Robotica (1984) volume 2, pp 17–25

Servoed world models as interfaces between robot control systems and sensory data*

Ernest W. Kent and James S. Albus†

National Bureau of Standards, Washington, D.C. 20234 (U.S.A.)

(Received: December 9, 1983)

SUMMARY

A hierarchical robot sensory system being developed for industrial robotics is described. At each level of the hierarchy, sensory interpretative processes are guided by expectancy-generating modeling processes. The modeling processes are driven by *a priori* knowledge (object prototypes), by knowledge of the robot's movements (feedforward from the control system), and by feedback from the interpretative processes (prior state of the sensory world). At the lowest level, the senses (vision, proximity, tactile, force, joint angle, etc.) are handled separately; above this level, they are integrated into a multi-model world model. At successively higher levels, the interpretative and modeling processes describe the world with successively higher order constructs, and over successively longer time periods. Each level of the modeling hierarchy provides output, in parallel, to guide the corresponding levels of a hierarchical robot control system.

INTRODUCTION

It is characteristic of robot applications that most of the system's sensory processing time will be spent on problems of sensory servoing, rather than on object identification. This is because almost all of the images or other sensory data with which it deals are encountered within an historical context. They are members of a sequential set successively altered by object and observer motion. In this respect, the problem domain is very similar to that of animal sensory processing. After objects are initially acquired by the sensory system, its principal job is to provide continuous sensory information to guide the control system as robot and object orientations change. Conversely, the sensory system may obtain information about the next viewing position from the robot control system. As a result of this continuity in the world being sensed, the sensory system can employ many kinds of context-dependent and context-independent knowledge to generate attention processes and expectancies which guide the processing of incoming data, and thus facilitate real-time operation.

In our approach, an internal model of the external world is maintained by continuously servoing this model to the sensory data. The model is a source of predictions about the incoming data, a subset of which can be selected for attention on the basis of optimal potential discriminability. The model is in turn servoed to the data by correcting it on the basis of comparison between the data and its predictions. All sensory information required by the control system is obtained from this internal model, which is always the system's "best guess" about the nature of external reality. The data being obtained from the model by the control system thus may be independent of the particular sensory data to which the sensory interpretative processes are attending as they servo the model to the external world. The control system is a source of feedforward information for the sensory system which causes the model to generate new predictions based on system goals, for example, a change of viewing position.

The world model thus functions as the interface between the sensory and control systems, transforming control actions into attention-generating sensory predictions on the one hand, and transforming sensory data into feedback for the guidance of control actions on the other. A major advantage obtained by interposing a modeling system between the sensory and control systems is that it permits a decoupling of the two systems. The sensory system's attention processes are no longer dictated by the specific data required by the control system, and the control system need not wait on the sensory analysis of recently sampled data. The model can in fact, act as a matched filter which integrates data over time so as to extract signal from noise. Once, the model is correlated with incoming sensory data in the time domain, the model can then be used as a predictor so that the control system can anticipate predictable events, and synchronize with periodicities in the environment.

The National Bureau of Standards' robot system consists of a multi-level, hierarchically ordered, computing structure (see Figure 1). The chain of control levels on the right acts as a task decomposition hierarchy. Each of these control levels is functionally a state machine, with input defined by a task specification from its superior, a status report from its subordinate, and a description of external conditions from the modeling and sensory processing hierarchies on the left. These inputs together with the internal state of the machine, define the address of a line in each control level's state transition table. This line points to a procedure which defines the control level's

* This paper was prepared in conjunction with the official duties of United States Government employees, and is not subject to United States copyright.

† The project described by the present authors represents the joint effort of many people. Among those who have contributed prominently are: A. Barbera, M. Fitzgerald, P. Mansbach, M. Nashman, L. Palombo, M. Shneier, and T. Wheatley.

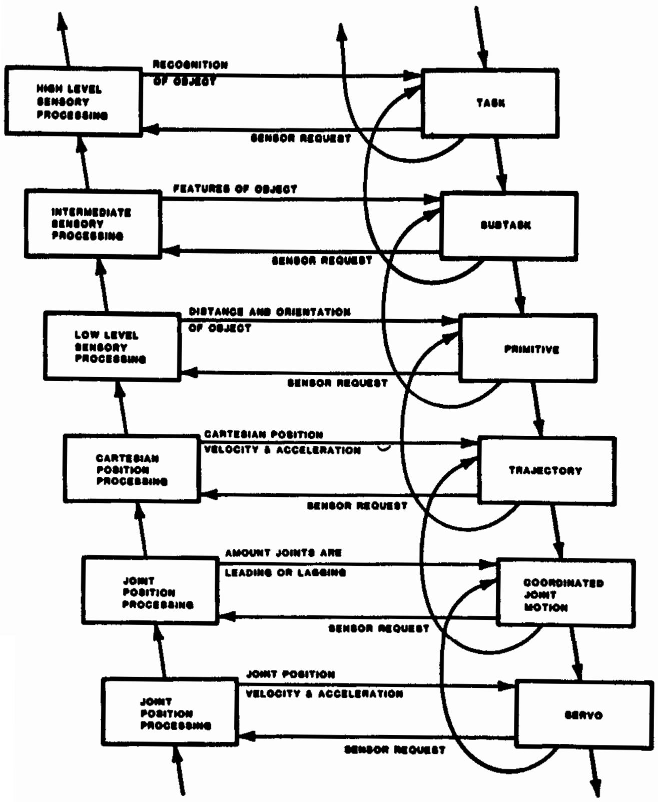
HIGH LEVEL SENSORY PROCESSING
RECOGNITION OF OBJECT
SENSOR REQUEST
TASK
INTERMEDIATE SENSORY PROCESSING
FEATURES OF OBJECT
SENSOR REQUEST
SUBTASK
LOW LEVEL SENSORY PROCESSING
DISTANCE AND ORIENTATION OF OBJECT
SENSOR REQUEST
PRIMITIVE
CARTESIAN POSITION PROCESSING
CARTESIAN POSITION VELOCITY & ACCELERATION
SENSOR REQUEST
TRAJECTORY
JOINT POSITION PROCESSING
AMOUNT JOINTS ARE LEADING OR LAGGING
SENSOR REQUEST
COORDINATED JOINT MOTION
JOINT POSITION PROCESSING
JOINT POSITION VELOCITY & ACCELERATION
SENSOR REQUEST
SERVO

Fig. 1. The overall architectural scheme of the NBS robot system. On the right, a control hierarchy is constructed from a chain of state machines performing task decomposition. At each level, the task input from above, the status feedback from below, and the state of the world from the left determine the current output from a state transition table. On the right, an ascending sensory hierarchy accepts interoceptive and exteroceptive sensory data and processes it to provide appropriate world-state information for each level of the control hierarchy.

output and selects its state for the next clock cycle. Typically, levels near the top of the hierarchy will change state slowly and each action selected will represent a step in a long term plan, while those levels near the bottom will change state rapidly and each action represents the next step in a task sequence, or a simple motion primitive such as next joint position.

The sensory hierarchy on the left in Figure 1 accepts sensor data at various levels. At the bottom level this consists only of immediate interoceptive feedback from force sensors, tachometers, or joint angle sensors. At higher levels, a variety of exteroceptive data are accepted, such as television frames, touch and proximity sensors. As data enters into and ascends through this sensory hierarchy, its progressive analysis makes information available to the control system at levels of descriptive complexity appropriate to the actions of the control levels receiving it. Since, in general, higher order descriptions of the sensory world require more processing, there is a general increase in the time taken for successive levels of the ascending sensory hierarchy to produce updated models of the world. The implementation of independent functional levels is selected to achieve approximate equality between the processing times for sensory analysis and the state changes in the corresponding levels of the control hierarchy. Figure 2 presents a more detailed view, in which the sensory side of the hierarchy is broken into its separate sensory-interpretative (G) and sensory-modelling (M) systems. The labeled arrows indicate the kinds of information which flow laterally within levels of the system, and their relation to ascending and descending information. Status feedback in the control (H) hierarchy is omitted here

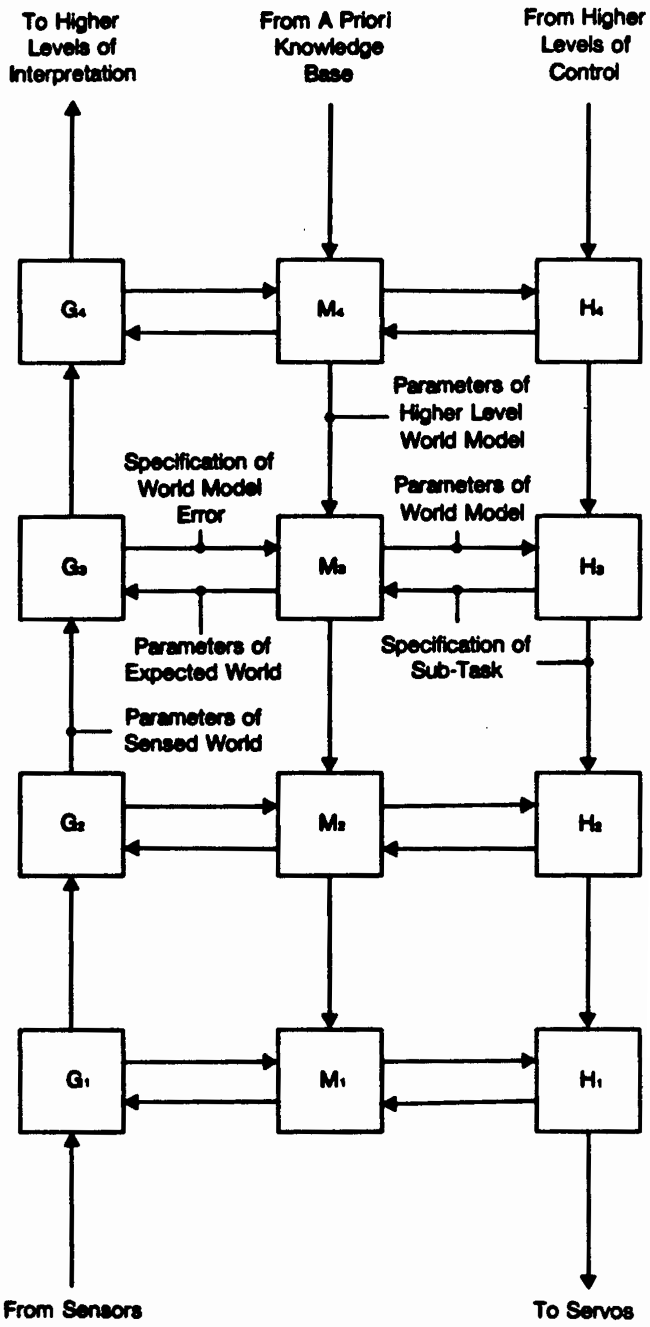
To Higher Levels of Interpretation
From A Priori Knowledge Base
From Higher Levels of Control
G₄
M₄
H₄
Parameters of Higher Level World Model
Specification of World Model Error
Parameters of World Model
G₃
M₃
H₃
Parameters of Expected World
Specification of Sub-Task
Parameters of Sensed World
G₂
M₂
H₂
G₁
M₁
H₁
From Sensors
To Servos

Fig. 2. In this diagram, the sensory hierarchy is shown decomposed into its interpretative (G) and modeling (M) components. The control hierarchy (H) communicates only with the modeling hierarchy. The communications occurring at each level are of the same type, and are indicated for a single level. (The status feedback loops are omitted from the control side.)

for clarity. In the present paper, we focus on the activities of the sensory portion of the system. The details of the control hierarchy have been presented elsewhere.[1]

As is apparent in these figures, each level of the sensory system presents its output to the robot control system in addition to passing information up to the next sensory processing level. These communications between the sensory and control systems occur in parallel at each level. This makes the lower level, more rapidly processed, information in the data available to the control system as rapidly as it is discovered. At higher levels, the sensory system describes all data entering the system in terms of constructs such as shape, extent, and orientation, which are needed by the control system for physical interaction with either known or unknown objects. At the highest levels, the task planning portions of the control system receive descriptions of objects and, when possible, this may include classifying objects into sets which are known by name. When the objects prove nameable both the control system and the sensory system can use stored information about characteristics common to all members of their set.

Descriptions of object relations, objects, and all of their parts are thus independently available, with the lower levels of description represented by the most current information. Since the simpler relations will usually be updated more frequently than the more complex ones in this multi-level system, a hierarchical control system can use these lower level data for rapid servoing at the lower levels of control. This is possible once their significance has been understood by the more slowly updated higher levels of the system. For example, at the lowest level of visual description the output is essentially only range and azimuth information about points. This information can easily be computed at frame rates, particularly once higher level information about object identity and geometric coordinate frames have been determined.

In the simplest mode of interaction between the sensory and control systems, the control system may request particular pieces of information at any level as it requires them. However, the internal attention functions of the sensory system include the ability to select specified windows and filter types based on information from the model. In this query-driven attention mode processing can be restricted to the items of interest over the lifetime of some phase of a control task, with a corresponding reduction in processing time. In general, information may either be requested about particular entities or types of entities, or it may be requested about the contents of particular space, time or frequency windows.

HIERARCHICAL WORLD MODELING

The principal internal activity of the robot sensory system is to build, modify, and maintain a world model embodying the most complete possible description of the environment. This world model is built with reference to all available sensory input, including a variety of senses for external data, such as vision, touch, and proximity, and senses for internal data, such as joint angle and force. A fund of internally stored knowledge is also employed in construction of the world model. This consists of general knowledge about the external environment (such as ideal, or characteristic, descriptions of classes of known objects) contained in a "knowledge base", and of particular knowledge, such as the names and locations of objects which are expected in the current context. A further category of knowledge is derived from knowledge of the state of the system, such as knowledge about the momentary context of the observation itself. It includes knowledge about the location and motion of the observer (or sensor) in space, as well as knowledge about such variables as conditions of illumination. This latter information is obtained from that part of the robot system which controls the sources of illumination.

The world model is constructed at many levels of description. These are roughly hierarchical in sequential order of construction, time of completion, and complexity of the elements described. The system continually uses all its computing resources by processing as many objects as possible through all levels of description. Where this is not possible, the priorities are set by the requirements of the task.

The general nature of this descriptive hierarchy is portrayed in Figure 3. The lowest level of description

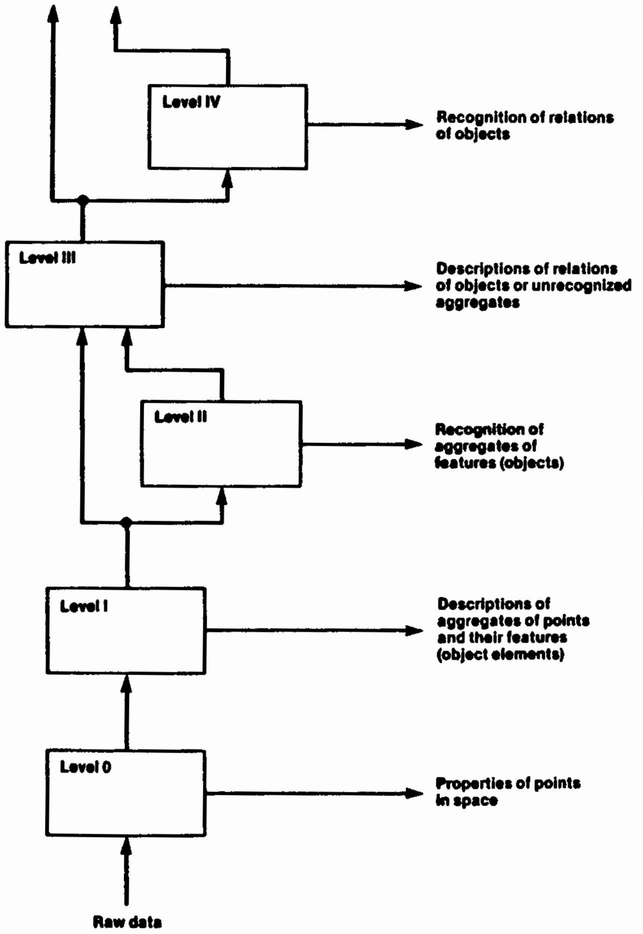
Level IV
Recognition of relations of objects
Level III
Descriptions of relations of objects or unrecognized aggregates
Level II
Recognition of aggregates of features (objects)
Level I
Descriptions of aggregates of points and their features (object elements)
Level 0
Properties of points in space
Raw data

Fig. 3. The levels of the sensory hierarchy, illustrating paths of information flow for cases where objects are or are not recognizable.

(Level 0) represents the world in terms of properties of individual units of data which the system can sense or derive from each sensory modality. For example, from the visual modality, intensity, color, range, and other features of individual pixels would be available. No attempt is made at this level to correlate data from different sensory modalities.

The next level of description (Level I), represents the data in terms of properties belonging to connected aggregates of data points. This is possible for most sensory information. Thus, at this level connected groups are identified, and their useful spatial and temporal features, such as corners and velocities, are catalogued. At this level, correlations are sought between data from all modalities, and the entities represented are, at least potentially, multi-modal.

These two levels of description are possible using only state-independent knowledge. However, state-dependent knowledge may also be invoked where available. For example, at level I, positions and orientations of edges, lines, or corner features may be found and described without invoking *a priori* knowledge about the current contents of the scene. But, if *a priori* knowledge about the kinds and positions of objects in the scene were available from the model, it might either guide the system to probable corner locations, or provide the system with a set of matched filters to determine if a corner was present, and what type it was.

The next level of description (Level II) attempts to aggregate clusters of features into more global entities. That is, particular relations between clusters of level I features may be recognized as objects which can be named by Level II. When this can be done, it permits access to knowledge in the model about the known objects and their parts. This knowledge can then loop back to improve the representation of the object and to resolve uncertainties in its description. It permits as yet unseen features of the object to be inferred. Naming the object also permits more concise description of items of information which vary together with changes in orientation of the object.

Achieving this level of description requires *a priori* state-independent knowledge, in the form of descriptions of object prototypes which the system "knows". Additionally, when available, state-dependent information about expected object types, orientations, and locations may be very usefully employed. Such information may come from earlier recognition operations performed by the system itself, or it may come from external sources such as an automated materials handling data base.

Level III (the fourth level of description in the world model) attempts to describe spatial and temporal relations of aggregates of the objects described at level II and/or level I. For example, the locations of object A and object B might be described at level II, but level III would contain the fact that the relationship "A is on B" exists. Temporal relations at this level might include relations such as "A approaching B". This level thus recognizes and describes relationships which are properties of the scene, while previous levels describe properties of objects in the scene.

This level of description is not limited to objects which can be named by Level II. It can also describe relations of objects or object elements described at Level I, even when they cannot be named. If objects have been successfully named, potent state-dependent knowledge may be available to assist interpretation at the Level III, and naming at Level II may also indicate that apparently separate entities are actually parts of the same recognizable object. Nonetheless, spatial and temporal relations may also be described between unclassified entities. This allows Level III to represent the spatial and temporal relationships between unidentified objects and other aspects of the scene.

This ability to describe scene contents at any level, irrespective of the ability to name things, is fundamental to a sensory system intended for sensory-servoed robot guidance. An understanding of the spatial and temporal structure of the environment is basic to the ability to physically act in it. It is only when the physical structure of the environment is understood by the system that it can act on the information gained by recognition of objects, and many operations may be required on unrecognized objects. The robot must at least avoid collision with unrecognized objects and maneuver around them. Additionally, it will usually need to inspect unknown objects in an attempt to identify them, and this may involve actually manipulating them. Unrecognizable objects may also require removal from the workspace or relocation within it. In general, the further we go up or down scale from the object level of organization, the less utility naming appears to have. In contrast, a model of the spatial and temporal structure of the environment is necessary to interact with it at all levels.

The final level shown in Figure 3, level IV, simply indicates the possibility for indefinite extension of this type of hierarchical structure. In this case, the level would again attempt to name aggregates of relations identified at the preceding level. Since these would be spatial or temporal relations of objects (A is on B, etc.) this level would be attempting to name collections of such relations, e.g. "Subassembly 23".

SERVOING MODEL TO DATA

Each level of the descriptive hierarchy except the lowest is composed of two major processes, an interpretative process and a modeling process. The interpretative processes attempt to classify and group data ascending from lower levels. The modeling processes generate testable expectations about the probable relations among the data received by the interpretative processes. Figure 4 shows the organization of these two major processes for two levels of the hierarchy. Within each level there exists a continuous interaction between the interpretative process, operating to discover properties of the data arriving from lower levels, and the modeling process operating to generate predictions of what those properties are expected to be. The interpretation of the scene at each level of description is based on the data presented from lower levels, on general knowledge about rules for interpreting the sensory world, and on the expectations generated by the modeling process. The expectation is, in

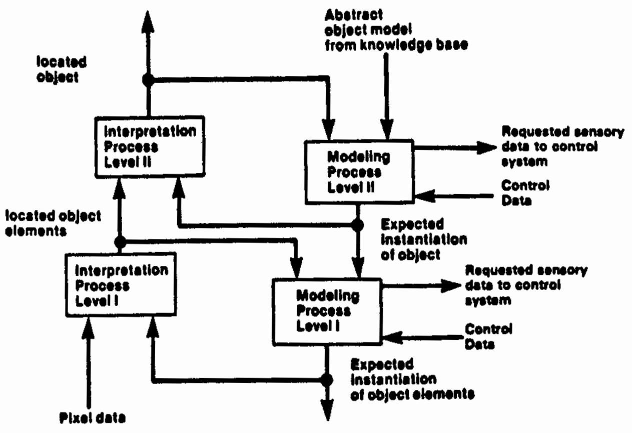
Abstract object model from knowledge base
located object
Interpretation Process Level II
Modeling Process Level II
Requested sensory data to control system
Control Data
located object elements
Expected instantiation of object
Interpretation Process Level I
Modeling Process Level I
Requested sensory data to control system
Control Data
Expected instantiation of object elements
Pixel data

Fig. 4. Two levels of the sensory hierarchy, showing the relations between the interpretative and modeling components. At each level, the current world model is servoed to the data being processed by the interpretative processes. The control system may request data from any aspect of the model, while the modeling process generates predictions based on current control actions.

turn, based upon knowledge of what the interpretative process has previously discovered or identified, upon *a priori* knowledge about the scene available from higher levels of modeling, and upon knowledge of changes in viewing position which have been commanded by the control system.

The interpretative processes form an ascending hierarchy beginning at the bottom with input from the sensors. The modeling processes form a descending hierarchy beginning at the top level with input from the "Knowledge Base", which contains the system's store of *a priori* knowledge about object types and expected scene contents. At each level of the descending modeling hierarchy, this information is transformed by application of other information to generate an expectation about the current appearance of the scene and objects in it. This transformation is guided by information about current decisions of the interpretative process at that level, and by input from the control system describing current actions of the robot which influence receptor orientation and operation. Successively lower levels of this modeling hierarchy operate on successively more detailed aspects of the scene.

At each level, the modeling process may also accept information from the interpretative process about discovered entities which are not contained in the world model received from higher levels. In this case, whatever description is possible for the new item is added to the world model at that level. The new item may then enter into the expectancy generating process insofar as changes in point of view may be expected to alter its apparent position, but not in other ways.

At every level, the interaction between the interpretative and modeling processes attempts to reconcile observation with expectation. This is accomplished by servoing the model to a best fit with the data, and by using the model to improve the description of the data. An example is shown in Figure 5. Here, the modeling hierarchy predicts a corner feature at a particular location, orientation and apex angle, and the interpretation algorithms compare this prediction with the edge data points arriving from the level below. If the fit is statistically good, the predicted feature is relocated to the best fit to the data. This ideal feature, relocated to the best fit position, is what is passed to the next level's interpretative process as data. The location discrepancy is fed back to the modeling process, which uses it to correct its predictions. At the next level up, predictions about entire objects are compared with descriptions of located features arriving from below. A best fit of the ideal object to the collection of features is performed, and a particular instantiation of the ideal object is passed as data to higher interpretative levels. The discrepancy between prediction and observation is again passed back to the modeling process at this level.

Within the modeling process, the position discrepancy fedback from the interpretative process is used to im-

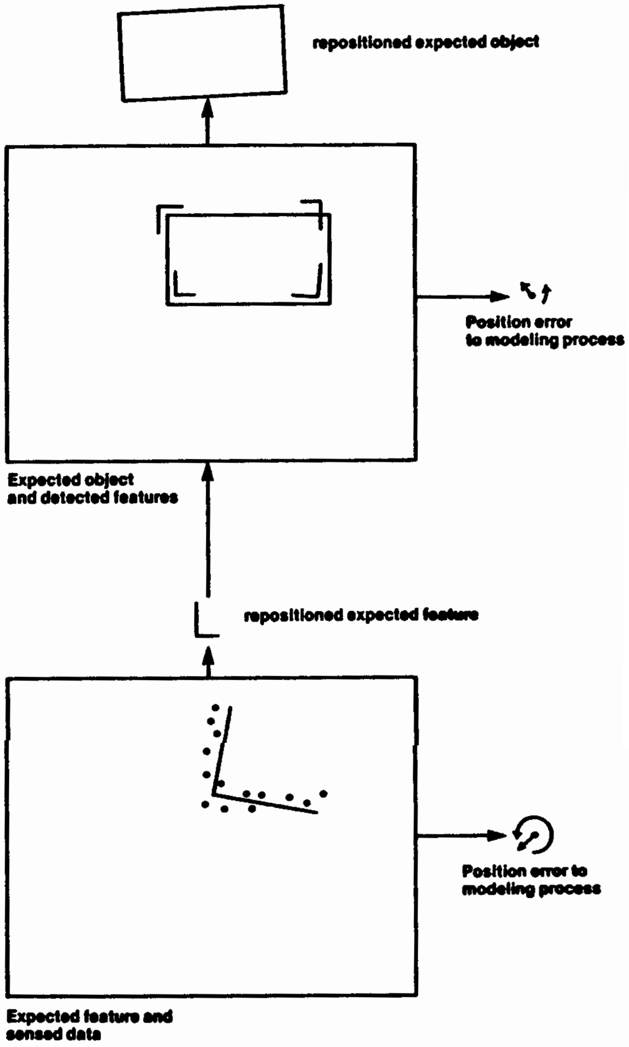
repositioned expected object
Position error to modeling process
Expected object and detected features
repositioned expected feature
Position error to modeling process
Expected feature and sensed data

Fig. 5. The interaction between data discovered by the system and predictions of the modeling process. At the lower level, a predicted corner feature is compared with observed edge points, the best-fit instantiation of the predicted corner is passed on to higher levels, and the error is passed back to the modeling process. At the upper level, a similar process takes place to fit a predicted rectangle to the discovered corner features.

prove the model of the object's position. This may cause the modeling algorithms to change the predicted description of the object as well as its predicted location. This might occur, for example, because new features become visible, or apparent angles change. On the next iteration, this may improve the fit between observation and expectation and in turn cause a refinement of the observed discrepancy in fit and location. Note that the interpreting process may change the position of the predicted ideal to accord with observation, but does not change its description. The modeling process on the other hand may change the description of the expected appearance of the object.

Assuming that the prediction provided by the modeling process has fit the observed data within prescribed limits of error, the interpretative process passes forward the repositioned ideal description of the expected element rather than the actual data. This follows from the assumption that (if the prediction is correct) the observed data will never be as good a description of the object as the *a priori* model. The function of the data in this case is only to form a basis for decision about the probability that the object seen is the object expected, and to generate statistics for reducing errors in the current orientation and position parameters of the model.

Another possible mode occurs when the interpretative algorithms discover features which do not correspond sufficiently to features predicted by the modeling process. This may occur, for example, when an object is an aberrant example of its type, or when the data are noisy. It may also occur when a feature is discovered where none was predicted; for example an unknown object. In this case a description of the feature, as a type with parameters derived directly from the data, is passed to superior levels and to the modeling process. When this occurs, certain orders of representation are denied to the object (for example, it may not be possible to name it), and the accuracy of its representation may be inferior. However, it can still enter into the expectancy generating process and be incorporated into the world model, thus serving as a basis for action by the control system.

The levels of the system are loosely coupled. The system as a whole is data-driven, and no particular synchronization of levels occurs, except at the lowest levels where there may be dependencies on receptor hardware timing. Time required to process a scene at each level may fluctuate widely and independently, due to a variety of factors. Thus, it may occur that a level will not find new data ready to use in servoing the current iteration of its model, while at the same time new expectancies are being generated by events such as movement of the robot. In this case, the expectancy of the current state of the model is fed forward unmodified as the "best guess" about the nature of the world. This is analogous to continuing to walk when your eyes are momentarily closed, based on your predictions about what you *would* see if they were open, and it serves essentially the same purpose in a robot system. When new data become available, any drift due to this feed-forward is corrected. The world model is continuously being corrected by the data, so that it always represents the current best guess. If the model includes temporal features such as velocities and accelerations, the dead-reckoning can remain accurate for longer periods.

Of course even with feed-forward capabilities provided by the model, the existence of computational delays in the sensory feedback loop raises the possibility of instability in the control system. This can be a significant problem at the lower levels where the robot may actually be tracking a target using visual, tactile, or force feedback. In this case classical methods of stability analysis must be applied. At the higher levels, where elements of a task sequence are being selected based on recognition of features and objects, the control system may simply wait in a "PAUSE" state until sufficient information is processed to allow it to proceed to the next step in its task sequence.

DATA STRUCTURES AND THE MODELING PROCESS

The "Knowledge Base" in the upper right of Figure 6 represents the robot's store of *a priori* knowledge, including the descriptions of known objects, the set of these which is initially expected, and their expected initial positions. Some of this is state-dependent information which originates in other parts of an automated factory; for example, from another robot loading a parts tray. To the extent that such initial expectations are incomplete or

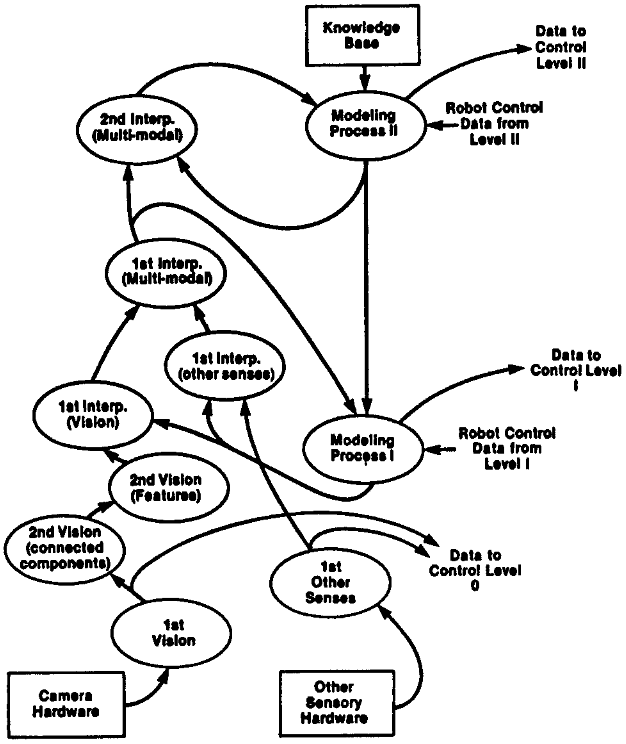

Fig. 6. The current version of the sensory system being developed at NBS to implement the general approach outlined in this paper. In this version, binary vision is employed together with a minimal set of other senses. After preprocessing, the data are carried through two levels of modeling-interpretative interaction.

inaccurate, the system must perform greater or lesser degrees of initial acquisition processing upon first seeing the parts.

The ideal prototypes of known objects, however, may be derived directly from CAD descriptions. These CAD descriptions are used to create three dimensional models of the objects in two very different descriptive formats. The first is a standard linked list representation of component parts such as edges and vertices, together with metric information concerning them. This representation is augmented by, and linked to, a representation of the object as a graph of "generic views", inspired by the work of Koenderink and van Doorn on aspect graphs.[2] An aspect graph is a representation of the relations between all of the regions ("parcels") of the viewing space within which the visible features of the object have constant topological relations. That is, the perspective projection may change as the observer moves within the region or "parcel" of space, but not the occlusion relations. The topologically constant view from such a region is an aspect, or "generic view". These generic viewing regions are derived by a process which permits the natural geometry of the object to define this parcellation of the viewing space.

Since the occlusion relations of the object's surface are invariant within any such parcel of space, it is possible to obtain a pre-solution of hidden surface problems for any given generic view. Links between points in the generic view and metric information in the remainder of the data structure permit calculation of changes in perspective as the camera moves within the parcel. Because motions of the camera within the parcel cannot generate discontinuities in these calculations, these changes can be conveniently approximated by simple parametric equations associated with the space parcel of each generic view. The graph structure having the generic views as nodes (the aspect graph) represents the topological relations among the generic views themselves, and permits the modeling process to anticipate transitions from one view to another as the camera moves through space.

These two types of representation are chosen to permit rapid construction of expected instantiations of each type of known object from the current viewing position. Once this is complete, the most useful features of the expected view may be identified. At each level, the modeling process applies algorithms which select for attention those features expected to be most discriminable by the sensors, while providing the best discrimination of the position of the expected objects. Thus, the "attention" of the interpretative processes may be directed first to those items most likely to disconfirm the expectancy. If these attended features are confirmed, the expectancy can be accepted without further processing. (In addition, the control or goal-directing systems may request attention to particular classes of objects or features, or to particular spatial and temporal windows of occurrence.)

The data collected by the sensory system is represented in a world model database designed to satisfy several requirements. It must represent all known information about the objects' locations, the volumes they occupy, and the uncertainties in their positions and sizes. It must be able to integrate information obtained by sensing the world with that obtained from CAD models and expectations. It must enable questions to be answered about free space as well as space that is occupied, and it must uniformly represent expected and discovered, known and unknown objects. We have chosen a dual representation consisting of an object model and a volume model. The object model consists of a set of tables describing the properties (including exact position and orientation) of every object in the workspace known to the system. These tables may be derived by instantiation of CAD models, and thus be quite complete and contain names. They may also be descriptions of features, positions, and other discovered information concerning as yet unnamed objects; in either case the format of representation is the same. The volume model is an octree representation of the workspace. The nodes of the octree indicate whether the volume that each represents is occupied, empty, or unexamined, and, by means of pointers to the tables, by what it is occupied.

This dual representation allows retrieval of information in either spatially indexed or feature indexed formats, and both may required in typical robotics applications. The octree structure permits variable resolution descriptions of the work space, so that large unoccupied volumes are compactly represented, and questions about the workspace can be answered by searching only to the level of resolution desired. Where very detailed information on locations of object features is required, the object description tables may be consulted, either initially or after reference through the volume model. Either of these representations may be filled in through the use of any sensory modality. Even within a single modality such as vision there may be different "channels" which require different types of representation. For example, in our present structured light visual system[3] there is a flood type of illumination which gives two dimensional information, but which covers the entire visual field, and a plane of light illumination which gives three dimensional position information about objects intersecting a pair of planes projected into the visual space.

The plane of light illumination is analyzed to yield information about object locations and orientations which is entered into the object and feature indexed tables. The flood illumination is used to build the spatially indexed volume model. This latter activity is greatly simplified by the fact that the camera rides on the robot's hand, and thus its position and orientation are always known absolutely. When a 2-D flood image is obtained, it is projected into the volume model to form a labeling of the octree which represents a generalized cone in the workspace. This indicates all the locations which might be occupied on the basis of that picture. As the robot moves, similar cones are generated by other views of the same object. The intersections of these cones are retained as "occupied" nodes in the volume model, while nodes that do not intersect are erased. In this fashion the system successively carves out a region of space containing the object. If needed, this representation may be refined by investigation on the part of the robot (i.e., it may move in for better resolution or move to look at an object from the side.) If and when it becomes possible to identify the object, the volume representation may be further refined by intersection with the instantiated CAD model. Whether or not the object can be identified, however, the volume model is a guide to physical interaction with objects and the space containing them.

CURRENT IMPLEMENTATION

This general design and rationale forms the basis for a series of systems currently being constructed at the National Bureau of Standards. A variety of types of vision and other senses potentially can be employed by this scheme. In our first version of the robot sensory system, we are employing a two frame structured light vision system,[3] and tactile, force/torque, and infra-red proximity transducers as "other" sensors. Later versions will employ advanced gray-scale image pipelines, and range-image systems currently under construction.

Algorithms for automatic derivation of generic views from CAD databases are being developed, but this is not yet possible for arbitrary objects. In the present stage of development, generic views restricted to perimeter features and constructed with operator interaction will be employed. At the same time, alternative algorithms for generating expected views using more traditional in-line hidden surface algorithms are also being used.

The actual functional levels and hardware of the first development stage appear in Figure 6. Beginning in the lower left, the Camera Hardware includes controls for the structured illumination and the interface and hardware pre-processing for the television camera. This hardware acquires images, creates binary images from them by thresholding, and converts the binary images to lists of run lengths between binary transitions. The two types of structured light used are a point source, which gives a two-dimensional outline of objects when thresholded, and a dual plane-of-light source which can be analysed to give both the range to a surface, and its orientation relative to the camera. The system can thus infer three dimensional interpretations of the two dimensional images acquired from the point (flood) flash frame. These two structured light types are used in alternate television frames. Between the two, the early stages of vision can deduce all six degrees of freedom (relative to the robot) of any surface which the double plane illumination strikes.

This data is transmitted to First Stage Vision, which cleans and selects the run length data, computes the visual information to be output to Control Level 0, and transmits all of the run-length data to the "Connected Components" section of Second Stage Vision. This stage finds connected components or "blobs" in the image and transmits a description of them, in chain code, to the next component of Second Stage Vision, which identifies boundary features of the blobs. Second Stage Vision also controls a number of operating parameters of the vision hardware, such as exposure values.

On the bottom right, Other Sensor Hardware represents all of the electronics associated with the proximity, tactile, and force/torque sensors. This information is received and processed by First Stage Other Senses, which computes information to be output to Control Level 0, and passes processed data to the next level. Owing to the current rudimentary state of the other senses in the system, there is no secondary pre-processing of this data as there is with vision.

Following these pre-processing paths, there are two complete levels of the sort described earlier, and diagrammed in Figures 3 and 4. These each consist of an interpretative process and a world modeling process. The first such level is unique in that the data from the various senses have not yet been correlated with one another. Thus, the interpretative process is split into three stages. The first two attempt, in parallel, to reconcile their data with the world modeling process' expectations for the various senses separately. This having been accomplished, a third part of the process attempts to correlate features described by the previous two into a unified multi-modal description of the data. It transmits this to the next level, and transmits discrepancies between actuality and prediction back to the world modeling process. Three dimensional information may or may not exist for individual features, but all expectancies are generated and compared as two dimensional projections.

At the next level, a single multi-modal interpretation process attempts to reconcile the located ideal features described by the first level with the expected features of known objects generated by the second level world modeling process. At this level, and at subsequent levels yet to be implemented, the modality or modalities of origin of the data are not represented in the model. The system hardware currently implemented consists of a hierarchy of parallel microprocessors. This permits us to take advantage of the parallelism and pipeline organization inherent in the system design. The essential autonomy of the various levels allows the component processors to run with little overhead other than handshaking to pass data. Save for the three sub-processes in the Level I interpretative process, each of the elements of Figure 6 resides in a functionally separate microprocessor-based computer. Currently, these are 8086/8087 pairs. Communication between these elements takes place over dedicated point-to-point data paths which are handled as buffered DMA transfers. A low bandwidth broadcast bus ties all the processors together for occasional system messages and initialization.

In the present form of the system, the timing of the sensory processing hierarchy ranges from frame-rate (30 msec.) data at the level of primitive features such as structured light range, to about 1 second for simple object recognition operations. Except for the lowest level, however, these times are data dependant, and will vary according to the complexity of the scene. In the present version of the system, this sensory hierarchy is mated to a hierarchical control system which can accept new data at all levels every 15 milliseconds.

References

1. J.S. Albus, A.J. Barbera and R.N. Nagel, Theory and Practice of Hierarchical Control. **In**: 23rd IEEE Computer Soc. International Conference (September, 1981) pp. 18–39.
2. J.J. Koenderink and A.J. van Doorn, The Internal Representation of Solid Shape with Respect to Vision. *Biol. Cybernetics* **32,** 211–216 (1979).
3. J.S. Albus, E.W. Kent, M. Nashman, P. Mansbach, L. Palombo and M. Shneier, A 6-D Vision System. *Proc. SPIE Technical Symposium*, Crystal City, VA. (May 1982).