International Journal of Computer Vision 45(3), 265–280, 2001

Motion Understanding: Task-Directed Attention and Representations that Link Perception with Action

JOHN K. TSOTSOS
Department of Computer Science, and Centre for Vision Research, York University, Toronto, Canada
tsotsos@cs.yorku.ca

Received November 13, 2000; Accepted May 29, 2001

Abstract. This short paper outlines my position on a future direction for computational research on visual motion understanding. The direction combines motion perception, visual attention, action representation and computational vision. Due the breadth of literature in these areas, the paper cannot present a comprehensive review of any one topic. The review is a selective one, a selection that attempts to make some particular points. I claim that task-directed attentive processing is a largely unexplored dimension in the computational motion field. I recount in the context of motion understanding a past argument that in order to make vision systems general, attention is one of the components of the strategy. No matter how sophisticated the methods become for extracting motion information from image sequences, it will not be possible to achieve the goal of human-like performance without integrating the optimization of processing that attention provides. Virtually all past surveys of computational models of motion processing completely ignore attention. However, the concept has crept into work over the years in a variety of ways. A second claim is that the biology of attention offers some interesting insights to guide future development. Many computational authors had previously commented that too little is known about how biological vision systems use task-directed attention in motion processing; this is no longer true. Here, I briefly summarize biological evidence that attentive processing affects all aspects of visual perception including motion, and again emphasize that this paper does not do justice to the breadth and depth of the field. New findings provide a critical link between the perception of visual actions and their execution. Together these findings point to a strategy for motion understanding closely related to that presented more than two decades ago.

Keywords: motion understanding, visual attention, task guidance, biological vision, representations

Introduction

This paper touches on several topics—visual attention, computer vision, motion perception, computational models of motion understanding. All are huge areas and cover several disciplines. One cannot do justice to this breadth and depth in a short paper. This paper does not present a unified nor complete review of visual attention, motion perception, or computational vision. It attempts to highlight certain recent findings, mostly neurobiological, and relate them to the history of high level motion understanding research in computer vision. As a result, the paper begins by tackling a beast of a problem, namely, a definition of attention. Almost certainly, Section 2.0 fails at this task, but each attempt hopefully will bring us closer to an acceptable definition. Section 3.0 overviews computational research on motion understanding during the past 25 years with a definite bias. The point this section tries to make is that the earliest work seems more in tune with what general purpose vision systems might require than more recent work. Section 4.0 provides a highly selective and extremely brief summary of experimental results in visual attention in humans and primates and on representations of actions with the goal of introducing these recent findings to the computer vision

community. Section 5.0 summarizes the selective tuning model for visual attention and provides links for its possible use in motion systems. The final section provides linkages among these diverse topics, suggesting directions for future research.

2. A Computational Perspective on Visual Attention

There is no wish to add to the confusion of buzzwords in computer vision here. Attention is a term not well understood by any discipline; yet, if we wish to have a meaningful discussion about it, some concreteness must be provided. The definition proposed here is a computational one, but one that may transcend disciplines because it was motivated by and captures the notions of processing cost and capacity limit so common in the attention literature. It is also important to distinguish attention here from closely related terms in computer vision such as 'active vision.' In doing so, it may be that the proposed definition is a common factor of other related terms.

2.1. Towards a Definition

What is 'attention'? My theoretical work initially addressed the question "Is there a computational justification for attentive selection?" The obvious answer that has been given many times that the brain is not large enough to process all the incoming stimuli, is hardly satisfactory (Tsotsos, 1987b). This answer is not quantitative and provides no constraints on what processing system might be sufficient. We have employed methods from computational complexity theory to formally prove for the first time that purely data-directed visual search in its most general form is an intractable problem in any realization (Tsotsos, 1989). There, it is claimed that visual search is ubiquitous in vision, and thus purely data-directed visual processing is also intractable in general. Those analyses provided important constraints on visual processing mechanisms and led to a specific (not necessarily unique or optimal) solution for visual perception. One of those constraints concerned the importance of attentive processing at all stages of analysis: the combinatorics of search are too large at each stage of analysis otherwise. Attentive selection based on task knowledge turns out to be a powerful heuristic to limit search and make the overall problem tractable (Tsotsos, 1990). Thus, I arrive at a particular proposal for a definition of attention: *Attention is a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception.*

This is of course not the only viewpoint possible and it is instructive to compare it to other computational approaches to attention and its related concepts, importantly, active vision. These are given in approximate chronological order and should not be considered as a complete list:

- Ullman (1984). Visual routines are composed of sequences of elemental operations, including shifting of the focus of attention, indexing to an odd-man-out location, boundary tracing, etc. The attention process itself is sketched as a competition among local signals, with no top-down component. It was elaborated in Koch and Ullman (1985).
- Koch and Ullman (1985). Attention selects single elements in the visual field from a saliency map, in order of saliency, for further processing.
- Bajcsy (1985). Active sensing is a control strategy applied to the data acquisition process that depends on the current state of the data interpretation including recognition.
- Aloimonos, Weiss and Bandyopadhay (1987). An observer is called active when engaged in some kind of activity whose purpose is to control the geometric parameters of the sensory apparatus.
- Burt (1988). Dynamic vision consists of foveation, tracking and high level interpretation, focussing the system's resources on selected areas of the scene.
- Clark and Ferrier (1988). The location with the highest saliency value is selected for the next fixation where saliency is computed dynamically as an application-specific weighted sum of the cross-correlations of the image with a set of known templates.
- Ballard (1991). The central asset of animate vision is gaze control. Animate vision systems can use physical search, make approximate camera movements, use exocentric coordinate frames, use qualitative algorithms, segment areas of interest precategorically, exploit environmental context and employ learning.
- Blake (1992). Active vision emphasizes the role of vision as a sense for robots and real-time perception systems, with advantages for structure from controlled motion, tracking, focussed attention and prediction.

- Pahlavan, Uhlin and Eklundh (1993). An active visual system is a system which is able to manipulate its visual parameters in a controlled manner in order to extract useful data about the scene in time and space.
- Olshausen, Andersen and Van Essen (1993). Attention solves the selection and routing problems, controlling the path of attended stimuli determined using the Koch and Ullman (1985) method, through the visual processing mechanism such that it is transformed into a canonical representation suitable for recognition from models.
- Ullman (1995). Within the sequence seeking process of information flow, attention can provide priming signals and context effects based on task using the feedback connections that are prominent throughout the visual cortex. The feedback projections either prime and modulate an ascending stream or directly activate a lower area. The counter streams structure activates bottom-up and top-down searches simultaneously, seeking to have them meet thus defining a sequence of mappings from input signal to target memory items.

It is important to note that although the concept of active perception is relatively new to computer vision, it is not new in the psychology of perception. In 1874, Franz Brentano introduced the concept of *act psychology*. He raised the possibility that a subject's actions play a role in perception and that perception and all conscious acts must be grounded by real objects. G.E. Müller defined his Komplextheorie of collective attention in 1904 where perception was based on actions. See Metzger (1974) for an overview of these and other early ideas.

The view of attention presented by my work, namely that attention optimizes search, can be thought of as a common factor among all of the above. Many (the active/animate vision researchers) seem to claim that attention and eye movements are one and the same; certainly none of the biological scientists working on this problem would agree. That one can attend to particular locations in the visual field without eye movements has been known since Helmholtz, but eye movements require visual attention to precede them to their goal (Hoffman, 1998 surveys relevant experimental work). Both selection goals are needed corresponding to overt and covert attentional fixations described in the perception literature. Active vision, as it has been proposed and used in computer vision, necessarily includes attention as a sub-problem.

2.2. Attention in Computer Vision

What is it about attention that makes it one of the easiest topics to neglect in computer vision? Take a look at the majority of books on computer vision and one will not find attention even mentioned. A telling example is found in the book *Active Vision* edited by Aloimonos (1993), where one would expect attention to be prominent. The index includes only one entry under the term 'attention.' That entry points to a paper by Sandini et al. and tells us that the task of tracking, or active control of fixation, requires as a fist step the detection of the target or focus of attention. That is all. How would one go about solving this? In Tsotsos (1989) (and the proof was confirmed using a totally different strategy by Rensink, 1989) it was shown that with no task knowledge and in a purely-data-directed manner, this sub-task of target detection is NP-Complete. Brooks also requires the detection of the trigger stimulus for his behaviors in the subsumption architecture (1986)—this too is NP-Complete (see Tsotsos, 1995b for discussion). It appears as if these authors are attempting to solve a problem that includes known intractable sub-problems. What conclusions can be drawn from such proposals? Is the problem thought to be irrelevant or is it somehow assumed away?

Several researchers study computational models of visual attention for their own sake (see Tsotsos, in press, for a review of those with biological relevance; Yeshurun, 1997 provides a brief summary of attention models in computer vision). Those who build complete vision application systems invoke attentional mechanisms because they must confront and defeat the computational load in order to achieve the goal of real-time processing (there are many examples, two of them being Baluja and Pomerleau, 1997 and Dickmanns, 1992). But the mainstream of computer vision does not give attentive processes, especially task-directed attention, much consideration.

Marr (1982) mentions attention, but only to discount it. Marr's influence seems to still be strong, even though on this point, we now know that his conclusions were incorrect. It is important to be clear about what Marr meant in his book. Several short excerpts are relevant:

p. 35. "The general trend in the computer vision community was to believe that recognition was so difficult that it required every possible kind of information"

p. 96. When describing grouping processes and the full primal sketch, he says, "our approach requires that the discrimination be made quickly—to be safe, in less than 160 ms—and that a clear psychophysical boundary be present"

p. 100. When describing the modular organization of the human visual processor,[1] he adds, "although some top-down information is sometimes used and necessary it is of only secondary importance ... evidence ... was willfully ignored by the computer vision community"

Marr argued that knowledge was not necessary in general to understand images. I feel he tried to draw too strong a conclusion from the experimental work of the period and underestimated the role of task-directed processes in vision. It is known now that attentive effects do not appear until at least 150 ms after the onset of a stimulus in IT cortex (Chelazzi et al., 1998) and 230 ms in area V1 (Roelfsema et al., 1998). Attentive influences appear *after* the time period Marr was considering. Experimental and theoretical work since then has provided a different perspective (see Tsotsos, 1990; Duncan and Desimone, 1995; Kastner and Ungerleider, 2000).

In Tsotsos (1992), I listed a spectrum of problems requiring attention: selection of objects, events, tasks relevant for domain, selection of world model, selection of visual field, selection of detailed sub-regions for analysis, selection of spatial and feature dimensions of interest, selection of operating parameters for low level operations. Take a look at this list and note how most research makes assumptions that reduce or eliminate the need for attention:

- Fixed camera systems negate the need for selection of visual field
- Pre-segmentation eliminates the need to select a region of interest
- 'Clean' backgrounds ameliorate the segmentation problem
- Assumptions about relevant features and the ranges of their values reduce their search ranges
- Knowledge of task domain negates the need to search a stored set of all domains
- Knowledge of which objects appear in scenes negates the need to search a stored set of all objects
- Knowledge of which events are of interest negates the need to search a stored set of all events

The point is that the extent of the search space is seriously reduced before the visual processing takes place, and most often even before the algorithms for solution are designed! However, it is clear that in everyday vision, and certainly in order to understand vision, these assumptions cannot be made.

3. Computational Research on Motion Understanding

3.1. Early Work

A great deal of research has appeared since the earliest days of computer vision and image processing, work that dealt with how a computer may extract time-varying properties from image sequences. However, I think it is correct to say that the very first major piece of research on the computer interpretation of motion from image sequences was the 1976 Ph.D. thesis of Norman Badler at the University of Toronto. I make this claim because he first connected motion features with complex, abstract, spatio-temporal concepts within an algorithm that was tested computationally. Thus, although many had used image sequences previously, the works that assume constancy of time-varying features over the sequences and were highly application dependent (such as cloud tracking) should be distinguished from those that consider discontinuities over time and hypothesize a general solution.

Badler described a methodology for producing conceptual descriptions of three-dimensional time-varying visual scenes. He began with two-dimensional coordinates of object features and described them at successively higher levels of abstraction. He developed a representation for objects and events based on adverbs and prepositions that characterize direction and higher order natural language terms to cover the notions of repetition and event sequences. At the highest level he had specific motion verbs. It is surprising how much of the vision community has interpreted his work as having more relevance for natural language than for visual motion understanding (HHN of course, did not make this mistake—see Nagel, 1981, 1988). The verbs are simply the most natural, best-understood and most readily accessible way humans have of communicating motion and time-varying concepts. The fact that no subsequent work has found any other high level representational concepts as replacement is testament to this. The use of verbs as descriptors has been 'independently' re-discovered several times since. Badler

considered both rigid and articulated motions. His solution was tested on synthetic image sequences. Algorithmically, he based his processes on the concept of a demon: small procedures that lie dormant until a certain set of assertions is recognized as true. The demon then becomes activate. To execute the demon, a second set of assertions or expectations must be also verified. Once the expectations are confirmed, then the demon operates on the database of facts and assertions, perhaps confirming a recognized instance. Demons have priority: lower level demons (those dealing with more primitive concepts) are executed first, followed by the higher level ones.

Both objects and events are represented using semantic networks, and properties such as type, parts, visibility, mobility, location, orientation, size are included. Location, for example, is further subdivided into adjacency, distance to other objects, contact, support, guided-by, connected-to, surrounded-by, contained-in covered-by, support-plane, elevation relationships. Events are represented using a set of nodes or cases, subject, agent, instrument, reference, direction, trajectory, velocity, axis, angular-velocity, next, start-time, end-time, repeat-path relationships. Adverbs and prepositions are also included similarly. Altogether, a rather rich descriptive vocabulary is defined. For each primitive concept that connects directly to image features or changes of image features, a demon is defined for its recognition. Higher level concepts also have their own demons, thus building upon concepts recognized at lower levels. The overall strategy is one that held great promise then and I feel still does. It certainly laid the foundation for all motion description work that followed.

Following Badler at the University of Toronto, I attempted to further his research but this time using real image sequences. Given that the work was done in the mid-to-late 1970's, with primitive computing facilities given today's systems, finding ready image sequences that would not require computational resources that were unavailable at the time, was critical. Although the work was presented as a general solution in the same sense that Badler's was, the test domain of cardiology took on a life of its own.[2]

I had defined a set of general motion concepts, organized in a hierarchy for ease of search, economy of definition, and explanatory power. Each concept corresponded roughly to a motion verb (not all were actual verbs, but rather convenient aggregate concepts useful for construction of the hierarchy). The hierarchy had several organization principles:

- generalization—organizing concepts from the specific to the generic;
- sub-part—organizing concepts depending on their components so different temporal and spatial granularity of concepts were included;
- similarity—representation based on differences between concepts, including methods for detecting the differences during matching and causing interpretation changes towards the concept most compatible with the image data;
- temporal precedence—representation that orders concepts in time, including cycles.

Motion expectations based on current hypotheses guided interpretation throughout. System resources were allocated based on a focus of attention, computed on strongest current hypotheses in space (image) and conceptual domains. Other contributions of that work (Tsotsos, 1977; Tsotsos et al., 1980; Tsotsos, 1980, 1985, 1987a) include:

1. Events were represented using packages of constraints on spatial and temporal characteristics. Cyclic events as well as phases within cycles were represented. These knowledge packages were organized into multiple intersecting hierarchies for ease of search and economy of representation using subpart, generalization, temporal precedence and similarity relationships. During recognition, a dynamically updated certainty factor was associated with each active hypothesis.
2. Integration of spatial information over time was accomplished by allowing hypotheses to support or compete with one another within a cooperative relaxation process that allowed their certainty factors to evolve over time. Certainty changes were dependent on matching with the image as well as context specified by relationship to other events, to their parts and to more generic events. The temporal sampling issues inherent in this problem were empirically included in order to tie system recognition performance to frame rate (as opposed to a frame rate that seems right for human observers).
3. Event boundaries were determined by detecting maximal differences of certainty factors over time. If hypothesis A has a decreasing certainty, hypothesis B has an increasing certainty, and both are tied to the same object, then hypothesis B begins and

hypothesis A ends when their certainties are maximally different.

4. A sophisticated hypothesis generation and refinement mechanism was defined that permitted new hypotheses to be activated depending on data-directed, model-directed and failure-directed influences. On hypothesis failure, alternate viable hypotheses were determined by either finding 'similar' alternates (move along similarity dimension) or relaxing constraints to a more general alternate (move upward in generalization hierarchy).

Both Badler and I agreed on one conclusion: it was clear that in order to push the methodology forward, low level computer vision must advance far beyond what was possible at that time and computational power would have to be significantly improved. In fact, that is exactly what has happened.

3.2. Current Research on Computational Event and Motion Recognition and the Role Attentive Processes have Played

One can survey the current literature and realize that attentive processing that utilizes task information to guide processing is not much of a concern. Many recent reviews of various aspects of motion understanding have not made any mention of attentive processing of any kind (Aggarwal et al., 1998; Shah and Jain, 1997; Cedras and Shah, 1994; Cedras and Shah, 1995; Hildreth and Royden, 1998). Well-known research such as Bobick (1997) claims to use knowledge but there is no attentive guidance of processing by that knowledge whatsoever. The review by Aggarwal and Cai (1999) includes one example of work that uses motion cues to segment an object and to affix an attentional window on to it. This is a data-directed attentional tool. Gavrila's (1999) review includes one example of where vision can provide an attentional cue for speech localization. Most of these cited papers make the claim that little or no work had been done on the topic of high level motion understanding previously (thus, the reintroduction of the previous section).

Many authors do not consider attention simply because they have assumed it away. As is also typical, there is always the accompanying justification that these tasks can easily be solved and they are not the focus of the main work. I can certainly appreciate this; I have used these phrases myself. It is true that good progress often comes by sacrificing some aspects of the problem. An example of the kinds of assumptions that are typically made even in the best work, is found in Siskind (1995). The input to his system must satisfy the following: a) all frames in a given movie must contain the same number of figures; b) the figures in each frame must be placed in a one-to-one correspondence with figures in adjacent frames; and, c) the system must be given this correspondence as input. Another example is Mann et al. (1997) who assume that their algorithm starts off by being given the region of interest that corresponds to each object that may be moving. The processing that ensues is perhaps the best of its kind currently, but the algorithm critically depends on reasonable regions of interest and is not designed to find that region of interest either independently or concurrently as it is understanding the events in the scene. A third example is the work of Pinhanez and Bobick (1997) who manually extract the values for their sensors by watching a video of the action and further, even determine the interval where every action and sub-action occurs. The problem is not that any one effort makes these assumptions; the problem lies in the fact that it is now almost universal to assume the unreasonable. Rather than think I am critical of these authors, the correct conclusion to draw from my comments is that I suggest a more balanced approach to the problem across the discipline, where at least some researchers study the attentive issues involved in a more general solution.

Attentive components have been included in systems not only through assumptions. At least three tools have appeared: the detection of salient tracking points/structures; search region predictions; and, Kalman filters and their extensions. Many examples have appeared; I only include a few illustrative ones.

HHN and his colleagues have had a long-standing interest in the detection of salient tracking points. The idea here is that if points or spatially-restricted structure can be found that correspond strongly to physical structure in the scene, then tracking those points alone rather than all points in the scene will both reduce the amount of computation needed plus yield object correspondence across images. For example, Dreschler and Nagel (1982) considered corner detectors since they seem to connect well with large corner-like structure on cars.

Several people have used predictions of where to search for corresponding structure image-to-image, an idea that appeared at least as early as Tsotsos (1980). Examples of its use have appeared in most works dealing with autonomous driving (such as Dickmanns and

Wünsche, 1999). In those systems, the search path corresponded to locations in the image where road edges might be found given the current car trajectory and location. In Tsotsos (1980) not only was a location prediction deployed to cut down the search region in an image but an additional mechanism was provided to deal with the real possibility that the prediction is wrong. Predictions were generated at several levels of abstraction. The most specific prediction was tried first; if in error, successively more abstract predictions were tried. For example, the most specific prediction might point to an exact location where an object will be found. If not found there, a more abstract prediction might provide a small region based on the object's known past motion history. If this too is incorrect, the next prediction might reflect simple forward motion. Finally the most abstract prediction would direct a search to a circular region around the objects' old position. Failure to find the object would be a cue for possible occlusion. This search reflects movement along the motion generalization hierarchy and is one of several dimensions of search employed by the system. HHN (1988) concludes that multiple representation dimensions should be useful for controlling search. Dickmanns and Wünsche later describe multiple hierarchies and multiple scales in space and time. The multiple scales are similar to those I used (moving up the part-whole hierarchy increases time scale as well as spatial scale depending on the type of concept represented).

Finally, the concept of a Kalman filter has been available for a long time (Kalman, 1960). Linear Kalman filters use an internal state representation of the moving object that includes a noise model and requires an explicit motion model that is assumed to be a constantly accelerated motion. They can be used to minimize the difference between the measured and predicted system state over time. They are suitable for a variety of tasks where the linearity assumption holds. Many people have used Kalman filters for motion processing with good success. They are an example of a well-defined mathematical construct that permits a connection between motion models and image predictions. A variety of extensions have appeared (for example, Wachter and Nagel, 1999; Dickmanns and Wünsche, 1999).

All are clearly strategies that help reduce search and fall within the proposed definition of attention; but task knowledge plays little or no role. As will hopefully become clear after the next section, attention can play a far larger role.

4. Attention and Representation in Biological Motion Perception

Attentive effects in motion perception have been known for some time (see for example, Wertheimer, 1912; Cavanagh, 1992; Lu and Sperling, 1995; Raymond, 2000). Further, the neurobiological evidence for attentive processes in vision, in both primates and humans, has been steadily increasing (see Desimone and Duncan, 1995; Kastner and Ungerleider, 2000). It is commonly accepted now that task can significantly modulate the responses of single neurons in many areas of visual cortex. This applies to spatial as well as to time-varying visual stimuli. Neural representations of motions and of actions have also been found in visual cortex. A very brief and selective summary follows. The purpose of this summary is not to comprehensively review the field,[3] but rather to introduce a computer vision audience to some recent experimental results that should have great impact on computational modeling of motion understanding.

4.1. Attentive Effects

In 1985, Moran and Desimone upset the bottom-up processing cart with a classic paper describing for the first time task-dependant, attentional modulation of V4 visual neurons in monkeys. Since then it has become well accepted that attentional modulation occurs in most visual areas of the brain for spatial tasks (Kastner and Ungerleider, 2000). Similarly, recent research in cognition, perception and neurophysiology provides evidence that behavioral tasks modulate the processing of visual motion. There is not much evidence for task-independent processing. Studies of human perception, measuring motion priming, motion aftereffects, uncertainty effects, and motion-interaction effects show that even simple aspects of motion processing may be affected by whether task motion is used or ignored by the perceiver (for reviews of this evidence see Watanabe and Miyauchi, 1998; Raymond, 2000). It is a nice convergence of results that this same dichotomy—whether the perceiver uses or ignores task information—is what led to the proofs in Tsotsos (1989, 1990, 1992) showing that when task information is used the resulting search space is reduced from exponential to linear size. Indeed, for some types of motion, selective attention can determine not only the perceived direction of motion and even whether or not any motion is perceived at all. Sperling and

Lu (1998) present an interesting functional architecture for the motion system that explicitly includes selective attention acting on saliency representations and that is claimed to account for several of these phenomena. Cavanagh et al. (2000) present an up-to-date version of Ullman's visual routines (1984). They considered common but complex motions such as a pencil bouncing on a table or a closing door. They propose that the perception of these motion patterns is mediated by attention as a high-level mental animation or *sprite*. Their experiments conclude that discrimination of even the simplest dynamic patterns demands attention support these specialized representations of action.

Additional contributions to understanding how attention and motion-processing brain mechanisms interact comes from single-unit studies in areas MT/MST of the monkey brain (Treue and Maunsell, 1996; Treue and Trujillo, 1999; Seidemann and Newsome, 1999), and in humans from functional imaging (Gandhi et al., 1999; Büchell et al., 1998; O'Craven et al., 1997; Beauchamp et al., 1997; Rees et al., 1997) and event-related potential studies (Valdes-Sosa et al., 1998). See also the review by Albright and Stoner (1995).

Let me provide just one example as illustration of the kind of experiment that leads to such a conclusion. Findings from single-unit studies in areas MT and MST of monkey show that when motion stimuli are presented inside a cell's receptive field, its response rate will depend on whether the animal is required to use the information in a concurrent task. The magnitude of attentional modulation may depend on the type of motion judgement required, whether motion is a target-defining feature in the stimulus configuration, and whether the target stimulus is inside or outside the cell's receptive field.

The stimuli used by Treue and Maunsell (1996) are shown in Fig. 1. One stationary square (target) was presented before the other and then both oscillated in counter-phase. Monkeys maintained central fixation and were rewarded for a key press whenever the target, but not the distractor, changed speed. In other words, they had a behavioral reason for responding to the motion of the target (the first of the two black squares presented). With the configuration shown in (a) neural firing rates were greater when the target, as opposed to the distractor, was within the cell's receptive field. For the configuration shown in (b) firing rates were maximal when the attended square moved in the cell's preferred direction and decreased substantially when the unattended square moved in the cell's preferred direction.

It seems a good hypothesis that motion-sensitive neurons (filters) at many levels of processing abstraction can be dynamically tuned to the task of the moment. The tuning takes the form of radical changes to the tuning properties of the receptive field that permit the neuron to be more selective to both the type of motion being attended as well as to its location.

Two examples of computational models that use some form of attentive processing are briefly mentioned, and others (that do not use attention) are overviewed in Hildreth and Royden (1998). Baloch and Grossberg (1997) present a model of high-level motion processing that includes some limited attentional processes. Within their scheme, top-down attentional signals play a priming role, priming recognition to motion continuity from image to image. This is not task

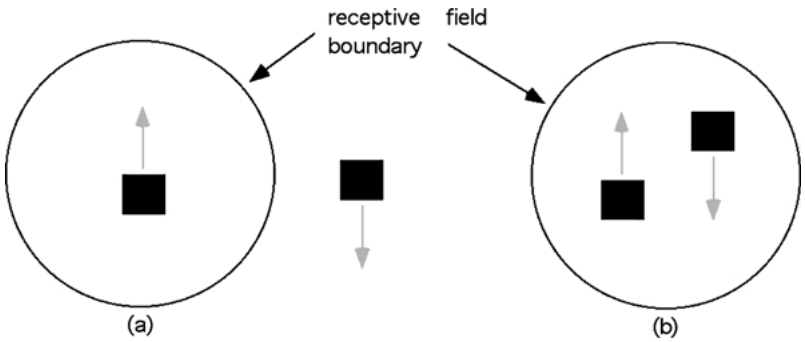

Figure 1. The two experimental setup used by Treue and Maunsell (1996) for the two key experiments. The circles represent the spatial extent of the receptive fields in area MT or MST from which recordings were taken. The black squares represent the actual stimuli used. The gray arrows denote the direction of motion of the black squares.

knowledge but is indeed a top-down effect. Nowlan and Sejnowksi (1995) employ selection of portions of visual field where velocity estimates are most reliable rather than process all locations in an image. This too is not task knowledge but does help limit the computational complexity of processing.

4.2. Representation of Observed and Executed Actions

Is there evidence for a representation of motion concepts in the brain? The reviews cited above provide evidence for a hierarchy of motion cortical areas, V1 neurons being sensitive to motion direction, MT neurons sensitive to direction and speed and MST neurons sensitive to patterns of motion. What is known for higher order motion concepts?

Perrett and colleagues (1990) describe neurons in monkey superior temporal sulcus (STS)[4] that become selectively active during the *sight* of hand actions. In area F5 in monkey,[5] neurons have been found to respond to specific actions such as grasping, holding, tearing (see Rizzolatti and Arbib for review, 1998). The actions are quite specific. For example, for a given neuron, the grasp must be with the index finger and thumb in order for that neuron to respond. The same observations have been made in humans using PET studies in Broca's area. A subset of these neurons also responds when the monkey is executing that action. In order words, a single neuron responds if the observer views a specific action, or if the observer executes that action. These neurons are called 'mirror' neurons. Such neurons may be important links between vision and action. Fadiga et al. (2000) go one step forward in suggesting that the collection of these neurons represent goal-directed 'motor words' and form the 'vocabulary' of actions the system can visually understand and can execute with its effectors. Rizzolatti and Arbib argue that individuals recognize actions made by others because the neural pattern elicited in their premotor areas during the observation of an action is similar to that internally generated to execute that action. Humphreys and Riddoch (2001) show that search for an object in a cluttered scene can be based not only on its perceptual properties but also on the intended actions using that object. These action templates can be used in visual search independently of perceptual properties and even object parts can cue those actions. These results further strengthen the proposal that there is a direct link between perception and action.

It does not seem to be much of a stretch to hypothesize that a hierarchy of motion selective neurons exists that codes simple direction at the early levels and complex motion at the highest levels (not unlike that motion hierarchy is Tsotsos, 1980). This hierarchy has the unique feature that it seems directly connected to action execution at its most abstract level.

5. The Selective Tuning Model for Visual Attention

5.1. Overview

Complexity analysis leads to the conclusion that attention must tune the visual processing architecture to permit task-directed processing (Tsotsos, 1990). Selective tuning takes two forms: *spatial* selection is realized by inhibiting task-irrelevant locations in the neural network, and *feature* selection is realized by inhibiting the neurons that represent task-irrelevant features. Only a brief summary is presented here since the model is detailed elsewhere (Tsotsos et al., 1995).

The spatial role of attention in the image domain is to localize a subset of the input image and its path through the processing hierarchy such as to minimize any interfering or corrupting signals. The visual processing architecture is a pyramidal network composed of units receiving both feed-forward and feedback connections. When a stimulus is first applied to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid to which it is connected. The result is the activation of an inverted sub-pyramid of units and we assume that the degree of unit activation reflects the goodness-of-match between the unit and the stimulus it represents.

Attentional selection relies on a hierarchy of winner-take-all (WTA) processes. WTA is a parallel algorithm for finding the maximum value in a set of variables, which was first proposed in this context by Koch and Ullman (1985). WTA can be steered to favor particular stimulus locations or features but in the absence of such guidance it operates independently. The processing of a visual input involves three main stages. During the first stage, a stimulus is applied to the input layer and activity propagates along feed-forward connections towards the output layer. The response of each unit depends on its particular selectivities, and perhaps also on a top-down bias for task-relevant qualities. During the second stage, a hierarchy of WTA processes is applied in a top-down, coarse-to-fine manner. The first

WTA process operates in the top layer and covers the entire visual field at the top layer: it computes the unit with the largest response in the output layer, that is, the *global winner*. In turn, the global winner activates a WTA amongst its input units in the layer immediately below. This localizes the largest response within the receptive field of the global winner. All of the connections of the visual pyramid that do not contribute to the winner are pruned (i.e., attenuated). This strategy of finding the winner within each receptive fields and then pruning away irrelevant connections, is applied recursively through the pyramid, layer by layer. Thus, the global winner in the output layer is eventually traced back to its perceptual origin in the input layer. The connections that remain (i.e., are not pruned) may be considered the *pass zone* of the attentional beam, while the pruned connections an *inhibitory zone* around that beam. While we are not claiming biological accuracy for the WTA process, we are claiming plausibility, as it does not violate biological connectivity or time constraints. During the third stage, the selected stimuli in the input layer re-propagate through the network, being processed by the same neurons but this time without distracting stimuli in each receptive field, as if they had been presented on a "blank" background.

5.2. *The Selective Tuning Model and Temporal Concepts*

The attention model considers two types of time-varying events so far: onset and offset of objects and motion segmentation using optical flow patterns (see Tsotsos et al., 1995 for details). It is clear that these just begin the integration of event and motion understanding within the model and there is a huge amount of work to be done.

The onset (appearance) of an object is a well-known attention capture mechanism and a good deal of psychophysical research supports its importance. Further, the onsets may occur in any representation (luminance, depth, motion, texture) and all have similar attention-grabbing effects. We defined a simple detection mechanism based on temporal differences of difference of spatial gaussian filters, with winner-take-all processes operating within scale and across scales to locate the strongest event. The location of this strongest event was the focus of attention. Both onsets and offsets can be handled; gradual onsets and offsets similarly are included in the method (slow lighting variations, for example).

Instantaneous full velocity optic flow patterns can be used, not to interpret the motion (i.e., extract motion parameters), but rather, in a fast method of localizing and labeling salient motion patterns. Once localized, they can be examined in more detail for motion parameters. We constructed templates for each of 16 types of motion pattern. The patterns fall into two categories: motion of the environment (full field motions such as 'approach' or 'rotation') and motion of objects in the visual field (such as translation, dilate, rotate). Goodness-of-fit measures for how well a template explains a given subset of optic flow in the image were computed using straightforward correlation. In each case, sub-regions of the image (hypothesized objects) are exhibiting this flow pattern. Winner-take-all processes determined the strongest matching location within each motion type, and then another WTA across all motions would determine the most salient motion. The system localizes the motion (segments the object exhibiting the motion) and provides the motion category. These two results seem to provide the bulk of the attention guidance most motion understanding methods need. It is not claimed that this is all there is to motion processing; rather, this simple scheme appears to be sufficient to detect, localize and label regions where salient motion is occurring so that further analysis may consider only that sub-image for detailed inspection. The templates are consistent with stimuli found effective for neurons in motion areas in monkey.

5.3. *Biological Predictions*

The selective tuning model was initially developed with the dual goals of computational utility and biological predictive power. The predictions for human and primate vision and supporting evidence are briefly described.

- An early prediction (Tsotsos, 1990) was that attention seems necessary at any level of processing where a many-to-one mapping between neural processes was found. Further, attention occurs in all the areas in a coordinated manner. The prediction was made at a time when good evidence for attentional modulation was known for area V4 only (Moran and Desimone, 1985). Since then, attentional modulation has been found in many other areas both earlier and later in the visual processing stream, and that it occurs in these areas simultaneously (Kastner et al., 1998). Vanduffel et al. (2000) have shown that attentional

modulation appears as early as the LGN. The prediction that attention modulates all cortical and even subcortical levels of processing has been borne out by recent work from several groups (e.g., Brefzynski and DeYoe, 1999; Ghandhi et al., 1999).

- The notions of competition between stimuli and of attentional modulation of this competition were also early components of the model (Tsotsos, 1990) and these too have gained substantial support over the years (Desimone and Duncan, 1995; Kastner et al., 1998; Reynolds et al., 1999).
- The model predicts an inhibitory surround that impairs perception around the focus of attention (Tsotsos, 1990) a prediction that seems to be gaining support, both psychophysically and neurophysiologically (Caputo and Guerra, 1998; Bahcall and Kowler, 1999; Vanduffel et al., 2000; Smith et al., 2000; Tsotsos et al., 2001).
- The model further implies that pre-attentive and attentive visual processing occur in the same neural substrate, which contrasts with the traditional view that these are wholly independent mechanisms. This point of view has also been gaining ground recently (Joseph et al., 1997; Yeshurun and Carrasco, 1999).
- A final prediction is that attentional guidance and control are integrated into the visual processing hierarchy, rather than being centralized in some external brain structure. This implies that the latency of attentional modulations *decreases* from lower to higher visual areas, and constitutes one of the strongest predictions of the model. This seems confirmed by experiment. Attentive effects do not appear until 150 ms after the onset of a stimulus in IT cortex (Chelazzi et al., 1998) while in V1 they appear after 230 ms (Roelfsema et al., 1998).

Additional predictions of the selective tuning model concern the form of spatial and temporal modulations of visual cortical responses around the focus of attention, and the existence of a WTA circuit connecting cortical columns of similar selectivity. The selective tuning model offers a principled solution to the fundamental problems of visual complexity, a detailed perceptual account of both the guidance and the consequences of visual attention, and a neurally plausible implementation as an integral part of the visual cortical hierarchy. Thus, the model "works" at three distinct levels—computational, perceptual, and neural—and offers a more concrete account, and far more specific predictions, than previous models limited to one of these levels. We are working to extend the model in several directions.

6. Discussion

I have tried to argue, in the brief exposition above, that motion understanding can benefit if attentive capabilities are added to current conceptions. I have tried to remind that early work made these claims long ago. In computer vision, many previously commented that not enough is known about how biology uses attention in motion processing (and in vision in general) and thus there is no value in trying to gain inspiration from biology; this is no longer true. Biological evidence has clearly made dramatic statements about the role of attention in visual perception. I draw two major conclusions here and address the role of attention first, and representations second.

In order to make the case for attentive processes in computer vision it is not enough to simply cite evidence that biology appears to depend strongly on attentive influences. One must argue convincingly that computer vision can benefit significantly from attentive processes. Why should it be the case that attention plays such a large role in biological or computational perception? Theoretically, we (Tsotsos, 1989, 1990, 1992; Parodi et al., 1998; Ye and Tsotsos, 1999) have shown that without attention, specifically without the use of task specific knowledge to guide processing, vision in the general case has exponential complexity. With the addition of even small amounts of knowledge to guide processing polynomial to linear complexity can be achieved. The results are mostly for worst case analysis; but note that the Parodi et al. paper deals with median case analysis. I therefore feel the argument against a purely data-driven approach to general purpose vision is as strong as it can possibly be. There is neither biological nor computational justification to support purely data-driven visual processing as a general strategy.

To date, motion systems have been single purpose with assumptions that reduce or eliminate the need for attention embedded implicitly or explicitly into the processing strategy. However, it is clear that in time systems will become multi-purpose. What would be a suitable strategy for their design? One could, of course, cobble together a number of single-purpose systems with a smart switching mechanism to apply the right one at the right time. As shown in Tsotsos (1995b), this will not scale well. The strategy that paper

recommends, and which the human brain seems to also employ, is to enable the co-ordinated optimization of the basic processing methods for the task at hand. This is the role of attention, and it has as significant utility in the image domain, in the temporal domain and in the conceptual domain.

We can enumerate some different attentional effects on motion processing and conclude that they are all forms of search optimization. What aspects of a motion task can be used as task guidance to optimize search?

- prediction in time about where objects might appear depending on speed and trajectory
- predictions on how to find objects if predictions fail using abstract concepts and relaxing the categories
- predicting the temporal sequence of concepts
- predicting change in appearance over time
- predicting interactions with other objects
- predicting the effects of gravity or other forces
- tuning the performance of individual filters (neurons) to improve their signal-to-noise ratio

In each case, without the attentive action, a search must be done over the space of possible explanations for the changes observed. As is true in human perception, a correct prediction (or 'cue') can lead to faster performance since the search process is short-circuited. An erroneous prediction leads to performance that is worse than without any cue at all. The burden then on the system is to provide good cues or predictions as a means of reducing or eliminating search. For example, the use of Newtonian mechanics explanations for motion as in Mann et al. (1997) can lead to well-reasoned predictions grounded in the forces and dynamics of particular object motion.

The second conclusion relates to the kinds of representations motion understanding might employ and that task-directed attention modulates. It was shown 20 years ago that attention, defined by competition among hypotheses that chooses the strongest one to focus on, can operate successfully over a complex hierarchy of motion concepts (Tsotsos, 1980) and can eliminate a significant amount of the search that a non-attentive process might otherwise require. Also, edge operators were tuned differentially across the image depending on expected orientations of local scene structure. That work did not include a 'natural' method for the system to acquire and represent the task nor were there any motion filters. Tuning can apply to motion filters as well, as has been discovered in the motion visual areas of monkey. This would greatly improve their specificity and reduce their computational complexity. The motion neurons described earlier seem to point to a representation that is hierarchical with single neurons representing complex motion concepts.[6] This is not unlike that used in my past work, although there is no reason to think that any level of my representation has any analogue to biological representations. The basic result points to hierarchical representations of simple to complex motion concepts. Tie this together with the selective tuning strategy for attentive selection and a new direction for motion understanding results, one that is biologically consistent with current primate and human observations. It also holds the promise of a tight coupling between the perception of action and their execution. Task can be communicated to the system simply by 'showing' the system the action to be performed. All of this is quite speculative; however, I hope to use the motivation that HHN instilled in me in 1981 when I sat down across from him, to pursue this direction with new vigor.

Acknowledgments

I thank David Fleet, Bill Freeman, Jeff Siskind and Florin Cutzu for several relevant literature pointers. I also thank Florin for useful comments on an early draft. Three anonymous reviewers provided much needed feedback and I thank them.

Hans-Hellmut Nagel: My Personal Notes

My first contact with HHN. In order to complete one's doctoral dissertation at the University of Toronto, the last requirement to be satisfied is a positive appraisal by a senior external researcher. My thesis committee selected HHN for this task. I have kept the original copy of my thesis that he returned to me. What an amazing effort! I keep this as a reminder of what it means to be thorough, careful and complete. I also use it often to show students who complain on seeing my markings on their documents! His efforts greatly improved my thesis document and research.

Sharing an Office with HHN. HHN invited me to spend part of my first post-PhD year at the University of Hamburg and it was a surprise and a thrill to share an office with him. A wonderful experience! I still recall my first day. I had not realized that I would share an office with him, an office with two large wooden desks facing each other over-looking Schluterstrasse on the

very corner where the famous Hamburg taxi-cab image sequence was created. He started laying out the plan for the summer for me on the chalkboard behind him, a literal wall of equations and ideas. Try as I might to absorb all of this, it took me some time to get up to a speed where I could trick myself into thinking that I could keep up with him (and I am not sure that even now I can accomplish this).

One of the reasons for the invitation to Hamburg was to see if I could re-implement the work done on my PhD on their network of mini-computers and using the programming language ADA. This was a tall enough order on its own, but what we did not realize was that HHN's perspective, originating from his training in physics, and mine, from primarily symbolic AI, were not very compatible. We spent a good deal of time trying to sort out what we really thought were the issues, the open problems and the solution methods. I feel I disappointed him because I never did complete that implementation. However, I wish to assure him that I learned a great deal from him and his perspective has helped guide my thinking ever since.

I fondly recall having dinner in his home one evening and then going for a walk along the river Elbe discussing vision, science and philosophy. This is what being an academic is all about, I thought—I was sold!

This Paper and HHN. When I was asked to contribute to this special issue I was honored, yet very concerned. HHN's contributions have been broad and deep: optical flow, motion understanding, applications such as walking people, traffic scenes, and more. I have been away from the motion field for a long time and even when active I concentrated on only one small aspect of what HHN worked on—motion understanding. What did I have to write about? Did I have a perspective on the field that was still valid? My current research focuses on biological visual attention and the development of a computational model that explains behavioral observations at both psychophysical and neurobiological levels. I have made only small excursions into the temporal domain and have focussed on spatial attention primarily. Could my attention research have relevance to motion understanding? I think it might.

As I was thinking about those pleasant days in Hamburg, I realized that the parts of my stay that I valued the most were my 'debates' with HHN. We had discussions on the nature of the motion field, what has been done, what should be done, what are the open problems, which are the valid methods, and so on. So, Hans-Hellmut, in your honor and to remind you of those productive discussions, I present here a personal view of the field, more of a position paper than a description of results, and of possible avenues for future work. When we were together we stirred the pot, as the saying goes, quite strongly and this paper is written with the goal of keeping the pot well-mixed! Knowing you as well as I do, I am fully expecting for you send me your concerns, refinements, questions, and doubts. After all, that is what we did years ago and it is the foundation for both my pleasant memories of you and for my respect for all you have taught me.

Notes

1. We also now know that the kind of modularity and processing independence that Marr hoped for are not found in primate or human vision (Felleman and Van Essen, 1991; Salin and Bullier, 1995).
2. It is both amusing and depressing in hindsight to note that in the late 70's and early 80's, the use of a medical problem as a test domain was the kiss of death with respect to how the community viewed the generality of the results. More recently, the use of a medical domain has been common and acceptable as a demonstration of one's results. Indeed, motion understanding research has used ballet, face gestures, hand gestures, traffic scenes, cooking shows, baseball all with claims of 'no loss of generality.' We may thus conclude there has been a significant change in the sociology of the discipline that has now made this claim acceptable.
3. Overviews of attention research in biological vision can be found in Pashler (1998), Styles (1998), Desimone and Duncan (1995), Kastner and Ungerleider (2000), Allport (1989).
4. The superior temporal sulcus includes portions of several visual areas (STP, FST and others); FST is the same level in the hierarchy as MST, the remainder are above.
5. Area F5 is a ventral premotor area in the monkey involved in the control of hand movements. Area 7b provides visual input to F5 and receives input from MST and FST, among others (Felleman and Van Essen, 1991).
6. This does indeed sound like 'grandmother cells' for motion. However, there was no suggestion that each motion concept is represented only by a unique neuron. It may be that there are several neurons each able to represent similar concepts.

References

Aggarwal, J.K. and Cai, Q. 1999. Human motion analysis: A Review. *Computer Vision and Image Understanding*, 73(3):428–440.

Aggarwal, J.K., Cai, Q., Liao, W., and Sabata, B. 1998. Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding*, 70(2):142–156.

Albright, T. and Stoner, G. 1995. Visual motion perception. *Proc. National Academy of Science, USA*, 92(7):2433–2440.

Allport, A. 1989. Visual attention. In *Foundations of Cognitive Science*, M.I. Posner (Ed.). MIT Press; Bradford Books: Cambridge, pp. 631–682.

Aloimonos, Y. (Ed.). 1993. *Active Perception*. Lawrence Erlbaum Associates Publ.

Aloimonos, Y., Weiss, I., and Bandyopadhay, A. 1987. Active vision. *Int. J. Computer Vision*, 1(4):333–356.

Badler, N.I. 1975. Temporal scene analysis: Conceptual descriptions of object movements. Ph.D. Thesis, Dept. of Computer Science, University of Toronto.

Bahcall, D. and Kowler, E. 1999. Attentional interference at small spatial separations. *Vision Research*, 39(1):71–86.

Bajcsy, R. 1985. Active perception vs passive perception. In *Proc. IEEE Workshop on Computer Vision: Representation and Control*, Oct., Bellaire, Mich., pp. 55–62.

Ballard, D. 1991. Animate vision. *Artificial Intelligence*, 48:57–86.

Baloch, A. and Grossberg, S. 1997. A neural model of high-level motion processing: Line motion and formoiton dynamics. *Vision Research*, 37(21):3037–3059.

Baluja, S. and Pomerleau, D.A. 1997. Expectation-based selective attention for the visual monitoring and control of a robot vehicle. *Robotics and Autonomous Systems Journal*, 22:329–344.

Beauchamp, M.S. et al. 1997. Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol*., 78:516–520.

Blake, A. 1992. Active vision. In *The Handbook of Brain Theory and Neural Networks*, M. Arbib (Ed.). MIT Press, pp. 61–63.

Bobick, A. 1997. Movement, activity, and action: The role of knowledge in the perception of motion. *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, London, England.

Brefczynski, J.A. and DeYoe, E.A. 1999. A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*., 2(4):370–374.

Brooks, R. 1986. A layered intelligent control system for a mobile robot. *IEEE Journal of Robotics and Automation RA-2*, April, 14–23.

Büchel, C. et al. 1998. The functional anatomy of attention to visual motion: A functional MRI study. *Brain*, 121:1281–1294.

Burt, P. 1988. Attention mechanism for vision in a dynamic world. In *Proc. 9th Int. Conf. on Pattern Recognition*, pp. 977–987.

Caputo, G. and Guerra, S. 1998. Attentional selection by distractor suppression. *Vision Research*, 38(5):669–689.

Cavanagh, P. 1992. Attention based motion processing. *Science*, 257:1563–1565.

Cavanagh, P., Labianca, A.T., and Thornton, I. M. 2000. Attention-based visual routines: Sprites. *Cognition*, in press.

Cedras, C. and Shah, M. 1994. A survey of motion analysis from moving light displays. *IEEE CVPR-94*, Seattle, Washington, pp. 214–221.

Cedras, C. and Shah, M. 1995. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155.

Chelazzi, L., Duncan, J., Miller, E., and Desimone, R. 1998. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiology*, 80:2918–2940.

Clark, J.J. and Ferrier, N. 1988. Modal control of an attentive vision system. In *Proc. ICCV*, Dec., Tarpon Springs Florida, pp. 514–523.

Desimone, R. and Duncan, J. 1995. Neural mechanisms of selective attention. *Annual Review of Neuroscience*, 18:193–222.

Dickmanns, E. 1992. Expectation-based dynamic scene understanding. In *Active Vision*, Blake and Yuille (Eds.), MIT Press: Cambridge, Massachusetts, pp. 303–334.

Dickmanns, E.D. and Wünsche, H.J. 1999. Dynamic vision for perception and control of motion. In *Handbook of Computer Vision and Applications Vol. 2*, B. Jahne, H. Haubeccker, and P. Geibler (Ed.). Academic Press.

Dreschler, L. and Nagel, H.H. 1982. On the selection of critical points and local curvature extrema of region boundaries for interframe matching. In *Proc. Int. Conf. Pattern Recognition*, Munich, pp. 542–544.

Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. 2000. Visuomotor neurons: Ambiguity of the discharge or 'motor' perception? *Int. J. Psychophysiology*, 35:165–177.

Felleman, D. and Van Essen, D. 1991. Distributed hierarchical processing in the primate visual cortex. *Cerebral Cortex*, 1:1–47.

Gandhi, S.P., Heeger, D.J., and Boynton, G.M. 1999. Spatial attention affects brain activity in human primary visual cortex. *Proc Natl Acad Sci USA*, 96(6):3314–3319.

Gavrila, D.M. 1999. The visual analysis of human movement: A Survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Hildreth, E. and Royden, C. 1995. Motion perception. In *The Handbook of Brain Theory and Neural Networks*, M. Arbib (Ed.). pp. 585–588. MIT Press.

Hoffman, J. 1998. Visual attention and eye movements. In *Attention*, H. Pashler (Ed.). *Psychology Press*, 119–154.

Humphreys, G. and Riddoch, M. 2001. Detection by action: Neuropsychological evidence for action-defined templates in search. *Nature Neuroscience*, 4:84–88.

Joseph, J., Chun, M., and Nakayama, K. 1997. Attentional requirements in a 'preattentive' feature search task. *Nature*, 387:805–807.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Trans. ASMEE J. Basic Eng*., 82:35Œ45.

Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L. 1998. Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, 282:108–111.

Kastner, S. and Ungerleider, L. 2000. Mechanisms of visual attention in the human cortex. *Annual Rev. Neuroscience*, 23:315–341.

Koch, C. and Ullman, S. 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiology*, 4:219–227.

Lu, Z. and Sperling, G. 1995. Attention-generated apparent motion. *Nature*, 377:237–239.

Mann, R., Jepson, A., and Siskind, J. 1997. The computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65(2):113–128.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

Martin, W. and Aggarwal, J. 1988.

Metzger, W. 1974. Consciousness, Perception and Action. In *Handbook of Perception vol. 1, Historical and Philosophical Roots of Perception*, Academic Press, pp. 109–125.

Moran, J. and Desimone, R. 1985. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784.

Nagel, H.H. 1981. Image sequence analysis: What can we learn from applications? In *Image Sequence Analysis*, T.S. Huang (Ed.). Springer Verlag: Berlin, pp. 19–228.

Nagel, H.H. 1988. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74.

Nowlan, S. and Sejnowski, T. 1995. A selection model for motion processing in area MT of primates. *The Journal of Neuroscience*, 15(2):1195–1214.

O'Craven, K.M. et al. 1997. Voluntary attention modulates fMRI activity in human MT-MST. *Neuron*, 18:591–598.

Olshausen, B. Anderson and C. Van Essen, D. 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. of Neuroscience*, 13(1):4700–4719.

Pahlavan, Uhlin and Eklundh. 1993. Active vision as a methodology. In *Active Perception*, Y. Aloimonos (Ed.). Lawrence Erlbaum Associates Publ., pp.19–46.

Parodi, P., Lanciwicki, R., Vijh, A., and Tsotsos J.K. 1998. Empirically-derived estimates of the complexity of labeling line drawings of polyhedral scenes. *Artificial Intelligence*, 105:47–75.

Pashler, H. 1997. *The Psychology of Attention*. MIT Press.

Perrett, D., Mistlin, A., Harries, M., and Chitty, A. 1990. Understanding the visual appearance and consequence of hand actions. In *Vision and Action: The Control of Grasping*, M. Goodale (Ed.). Ablex Publishing Corp., pp. 163–180.

Pinhanez, C. and Bobick, A. 1997. Human action detection using PNF propagation of temporal constraints. MIT Media Lab TR 423.

Raymond, J. 2000. Attentional modulation of visual motion perception. *Trends in Cognitive Sciences*, 4(2):42–50.

Rees, G. et al. 1997. Modulation of irrelevant motion perception by varying load in an unrelated task. *Science*, 278:1616–1619.

Rensink, R. 1989. A new proof of the NP-completeness of visual match. TR 89-22, Dept. of Computer Science, University of British Columbia.

Reynolds, J., Chelazzi, L., and Desimone, R. 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of Neuroscience*, 19(5):1736–1753.

Rizzolatti, G. and Arbib, M. 1998. Language within our grasp. *Trends in Neuroscience*, 21(5):188–194.

Roelfsema, P., Lamme, V., and Spekreijse, H. 1998. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395:376–380.

Salin, P.A. and Bullier, J. 1995. Corticocortical connections in the visual system: Structure and function. *Physiological Reviews*, 75(1):107–154.

Sandini, G., Gandolfo, F., Grosso, E., and Tistarelli, M. 1993. Vision during action. In *Active Perception*, Y. Aloimonos (Ed.). Lawrence Erlbaum Associates Publ.

Seidemann, E. and Newsome, W.T. 1999. Effect of spatial attention on the responses of area MT neurons. *J. Neurophysiol.*, 81:1783–1794.

Shah, M. and Jain, R. 1997. Visual recognition of activities, gestures, facial expressions and speech: An introduction and a perspective. In *Motion-Based Recognition*, M. Shah and R. Jain (Ed.). Kluwer Academic Publishers.

Siskind, J.M. 1995. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391.

Sperling, G. and Lu, Z.-L. 1998. A systems analysis of visual motion perception. In *High-Level Motion Processing*, T. Watanabe (Ed.). MIT Press. pp. 153–183.

Styles, E. 1997. *The Psychology of Attention*. Psychology Press.

Treue, S. and Maunsell, J.H.R. 1996. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, 382:539–541.

Treue, S. and Trujillo, J.C.M. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–579.

Tsotsos, J.K. 1977. Some notes on motion understanding. In *Proc. Fifth International Joint Conference on Artificial Intelligence*, MIT, Cambridge: Mass. p. 611.

Tsotsos, J.K. 1980. A framework for visual motion understanding. Ph.D. Thesis, Dept. of Computer Science, University of Toronto.

Tsotsos, J.K. 1985. The role of knowledge organization in representation and interpretation of time-varying data: The ALVEN system. *Computational Intelligence*, 1(1):16–32.

Tsotsos, J.K. 1987a. Representational axes and temporal cooperative processes. In *Vision, Brain and Cooperative Computation*, M. Arbib and A. Hanson (Ed.). MIT Press/Bradford Books, pp. 361–418.

Tsotsos, J.K. 1987b. A 'complexity level' analysis of vision. In *Proc. International Conference on Computer Vision*, London, England.

Tsotsos, J.K. 1989. The complexity of perceptual search tasks. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan, pp. 1571–1577.

Tsotsos, J.K. 1990. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–445.

Tsotsos, J.K. 1992. On the relative complexity of passive vs active visual search. *International Journal of Computer Vision*, 7(2):127–141.

Tsotsos, J.K. 1995b. On behaviorist intelligence and the scaling problem. *Artificial Intelligence*, 75:135–160.

Tsotsos, J.K. in press. Neurobiological models of visual attention. *5th Course of International Summer School "Neural Nets E.R. Caianiello" on Visual Attention Mechanisms*.

Tsotsos, J.K., Culhane, S., and Cutzu, F. 2001. From theoretical foundations to a hierarchical circuit for selective attention. In *Visual Attention and Cortical Circuits*, J. Braun, C. Koch, and J. Davis, (Ed.). MIT Press, pp. 285–306.

Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., and Nuflo, F. 1995. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1–2):507–547.

Tsotsos, J., Mylopoulos, J., Covvey, H.D., and Zucker, S.W. 1980. A framework for visual motion understanding. *IEEE Pattern Analysis and Machine Intelligence*, Special Issue on Computer Analysis of Time-Varying Imagery, pp. 563–573.

Ullman S. 1984. Visual routines. In *Visual Cognition*, S. Pinker (Ed.). MIT Press, pp. 97–160.

Ullman, S. 1995. Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, 1:1–11.

Valdes-Sosa, M. et al. 1998. Switching attention without shifting the spotlight: Object-based attentional modulation of brain potentials. *J. Cogn. Neurosci.*, 10:137–151.

Vanduffel, W., Tootell, R., and Orban, G. 2000. Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system. *Cerebral Cortex*, 10:109–126.

Wachter, S. and Nagel, H.H. 1999. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192.

Watanabe, T. and Miyauchi, S. 1998. Roles of attention and form in visual motion processing: Psychophysical and brain-imaging studies. In *High-Level Motion Processing*, T. Watanabe (Ed.). MIT Press, pp. 95–113.

Wertheimer, M. 1912, 1961. Experimentelle studien über das sehen von bewegung. In *Classics in Psychology*, T. Shipley (Ed.). Philosophical Library: New York, Vol. 61, pp. 161–265.

Ye, Y. and Tsotsos, J.K. 1999. Sensor planning for object search. *Computer Vision and Image Understanding*, 73(2):145–168.

Yeshurun, Y. 1997. Attentional mechanisms in computer vision. In *Artificial Vision: Image Description, Recognition and Communication*, V. Cantoni, S. Levialdi, and V. Roberto (Ed.). Academic Press, pp. 43–52.

Yeshurun, Y. and Carrasco, M. 1999. Spatial attention improves performance in spatial resolution tasks. *Vision Research*, 39(2):293–306.