Temporal Segmentation for Faster and Better Learning

Brad Wyble (bwyble@gmail.com)
Pennsylvania State University Psychology Department
University Park, PA, 16802 USA

Howard Bowman (hb5@kent.ac.uk)
University of Birmingham, School of Psychology; also University of Kent, School of Computing
Birmingham, UK; Canterbury, UK

Abstract:

The human visual system faces an extraordinary challenge in building memories from a continuous stream of visual data without the opportunity to store it and process offline. This suggests a crucial role for visual attention in attending to specific moments in time. This project outlines key data and theories related to human temporal attention. The focus of this submission is on bridging the divide between the human visual system and artificial models by explicating segmentation mechanisms that accelerate the ability to learn the structure of the world through continuous vision. A computational model that simulates the temporal dynamics of visual attention in humans indicates a role for attention in on-demand temporal segmentation of incoming information at the sub-second scale. We predict that in humans this segmentation plays a key role in 1) simplifying visual information for learning about object kinds through compression 2) segmenting information from neighboring fixations and 3) encoding the temporal sequence of events. Such segmentation is likely to also play a key role in allowing artificial systems to learn from visual input in an online fashion, even though their specific temporal constraints are not shared with biology.

Keywords: temporal segmentation; visual attention; video processing; attentional blink

Temporal Attention in Humans

One of the greatest challenges facing a cognitive system is the demand to focus a limited set of decoding, decision-making and memory-encoding processing resources effectively on the most valuable subset of data. This is the attention dilemma; it is ubiquitous across cognitive systems and exists at multiple layers within a given problem. In the context of cognition, attention is most clearly defined as the set of processes that determine *what information should be discarded*. Attention is crucial for allowing perception and decision-making systems to keep pace with incoming sense data but also, as will be argued here, to accelerate the long-term acquisition of knowledge about *real kinds* and causational interactions between them.

It is necessary to consider attention as having a myriad of manifestations in a cognitive system. For example in human vision, starting from the retina, the concentration of cones at the fovea is a form of attention that privileges the reception of photons at that location and enables the use of eye movements to select information at a time scale of 1-5 samples per second. This physical instantiation of visual attention is complemented by covert attentional mechanisms within the brain itself (i.e. in ‘software’). These mechanisms can be further divided according to whether they select spatial regions of the visual field, or emphasize particular moments in time.

The emphasis here will be on the evidence for temporal attention in humans and discussion of its computational advantages for learning in both natural and artificial systems. While there has been a tranche of attention models in computer vision in recent years, they have, but for a few exceptions, focused on the problem of spatial attention, i.e. at any given moment, what spatial regions of an image plane should be subject to enhanced processing. Models that utilize temporal attention typically extract ‘key frames’ from a short video sequence to increase the accuracy of video decoding during supervised training. While useful, these techniques lack application to the problem faced by a cognitive agent, which is how to focus attention in real-time processing during a continuous and unending stream of visual input. This is the problem faced by the human visual system, particularly during developmental years when the visual system is being tuned to learn the constellations of features that correspond to object kinds. To this end, research on human psychophysics has identified empirical phenomena that are related to the problem of deploying attention to, and withholding attention from, particular moments in time.

The Timing of Visual Experience

When visual input enters the retina, it is transmitted to the brain through subcortical structures in a span of 50ms. Beyond this point, information perfuses through cortical and subcortical routes that transform it into a variety of representations, storing some of them into memory, and also driving decisions and actions. Detection of a highly pertinent stimulus evokes a coordination of activity across widespread regions of cortex, and also occupies a brief focus in visual awareness. It is thought that this moment of awareness enables the updating of online memory representations (Polich 2007).

The duration of this coordinated event has been measured with scalp recordings of the EEG, revealing the most canonical of brainwave findings, the P3 or P300 (Figure 1). This electrical waveform occurs at roughly 300-800ms following the visual onset of an oddball stimulus (i.e. a stimulus that stands out from the background sensory data by virtue of its physical characteristics) or a stimulus that matches current goals. It is interesting that the duration of the physical stimulus that evokes a P3 has little effect on the duration of the P3 itself. Even a 30ms duration event elicits a P3 that lasts for about half a second. The P3 shown here is in response to a single letter target, and similar findings are obtained from nearly any visual stimulus that is presented in such a way as to evoke a moment of visual focus and memory encoding.

The Attentional Blink

Along with the P3, visual targets evoke a corresponding effect in behavior. Detecting a target in a visual search evokes a temporal gap in the ongoing deployment of attention, which is referred to as an attentional blink (AB; Raymond, Shapiro & Arnell 1992; Chun & Potter 1995). The gap can be measured by placing two targets in sequence, and observing that the second target is hard to detect when it follows the first in a particular temporal window (Figure 2) or fixation sequence (Adamo, Cain & Mitroff 2013). The width of this window is typically on the order of 300ms, but extends up to a second in duration for targets that bear more information (Ouimet & Jolicoeur 2007). The AB is one of the most robust effects in cognitive psychology, having been replicated many times, and across many visual stimulus types including letters, shapes, words, colors, and images.

What is paradoxical about the AB is that the second target is easy to see when it occurs within about 100ms following the first item, an effect known as lag-1 ***sparing***. Sparing reflects a temporal window in attention that allows multiple items to be encoded into memory provided that they occur within a rapid temporal sequence. Sparing is theoretically important because it shows that the attentional blink is not an effect of depleted cognitive resources, but may instead be the signature of temporal segmentation, i.e. a means to insert temporal gaps into the processing pipeline to discretize the continuous stream of visual sense data.

Computational theories of Segmentation

Computational models have played a crucial role in formulating Insights from data regarding temporal attention. In the eSTST model (Wyble, Bowman & Nieuwenstein 2009), the attentional blink results from an inhibitory process that throttles the ongoing

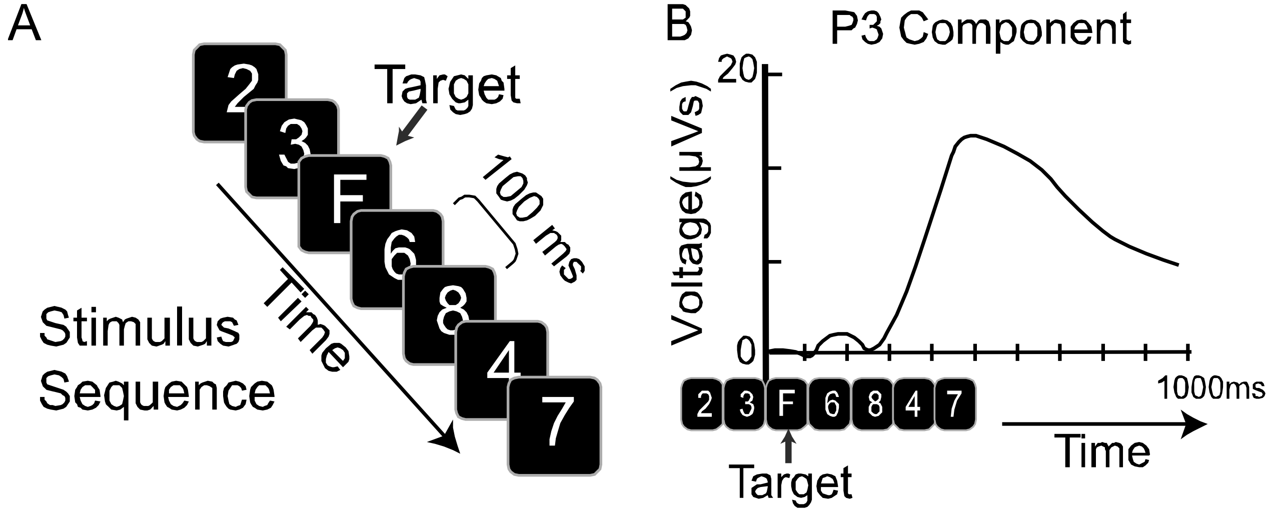

Figure 1. **a.** Example experimental paradigm in which subjects view a sequence of stimuli looking for targets with each display replacing the previous at 100ms. In this case, subjects would look for a letter among digit distractors although nearly any kind of intuitive task can be used (e.g. looking for images of vehicles, or specific animals). **b.** Representative example of P3 component elicited by the paradigm shown above. The line depicts the average voltage on a scalp electrode near parietal cortex. At about 300ms after target onset there is a sharp increase in voltage with a slowly decaying envelope that is thought to reflect the update of memory representations. Note that the stimulus has been removed from the display 200ms prior to the onset of the P3.

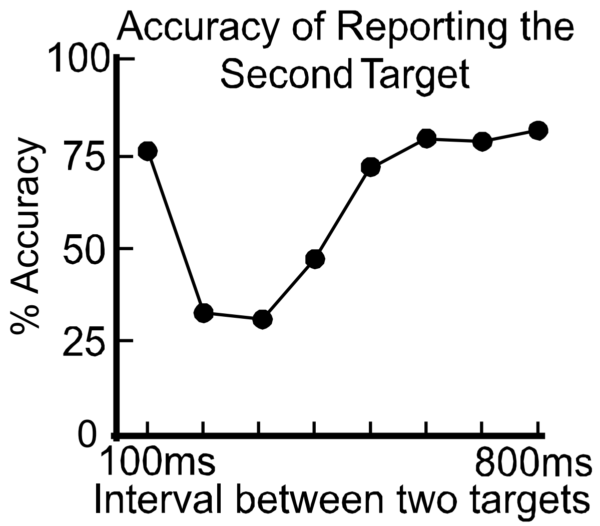

Figure 2. Classic example of the *attentional blink* (Chun & Potter 1995). The paradigm is like Figure 1a except that two targets are shown, separated by 0-7 distractors. A 100ms interval means that the two targets are directly adjacent (since the first target is 100ms in duration). Shown is the mean accuracy of correctly reporting the 2nd target's identity from all trials in which the first target was also reported correctly.

deployment of attention while the mind is actively engaged with creating a memory. Perceiving a target triggers an attentional window that initiates the construction of a memory by copying information from a sensory register, such as visual cortex, into a general purpose memory store (i.e. working memory). It is also proposed that such memory encoding is the neural generator of the P3. In the model, the P3 component can be simulated as the sum of (post-synaptic) neural activity within a distributed pool of gating neurons

At the heart of the model is a competitive inhibition circuit that controls the timing of attentional windows. When targets appear in rapid sequence, the circuit simulates a window of attention that encompasses the targets. During this brief episode, attention is strongly engaged, allowing multiple pieces of information to be rapidly stored in memory, although at the cost of lost temporal order information.

However,when targets are separated by a brief gap, the circuit initiates an attentional blink, which is a temporal interval in which attention is sluggish. This delays the encoding of subsequent information, to keep it temporally distinct from the preceding information.

Models such as the eSTST explicate mechanisms for shaping the temporal profile of attention. When two items are presented in close temporal proximity, they are stored simultaneously into a joint representation that sacrifices their individuality. When they are separated by 200ms or more, the suppression of attention delays the encoding of the second target. Thus, the attentional blink reflects a mechanism that either groups, or separates sense data impressions, according to their temporal separation.

Temporal Segmentation and Learning

eSTST provided a series of predictions about how well participants will remember the identity and temporal order of a series of visual stimuli, many of which have been confirmed experimentally. However the more important predictions of this theory relate to learning, and these have been untestable using experimental methods in the laboratory. The ***big idea*** here is that temporal segmentation is advantageous during learning, particularly for children. In other words, it may be that the primary role of temporal segmentation in vision is not to improve perception in the moment, but rather to accelerate the learning of regularities about the visual world at the developmental time scale.

To understand this idea it is important to remember that the set of visual patterns reflected by a given kind of object have clear regularities that cluster tightly in the space of possible sensory signals. It is these regularities that drive deep learning's effectiveness at mapping pixels onto categories. When a given object is visually perceived, it will project a representation into the visual system, and the brain must learn the patterns of regularities that coincide with each kind of object.

The second thing to remember is that the visual system samples visual regions in rapid succession, with eye movements occurring at approximately 200-500ms intervals. Thus, clusters of sense data from objects in the environment are sampled sequentially and this sequential sampling is, presumably, the set of regularities that shape visual object fluency in the developing visual system.

In traditional DNN training algorithms, segmentation is provided by training on isolated images, but real-world visual input is temporally continuous. Visual saccades provide some degree of segmentation, but their frequency is likely too fast to eliminate interference between distinct representations. As described above, processing even a simple stimulus event requires on the order of 500ms to complete. Therefore, information from neighboring visual fixations would likely intersect, reducing the ability to learn which clusters of features go together. Even when the world is relatively static, the sequence of eye movements will project a rapid series of distinct object representations onto the visual system, which is likely to impair the ability to learn the boundaries between those representations.

The other key challenge faced by the developing visual brain is that information must be decoded in an on-demand fashion. Decisions of what information should be stored vs discarded, and how to separate information into discrete events must be made nearly immediately, rather than in an offline analysis. It is our guess that the attentional blink reflects an on-demand temporal segmentation process. This approach allows saccades to sample information rapidly without being constrained by the slower processing of information of higher order cognitive mechanisms. Thus, the execution of visual saccades can be partially decoupled from the rate limit of higher order processing.

Our modelling work suggests that temporal attention provides the following benefits for the learning of how visual features map onto specific object representations and also to learn temporal sequence information.

Information Compression: The theory implies that human vision bundles information acquired from an attentional window, collapsing features across time in a temporal analog of a convolutional network's spatial pooling function to reduce the dimensionality of the learning process. A further compression is achieved in that information from many fixations is identified but discarded without being encoded into memory at all. *Thus, segmentation makes the learning problem more tractable by reducing stored data.*

Temporal Segmentation: The mechanism that produces the AB will reduce the overlap between sequential samples from the environment. This allows the eyes to sample new information while higher-order cognitive processes are still processing data from previous fixations. The new samples will typically be unable to enter higher order processing until the previous information has been completely processed. *Thus, segmentation makes the learning problem more tractable by a form of temporal pattern separation.*

Temporal Sequencing: When two visual events occur sequentially at the same spatial location and very closely in time, people tend to merge them into a single episodic representation that ignores temporal order. However at a temporal separation of 200ms or more, the visual system will accurately encode their temporal sequence. This sensitivity to a temporal boundary for accurate sequence memory presumably reflects a compromise between information compression, and the need to represent causality of events at behaviorally relevant time scales. *Thus, segmentation accelerates the learning of causality by enhancing sequence information.*

Development of more advanced video-learning architectures will, at some point, be able to test these predictions by simulating human learning and comparing cases with or without brain-inspired segmentation algorithms to empirically measure their effectiveness. Such work will forge a new link across the areas of cognitive, developmental and computational areas of psychology. Moreover it is predicted that analogs of these segmentation algorithms will ultimately be necessary for developing artificial intelligence systems that can learn through environmental immersion. Regardless of computational capability (within reasonable boundaries), the processing of real-time information will serve as a crucial limit in cognitive architectures. For any amount of computation, it will almost always be more efficient to focus processing on information from specific moments in time, than to distribute that processing uniformly. The brain provides a roadmap for thinking about how to make such decisions in real-time.

Acknowledgments

This work was supported by NSF Grant 173422

References

Adamo, S. H., Cain, M. S., & Mitroff, S. R. (2013). Self-induced attentional blink: A cause of errors in multiple-target search. *Psychological science*, 24(12), 2569-2574.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human perception and performance*, *21*(1), 109.

Ouimet, C., & Jolicœur, P. (2007). Beyond Task 1 difficulty: The duration of T1 encoding modulates the attentional blink. *Visual Cognition*, *15*(3), 290-304.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, *118*(10), 2128-2148.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, *18*(3), 849.

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The attentional blink provides episodic distinctiveness: sparing at a cost. *Journal of experimental psychology: Human perception and performance*, *35*(3), 787