of sufficient amplitude must reach the sensor array. This can occur if a smooth surface presents a sufficiently large area perpendicular to the direction of the sensor, or an edge is sufficiently close to diffract back enough energy or a right-angled concave corner is within the beam width of the sensor. As a consequence, the sonar sensor usually detects relatively few targets compared to a vision system. The processing of the sonar data is fast compared to vision due to the sparseness of the sonar data. The possibility then arises for sonar to speed up the processing of the visual data by selecting corresponding regions of interest in the raw visual data, or conversely the dense visual data can enrich the sparse sonar data. We take the latter approach in this paper.

Little research has been reported on the fusion of visual and sonar data. The fusion of monocular image data and sonar scans is discussed by Jones *et al* [9], where the low azimuth resolution sonar range data is combined with vertical edge features from a camera to provide an environmental model. Matthies and Elfes [11] describe the integration of sonar and scanline stereo in creating an occupancy grid for mobile robot mapping. In that system,

0-7803-2988-4/96 $4.00 © 1996 IEEE

the complementary nature of the information provided by scanline stereo and sonar is exploited. The scanline stereo detects and localizes boundaries but misses unmarked surfaces. On the other hand, the sonar sensor detects broad surfaces and misses boundaries. Abidi [2] describes the fusion of stereo vision and ultrasonic range. The greatest error in stereo measurements are in the range, whilst the ultrasonic range data has good range measurements, but lacks lateral resolution due its wide beam pattern. These two complementary information sources are combined to improve the 3D coordinate estimate. The work presented here uses more refined sonar sensing techniques which provide accurate range, bearing and target type estimates. Thus the sonar sensing allows for more robust and accurate sensor data fusion with vision than previous techniques.

2. Environmental Modeling Approach

In this section, we present an overview of the 3D environmental modeling approach using sonar and vision. As most indoor environments contain an abundance of zontal and vertical lines, the environmental model is characterized using these features.

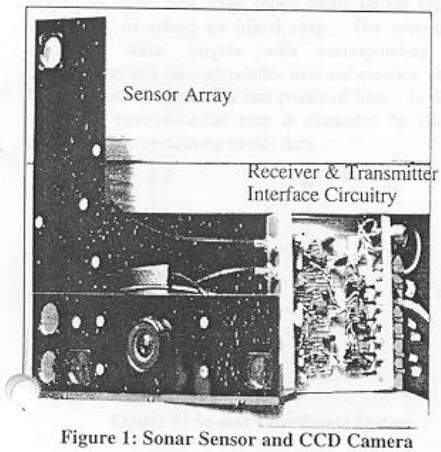


Figure 1: Sonar Sensor and CCD Camera

2.1. Sonar and Vision Pre-processing

A sonar sensor has been developed to localise and classify targets into 16 different target types [4,7,8], including targets seen by a camera as horizontal lines, vertical lines and points in an indoor environment. The sonar sensor consists of three receivers (bottom left in Figure 1) and three transmitters (transceiver bottom left, top left and bottom right in Figure 1). The three receivers enable 3D location estimation by providing range plus vertical and horizontal bearing information. The three transmitters provide complementary target information allowing target classification.

Figure 2 shows a simplified room scenario with an instance of each of target type that the sonar sensor is able to distinguish. Importantly, it classifies targets seen as visual lines into one of four categories: Horizontal Line Corner, Vertical Line Corner, Horizontal Line Edge and Vertical Line Edge. Furthermore, it classifies visual points into one of three categories: Point Corner, Point Edge and Complex Point.

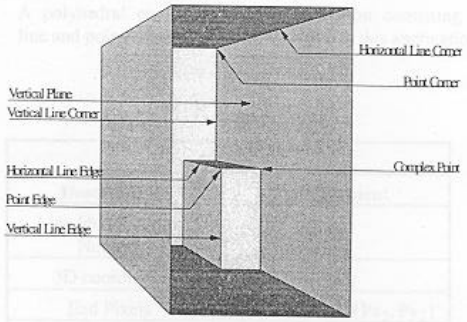


Figure 2: Simplified Indoor Scene

With the line type targets, the sonar sensor locates a 3D point in the direction normal to the line, defined by range, r, vertical bearing, α, horizontal bearing γ on the line. For the point type targets, the sonar sensor locates the 3D position of the point (r, α, γ). The coordinate system is illustrated in Figure 4.

The visual data is obtained using a greyscale CCD camera and is processed using a Hough transform to extract a set of equations of all lines that occur in the image. To facilitate data fusion, a relationship between a sonar 3D point, (r, α, γ), and the image pixel (Px, Py) is established. Consequently, a sonar point can be transformed into a pixel location on the camera view. Conversely, a pixel can be transformed to a line in 3D sonar coordinate space.

An outline of the fusion process is shown in Figure 3. Before attempting any fusion, the raw sonar data and the raw visual data are processed individually to a high level. The high level sonar data consists of 3D localization and classification information of detected targets. High level visual data consists of bounded straight lines. An ***unbounded*** line descriptor is one which identifies all points on a line. A ***bounded*** descriptor is an unbounded descriptor which also includes end points of the line.

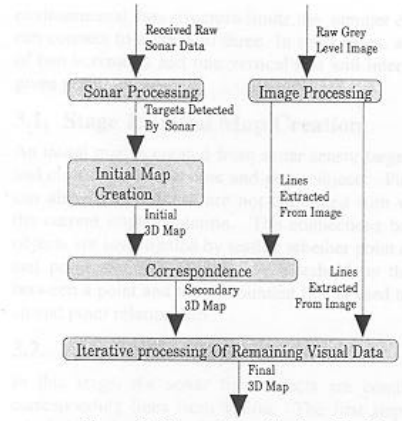


Figure 3: Sensor Data Fusion Model

The environmental modeling starts with the processed sonar data and successively incorporates straight lines from the high level visual data. Three stages are employed: The first stage relies solely on the processed sonar data to create an initial map. The second stage associates sonar targets with corresponding visual counterparts and this introduces new information provided by the visual data, such as end points of lines. In the final stage, the environmental map is expanded by iteratively combining the remaining visual data.

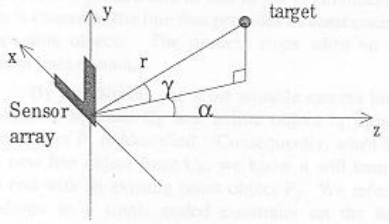


Figure 4: Sensor Coordinate System

2.2. Modeling Assumptions

An indoor environment composed of mainly horizontal and vertical lines is assumed in this paper, such as an office or warehouse.

The modeling system often encounters redundant location information for the same target from sonar and vision. For example, a point target has an estimated 3D sonar location $\{r, \alpha, \gamma\}$ and vision location at pixel (Px_1, Py_1). The 3D position of the point can be estimated from the sonar data alone. The pixel (Px_1, Py_1) provides redundant information and could be incorporated into the position estimate using techniques such as Bayesian statistics or Kalman filtering. However the accuracy of the sonar bearing angles has been established to be less than 0.02° [4,8] compared to a pixel bearing angle error of half the angular arc of a pixel, or 0.16°. In practice the pixel bearing is worse still due to noise in estimating edges in an image. Consequently data fusing will be heavily biased towards the sonar data and we simply choose the sonar data in preference to the vision where redundancy occurs.

3. Environmental Representation

A polyhedral environmental representation consisting of line and point objects is a natural choice in this application.

Table 2: Object Types

Line Object Type	
Description	**Data Element**
Line Identification Number	L
3D coordinate	r, α, γ
End Pixels	$\{Px_1, Py_1\}$ and $\{Px_2, Py_2\}$
Ultrasound Definitions	$Def_{horizontal}$ and $Def_{vertical}$
Simple Definition	Def_{simple}
Termination Points	P_1 and P_2

Point Object Type	
Description	**Data Element**
Point Identification Number	P
3D coordinate	r, α, γ
Pixel	$\{Px, Py\}$
Ultrasound Definitions	$Def_{horizontal}$ and $Def_{vertical}$
Connecting Lines	L_1, L_2 and L_3

The data structures for both line and point objects are shown in Table 2. A unique identification number is assigned to each line object and to each point object. The facility of maintaining the object's 3D location and the object's position on the image is provided by separate fields in the data structures. The line object data structure contains a field to register its simple definition which catagorizes the line as lying in a horizontal plane, being a vertical line or an anomalous line.

The connecting lines of a point object identify the lines which end at that point. This facilitates establishment and maintenance of connections between objects. Because only horizontal and vertical lines are represented, the

environmental data structure limits the number of lines that can connect to a point to three. In most cases, a maximum of two horizontal and one vertical line will intersect at any given point.

3.1. Stage 1: Initial Map Creation

An initial map is created from sonar sensor target positions and classifications for line and point objects. Plane targets can also be stored, but are not combined with vision with the current implementation. The connections between the objects are investigated by testing whether point objects are end points for line objects. A threshold on the distance between a point and the unbounded line is used to establish an end point relationship.

3.2. Stage 2: Correspondence

In this stage, the sonar line objects are combined with corresponding lines from vision. The first step identifies which bounded vision lines correspond to each 3D sonar line. This is the only stage where redundancy exists between vision and sonar. The map is refined with the end points obtained from vision by projecting them onto the 3D sonar line.

3.3. Stage 3: Iterative Processing of Remaining Visual Data

In this stage suitable camera lines are iteratively incorporated into the environmental model. At least one line object merged from sonar and vision must be present to start adding camera lines. The first step is to identify the most suitable camera line to add to the environmental map. This is chosen as the line that provides the best connectivity to existing objects. The process stops when no suitable camera lines remain.

By establishing the most suitable camera line C_k, a connection between C_k and a line object L_i through the point object P_j is identified. Consequently, when creating this new line object from C_k, we know it will terminate at one end with an existing point object P_j. We refer to this condition as a single ended constraint on the new line object. Furthermore, we must determine if the second end of the new line object is constrained. The connectivity of the camera line C_k is tested with all other existing line objects. If a line object L_m is found such that it connects with the other end of the camera line C_k at the end point object P_n (where $P_j \neq P_n$) then the two ends of the camera line are constrained. We refer to this condition as a double ended constraint on the new line object. The process of incorporating a new line object into the environmental model depends on C_k having a single or double ended constraint.

With a double ended constraint on the camera line C_k, both end points already exist and a new line object only is added. If C_k possess a single ended constraint, then an point object is also created corresponding to the end of the line. Note that this requires the simple definition (vertical or horizontal) of the line to establish the 3D location of the new objects using only the 2D camera line C_k. The simple definition of the line must be inferred from the 2D camera line C_k. We now derive a method for achieving the simple definition.

The orientation of camera lines generated by vertical and horizontal 3D lines are first derived. Using the coordinate from of Figure 4, the parametric equation of a vertical line is:

$$x = x_1;\ y = (y_2 - y_1)t + y_1;\ z = z_1 \tag{1}$$

The gradient of the line projected onto the camera plane, dx_p/dy_p, is zero, as shown below:

$$x_p = \frac{x_1}{z_1} z_p \Rightarrow \frac{dx_p}{dt} = 0$$

$$y_p = \frac{(y_2 - y_1)t + y_1}{z_1} z_p \Rightarrow \frac{dy_p}{dt} = \frac{(y_2 - y_1)}{z_1} z_p \tag{2}$$

$$\Rightarrow \frac{dx_p}{dy_p} = 0$$

Thus any 3D vertical line will always appear vertical on the camera projection. The gradient dx_p/dy_p of a 3D horizontal line on the camera projection plane can take any value including vertical as occurs when the horizontal line passes through the camera focal point. The gradient can be zero only when either y_1 is zero or it approaches infinity. This implies that a horizontal line can never cross the line $y_p=0$. These results show that the definition of the new line object as horizontal or vertical from its camera line C_k alone can lead to errors. A heuristic is employed to categorize the new line object in most situations:

1. *If C_k is vertical then the new line object is defined as **vertical**.*
2. *Otherwise, If C_k does **not** pass through the line $y_p=0$, then define as **horizontal**.*
3. *Otherwise define as **anomalous**.*

If indoor environments consist predominantly of vertical or horizontal lines, the rules will proficiently evaluate the definition of the new line object. However, errors in definition can occur. and are discussed in [4].

Once a definition for the most suitable camera line C_k is made, we are able to establish the 3D location of the new objects. The procedure iterates until no more suitable camera lines remain.

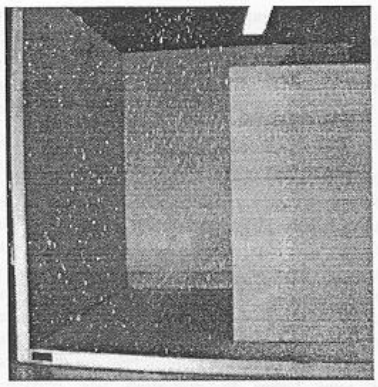
3.4. Errors and Limitations of Stage 3

Errors can occur as a result of the heuristic procedure to infer 3D classification from 2D camera data. Also it is

possible from a camera view to falsely connect two objects which are really at different ranges. Both these classes of errors can be quickly and easily detected and corrected by moving the sensor as described in detail in [4].

4. Experimental Results

A corner object with an overlapping plane is shown in Figure 5. The sonar sensor locates the vertical line corner of this structure. The lines extracted from the image are shown in Figure 5. Lines labelled 1 to 4 are not included in the final environmental model since they are not connected directly or indirectly to a sonar line or point.



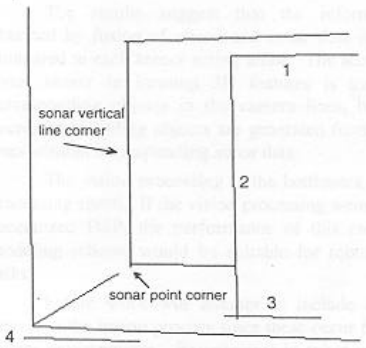
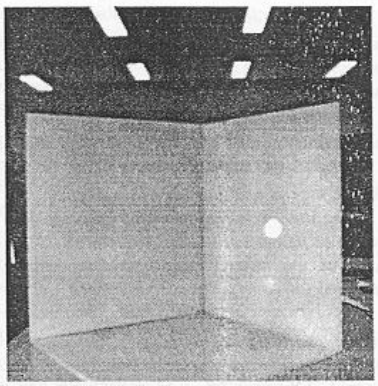


Figure 5: Original Image and Lines Extracted by Image Processing

The remaining lines in Figure 5 are correctly established in their 3D positions.

A second example illustrates the accuracy of the sensor with results from the target shown in Figure 6. The lines extracted from the image are labelled with numbers 1 to 7. Data is collected from 10 measurements. By physically measuring the lengths of lines and comparing them with the experimentally produced values, the errors list in table 2 is obtained.



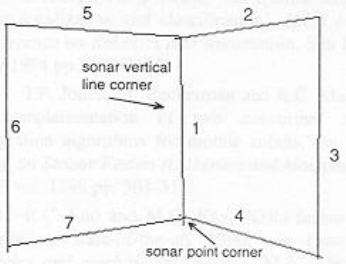


Figure 6: Camera View and Lines Extracted

The worst case error in measuring the length of lines is less than 4.2%. As expected, this error occurs far away from the original sonar data. The worst case standard deviation in measuring the length of a line is less than 2.6 cm or 5% of the line length.

Table 2: Statistics of Measuring the Length of a Line from 10 Samples.

Line	Physically Measured	Mean from Sensor	St.. Dev. Sensor	Error %
1	0.499 m	0.500 m	0.0075 m	0.213
2	0.493 m	0.498 m	0.0131 m	0.927
3	0.498 m	0.484 m	0.0104 m	-2.767
4	0.493 m	0.502 m	0.0141 m	1.760
5	0.496 m	0.504 m	0.0235 m	1.416
6	0.499 m	0.478 m	0.0258 m	-4.131
7	0.496 m	0.509 m	0.0255 m	2.583

With a 486 66 MHz processor, processing time is of the order of 90 seconds for this 3D reconstruction task.. The major part of this time is spent in vision processing. The time taken for the three stages of fusion is negligible compared to the vision processing time.

5. Conclusions and Extensions

In this paper, we have described a technique that fuses sonar and visual data to create a 3D environmental model with applications to robot navigation. The model characterizes the environment as a set of connected horizontal and vertical lines. Starting with sparse sonar observations, the environmental model is expanded successively to include lines seen in the visual data. The calculations from 2D vision to produce 3D environmental information are performed by making deductions as to the connection and definition of visual lines from the insufficient 2D data. Errors resulting from wrong deductions can be detected by moving the robot to second location.

The results suggest that the information gain obtained by fusion of visual and sonar data is significant compared to each sensor acting alone. The accuracy of the sonar sensor in locating 3D features is transferred to corresponding objects in the camera lines, but accuracy decreases as sibling objects are generated from the camera lines without corresponding sonar data.

The vision processing is the bottleneck in terms of processing speed. If the vision processing were moved to a specialized DSP, the performance of this environmental modeling scheme would be suitable for robot navigation tasks.

Future work will attempt to include sonar plane targets in the fusion process since these occur frequently in indoor environments. Experiments in robot map building using the sensor are envisaged.

References

[1] M.A. Abidi and R.C. Gonzalez, *Data fusion in robotics and machine intelligence*, San Diego: Academic Press, 1992

[2] M.A. Abidi, "Fusion of multi-dimensional data using regularization," in *Data fusion in robotics and machine intelligence*, M.A. Abidi and R.C. Gonzalez Eds., San Diego: Academic Press, ch. 10, 1992.

[3] H. Akbarally and L. Kleeman, "A sonar sensor for accurate 3D target localisation and classification", *IEEE International Conference on Robotics and Automation* 1995, Nagoya, Japan, May 1995 pp. 3003-3008.

[4] H. Akbarally, ***A vision supplemented sonar sensor for mobile robotics***, Masters by Research Thesis, Department of Electrical and Computer Engineering, Monash University 1995.

[5] M. Beckerman, "A Bayes-maximum entropy method for multi-sensor data fusion," in *Proc of the 1992 IEEE Int. Conf. of Robotics and Automation,* Nice France, May 1992 pp. 1668-1674

[6] A. Elfes, "Multi-source spatial data fusion using Bayesian reasoning," in *Data fusion in robotics and machine intelligence*, M.A. Abidi and R.C. Gonzalez Eds., San Diego: Academic Press, ch. 3, 1992.

[7] L. Kleeman and R. Kuc, "Mobile robot sonar for target localization and classification", International Journal of Robotics Research, Volume 14, Number 4, August 1995, pp 295-318.

[8] L.Kleeman and R.Kuc, "An optimal sonar array for target localization and classification", *IEEE International Conference on Robotics and Automation*, San Diego USA, May 1994 pp 3130-3135

[9] J.P. Jones, M. Beckerman and R.C. Mann, "Design and implementation of two concurrent multi-sensor integration algorithms for mobile robots," in *Proc. SPIE Conf. on Sensor Fusion II: Human and Machine Strategies,* 1989 vol. 1198 pp. 301-311.

[10] R.C. Luo and M.G. Kay, "Data fusion and sensor integration: state-of-the-art 1990," in *Data fusion in robotics and machine intelligence*, M.A. Abidi and R.C. Gonzalez Eds., San Diego: Academic Press, ch. 2, 1992.

[11] L. Matthies and A. Elfes, "Integration of sonar and stereo range data using a grid-based representation," in *Proc of the 1988 IEEE Int. Conf. on Robotics and Automation*, Philadelphia, USA, Apr. 1988.