Hierarchical Interaction Between Sensory Processing and World Modeling in Intelligent Systems

James S. Albus

Robot Systems Division
National Institute of Standards and Technology

Abstract

The role of the sensory processing system is to transform sensory maps into world model maps and to extract entity attributes and states. The role of the world modeling system is to store best estimates of the extracted values, and to generate from them, predicted inputs. There is a tight coupling between sensory processing and world modeling. The world model generates predicted sensory input from stored state estimates. The sensory processing system compares predictions with observations, and returns the difference signals to the world model for updating state estimators. The sensory processing system also performs spatial and temporal integration on various combinations of observed and predicted data, and compares the resulting correlation function against thresholds of recognition and detection.

Introduction

Sensory processing is the mechanism of perception. Perception is the establishment and maintenance of correspondence between the internal world model and the external real world. The function of sensory processing is to extract information about surfaces, entities, events, states, and relationships in the external world, so as keep the world model accurate and up to date.

The intelligent control system architecture described in [1] consists of a hierarchy of task decomposition, world modeling, sensory processing, and value judgment modules. In [2], the world model is defined in terms of maps, entities, and states. This paper describes the sensory processing system and its relationship to the world model.

Sensory Processing SP Modules

At each level of the architecture proposed in [1], there exist a number of sensory processing (SP) modules. Each SP module consists of four sublevels, as shown in Figure 1.

Sublevel 1 -- comparison

Each comparison submodule matches an observed sensory variable with a world model prediction of that variable. This comparison typically involves an arithmetic operation, such as multiplication or subtraction, which yields a measure of similarity and difference between an observed variable and a predicted variable. Similarities indicate the degree to which the WM predictions are correct, and hence are a measure of the correspondence between the world model and reality. Differences indicate a lack of correspondence between world model predictions and sensory observations. Differences imply that either the sensor data or world model is incorrect.

Output from the comparison submodules goes two places. First, it is returned directly to the WM module to update the world model prediction. Second, it is transmitted upward to sublevels 2 and 3 for temporal and spatial integration.

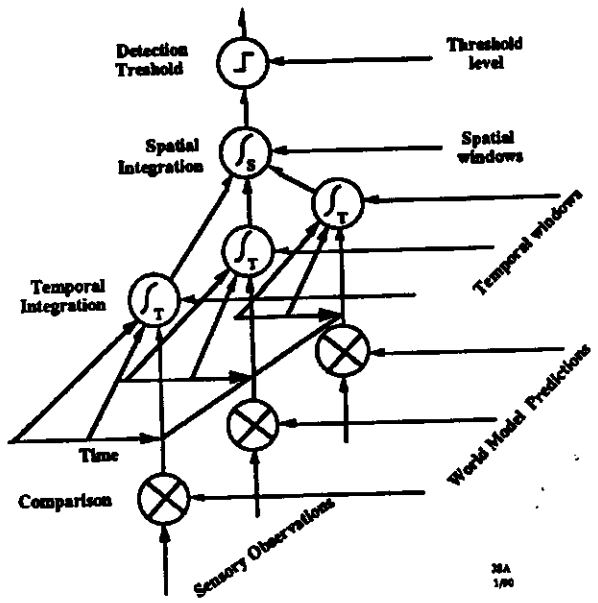

Figure 1. Each sensory processing SP module consists of: 1) a set of comparators that compare sensory observations with world model predictions, 2) a set of temporal integrators that integrate similarities and differences, 3) a set of spatial integrators that fuse information from different sensory data streams, and 4) a set of threshold detectors that recognize entities and detect events.

Sublevel 2 -- temporal integration

Temporal integrator submodules operating on sensory data alone may produce a summary, such as a total, or average, of sensory information over a given time window. Temporal integrator submodules operating on the similarity and difference values computed by comparison submodules may produce temporal cross-correlation and error functions between the model and the observed data. These correlation and error functions are measures of how well the dynamic properties of the world model entity match those of the real world entity. The boundries of the temporal integration window may be derived from WM representations of task decomposition parameters such as sensor fixation periods. Temporal integration produces a temporal "chunking", or clustering, of information.

Proceedings of the Fifth IEEE International Symposium on Intelligent Control, Philadelphia, PA, September 5-7, 1990

Sublevel 3 -- spatial integration

Spatial integrator submodules integrate similarities and differences between predictions and observations over regions of space. This produces spatial cross-correlation or convolution functions between the model and the observed data. Spatial integration summarizes sensory information from multiple sources at a single point in time. It determines whether the geometric properties of a world model entity match those of a real world entity. For example, the product of an edge operator and an input image may be summed (i.e. integrated) over the area of the operator to obtain the correlation between the image and the edge operator at a point. The limits of the spatial integration window may be determined by world model predictions of entity size. Spatial integration produces a spatial "chunking", or clustering, of information.

In some cases, the order of temporal and spatial integration may be reversed, or interleaved. Both temporal and spatial integration of correlations and differences may be used by the value judgment system to compute confidence and believability factors.

Sublevel 4 -- recognition/detection threshold

When the spatio-temporal correlation function exceeds some threshold, object recognition (or event detection) occurs. For example, if the spatio-temporal summation over the area of an edge operator exceeds threshold, an edge is said to be detected at the center of the area.

Figure 2 illustrates the nature of the SP-WM interactions between an intelligent vision system and the world model at one level. On the left of Figure 2, the world of reality is viewed through the window of an egosphere such as exists in the primary visual cortex. On the right is a world model consisting of: 1) a symbolic entity frame in which entity attributes are stored, and 2) an iconic predicted image that is registered in real-time with the sensory image. The predicted image is initialized by a graphics engine operating on the symbolic entity frame, and updated by observed differences between itself and the observed sensory input. In the center of Figure 2 is a comparator where the expected image is subtracted from (or otherwise compared with) the observed image.

Difference images from the comparator go two places:

1) They are returned directly to the WM for real-time local pixel attribute updates. This produces a tight feedback loop whereby the world model predicted image becomes an array of Kalman filter state estimators. Difference images are thus error signals by which each pixel of the predicted image can be "servoed" into correspondence with current sensory input.

2) They are also transmitted upward to the integration sublevels where they are integrated over time and space in order to recognize and detect global entity attributes. This integration constitutes a "chunking" of sensory data into entities. At each level, lower order entities are "chunked" into higher order entities, i.e. points into lines, lines into surfaces, surfaces into objects, objects into groups, etc.

Recognized entities output from level(i) are entered into the entity database at level(i+1) of the world model. This closes a slower, more global, servo loop between WM and SP modules through the symbolic entity frames. Many examples of this type of slow looping interaction can be found in the model matching and model based recognition literature for static images [3]. Application of tight closed loop concepts to dynamic systems modeling has been used in aircraft flight control systems for years, and has recently been demonstrated in a control system for high speed visual guidance of autonomous road vehicles [4].

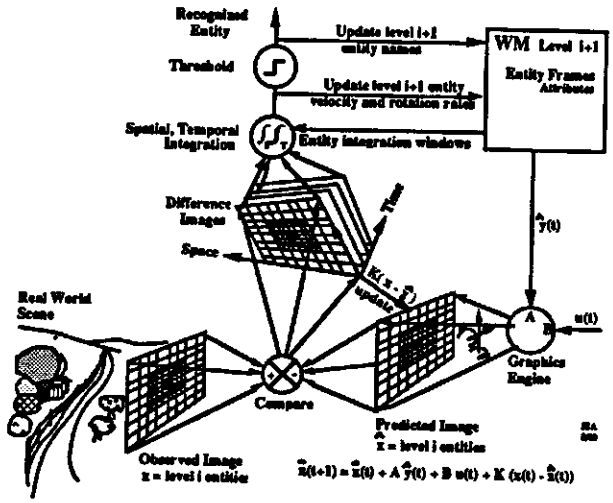

Figure 2. Interaction between world model and sensory processing. Difference images are generated by comparing predicted images with observed images. Feedback of difference images produces a Kalman filter best estimate for each data variable in the world model. Spatial and temporal integration produce cross-correlation functions between the estimated attributes in the world model and the real world attributes measured in the observed image. When the correlation exceeds threshold, entity recognition occurs.

In biological or neural network implementations, SP modules may contain thousands, even millions, of comparison submodules, temporal and spatial integrators, and threshold submodules. Maps with many different overlays, as well as entire lists of symbolic attributes, can be processed in parallel in real-time. Multiple hypotheses may be compared with the sensory data stream at multiple hierarchical levels, all simultaneously.

World Model Update

Attributes in the world model predicted image may be updated by a formula of the form

$$\hat{\mathbf{x}}(t+1) = \hat{\mathbf{x}}(t) + \mathbf{A}\,\hat{\mathbf{y}}(t) + \mathbf{B}\,\mathbf{u}(t) + \mathbf{K}(t)\,[\mathbf{x}(t) - \hat{\mathbf{x}}(t)]$$

where $\hat{\mathbf{x}}(t)$ is the best estimate vector of world model i-order entity attributes at time t

A is a matrix that computes the expected rate of change of $\hat{\mathbf{x}}(t)$ given the current best estimate of the i+1 order entity attribute vector $\hat{\mathbf{y}}(t)$

B is a matrix that computes the expected rate of change of $\hat{\mathbf{x}}(t)$ due to external input $\mathbf{u}(t)$

$K(t)$ is a confidence factor vector for updating $\hat{x}(t)$

The value of $K(t)$ may be computed by a formula of the form

$$K(t) = K_s(j,t)\,[1 - K_m(j,t)]$$

where $K_s(j,t)$ is the confidence in the sensory observation of the j-th real world attribute $x(j,t)$ at time t

$K_m(j,t)$ is the confidence in the world model prediction of the j-th attribute $\hat{x}(j,t)$ at time t

The confidence factors (K_m and K_s) in the above formulae may depend on the statistics of the correspondence between the world model entity and the real world entity (e.g. the number of data samples, the mean and standard deviation of $[x(t) - \hat{x}(t)]$, etc.). A high degree of correlation between $x(t)$ and $\hat{x}(t)$ in both temporal and spatial domains indicates that entities or events have been correctly recognized, and states and attributes of entities and events in the world model correspond to those in the real world environment. World model data elements that match observed sensory data elements are reinforced by increasing the confidence, or believability factor, $K_m(j,t)$ for the entity or state at location j in the world model attribute lists. World model entities and states that fail to match sensory observations have their confidence factors $K_m(j,t)$ reduced. The confidence factor $K_s(j,t)$ may be derived from the signal-to-noise ratio of the j-th sensory data stream.

The numerical value of the confidence factors may be computed by a variety of statistical methods such Baysian or Dempster-Shafer statistics, or Kalman filter methods.

Measurement of Surfaces

World model maps are updated by sensory measurement of surfaces. Such information is usually derived from vision or touch sensors, although some intelligent systems may derive surface information from sonar, radar, or laser sensors.

The most direct method of measuring surfaces is through touch. Many creatures, from insects to mammals, have antennae or whiskers that are used to measure the position and orientation of surfaces in the environment. Virtually all creatures have tactile sensors in the skin, particularly in the digits, lips, and tongue. Proprioceptive sensors indicate the position of the feeler or tactile sensor relative to the self when contact is made with an external surface. This, combined with knowledge of the kinematic position of the feeler endpoint, provides the information necessary to compute the position on the egosphere of each point contacted. A series of felt points defines a surface on the egosphere.

Another primitive measure of surface orientation and depth is available from image flow (i.e. motion of an image on the retina of the eye). Image flow may be caused either by motion of objects in the world, or by motion of the eye through the world. The image flow of stationary objects caused by translation of the eye is inversely proportional to the distance from the eye to the point being observed. Thus, if eye rotation is zero, and the translational velocity of the eye is known, range to any stationary point in the world can be computed directly from image flow.

Knowledge of eye velocity, both translational and rotational, may be computed by the vestibular system, the locomotion system, and/or high levels of the vision system. Depth from image flow enables creatures of nature, from fish and insects to birds and mammals, to maneuver rapidly through natural environments filled with complex obstacles without collision. Moving objects can be segmented from stationary by their failure to match world model predictions for stationary objects. Near objects can be segmented from distant by their differential flow rates.

Distance to surfaces may also be computed from stereovision. The angular disparity between images in two eyes separated by a known distance can be used to compute range. Depth from stereo is more complex than depth from image flow in that it requires identification of corresponding points in images from different eyes. Hence it cannot be computed locally. However, stereo is simpler than image flow in that it does not require eye translation and is not confounded by eye rotation or by moving objects in the world.

Distance to surfaces may also be computed from sonar by measuring the time delay between emitting radiation and receiving an echo. Difficulties arise from poor angular resolution and from a variety of sensitivity, scattering, and multipath problems. Creatures such as bats and marine mammals use multispectral signals such as chirps and clicks to minimize confusion from these effects.

All of the above methods for deriving surfaces are primitive in the sense that they compute directly from sensory input without recognizing entities or understanding anything about the scene. Depth measurements from primitive processes can immediately generate maps that can be used directly by the lower levels of the task decomposition hierarchy to avoid obstacles and approach surfaces.

Additional information about surface position and orientation may also be computed from shading, shadows, and texture gradients. These methods typically depend on higher levels of visual perception such as geometric reasoning, recognition of objects, detection of events and states, and the understanding of scenes.

Recognition and Detection

Recognition is the establishment of a one-to-one match, or correspondence, between a real world entity and a world model entity .

For each entity in the world model, there exists a frame filled with information that can be used to predict attributes of corresponding entities observed in the world. The process of recognition begins by hypothesizing a world model entity and comparing its predicted attributes with those of the observed entity. When the similarities and differences between the world model prediction and the sensory observation are integrated over a space-time window covering an entity, a matching, or cross-correlation value is computed between the entity and the model. If the correlation value rises above a selected threshold, the entity is said to be recognized. If not, the hypothesized entity is rejected and another tried. The recognition process proceeds until a match is found, or the list

of world model entities is exhausted. Many matching processes may take place in parallel.

If a SP module recognizes a specific entity, the WM at that level updates the attributes in the frame of that specific WM entity with information from the sensory system.

If the SP module fails to recognize a specific entity, but instead achieves a match between the sensory input and a generic model entity, a new specific WM entity will be created with a frame that initially inherits the features of the generic entity. Slots in the specific entity frame can then be updated with information from the senosry input.

If the SP module fails to recognize either a specific or a generic entity, the WM may create an "unidentified" entity with an empty frame. This may then be filled with information gathered from the sensory input.

When an unidentified entity occurs in the world model, the task decomposition system may (depending on other priorities) select a new goal to <identify the unidentified entity>. This may initiate an exploration task that positions and focuses the sensor systems on the unidentified entity, and possibly even probes and manipulates it, until a world model frame is constructed that adequately describes the entity. The sophistication and complexity of the exploration task depends on task knowledge about exploring things. Such knowledge may be very advanced and include sophisticated tools and procedures, or very primitive. Entities may, of course, simply remain labeled as "unidentified".

Event detection is analogous to entity recognition. Observed states of the real world are compared with states predicted by the world model. Similarities and differences are integrated over an event space-time window, and a matching, or cross-correlation value is computed between the observed event and the model event. When the cross-correlation value rises above a given threshold, the event is detected.

The Context of Perception

If, as suggested in [1], there exists in the world model at every hierarchical level a short term memory in which is stored a temporal history consisting of a series of past values of time dependent entity and event attributes and states, it can be assumed that at any point in time, an intelligent system has in its short term memory a record of how it reached its current state. [1] also suggests that, for every planner in each task decomposition (TD) module at each level, there exists a plan, and that each executor is currently executing the first step in its respective plan. Finally, it can be assumed that the knowledge in all these plans and temporal histories, and all the task, entity, and event frames referenced by them, is available in the world model.

The effect is that the intelligent system almost always knows where it is on a world map, knows how it got there, where it is going, what it is doing, and has a current list of entities of attention, each of which has a frame of state variables that describe their recent past, and provide a basis for predicting their future states. This includes a prediction of where and how object surfaces will appear, and which surface boundaries, vertices, and points will be visible in the image produced by the sensor system. It also means that the position and motion of the eyes, ears, and tactile sensors relative to surfaces and objects in the world are known, and this knowledge is available to be used by the sensory processing system for constructing maps and overlays, recognizing entities, and detecting events.

Were the above not the case, the intelligent system would exist in a situation analogous to a person who suddenly awakens at an unknown point in space and time. In such cases, it typically is necessary for humans to perform a series of tasks designed to "regain their bearings", i.e. to bring their world model into correspondence with the state of the external world, and to initialize plans and system state variables.

It is, of course, possible for an intelligent creature to function in a totally unknown environment, but not well, and not for long. Not well, because every intelligent creature makes much good use of the a priori information that forms the historical context of its current task. Without information about where it is, and what is going on, even the most intelligent creature is severely handicapped. Not for long, because the world model continuously accumulates information about the current situation and its recent historical development, so that for example, within a few seconds a functionally adequate map of the surrounding space can usually be acquired.

The Mechanisms of Attention

In any intelligent system, sensory processing is an active process which is directed by goals and priorities generated in the task decomposition system.

In each node of the intelligent system hierarchy, the task decomposition TD modules request information needed for the current task from sensory processing SP modules. By means of such requests, the TD modules control the processing of sensory information and focus the attention of the WM and SP modules on the entities and regions of space that are important to the success of the intelligent system in achieving its behavioral goals. Requests by TD modules for specific types of information cause SP modules to select particular sensory processing masks and filters to apply to the incoming sensory data. Requests from TD modules enable the WM to select which world model data to use for predictions, and which prediction algorithm to apply to the world model data. TD requests also define which correlation and differencing operators to use, and which spatial and temporal integration windows and detection thresholds to apply.

Task decomposition TD modules in the attention subsystem thus actively point the eyes and ears, and direct the tactile sensors of antennae, fingers, tongue, lips, and teeth toward objects of attention. TD modules in the vision subsystem control the motion of the eyes, adjust the iris and focus, and actively point the fovea to probe the environment for the visual information needed to pursue behavioral goals [5,6]. Similarly, TD modules in the auditory subsystem actively direct the ears and tune audio filters to mask background noises and discriminate in favor of acoustic signals of importance to behavioral goals.

The act of perception involves both sequential and parallel operations. For example, the eye consists of an array of rods and cones that collect the image in parallel, but the photodectors on the retina are not uniformly distributed. They are focused into a high resolution fovea that is typically scanned sequentially over points of attention in the visual field [5]. While this sequential scanning is going on, parallel recognition processes hypothesize and compare entities at all levels simultaneously.

The Sensory Processing Hierarchy

It has long been recognized that sensory processing occurs in a hierarchy of processing modules, and that perception proceeds by "chunking", i.e. by recognizing patterns, groups, strings, or clusters of points at one level as a single feature, or point in a higher level, more abstract space. It also has been observed that this chunking process proceeds by about an order of magnitude per level, both spatially and temporally [7]. Thus, at each level in the proposed architecture, as SP modules extract from the sensory data stream the information needed to update the world model and service the task decomposition modules at that level, they also integrate, or chunk, information over space and time by about an order of magnitude.

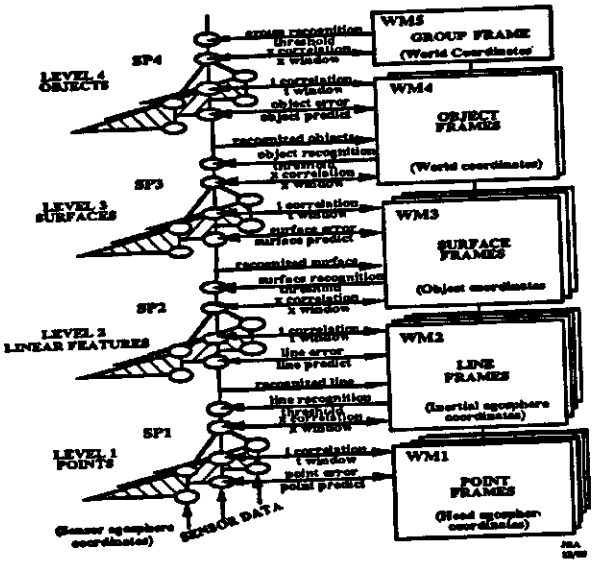

Figure 3. The nature of the interactions that take place between the world model and sensory processing modules. At each level, predicted entities are compared with observed. Errors are returned directly to the world model to update the model, as well as forwarded upward to be integrated over time and space windows provided by the world model. Correlations that exceed threshold are recognized as entities.

Figure 3 describes the nature of the interactions that are hypothesized to take place between the sensory processing and world modeling modules at the first four levels, as the recognition process proceeds. The functional properties of the SP modules are coupled to, and determined by, the requirements of the TD and WM modules in their respective processing nodes.

It can be hypothesized that there exists a correspondence between the egosphere and world model map representations suggested here, and hierarchical levels known to exist in the mammalian vision system. Figure 4 summarizes this hypothesis.

On the retina, visual input data variables consist of photometric intensities I(k, AX, El, t) measured by rods and cones, where k is the sensor egosphere index, AZ is the azimuth and EL is the elevation of the pixel on the egosphere where intensity I is observed at time t. Retinal sensor signals may vary over time intervals on the order of a millisecond or less.

On the retina, image preprocessing is performed by horizontal, bipolar, amacrine, and ganglion cells. Output data variables from the retina are spatial and temporal gradients carried by ganglion cell axons. Center-surround receptive fields of "on-center" and "off-center" detect both spatial and temporal derivatives at each point in the visual field. Outputs from the retina become input to sensory processing level 1 as shown in Figure 4. Level 1 inputs correspond to events of a few milliseconds duration.

In the brain, the level 1 vision processing module consists of the neurons in the lateral geniculate bodies and the tectum (or superior colliculus). Optic nerve inputs from two eyes are overlaid such that the visual fields from left and right eyes are in registration. Data from stretch sensors in the occular muscles provides information to the tectum (or superior colliculus), and possibly also to the lateral geniculate, about eye convergence, pan, tilt, and roll of the retina relative to the head. This allows map points in retinal coordinates to be transformed into head coordinates (or vice versa) so that visual and acoustic position data can be registered and fused [8,9]. It also provides the basis for range from stereo to be computed for points. However, experimental evidence seems to suggest that range from stereo is not computed before level 2 [10].

Additional data from the vestibular system and the locomotion systems provides the basis for estimating rotary and linear head and eye velocities and hence image flow direction. Center-surround receptive fields can be subtracted to derive spatial derivatives along image flow lines. At each point in the image where the spatial derivative along the flow direction is non-zero, spatial and temporal derivatives can be combined with eye velocity to compute the image flow rate. Range to each pixel can thus be computed directly and in parallel from local image data.

Level 1 inputs are integrated into level 1 output events spanning a few tens of milliseconds.

The level 2 vision processing module is hypothesized to consist of neurons in the primary visual cortex (area 17). At level 2, point entities are clustered and connected to form linear entities. Individual neurons detect edges and lines at particular orientations. Other neurons detect edge points, curves, trajectories, vertices, and boundaries. Level 2 inputs are integrated into level 2 output events spanning a few hundreds of milliseconds.

Input to the world model from the vestibular system as to the direction of gravity and the rotation of the head, allows the level 2 world model to transform head egosphere representations into inertial egosphere coordinates where the world is perceived to be stationary.

The level 3 vision processing module is hypothesized to reside in the secondary visual cortex (area 18). Clusters of linear entities are recognized as surface entities. Cells that detect motion of edges in specific directions provide indication of surface boundaries and depth discontinuities. Correlations and differences between world model predictions and sensory observations of surfaces give rise to meaningful image segmentation and recognition of surfaces. World model knowledge of lighting and texture allow computation of surface orientation, discontinuities, boundaries, and physical properties.

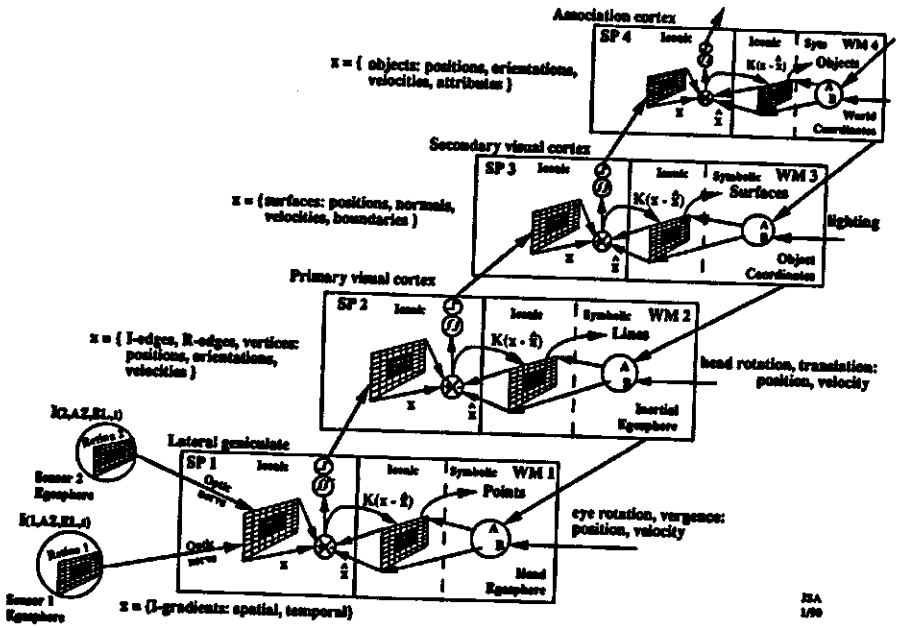

Figure 4. Hypothesized correspondence between levels in the proposed model and neuranatomical structures in the human vision system. At each level, the observed image is compared with a predicted image. At each level, additional map overlays are computed. There may be no reduction in resolution in the iconic representations of the first four levels; simply aggregation of pixels into more global entities.

Strings of level 2 events are integrated into level 3 events spanning a few seconds. (This does not imply that it necessarily takes seconds to recognize surfaces, but that both surfaces and patterns of motion that occupy a few seconds are recognized at level 3. For example, the recognition of a gesture that involves the tracking of the surface of an arm for a period of a second or more might occur at this level.)

World model knowledge of the position of the self relative to surfaces enables level 3 to transform inertial egosphere representations into object coordinates.

The level 4 vision processing module is hypothesized to reside in the tertiary visual cortex. At level 4, contiguous surfaces are recognized as objects. Correlations and differences between world model predictions and sensory observations of objects allows sets of surfaces to be clustered and recognized as objects and volumes. Strings of level 3 events are grouped into level 4 events spanning a few tens of seconds.

World model input from the locomotion and navigation systems allow level 4 to transform object coordinates into world coordinates.

Level 5 vision is hypothesized to reside in the association areas of the parietal and temporal cortex. At level 5, clusters of objects are recognized as groups. Strings of level 4 events are grouped into level 5 events spanning a few minutes. Outputs are transformed into a level 6 scale world coordinate map with larger span and lower resolution than level 5.

At higher levels, clusters of groups are recognized as higher level groups, and strings of events are grouped into higher level events spanning longer time intervals. World maps have larger spans and lower resolution.

Gestalt effects

When an observed entity is recognized at a particular hierarchical level, its entry into the WM provides predictive support to the level below. The recognition is also passed upward so as to prune the search tree at the level above. For example, a linear feature recognized and entered into the world model at level 2, can be used to generate expected points at level 1. It can also be used to narrow the search at level 3 to entities that contain that particular type of linear feature. Similarly, surface features recognized at level 3 can generate specific expected linear features at level 2, and limit the search at level 4 to objects with such surfaces, etc. The recognition of an entity at any level thus provides to both lower and higher levels information that is useful in selecting processing algorithms and setting spatial and temporal integration windows to integrate lower level features into higher level chunks.

If the correlation function at any level falls below threshold, the current world model entity or event at that level will be rejected, and others tried. When an entity or event is rejected, the rejection also propagates both upward and downward, broadening the search space at both higher and lower levels.

At each level, the SP and WM modules are embedded in a feedback loop that has the properties of a relaxation process, or phase-lock loop. WM predictions are compared with SP observations, and the correlations and differences are fed back to modify subsequent WM predictions. WM predictions can thus be "servoed" into correspondence with the SP observations. Such looping interactions will either converge to a tight correspondence between predictions and observations, or will diverge to produce a definitive set of irreconcilable differences.

Perception is complete only when the correlation functions at all levels exceed threshold simultaneously. It is the nature of closed loop processes for lock-on to occur with a positive "snap". This is especially pronounced in systems with many coupled loops that lock on in quick succession. The result is a gestalt "aha" effect that is characteristic of many human perceptions.

Flywheeling, Hysteresis, and Illusion

Once recognition occurs, the looping process between SP and WM acts as a tracking filter. This enables the WM model predictions to track real world entities through noise, data dropouts, and occlusions.

In the system described above, recognition will occur when the first hypothesized entity exceeds threshold. Once recognition occurs, the search process is terminated, and the thresholds for all competing recognition hypotheses are effectively raised. This creates a hysteresis effect that tends to keep the WM predictions locked onto sensory input during the tracking mode. It may also produce undesirable side effects, such as a tendency to perceive only what is expected, and a tendency to ignore what does not fit preconceived models of the world.

In cases where sensory data is ambiguous, there is more than one model that can match a particular observed object. The first model that matches will be recognized, and other models will be suppressed. This explains the effects produced by ambiguous figures such as the wire-frame cube.

Once an entity has been recognized, the world model projects its predicted appearance so that it can be compared with the sensory input. If this predicted information is added to sensory input (or multiplied by a positive bias), perception at higher levels, will be based on a mix of sensory observations and world model predictions. By this mechanism, the world model may fill in sensory data that is missing, and provide information that may be left out of the sensory data. For example, it is well known that the audio system routinely "flywheels" through interruptions in speech data, and fills-in over noise bursts.

This merging of world model predictions with sensory observations may account for many familiar optical illusions. In pathological cases, it may also account for visions and voices, and an inability to distinguish between reality and imagination.

Summary and Conclusions

Both the sensory processing and world modeling systems are hypothesized to be hierarchical, with a tight coupling between SP and WM modules at each level of the hierarchy. At every level, world model predictions are compared with sensory observations. Correlations and differences are used to improve estimated values stored in the world model. The world model consists of both symbolic and iconic representations, and mechanisms for transforming from one to the other. The sensory processing system consists of iconic images, coordinate transformers, and mechanisms for comparison, correlation, recognition, and detection. The close coupling between sensory processing and world modeling produces the phenomenal capacity of an intelligent system to recognize objects, detect events, and perceive what is important for successful behavior in the natural world.

REFERENCES

1. Albus, J.S., "A Theory of Intelligent Systems", Proceedings IEEE Conference on Intelligent Control, Sept 1990, Philadelphia, PA

2. Albus, J.S., "The Role of World Modeling and Value Judgment in Perception", Proceedings IEEE Conference on Intelligent Control, Sept 1990, Philadelphia, PA

3. Kent, E., and Albus, J.S., "Servoed World Models as Interfaces between Robot Control Systems and Sensory Data", Robotica, Vol. 2, pags. 17-25, 1984.

4. Dickmanns, E.D., Th. Christians, "Relative 3D-State Estimation for Autonomous Visual Guidance of Road Vehicles", Intelligent Autonomous System 2(IAS-2), Amsterdam, 11-14 December, 1989

5. Yarbus, A.L., *Eye Movements and Vision*, Plenum Press, 1967

6. Bajcsy, R., "Passive perception vs. active perception" Proc, IEEE Workshop on Computer Vision, Ann Arbor, 1986

7. Miller, G. A., "The Magical Number Seven, Plus or Minus T: Some Limits on Our Capacity for Processing Information", The Psychological Review, 63, pp.71-97, 1956.

8. Pearson, J.C., J. Gelfand, W.Sullivan, R. Peterson, and C. Spence, "A neural network approach to sensory fusion", Proceedings SPIE Sensor Fusion Conference, Orlando, 1988

9. Sparks, D.L. and M. Jay, "The Role of the Primate Superior Colliculus in SensorimotorIntegration", In *Vision, Brain, and Cooperative Computation* (Arbib and Hanson, Eds.) MIT Press, Cambridge, 1987

10. Hubel, D., *Eye, Brain, and Vision*, Scientific American Library, Freeman, N.Y., 1988.