YORK
UNIVERSITÉ
UNIVERSITY

redefine THE POSSIBLE.

Re-Visiting Visual Routines: A White Paper

John K. Tsotsos

Technical Report CSE-2010-11

October 1 2010

Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Re-Visiting Visual Routines: A White Paper

John K. Tsotsos[1]

Dept. of Computer Science and Engineering
York University
Toronto, Canada

October 1, 2010

Abstract

This White Paper *lays out a set of new research objectives and the skeleton of a plan on how to achieve these. The objectives were motivated by the desire to take the successful Selective Tuning model of visual attention (ST) and move it to the next stage of its natural development. A key missing element of ST is an executive controller, a component that uses the attentional processes that Selective Tuning provides in solving real problems associated with visual perception, visual cognition and reasoning, including the use of vision for the guidance of action. To this end, the conceptualization of Ullman's Visual Routines seems to provide the best starting point. This extremely brief* White Paper *presents a proposal for how Ullman's work may be re-examined in the light of an up-to-date understanding of visual attention and visual processing more generally. Conceptually, Ullman's contribution was significant but he left most of the details unspecified. This proposal suggests ways to update Ullman's visual routines concept given modern views on vision and attention, and moves to subsequently revise, extend and flesh-out the ideas in order to provide the functionality required to develop an executive controller for ST.*

[1] Also, Centre for Vision Research, York University; author contact: tsotsos@cse.yorku.ca

Research Objectives

The dream of useful robots, such as companion robots for the elderly, autonomous vehicles or flexible manufacturing robots that interact with human workers, requires major advances in the computational realization of vision and visual cognition. Pessig & Tarr (2007) conclude that behavioral and neuroscientific research must be better grounded in models of object recognition and computational models must be better grounded in empirical data. Dickinson (2009) reaches a similar conclusion specifically advocating a stronger emphasis on shape, shape abstraction and viewpoint control within recognition paradigms. Such a broad interdisciplinary interplay has been at the heart of our Selective Tuning model of visual attention which has now received strong experimental support. Further, it has been extended to show how attention interacts with recognition and with feature binding.

A proposal on the inter-relationships among visual attention, object (and event) recognition and visual feature binding within an overall framework for visual perception was presented in Tsotsos et al. (2008) and in Tsotsos (2011). Key elements include:

a) attention is a set of mechanisms that tune the search processes in vision to achieve their best performance to a given task (even in free viewing);

b) an overall visual processing network that can solve a particular class of vision problems very quickly but can be tuned dynamically to adapt its performance to the remaining sub-classes of vision problems but at a cost of greater time to process;

c) the greater processing time is realized by iterative feed-forward and recurrent processing passes through the network, each primed by task or sub-task information to achieve particular goals or sub-goals;

d) top-down tracing of neural activations supplemented by local, within receptive field, surround suppression throughout the network plays the role of localization of an attended stimulus.

This Selective Tuning (ST) theory (Tsotsos et al. 1995), has received strong experimental support (summarized in Tsotsos et al. 2008, Hopf et al. 2010; Tsotsos 2011). However, on its own it is certainly insufficient. A critical component that is missing is the 'executive control' function for attention, and indeed for vision as a whole. ST shows how intermediate representations may be created, it details a set of mechanisms that together comprise what is commonly called attention, but does not show how they may be used in visual reasoning or in solving the kinds of tasks an autonomous agent may face. How can an agent use task requirements and construct the proper sequence of operations (sensing, gaze and viewpoint changes, attention, classification, identification, localization, etc.) in order to fulfill that task? This exact problem has been considered previously in the concept of a visual routine (Ullman 1985).

Ullman's visual routine (VR) idea has been very influential. However, Ullman based his conceptualization on knowledge of human vision and attention of the early 1980's, and most followers of VRs do the same. Among them, Ballard & Hayhoe make strong arguments about the need for non-saliency methods (the ones Ullman used) for attention but do not propose an alternative (Ballard & Hayhoe 2009).

It is important at this point to stress that our perspective is to consider vision as much more than a classifier. Classifiers as they are commonly used in computer vision may be useful for solving what I call the 'at a glance' visual problems, such as detection or categorization in uncluttered scenes. But everyday vision as we all experience it involves much more. We may call these the 'more than a glance' problems (problems where simple detection or naming of a single item in a scene does not suffice, eg., tracking objects in time-varying scenes, questions about image contents, Non-pop-out visual search, where a behavior is required such as an eye movement or a manipulation). It is not an exaggeration to claim that most of our visual behavior is of the more than a glance variety.

The Original Visual Routines

Ullman suggested a strategy for how human vision extracts shape and spatial relations. He proposed:

- VRs compute spatial properties and relations from *base representations* to produce *incremental representations*;
- VRs are assembled from *elemental operations*;
- new routines can be assembled to meet processing goals;
- different VRs share elemental operations;
- a VR can be applied to different spatial locations;
- mechanisms are required for sequencing elemental operations and selecting the locations at which VRs are applied;
- VR's can be applied to both base and incremental representations.

Ullman proposed that *universal routines* operate in the absence of prior knowledge. His elemental operations were: shift of processing focus, indexing, boundary tracing, marking and bounded activation. The base representations are derived automatically, are assumed correct, and describe local image properties such as color, orientation, motion and depth (Marr's 2½D Sketch, 1982). In this regard, the work is based completely on how Marr viewed visual processing in the brain. Attentive operations are critical and based on Koch & Ullman (1985). As result, the saliency map ideas play a central role for Ullman, understandably. Ullman outlined 4 aspects: elemental operations, how they are integrated into routines, VR control; creation of new task-based VRs. He only detailed the first.

How we Understood Attention and Vision in 1984

The dominant relevant papers were Feature Integration Theory (Treisman & Gelade 1980), Marr's theory of vision (1982) and The Saliency Map model of visual attention (Koch & Ullman 1985), itself strongly motivated by Feature Integration Theory. Attention involved a spotlight that selected regions of interest in images based on peaks in a conspicuity representation. With respect to visual perception, Marr claimed that feedforward processing, through a set of independent modules, suffices and results in representations termed the primal and 2½D sketches. He believed these reflected the result of the first 160ms of human vision; he did not consider processes beyond this time limit. Marr's theory makes a strong assumption: that images be quickly and easily segmentable.

VR's Since Ullman

A number of researchers have pursued the visual routines concept, but not as many as one might think. Johnson (1993) and McCallum (1996) looked into how VRs may be learned, using genetic programming and reinforcement learning. Horswill (1995) developed a system that performs visual search to answer queries in a blocks world. He included a set of task-specific weights to compute a saliency map, a set of markers that hold the centroids of regions, and a return inhibition map that masks out regions that should not be selected. Brunnström et al. (1996) propose an active approach including an attentional mechanism and selective fixation. They define VRs that can rapidly acquire information to detect, localize and characterize features. Ballard et al. (1997) emphasize the need for an attentive 'pointing device' in visual reasoning. Rao's (1998) primitive VR operations are: shift of focus of attention; operations for establishing properties at the focus; location of interest selection. These enable VRs for many visuospatial tasks. Ballard & Hayhoe (2009) describe a gaze control model for event sequence recognition. They highlight problems with saliency map methods for task-based gaze control.

Biological research has also embraced VRs. Roelfsema et al. (2000; 2003) and Roelfsema (2005) have provided neurophysiologic support. They discovered neurons in motor cortex selective for movement sequences. They also monitored the progression of a sequence by recording activity of neurons in early visual cortex, associating elemental operations with changes in neuron response. They thus suggested an enhanced set of VRs: visual search, cuing, trace, region filling, association, working memory, suppression, matching, motor acts. Cavanaugh et al. (2001) found that discrimination of motion patterns demands attention. They consider a *sprite* as the set of routines that detects specific motion.

VRs also found utility in practical domains: control of humanoids (Sprague & Ballard 2001); autonomous driving (Salgian & Ballard 1998); natural language interpretation and motor control (Horwsill 1995); control of a robot camera system (Clark & Ferrier 1988).

Methods and Proposed Approach

Notwithstanding the previous briefly mentioned works, progress on systems with visual cognition that visual routines may enable remains elusive. Perhaps the lack of progress is due to: 1) all authors conclude that attention is required for VR's however the conceptualization of attention used is the saliency map idea that accounts only for gaze change, not the full breadth of attention - Ballard & Hayhoe make this point too; 2) the premise that one feed-forward pass through the visual hierarchy is sufficient to provide all required information is pervasive, yet true only for special cases and clearly not for scenes where viewpoint change or visual search are required (Tsotsos et al. 2008). Our first step is to see how these two issues impact the concept of VR's. In the following, the broader VR functionality suggested by the union of Ullman, Roelfesma, Cavanaugh and Ballard & Hayhoe is used. This is a novel and innovative perspective, based on our ST model, an approach to VR's not previously considered.

Almost everything has changed in our knowledge of vision and attention since 1984. We know that attention is more complex than region-of-interest selection for gaze change. It also involves top-down priming of early visual computations, feedback processing, imposes a suppressive surround around attended items to ignore background clutter and modulates individual neurons to optimize them for the task at hand both before the stimulus is presented as well as during its perception. Attentive modulation changes the operating characteristics of single neurons virtually everywhere in the visual cortex (see Itti et al. 2005). None of these are seen in the previous computational research on VR's (but all exist within ST). Moreover, we know the time course of attentive effects differs depending on task; attentional effects are seen *after* Marr's limit of 160ms. Further we now know there are no independent modules, as Marr beleived, because most neurons are sensitive to more than one visual modality/feature. We also know that the feedforward pass of the visual cortex has strict limits on what can and cannot be processed. It is not the case that this feedforward pass, as Marr had thought, suffices to compute a complete base representation on which any additional reasoning can take place. If anything, that feedforward pass is only the beginning of the act of perception. Basic feedforward visual tasks have been proved intractable in their most general definition refuting Marr's belief in a complete feedforward early visual process and a passive, fixed visual cortex has been refuted. (For justification of these claims see Tsotsos & Bruce 2008, Tsotsos et al. 2008, Pessig & Tarr 2007, Dickinson et al. 2009 for reviews). This research will re-visit VR's in this context. We address the following in this research program:

1) How do Ullman's VR's fare with a current understanding of human visual processing and attention?

We will first examine how the conception of attention in VR's (recall, we mean the broader concept of VR) changes with a current view of attention and vision processing. Simultaneously, we will determine the dependence of his routines on a passive, complete and constant base representation (in the Marr sense). This step highlights changes that must be made to the general concept of the visual routine.

2) Can ST play the role of attention as Ballard & Hayhoe describe?

Spatial attention in the form of a location of interest plays a central role in VR's. Ballard & Hayhoe argue that the saliency map view of attention is not the appropriate one, but do not say what an alternate choice might be. Our ST model seems to provide the elements required but does ST's conceptualization of attention fit the visual routines functionality and if not, what must change?

3) How must VR's be updated given current understanding of vision and attention?

Given the changes steps 1 and 2 reveal, we will take the general statements of the various VR's and update them with current best knowledge of vision and attention providing detailed definitions for all of the VR's mentioned above. For example, in a multi-step routine for, say, boundary tracing, given that we know that top-down priming is possible, how can the VR be primed dynamically depending on its current state? VR definitions will be modified as needed.

4) What is the best representation for a VR?

Once the VR's have been redefined, the next step is to determine how to best represent these. A start on needed representational concepts follows. The *incremental representations* will be the elements that populate *working memory*. Previous conceptualizations have not included a working memory but it is an important element for a full attentive vision system. Ullman's universal routines will be termed *methods* and when tuned to specific tasks, *scripts*. The routines will be of four classes: *task*, *sensing*, *motor*, and *reasoning* in order to cover the breadth the above authors have presented. A *task method* creates the information needed by a task script from a given task specification. *Task scripts* are responsible for tuning other methods to particular tasks and producing reasoning, sensing or motor scripts. They may employ the Restriction and Suppression mechanisms of ST. A *sensing method* represents an un-tuned (without any priming) feed-forward pass through the visual processing hierarchy. A *sensing script* represents a tuned feed-forward pass. A *motor method* is a ballistic action to move the agent (or one of its components) from point A to B, while a *motor script* is tuned by information about path, obstacles, or other motion constraints. A *reasoning method* provides a generic answer to a question about a stimulus: location, feature composition, size, shape, spatial relations, and so on. To do so, it may embody all of the attentional mechanisms of ST. A *reasoning script* is a reasoning method tuned to the task at hand. These are intended to be the direct successors to the VR's Ullman proposed for spatial relations and shape properties. Scripts and methods may link to others of the same or different class. A key question is how is it determined which methods/scripts to apply at any given time. Methods and scripts each have *triggers*, elements of incremental representations created by other methods or scripts. That is, a VR may be activated when one of its trigger elements reaches a strong enough response in working memory. Because VR's represent action sequences, a graphical model is the right representational device (see Bishop 2006, for review). *Markov Decision Processes* (MDPs), *Partially Observable MDP*'s (POMDP's) and *Dynamic Bayes Nets* (DBNs) will form the starting points of investigation. Ballard & Hayhoe (2009) and Yi & Ballard (2009) argue for the effectiveness of DBN's to represent VR's, detailing how human experimental data can program a DBN which can recognize new instances.

5) Given the new definition, can useful routines be developed and tested, and shown to have robust properties?

We will hand-code a set of VR's in order to test their operation. This test requires a sufficiently complex visual representation to replace the base representation Ullman had conceived. Fortunately we have developments on both motion and shape detection within an attentive framework (Tsotsos et al. 2005; Rodriguez-Sanchez 2010). We will follow the methodology of Yi & Ballard (2009) as a first step.

6) Can these new VR's be learned and if so how?

Ghahramani (1997) lays out several approaches to learning DBNs. Reinforcement learning (RL) seems an ideal framework in which to learn these sequences of operations because its strength is its ability to address the issues of choosing sequences of actions (Sutton & Barto 1998). Others have used RL to learn VR's in the past (eg. McCallum 1993). Mugan & Kuipers (2009) use RL for their robot control task resulting in an agent that learns a hierarchy of actions in a continuous environment; this seems closest to our needs. Our requirements include: the model structure is unknown; we can observe external actions of the agent (visual gaze changes for example) but not the internal ones (visual reasoning for example); the agent can interact with the environment (move about, take different views of objects, move things around); the agent needs to not only recognize sequences of actions that it sees but also to generate control signals from learned sequences, appropriate for the current task. We will begin with the Mugan & Kuipers work, evaluate its suitability, and extend or modify as required.

7) Can the set of new VR's be mapped onto an executive controller for ST?

In order to design an executive controller for ST using VR's, the first question that must be addressed is how is the right VR chosen at the right time? The executive may be primed by task but task progress would dictate next steps in a sequence. This element was not included in past works mostly because most past work has not considered a number of VR's working in concert. Generally, the action of the executive controller is to select/assemble the proper set of VR's, to set up the processes to tune them, to execute them while monitoring task progress, and make modifications along the way in order to fulfill the task. It is one of the key elements of VR's Ullman described but has not been yet addressed.

Evaluation and Milestones

No existing video dataset can provide a test domain for this work and thus we will create our own. Our past developments on methods for detecting, attending, classifying and localizing motion and shape will form the basis for the tests. We will develop software that generates synthetic video sequences (we have previously done this for random polyhedral scenes with excellent results (Parodi et al 1998) and will follow that methodology). The idea is this: place random 2D coloured shapes randomly in a scene, with visual field processed being a small subset of the full scene to enable gaze shifts; occlusion will be permitted but not transparency; the shapes will each exhibit a randomly chosen affine motion with a randomly chosen duration; images may be degraded by noise and/or contrast manipulations; we will follow Parodi et al. (1998) in order to ensure the statistical properties of the generator. This generator will be made public. Of this dataset, we can ask questions such as: how many shapes in an image? how many similar shapes? how many shapes of a particular type? how many near a specific one? any odd-man-outs (shape, size, speed, direction)? any above another (below, next-to)? how many of a particular type are moving? how many moving similarly? One might imagine a much larger set of similar queries that may be investigated.

The evaluation procedure will follow the strategy of Parodi et al. (1998). Random videos will be generated and grouped depending on the number of shapes in the video. The system would then answer each of N questions for each video. Ground truth can be hand-coded for spatial relations; numbers/shape types and velocities can be attached by the generating program. For the responses we will tabulate %correct, false positives/negatives, distance from correct answer, all with respect to input set size**.**

If the above evaluation is successful, more complex queries and tasks will be considered. The above will exercise the general framework but do not test the motor methods/scripts. This extension can occur within a real laboratory environment and with real robots and involve search for objects (as in Shubina & Tsotsos 2010; Andreopoulos et al. 2011) or other complex tasks involving locomotion, navigation and manipulation, such as those a wheelchair user might require (Tsotsos et al. 1998; Andreopoulos & Tsotsos 2007).

Significance

Recently, the quest to develop effective methods for categorizing objects, events or scenes from video has dominated computer vision. Progress has been tremendous (Dickinson et al. 2009); however that progress should not be considered as nearing the solution to computer vision. The role of visual perception, for both humans and robots, is not to simply categorize images or to detect a particular object. It is to enable an understanding of the physical world and to guide action and behaviour. Without the ability to visually reason about what is seen, this goal cannot be realized. In natural human behavior, we need to know not only what is in a scene, but also where it is, how is it related to the other elements of the scene, how things change over time, what would the effects of a manipulation be, how to achieve a behavior within that scene, etc. How do we know where on a table to put down a cup, select the shortest checkout queue in a store, look for moving vehicles before we cross a road? These questions require a capacity to reason about the scene, a capacity that must be integrated within a system that can detect and analyze scene elements in the context of the current task and the agent's world knowledge. This research contributes to this goal in a totally novel manner. Applications in surveillance, autonomous mobile robotics, human-robot cooperative assembly, etc. are apparent.

Summary

This is a virtually identical document to a grant proposal submitted to the NSERC Discovery Grant Program in October 2010 (Title: Re-Visiting Ullman's Visual Routines, Appl. #4557-2011). The funding request for this research program was fully granted for a 5-year period and work began immediately.

As a consequence, in the short-term, we seek answers to:

1) How do Ullman's VR's fare with a current understanding of human visual processing and attention?
2) Can ST play the role of attention as Ballard & Hayhoe describe?
3) How must VR's be updated given current understanding of vision and attention?
4) What is the best representation for a VR?
5) Given the new definition, can useful routines be developed and tested, and shown to have robust properties?
6) Can these new VR's be learned and if so how?
7) Can the set of new VR's be mapped onto an executive controller for ST?

In the long-term, we seek to impact our understanding of both computational and human vision, with new concepts that lead to more general and effective computer vision systems and experimental predictions that add to knowledge of human vision.

References

Andreopoulos, A. (2010). Active Object Recognition in Theory an Practice, PhD Thesis, Dept. of Computer Science & Engineering, York University.

Andreopoulos, A., Wersing, H., Janssen, H., Hasler, S., Tsotsos, J.K., Körner, E. (2011). Active 3D Object Localization using a Humanoid Robot, *IEEE Transactions on Robotics*, 27(1), p47-64.

Andreopoulos, A., Tsotsos, J.K. (2007). A Framework for Door Localization and Door Opening Using a Robotic Wheelchair for People Living with Mobility Impairments, RSS 2007 Manipulation Workshop: Sensing and Adapting to the Real World, Atlanta, Jun. 30.

Ballard,D.,Hayhoe,M.(2009).Modeling the role of task in the control of gaze,*Vis.Cogn.*17(6),1185-1204.

Ballard, D. et al. (1997). Deictic codes of the embodiment of cognition, *Beh. Brain Sci.* 20, p723-767

Bishop, C. (2006). **Pattern Recognition and Machine Learning**, Springer NY.

Bruce, N.D.B., Tsotsos, J.K., Saliency, Attention, and Visual Search: An Information Theoretic Approach, *Journal of Vision*, 9:3, p1-24, 2009.

Brunnström, K. et al.(1996). Active fixation for scene exploration, *Int.J. Computer Vision*17, p137-162.

Cavanagh, P. et al. (2001). Attention-based visual routines: Sprites, *Cognition* 80, p47-60.

Clark, J.J., Ferrier, N. (1988). control of an attentive vision system, Int. Conf. Computer Vision, Tarpon Springs FL, p. 514 - 523.

Dickinson, S. (2009). The Evolution of Object Categorization and the Challenge of Image Abstraction. In S. Dickinson et al. (eds.), **Object Categorization**, p. 1–37, Cambridge University Press.

Dickinson, S., Leonardis, A., Schiele, B., Tarr, M.(eds.), **Object Categorization**, Cambridge Uni. Press.

Ghahramani, Z. (1997). Learning Dynamic Bayesian Networks, in **Adaptive Processing of Temporal Information**, Ed. by C. Giles & M. Gori, Springer-Verlag.

Hopf, J.-M., Boehler, N., Schoenfeld, M., Heinze, H.-J., Tsotsos, J.K., The spatial profile of the focus of attention in visual search: Insights from MEG recordings, *Vision Research* 50(14), p.1312-1320, 2010.

Horswill, I. (1995). Visual routines and visual search: a real-time implementation and an automata-theoretic analysis, Proc. Int. J. Conf. on Artificial Intelligence, Montreal, p56-62

Johnson, M.P. (1993). Evolving Visual Routines, B.S., Computer Science, MIT, Cambridge, MA June.

Koch, C., Ullman, S. (1985). Shifts in selective visual attention, *Hum. Neurobiol*. 4, p.219–227.

Marr, D., (1982). **Vision**, Henry Holt and Co., New York.

McCallum,A.(1996).Learning Visual Routines with Reinforcement Learning,AAAI TRFS-96-02,82-86.

Mekuz, N. (2010). Learning Geometric Local Appearance Pairs for Object Categorization, PhD Thesis, Dept. of Computer Science & Engineering, York University.

Mugan, J., Kuipers, B. (2009). Autonomously learning an action hierarchy using a learned qualitative state representation, Proc. Int. Joint Conf. on Artificial Intelligence, 2009

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

Peissig, J. Tarr, M. (2007). Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago? *Annual Review of Psychology* 58, p.75-96.

Roelfsema, P. et al. (2000). The implementation of visual routines, *Vision Research* 40, p1385-1411.

Roelfsema, P. (2005). Elemental operations in vision, *TRENDS in Cognitive Sciences* 9(5), p. 226-233.

Roelfsema, P. Kyahat, P. Spekreijse, H. (2003) Subtask sequencing in the primary visual cortex, *Proc. National Academy of Sciences* 100(9), p. 5467-5472.

Rodriguez-Sanchez, A.J. (2010). Intermediate Visual Representations for Attentive Recognition Systems, PhD Thesis, Dept. of Computer Science & Engineering, York University.

Rothenstein, A.L. (2010). Beyond the Limits of Feed-Forward Processing: Visual Feature Binding and Object Recognition, PhD Thesis, Dept. of Computer Science & Engineering, York University.

Sprague, N., Ballard, D. (2001). A Visual Control Architecture for a Virtual Humanoid, Proc. IEEE RAS International Conference on Humanoid Robots, Tokyo.

Salgian, G., Ballard, D. (1998). Visual Routines for Autonomous Driving, Int. Conf. Computer Vision, Bombay, p. 876 - 882.

Shubina, K., Tsotsos, J.K. (2010). Visual Search for an Object in a 3D Environment using a Mobile Robot, *Computer Vision and Image Understanding*, 114, p535-547.

Sutton, R., Barto, A. (1998). **Reinforcement Learning: An Introduction**, MIT Press, Cambridge MA.

Rao, S. (1998). **Visual Routines and Attention**, PhD Dissertation, MIT, February.

Treisman, A.,Gelade, G.(1980). A feature-integration theory of attention. *Cogn. Psychol.*12(1), p97-136.

Tsotsos, J.K., Bruce, N.D.B., Computational foundations for attentive processes, *Scholarpedia* 3(12):6545, 2008.

Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F. (1995). Modeling visual attention via selective tuning, *Artificial Intelligence*, 78(1-2), p507-547.

Tsotsos, J.K., Verghese, G., Dickinson, S., Jenkin, M., Jepson, A., Milios, E., Nuflo, F., Stevenson, S., Black, M., Metaxas, D., Culhane, S., Ye, Y., Mann, R. (1998). PLAYBOT: A visually-guided robot to assist physically disabled children in play, *Image & Vision Computing Journal*, 16, p275-292.

Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K., Attending to Visual Motion, *Computer Vision and Image Understanding* 100(1-2), p 3 - 40, Oct. 2005.

Tsotsos, J.K. Rodriguez-Sanchez, A., Rothenstein, A., Simine, E., Different Binding Strategies for the Different Stages of Visual Recognition, *Brain Research* 1225, p119-132, 2008.

Tsotsos, J.K. (2011). **A Computational Perspective on Visual Attention**, The MIT Press.

Ullman, S. (1985). Visual Routines, *Cognition* 18, p97-159.

Yi, W., Ballard, D. (2009). Recognizing Behavior in Hand-Eye Coordination Patterns, *Int. J. Humanoid Robotics* 6(3), p. 337-359.