14

MODELING OF MULTISENSORY ROBOTIC SYSTEMS WITH FAILURE DIAGNOSTIC CAPABILITIES

Guna Seetharaman and Kimon P. Valavanis

The Center for Advanced Computer Studies
University of Southwestern Louisiana
Lafayette, LA 70504-4330
guna, kimon @cacs.usl.edu

ABSTRACT

A multisensory robotic system (MRS) consists of a central high-level computer, one or more robotic manipulators with dedicated computer controllers and a set of diverse visual and non visual sensors. The intelligent, adaptive and autonomous behaviour of an MRS depends heavily on its ability to perceive and respond to the dynamic events that take place in its work environment. At any given instance, various factors, such as payload variations, the position, shape, orientation and motion of independently moving objects may affect the course of action taken by the MRS. The information required to detect potential failures, to distinguish between temporary failures (hard or soft), and to accommodate failures, is extracted from a diverse set of data. Complete perception is made possible through sensor fusion of the data (information) derived from the system's diverse set of sensors.

The chapter models the MRS as a hierarchical system with bidirectional interaction and focusses on the function and complexity of the vision subsystems. Various conditions that may cause the vision system to fail are illustrated. The problems involved in fusing (and registering) multisensory data are explained. The design of a new hybrid range and intensity sensor is explained. A VLSI architecture suitable for an MRS is also described.

1. INTRODUCTION

A multisensory robotic system may be modeled as a three interactive level system of organization, coordination and execution of tasks, a common structure of hierarchical systems [1]. The communication within the hierarchy is kept bidirectional to facilitate processing of the feedback signals. Given a user command, the system formulates plan candidates based on prior experience and information through various sensors, in order to evaluate the dynamic state of its workspace and adapt (if necessary) its course of action. The on-line dynamic interaction of the system with its environment of operation may dictate modifications in the execution of a specific task, or accommodation of local failures due to unexpected events.

The hierarchical structure of the system, and in accordance with previous studies [1], dictates that the organization level deals with off-line system functions while the coordination and execution levels deal with real-time, on-line dynamic situations occurring during the execution of a specific plan scenario. It is, therefore, the objective of the coordination level to develop specific execution scenarios and detect, identify, isolate and accommodate potential (local) failures related to the mechanical components of the system.

The coordination level is composed of a specific number of coordinators of a fixed structure, each performing a set of specific functions. For an MRS, these coordinators are defined to be: i) the vision system coordinator, ii) the motion system coordinator, iii) the gripper system coordinator, and, iv) the (non visual) sensor system coordinator.

Specific execution devices are associated with each coordinator, which executes specific tasks that the coordinator is being assigned. The coordinators do not communicate with each other (serially) directly; however, sharing and exchange of data between the coordinators is made possible by a dispatcher, common to all coordinators, the variable structure of which is dictated by the organization level [2].

This chapter concentrates on failures due to the vision subsystem. Methods are suggested to overcome several potential (soft) failures to enhance the flexibility of the vision system coordinator. The hardware mechanisms to be built are also related to the vision system coordinator components. Therefore, none of the other system components is affected. The organization level remains unchanged, too. However, the overall system performance is enhanced.

Vision (video) sensors provide a wealth of information that may be used by the system in several ways. The operational complexity of the vision subsystem in a multi sensory robotic system varies vastly. For example, in a simple situation, the robot may require information related to the presence/absence of any obstacle within its predefined path of motion, while in more complex situations, the position, orientation and the surface structure of a totally unknown object (kept in the workspace) must be understood in order to generate an acceptable *path of motion.* The 'path of motion' refers to the exact sequence of movements a robot manipulator follows in order to pickup an object for further manipulation. Consider for example a scenario where a robot manipulator must pickup an object A from location L_A and move it to a location L_B. A potential failure occurs, if the object is dropped by the manipulator while moving from L_A to L_B. Another potential failure occurs when the vision subsystem fails to recognize a known object (possibly due to noisy data) or an 'unknown' object entering the workspace environment. In all cases, the vision subsystem plays a dominant role in failure recovery.

Reflecting this large variation in the functional demands, the vision subsystem is required to operate over a large dynamic range of underlying complexity, resorting to simple, fast methods wherever and whenever it is sufficient to do so. The vision subsystem should operate under at least two different modes of operation: i) acquire coarse and fast measurements under normal conditions suitable for most model based vision applications, and, ii) acquire more accurate, complete and perhaps slow (not significantly) measurements required for failure prone conditions. When the vision subsystem finds itself inadequate to resolve the signals it should advice the co-ordinator (level) module which in turn will activate other (nonvisual) sensors to further resolve the scene using complex methods suitable for unstructured scenes.

Section 2 explains various aspects of the vision subsystem. The discussion includes the factors that could challenge the proper operation of the vision subsystem. It emphasizes the nature and difficulties involved in sensor fusion. The design of a hybrid range intensity sensor is described in section 3. The theory and operation of the sensor is covered in detail. The barrier removed by this sensor is emphasized. A VLSI implementation of the sensor is proposed. Section 4 concludes the chapter.

2. ROLE OF THE VISION SUBSYSTEM

Applications of three dimensional (3-D) machine perception techniques for autonomous systems have become very important in recent years. It has been demonstrated that the effectiveness and reliability of robotic assembly (RA) systems [3,4] and combat-oriented target identification systems [5], are significantly enhanced when they are endowed with 3-D visual (perception) feedback. Research on 3-D perception may be broadly classified into: i) understanding of the 3-D state of nature of a (structured) scene consisting of a known class of objects and, ii) understanding the 3-D state of nature of an (unstructured) environment where the presence of alien (unknown) objects is inevitable. Most of the DARPA lead research on image understanding [6, 7, 8] has been focussed on problems related to structured scenes with known objects.

3-D perception systems reported in literature [3, 4] are capable of perceiving the 3-D shape, orientation and location of objects within *static* as well as *dynamic* (slowly varying) scenes in the realm of structured/controlled environments. Published techniques may be broadly categorized into: i) passive monocular techniques (shape from shading [9], occlusion clues [8], surface orientation [10], and geometrical clues [11], [12]), ii) passive binocular techniques using photogrammetry [9], iii) dynamic scene analysis of monocular image sequences (motion-based techniques for objects with planar [13] and quadratic surfaces [14]) and, iv) fusion of images derived from multiple views [15], and multiple sensors (stereo analysis of intensity and range images [16]). Contributions made in the first three categories have made it possible, to a large extent, to solve many real-world applications where the scene is structured (or slightly unstructured).

2.1. Open Problems in Designing an Ideal Vision System

The fundamental problems in vision systems are generally associated with the many-to-one transformation that takes place during the image formation. Factors contributing to fundamental problems include:

1) Regardless of the sensor and the sensing methods used, the data suffer from a limitation called *finite volumetric aperture.* The objects self occlude themselves and prevent their back surface from being visible.

2) Depth ambiguities in orthographic images and scale ambiguities in perspective projections are inherent.

3) When more than one object are in the scene, critical parts of a specific object may be occluded by one or more objects making the recognition of the specific object almost impossible. Situations may occur where all the clues which facilitate unique identification of the specific object have been occluded by other parts in the scene to the extent that a known object is marked "unknown."

To illustrate further the above problems, consider the smallest sphere that completely encloses the object space to be monitored. A finite number of cameras may be positioned in orbits around this sphere to collect images from distinct vantage points in order to cover all of the 4π steradians possible views. However, physical imaging conditions require a surface of support for the objects; hence, cuts down the field of view to 2π steradians. Based on these restriction, some other sensors like tactile sensors may be used on or behind these surfaces to collect data. Therefore, i) images have to be registered somehow and, ii) while self occlusion is completely dealt with in the case of single objects, this approach is not a solution in the case of scenes with multiple objects.

The interpretation of 3-D information from 2-D images is similar to solving any other *ill posed inversion* problems. *Ill posed* problems are broadly divided into three groups: i) those with no solution at all, ii) those with no-unique solution and, iii) systems that do not depend continuously on initial data. It is apparent that we are dealing with the second group of problems. The general approach to such problems is to devise a set of consistency tests (functions) based on *a priori* knowledge of the solution space. That is, the problem is *regularized* by imposing a set of appropriate constraint in order to narrow the class of feasible solutions.

2.1.1. Principles of Model Based Vision

The process of *regularization* invariably involves minimizing some disparity functions, and/or energy functions. Methods that follow the *hypothesize and verify* approach tend to back project what was understood of the scene onto the image by first reconstructing the 3-D scene (hypothesized version) and then comparing its predicted image to the data, thereby minimizing certain regularity function. Least squared error functions are used in general. Situations do arise wherein the *visual perception* is meaningless while the *algebraic perception* is stable, – at least in the least squared sense.

One possibility is to take into account the image spatial structure of the error (disparity) image. The weighted structure-based error is interpreted in such a way that erroneous patterns which are more intolerable are assigned high cost functions. This leads to model based vision as a potential solution. The emphasis is on the underlying 2-D structure present in the 2-D image, from which the strong clues about the 3-D structure of the objects may be recovered. The images are segmented and described by a graph structure called *region adjacency graph (RAG).*

The 3-D perception problem reduces to finding a subgraph isomorphism between various RAGs and the anticipated 2-D structures of a 3-D object. The use of range images has been shown to accelerate the computation [3] and to increase the robustness. Consider the representation of the intensity image of a ball. The representation of the segmented image may indicate two patches. Albedos, the characters written on the surface of objects is another problem in representing the objects. Examples indicate that the validity of the 2-D RAG structure is critical. This may be achieved by having a coarsely sampled range image, or, by stereo vision.

2.1.2. *Introspective Vision: An effective Paradigm*

A major class of vision applications is related to ***introspective vision systems.*** An introspective vision system examines by definition a scene very thoroughly when necessary and plays a less significant role when everything in the scene conforms to what is expected of the scene. Upon identifying an event of importance in the scene, the vision system can specifically focus on to that location. For example, consider a model based vision algorithm devised to detect spheres. If an alien object is placed in the scene, the iterative computation may not converge. Eventually, the iterative algorithm would terminate saying that the data is ill conditioned. The objective of the introspective vision is to then gather adequate information and help the recovery process by a set of more complex algorithms designed to deal with alien but tractable objects. Generally speaking, introspective vision is highly directional, sensitive, and is nonuniform in nature.

In principle, mobile robotic systems are required to operate in dynamic unstructured environments. Such systems are equipped with binocular vision in order to detect 3-D objects and hence prevent collision. Fast response is required, and simplifying assumptions are necessary to adapt to any changes in the environment. Both binocular vision (spatial aggregation) and dynamic vision (temporal aggregation) techniques may be used to enhance the system flexibility and adaptability. Introspective vision requires that the robot be able to focus on every point in its workspace with almost equal sensitivity. It becomes necessary to dynamically alter the camera parameters to meet such specifications.

2.2. Sensor Fusion: An Alternative to Vision

Sensor fusion attempts to integrate information derived from two or more sensors of different modalities. The simplest application includes at least one *range image* and an *intensity image* of a scene recorded by a video camera and a depth sensor respectively. The objective is to measure those features (such as a spherical surface) using range images, albedo features (the identification labels or written text) of the surfaces by intensity based methods. The physical features such as the size and mounting hardware of these sensors (cameras) require that the sensors be placed apart in the 3-D space. Thus, each image contains certain information that may not be visible from the vantage point of the other sensors. The task is to integrate information from a set of (two or more) views of a 3-D scene in which each view is either a *range image* or an *intensity image.* Theoretical results in this area indicate potential for solving the complex problem of 3-D perception in an unstructured environment. To emphasize the difficulties involved, a description of the registration process is given in the next section.

2.3. Registration of Multi Sensory Images

Consider a *multi sensory* robotic system whose operation involves the 3-D perception of its workspace environment. The problem involves integration of information derived from: i) multiple video images and/or ii) multiple data sets where each data set is derived from a different sensor.

Let $\Phi_1(\boldsymbol{X};t)$ and $\Phi_2(\boldsymbol{X};t)$ be two distinctly different characteristics of the scene that are measured in a multi sensory system by two sensors $f_1(.)$ and $f_2(.)$, respectively. Also, let the two measurements $f_1(\pi)$, $f_2(\rho)$ be made available in two entirely different domains $\boldsymbol{\Pi}$ and $\boldsymbol{R}$ respectively. It is required to *register* the images by identifying the intrinsic relationship between these spaces so that the measured signals can be grouped easily. The complexity of the registration is determined by the nature of the $\boldsymbol{X} \rightarrow \boldsymbol{\Pi}$ and $\boldsymbol{X} \rightarrow \boldsymbol{R}$ mappings each of which may be *many to one* and *non invertible* in the worst case.

Consider a point $\boldsymbol{x'}_1 \in \boldsymbol{\Pi}$. Let f_1 and f_2 be a pair of intensity (video) and thermal (infrared) images. Then registration identifies a point $\boldsymbol{x'}_2 \in \boldsymbol{R}$ that corresponds to the given point $\boldsymbol{x'}_1$, so that the observed image-intensity values $f_1(\boldsymbol{x'}_1)$ and $f_2(\boldsymbol{x'}_2)$ may be grouped in the perception process. The points $\boldsymbol{x'}_1$ and $\boldsymbol{x'}_2$ are said to form *registration* or *point correspondence,* if they indeed represent the same physical point located in the scene. The example deserves a further comment in that both $\boldsymbol{X} \rightarrow \boldsymbol{x'}_1$ and $\boldsymbol{X} \rightarrow \boldsymbol{x'}_2$ are *many to one* and *noninvertible.* Therefore, given a point $\boldsymbol{x'}_1 \in \boldsymbol{\Pi}$ and a point $\boldsymbol{x'}_2 \in \boldsymbol{R}$ it is not possible to uniquely determine $\boldsymbol{X}$; hence, there is no direct procedure test if they form a registered pair. It is sufficient that at least one of the spaces $\boldsymbol{\Pi}$ or $\boldsymbol{R}$ be invertible.

When the overall objective is to monitor the workspace, one can assume specific geometric knowledge (to a certain extent) of the workspace. Then, at least in principle, for every point $\boldsymbol{X}$ in the workspace one may first compute its location in each image (or sensor domain) and then aggregate the information across many sensors. That is, for every $\boldsymbol{X}$ in the workspace, first compute $\boldsymbol{X} \rightarrow \boldsymbol{x'}_1$ and $\boldsymbol{X} \rightarrow \boldsymbol{x'}_2$ and then use $f_1(\boldsymbol{x'}_1)$ and $f_2(\boldsymbol{x'}_2)$ for fusion. Such applications are said to operate in a *structured environment* in that the 3-D structure of the objects and the position as well as the orientation of the cameras in the scene are known *a priori.*

Real world applications, however, are more complicated. Most systems, in fact, are required to operate in *unstructured environments* where the 3-D geometrical (spatial) structure can not be assumed explicitly *a priori.* The processes of registration, recognition, as well as localization tasks are indirectly related. The necessary condition for registration is that: at least one of the sensors, say $f_i(.)$, must have a *one to one* and *invertible* mapping $\boldsymbol{X} \rightarrow \boldsymbol{x'}_i$ which permits to compute $\boldsymbol{x'}_i \rightarrow \boldsymbol{X}$ uniquely. The registration is further complicated due to the discretization of the $\boldsymbol{\Pi}, \boldsymbol{R}, \cdots,$ spaces as a result of the sampling process.

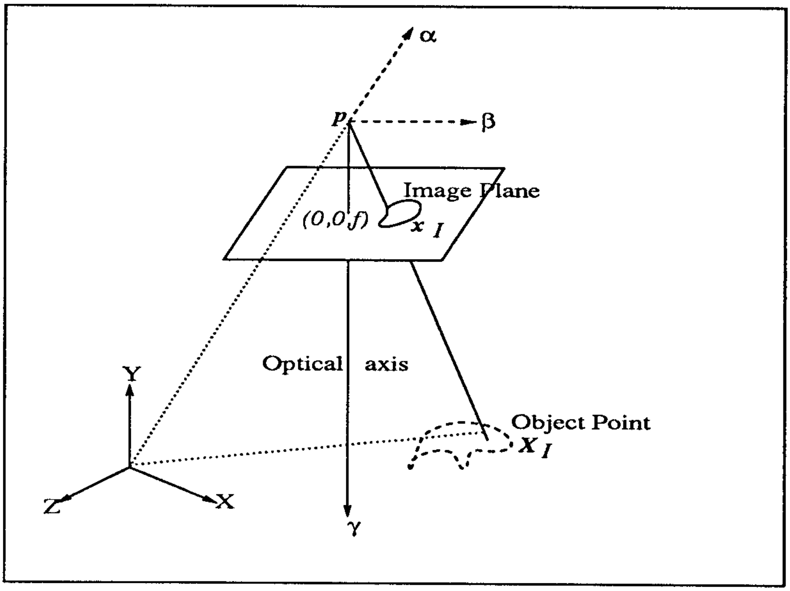

Figure 1. A simple perspective imaging system with its origin located at *p*.

2.3.1. *Loss of Depth in Perspective Imaging*

The intrinsic geometric model of an intensity camera is illustrated in Figure 1. $\boldsymbol{X}_I$, $(X,Y,Z)_I$ and/ or (X_I, Y_I, Z_I) are used to represent the position (X_I, Y_I, Z_I) of an arbitrary point $\boldsymbol{X}$ measured with respect to the camera coordinate system I. In general, the intensity camera projects a certain point $\boldsymbol{X}_I$ located on the surface of an opaque object onto an image point $\boldsymbol{x}_I = (x, y, z = f)_I$ located on the image plane. The image plane is uniquely determined by the focal length f of the camera, and satisfies the equality $Z_I = f$. An irreversible loss of depth information is introduced by the underlying perspective projection expressed as:

$$\begin{bmatrix} x \\ y \\ f \end{bmatrix}_I = [\boldsymbol{P}] \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_I \tag{1}$$

where,

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{\lambda} & 0 & 0 \\ 0 & \frac{1}{\lambda} & 0 \\ 0 & 0 & \frac{1}{\lambda} \end{bmatrix} \qquad \text{with,} \quad \lambda = \frac{Z_I}{f} \qquad \text{and} \quad \lambda > 1 \tag{2}$$

That is, both $\boldsymbol{X}_I$ and $a\boldsymbol{X}_I$, where $a \neq 0$, result in the same image point. Therefore, (1) is noninvertible in that given $\boldsymbol{X}_I$ one can determine $\boldsymbol{x}_I$ but not the opposite. However, given a point $\boldsymbol{x}_I$ on the intensity image, $\boldsymbol{X}_I$ is constrained to a line (of points) passing through the focal point $(0,0,0)_I$ and the image point $(x, y, z = f)_I$.

Given the absolute position $\boldsymbol{X}$ of a point (with respect to the world coordinate system), both $\boldsymbol{X}_I$ and hence $\boldsymbol{x}_I$ are described as follows:

$$\begin{bmatrix} X_I \\ Y_I \\ Z_I \\ 1 \end{bmatrix} = [\boldsymbol{T}] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3a}$$

where,

$$\boldsymbol{T} = \begin{bmatrix} \alpha_x & \alpha_y & \alpha_z & \vdots & -\boldsymbol{\alpha}^T.\boldsymbol{p} \\ \beta_x & \beta_y & \beta_z & \vdots & -\boldsymbol{\beta}^T.\boldsymbol{p} \\ \gamma_x & \gamma_y & \gamma_z & \vdots & -\boldsymbol{\gamma}^T.\boldsymbol{p} \\ \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ 0 & 0 & 0 & \vdots & 1 \end{bmatrix} \tag{3b}$$

$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ = direction cosines of the camera's X,Y and Z axes,

$\boldsymbol{p}$ = vector position of the origin of the camera coordinate system.

The matrix $\boldsymbol{T}$ is uniquely characterized by six parameters, and is always invertible. These parameters are easily calculated when the camera position and orientation are known; also, in principle, these parameters can be experimentally estimated by some calibration techniques. From equations (1), (2) and (3) it follows that:

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = [\boldsymbol{T}]^{-1} \begin{bmatrix} \lambda x_I \\ \lambda y_I \\ \lambda f \\ 1 \end{bmatrix} \equiv [\boldsymbol{T}]^{-1} \begin{bmatrix} \lambda \boldsymbol{x}_I \\ -- \\ 1 \end{bmatrix}. \tag{4}$$

Thus, the absolute position $\boldsymbol{X}$ of a point is constrained to a line by its image $\boldsymbol{x}_I$.

2.3.2. Recovery of Depth from Stereo Images

Consider a multi sensory system consisting of two intensity cameras, called L and R. These cameras will also be referred to as left and right cameras respectively. The objective is to extract the depth of the observed object points by using the left and right images. The notations, $\boldsymbol{X}_L, (X, Y, Z)_L$ and/or (X_L, Y_L, Z_L) are used to represent the position (X_L, Y_L, Z_L) of an arbitrary point $\boldsymbol{X}$ measured with respect to the coordinate system L. Let the focal length of the left camera L be f_L, and the image of a point $\boldsymbol{X}_L$ be defined by $\boldsymbol{x}_L$ in a manner consistent with previous definitions. Similar definitions hold for the right camera R. Let $\boldsymbol{X}$ be an object point whose image is at $\boldsymbol{x}_R$ and $\boldsymbol{x}_L$ from the right and left images respectively. The points $\boldsymbol{x}_R$ and $\boldsymbol{x}_L$ form a *registered pair* or a *point correspondence.* From (4) it is concluded that:

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = [\boldsymbol{T}_R]^{-1} \begin{bmatrix} \lambda_R x_R \\ \lambda_R y_R \\ \lambda_R f_R \\ 1 \end{bmatrix} = [\boldsymbol{T}_L]^{-1} \begin{bmatrix} \lambda_L x_L \\ \lambda_L y_L \\ \lambda_L f_L \\ 1 \end{bmatrix} \tag{5}$$

where both λ_R and λ_L are unknown, positive real numbers greater than unity. By equating the corresponding entries, three equations (6) in the two unknowns are obtained and solved uniquely for λ_R and/or λ_L hence the absolute position of the object point $\boldsymbol{X}$. The equations for this process are:

$$\begin{aligned} x_R\lambda_R - (r_{11}x_L + r_{12}y_L + r_{13}f_L)\lambda_L &= t_x \\ y_R\lambda_R - (r_{21}x_L + r_{22}y_L + r_{23}f_L)\lambda_L &= t_y \\ f_R\lambda_R - (r_{31}x_L + r_{32}y_L + r_{33}f_L)\lambda_L &= t_z \end{aligned} \tag{6}$$

where,

$$\left[\widehat{\boldsymbol{T}}\right] = [\boldsymbol{T}_R]\ [\boldsymbol{T}_L]^{-1} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

When x_L and x_R are known, the depth of $\boldsymbol{X}$ can be computed by solving (7) for λ_L as:

$$\lambda_L = \frac{f_R t_x - x_R t_z}{(r_{31}x_L + r_{32}y_L + r_{33}f_L)x_R - (r_{11}x_L + r_{12}y_L + r_{13}f_L)f_R} \tag{8}$$

2.3.3. Registration of Stereo Images

The major problem in stereo vision is with establishing the point correspondence, *i.e.*, identification of the pairs x_L and x_R. A large number of these pairs are required to compute a densely sampled depth image of the 3-D workspace.

Consider a problem instance where x_L is known and it is required to uniquely determine the corresponding point x_R. Further inspection of (6) reveals that there are three equations in four unknowns, namely x_R, y_R, λ_R and λ_L. If either λ_L or λ_R is known, then one could solve for x_R. However, the very objective is to compute λ_L and/or λ_R. Eliminating λ_L and λ_R in (6) results in [**14**]:

$$(x_R, y_R, f_R) \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \begin{bmatrix} x_L \\ y_L \\ f_L \end{bmatrix} = 0 \tag{9}$$

where:

$$\begin{aligned} e_{11} &= (r_{21}t_z - r_{31}t_y) & e_{12} &= (r_{22}t_z - r_{32}t_y) & e_{13} &= (r_{23}t_z - r_{33}t_y) \\ e_{21} &= (r_{31}t_x - r_{11}t_z) & e_{22} &= (r_{32}t_x - r_{12}t_z) & e_{23} &= (r_{33}t_x - r_{13}t_z) \\ e_{31} &= (r_{11}t_y - r_{21}t_x) & e_{32} &= (r_{12}t_y - r_{22}t_x) & e_{33} &= (r_{13}t_y - r_{23}t_x) \end{aligned} \tag{10}$$

The interpretation of (9) is that, given $\boldsymbol{x}_L$ the expected value of $\boldsymbol{x}_R$ is constrained to a line. Thus, given a pair of points, one can test if they form a point correspondence. Different values of $\boldsymbol{x}_L$ generate distinctly different lines. All of these lines pass through a same point in the $(x, y, z = f)_R$ plane. These lines are called *epipolar lines* and they all concur at a point called *epicenter.* The *epicenter* is actually the image of $\boldsymbol{X}_L = 0$ imaged on the image plane of the camera R. It is not possible to identify the desired $\boldsymbol{x}_R$ even though $\widehat{\boldsymbol{T}}$ and $\boldsymbol{x}_L$ are known. Additional information is necessary to uniquely determine the corresponding point $\boldsymbol{x}_R$ when $\boldsymbol{x}_L$ is given.

2.3.4. Registration in Structured Environments

It is instructive to examine if the problem may be simplified in structured environments. When the parametric form of the surface is known one gets an additional constraint to solve, the equation (9). It can be shown that the problem is still complex in that the knowledge is insufficient even when the structure (orientation) is fully known.

To prove the statement assume that the object point is located on a planar surface. The objective is to solve for x_R, y_R when (x_L, y_L, f_L) is known. First, the equation of the plane is expressed conveniently in the form:

$$Z = -pX - qY - s \tag{11}$$

where (p, q) uniquely define the orientation of the planar face of the object and s is a parameter that fixes (uniquely) the object face. Let (p_L, q_L, s_L) and (p_R, q_R, s_R) describe the plane uniquely with respect to the cameras L and R. Actually $(p, q, s)_R$ may be computed from $(p, q, s)_L$ and $\widehat{\boldsymbol{T}}$. By substituting $\boldsymbol{X}_L = \lambda_L \boldsymbol{x}_L$, in (11) we get:

$$\lambda_L = -\frac{s_L}{(p_L x_L + q_L y_L + f_L)} \tag{12}$$

From (11), (12) and (6) one could show that s_L is required to uniquely determine x_R from x_L. That is, p_L, q_L and s_L must all be known to uniquely compute x_R. In effect, we require that the 3-D location and orientation of the planar face be known *a priori.* However, the very objective of the stereo vision system is to locate the object. It is permissible to assume only (p, q) as known and not s. Thus it is clear that x_R and y_R can not be determined uniquely, in the absence of λ_L and λ_R.

2.3.5. Registration in Unstructured Environments

If the underlying problem is the recovery of 3-D shape of an unstructured scene, the point correspondence has to be established by clues that do not in anyway restrict the geometrical shape or state of the scene, for example color or spatial signatures.

One practical approach is to extract a number of candidate points ξ_R and ξ_L from the $(Z = f)_R$ and $(Z = f)_L$ image planes respectively, where the observed image indicates distinct features. Then, for each point in ξ_R, its potential match is expected (most likely) to be present in ξ_L. Certain correlation operators may be applied to evaluate the likelihood of a match.

It is clear, that we are confronted with a fundamental issue in that, 1) we need 3-D position of the object point in order to extract the point correspondence; 2) the very objective of establishing point correspondence is to extract the 3-D position of these points.

2.3.6. Registration of Two Image Sequences

Temporal variations in an image sequence are isotropic features that are easily measured and processed. In principle, it is possible to extract the time varying nature of intensity at each pixel in each image sequence, and be able to assign the pixels to one of many classes. For each pixel with a particular temporal signature in a particular sequence, one could expect its corresponding pixel in the other image sequences to indicate the same. Hence the potential match can be found using a finite search. The only requirement is that the object must move in space and time and/or the scene must be dynamic in nature.

Recovery of 3-D motion and orientation of objects from an image sequence is a problem that has gained attention in the past decade [13]. Several approaches have been proposed to recover both the shape and orientation of the objects in addition to recovering the motion parameters, $\widehat{T}$. The derivation of (9) is taken from the point correspondence approach due to [14]. The method relies on external aid to establish the point correspondence. A line correspondence approach due to [17] indicates some improvement, nonetheless we are confronted with identifying lines in both images that correspond to each other. The method may not be suitable for scenes where polyhedral objects are least likely. A region-based region-correspondence approach developed in [13] reduces the burden in establishing the correspondence. However, the method is applicable to planar faced objects only. All of these methods can be utilized fully, if the temporal signatures were taken into account for establishing the match first, and then the recovery of the motion parameters of the objects next. These techniques were primarily developed for monocular image sequences, and are easily applied to provide registered stereo image sequences.

2.4. Sensor Fusion between Range and Intensity Images

Several sensors are used to constantly monitor the environment in order to detect and respond to the dynamic changes in the scene. These sensors measure different characteristics of the workspace and provide information of complementary nature. Availability of range and intensity images has been shown to simplify certain robotic tasks [18]. However, there is a bottleneck in these applications with image registration [16, 19]. The resolution, sensitivity, mechanical characteristics and dynamic range of each sensor vary considerably from that of the other sensors. Such variations in the resolution make it difficult to establish registration and thus restrict potential applications. In the limiting case all of the sensors are considered to be identical in that all sensors are *intensity cameras* or *range cameras*. A multi sensory system should include at least one sensor of each type. The objective behind fusing information from multisensory images is to achieve a 3-D perception of complex scenes [4]. The complexity of the recognition task is directly influenced by real-time performance requirements and the degree to which the workspace is kept free of foreign objects.

3. SENSOR FUSION: A HYBRID RANGE-INTENSITY SENSOR

The design of a low cost, hybrid *range-intensity* sensor is described in this section. It is expected to promote significantly the implementation of sensor fusion and contribute to the advancement of research in this area. The salient features of the sensor include: i) low-cost, ii) reliability and less sensitivity to the misalignment of moving parts, and, iii) VLSI implementation. The sensor offers *six* different operational modes to provide: i) two (binocular) intensity images, ii) two (binocular) range images observed from two vantage points, iii) registered pairs of *range and intensity* images, iv) binocular intensity image sequences; v) binocular range image sequences, and. vi) multi sensor, multiview image sequences respectively.

In one mode of operation the sensor is viewed as a pair of intensity cameras operating under a stereoscopic configuration. When the scene is completely structured, this mode facilitates the use of dedicated, inexpensive intensity-based image processing hardware. In the second mode, the system operates as a range-intensity sensor, and delivers two range image sequences and two intensity image sequences.

The *range and intensity* sensor offers the following advantages: i) easily alterable camera /sensor parameters with extended dynamic range to support introspective vision, ii) registered pairs of intensity and range images with accuracies that are of the same order, iii) fast acquisition times including real-time embedded control algorithms to facilitate the use of this sensor in closed loop applications, iv) less sensitivity to positioning errors, easy calibration as well as linearization, and/or compensation of spatial disorders, and, v) partial implementation of steps iii) and iv) in VLSI.

The specific design details pertaining to a prototype sensor involving two cameras, as well the design of a VLSI based Radon Transform Processor are now explained in detail.

3.1. The Principle of Operation

The basic structure of the hybrid sensor is illustrated in Figure 2. The sensor consists of four major components:

i) A pair of video cameras that can monitor the scene under existing lighting conditions.

ii) A laser beam-spreader that generates and steers (deflects) a planar sheet of laser-beam as shown in Figure 2. When the laser beam is made incident on the object surface it generates a contour which is referred to as laser induced contour (LIC). The geometric nature of the LIC is a function of the surface parameters. LIC is a straight line for planar surfaces and it is a conic for quadratic surfaces.

iii) Control and coordination system to position these two cameras in a desired geometrical relationship.

iv) Hardware components to extract the laser induced contours (LIC) in each image to calculate /identify point correspondence of the points located on these LICs and to recover the depth image from these corresponding points.

The design of the proposed sensor takes advantage of the stereopsis between the cameras to recover a registered pair of intensity and depth images. The registration problem is trivially solved by extracting the LICs from each image and making use of (9). Given a point $\boldsymbol{x}_L$ on the LIC observed in the left image, it is clear that its corresponding point must satisfy (9) and must be also located on the LIC extracted from the right image. The orientation of the plane of laser sheet is not required since the sensor operation does not use that information.

3.2. Image Registration in Real Time

In real applications this results in attractive hardware solutions. For the y^{th} row in the left image one may first find the position of the point at which the LIC intersects that row, by fitting a gaussian. A patented algorithm involving two fast adders and one multiplier/divider has been used in [20]. Given the value of $(x, y, f)_L$, there may be a digital differential analyzer (DDA) that will generate a set of points in the second image (*i.e.*, right image plane) located on the line defined by (9) to facilitate a search for the point where this line

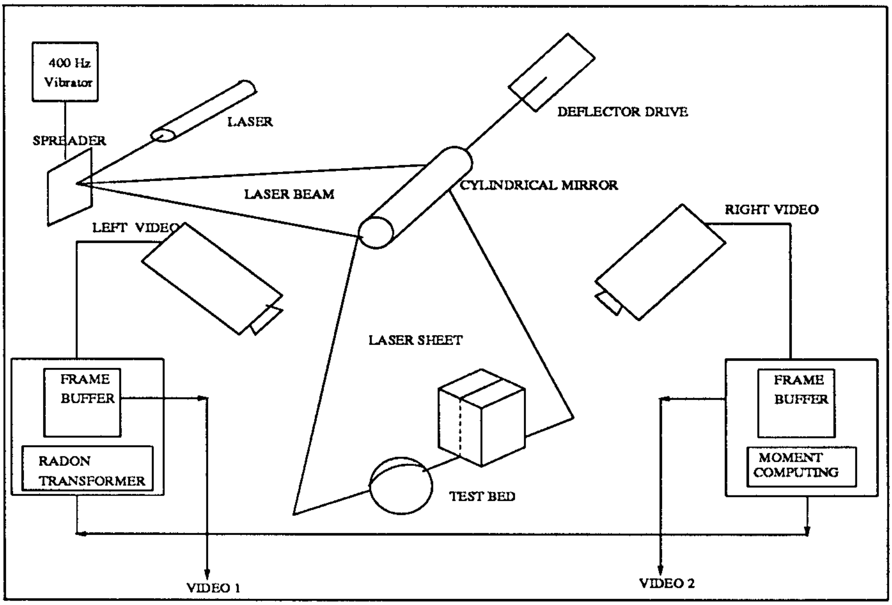

Figure 2. The functional components of the Range-Intensity Sensor in a object recognition application.

intersects the image of the LIC. The line is in fact determined by the values of each $(x, y, f)_L$ and the T_{IR} parameters as given in (9). Exactly one DDA is required to identify each point correspondence. It is known that each DDA would detect exactly one match, within a finite amount of time. A set of DDAs operating in parallel may be designed making a VLSI solution feasible. In fact, exactly N_L number of DDAs will be required, where N_L is the number of raster lines derived from the left camera. For commercially available cameras this number varies from 480 to 625 considering NTSC and PAL systems as extreme cases.

The point $(x, y, f)_L$ in the left image is given at a subpixel accuracy since it is computed based on moments[20]. The DDAs on the other hand generate the line of search (within the right image plane) at the pixel resolution. One way to handle this disparity is to fit a gaussian pattern along this line, too. The model is justified since an oblique cross-section of gaussian hill is also a gaussian with its centroid in place. Thus, an integration (or summation) operator is conducted over the line of points generated by each DDA. This is in fact one component of the *radon transform* of the right image. The general form of radon transform is defined as follows:

$$R(\rho, \theta) = \int_x \int_y K(x, y, \rho, \theta)\, f(x, y)\, \delta(x \cos\theta + y \sin\theta - \rho)\, dx dy \qquad (13)$$

where K is called a kernel function of the transformation, ρ and θ define a line and δ is the Kronecker function.

At any given instance, the parameters (ρ, θ) that determine the line in (13) are considered available through (9) and (10) since $T_{LR}, \boldsymbol{x}_L$ are known. Hence, the line of integration is uniquely fixed by $(x, y, f)_L$. The values (ρ, θ) which determine these families of lines on the right image plane may be precomputed and stored in a lookup table. The memory used to store $\rho\,(x, y, f)_L$ and $\theta\,(x, y, f)_L$ can be updated if the relative orientation is changed over time. Only 512 programmable radon-transform processors may be required. Although (13) appears to contain multiplication, one can still use double adder mechanisms for this specific application. The parallel operations of these radon transform processors do however introduce interesting problems related to the memory organization of the second image. A busy traffic (in terms of memory access) is expected for pixels $\boldsymbol{x}_R$ that are closer to the *epicenter* in the right image plane. The design of the parallel radon-transform processors must account for the potentially simultaneous access to the gray level value stored at these pixels.

3.3. The Architecture of Sensor Controller

The architecture of the overall sensor is illustrated in Figure 3 for a configuration containing two cameras. It is assumed that both of the cameras are synchronized. A frame buffer is essentially a high speed random access memory (RAM), which is scanned /accessed in a particular way. The processors are organized in a specific manner so that each processor can easily access the image stored in the RAM of any other processor.

One of the video processors, say the primary L, includes a real-time moment computing circuit, capable of computing the first and second moments of the pixel intensity values along each horizontal line of its image. The computed moments are then processed to locate the point $(x, y, f)_L$ where LIC intersects that line. Estimation of $(x, y, f)_L$ based on the moments has to be performed within 63μs (assuming RS-170 standards) before the end of the next horizontal scan. This point is then read by the central computer, which is then used to initiate the search for corresponding points in the other video images.

The video processor of each secondary camera, for example R, contains a set of Radon Transform Processor Elements (RPE). The RPEs facilitate the search for the point $(x, y, f)_R$ located on the line defined by (9) where it intersects the LIC. The estimation of $(x, y, f)_R$ is based on the first and second moments computed along the line, similar to the process used in the estimation of $(x, y, f)_L$. At the end of each horizontal trace, exactly one radon processor in each secondary video processor will be loaded with the necessary information (ρ, θ) to conduct the search for corresponding points. The exact values are

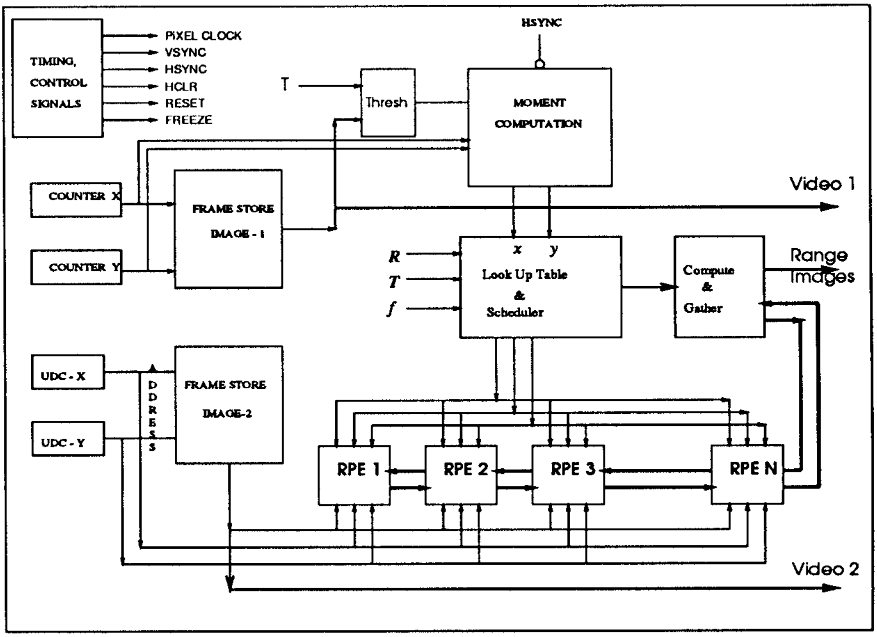

Figure 3. The architecture of the range-intensity sensor.

computed by using the most recently extracted x_L in (9) as follows:

$$\begin{aligned} \theta &= \arctan\left[\frac{e_{11}x_L + e_{12}y_L + e_{13}f_L}{e_{21}x_L + e_{22}y_L + e_{23}f_L}\right] \\ \rho &= \frac{e_{31}x_L + e_{32}y_L + e_{33}f_L}{\sqrt{(e_{11}x_L + e_{12}y_L + e_{13}f_L)^2 + (e_{21}x_L + e_{22}y_L + e_{23}f_L)^2}} \end{aligned} \tag{14}$$

Computation of (14), and the initialization of necessary constants are to be performed by the central controller; these operations are condensed into a block called look up tables and scheduling in Figure 3.

It is desirable to operate all the RPEs to compute the radon transform using maximum parallelism. Two fundamentally different scheduling strategies may be considered for scheduling to facilitate the concurrent computation of moments by all the RPEs. The first scheduling strategy is to activate each RPE as soon as the necessary parameters have been downloaded. This implies that a large number of loosely coupled RPEs computing moments along distinct lines, with the pixel values being fetched (read only) from a shared memory. This approach poses severe constraints on the design of the memory. The second strategy is to operate the RPEs under a data parallel SIMD configuration with no conflicts in memory access. The principle of this operation is as follows: the contents of the RAM (image plane) will be traversed in a specific order and broadcast it to all the RPEs within that video processor; each RPE must then test if the pixel lies on the line of its interest then capture accordingly both the coordinates and the value of the pixel for computing the moments; the entire operation takes $O(n^2)$ time for an image of $n \times n$ pixels.

The data parallel approach was chosen to suit the nature of the problem; however, the video processors can also be used for computing general purpose radon transforms. For a given LIC in space, that is for a fixed position of the laser beam deflector, the total time taken for establishing all possible point correspondence is expected to be an integral multiple of the time required to traverse all the pixels. Consequently, the sweep rate of the cylindrical mirror is determined by this time as well.

The programming model of the RPE chip, including the important signals is described in[21] The adders and comparators are implemented in bit serial logic. Current projections indicate that at least 16 of these RPEs will be integrated into one VLSI chip. Each VLSI chip will also include a pipelined multiplier, and a realtime serial to parallel interface transposer (SPINT) network, to facilitate fast fixed-point multiplications. The proposed linearly connected, RPE array is easily adapted for both forward and inverse Radon Transformation of general purpose kernels. For a detailed discussion, the reader is referred to [21] and [22].

4. CONCLUSION

The proposed sensor significantly reduces, if not eliminates, the problem of registering two or more images of a scene viewed from different vantage points. No explicit assumption was made about the object surfaces. The result is a simple, robust sensor capable of recording multiview video images, and densely sampled range (depth) scene images. Each video camera in the sensor provides a video image sequence over time. Thus, the sensor facilitates the application of dynamic scene analysis, such as the recovery of 3-D motion parameters, shape as well as object orientation. In particular, this is very useful in autonomous land vehicles.

Traditionally there are two ways of perceiving a scene. The first is to get a number of images from many different vantage points. Any conclusion reached from these images is said to follow *spatial clues*. There is evidence in human perception that spatial cues play an important role in perceiving static scenes. The mechanism assumes that extreme conditions in the images represent extreme conditions in the scene. The success depends on the structure of the scene. Spatial inference is made possible if the objects are rich in features.

The second approach to perceiving a scene is called temporal surveillance, in which the aggregation is somewhat spatio(local)-temporal. A hypothesis of the spatial configuration of the scene is then formed based on how the local evidence (temporal information) vary (spatially) between the images. This is called time-aperture of the sensing phase. No initial knowledge is necessary except that one has to rely on the mobility of the subjects within the field of view. This corresponds to unstructured dynamic environment. The stereo image sequences derived from the sensor facilitate such analysis.

The proposed sensor is expected to ease the barriers for research in sensor fusion, and integration of spatial and temporal scene analysis of ***unstructured*** *scenes.* Adaptive intelligent systems, capable of operating (primarily) based on model-based vision algorithms may now detect and gracefully switch the mode into complex vision algorithms for unstructured environments. Insight gained may further our knowledge of integrating multisensory, spatio-temporal image sequences.

The present sensor may be applied in: 1) a mobile robotic system; 2) multi-armed multisensory robots; 3) aerial sensors for reconnaissance. The VLSI circuitry developed related to this sensor may be used in synthetic aperture radar systems.

A multi sensory robotic system equipped with the proposed sensor will have enhanced visual capabilities and will be able to recover from local failures related not only to model based vision but also to changes within the dynamic (unstructured) environment of operation.

5. BIBLIOGRAPHY

[1] Kimon P. Valavanis. *A Mathematical Formulation for the Analytical Design of Intelligent Machines.* PhD thesis, Rensselaer Polytechnic Institute, Troy, N.Y., 1986.

[2] Kimon P. Valavanis and George N. Saridis. *Intelligent Robotic Systems: Theory, Design and Application.* Kluwer Academic, 1992.

[3] Robert C. Bolles, Patrice Horaud, and Marsha Jo Hannah. 3DPO: A three dimensional part orientation system. In *Readings in Computer Vision*, pages pp. 355–359. Academic Press, 1989.

[4] Roland T. Chin and Charles R. Dyer. Model based recognition in robot vision. *ACM Computing Surveys*, Vol(18):68–108, March 1986.

[5] Department of Army and The Department of Research. *Proceedings of the Conference on: Pattern Recognition for Advanced Missile Systems, Huntsville, Alabama.* Department of Defense, 14-15 November 1988.

[6] Martin A. Fischler and Oscar Firschein, editors. *Readings in Computer Vision: Issuses, Problems, and Paradigms.* Morgan Kaufmann, 1987.

[7] Defence Advanced Research Projects Agency. *DARPA Neural Network Study.* AFCEA Press, Fairfax VA, November 1988.

[8] Michael Brady. Computational approaches to image understanding. *ACM Computing Surveys*, Vol. 14(No. 1):pp. 3–72, March 1982.

[9] Berthold K. P. Horn. *Robot Vision.* The MIT Press, Cambridge, MA., 1987.

[10] Takeo Kanade. Recovery of the 3d shape of an object from a single view. *Artificial Intelligence*, Vol.17:409–460, 1981.

[11] Tzay Y. Young, Guna S. Seetharaman, and Wasim J. Shomar. A rule-based system for 3-D shape recovery from a single persspective view. In *Proceedings of The SPIE Conf. on Applications of Artificial Intelligence, VI, Orlando*, pages pp.294–302, Vol–937, 4-6 April 1988.

[12] Stephen T. Barnard. Interpreting perspective images. *Artificial Intelligence*, AI-21:435–462, 1983.

[13] Guna Seetharaman. *Estimation of 3-D Motion and Orientation of Rigid Objects from an Image Sequence: A Region Correspondence Approach.* PhD thesis, University of Miami, Coral Gables, Miami, August 1988.

[14] Roger Y. Tsai and Thomas S. Huang. Uniqueness and estimation of three dimensional motion parameters of rigid objects with curved surfaces. *IEEE Trans. on Pattern Analysis and Machine Intell.*, PAMI(6):545–554, 1984.

[15] P. G. Mulgaonkar. Multiview image acquisition for postal parcels. *Advanced Imaging*, (No. 2):pp. 44, Feb 1991.

[16] R. O. Duda, D. Nitzan, and P. Barrett. Use of range and reflectance data to find planar surfaces. *IEEE Trans. on Pattern Analysis and Mach. Intell.*, PAMI(1):259–271, July 1979.

[17] J. Weng, Y. Liu, T.S. Huang, and N. Ahuja. Determining motion/structure from line correspondences: A robust linear algorithm and unqueness theorems. Technical Report ISP-315, University of Illinois, Urbana, IL, June, 15 1987.

[18] Paul J. Besl. *Surfaces in Range Image Understanding.* Springer Verlag, New York, 1988.

[19] Gerard Medioni and Ramakant Nevatia. Segment-based stereo matching. *Computer Vision Graphics and Image Processing,* CVGIP-31(No. 1):pp. 2–18, 1985.

[20] Stephen White. *100X White Scanner User's Manual.* Technical Arts Corp., Seattle, WA, 1980.

[21] Guna Seetharaman, Kimon Valavani, Magdy Bayoumi, and Michael Mulder. A hybrid range-intensity sensor for dynamic scene analysis and sensor-fusion. Technical Report TR 91-1-12, The Center for Advanced Computer Studies, Univ. of Southwestern Louisiana, May 1991.

[22] Guna Seetharaman, Magdy Bayoumi, Kimon Valavanis, and Michael Mulder. A VLSI architecture for stereo image sensors. In *Proceedings of the Workshop on Computer Architecture for Machine Perception. Paris.,* 1991.