Computer Science Department Technical Report
Machine Perception Laboratory
University of California
Los Angeles, CA 90024-1596

NEURAL SPECIFICATION OF A GENERAL PURPOSE VISION SYSTEM

Josef Skrzypek

December 1989
CSD-890072

MPL

Machine
Perception
Lab

UCLA
Computer Science
Department

Neural Specification of a General Purpose Vision System

Josef Skrzypek

TR 89-9

Sept. 5 1989

MPL

Machine
Perception
Lab

top-down

bottom-up

Neural Specification of a General Purpose Vision System

Josef Skrzypek
Machine Perception Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024

Abstract

The unpredictability of events in an unconstrained environment implies that an autonomous land vehicle (ALV) must be equipped with robust, "real-time" perceptual system. To realize such a system means to program it based on the expectation of future, unconstrained events. Hence, there is a need to understand how to specify a "real-time general purpose" machine vision (GPV) system that is capable of PERCEIVING and UNDERSTANDING images in an unconstrained environment.

The research undertaken at the UCLA Machine Perception Laboratory addresses this need by focusing on two specific issues: *1) functional specification of GPV in the domain of computational neuroscience; the long term goals for machine vision research as a joint effort between the neurosciences and computer science; and 2) a framework for evaluating progress in machine vision.* The primary motivation behind our approach is that the human visual system is the only existing example of a "general purpose" vision system which uses a neural computing substrate to complete all necessary visual tasks in "real-time".

1 Introduction

The long term goal of the machine vision effort is to synthesize a real-time general purpose system that can perceive and understand images in an unconstrained environment. Our approach towards this goal is through interdisciplinary research involving artificial intelligence and neuroscience. From the AI perspective, we are using computational modeling as a domain in which to study fundamental problems of image understanding including knowledge representation, reasoning, problem solving, image processing and analysis. From the neuroscience perspective we are using AI techniques to model and understand visual functions and their computational substrate as discovered in natural systems.

Even the most optimistic guess leads to the conclusion that it will be a long time before we can fully duplicate human vision abilities. Therefore, our near-term goal is to better understand the meaning of general purpose and to define a functional block diagram of a GPV system. Since we assume the process to be evolutionary, we expect that our initial definition of "general purposeness" will improve with future refinements.

The goal of this paper is to outline our vision of GPV with emphasis focused on five specific problems: 1) definition of "general purpose" vision in terms of a kernel of visual tasks that underlies all visually guided human behavior, 2) breaking down tasks into routines and algorithms, 3) mapping routines and task to specific functional modules as constrained by information from neuroscience, 4) specifying connectivity between modules, 5) synthesize tasks computed by modules in terms of neural structures.

1.1 Interdisciplinary approach to computer vision

In the past, vision research has been carried out disjointly by various disciplines including the neurosciences, psychology, optics, computer science, and electrical engineering. Recently, attempts have been made to combine knowledge from these disparate fields. For example, Kosslyn [1,2] and Shwartz's [3] approach to the investigation of visual processes at the very high level, combines cognitive psychology, neuropsychology, and aspects of computer science. In his view, each effort has advantages and disadvantages, but the integration of these approaches provides an opportunity to use a particular field's strengths while perhaps circumventing another's weaknesses. His theory attempts to mimic the mental events that occur when humans generate and use mental images without detailed consideration of underlying neuronal mechanisms or the reality of the images. The main claim is that brain imaging functions are divided into two classes: knowledge structures (active and long term memories) and processes (IMAGE, PICTURE, PUT, FIND, etc.). Active memory maintains the image being operated on, while long term memory maintains visual information in terms of images along with nonvisual properties of the images (a hierarchical description of objects). Processes operate on these structures to generate, transform, or inspect mental images. This aspect of human image understanding, involves processes of *imagining* objects and scenes from memory. Our primary interest is in the phenomenon of visual recognition which can be defined as the ability to extract from the 2D retinal input, all meaningfull patterns of light, and to interpret them as representation of objects, scenes and their properties

Marr [4] offers another example of elucidating, potentialy useful principles for machine vision by constructing computational models of some low-level visual functions of biological systems. His first work in the area of vision was an attempt to develop theory based on neurophysiology of the retina [5]. Despite some wrong conclusions about physiology the approach was to use neural structures to constrain the mathematical model. However in his later work he switched to viewing perception as an information processing problem where algorithms must be invented before attempting to understand neuronal structure.

Our approach incorporates elements of both, early stages of visual processing as well as some cognitive aspects of vision. We want to enumerate all visual tasks as high-level processes that comprise the phenomenon of unconstrained perception and we want to specify the resulting functional modules in terms of underlying neuronal structures that obey rules of connectivity and interactions.

There are many ways to synthesize models of brain functions. One approach, based on mathematical methods tries to understand from the mathematical view point the capabilities of neural networks, that might have no correspondence in biological systems. Another way of theorizing about the brain function is exemplified by artificial intelligence. This is a top-down approach, where hypothesized brain functions are simulated as "intelligent behavior" in the form of a computer program without reference to any underlying neural mechanism. Our approach is different from above two methods in that our neural net models of specific brain functions adheres to current physiological and psychophysical data. Although the model is not identical in details to the reality, we can learn a lot from achieving functional equivalence [6]. For example, computationally, there are two aspect of segmentation: 1)physical and geometric constraints of the world and 2)the specification of the computing substrate. The constraints must be discovered and posed in some mathematical formalism in order to be solved. However, the selection of the formalisms is constrained itself by the underlying computing architecture. In this sense our approach is different from the one popularized by Marr [4] where he advocates algorithmic approach to vision; global brain phenomena, captured by psychology need only to be model by mathematical formalisms, and once in this form they can be computed by any substrate. We believe that mathematics constrained by the specifics of neural structures leads to heuristic solutions that are neither mathematically elegant nor computationally optimal.

To simplify the simulation of some of the brain functions we reduce our modeling to static processes. Since neural events are dynamic the best modeling approach is to use differential equations. This gives maximal information when the dynamic system is away from equilibrium. Omitting the time component can significantly degrade our understanding of the computation performed by the neural structure. However, many problems in early vision can be initially analysed in static form. Because our models are expressed in terms of specific neural architectures we can characterize our approach as mostly bottom-up. The key principle here is that given the behavior of components and how they are interconnected, it should be possible by simulation to study and to specify a global behavior of the system. Following general rules about local connectivity it is possible to achieve the emergence of order out of seemingly chaotic organization [7]. One benefit of this top-down approach is that by explicitelly specifying the modeling paradigm we are forced to address many issues that would otherwise be neglected. By studying computer simulations of the model, we can predict neural phenomena of interest to neurophysiologists and we can investigate how faithfully the model generates emergent global functions that are analogous to psychophysical data. This will help to extract most usefull principle for synthesizing computer vision.

It is conceivable that complexity of the neural structure in the human perceptual function does not have the same meaning as in machine perception. Namely, biological vision system might use a heuristic solution in the form of structural additions to genetically inherited architectures which are not decomposable into simpler computing units. Furthermore, it appears that complete reliance on the neurophysiology of vision when specifying GPV is subject to the basic "FUNCTION FROM STRUCTURE" problem of neurosciences, namely how to relate the neuronal computing substrate to the function that it performs. Considering that the brain consists of billions of neurons, each with perhaps 1000 to 10000 synapses, arranged into varying, task dependent architectures, it is doubtfull that neurosciences will be able to study in deterministic manner every synapse. Therefore, modelling studies by computer simulations seem to be all the more significant and perhaps the only reasonable approach to the problem.

Can we synthesize a general purpose machine vision system without taking advantage of knowledge transfer from psychology or neurophysiology of vision? Slow progress of the past thirty years suggests that it would be an extremely difficult undertaking. Which aspect of biological knowledge would be most usefull? Phenomenological description of visual processes at the global level as offered by psychology or detailed outline of local neural architectures as described by neurophysiology, or both? One argument posed by psychology (Hochberg [8]) is that machine vision systems should not be designed to emulate human vision, but it would help if machines *knew* how people see; they should not suffer from visual inconsistencies due to depth reversals, illusions, or apparent motion. The implication is that such illusions represent unwanted byproducts of a very complex and "general purpose" system. However, consider the role played by illusory (subjective) contours when we recognize an object presented as an incomplete figure; illusory contours aid us in completing the missing data and in fact neurons have been found that specialize in detecting subjective contours [9]. This implies that our ability to perceive illusory contours is not just a result of visual inconsistencies, but rather a purposeful feature of the system, designed to aid in our perception of the environment.

Another implication of Hochberg's viewpoint is that we should keep developing vision systems specifically tailored to handle the particular tasks. In general, it is virtually impossible to apply these highly specialized systems to new tasks without major redesign. In the past thirty years, this approach has resulted in many excellent dedicated imaging systems but it did not provide much usefull knowledge on how to build a more robust vision system; knowing details of an insect vision specialized for detecting ultraviolet radiation in navigation tasks might not help much in understanding problems of higher level vision in human system.

1.2 A new proposal for a GPV machine implies ability to evaluate existing systems

The current state of affairs in computer vision research is analogous to the "catch-22" situation. There are many different vision systems, realized or proposed that aspire to become a GPV's. However, the definition of "general purposeness" is lacking; there are no guidelines enumerating all visual tasks that GPV must be able to perform. In other words, we are missing the complete list of goals for GPV. In absence of such definition it is very difficult to derive common metrics for evaluation of the various systems. And without evaluation it is difficult to propose new, improved and more general systems. To define generality we need to evaluate current systems and to evaluate them we need to have a common definition of generality, hence catch-22.

Every existing machine vision systems has been designed for a specific purpose. This means that they can not perform the same perceptual or categorization tasks. Consequently, evaluating their performance based on a fixed set of input images is almost impossible. It does not make much sense to evaluate vision systems using such techniques as figure-of-merits (FOM) [10]; FOM derived from weighted combinations of measures such as speed, reliability, and accuracy, are not likely to help in deciding which visual tasks are important or generic. After all, using a FOM to compare a system designed to locate three types of industrial parts in a highly constrained environment with another system designed to locate a camouflaged target in an unconstrained environment is meaningless. Such measures might be useful in the image processing domain where for example convolving an image with some kernel is a well defined procedure, but in vision the result of using FOM for general evaluation would most likely be misleading. Therefore, one of our research goals is to enumerate and categorize all visual tasks that a GPV system must be able to perform. This might lead to a framework for evaluating progress in machine vision systems, useful not so much to judge other systems, but rather to discover what each system has proposed and where to direct future efforts.

The remainder of this report begins with a cursory review of fifteen machine vision systems developed during the last decade in order to elucidate possible categories along which machine vision systems may be evaluated. This analysis is also intended to search for common, fundamental computational principles used by the different systems. We would like to endow our general vision system with these same principles. In the following section we attempt to justify the use of the human visual system as a model for GPVS. Next, we discuss the problem of comparing the visual performance of humans and machines, and which visual tasks can be used to measure the obvious gap. The report concludes with an outline of a proposed general purpose machine vision system derived from integrated analysis of current data in neurosciences.

2 Review of Selected Systems

In an attempt to elucidate principles comprising the definition of a "general purpose" system we first analyzed fifteen systems built during the last fifteen years (Table 1). We have based our analysis on five dimensions: (1) image attributes; (2) perceptual primitives; (3) knowledge base; (4) object representation; and (5) control strategy. For the most part, we tried to use only systems that have either been proposed and partially realized or completely built.

Most of the systems use one or two image attributes such as edges and perhaps color. Some of them use higher-level attributes such as texture while a few of the systems use motion. Image attributes can be defined as the most basic elements of pictorial representation which carry nonredundant information. By independent information we mean for example that color information can not be obtained from texture or motion. It is clear that a general purpose machine vision system requires the exploitation of all of the image attributes (Table 1).

Perceptual primitives represent the second comparison dimension. A working definition of a perceptual primitive might be the abstraction of image attributes into higher level data structures. This abstraction may employ the use of Gestalt laws of organization [11] such as closure, similarity, proximity, collinearity, common fate, etc. Gestalt effects can be considered as mechanisms that group regions based on the idea of uniformity.

Analyzed systems use very simplistic perceptual primitives which display very little abstraction and are intimately related to the original image attributes. Two of the most often used primitives are lines and region, directly related to contours. Most often, image attributes are simply integrated into new data structures. None of the systems uses more abstract perceptual primitives such as illusory contours.

A third way of looking at these systems is to examine the type of knowledge they use and the way the knowledge is represented and manipulated. In all of the systems, knowledge is constrained to a very narrow application domain. It is not immediately obvious how to compile all of the knowledge that relates to unconstrained environments.

We also examined the use and representation of objects. All of the systems use objects almost directly related to the very lowest levels of image representation. It would seem that some more complex (symbolic) representations, abstracted from combinations of those primitives would be more beneficial. All of the available representational schemes such as graphs, frames, production rules, 3D CAD models, and generalized cones are used in various systems. For example, Hanson and Riseman's VISIONS schema structure provide a rich symbolic mechanism for the hierarchical representation of objects. In their method an object could range from being an urban schema,

Principal Investigators (Year)	Image Attributes	Perceptual Primitives	Knowledge Base	Object	Represen-tation	Control Strategy	Implemen-tation
Ballard, Brown, & Feldman (1978)	1, 2	1, 2	1	1,2	1	1	2b
Hanson & Riseman (1978)	3a, 2, 3	1, 2	1	1,2	1d	1,2	1
Nagao & Matsuyama (1978)	1	2	1	1	1a	1,2	1
Rubin (1978)	2, 3	2	1	3	1a	1,2	3
Shirai (1978)	1	1, 1a	1	4	4	1,2	2
Ohta (1980)	3a, 2, 3	2	1	1	1a	1,2	1a
Brooks (1981)	1	2a, 1b	1	5	2	1,2	4
Nevatia & Price (1982)	1, 3a, 3	1, 2	1,2	1,2	1b	1	3
Shneier, Lumia, & Kent (1982)	3a,1,3	1,1c	1	5a	5	1a,2a,2b,3	4
Bolles, Houraud, & Hannah (1983)	1	2	1,3	5a	5	1,2	4
Levine & Nazif (1984)	3a,3b,3c,2	1,2,2b	1	1,2	3	1a,2a	4
Herman & Kanade (1984)	1,1a,4	1e,2c,1d	1	6	1c	2	3
McKeown (1985)	2,3,4	1,2	1	3a	7	1,2	1
Tsotsos (1985)	1,5	1f	1	4a	2	1a,2a,2b,3	5
Davis & Kushner (1986)	1,1a,3,5	1,2	1	3b	6	1a,2b,3,2a	2a

Image attributes: Edges = 1; Point = 1a; Texture = 2; Color = 3; Intensity = 3a; Hue = 3b; Saturation = 3c; Depth = 4; Motion = 5

Perceptual primitives: Line Segments = 1; Curves = 1a; Ribbons = 1b; Corners = 1c; Junctions = 1d; Verticies = 1e; Markers = 1f Regions = 2; Ellipses = 2a; Areas = 2b; Faces = 2c;

Knowledge base: Domain constrained = 1; Context constrained = 2; Location constrained = 3

Object: Regions = 1; Lines = 2; 3D map = 3; Airports = 3a; Roads = 3b; Feature values = 4; Heart marker = 4a; Generalized cones =5; Parts = 5a; 3D wireframe = 6

Object representation: Constraints graph = 1; Relational graph = 1a; Schematic network =1b; Structure graph = 1c; Schemas = 1d; Frames = 2; Rules = 3; Procedures =4; 3D CAD model = 5; ALV terrain model = 6; MAPS database = 7

Control strategy: Top-down = 1; Data driven control = 1a; Bottom-up = 2; Goal driven control = 2a; Model driven control = 2b; Temporal driven control = 3

Control implementation: Rule based system = 1; Production system = 1a; Monitor program = 2; Vision executive = 2a; User executive program = 2b; Procedural = 3; Prediction system = 4; Frames = 5

Table 1: Comparison of Systems Along Image Attribute, Perceptual Primitives, Knowledge Base Dimensions, Object and its Representation and Control.

which is composed of lower level subschemas such as house and road schemas, to simply a car schema. This type of rich symbolic image data structures is surely required for the development of a general purpose vision system, although the implementation may vary significantly.

Finally, we looked at the control strategy that is used to process the data within those systems. All of the systems use objects almost directly related to the very lowest levels of image representation. In most cases both, bottom-up and top-down strategies are used but only a few systems have incorporated temporal aspects of the environment or goal-driven strategies to process the information.

It is extremely difficult to derive detailed conclusions from our preliminary comparisons of existing systems. In the absence of a common definition and specification for all these systems, and the lack of a good definition for general purpose vision, comparisons between these specialized systems can be cursory at best. Clearly, these systems do not use all of the available information from the early stages of processing. The knowledge domain in all of the systems is highly constrained. The high-level vision components in all examined systems are rather weak and very much ad hoc. All of the high-level processes are derived from established artificial intelligence concepts (see [12]) that have been borrowed from natural language processing research. It is not clear that processes underlying the higher levels of vision are identical to processes involved in language understanding; most probably this is not the case. Although all systems are (implicitely or explicitely) able to perform object recognition task, the evaluation of this performance by the inventors is inconsistent or completely lacking. Some systems can operate only on presegmented images of one object only. Others can "interpret" multiple objects but only if certain conditions are satisfied. These might include: well controlled illumination of the scene, object placed in particular position, object models developed independently of visual input, etc. In general, the task of object recognition is not defined and only very simple object are considered. For the most part, temporal aspects of the sensory environment have not been addressed by these systems. Finally, although every system starts out with the goal to build a general purpose vision system, either explicitly or by extension, for example see [13], little attempt was made to define the meaning of a general purpose system, much less the development of such a system. Perhaps this difficulty lies in the fact that we do not now what a general vision system is and we have never attempted to define strictly what is vision.

None of the machine vision systems presented in the previous section has the capability to perceive and understand images in an unconstrained environment. It is conceivable that many problems could be resolved if we had a specification for a general purpose system. Attempts to build systems using physical laws such as the laws governing image formation, have proven to be intractable. Another promising approach to ill-posed vision problems is to reformulate them into problems which either have already been solved or for which a solution can be discovered. The reformulation can be realized only through the introduction of numerous constraints, thereby making the problem more tractable. However, the constraints often change the problem beyond its original specification. While this approach has met with some success in low level visual tasks such as edge detection [14], its application to higher level visual tasks, such as shape from shading [15] has failed.

All of the realized systems are unique, special case systems that work well in their dedicated application domain. Hence, it is very difficult or impossible to have a meaningful comparison of a system against its competitors or different approaches (see also [16]. It is even difficult to decide if all of the domains that have been selected are the right ones, or even significant ones, or how they contribute to our overall desire to understand and implement a general purpose system. In this sense the review of existing machine vision vision system did not help in explaining what general purpose vision is. However, it is expected that machine vision systems under development for operation in unconstrained environments, such as the autonomous land vehicle research effort, will contribute significantly to understanding these problems.

3 Human Vision as a Model of General Purpose Vision

Our analysis of 15 existing machine vision systems did not help to better understand the meaning of the term *general purpose*. The list of the manifest properties of a general purpose system has never been compiled in the past. Which characteristics of perception would qualify the system for successfull operation in an unconstrained environment? Clearly, it must be able to handle a wide assortment of visual tasks, some of which we have considered in previous sections. There are no rigorous studies that attempt to enumerate and define all visual tasks. Consequently this is one of our goals in this project.

What can be gained from the analysis of the human visual system? Clearly, it serves as an existence proof of a *general* purpose vision system, capable of adapting to the requirements posed by the unconstrained environment in which we live. Computer vision literature is full of examples where the limits for machine vision performance are modeled after the limits of human vision [16]. By analyzing the human visual system, the **only** vision system that works, we can have a better understanding of what vision is and what are the critical visual tasks that must be performed by a *general* purpose system. Along this line, the analysis of visual deficits in the human visual system may shed some light on the functional organization and the neural mechanisms underlying the performance of particular visual tasks.

3.0.1 Human visual tasks

We assume that the human perceiver represents *general* purpose vision system for which we are able to enumerate most of the visual tasks. Some of these tasks might be very difficult for current implementation of machine vision systems, others might be trivial. We pose that in human perception, there exists a kernel of visual tasks, a subset of which underlies most of the visually guided behavior in an unconstrained environment. Having this list of visual

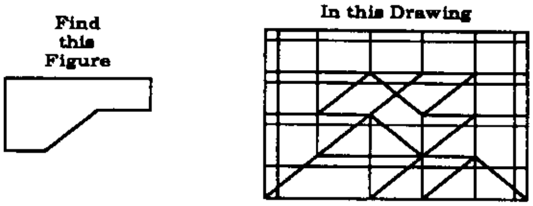

Figure 1: Sample Question from the Hidden Figure Test

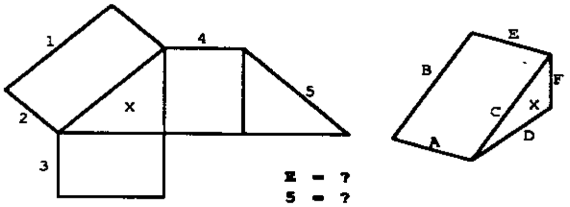

Figure 2: Sample Question from the Surface Development Test

functions would simplify our task to develop specifications of a "general" purpose machine vision.

One compilation of visual tasks has been developed by the Education Testing Service (ETS) for the measurement of childhood cognitive development. The underlying mechanisms which permit human vision to succeed in such tasks are not completely understood. This is an area of active investigation within neuropsychology and cognitive psychology. Knowing these mechanisms would perhaps allow us to define some of the underlying primitive visual functions that could be transferred to machine vision. The problem is that we need to select tasks that allow meaningful transfer from human performance to machine vision.

The Hidden Figures Test CF-1 [17], is an example of a visual task that requires the participant to find a given shape in some complex picture (the shape undergoes no rotation or size changes when placed in the picture, see Figure 1). This appears difficult for us because of the cluttered background. Using perhaps very high level processes of visual cognition, we can usually solve this visual task after starring at it for while. In the process of looking at it we probable employ such subtasks a boundary tracking, symmetry finding, matching etc. A computer program, on the other hand, using very low-level processes such as template match on a run-length encoded image could complete the task in a fraction of a second. Therefore, a comparison between human and machine performance based on this task entails computation at different levels of processing space and is not be very revealing .

Another example test is called the Surface Development Test VZ-3 [17]. In this case, the task is to fit a given surface together into a line drawing representation of the three dimensional object (Figure 2). Some tasks of this type are easy while others are very difficult and might involve very different problems of selecting and representing models, manipulating internal models of complex three dimensional objects, and finally matching the models against the data. Some of these problems, such as geometric reasoning, pose a great challenge to AI in general although they are relatively routine for humans.

Another task to consider is the Gestalt Completion Test CS-1 [17] which might be closely related to segmentation. In this task, a subject is given incomplete data such as that presented in Figure 3. The subject must determine what the data represents. Problems posed by this test include how to match the incomplete data against a set of models that one has acquired over time. There are many difficult questions to address. What is a model? How are models represented? How much data is needed? How are models matched against incomplete data? How do we implement something like illusory contours that would help in the performance of this type of task? Do we first complete the data by filling in the illusory contours? Biological vision has evolved specialized neurons that sense illusory contours after proper context has been analyzed. On the other hand computer vision has yet to demonstrate an algorithm that even can detect the possibility of a contour completion in absence of data.

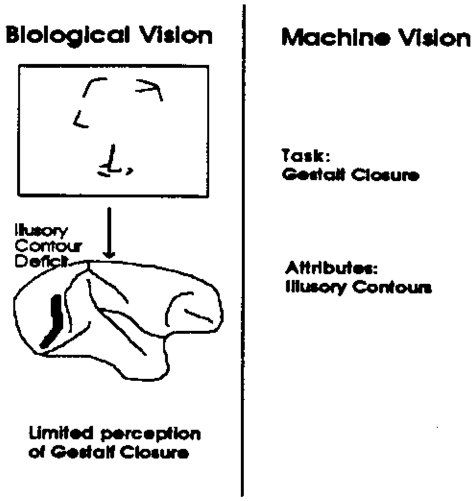

Figure 3: Relating a Computational Task with its Computational Substrate

We have begun developing a prototype of a visual task sourcebook. The selected visual tasks range in difficulty and in their association with low-level or high-level visual processing. Although the sourcebook is far from complete, it does provide a more extensive set of visual tasks which should conceivably be handled by a general purpose machine vision system. An incomplete listing of visual task includes: object recognition, selective attention, emergent feature configuration, divided attention, visually guided manipulation (pick and place; assembly), visual inspection (surface; missing parts; relative dimension; sorting; labeling), visually guided navigation (road following, obstacle avoidance, landmark recognition, cross-country navigation), figure-ground discrimination, spatial relationships between shapes and parts of shapes, inside-outside relationship, Gestalt groupings (good continuity, proximity, similarity, symmetry, common faith, closure), expectation based on context, boundary tracking, context analysis, visual counting, model building - intermediate representation, illusory contour detection, and others. Having this list of visual task required of future GPV we would like to relate them to specific modules the could support such computation.

3.1 From visual task to neuronal architecture

How can we judge that a selected visual task is fundamental to human visual performance? The answer to this question requires the combination of knowledge from two subfields of neuroscience. We need to combine our knowledge of functional neuroanatomy and physiology of the visual system with clinical neuropsychology. The former provides bottom-up, detailed information about the anatomy and function of neural architectures underlying a particular visual task. The latter on the other hand generates a top-down, global relationship between various aspects of the visual task and the coarse structures of the visual system.

All of these tasks at some lower level derive from primitive image attributes that include motion, texture, color, stereo and perhaps edges and lightness. Specific information about depth for example, can be extracted from a combination of stereo, motion and lightness or shading. Motion in depth could be obtained from combining motion, edges and stereo. Lightness, color, depth and motion can tell us about light sources. In this way we should be able to enumerate all information at different levels that is necessary to support a high level visual task such as obstacle avoidance. This exercise will also specify, in part, the connectivity between different modules. More significantly this approach might also eliminate the need for computationally expenssive approaches such as representing visual functions as mathematically ill-posed problems [18].

To illustrate the relationship between a visual task and its computational substrate, consider the visual task of Gestalt closure (see Figure 3). We know from studying human vision that a visual deficit in illusory contours, which may be related to the pathology of area V2, is manifested by a reduced or limited performance in Gestalt closure

Visual Deficit	Symptom	Damage
Autopagnosia [19]	Impaired recognition of body parts	Parietal lobe
Simultanagnosia [20]	Only one aspect of an object can be appreciated at a time, e.g. color or shape	Left hemisphere of Occipital Lobe
Balint's Syndrome [20]	Inability to visually localize objects in space	Occipital Lobe
Object Mirror Reversals [21]	Use common objects upside down or backwards	Area 39 - Parietal-Temporal-Occipital Association Cortex
Visuoimaginal Constructional Apraxia [22]	Unable to draw a simple object without a model	Area 37 - Temporal Posterior Cortex
Charcot-Wilbrand Syndrome [19]	Unable to revisualize images	Posterior Temporal Cortex
Anton's Syndrome [19]	Agnosia where a patient denies that they are blind	Area 7 - Parietal Lobe
Gerstmann's Syndrome [19]	Disability to calculate, right-left disorientation	Area 39 - Parietal Temporal Association Cortex
Prosopagnosia [20,23,24]	Failure to recognize faces and complex objects	Bilateral Occipital association areas

Table 2: Visual Deficits and Corresponding Lesions

tasks. Inability to do Gestalt closure leads to other failures in visual performance related to boundary finding. Because V2 is considered to be early on in the hierarchy of visual processes this suggests that the Gestalt closure visual task seems fundamental to many other visual functions. The ability to perform tasks based on Gestalt closure, translates into the development of a much more general machine vision system.

The end result of our preliminary research is a list of functions that underly the performance of selected visual tasks. Table 2 presents neuropsychological results that aid us in relating functions and tasks to specific anatomical structures. The association of visual tasks with their possible neuronal substrate provides us with a methodology for synthesizing a framework for a general purpose machine vision system. Our map of the functional areas of the brain is a combination of the results of many studies [25,26,27,28,29,30]. However, since the functional neuroanatomy of the Macaque monkey is not easily interpreted within the realm of computer science, we simplified this task by constructing functional diagrams like the one presented in Figure 4. Interestingly, there are multiple hierarchies of processing stages and the connectivity among various modules, although nonrandom, allows for direct vertical and horizontal interactions between processing stages which are seemingly distant in function. This is unlike the clear, single channel, sequential structures of GPV proposed within Computer Vision where only immediately neighboring stages withing the hierarchy can communicate with each other [13].

4 Elements of GPV

It is impossible to completely specify GPV at the present time. Hence, we will be concerned only with some elements of GPV which seem to require more attention if progress is to be made. To simplify our consideration we can make certain "common sense" assumptions about a natural environment. For example, matter is cohesive, therefore, adjacent regions stick together in space and time. Most of the object that we manipulate and interact with are solid. As such one object can occupy only one point in space at one time. Many objects are symmetrical and they are usually attached (stand) to some surface. In comparison to background, objects are small and surface properties are similar within the bounds of the object. To simplify the problem further we can make some reasonable assumptions about a GPV. Multiple views of the object or a scene are available during the model formation (learning) stage. Vision consists of various complex functions including recognition, model formation, scene reconstruction, visual task planning, and others. We will emphasize visual recognition which simplfies the problem by presupposing the existence of memories and world model. Human visual system is able to recognize objects of various positions/orientations and to describe highly complex scenes depicted in black and white photographs. This suggests that all of the information necessary for recognition can be encoded in only one variable - intensity

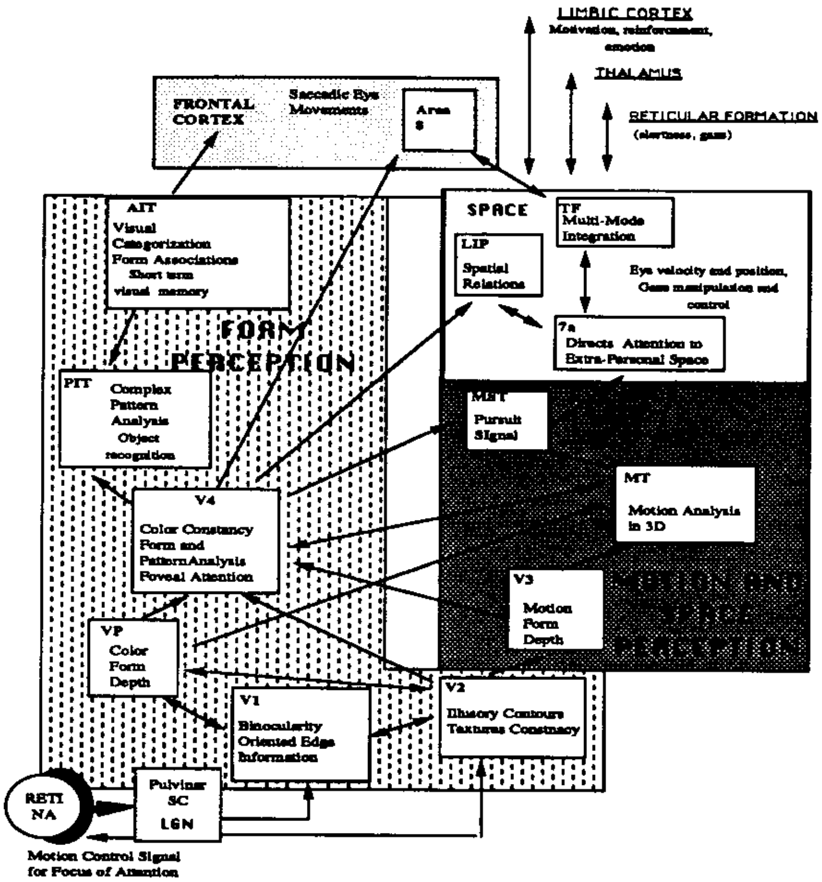

Figure 4: Functional Block Diagram of the Visual System

changes. For computer vision, this is one of the most compelling justification for using human vision as a model of GPV.

Our overall philosophical approach to visual recognition problem is that the system consists of hierarchically arranged modules that incorporate at least three different theories of shape recognition: feature detectors, template matching and symbolic manipulation. Our approach thus separates the process of visual recognition into partially overlapping but distinct stages. For example, the lowest levels will always produce boundaries regardless of the higher processes that recognize boundaries as contours of the shape. Depending on the motivation, some of the boundaries may be considered to be irrelevant. It is our intent not to assume the availability of the input without specifying and generating the output from the preceeding stage and to account for every processing step in terms of an explanation based on neural architectures.

Simple visual functions such as edge detection and textural segmentation [31,32], color and motion processing are carried out by the early stages. Complex pattern analysis, perceptual organization, short term memory, attention, spatial perception, context analysis and the formation of visual categories are carried out at the intermediate levels. The highest levels are concerned with generating goals, specifying tasks, planning their execution, long-term memory, multisensory interpretation, etc. The whole system is loosely modeled on the two-channel flow of information from the retina to higher centers. The occipito-temporal path deals with identification of shape - what is object? The other, occipito-parietal visual path is concerned with spatial perception - where is object?

At the lowest level we have feature detectors [33,34] that produce responses to all image attributes. For example, V1 contains cells responsible for binocular convergence, color, spatial frequency filtering, pattern specific adaptation, orientation and direction selectivity. Here, firing of a cell or a group of cells is a neural representation of some aspect of the scene. A vector of weights representing various features serves as an intermediate representation for parts of shapes. At this level spatial relationships among features are fixed by the location of detectors within neural networks. Some functions computed at this level include color [35], lightness [36,37]and size invariance [38], optical flow, some texture, stereo and motion segmentation, data abstraction as well as illusory contours.

The output of the earlier stages results in new representations in the form of perceptual primitives or their combinations that resemble parts of shape. The abstractions of perceptual primitives out of image attributes involves Gestalt laws of grouping. Gestalt effects are perhaps preattentive parallel mechanisms that develop through interactions with environment very early on in life. At the core of all Gestalt laws is a process of grouping, the result of which may lead to emergence of a new configuration of existing or new shape features. In general, any Gestalt grouping law implies a low level process without the need for attention. On the other hand perceptual assembly without grouping requires selective attention for examination [39]. It is possible that some of the grouping mechanisms extend into higher levels.

The next several levels in our hierarchy can be losely compared to the template theory [40] of shape recognition. Simply, there exist memories of prior retinal stimulation patterns and they can be used as templates. Superimposing a memory (template) on the incoming pattern of activity generated by feature detectors or their Gestalt groupings would result in a match in the presence of the expected shape. Problems due to subjective contours, partial matches due to changes in distance etc., have been already addressed at the lower levels. For templates to be useful, they must be manipulated in 3D in order to account for missing surfaces of rotated 3D objects that are projected onto 2D retina. The information as to which direction to rotate can be precomputed from the response of feature detectors [41]. Also, before templates are applied some context information must be analyzed [42] and the figure-ground segregation must be completed to some degree. These two functions are closely related and are in part responsible for the so called image segmentation problem which remains unsolved.

Finally, at the highest levels of the recognition hierarchy, we assume the existence of a symbolic scene representation that is the precursor of a verbal description. The AI community has successfully championed a localist symbolic approach for high-level visual reasoning and problem solving. Our symbolic representation differs from traditional AI by being distributed across the dynamic patterns of neural activity. In other words, visual information from the real world is modeled in the excitation patterns coming from widely distributed, lower-level neural pathways. Using the convergent properties of self-organizing, adaptable neural architectures, symbols are distilled from the incoming patterned activity to produce a structural and functional description of the scene. The components of this symbolic representation come from lower levels and include categories of templates or short-term memories of perceived objects, expressed in a behavioral context. Long term memory is the repository of distributed symbols measuring meaningful perceptions, each a stored representation of the activities corresponding to the patterned activity of lower modules. Operationally, the highest level uses symbols residing as prior memory patterns to modulate the activity of neural structures and set up expectations of incoming data. Behavioral motivation regulates the combining of distributed memory primitives to assemble complex expectations and procedures in support of higher-level visual tasks such as goal directed scanning, sorting, and occluded object recognition. In this approach, perception is the byproduct of matching the current state of neural structures to an expectation [43].

4.1 Control and hypothesis formation

In case of a goal-directed visual task, the processing within GPV could begin with the acquisition of the model of the expected object from the world knowledge base. Model-directed control refines initial hypothesis and some instance of it perhaps reconfigures the sensory system in order to find image-based evidence for the objects. Continuous refinement of hypotheses can be used to make predictions about the next data sample. As new sensory data is acquired, expectancies (current best guess) can be updated [44]. The expectancies derived from 3D models can be used to generate predictions about 2D projections of objects which in turn predict the expected image features. Most of the "high-level" processes are couched in AI terminology because the neurosciences can not at the present time specify the location or the architectures of neural structures underlying hypothesis formation or other aspects of model-driven control. Inferotemporal cortex cells, known to display longer response latencies are perhaps involved in hypothesis formation [42].

Since the system consists of many knowledge sources which specialize in searching/interpreting objects and events in environment, we need a mechanism that allows the integration of their outputs into a coherent solution. Various forms of "winner take all" networks have been proposed. Symbol-based approaches to vision use blackboard mechanisms [45,46] to form hypotheses about potential objects and or relationship between them. In biological systems such notions might be based on feedback loops from a higher level to a lower level, although not all top-down pathways are well traced and understood. One conceivable principle of operation involves the existence of a feedback loop between function that is computed and the neural structure/architecture that computes this function. Expectation from higher levels act as templates which modulate feedback loops thus filtering out only relevant information.

Not all of the control issues pertain only to top-down information. Some control can be resolved locally by specifying rules of interactions between neuronal elements. Since the sensory system encodes relative information, the signaling is done as "above" or "below" an average level. For a system to avoid saturattion requires the balancing of inhibitory (OFF) and excitatory (ON) channels. This dichotomy can also provide a phase information. The redundancy of information is also dictated by the inability of a neuron to generate negative spikes however, this argument does not apply to graded potential cells. Other local control strategies include time-limited "winner-take-all" networks, automatic gain control via lateral inhibitory pathways, and others.

4.2 Model manipulation and matching

Matching models with incoming and possibly incomplete image data entails the ability to manipulate object models. GPV must possess mechanisms with which to maintain and manipulate hierarchical descriptions of object models. The object models have been represented using semantic networks with nodes representing objects and links representing constraint relations [47]. This approach is repeatedly used in many of the existing systems and it has an intuitively obvious mapping onto neural networks. Although, it is not clear whether semantic nets are the best representation for visual information, especially at the early and middle levels of processing, symbolic nodes could perhaps be implemented as small neural networks. Even in this case some of the nodes (neural nets) must be able to represent various aspects of objects in different contexts.

Matching models to data has been attempted in the past using rule-based systems. This is perhaps appropriate at higher levels of processing. Neural networks can be set up as a rule-based system where connectivity represents a prediction graph depicting expected objects or their representation in terms of some primitives. Incoming data then activates all neuronal feature detectors, but those that have been primed by the signal from the prediction graph generate highest level of activities. Similar notion has been introduced as "spreading activation" models. Highest activities in turn signify the correct matches. Thus matching is reduced to looking for the maximum cross section between activities in neuronal subgraphs representing image data and predicted object features. It is conceivable that some variation of this type of matching mechanism operates in a distributed fashion throughout all the levels of visual structure. In other words, most of the matching is done perhaps on the local scale using feedback pathways and error detection implemented by networks of local, graded-potential neurons. This would mean that in our functional model there is no identifiable module responsible for matching but some abstratcted results of all matches are recognized as compatible with expectancies derived from the long term memory.

A distributed matching mechanism would help to avoid many difficulties encountered by traditional AI when attempting to solve the problems of recognition, hypothesis formation, goal seeking, belief maintenance, etc. In all of these problems final interpretation of the scene depends on matching to establish correspondence between various representations. The goodness of a match depends on number of features, labels and attributes that are available for matching. Since we lack a strict definition of a shape, object or a scene, exhaustive search of matching features can lead to combinatorial explosion. Hence, in traditional AI, to speed up the computation of correspondence, matching tends to be based on very few key features. In general, it has not been possible to enumerate all such features for all objects. However, using local matching within the distributed neural network, it could be possible to discover these features by abstracting from more primitive attributes. The solution to the category formation problem might be crucial in this task.

4.3 Adaptation, learning and categorization

Our visual experience is not just the perception of things but rather the perception of categorical meanings such as cars, birds, flowers etc. Some aspect of recognition perhaps begin at the "basic level category" [48] when a prototype is selected from initial instances of data. In general the problem of prototypes remains unsolved (but see Grossberg [49]. Simply, we dont understand which critical features of shape or objects constitute prototypes. Hence, the difficulty in data-driven selection of prototypes that would represent categories of objects. Another problem with forming categories is that even if we knew which features/parts should be selected first we still dont know how should they be organized? Some of these selections are performed by feature detectors and other architectural structures of the neural network. In terms of neural structures, categorization is a process of adaptive selection of critical features or patterns of activities which are then stored in short term memory. These patterns of activities are then recognized by matching them with expectancies. Familiar entities result in successful matches while novel entities must be formed (via learning) into new categories of knowledge so that new experiences (i.e., new objects) can be retained for future use.

GPV should have the ability to continually adapt its behavior through its interactions with the environment; we cannot synthesize a system with a priori knowledge of all possible scenarios that it may encounter. This implies the ability to learn. Consequently, perception can not just simply be a filtering of incoming data but must be an active process of continuously matching incoming data to internally represented expectations. This implies that GPV must be at times data-driven to permit instant response in certain situations, while at other times goal-driven to permit the execution of requested tasks. ALVEN [50] is an example of a system where the search through its knowledge base composed of object classes and relationships between them can be: goal-directed; model-directed; and/or data-directed. In other words, a GPV has to be able to select a prototype from input data, as well as descend down the categorical hierarchy under top-down control to instantiate the expected perceptual data.

Most of the existing neural models of learning require external teacher. These include Rumelhart et al "Back-propagation", Widrow's "Adaline",and Fukushima's "Neocognitron". However, biological systems, especially during early stages of development, can perceive patterns (see section on Shape Perception) before they are capable of recognizing or communicating with a teacher. Hence, self-organization must be one of the principal rules of learning and adaptation [51,52,7,53,54,55]. The self-organization and self-regulation are well recognized properties of the brain that have not been captured in any form by any existing machine vision system. The mechanisms underlying such ability derive from plasticity of the neural nets which in absence of external teacher must involve some internal controls such as adaptive error control, sensitization based on neighborhood activities, relative instead of absolute encoding, redundancy of identical information in various representations and Hebbian rules of synaptic interactions. The organizing principles are probably identical throughout the neocortex, which displays almost identical coarse architecture from one functional module to another. The flexibility of the cortex to develop feature detectors in response to incoming sensory information delivered to any specific part of the brain during some critical period has been confirmed by Sur et al [56].

How could an existing neural structure respond to novel patterns? Network plasticity, feedback and short term memory seem to underly this ability in biological systems. Patterns of activities generated by a hierarchy of feature detectors can activate dedicated neural structures (AIT or PIT) in a specific manner. Some of the neurons in these structures would respond in general to the presented input and because of inhibitory links they would inhibit other neurons over the duration of their refractory period. After this time other sets of neuron could respond to the persisting input and so on. This implies that general, bottom-up rules on interconnectivity and synaptic interactions lead to modes of activities in the net that represent the input pattern. These rules could include, final number of synapses per neuron, local memory through long term potentiation, coincidence of synaptic inputs from different neurons, gating some synaptic pathways by signals from long-term memories (expectancies), exponential decay of neuronal potential, time-limited "winner takes all" balance between inhibitory and excitatory inputs and others. In our scheme, over the course of stimulus presentation, various spatial sets of neurons would respond to the input. This spatio-temporal pattern of activities could be at the basis of categorization of the inputs. The final result would be that a network learned to categorize and input pattern by having a set of neurons that consistently display higher activity to this pattern.

Another way to imagine the process of categorization is to picture the pattern of activities in a sheet of neurons. In absence of input the activity is on the average unifrom. In presence of an input many neurons initially respond in a non-unique fashion. This would cause spatial, rounded-hill like bumps in the activity profile of the sheet. Because of various synaptic mechanism, in time, some of these bumps would disappear while in other places more pronounced, sharp-peak profiles would appear. The initial rounded bumps represent general categorization while later specific peaks represent detail recognition of input. This is consistent with the notion that context analysis must preceed final recognition [42].

Multiple presentations of the same input could lead to architectural changes in the net and perhaps short term memory. Thus few presentation of the input could lead to very detailed categorization that is temporarily stored in the form of structural changes of the net [57,58]. A neural mechanism that could support the requirement of memorizing input, presented only a few times in association with other neural activites is long term potentiation. A network of neurons displaying LTP is also the most probable structure underlying categorization [59]. LTP is a cooperative mechanism that shows persistent increase in synaptic efficacy. LTP can be induced rapidly and it is known to outlast other forms of synaptic enhancement. LTP can be found in many areas of the brain, it can last for weeks [60], appears to be the result of increased synaptic contacts [61], and can take place only when presynaptic activities are associated with postsynaptic depolarization [62].

Finally, a GPV must be able to learn and categorize object based on limited number of presentation [58]. This suggests that GPV must be able to analyze context, in parallel with or even perhaps before the final categorization of the input stimulus. It does not seem that a computation of a context is a trivial problem and possibly it requires computing resources comparable to the ones needed for the analysis of the target. This problem of maintaining correspondence between the data and its modelbecomes clear when considering recognition and categorization of a moving object; besides the possible change in the appearance of an object due to rotation, translation, foreshortening etc., its background from one frame to another might also undergo a drastic change. For these reasons it seems desirable to have a match based on a few invariant, critical features. However, since we cant enumerate all such critical features under all possible conditions, perhaps some form of adaptive learning is involved in selection of category prototypes.

4.4 Context analysis

One of the principles of processing is that visual system seems to incorporate a primitive but general function that permits the active investigation of a scene's topological organization before the analysis of details. When visual conditions are poor, figure-ground discrimination is easier to perform than the extraction of detailed shape information. This is consistent with the view that perception of details is easier in presence of context. As such, context must be computed before the attention to details [42]. The context might include orientation of the figure which in turn affects the perception of its 3-D surfaces. In tasks that depend on texture discrimination local processes might be used to find discontinuities [32] but to complete the object's boundaries the system seems to utilize global configuration clues; detailed featural analysis of a shape takes place after top-down processes uncover the global configuration of a pattern. Why is this global structure so important? It seems that this information must be available to preattentive processes so that selective attention may be used in decisions such as where to position the next saccade, when the system must attend to and at least coarsely analyze the image structure outside of the fovea in order to decide the future foveal fixation points. It is possible that some global analysis of the image, including perhaps figure-ground discrimination is performed by the low-resolution, phylogenetically older, pulvinar-striate cortex path [63]. The geniculo-striate pathway, phylogenetically more recent, is dedicated to high-resolution analysis of the shape, once the target has been selected. This implies that perceptual features are hierarchically organized and processed into objects beginning with figure-ground discrimination that perhaps drives selective attention. This is followed by the analysis boundary information and finally details existing inside of the boundaries.

Perception must involve processes such as discrimination, identification, matching; all performed in some motivational context. The context derives from correlates of expectations and resulting actions. It is not difficult to imagine that one part of the perceptual system must be dedicated to internal representation of the external world. This consists of known cortical areas including V1 through V5, where features corresponding to instances of external world are adaptively organized into some representations. This part of the perceptual system receives bottom-up sensory input and is continuously under top-down control of processes concerned with goals and plans set by the system. A difficult task in analysing perceptual system is discovering what kind of neural structures would support the processes of planning and goal setting. In our approach we subscribe to the idea that some part of the visual system is involved in continuous model building activities of the external world [64]. The activities of planning, task specification and goal setting are part of these processes. Contrary to the traditional practice in the AI community, it is probably incorrect to arbitrarily divide the phenomenon of vision into high-level cognitive processes that can be investigated independently of early stages of visual processing. Meaningfull perception at the higher level implies proper architectures and activities at the lower levels. Higher-levels dont deal with external world the way we see it but they operate on a representation of this world, derived from patterns of activities in some specific neural structures of earlier stages. And the modeling and interpretation that higher levels perform is realized in terms of manipulating activities and connectivity patterns of lower-levels neural nets.

4.5 Attention

At the highest levels, the cooperation between the task specifier, goal generator and planner results in the motivationally selected design of routines necessary for the completion of a task. This "vision executive" [65] controls the focus of attention mechanism for sensory data processing and attempts to verify and extend the sensor-based model of the environment (the scene model). The planner is also responsible for strategy selection, i.e. the ordering of focus of attention rules and monitoring performance of the lower-level processes, such as segmentation. The physiology of these very "high-level" functions of planing and specifying goals and tasks is not known except for the possibility that the frontal cortex might be involved. Attention on the other hand has been analyzed in more detail.

One top down control process underlying attention operates by feedback pathways from higher levels where signal changes are gated by adjusting properties/shape of the receptive fields at lower level neurons [66]. A general principle of operation is that in a free running mode, the system always has some expectancy about the environment. This is one input to the mechanism that controls focus of attention. Any changes in the environment that are not predicted by the expected representation must be attended to. The attention mechanism can be focused on a small region of the environment without loosing data. Simply our world is highly structured, events are highly localized in time and therefore a few "well chosen" (by adaptive learning) samples of the environment can provide all of the information necessary to maintain the match with the predicted model.

The attention mechanism could operate according to two functional principles: it must actively reduce the irrelevant information and it must purposelly direct the focus of attention to next important sample. This is done by changing the resolution of the incomming representation. Minimal resolution that is sufficient for a given task reduces the amount of information to be analyzed. Another method useful in reducing the amount of the processing is the use of variable diameter window of analysis; not all of the scene needs to be viewed at all times. Abstraction in the form of gestalt groupings is an additional measure that further minimizes the amount of data. Finally, attentional mechanisms do not work directly on the raw image but rather on the higher level representations. These might include intrinsic images, perceptual primitives, perhaps short-term memories as well as some integrated representations that include information from other sensory modalities.

Variable diameter analysis window principle is based on the comparison of size of the viewed object size versus the size of selective attention field [67,68]. Analogous to a variable diameter spotlight beam, focus of attention can be narrowed to view only one object with very high resolution or to view the whole field at low resolution. A small object can be viewed by the combination of appropriate receptive fields. Active exploration (scanning) is activated only when size of an object exceeds the size of the high-resolution area at the center of gaze. During active exploration information about shape could come from the centers which control eye motion by integrating information about successive foveations ("where I was"). However, it is possible to redirect visual attention while eyes remain stationary. These attentional mechanisms for analysis for visual patterns are possible in the context of different strategies such as: 1) discriminatory learning of pattern cues, 2)context driven selection of cues, and

3)selective attention to cues regardless of background or visual space. First two strategies have been localized to temporal lobe while the third seems to be associated with parietal neurons [69]. It is conceivable that there are some contributions due to a secondary, phylogenetically older pathway for processing form that involves Superior Colliculus to the Pulvinar to Inferotemporal Cortex.

The focus of attention module manages the task of selecting the particular visual target or data against which the knowledge is to be matched, and determines the ordering of knowledge rules for various regions of analysis. The selection of targets must be done in a proper coordinate system with respect to the spatial representation of the environment. The GPV should be able to perceive an object in environmental coordinates. This means that in some situations a GPV on a mobile platform must be able to discount image motion due to retina-motion or self-motion. Hence, a subsystem is needed that translates the information from retinal coordinates to egocentric coordinates and finally into environmental coordinates. A mechanism of selective attention based on target motion or spatial position seems to be closely related to the physical characteristics of the stimulus. These functions are computed by a dedicated module that attends to "extrapersonal" space. Selective attention based on the content of the stimulus (foveal attention) seems to be a more difficult problem that requires at least partial prior solution to shape recognition and context analysis.

How should a battery of high resolution sensors be directed to a target and how is it decided what is important to look at? One strategy is to look only at changes in time and space. Static information is redundant. For this reason peripheral vision must be able to detect temporal changes and send information about spatial changes to processing centers that could foveate on the change. It is also important that attention be paid to the most critical events first. Peripheral motion in fronto-parallel plane is less critical then object motion toward the viewer. The top-down control should operate on a preselected model, using the current hypothesis about the scene to direct the next area to view. The strategy here is to initially perform coarse processing and if warranted, follow it up with fine analysis. Thus the foveation decision is in large part a function of bottom-up, low resolution information from peripheral vision that is supplemented by high level information from the planning system in conjunction with short term memory of previous foveations.

What is the best representation to drive the attention system? It must include only relevant information; it cannot be low level image attributes. The information should, for the most part, be created by bottom-up processes, that enhance the areas of some average or minimal complexity. These areas should attract attention, while areas of extreme complexity should be omited. Additionally, features that "stick out" should be attended to, such as yellow banana in a bowl full of blueberries. Features that are completely novel and dont match any existing models or appear to be out of context, such as seal sunbathing in the middle of Sahara desert should have priority of attentional mechanism. Under some conditions such as while concentrating on the task at hand, irrelevant peripheral information should be suppressed by higher centers so as not to distract from the central problem. This is perhaps controlled by the Limbic system which modulates the general level of arousal. Another suppressive strategy is habituation which allows a system to discount repeated events while simultaneously drawing attention to rare occurences.

Many models of attention have been studied in the past [68,70,71] and our model no doubt incorporates various features from these proposal. However our most important, distinguishing feature is consistency with the neurophysiological findings that context must be processed before attention to details. It is proposed that at any time, the surrounding environment is represented as a spatial map at some higher level(s) in parietal cortex. This map is continuously updated with information about past foveations and saccades. These, in turn, are associated with objects (scene) details, as represented in the association areas of InferoTemporal Cortex. The details about shapes, etc. in the IT are integrated with data available in intrinsic images or feature maps of Barlow (see also [72]. Thus, the attention mechanism, driven by input from either short or long term memory, specifies the next location on the spatial map to process and only then are details available from the specific intrinsic maps. Integration of these features and their subseequent matching to some expectancies in long term memory represents the final step of perception (recognition). Of course not all visual tasks require attention. Preattentive processes which produce intrinsic maps are performed in parallel and are not affected strongly by motivation or behavior. Visual attention on the other hand is a top-down process that influences how sensory information is to be processed by determining the view/target priorities. At minimum, this requires short term memory, and cells have been found in the IT that remember behaviorally significant sensory information [42].

4.6 Space perception

From the very start, the visual system seems to be divided into two specialized and parallel channels of processing information peripheral vision, which computes localization of object and provides information that directs the fovea to the target and the foveal system which is concerned with shape recognition. The flow of information is not continuous. There are saccades (see [73]), lasting about 100 ms during which the eye executes the movement to the next target and the information flow to higher centers is suppressed. The target is then viewed for about 2000msec. So that new portion of the visual space is sampled every 300msec. What directs the eye to any particular point in space and how are these samples put together to recreate a percept of an object? These questions remain to be answered but we do know that the sequence of scanning a shape is often very similar for different observers [74]. It is conceivable that each glance produces a partial description of an object in the short-term memory and only after a while can all parts be put together into a unified picture. This would support the idea that objects are classified into categories, implemented as spatio-temporal activities of neural nets [58]. The sequence of samples are integrated perhaps by using information about previous saccades and fixations, the topology of which are preserved in a spatial map maintained within the parietal cortex. Markers from such a map represent information like "where was I", which in many instances may be sufficient to recover shape.

Parietal cortex involvement in space perception includes: spatial relations between objects, movement in space, spatial representation of the environment and command/control of all motor activity (for review see [75]). Neurons in the Inferior Parietal Lobe are active during, looking at, detecting, reaching for, a motivationally relevant object [76]. The posterior parietal area has "motivational" inputs from the Limbic areas, sensory inputs from the association cortex and pulvinar and attentional inputs from the reticular formation that regulates cortical activation according to sleep/awake states. The middle parietal cortex is involved in the integration of multisensory information and memory as well as some verbal processes related to description of spatial relations. The Frontal Eye Fields and Striatum is crucial for visual scanning, visual orienting, exploration with head/eye system and reaching for objects. Area 8 of the frontal lobe is involved in saccadic eye movements. Combinations of Limbic and Sensory inputs permit motivational significance to be associated with sensory events as for example, increased neural activity to danger when being pursued. Finally, the addition of reticular input may permit regulation of vigilance. Thus the distribution of attention is regulated by three representations of extrapersonal space: First, the sensory map located in posterior parietal cortex, second, the motor map for scanning, orienting and exploring located in FEF and finally, the motivational map in cingulate cortex.

There are three sources of information about depth in space: binocularity, stationary cues and motion. Binocular disparity perspective has limited depth range and the reference point is the point of fixation. Motion perspective comes from image motion due to the eye/head system and object motion. Eye movement is important for searching of the scene, despite the fact that the same point can be scanned bydifferent combinations of receptive fields. Head movement, on the other hand, adds motion perspective information; objects closer to the retina move across faster than objects further away. Stationary cues about perspective come from sources such as textural segmentation where surfaces have a definite texture gradient, as for example ground receeding into distance. This information about perspective is of paramount importance because almost all projections on the retina are subject to perspective transformation. However, if we are viewing a flat photograph, we can still perceive the perspective although motion information and the binocular inputs dont tell us anything about perspective. In this case, all of the information about depth comes from local cues, derived from the arrangements of objects/surfaces in the picture. Since the system can compute correspondence from motion information, it is possible that binocular vision is a recent neural structure for very precise computation of point by point correlation between two images and only within very limited range of depth.

One of the principal problems in space perception is to compute motion/location of objects in their proper spatial relationship to the rest of the scene. The visual system seems able to detect point to point correlation between images to give local apparent motion. There is also evidence for another strategy which avoids pixel by pixel correlation; distinguishing features such as boundaries are used to capture pixels belonging to otherwise featureless regions [77]. The motion aftereffects can be made contingent on color, intensity, pattern, etc. Instead of processing absolute values the system operates on relative values by utilizing various forms of simultaneous and/or successive contrasts. The former operates on qualities that can be compared/differentiated in space at one point in time; enhance edges, compress dynamic range of the stimuli, remove redundancies, encode efficiently. The later operates in the temporal domain and can be demonstrated through various aftereffects.

A depth analysis module included in our GPV system, has inputs from the primitive Feature Processing Module, from the Motion Form Analysis and Illusory Contours and it can compute range information by integrating input sources from "Depth from X" (X = texture, motion, shading, occlusion) and binocular vision. It is not clear how the information from various sources can be best integrated but some abvious strategies are available from examples of neural interactions in other areas of the nervous system. Noisy and weak signals among many neurons/modules can be agonistically averaged thus eliminating random noise and enhancing signal. Weaker signals such as depth from shading can be inhibited by a strong signal from the stereo neurons. Ambiguity about depth between conflicting signals from other sources such as motion and texture can be resolved locally by information from occlusion and/or stereo. Synaptic transmission carrying a signal about depth from, for example, texture could be potentiated by similar depth information from other source like motion.

4.7 Shape perception - perceptual organization and segmentation

The ability to recognize, classify and identify objects from projections on the retina is a process that develops over time and it involves learning and structural changes to neural architectures. For example, one month old infants prefer to look at grating patterns, and this preference changes after two months to bulls eye pattern [78]. With time, more and more complex patterns are prefered, which implies increasing ability to process details with higher spatial frequencies. The principles behind the development of mechanisms that underly perceptual organization, i.e. the organization into meaningful segments of an image are not well understood but it appears that most of the Gestalt principles are fully developed within one year after birth. For example proximity (neighboring element most-likely represent the same surface) develops at seven month of age, common fate (neighboring elements that move in the same direction belong to the same rigid(?) body) is present at one month, subjective contour perception is detectable at four months[79] and symmetry (grouping elements that are symmetrical) is detectable at five months [80]. This suggests that perceptual organization is not innate but must be learned through interactions with the environment. Similar conclusion can be extended to generalization. For example, the ability to perceive constant shape of varying size develops within the first year of life [81]. Furthermore, this implies that transformation from the viewer center coordinates to an object centered representation, is a necessary prerequisite to discount object changes as the viewer moves around the environment.

Object recognition is one of the fundamental tasks in perception and yet we lack an acceptable definition of an object. Various proposals for representing objects are abundant however [16,12]. Most representations developed within Computer Vision are not general enough to allow easy description of sculpted surfaces. One exception is so called Surface Boundry Representation (SBR) [12] which conveys the information about a 3D surface in the form of triangle faced polyhedrons. More complex techniques using quadrics and higher order polynomials have also been developed. It is conceivable that some variation of this representation would be well suited for neural network architecture.

It appears that we easily categorize perceived things although it is not clear how categories are formed. One intuitively obvious approach is to define members of a category by their parts and their spatial relationships. How parts are joined together to form a notion of an object? This problem is not just image segmentation resulting in different, coherent regions of the scene. Grouping parts into objects must be somehow guided by meaningful relationships between parts that may or may not be distinct regions of an image. It is possible that this process is partially guided by expectations. In this case recognition of one part of an object could direct the attention to another expected and meaningful part. Another strategy is to complete the analysis of the context which perhaps contains information about the regularities or constraints of our world. These in turn could direct the grouping of parts into meaningful objects.

Knowledge about the physiology of shape perception is limited to early stages of visual processing up to V1. At higher levels our knowledge becomes more general. For example, we know in general that lesions of upper parts of V2 and V3 degrades pattern discrimination and visual acuity, while lesions in lower part of these structures affects recognition of patterns and objects. However, we lack the detailed knowledge of receptive fields and synaptic interactions at this level. We know that at the level of V4 and V5 retinotopic mapping is encoded in the cell response. All higher level areas are specialized for nontopographic processing of separate attributes of the image [82]. Some form of linking based on similarity of information in texture, color, motion, collineraity, disparity, figure/ground and others must also take place in nontopographic representation. The underlying principles of linking are not well understood. Linking is important in the segmentation process because often, specific intrinsic images might not have enough information to be interpretable.

The fundamental problem of segmentation is that it cannot be considered as only a bottom-up process or as only top-down process. Segmentation seems to involve both strategies, continuously interacting and penetrating each other to different depths, depending on the task. Can local information be sufficient for segmentation and interpretation? It is probably sufficient for partially guiding segmentation. To have the interpretation we need also to include global information. The difficulty is that except for trivial cases, it is not clear what makes up the global information is and how to compute it. For example, segmentation based on the similarity of features within a region can be completed to a degree using only local information. However, it is not completely clear how similar the regions must be. Considering camouflage, similarity is not equivalent to identity. Hence, the question of which similarity parameters are most important in any specific situation might be guided by some global information.

Robust image segmentation will also reduce potential errors in higher level processes like planing and matching. Image segmentation can be based on cooperative/competitive relaxation algorithms [83] applied to all image attributes as well as to integrated representations of intrinsic images. The segmentation process must take advantage of all available top-down strategies based on applicable knowledge. Some initial plans can be generated by bottom-up, coarse region segmentation. Such a plan could produce a set of large areas within each intrinsic image that become refined by integration with information from different image attributes. Eventually, a top-down process, initiated by the focus of attention module, segments the scene into background and target objects which can be examined for detailed structure in the context of large patches that might have already been interpreted as background.

4.8 Constancies and generalization

Why are constancies important? All of the listed problem can be simplified if the system has the ability to deal with color, motion, shape, size, and lightness constancies. In some respect we can view constancy mechanisms as precursor of generalization. For example light and color constancy helps to generalize across variations of illuminants. Size constancy allows to generalize across varying shape sizes. Having constancy built into neural network reduces the complexity of object recognition and minimizes storage requirement. The implication of this is that the process of categorization is simplified. It is conceivable that this form of generalization applies also to other attributes of shapes such as motion and texture. The concept of constancy also applies to shape but at a higher functional level. Thus we are able to generalize across the birds or vehicles despite their often drastic differences in details. It is conceivable that at this level the constancy might apply to spatial relationships among parts of the category members. In all of these cases the underlying concept is that constancy is a form of generalization that discounts the variation in the object caused by environmental changes. It seems intuitively clear how to implement some constancies with neural elements for example lightness [84] and color constqancy [35]. The basic principle here is to enhance and compare variation that are above or below the average of the neighborhood. Shape constancy however, is a difficult problem that might require cooperation among many complex distributed processes, including memory in order to produce the final percept.

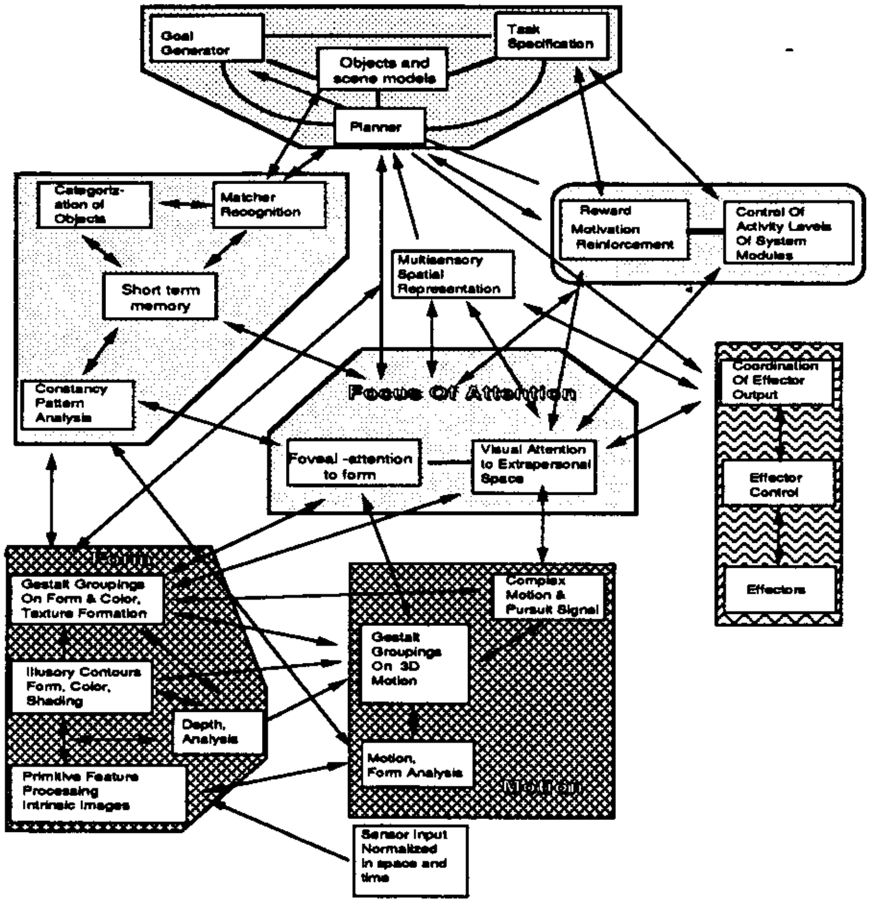
Goal Generator
Task Specification
Objects and scene models
Planner
Categoriz- ation of Objects
Matcher Recognition
Short term memory
Constancy Pattern Analysis
Multisensory Spatial Representation
Reward Motivation Reinforcement
Control Of Activity Levels Of System Modules
Focus Of Attention
Foveal -attention to form
Visual Attention to Extrapersonal Space
Coordination Of Effector Output
Effector Control
Effectors
Form
Gestalt Groupings On Form & Color, Texture Formation
Illusory Contours Form, Color, Shading
Depth, Analysis
Primitive Feature Processing Intrinsic Images
Complex Motion & Pursuit Signal
Gestalt Groupings On 3D Motion
Motion, Form Analysis
Motion
Sensor Input Normalized in space and time

Figure 5: Proposed Framework for a General Purpose Machine Vision System

5 Conclusions

The goal of developing a *general* purpose machine vision system has been decomposed into the series of subgoals. The analysis of existing computer vision systems elucidated not only the lack of truly robust machine vision, but also the need for a good working definition of a general purpose vision system. This analysis also demonstrated the need for collaboration between the neurosciences, computer science, and psychology. We have looked at the human visual system for hints regarding the underlying mechanisms necessary for the development of general purpose vision. A collection of visual tasks was generated and whenever possible, their corresponding neuronal substrates were noted. Although this list is far from complete, it serves to illustrate the tremendous scope of problems that a general purpose vision system must not only address but also solve.

Starting with existing data from functional neuroanatomy, we synthesized a prototype framework for a general purpose vision system. This framework was then used to synthesize a more elaborate functional block diagram for a general purpose machine vision system (see Figure 5). Two strongly shaded areas are emphasized because we know more about them from neurophysiology and we also have some understanding of how to model their functionality in computer science. However, other areas are less well known. For example, categorization, constancy, matcher/recognition, and short-term memory, are perhaps equivalent to visual areas like the posterior inferotemporal cortex (PIT) and the anterior inferotemporal cortex (AIT). A difficulty that remains in this attempt to synthesize a general purpose machine vision system is to determine where to place the homunculus that will finally decide what is being seen. Currently, our homunculus sits in the boxes labeled goal generator, task specification, and planner. These functions have received more attention in AI studies and very little is known about their physiology. Some of their aspects may be analogous to parts of the frontal cortex and the limbic system. Our model is evolutionary and will improve with the collection of new experimental data from the simulations in conjunction with a continuous analysis of neuroscience literature.

Having a specification for a general purpose vision system will hopefully permit us to develop successful evaluation methods that will perhaps be used as standards or guidelines for proposing new machine vision systems. It is clear that such a development of such specifications will require collaborative efforts among neurophysiology, neuropsychology, computer science and cognitive psychology.

Acknowledgements

Support for the UCLA Machine Perception Laboratory environment is provided in part by generous grants from IBM and Hewlett Packard. We sincerely acknowledge support by ARCO-UCLA Grant #1, MICRO-Hughes grant #541122-57442, ONR grant #N00014-86-K-0395, ARO grant DAAL03-88-K-0052 and PMTC-ATI grant #N00123-87-D-0364. Special thanks to D. Gungner, E. Mesrobian and other students of the MPL for critical reading of previous versions of this manuscript.

References

[1] S. M. Kosslyn. *Toward a Computational Neuropsychology of High-Level Vision.* Technical Report NTIS Order No. AD-A145711, National Technical Information Service, 1984.

[2] S. M. Kosslyn. *Ghosts in the Mind's Machine.* W. W. Norton, New York, 1983.

[3] S. M. Kosslyn and S. P. Shwartz. A simulation of visual imagery. *Cognitive Science*, 1:265–295, 1977.

[4] D. Marr. *Vision.* Lange Medical Publications, Los Altos, California, 1 edition, 1982.

[5] D. Marr. *Vision Research*, 331–341, 1976.

[6] L. Harmon and E. R. Lewis. Neural modeling. *Physiological Reviews*, 46:513–591, 1966.

[7] R. Linsker. From basic network principles to neural architectures: emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci. USA*, 83:7508–7512, 1986.

[8] J. Hochberg. Machines should not see as people do, but must know how people see. *Computer Vision, Graphics, and Image Processing*, 37:221–237, 1987.

[9] G. Baumgartner R. von der Heydt, E. Peterhans. Illusory contours and cortical neuron responses. *Science*, 224:1260–1262, 1984.

[10] C. C. Gotlieb. *The economics of computers: Cost, benefits and strategies.* Prentics Hall, New York, 1985.

[11] M. Wertheimer. Untersuchungen zur lehre von der gestalt. *Psychologische Forschung*, 4:301–350, 1923. Translation in A Source Book of Gestalt Psychology, W. D. Ellis, ed., New York: Harcourt, Brace, 1938.

[12] D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice Hall, New Jersey, 1982.

[13] A.P.Witkin and J.M. Tenenbaum. On the role of structure in vision. In B.Hope J.Beck and A. Rozenfeld, editors, *Human and Machine Vision*, pages 481–544, Academic Press, New York, 1983.

[14] V.Torre and T. Poggio. *On edge detection.* Technical Report A.I.Memo 768, Masachusetts Institute of Technology, August 1984.

[15] K. Ikeuchi. Recognition 3d objects using the extended gaussian imge. In *Proceeding of the 7th International Joint Conference on Artificial intelligence*, pages 24–28, 1981.

[16] P.J. Besl and R.C. Jain. Three-dimensional object recognition. *ACM Computing SurveysVision*, 17:75–145, 1985.

[17] R. B. Ekstrom, J. W. French, H. H. Harman, and D. Dermen. *Kit Of Factor-Referenced Cognitive Tests.* Educational Testing Service, Princeton, 1976.

[18] 1985.

[19] J. G. Chusid. *Correlative NeuroAnatomy and Functional Neurology.* Lange Medical Publications, Los Altos, California, 19 edition, 1985.

[20] J. K. Roberts. *Differential Diagnosis in Neuropsychiatry.* Wiley, New York, 1984.

[21] Feinberg and Jones. Object reversals after parietal lobe infarction. *Cortex*, 261–271, 1985.

[22] D. Grossi, A. Orsini, A. Modafferi, and M. Liotti. Visuoimaginal constructional apraxia: on a case of selective deficit of imagery. *Brain and Cognition*, 5:255–267, 1986.

[23] D. F. Benson, J. Segarra, and M. Albert. Visual agnosia-prosopagnosia, a clinicopathological correlation. *Archives of Neurology*, 30:307–310, 1974.

[24] A. R. Damasio, H. Damasio, and G.W. van Hoesen. Prosopagnosia: anatomic basis and behavioural mechanisms. *Neurology (NY)*, 32:331–341, 1982.

[25] D. H. Hubel and T. N. Wiesel. Shape and arrangement of columns in the cat's striate cortex. *Journal of Physiology (London)*, 165(3):559–568, 1963.

[26] D. C. Van Essen. *Functional Organization of Primate Visual Cortex*, chapter 7, pages 259–329. Plenum, New York, 1985.

[27] D. C. Van Essen, W. T. Newsome, J. H. R. Maunsell, and J. L. Bixby. The projections from striate cortex (v1) to areas v2 and v3 in the macaque monkey: asymmetries, areal boundaries, and patchy connections. *The Journal of Comparative Neurology*, 244:451–480, 1986.

[28] W. T. Newsome, J. H. R. Maunsell, and D. C. Van Essen. Ventral posterior visual area of the macaque: visual topography and areal boundaries. *The Journal of Comparative Neurology*, 252:139–153, 1986.

[29] D. C. Van Essen, W. T. Newsome, and J. H. R. Maunsell. The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotopies, and individual variability. *Vision Research*, 24(5):429–448, 1984.

[30] M. Connolly and D. C. Van Essen. The representation of the visual field in parvicellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey. *The Journal of Comparative Neurology*, 226:544–564, 1984.

[31] E. Mesrobian and J. Skrzypek. A connectionist architecture for computing textural segmentation. In *Proceedings of the SPIE Conference on Image Understanding and the Man-Machine Interface*, Los Angeles, California, 1987.

[32] J. Skrzypek and E. Mesrobian. Textural segmentation: gestalt heuristics as a connectionist hierarchy of feature detectors. In *Proceedings of the IEEE Conference of the Engineering in Medicine and Bilogy*, Boston, Mass, November 1987.

[33] U. Neisser O. G. Selfridge. Pattern recognition by machine. *Scientific American*, 203:60–68, 1960.

[34] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's cortex. *Journal of Physiology (London)*, 160:106–154, 1962.

[35] I. Heisey and J. Skrzypek. Color constancy and early vision: A connectionist model. In *Proceedings of the IEEE First Annual International Conference on Neural Networks*, June 1987. San Diego, California.

[36] J. Skrzypek. Lightness constancy: neural network architecture for controlling sensitivity. *submitted to IEEE SMC*, 1989.

[37] J. Skrzypek. *Lightness constancy architecture - theory and simulation.* Technical Report UCLA-MPL-TR 89-5, University of California Los Angeles Machine Perception Lab., March 1989.

[38] E. L. Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, 20:645–669, 1980.

[39] J. R. Pomerantz. *Visual Form Perception: An Overview*, chapter 7, pages 1–29. Academic Press, New York, 1986.

[40] A. Trehub. Neuronal models for cognitive processes: networks for learning, peception and imagination. *J. Theor. Biol*, 65:141–161, 1977.

[41] D. Gungner and J. Skrzypek. A connectionist architecture for matching 3-D models to moving edge features. In *Proceedings of SPIE Conference on Intelligent Robots and Computer Vision*, 1986.

[42] J. M. Fuster. Inferotemporal units in selective visual attention and short term memory. *submitted*, 1989.

[43] D.M. MacKay. *Ways of looking at perception.*, pages 25–43. MIT Press, Boston, 1967.

[44] M. O. Shneier, R. Lumia, and Kent E. W. Model-based strategies for high-level robot vision. *Computer Vision, Graphics, and Image Processing*, 33:293–306, 1986.

[45] M. Nagao, T. Matsuyama, and Y. Ikeda. Region extraction and shape analysis of aerial photographs. In *Proceedings of the 4th International Conference on Pattern Recognition*, page 620, 1978.

[46] M. Nagao and T. Matsuyama. *Structural Analysis of Complex Aerial Photographs.* Plenum, New York, 1980.

[47] D. H. Ballard, C. M. Brown, and J. A. Feldman. An approach to knowledge-directed image analysis. In A. R. Hanson and E. M. Riseman, editors, *Computer Vision Systems*, pages 271-282, Academic Press, New York, 1978.

[48] E. Rosch and B. Loyd. *Cognition and Categorization.* Earlbaum Associates, Hillsdale, N.J., 1978.

[49] S. Grossberg. Competitive learning; from interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63, 1987.

[50] J. K. Tsotsos. Knowledge organization and its role in representation and interpretation for time-varying data: the alven system. *Computational Intelligence*, 1:16–32, 1985.

[51] J. L. Adams. *Principles of Complementarity, Cooperativity, and Adaptive Error Control in Pattern Learning and Recognition: A Physiological Neural Network Model Tested by Computer Simulation.* PhD thesis, University of California, Los Angeles, Dept of Neuroscience, 1989.

[52] G.A. Carpenter and S. Grossberg. The art of adaptive pattern recognition by self-organizing neural network. *Computer, IEEE*, 21:77–88, 1988.

[53] R. Linsker. From basic network principles to neural architectures: emergence of orientation-selective cells. *Proc. Natl. Acad. Sci. USA*, 83:8390–8394, 1986.

[54] R. Linsker. From basic network principles to neural architectures: emergence of orientation columns. *Proc. Natl. Acad. Sci. USA*, 83:8779–8783, 1986.

[55] C. von der Marlsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100, 1973.

[56] P. E. Garrahty M. Sur and A.W.Roe. Experimentally induced visual projections into auditory thalamus and cortex. *Science*, 242:1437–1441, 1988.

[57] J.Ambros-Ingerson R.Granger and G. Lynch. Derivation of encoding characteristics of layerii cerebral cortex. *Cognitive Neuroscience.*, 1:61–87, 1989.

[58] M Baudry G. Lynch, R. Grangner and J. Larson. *Cortical encoding of memory: Hypotheses derived from analysis and simulation of physiological learning rules and anatomical structures.*, chapter , pages 247–289. MIT Press, Cambridge, 1988.

[59] E.W. Kairiss T.H. Brown, P.F. Chapman and C.L. Keenan. Long-term synaptic potentiation. *Science*, 242:724–728, 1988.

[60] R.J. Racine and M. deJong. chapter , page . Liss, New York, 1988.

[61] W.T. Greenough and C.H. Bailey. The anatomy of memory: convergence of results across a diversity of tests. *Trends in Neuroscience*, 11:142–147, 1988.

[62] A.H.Ganong S.R. Kelso and T.H. Brown. *Proc. Natl. Acad. Sci*, 83:5326–5332, 1986.

[63] I. T. Diamond and W. C. Hall. *Science*, 164:251–262, 1969.

[64] H.B. Barlow. *Cerebral Cortex as Model Builder*, chapter 4, pages 37–47. J. Wiley & Sons, New York, 1985.

[65] L. S. Davis and T. R. Kushner. Vision-based navigation: a status report. In *Proceedings of DARPA Image Understanding Workshop*, pages 153–169, Feb. 1987.

[66] S.P. Wise and R. Desimone. Behavioral neurophysiology: insights into seeing and grasping. *Science*, 242:736–739, 1988.

[67] F. Crick. The function of the thalamic reticular complex: the searchlight hypothesis. *Proc. Natl. Acad. Sci.*, 81:4586–4590, 1984.

[68] A. Treisman. Features and objects. *The Quarterly J. of Exp. Psychology*, 40:201–37, 1988.

[69] E. Iwai. Neurophysiological basis of pattern vision in macaque monkeys. *Vis. Res.*, 25:425–439, 1985.

[70] K¿ Fukushima. A neural network model for selective attention. *Biol. Cybernetics*, 55:5–15, 1986.

[71] M.C. Mozer. *A connectionist model of selective attention in visual perception.* Technical Report CRG-TR-88-4, University of Toronto, Computer Science, July 1988.

[72] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. In A. R. Hanson and E. M. Riseman, editors, *Computer Vision Systems*, pages 3–26, Academic Press, New York, 1978.

[73] P Bach-y-Rita and G Lernerstrand. *Basic mechanisms of occular motility and their clinical implications.* Pergamon Press, Oxford, 1 edition, 1975.

[74] D. Noton and L. Stark. Eye movement and visual perception. *Sci. Amer.*, 6:, 1971.

[75] W. Richards. *Visual space perception*, pages 351–386. Academic Press, New York, 1975.

[76] M.M. Mesulam. The functional anatomy and hemispheric specialization for direcxted attention, the role of the parietal lobe and its connectivity. *Trends in Neurosciences*, 6:384–387, 1983.

[77] V.S. Ramachandra and S.M. Anstis. The perception of apparent motion. *Sci. Amer.*, 254:102–109, 1986.

[78] R.L. Fantz and S. Nevis. Pattern preferences and perceptual-cognitive development in early infancy. *Merril-Palmer Quarterly*, 13:77–108, 1967.

[79] J. J. Campos B.I. Bertenthal and M.M. Haith. Development of visual organization: the perception of subjective contours. *Child Development*, 51:1072–1080, 1980.

[80] K. Ferdinandsen C.B. Fisher and M.H. Bornstein. The role of symmetry in infant form discrimination. *Child Development*, 52:457–462, 1981.

[81] H.A. Ruff. Infant recognition of invariant form of objects. *Child Development*, 49:293–306, 1978.

[82] H.B. Barlow. *General principles: The senses considered as physical instruments*, pages 1–33. Cambridge University Press, Cambridge, 1982.

[83] A. R. Hanson and E. M. Riseman. Visions: a computer system for interpreting scenes. In A. R. Hanson and E. M. Riseman, editors, *Computer Vision Systems*, pages 303–334, Academic Press, New York, 1978.

[84] J. Skrzypek. In preparation. 1989.