A Connectionist Model for Visual Search via Evidence Accumulation

Kevin Leung

Stanford University

Abstract

Most models for visual search include a serial component to account for the linear reaction times in conjunction search. An alternative model for visual search is as evidence accumulation, similar to leaky competing accumulators. This paper presents such a model with time as another competing unit. Although the model shows some promise, the data shows potentially exponential growth in reaction times when time competes with the other units. Even so, the model does exhibit increasing reaction times out of entirely parallel computation.

A Connectionist Model for Visual Search via Evidence Accumulation

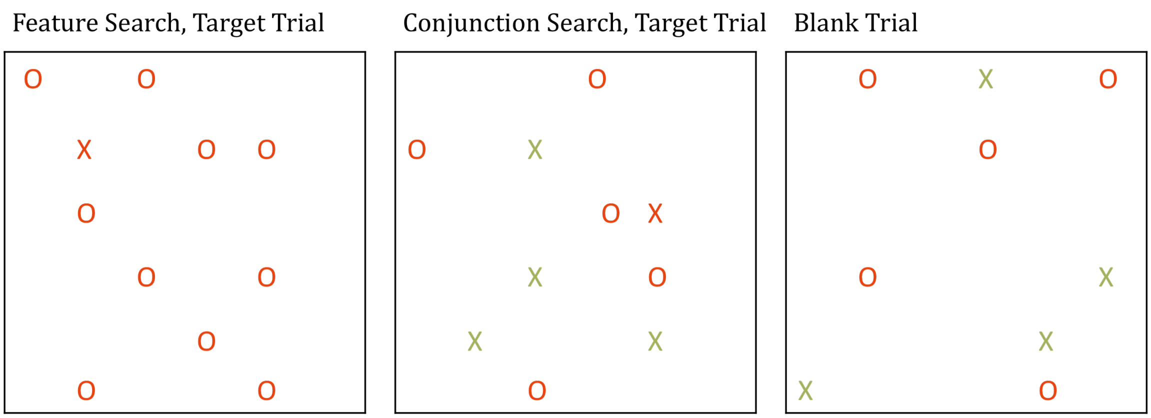
Having spent some time at university, I have met absent-minded professors and realized that they're just as bad as the stereotype. Whether it's forgetting to bring handouts to class or not being able to find their glasses perched on their forehead, they demonstrate an alarming inability to remember simple tasks and find common objects needed for everyday life. Although the memory problem belongs in a separate paper, the latter task of finding objects is an example of a common visual search task: given some area, find the stimulus that matches an intended target.

When we're looking for an item in a room, we often take a very methodical approach to finding it: start with the desk, then move onto the drawers of the dresser. This sort of serial search makes sense, and this analogy has been extended into models for visual search tasks in a single visual field. Although this serial search matches our intuition, it doesn't quite match the subjective experience of the target appearing in less than half a second without even enough time to consciously move our gaze. This paper presents an alternate model for how we perform visual search tasks without any serial mechanism.

Background

First, let's review the task that is most commonly performed in visual search and is being modeled here. Subjects see a visual field where multiple items are presented, and they must indicate whether a given target item is present or not. Although they are not required to fixate on any location, the entire visual field is visible simultaneously, and trials are often fast enough so that voluntary eye movement isn't necessary. Subjects are told to be as accurate as possible but are measured primarily on reaction time to a button press for whether the target is present. In blank trials, the visual field is filled with only distractors without a target.

The items are often colored letters or other simple shapes. For example, subjects may need to determine if a red X is present. In feature search, only one feature differs between the target and distractors. For example, the red X may only be among red Os. In conjunction search, distractors can differ in multiple features so the red X may be among red Os and green Xs.


Feature Search, Target Trial

Conjunction Search, Target Trial

Blank Trial

The most common phenomena found in visual search began with Anne Treisman's Feature Integration Theory (FIT) (Treisman & Gelade 1980). They discovered two important findings. First, in feature search, people have constant reaction times regardless of the number of distractors, a result typically attributed to a pop-out factor for the target. In conjunction search, reaction times are linear with respect to the number of distractors. Second, blank trials tend to exhibit a 2-to-1 slope ratio to target trials. This finding has been challenged by other experiments where the difficulty of the task affects relative slopes, though blank trials remain slower than target trials. FIT was the initial model presented with 2 stages of search. In pre-attentive search, all objects are mapped on one feature in parallel, and if a match is found, identification happens immediately. If that fails to identify a unique target, however, attentive search begins with serial attention to each target individually, rejecting items until the target is found.

Jeremy Wolfe later presented Guided Search (GS), which has come in several iterations (Wolfe, Cave, & Franzel 1989). In its most basic form, GS is similar to FIT in that it has both pre-attentive and attentive stages. The difference is that the information from pre-attentive search is used to guide search in the attentive stage. For example, in the example above, pre-attentive search generates an activation map for colors and another for shape, neither of which alone can find the target. These activation maps are combined so that in attentive search, the most likely candidates are attended to first.

These models, as well as others such as SERR and SLAM, all explain the linear reaction times with serial search, which seems like a natural explanation: we take more time when there are more distractors because we need to attend to each of them in turn (Humphreys & Müller 1993; Phaf, Van Der Heijden, Hudson 1990). This commonality between models is the first motivation for this particular model. Another possible explanation is that search is still happening in parallel, yet we reliably see an increase in reaction times from the additional load of each distractor. Thus, this model seeks to integrate all information in parallel. Previous work has shown that probabilistic parallel models can potentially better model visual search tasks (Dosher, Han, & Lu 2004).

Second, errors are often more difficult to account for. Given an input, one can generate an algorithm to find the target, but humans make mistakes in these tasks, and errors might be even harder to model than correct responses (Wolfe 2007, p. 112). Typically, the misses far outweigh the false alarms, though both occur. In many models, each item, when attended, is explicitly judged for being the target. Although this behavior makes sense at a high level, being able to reliably put together all evidence on the spot seems less plausible. Moreover, it also doesn’t allow for the chance to make errors unless the model randomly makes mistakes with some small chance, which begs the question about errors (Wolfe 1994, p. 210). That, however, is not particularly satisfying, so another goal of the model is to provide a reason for why errors occur.

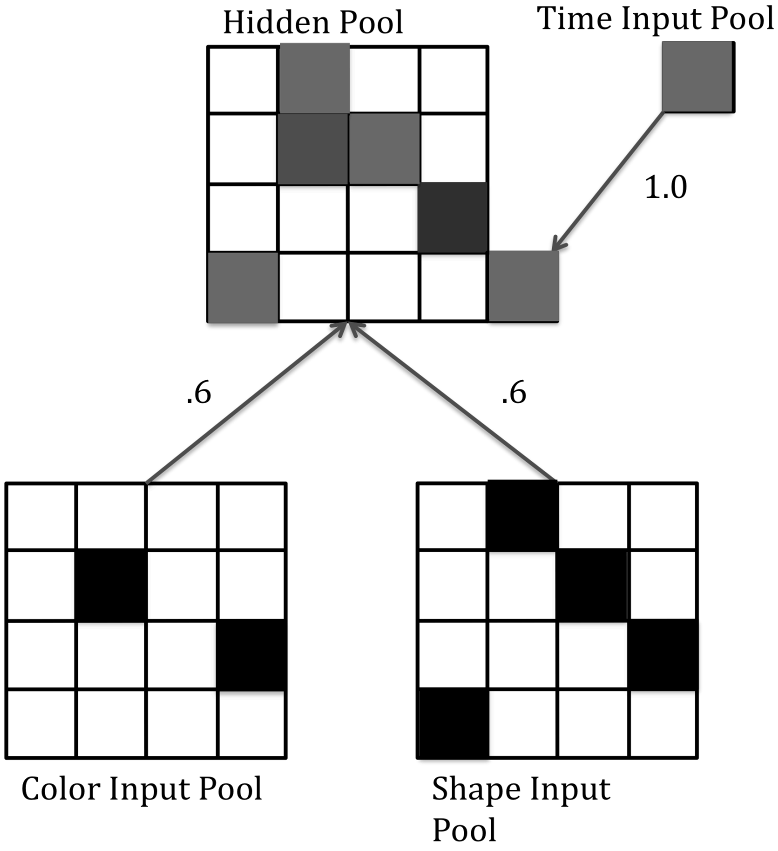
Finally, time is sometimes treated strangely in these models. Often, time becomes an outside mechanism that is forced on the model to make it fit. For example, some models terminate after all objects are inspected, which presumes a perfect memory. The third goal of the model is to integrate time as just another part of the model.

Given these constraints, the model has taken a form inspired by leaky competing accumulators (LCAs) (Usher & McClelland 2001). LCAs model decision-making by slowly accumulating information with a few important qualities in how this occurs. First, the accumulation is noisy, like other stochastic models that match intrinsic variability in firing and human behavior. Second, there’s information leakage as we don’t maintain perfect memory for all past actions and relevant data. Third, different accumulators will compete with each other through inhibition in a race to reach some threshold first. Fortunately, LCAs have a natural interpretation in visual search tasks. Each item in the visual field is one alternative, and as one looks at the image, he or she slowly accumulate information about whether each item is the target or not. Since the target has the most desirable qualities, it will hopefully accumulate evidence the fastest and be the first to reach some threshold of acceptance. This process accumulates evidence over all items simultaneously, and, because of noise, may also make a mistake, accounting for two of the three goals for the model.

The last goal was to integrate time. In unpublished work, Dufau, Grainger, and Ziegler modeled the lexical decision task using 2 LCAs. One was the accumulation of positive evidence, such as lexical access and “wordiness”, for acceptance, and the other was for rejection. Evidence for rejection was presented simply as time: as time passes, one becomes skeptical. With a similar scheme, rejection can also be another accumulator in the model with time inputted as evidence against the presence of a target.

Methods

Given the background and principles above, the architecture of the model borrows mechanisms from different models to become what it is. In addition to the goals mentioned above, there were a few principles in design. First, time and visual evidence should interact directly. Second, reaction times should naturally follow the reaction times without mechanism specifically to make it happen. Finally, both misses and false alarms should occur naturally.



The model was built using the Interactive Activation Model, similar to the Jets-Sharks model by McClelland and Rumelhart (1985). The visual field is represented by a 4x4 grid of possible locations for items. As visual input, there are 2 pools, one for shape and one for color. Both abstractly represent whether a given item has a feature in common with the target and receives an input of 1 if it corresponds and 0 otherwise. These two pools, together, map the visual field to search. For example, in the figure above, there are 5 items with the target to the far right.

Both of these are inputs to the activation map, combining the inputs into a single pool. 16 of the 17 units in this pool represent corresponding locations, so empty locations have no input, distractors have partial input, and the target has the highest input. The last unit represents time, which receives input from a single unit pool. That pool receives constantly increasing input over each time cycle.

The main activity of the models occurs in the activation map, which has all of the qualities described above. Each unit roughly represents an evidence accumulator, with its activation slowly rising over time. There’s a small amount of noise injected directly into the activation of each units independently of the other units. The activations also decay over time representing the information loss, though the decay is integrated into general maintenance with recurrent connections. An important difference between these units and LCAs is that these units have a maximum activation, introducing an additional complexity in units perhaps never reaching a given threshold as activation levels off at some value.

Mutual inhibition exists between all units in the combined activation map, including between the time unit and visual units. The interpretation for inhibition from the visual units to time is that with increasing evidence from the visual field, one should take more time to ensure that they look at all items in view, and inhibition from time to the visual units represents skepticism.

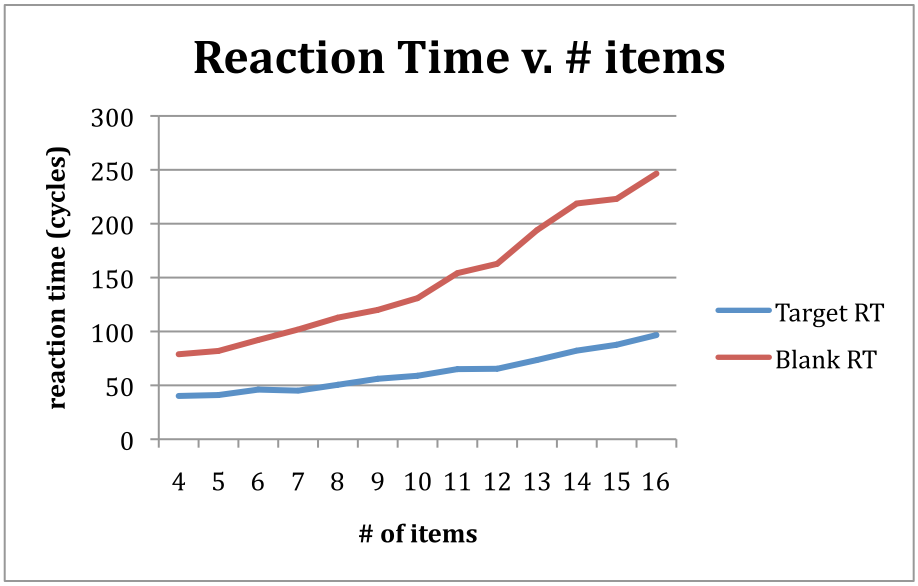
A trial terminates when the activation of a unit goes over a determined threshold. On target trials, if the target unit wins, it is a hit; otherwise, it is a miss. Importantly, activation of the wrong visual unit is a miss. Although this choice doesn’t reflect the actual method for real experiments that only require a yes or no response, false alarms in the wrong place in target trials are rare enough to not significantly affect results. In previous unpublished work with an even more stochastic model, these errors occurred in no more than 5% of the trials. From a pragmatic perspective, this distinction also ensures that the model isn't simply aggregating visual evidence from all locations to "find" the target. On blank trials, if the time unit in the activation map wins, it is marked as a correct rejection; otherwise, it is marked as a false alarm.

As a method of learning and the sort of feedback a subject would receive, the threshold is adaptive. The threshold increases slightly on false alarms to be more certain before responding. The threshold decreases slightly on misses because the model ahs become too strict and should accept with less evidence.

The experiments were run between 4 and 16 items in the visual field with 1000 randomly mixed target and blank trials for each number of items. Reaction time was counted in cycles, and accuracy was also calculated.

Results

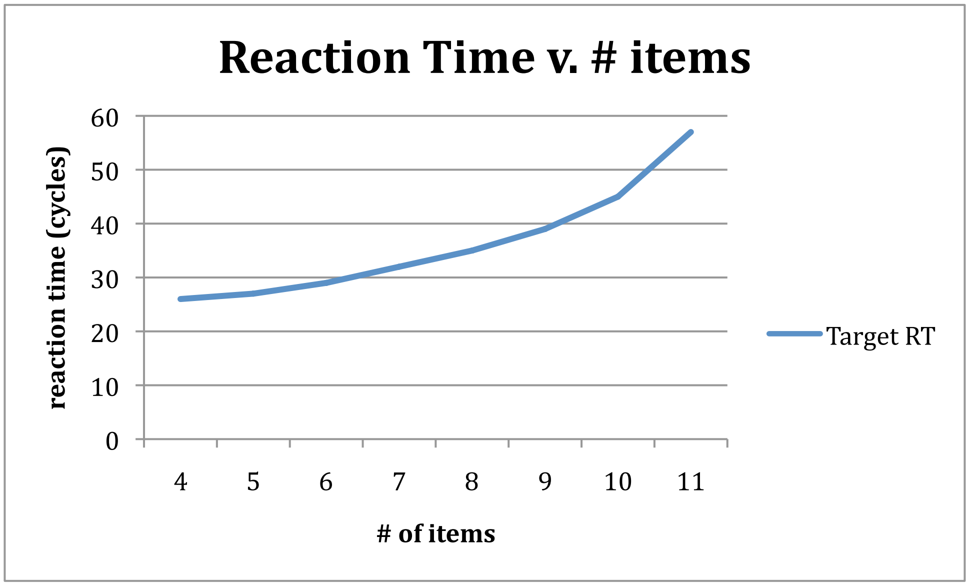
The results of the average reaction times are below:



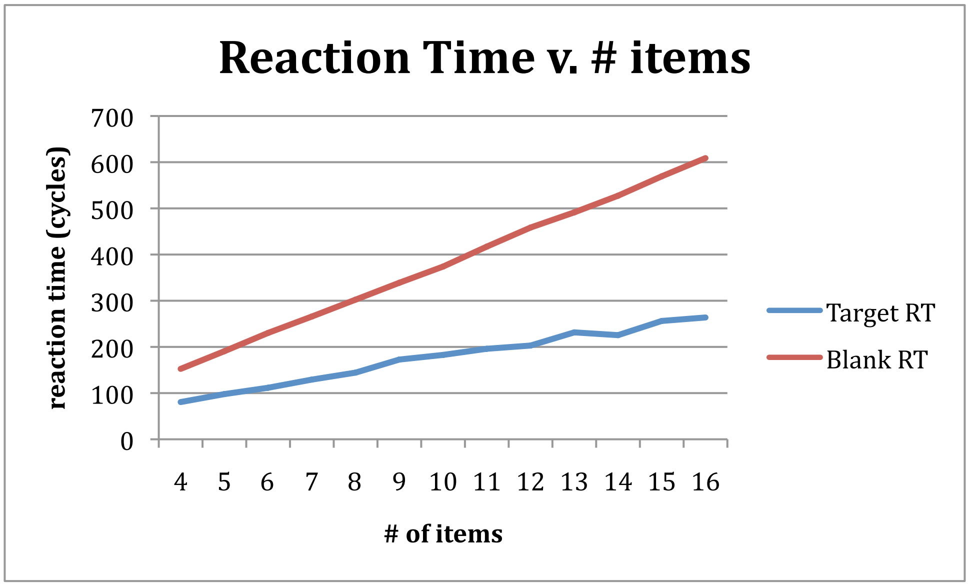
With least-squares linear regression, the reaction times for target trials are y=(4.62)*x+(15.97), r = .96 and for blank trials, y=(14.44)*x+(3.08), r = .96.

For these trials, the threshold was fixed instead of using the adaptive threshold. Although the adaptive threshold was useful in finding parameters, I found that with time included, it tended to introduce too much variation to create reasonable curves, and this particular parameter setting was highly sensitive to threshold changes. The model had a miss rate of approximately 17% and a false alarm rate of approximately 8% for a total accuracy of 84.7%. These error rates were consistent across all numbers of distractors. Although these error rates seem high, they aren't unreasonable for difficult tasks where error rates can go above 20% (Treisman & Gelade 1980, p. 108).

Although the data appears close to linear, the reaction times also appear to grow somewhat exponentially. To test this hypothesis, I removed time as a factor and eliminated all noise. To account for blank trials without time, I added a strict upper bound on the number of cycles equal to 40 * # of items. These experiments resulted in the curve below:



Further reaction times weren't possible because the rest of the target trials were unable to go over the required threshold as a limitation of the maximum activation in IAC models. I'm uncertain whether the exponential growth in the reaction times is strictly a property of having more LCAs or if this only occurs from particular parameter settings. Interestingly, with the addition of noise and the adaptive threshold, we get the following result:



With linear least squares regression, y=(15.35)*x+(23.05), r=.99. The adaptive threshold, which was problematic in the other model, is essential to this model. Moreover, the threshold was uncorrelated to the number of distractors. The accuracy was 88.6%, with target trials at 88.8% and blank trials at 88.5%. Because the adaptive threshold adjusts for every miss and false alarm, it also has the unintended consequence of often balancing the number of mistakes of each type. It's notable that this doesn't reflect actual data where misses far outweigh false alarms.

Given the success of this particular parameter setting, I attempted to extend the model with time from it while making as few changes as possible. I changed the parameters for time so that the threshold matched and minimized errors, though the error rates still drove the threshold down with more distractors. Another change I made was to disable inhibition from the time unit to the visual units to try to leave the visual-only system intact, but this change was also ineffective. Finally, I tried using a different adaptive setting so the threshold was only changed when error rates when above certain levels. This change was also ineffective, and I am still unsure how the adaptive threshold affects the reaction times in this model. Even so, I did make several other discoveries.

First, noise was critical to the functionality of the model. Without noise, the model is deterministic and doesn't yield particularly interesting results. Even more interesting was how small changes in the amount of noise greatly affected how the model performed. It's possible that noise covered up some of the problems with parameter settings. Although I worked to align time and visual input as best as possible, the model is sensitive to how quickly evidence accumulates for one relative to the other.

The noise also helps to keep the threshold relatively stable. In the model without time, the threshold was relatively stable, though in testing without noise, the threshold decreased as the number of items increased. With fewer items, the model can be more tolerant of a higher threshold. Because of the maximum activation and only partial accumulation, all activation stabilize at some level, and with more distractors, the maximum activation of the target is lower because of inhibition. Without noise, the target will always reach some threshold faster than any of the distractors, and as long as that threshold is lower than its equilibrium value, it will succeed. Only when the threshold is above its equilibrium value will the model miss the target and lower the threshold. Noise, however, appears to allow the model to maintain a relatively high threshold relative to the target's equilibrium value.

Discussion

Although the results currently don't exactly match empirical results, the large space of potential models keeps me optimistic that some parameter setting can match most of the most significant trends in the data. With more than 10 free parameters that have largely been explored only through manual testing and complex interactions between various settings, there are many models still to be tested.

Earlier, three factors from LCAs were emphasized, and all turned out to be critical in making the model work. Noise was addressed above in the results section, and accumulating evidence with leakage produces the reaction times. Inhibition was important as well in ensuring that the target remained more prominent in spite of growing numbers of distractors. In previous work, errors rates increased dramatically with the number of items because the difference between the target and distractors was shrinking as distractors increased. With inhibition, the target keeps the activation of other units down. The inhibition only slows down the rate at which the target is overwhelmed by distractors, and asymptotically, the target will again be drowned out. On the scale of only 16 items, however, inhibition works well enough. Additionally, studies have found a small but significant correlation between errors and the number of items, so this behavior is allowable (Wolfe, 1989, p. 421).

Since this LCA approach originated in other tasks, we can also apply some of the same ideas in re-evaluating how visual search tasks are viewed. Like the lexical decision task or mental rotation tasks, subjects are supposed to react as quickly as possible while maintaining accuracy. Similar questions come up, such as when knows when to stop the trial or how to react to errors. Hopefully this model shows that this paradigm applies equally well in visual search as in other domains.

The first next steps with the model would be to continue to tweak parameters to better fit the data. Beyond the many open questions in the results section, the model has only generally been used to fit the trends in data without matching specific reaction times or error rates. Being able to map the number of cycles onto actual reaction times instead of comparing shapes and slope is necessary, and I also want to compare the variance in the data. There are, however, other interesting directions in changes to the model.

First, the input into this model is very naïve. All distractors are equivalent in the model, and there are truly only 3 different states for inputs into the activation map: target, distractor, and no item. Other models have included more complex signals accounting for nonlinearities in features, such as the angle of lines, the locality of various targets, and both endogenous and exogenous attentional effects. Fortunately, because the core of the model is the processing in the activation map, the model can accept any sort of input to be fed as evidence for the activation map. Hopefully, more complex inputs could also produce more interesting effects in the data. Currently, the model depends on large noise to account for variation in the data, and more complex stimuli may help to create the same effects.

Second, I tried two different algorithms for adaptive thresholds, both fairly simple and neither based on actual data. Given how significant the adaptive threshold was in my results, it seems worth investigating both in the model and empirically. Actual subjects do respond to the errors that they make and will often slow down immediately after one (the effects are significant enough that the trials immediately after errors are often discarded). The natural interpretation of this is in evidence accumulation is that the subject requires more evidence to be certain about their decision.

On a similar note, one would predict that subjects speed up when they get several consecutive trials correct. In simpler terms, when subjects get cocky, they will tend to rush. A mechanism for decreasing the threshold marginally would also be useful. Such a mechanism was used in GS 2.0, though I am uncertain what the actual data is for reaction times over the course of a block of trials (Wolfe, 1994, 210). I predict that the variation between trials would significantly outweigh small changes because of cockiness. Combined with the large variation between subjects and the interruptions of errors, concrete data on this subject might be difficult to come by.

An important prediction of this model is that the evidence accumulation process leads to a fundamentally non-linear increase in reaction times. Although noise can mask these effects to a point, over enough trials and extending the number of distractors far enough, the reaction times should begin to increase faster and faster. This sort of response is also difficult to test. Visual search tasks largely reflect covert attention because the entire visual area can be seen simultaneously. With too many items in the visual field at once, however, reaction times should rise enough that overt eye movements are possible, which likely has a separate mechanism for search. At that point, cognitive control becomes a factor as attention is voluntarily directed to various parts of the visual field to analyze smaller clusters at once. This sort of predicted behavior may in fact be completely unused by the human visual system because of this design, though such an experiment could be interesting.

Overall, this model takes a fairly high-level approach to visual search. Since visual search became an important task in measuring visual attention, many more specific mechanisms have been identified, such as inhibition of return, the effect of saccades and microsaccades, feature saliency, and search asymmetry, to name a few. This approach also means that the model doesn't explain or even begin to integrate these effects.

This model seeks more to show what isn't necessary in modeling visual search tasks. As discussed, models for visual search tend to include serial attentive search to account for reaction times and specific changes to create errors. Even limited diffusion models, such as GS 4.0, are not strictly necessary to match the data (Wolfe, 2007). Although subjective experience can be hard to discriminate on such short time scales, a massively parallel search like this model matches the sense of being able to suddenly see the target. Also, this model unifies pre-attentive and attentive search into a single stage for processin and presents those as emergent properties of the model from its design.

References

Dosher, B. A., Han, S., & Lu, Z.-L. (2004). Parallel Processing in Visual Search Asymmetry. *Journal of Experimental Psychology: Human Perception and Performance*, 30(1), 3-27. doi:10.1037/0096-1523.30.1.3

Humphreys, G. W., & Müller, H. J. (1993). SEarch via Recursive Rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, 25(1), 43-110. doi: 10.1006/cogp.1993.1002

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114, 159–188.

Phaf, R. H., Van Der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A Connectionist Model for Attention in Visual Selection Tasks. *Cognitive Psychology*, 22, 273-341.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97-136. doi:10.1016/0010-0285(80)90005-5

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, 108, 550-592.

Wolfe, J.M. (1994) Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review*, 1(2): 202-238.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.

Wolfe, J.M., Cave, K. R., Franzel, S.L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 419-433.