

Chapter 7

Conclusion

Fixation locations are not completely deterministic, yet they are also not made to random locations in the field. This fact is obvious even without extensive eye-tracking studies, however, little work has been done to discover how exogenous and endogenous factors interact to determine the target of the next saccade. This is particularly true when considering high-level factors such as motivation, prior learning, and experience, where the visual system is used primarily as a tool to monitor, assist, and assess the immediate environment during ongoing activity. For this reason, an accurate description of the high-level parameters that determine how attentional resources will be allocated cannot be considered in isolation of the environment, or of the current task under execution.

Land *et al.* (1999) described visual behavior during a tea-making activity, and found that the eyes monitor and guide virtually every action that is necessary to complete the task of making tea. The visual salience (*i.e.*, color, luminance contrast, texture) of

objects was not important for determining fixation locations. Rather it was the object's relevance to the task that accurately predicted the saccadic landing position. This is a consequence of the goal-driven behavior of people, as Yarbus (1967) noted earlier. The visually salient properties of objects may be important for determining which object is fixated next for free-viewing scenes, but not for strategic behavior that requires formulating a plan of action.

Chapter Four found that certain simple visual routines can be characterized by low-level eye movement metrics such as fixation duration and saccade length. Reading and counting, which are characterized by short fixations and short saccades, are active tasks that require a high level of visual engagement with the environment, but little strategic planning. Having a conversation, which is characterized by long fixations and intermediate saccades, is an active task that allows for disengagement from the surrounding and requires no planning. Sorting is an active task that requires both visual engagement as well as formulating a plan of action. Intermediate fixations and intermediate to large saccades characterize this type of routine.

The results from Chapter 4 emphasize the need to study active visual tasks in the context of a real and extended environment, rather than a spatially restricted environment such as a laboratory setting. The finding that walking along a hallway elicits different eye movement characteristics from reading text is an indication that eye movements are closely tied to the type of task under execution. Therefore, any attempt to draw conclusions about oculomotor behavior must necessarily include information about the task, as well as the environment. It is likely that results from eye movement studies performed within the confines of the laboratory reflect the conditions and restraints

imposed upon the subject during the experiment, and are indicative of the experimental design, rather eye movement behavior in general. Bringing the subject out of the laboratory and into the real world provides the opportunity to study human vision as an evolutionarily useful sense, an aide for survival, *i.e.*, as a tool, rather than a task.

A two-dimensional feature vector was proposed in Chapter 4 to classify simple visual routines according to the level of visual engagement required and the need for strategic planning. This classification scheme allows one to predict fixation duration and saccade amplitude based on the requirements of the task. For example, having a conversation is an active visual task that requires little visual engagement with the surroundings and no strategic planning. Most of the time the eyes are directed toward either the face of the conversational partner, or irrelevant objects in the surrounding area. Disengagement from the surroundings allows the mind to pursue thoughts related to the conversation, rather than to details of the surroundings. The consequence of disengagement is that the eyes can dwell longer on non-essential objects and saccades can be directed toward irrelevant or random locations in the scene. This is in agreement with the findings of Chapter 4, which showed that fixation durations are long and saccade amplitudes range from medium to large during conversation. This is the case for both having a telephone conversation and having a face-to-face conversation.

At the other end of the engagement spectrum, both reading and counting are active visual tasks that require constant visual engagement with the environment, yet similar to conversation, they require no strategic planning. Short fixations and small saccade amplitudes characterize both of these tasks, possibly reflecting the ‘pre-programmed’ nature of these tasks.

Sorting is an active visual task that requires both visual engagement as well as strategic planning. There is little time for contemplative thought, yet the task requires formulating a plan of action concurrent with task execution. It is possible that intermediate fixation durations are indicated here because short fixations will not allow enough time for awareness of relevant object attributes, yet long fixations would distract from the efficient completion of the task.

Table 7-1 shows how the tasks are classified into the two-dimensional feature vector, and the associated characterization in terms of fixation duration and saccade amplitude.

	Engaged	Disengaged
Planning	→ FixDur ↑ SaccAmp Sorting	↑ FixDur ↑ SaccAmp ?
No Planning	▼ FixDur ▼ SaccAmp Counting Reading	↑ FixDur → SaccAmp Conversation Hallway walking

Table 7-1 Classification of tasks into feature vector corresponding to both the level of visual engagement with the environment and amount of strategic planning required.

The mean values of fixation durations and saccade amplitudes from Chapter 4 characterize the tasks, and correlate to the placement of the tasks in the feature vector. Up arrows in Figure 7-1 indicate a high value, down arrows indicate a low value, and sideways arrows indicate intermediate values.

A prediction of the proposed classification scheme is that a task that allows for disengagement with the environment, yet requires strategic planning (upper right box in Table 7-1) would be associated with long fixations and large saccade amplitudes. It would be similar to both sorting and having a conversation in those respects, yet opposite from counting and reading. The verification of this prediction is a topic for future study.

The higher level aspects of visual perception were considered in Chapter 5. This was accomplished by bringing the subject even further out of the laboratory and into the real-world environment. The results from the experiment discussed in Chapter 4 were used and analyzed for higher-order eye movement metrics, and prompted the design and execution of a second eye-tracking experiment. The second experiment was an attempt to understand how the task instruction alters the perceived conspicuity of objects in the scene, especially in the context of a real-world activity.

The sorting blocks and sorting cards tasks from the experiment of Chapter 4 were analyzed, as were the two additional tasks of copying a block model when the resource, model, and workspace were located in the same room, and copying the same model when the resource and workspace areas were located in one room, and the model was located in a different room. In general, it was found that the percentage of time spent looking in a particular region of a scene depends upon the task.

During sorting blocks the subjects spent a much higher percentage of time looking at the resource area than for any of the other tasks, perhaps as a result of the ability to make use of low-resolution peripheral vision to acquire information about the workspace. The workspace was fixated primarily on the occasion of forming a new group of sorted blocks.

During sorting cards, on the other hand, subjects spent a higher percentage of time looking at the workspace area than for any of the other tasks. This is probably a reflection of the need for strategic planning and decision making for this type of task, as well as the need for more physical manipulations of the cards in the workspace area.

For both of the copy-model tasks, subjects spent approximately an equal amount of time looking in the workspace area and the model area, and looked in the resource area the least. The high memory demands placed upon the subjects during the different-room-copy task was reflected in the finding that they spent more time looking in the workspace and model area, and less time looking at the resources, than for the same-room-copy task.

The second experiment of Chapter 5 was motivated by a desire to perform a similar analysis in a real-world setting. The percentage of time spent looking at particular objects in the environment was used as the basis for the analysis. For this experiment, task-differences were in the form of different instructions imposed upon the subject in a particular environment. Each of the four real-world environments (washroom, hallway, office, and vending machine) showed the same trend – a single task-relevant object dominated the total fixation time, followed by other task-relevant and future-task-relevant objects. Non-task relevant objects were occasionally fixated, for example, the floor was fixated 13 %- 17% of the time for each of the hallway tasks. The floor is neither task-relevant nor is it visually salient, yet subject still spent 1/6th of their time looking there. This is in agreement with the findings of Chapter 4 that suggest that hallway walking is a visually disengaged activity that requires little or no planning; perhaps this is true even when a task has been imposed.

People have a strong tendency to fixate regions of the visual field that are relevant for a current or future goal. Moreover, fixation locations are not only scene relevant, but they are highly task-relevant as well. Altering the instruction while the visual scene remains the same produces a dramatic change in the percentages of fixations on certain objects in the environment. These findings lead one to conclude that certain objects in the environment have a heightened perceptual conspicuity due to their perceived relevance, or importance, to the current task at hand. This is true for objects of current interest, as well as for objects of potentially future interest and objects that might be mistaken for relevant objects. The task-relevancy of perceptual conspicuity contributes to the modulation of visual saliency due to task demands. It is hypothesized here that modulated saliency due to task-relevancy is a far better indicator of where people will look in a scene than non-modulated saliency from the purely physical properties of the scene alone.

Are the eyes essentially passive “cameras,” capturing images of the world onto the retina and passing this information along to the brain? The evidence presented in Chapters 4 and 5 strongly suggests that the visual system is not passive, nor is it general-purpose, but rather it is active and specific, tightly coupled to the requirements of planned behavior and action. Based on this conclusion, one may speculate that humans have evolved an active, task-specific visual system for the purpose of enhancing the probability of survival in a complex, continually changing environment under the restrictions of limited neural processing capabilities and an imperfect sensorimotor system.

One implication of describing the human visual system as active and task-specific is that the design can be replicated on an artificial visual system, with large potential savings in computational efficiency. Developers of artificial vision systems for robotics and surveillance applications must contend with the limitations imposed upon their designs by currently available technology. Hardware constraints include a finite memory capacity and cache size, finite processor cycle speeds, and bandwidth limitations for networked systems. Software constraints include program development costs and time, code maintenance, error detection, security issues, and platform portability. Large scale video image processing coupled with the requirement of a real-time response from the processor can make any but the simplest of visual capabilities prohibitive. These limitations underscore the need to simplify computations as much as possible and develop highly efficient algorithms for implementing visual systems in machines.

One way to reduce the amount of required computation is to emulate the human visual system and eschew an explicit and detailed representation of the environment. Evidence gathered from this research effort supports the conclusion that the human visual system makes extensive use of strategies that simplify and reduce the cognitive load required for visual processing. These strategies can form the basis for a biologically-inspired computational model of visual perception that selectively reduces the input to only that which is necessary for any given task.

Representing the scene as a topographic map of relative perceptual conspicuity is a means for providing a selective mechanism for an active artificial visual system. The advantage of using a topographic map of conspicuity values, rather than the actual scene, is that the scene can be represented compactly, with only the most highly conspicuous

regions being selected for further processing. Once the highly conspicuous regions have been selected, processing can proceed to object identification and image interpretation. The goal is to make optimal use of limited processing resources, with little to no loss of accuracy in the result.

There are clear advantages to the idea of using conspicuity maps at the pre-processing stage of image understanding for artificial systems. Is there physiological evidence to support the hypothesis that humans make use of a pre-cognitive topographic map of conspicuity values? In support of this idea, Moran and Desimone (1985) provide physiological evidence showing that when subjects do not attend to effective stimuli (effective in the sense that the stimuli produce a high response when presented in the neuron's receptive field) the neuron does not fire. This is surprising because an effective stimulus presented within the receptive field should cause the cell to fire – that is why it is termed an effective stimulus. The authors conclude that the failure to attend to the effective stimulus de-activated the neuron's typical response pattern.

In terms of conspicuity maps, this finding can be interpreted as indicating that failure to attend to the highly salient low-level properties of the scene will prevent the activation of the “inherent” salience of those regions. The absence of focused attention on those regions essentially de-activates their saliency. Focused attention on task-relevant objects may be the “glue” that binds together the low-level salient features of the scene with awareness to facilitate object identification and understanding. This ultimately promotes our perception of a continuous, coherent reality in the presence of an overabundance of visual stimuli coupled with limited processing capabilities.

The main objective of Chapter 6 was to introduce a biologically-plausible model of selective visual perception that correlates well to fixation locations in natural scenes. A low-level saliency map, modeled after the saliency maps of Itti & Koch (2000), and Parkhurst, Law, and Niebur (2002), does not produce a strong correlation to fixation locations when used alone. A higher-level proto-object map that identifies regions in the image that contain potentially useful objects was shown to have a much higher correlation. A perceptual conspicuity map that merges together the low-level saliency map with the higher-level proto-object map and includes an inherent location bias was found to have the highest correlation of any of the maps considered. The location bias promotes the central region of an image only when such a bias is warranted, *i.e.*, only when objects are located there, and demotes the central region in the absence of centrally located objects.

Chapter 6 showed that there is virtually no correlation between the low-level salient properties of natural images and fixation locations. People simply do not look at something unless there is a need to do so. In terms of the activation theory described above, the salient properties of any particular region are promoted to the level of awareness only after they have been coupled to a desirable object. Even though feature salience is an inherent property of the image or scene, it is the location of the object that determines if and how the salience is used. Thus, perceptual conspicuity can be described as the modulation of saliency due to task preference for certain objects.

Extensions to the perceptual conspicuity model introduced here should include a spatial frequency reduction toward the periphery of the image when the central bias is warranted, to simulate the fall-off in visual acuity in the non-foveal regions of the retina.

The central-bias issue should also be examined with a broader range of images that include more non-centrally located objects and a wider range of scenes. Variable block sizes for the convolution kernels, as well as an adaptive window size for the expected location maps would also warrant further study. In addition, the size of the fixation location window used for the calculation of the F/M ratio should be allowed to vary over a greater spatial range in the image in order to promote flexibility across a wider range of image content. An enhancement of the top-down expected location module could include a Bayesian network that takes into account evidence from the scene and prior knowledge about the imposed task to reason about a fixation strategy.

In conclusion, this thesis found that locating highly conspicuous regions of an image or scene must ultimately take into consideration the implicit semantics of that scene – that is, the “meaningfulness” of the contents for the person viewing that scene. This thesis suggests that objects and their locations in the scene play an important role in determining meaningfulness in natural, task-oriented environments, especially when the environment is approached with a required action or an action-implied imperative.

Successfully predicting fixation densities in images requires computational algorithms that combine bottom-up processing with top-down constraints in a way that is task-relevant, goal-oriented, and ultimately most meaningful for the viewer. An artificial visual system that emulates the capabilities of the human visual system can take advantage of the ability to successfully predict fixation locations by eschewing an explicit representation of the scene in favor of a limited representation that takes into account information mostly from the fixation location. Overall, the advantages are compelling to warrant further study.

