DETE: CONNECTIONIST/SYMBOLIC MODEL OF VISUAL AND VERBAL ASSOCIATION

Valeriy I. Nenov
Michael G. Dyer

April 1988
CSD-880027

DETE: Connectionist / Symbolic Model of Visual and Verbal Association*

Valeriy I. Nenov
Michael G. Dyer

Artificial Intelligence Laboratory, Computer Science Department,
University of California Los Angeles, CA 90024, USA

Abstract

This paper describes a neurally inspired, computational model of the associative interactions between two types of cognitive functions in humans -- the processing of visual and verbal information. DETE -- a computer implementation of our model, is designed to learn to associate language descriptions of objects moving in a visual field with those objects, and thus explore the grounding problem -- i.e. how language semantics maps to sensory experiences. The model consists of connectionist and micro-symbolic modules responsible for processing of the visual and verbal inputs and interacting through an associative module coupled with a mechanism for selective attention. DETE accepts input from two modalities: (1) visual, which consists of a continuous sequence of visual scenes showing the behavior of simple 2-D shaped objects in a square region and (2) verbal, occasional streams of English sentences describing the everchanging visual scenes. The overt behavior of the system is: language generation and visual imagination. In grounding its primitive symbols in sensory categories DETE differs from pure, autonomous, top-down symbol systems in which the primitive symbols are merely arbitrary, undefined atomic tokens. The relation between the symbolic and non-symbolic sub-systems is a bottom-up complementary one. DETE is being implemented on the Connection Machine (CM2) in *LISP -- a data parallel language extension of Common Lisp, designed to program the CM2.

1. INTRODUCTION

Human information interactions with the environment, such as reading and writing of text as well as visual perception of ever changing scenes, are effectively sequential in nature. On the other hand, the nature of the internal processing of these phenomena is intrinsically parallel and distributed, and is embedded in a highly parallel architecture -- the human brain. The design of systems with externally sequential behavior produced by parallel architectures is a challenging problem [Rume86], [Gall87]. The objective of this research is to construct a learning system which has a set of given cognitive abilities springing from its structural organization, and which is allowed to adapt by being immersed in a quite simplified but realistic environment consisting of visual and verbal stimuli. The system accepts two types of inputs: (1) continuous visual input, composed of series of images (movie frames) containing simple 2-D geometrical objects, for instance circles, triangles or squares, and (2) occasional verbal input, containing descriptions of current, past or future visual scenes, or questions demanding verbal (description of scenes) or motor (manipulation of objects in the scene) responses. The main issues which we encounter in this research include: (1) the learning of language necessary for descriptions of the visual images, and (2) ways the visual system interacts with language learning.

2. TASK

Imagine a square region full of small objects of different two dimensional shapes and sizes moving around with different speeds and directions; bouncing off the walls of the box and with each other.

* This research is supported in part by a Hughes AI Center grant and a contract with the JTF program of the DoD, monitored by JPL.

What is the language that we as humans would use to describe what is happening in this microworld? How can such a language be learned by an Artificial Neural System? The task given to DETE* is to learn to describe this microworld of objects by observing their behavior and listening to verbal descriptions. To delimit the vocabulary that a human would use to describe this microworld, we analyze the various degrees of freedom of the system, i.e. feature dimensions/relationships and the possible values of each. Following is a summary of all possible dimensions/relations and some of the values they can take, given in parentheses.

- Individual object features (relative to the box, obstacles in the box or other object).
  - size (small, medium, large,...)
  - shape (circle, square, triangle, same, different, similar,...)
  - location (up, down, left, right, medial, lateral, touching, close, far, inside, outside, between, behind, in-front, below,...)
  - color (red, green, blue, same, different,...)
  - movement (moving, standstill)
    - direction (North, East, same, opposite, different, parallel, towards, along,...)
    - speed (slow, fast, zero,...)
  - rotation (rotating, not rotating)
    - direction (clockwise, counterclockwise)
    - speed (fast, slow,...)
  - intensity (dark, light, medium,...)
- Set relationships -- additional feature dimensions relevant when we have not one, but a number of objects in the box. These include:
  - number of objects (1, 2, ..., some, few, many, none,...)
  - grouping according: size, shape, movement, etc.
- Time relationships (relative and absolute):
  - present (now,...)
  - future (next, after, later,...)
  - past (before, long-ago,...)
  - modifiers (soon, immediately, just, ...)
- Object interactions. In our box-world we allow currently interactions of the following types:
  - collisions (bounce, enter/exit, stop, push,...)
  - containment (enter, exit,...)
  - relative movement (bypass, slide-along,...)

For the purpose of simplicity, we reduce this feature set. We assume that there are no field forces (gravitational, electromagnetic, etc.) that could effect the movement of the objects. Hence, all trajectories are straight lines which are broken during collisions or rebounds of objects with the walls or between each other. Stationary objects may reside in the microworld, and serve as containers or objects of reference.

Humans use language to describe changes in the environment, or if placed in a new situation, to provide an initial description of it. Once the environment has been described, and if it remains static, we usually do not bother to describe it again and again. Thus, our verbal input to DETE is not continuous but rather accidental. It is up to us -- the "parents" to decide when, what, and how much verbal input to provide. A general strategy for teaching DETE is to test its knowledge along the way so that we make sure DETE has learned what it has already seen, before we present it with more complex situations. It is important to design proper testing criteria. Our advantage, as compared to real humans, is that we have the option to freeze the system and the learning process at any time and, examine the content of what has been learned so far.

---

* DETE pronounced [dɛ'tɛ] stands for "child" in Bulgarian, the native language of the first author.

The world of simple shaped objects in a box, as described above, is narrow but quite rich. Examples of descriptions of visual scenes of increasing complexity are given below:

```
... A small ball is moving towards the left wall...

... It bounced off the wall and is moving up towards a large, dark square...

... They are about to collide...

... The ball stopped while the square started moving...
```

3. DETE's ARCHITECTURE and FUNCTION

DETE is composed of four main modules: (1) Visual subsystem (2) Verbal subsystem, (3) Association Module, and (4) Selective attention module. The network consists of neuron-like elements (nodes or neurons) with relatively simple properties and connections (links or synapses) between them. The nodes have input/output behavior which in many respects is similar to the overt behavior of neurons in the central nervous system (firing properties, membrane potential, threshold, etc). The links simulate synaptic connections between neurons and their main function is to carry information in the form of excitation or inhibition between neurons. The neurons are organized in clusters (nuclei or layers) according to the topological patterns of their natural connectivity and communicate among each other within and between these clusters. The information among neurons is passed in the form of spikes and is encoded in their spatial and temporal frequency pattern. DETE's manner of information transfer is more realistic than transmitting averaged frequency values -- a widely used technique in current connectionist modeling. An overview of DETE's architecture is given in Figure 1.

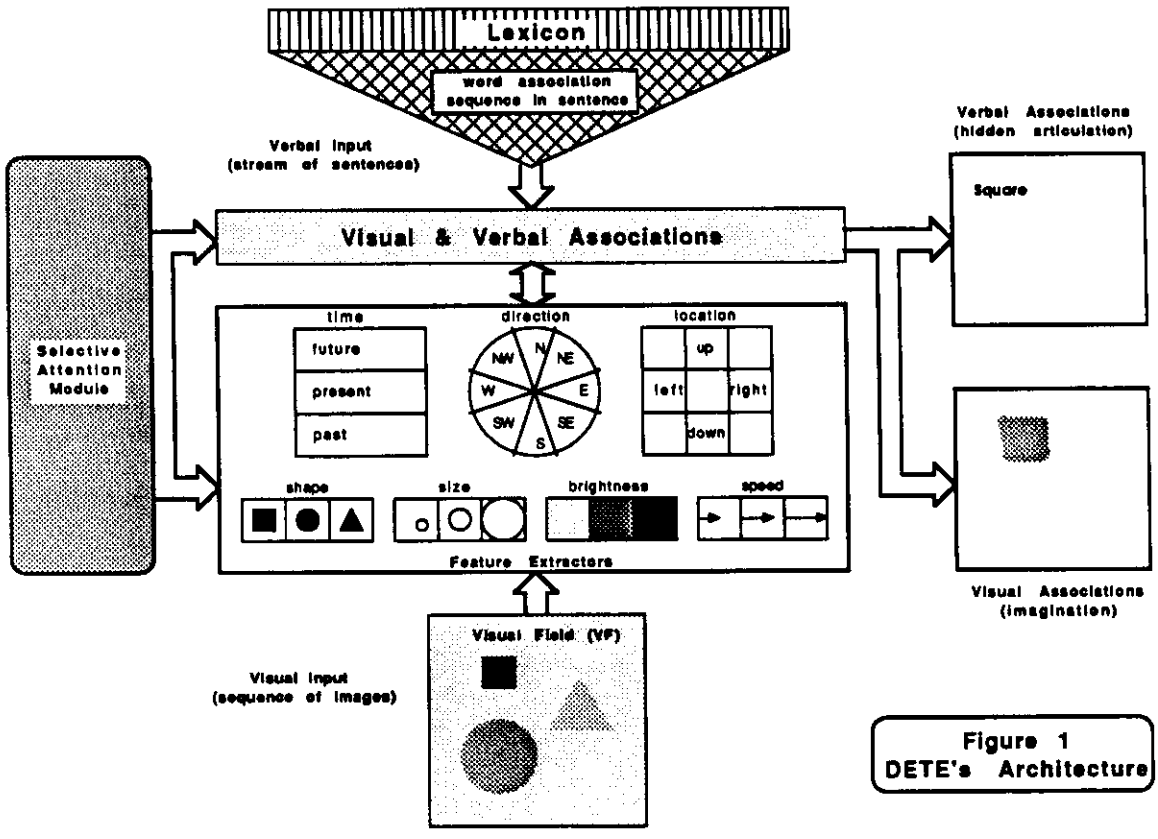

Figure 1
DETE's Architecture

3.1. The Visual Subsystem

Recent findings [Livi88] suggest that visual signals are segregated and processed by at least three separate processing systems in the brain, each with its own distinct function. One system processes information about shape perception; a second about color and third about movement, location and spatial organization. These three systems can be physiologically segregated at the level of the lateral geniculate bodies and after a sequence of semi-independent processing stages [VanE83], they converge at the level of the visual association cortices (VAC). These modules apply a given transformation function (filter) to the incoming signal to extract e.g. the location, movement, and size. Their functions are provided by a pre-wired architecture (inherited connectivity). The rest of the modules in the visual system have dynamic functions caused by modifiable synaptic weights. As a result they are able to learn their functions.

The visual subsystem of DETE consists of an input visual field (VF or retina) the output of which is presented in parallel to several modules (multi-layered connectionist networks) of specialized feature extractors for: shape; size; location; brightness; motion (direction and speed), and time. The retina is a 512 by 512 array of neurons. There are three cell types of retinal cells [Hart37]. These types are functionally characterized as follows: (1) Stimulation of the receptive field produces initial burst of impulses at high frequency, followed by a steady discharge at lower frequency lasting throughout the stimulation (20% of the cells). (2) Short burst of spikes at stimulus onset arise, followed by silence and another short burst when the stimulus is turned off (50% of the cells). (3) Bursts of spikes on stimulus offset appear with a gradually decreasing discharge frequency (30%). The neurons of each of these three cellular types are further subdivided to center-ON and center-OFF subsets. These choices of cellular types in the retina, and the appropriate segregation of their projections to the deeper structures, allow DETE's visual subsystem to segregate the desired feature dimensions. Following are brief descriptions of the feature extractor functions used in DETE.

Segmentation -- figure/ground separation: Before we can recognize an object, "figure" must be segregated from "ground". One must somehow pick out regions that are likely to correspond to distinct objects. A figure must be selected on the basis of physical properties of the input, such as regions of homogeneous color or texture, or contiguous zero-crossings in the second derivative of the function relating intensity to position (which occur at the edges of objects). There are numerous proposals in the computer vision literature for ways of organizing input into regions likely to correspond to figures [Ball82].

Shape: The visual system processes input at different spatial frequency bandwidths. Higher spatial frequencies correspond to more light/dark alternations per degree of visual angle; thus, higher resolution is required to detect higher spatial frequencies. The shape extracting module of the visual system can be described as having a number of different "channels," each differing in resolution. At average viewing distances, the lowest spatial frequency channel produces an output that will often correspond to the general shape envelope of an object. The architecture of DETE's shape recognition module was inspired by Fukushima's Neocognitron model [Fuku80]. This system allowed shape recognition independent of both object position on the retina and shape distortions.

Size: Computation of object size in DETE is based on its ability to learn to classify objects as blobs i.e. depending on the percentage of retinal space they occupy.

Location (position) Variability: A number of mechanisms have been proposed to solve the problem of position variability. Marr suggested that the appearance of objects is stored in object centered representations [Marr82]. In such representations, the locations of parts of objects are specified relative to other parts, not to positions in space. The solution adopted by the primate visual system to the problem of position variability is evident in the neurophysiological literature. It

has been found in primates that the visual cells in area TE (near the anterior end of the inferior temporal lobe) have very large receptive fields, and respond when patterns are present over a large range of positions. This area of the brain has been shown to be critically involved in recognition per se [Mish82]. Thus, primates rely on not representing the position of a pattern in the high-level shape representation system. One implication of this solution is that only one shape can be recognized at a time (although we could rapidly switch back and forth between stimuli, only one would be processed at any given instance). If multiple stimuli are being processed simultaneously, the large receptive fields would result in a system's not being able to tell if there is one stimulus or two of the same stimuli being presented in different locations. Hence, figure/ground segregation is necessary to isolate individual objects before they can be processed further.

Motion perception: Most of the literature on neural networks emphasizes spatial computation in networks by modifying weights. An important view upon neural computations is the relative timing and phase in neural networks. The information about visual motion in primates at the level of the retina is represented as time differences in the firing pattern of axons in the optic nerve. At the level of the visual cortex, the relative timing information is used to drive cells that respond best to edges that are moving in particular direction [Koch85]. DETE is being designed to use the information conveyed from the type 1 and 3 retinal cells to compute the motion of the objects.

3.2. The Verbal Subsystem

DETE is being designed to learn the syntax of the language used to describe the microworld, as well as its semantics. The words in the language serve as the primitive atomic symbols of a symbolic system that can be further elaborated with systematic compositionality, explicit rules, etc. Each word-node in our dictionary corresponds to one English word or syntactic delimiter (whitespace, comma, period, etc.) used in the description of the visual scenes. Further along, each word node is connected to any other word node through a matrix of hidden unit and thus, our neural architecture allows each word to be temporally associated with any other word within a sentence. Initially we will consider that during the time when a particular sentence is being processed, subnetworks corresponding to the particular words in the sentence are activated and the time courses of their activation overlap in time. Associations are made between the words and the corresponding word meanings in the association cortex. As soon as the input of a sentence comes to a stop, the previously activated word-networks are inhibited and the processing of a new sentence can be initiated. In Figure 2 we show how three different sentences can be encoded. Word nodes at the top level (an ordered sequence) are clamped sequentially. Each pair of nodes is connected to a 2 by N matrix of nodes where N is the maximum number of words in a sentence (can be chosen arbitrarily). Which column of each matrix (left or right) gets activated depends on the order of the words in the sentence and the sequential order of word entries in the layer of word nodes. Which node in a column gets activated depends on the sentence position of the word pair.

For instance, the sentence "Move small square left." causes the following sequence of nodes to be clamped: (1) The right top row node of the matrix common to the words "move" and "small". (2) The left, 2nd row node of the matrix of the pair "small" and "square". (3) The right, 3rd row node of the matrix between "square" and "left". This type of representation allows us to encode, in a pairwise manner, the word order in a sentence as well as the position of the pair in the sentence. The nodes in each word-pair matrix project to the layers of neurons in the association cortex.

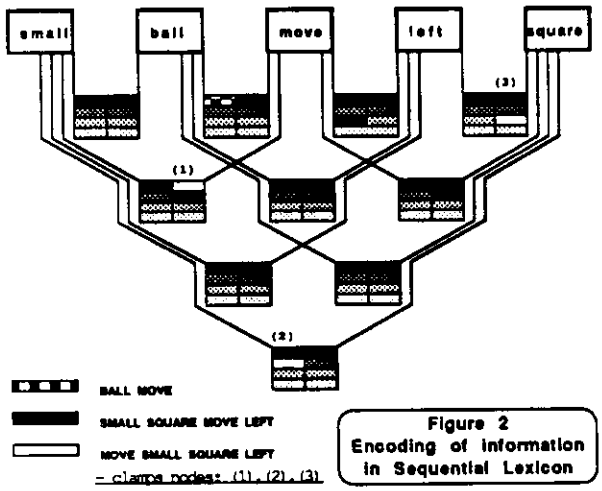

Figure 2
Encoding of Information in Sequential Lexicon

Reading and generation of text corresponds to sequential clamping of the corresponding word nodes. In this network we differentiate between clamping and firing of nodes. Only word nodes can be clamped one at a time. A word node is usually clamped for a number of life cycles of the network. Clamping a node causes it to generate a burst of spikes. Each word node is clamped temporarily and the duration of clamping is of particular importance. The verbal behavior of DETE consists of gradual learning to recognize the objects in a scene and (spontaneously or on demand) to name them. The system is also being designed to learn to generate sequences of lexical items (sentences) describing the current or past visual events, and to predict future scenes. The language generation can be monitored at various structural levels, where it appears in the form of hidden [Soko72] or articulated speech.

3.3 The Association Areas and the Mechanism of Selective Attention

Attention is a mental process through which we avoid distraction by irrelevant stimuli while seeking out and focusing on those that are behaviorally important. This complex psychological phenomena contains several component processes such as vigilance, concentration, focalization, scanning and exploration [Mars83]. Attention is directly related to motivation and volition. In the human mind, focus of attention is under dual control: (1) Sensory -- incoming sufficiently strong, unexpected or unusual sensory stimuli from the external environment can temporarily attract our attention. (2) Voluntary -- we are able to direct our attention as a result of certain mental processes. Attention is modality specific. We can focus our attention to the verbal or to the visual input. The function of the association cortex is to bind together distributed circuits, that represent a set of facts, by their simultaneous activation. Crick [Cric84] has suggested a scheme for obtaining such type of temporal binding. The binding occurs in intervals of about 50 ms during which bursts of impulses, produced by "searchlights" of attention in the thalamus, provide the signal for rapid synaptic changes. This hypothesis seems to have electrophysiological support, for instance Dempsey and Morison [Demp42] observed "augmenting" and "recruiting" waves spreading to the cortex after thalamic stimulation. In our system we are modeling certain thalamic functions, like the temporal binding of neural activity in different parts of the cortex. The advantage of this approach is that the representations are distributed over a population of neurons and that simultaneous co-activation of a group of neurons in one area (for instance the verbal association cortex) will impose simultaneous co-activation of a group of neurons in another area (like the visual association cortex) that receives a projection from the first. We also make use of Fukushima's model of selective attention [Fuku87].

4. THE ORDER OF LEARNING IN DETE

Teaching DETE the microworld language requires a consistent strategy. At the behavioral level (or the learning phase) we need to have consistency between verbal and visual input (i.e. we should not show to the system a small ball and call it "large ball"). DETE goes progressively through several stages of learning. Initially, verbal descriptions will only be gibberish to the system. The task facing DETE is to find the associations between the text and the visual scenes and to learn them in the following order: (1) Learn to recognize that there is an object in the VF, as opposed to an empty VF (i.e. object from surround). This is done on the basis of computing the difference between the total intensity of the input in resting and stimulated state. (2) Learn the meaning of a single concept. A general strategy allowing DETE to learn a single concept, for instance the concept of a circular shape, is to "clamp" this conceptual dimension in the visual field while letting the rest of the dimensions change with time. Figure 3 demonstrates graphically the learning of the concept of a circle (circular shape). Assume for a moment that the small square is not present on the VF. Currently DETE recognizes six feature dimensions including: shape, intensity, location, size, speed and direction of movement. First, we let the circle move and concurrently clamp the word "circle" at the verbal layer. As a result, the following changes happen in the association cortex:

One the one hand, the feature extractors have activated the appropriate feature values of the feature dimensions (shape "circle" -- left bottom of Figure 3; intensity "dark" -- left top; location "Left Up" -- middle top; etc.). On the other hand, the word "circle" (clamped simultaneously with the presentation of a circle on the VF) is continuously sending activation along its links to the association cortex (fully interconnected). Using the mechanism of selective attention described before, only the links to the association neurons, which are simultaneously activated through the visual input, are strengthened [Hebb49]. Assume that the circle in the visual field moves in different directions and speeds, shrinks in diameter or becomes darker or brighter (i.e. variation occurs along all feature dimensions but the shape dimension). As a result, the link between the word node "circle" and the association cortex node that will become strengthened maximally is the one (thicker line) going to the "circle" node (denoted by "C") of the shape dimension. Similarly, DETE will learn the meaning of any concept that has some visual embodyment.

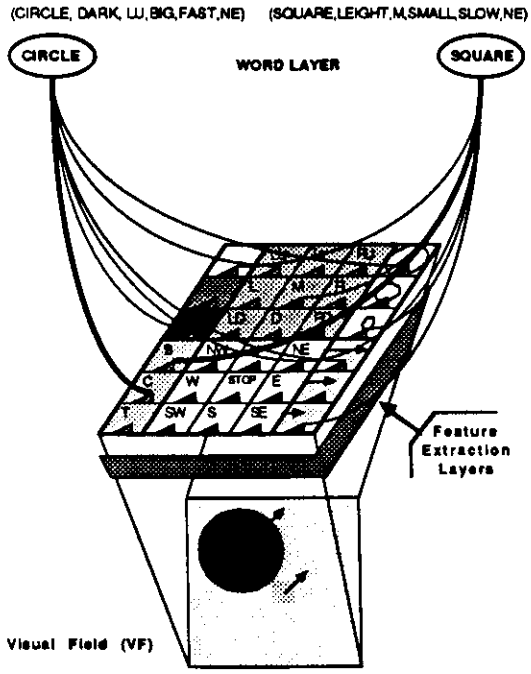

Figure 3
Learning word meanings

(3) Learning to understand the meaning of a description. The meaning of each sentence is represented by a pattern of activation spread over a set of nodes of the verbal subsystem and the associative cortex. Each consecutive sentence activates a different pattern and the activation patterns caused by previous sentences are suppressed.

5. CURRENT STATUS, ISSUES AND LONG-TERM GOALS

DETE is being written in *LISP -- a data parallel language used for programming the Connection Machine. For development purposes the model runs on the *LISP simulator on an Apollo DN4000. The neuronal layers, from which DETE's architecture is composed, are implemented as sets of parallel variables (pvars in *LISP). Currently, the visual field and object motions (w/ rebounding) are implemented as well as the Sequential Lexicon (SL) and also several of the feature extracting modules.

Some of the language issues which we face in the design of DETE include: (1) Learning syntactic patterns: e.g. *"the circle moves"* or *"the square moves"*. As a result, DETE should learn the pattern <object moves>. (2) Learning semantic relationships: for instance, how will DETE figure out that a description is referring to a future event? E.g., if it already knows that "circle" is circle and "collides" is a collision then, if it gets: "will collide" and sees a delayed collision, it can begin to learn about future tense.

The early stages of DETE's development involve extremely simple syntactic forms (no pronouns, gerunds, conjuncts, etc.). At present it only associates names with objects, no more complex descriptions are yet being learned. The next step is getting from naming an object to more complex sentences. For instance, learn a <motion direction> description. e.g. "moving up", "moving down", "standing", etc.

6. SUMMARY

The development of the DETE project has been inspired by recent advances in the neurosciences and the new insights obtained from computational modeling of real neural systems [Sejn86]. However, DETE differs in several aspects from current neural models. For instance, the neural elements used in the model are more realistic than in most connectionist models which gives us a number of advantages. One of them is in representation of timing relationships and especially in handling motion of objects in the VF. Another difference is the Sequential Lexicon (figure 2). This is a novel approach to handling sequential input, (e.g. linguistic input) which provides an efficient method for making temporally bound associations between pairs in sequential input.

Human infants learn much of language by associating the verbal utterances of other agents in the environment with changes in their visual fields. How does this visual information affect the acquisition of syntax and the formation of symbolic, language-based concepts? Current natural language processing (NLP) systems in AI do not ground their symbols in sensory experience. This is not surprising since computational approaches to the grounding problem are largely unexplored in AI. DETE is a good computational research environment in which to begin to address these issues.

References

[Ball82] Ballard, D.H., & Brown, C.M., (1982). *Computer Vision*, NY, Prentice Hall.

[Cric84] Crick, F.H.C., (1984) The function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences*, 81, 4586-4590.

[Demp42] Dempsey, E.W., & Morison, R.S., (1942). The production of rhythmically recurrent cortical potentials after localized thalamic stimulation. *American Journal of Physiology*. Vol. 135, pp. 293-300.

[Fuku80] Fukushima, K., (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *J Biological Cybernetics*. Vol. 36. pp. 193-202. Springer-Verlag.

[Fuku87] Fukushima, K., (1987). A Neural Network Model for Selective Attention. In *Proceedings of the IEEE First International Conference on Neural Networks*, San Diego, CA, Vol. 2, pp. 11-18.

[Gall87] Gallant, S., (1987). *Sequential Associative Memories*. Technical Report NU-CCS-87-20. Northeastern University, Cambridge, MA.

[Hart37] Hartline, H.K., (1837). The Response of Single Optic Nerve Fibers of the Vertebrate Eye to Illumination of the Retina. *The American Journal of Physiology*. 121(2): 400-415.

[Hebb49] Hebb, D. O., (1949). *The organization of behavior*. J. Wiley.

[Koch85] Koch, C., & Poggio, T., (1985). Biophysics of Computation: Nerves, synapses, and membranes. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *New insights into synaptic function*. New York: Neuroscience Research Foundation, 10, 616-637.

[Livi88] Livingstone, M.S., (1988). Art, Illusion and the Visual System. *Scientific American* Vol. 258(1). pp 78- 85.

[Marr82] Marr, D., (1982) *Vision*. San Fransisco, CA, W.H. Freeman.

[Mars83] M.-Marsel Mesulam, (1983 September). The Functional Anatomy and Hemispheric Specialization for Directed Attention. *TINS* pp.384-387.

[Mish82] Mishkin, M., (1982). *A Memory System in the Monkey*. Philosophical Transactions of the Royal Society of London B, Vol. 298, pp. 85-95.

[Rume86] Rumelhart, D.E., Smolensky, P., McClelland, J.L. and Hinton, G.E., (1986). Schemata and Sequential Thought Processes in PDP Models. In McClelland, J.L., Rumelhart, D.E. and the PDP Research Group, *Parallel Distributed Processing, Exploration in the Microstructure of Cognition*. 2, Chapter 14. The MIT Press, Cambridge, MA

[Sejn86] Sejnowski, T.J., (1986). Open Questions About Computation in Cerebral Cortex. *In Parallel Distributed Processing -- Explorations in the Microstructure of Cognition*. The MIT Press, Cambridge, MA, Vol. 2, pp 373-389.

[Soko72] Sokolov, A.N., (1972). *Inner Speech and thought*, Plenum Press, New York.

[VanE83] Van Essen, D.C. & Maunsell, J.H.R., (1983 September). Hierarchical organization and functional streams in the visual cortex. *TINS* pp 370-387.