An Embodied Real-Time Model of Language-Guided Incremental Visual Search

Matthias Scheutz and Evan Krause and Sepideh Sadeghi
{matthias.scheutz,evan.krause,sepideh.sadeghi}@tufts.edu
Human-Robot Interaction Laboratory, Department of Computer Science, Tufts University
Medford, MA 02155, USA

Abstract

A recent body of work has demonstrated that the incremental presentations of linguistic search cues can speed up visual processing in conjunctive visual search. In this paper, we investigate different processing configurations using a real-time embodied computational model and demonstrate that, different from previous hypotheses, the same incremental processing configuration can explain all experimental conditions.

Keywords: Incremental interactive processing, embodied real-time model, natural language and vision interaction

Introduction

A large body of work in cognitive science has demonstrated that much of human information processing is *interactive* and *incremental*: "interactive" means that information is exchanged and shared among multiple processes; "incremental" means that the received information is integrated as it becomes available. Hence, interactive incremental processing modules can incorporate information from other modules as constraints in their own processing and thus potentially finish their processing sooner.

A well-studied case of such interactive incremental processing is the interaction between visual and natural language processes. Converging evidence from studies using, in particular, the "visual words paradigm" demonstrate that vision and natural language processing in humans are highly interactive and incremental, being able to utilize constraints from the other modality to reduce processing effort and improve processing performance (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). For example, a visual search process attempting to find a target object in a visual scene such as a particular pen on a cluttered desk can be modulated through natural language instructions that provide additional information about the object (e.g., "small black"), leading to a more targeted, faster search (Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Krause, Cantrell, Potapova, Zillich, & Scheutz, 2013). Conversely, visual processing of a scene can influence natural language processing by helping to disambiguate otherwise ambiguous referential phrases such as the syntactic ambiguity due to different possible prepositional attachments in "put the black pen on the book on the table" where the black pen could be put either on the table or on the book that is on the table (Eberhard et al., 1995; Scheutz, Eberhard, & Andronache, 2004; Brick & Scheutz, 2007).

While various theoretically motivated hypotheses have been proposed about an underlying processing architecture that could enable such incremental natural language and vision interaction and information integration, only a few computational models actually demonstrate possible computational mechanisms (Scheutz et al., 2004; Hamker, 2004; Brick & Scheutz, 2007; Chiu & Spivey, 2012; Krause et al., 2013). However, computational models are often necessary to show that conclusions drawn about the processing architecture based on experimental evidence or theory alone might not be warranted.

In this paper, we present an embodied real-time model of interactive incremental vision and natural language processing that can explain previous experimental findings in a novel way by showing that divergent results found in different experimental conditions by Spivey et al. (2001) might not be due to differences in processing configurations (such as serial vs. parallel), but rather the specific effects of these experimental manipulations on the *same processing configuration*.

We start by reviewing some of the empirical evidence for the hypothesis that natural language can incrementally constrain vision processing and describe, in particular, the experiments in (Spivey et al., 2001) which we use for our model simulations. Next, we introduce the model architecture and provide a more detailed description of its vision system which is critical for the replication of the human data. We then specify the simulation setup, which used the same human stimuli as Spivey et al. (2001), and report the results from extensive simulation experiments with different configurations of the processing system. The analysis of the simulation data confirms many of the expected properties, but also shows that the same configuration can explain different experimental conditions that have been assumed to be the result of different processing configurations. This point is further elaborated in the subsequent discussion section and summarized in the conclusion which also points to future work.

Background and Motivation

It has long been hypothesized that early stages of bottom-up visual processing are highly parallel as single-feature visual search is not affected by the number of co-present distractors, while later stages must include a "serial bottleneck" since conjunctive visual search (assumed to tap into later processing stages) takes longer as the number of distractors increases (Wolfe, 2007). While various stimuli properties can affect search speed in conjunctive search, Spivey and colleagues demonstrated in a series of experiments that the incremental presentation of linguistic search cues can reduce the effect of distractors in the visual search process (Spivey et al., 2001; Reali, Spivey, Tyler, & Terranova, 2006; Chiu & Spivey, 2011, 2012). They hypothesized that the incremental presentation of search cues (which is natural in spoken language) enforced a serialization of the search process, allowing search results based on the first cue to be utilized in

the second search, thus shortening its duration. In contrast, no such incremental processing was hypothesized in experimental conditions where both search dimensions were simultaneously presented.

Spivey et al. (2001) used a standard conjunctive visual search paradigm where the presence or absence of a colored bar (red or green) in a particular orientation (horizontal or vertical) has to be detected. For the task presentation, they considered two linguistic conditions: in the "audio first" (A1st) condition the auditory target cue precedes the onset of the visual stimulus, while in the the concurrent "audio-vision" condition (A/V) the auditory cue and visual scene have the same onset. Consistent with the visual search literature, the results showed that the slopes of the best fitting lines relating response times (RTs) to stimulus set size for both target absent and target present cases were positive, with larger slopes for target absent compared to target present cases. Critically, the slopes in the A/V concurrent conditions were smaller than the slopes in the A1st conditions. The effect persisted when the cue order was altered and when stimuli were presented visually in the A/V condition (instead of through a preceding natural language instruction).

Exploring alternative explanations, Reali et al. (2006) replicated the findings from Spivey et al. (2001) with mixed A/V and A1st trials in random order. They also mixed color first conjunction searches with orientation first conjunction searches to show that the search improvement in the A/V condition is not due to subjects' listening strategies, nor does it disappear as the complexity of the utterances increases. Moreover, slopes are still smaller for A/V concurrent conditions than in the A1st conditions in triple visual searches.

Finally, Chiu and Spivey (2011) argued that the visual search utilizes a combination of serial and parallel strategies as opposed to purely parallel or serial strategies. By manipulating the stimulus onset asynchrony (SOA) between the end of the first and the onset of the second cue (for 0, 200, 400, and 600ms SOAs), they found significantly shallower slopes for the A/V compared to the A1st conditions when the SOA was equal to 400 or 600 ms. However, when the SOA was 0 or 200 ms, they did not observe significantly shallower slopes for the A/V compared to A1st conditions, indicating that the search improvement in the A/V condition is dependent on the SOA which acts as a "buffer" for the end of first cue search process.

The overarching question posed by this whole line of research then is why incremental processing should only occur in the A1st condition and not also in the A/V condition. To answer this question, we investigated different processing configurations of a computational model, described next, that can perform the task from Spivey et al. (2001) to find the configurations that most closely matched the human data in both A1st and A/V conditions.[1]

[1]We will restrict our modeling efforts here on Spivey et al. (2001) for lack of space, but note that the model generalizes to different variations of the experiments.

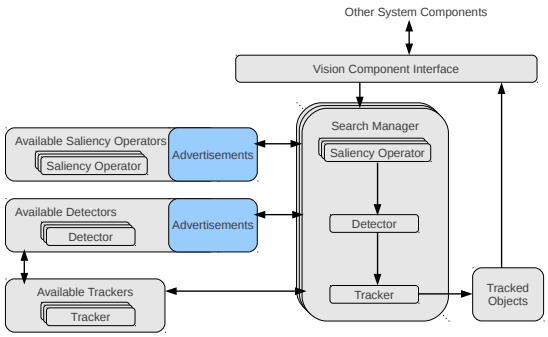

Figure 1: The processing architecture of the vision system.

The Embodied Real-Time Model

Over the last decade we have developed a complex integrated embodied cognitive architecture called "DIARC" (Scheutz, Schermerhorn, Kramer, & Anderson, 2007) which has been shown to exhibit several qualitative features of human-like natural language and vision processing (e.g., incremental reference resolution, Scheutz et al., 2004; incremental information integration, Brick & Scheutz, 2007; dialogue-based constraints on speech recognition, Veale, Briggs, & Scheutz, 2013; visual search constrained by linguistic expressions, Krause et al., 2013; and others). Here we will use DIARC for the first time to investigate potential *quantitative models of human performance* that differ only in their processing configuration in order to evaluate hypotheses about processing modes in linguistic-guided visual search. We depart from the requirement that our computational model be able to use the same real-world linguistic and visual input as in Spivey et al. (2001) for two reasons: (1) to be able to model human performance of real-time incremental interactive information processing and (2) to avoid complications about subtle timing effects that can arise with discrete-event simulation models that only have simulated paralellism.[2] Since the focus of the model is on configurations of visual search, we skip the overview of the natural language subsystem which has been described in detail elsewhere (Cantrell et al., 2010; Krause et al., 2013) and focus on the visual subsystem (only describing those parts of the natural language subsystem necessary for understanding the model configurations and runs).

The vision system is consistent with empirically grounded views on guided visual search in humans (e.g., Wolfe, 2007) and consists of three main components: *saliency operators* that compute different types of saliency maps, *object detectors* that can use various features (including information from saliency maps, textures, shapes, etc.) to detect objects, and *object trackers* that can track previously detected objects over

[2]An additional reason, not directly relevant to the effort in this paper, is our aim to run those models on robots in the context of human-robot interaction scenarios where robots have to respect human timing and modes of information processing in natural language dialogues (Scheutz et al., 2007; Cantrell, Scheutz, Schermerhorn, & Wu, 2010).

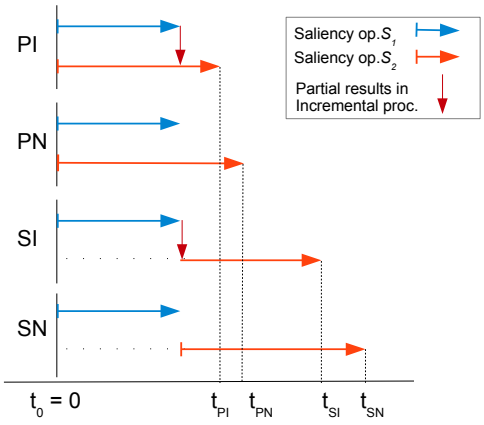

Figure 2: Four saliency operator configurations for conjunctive visual search and their relative times to completion.

time (a function not required for the current paper).[3]

Saliency operators. Saliency operators are massively parallel, computationally cheap bottom-up processes that operate directly on regions of the input image, thus constituting the first visual processing stage. They are used to extract basic visual features such as color, size, orientation, and motion (analogous to the system in Itti & Koch, 2001). The result of applying a saliency operator to a region in the input image is a (partial) saliency map with values between 0 (no salience) and 1 (maximal salience) for each pixel. Multiple saliency operators can be configured to perform computations either in *parallel* (P) or *serially* (S) (i.e., one map at a time, or multiple maps simultaneously even though they might take different times to compute). Different from other bottom-up saliency models where computations for individual maps are modular, it is possible in both cases for saliency operators to interact by using values from existing saliency maps (generated by other operators) to modulate the saliency computation in their own map – this *incremental* (I) mode of operation is contrasted with the *non-incremental* (N) mode of only performing calculations based on the input image. Hence, saliency operators can interact in four ways based on their configuration in ascending order of processing efficiency: PI, PN, SI, SN (see Figure 2).

Object detection. Objects are detected by segmenting regions in the input image based on their saliency as determined in (possibly combined) saliency maps. Different from saliency computations, this is a serial process where regions with the highest saliency are considered first. Segmentation is performed using a simple background model to distinguish background pixels from object pixels (i.e., modeling background pixels as a particular RGB value), and Euclidean clustering to grow an "object pixel cluster" (from a single pixel) corresponding to the most salient image region. Once a candidate object has been segmented, it is checked against the individual saliency maps in order to confirm that it has all necessary salient features. Thus, the most salient objects are detected first and are immediately available as "target objects" in the search, while less salient objects follow later. In target detection search it is thus possible to terminate the visual search early (compared to a search requiring a count of all target objects, say).

Simulation Results

The goal of the model simulations was to find the model configuration that most closely matched the data from Experiment 1 in Spivey et al. (2001), i.e., the differences in response times over stimuli sets with an increasing number of items in the target absence vs. target presence conditions in both the audio first vs. audio-vision conditions. We thus defined four different processing configurations based on two processing dimensions: *incremental* (I) vs. *non-incremental* (N), and *serial* (S) vs. *parallel* (P). We used the four sets of 32 image stimuli from (Spivey et al., 2001) which contain 5, 10, 15, and 20 vertical/horizontal and red/green bars, respectively (see Figure 3).

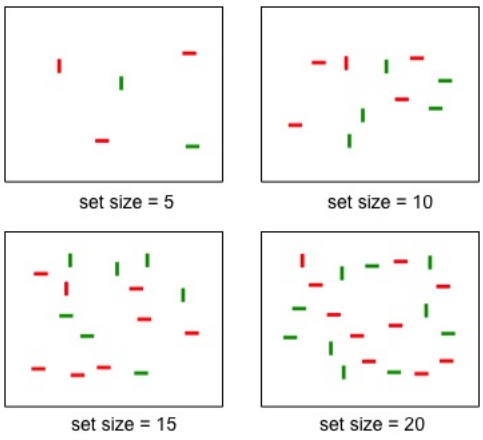

Figure 3: Examples of stimulus images for each of the four item set sizes.

We ran 100 replications of each visual stimulus with color terms followed by orientation terms for the four combinations of color ("red" vs. "green") and orientation ("vertical" vs. "horizontal") for a total of $100 \cdot 32 \cdot 4 = 12800$ runs for each of the four model configurations (i.e., over 50000 runs total) in the "audio first" condition. [4] For each run, we measured

[3]Note that the vision system does not realize biologically plausible computations "all the way down to individual neurons", but allows for different sequencing and interactions of processing modules, which is necessary for investigating human processing configurations.

[4]For the large evaluation, we did not run the complete architecture but only the vision subsystem because we were only interested in the response time from the onset of the visual stimulus in this condition. Otherwise, the model can perform the whole experiment in the same setup as human would (with real-time audio and video). Note that replications are important because computational

the processing duration for each feature (color and orientation) as well as the time required for information integration and decision-making in the object detector (the vision system was especially instrumented with time measurement code for that purpose). Since we were not interested in examining performance errors (such as false starts and wrong outputs), we set parameters in the vision system in a way that the model had perfect performance in all runs.

Instead of running separate simulations for the audio-vision conditions, we were able to reuse the data from the audio-first condition. Recall that in the audio-first condition, both visual search features have been already determined before the onset of the image and thus the visual search can either be carried out in parallel or serially. In contrast, in the audio-vision condition the visual features are given sequentially while the target image is already present. Since the vision system always completes the color processing of any stimuli in less time than it takes to pronounce the corresponding color terms, processing in the audio-vision condition is always be serial regardless of the model configuration (serial or parallel). Therefore, instead of running separate simulations for the "audio-vision" conditions, we were able to reuse the data from the audio-first condition by taking the duration from the onset of the image (which is also that of the first linguistic cue) and the onset of the second linguistic cue, and then adding the model's response time measured from the onset of the second cue until the decision (target or no-target) is reached.

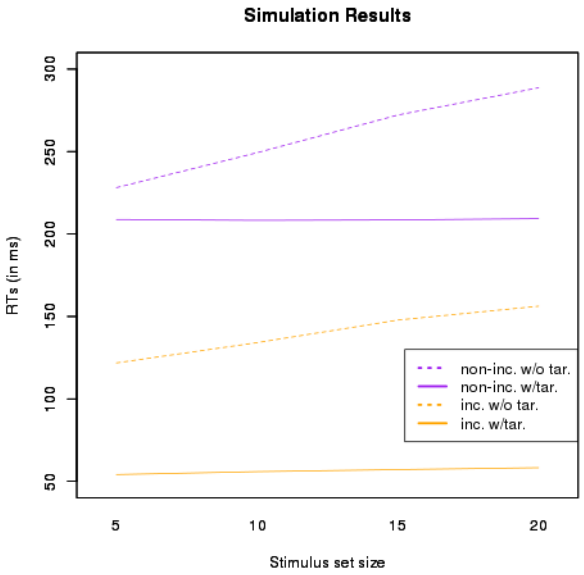

Figure 4: The three-way interaction between data size, incremental processing and target presence/absence conditions.

processes, even though specified by deterministic programs, have stochastic run-times given the many concurrently running processes in the Ubuntu Linux operating system on the employed quadcore PC with Intel i7-3820 CPU at 3.60GHz.

We performed a 2x2x2x4 ANOVA with *target condition* (present vs. absent), *integration mode* (incremental vs. non-incremental), *processing mode* (parallel vs. serial), and *item set size* (5, 10, 15, or 20 items) as independent, and *total time* (to processing completion from visual stimulus onset) as dependent variables. We found highly significant main effects (all $F(1,12784) > 100, p < .001$) on all four independent variables as expected: the absence of the target, non-incremental processing, serial processing, and increase in item set size all lead to longer RTs. In addition, we found significant two-way interactions (all $F(1,12784) > 100, p < .001$ except for the first with $F(1,12784) > 5, p = .024$): between processing mode and target presence/absence indicating that the difference in RTs between target absence and presence are increasing when processing is serial; between item set size and incremental processing indicating that as item size increases the advantage of incremental processing increases too; between item set size and target presence/absence indicating that item set size increases in the target absence condition increase the RTs massively while RTs in the target presence conditions show only a moderate increase; between incremental processing and target presence/absence indicating that the advantage of incremental processing in the target presence condition is greater than in the target absence condition. The last three two-way interactions are explained by a significant three-way interaction ($F(1,12784) > 100, p < .001$) between item set size, incremental processing and target presence/absence which corroborates the human data: increases in item set size lead to larger increases in RTs in the non-incremental compared to the incremental processing configuration thus closing the initially wider gap between the target presence vs. target absence conditions relative to the overall differences between incremental and non-incremental processing (see Figure 4).

Figure 5 then shows the best fitting lines for the four model configurations and Table 1 shows the intercepts and slopes for the models compared to the linear fits from the human data in the A1st vs. A/V conditions in (Spivey et al., 2001). The different intercepts and slopes are indicative of both differences in processing style but also differences in the duration of individual subcomponent processes (e.g., the time it takes to compute a saliency map). Overall, the fit lines confirm the results from the previous analysis (in part shown in Figure 4) that parallel processing is faster than serial, that incremental processing is much faster than non-incremental processing, and that the target-presence conditions scale much better over item set size compared to the target-absence conditions, all of which is in line with the human data.

To be able to directly compare these linear fits while taking into account the differences in processing times in individual human and model subsystems (as we were not attempting to model the details of human vision and natural language processing, but rather the overall processing configuration), we use a relational comparison. Following Spivey et al. (2001), we consider the ratio of slopes in the human target-absent to

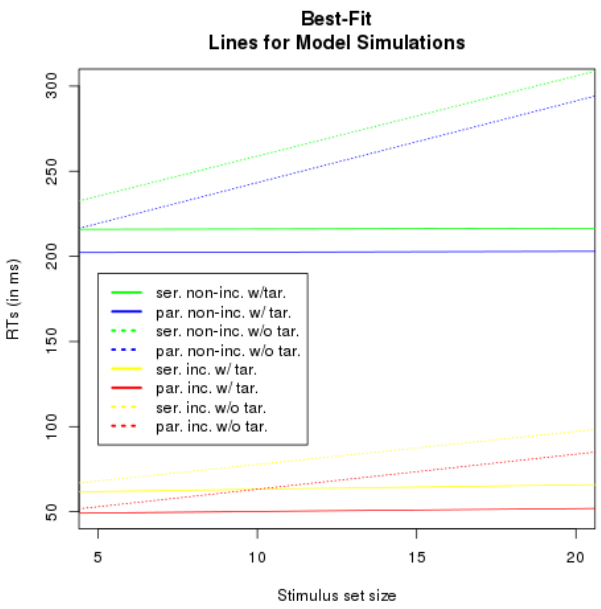

Figure 5: The best fit lines for all four models.

Cond.	target present		target absent	
	intercept	slope	intercept	slope
I-P	48.409	0.173	42.690	2.058
I-S	60.549	0.264	58.397	1.941
N-P	202.056	0.045	195.379	4.797
N-S	215.726	0.039	211.851	4.706
A1st	830.000	19.800	911.000	31.400
A/V	1539.000	7.700	1628.000	22.700

Table 1: Summary of the regression lines from four sets of model simulations and the data from Experiment 1 (A/V vs. A1st) in Spivey et al. (2001) (see text for explanation).

target-present conditions $S_{h,p}/S_{h,a}$ and now compare it to the same ratio for the model $S_{m,p}/S_{m,a}$, which – in the ideal case – should be identical.[5] Table 2 shows the calculated "model proximity values" $\frac{S_{h,p} \cdot S_{m,a}}{S_{h,a} \cdot S_{m,p}}$ (closer to 1 is better) for each of the four models in the "audio first" (A1st) and the two models in the "audio-vision" (A/V) conditions.

As can be seen from the model proximity values, the serial incremental models fare best in both instruction conditions. While it is premature to draw conclusions about the difference between sequential and parallel search given the close

[5]Note that comparing *slope ratios* between target absent and target present conditions addresses the data comparison problem raised by the fact that intercepts and slopes are different. If we had attempted an individual comparison of best fitting lines from model and human data, we would have had to scale the model intercept I_m to the human level I_h by the factor $\lambda = I_m/I_h$ and then adjust the model slope S_m accordingly: $S_m \cdot \lambda$. Note that scaling factors cancel out in a ratio comparison, hence no scaling is necessary for comparing ratios of human and model data.

Instr./ Mode	A/V		A1st	
	incr.	non-incr.	incr.	non-incr.
serial	2.492462	40.87944	4.633394	75.99336
parallel	4.036554	36.40101	N/A	N/A

Table 2: Model proximity values for the six model conditions (see text for explanation).

numeric results in both configurations, it is clear that incremental processing reflects the human data much better compared to non-incremental processing.

Discussion

The fact that both instruction conditions seem to utilize the same processing configuration is quite surprising at first glance, because it seems that at the very least in the A1st condition the parallelism of the feature search should be exploitable. And, indeed, the results in Table 1 confirm that parallel incremental search is the fastest in this condition. Spivey et al. (2001) argue that "in the auditory-first condition, the search process may employ a conjunction template to find the target, thus forcing a serial-like process akin to sequentially comparing each object with the target template. However, in the A/V-concurrent condition, it appears that the incremental nature of the speech input allows the search process to begin when only a single feature of the target identity has been heard [which then] proceeds in a more parallel fashion (with the second-mentioned target feature being used to find the target amidst an attended subset)." Based on our modeling results, we would like to propose an alternative explanation that is consistent with the experimental differences observed by Spivey et al. (2001) and does not require the stipulation of a different processing configuration.

The explanation rests on the assumption that with none of the visual stimuli it took humans longer to process the color cue than the time it took to pronounce the color word in the A/V condition, see Figure 6 for an illustration which shows the time course for processing the two visual cues in the A/V and A1st conditions for parallel and serial configurations (cue integration time is absorbed in the second cue processing for simplicity). Note that there is no chance to exploit parallelism at any point in the A/V condition precisely because the vision system will have finished processing the color of the items before the orientation cue occurs (although those processes take increasingly longer as the set size increases, hence there will be a set size where visual processing will exceed the duration of the spoken color cue). Thus, both serial and parallel processing configurations require the same overall processing time in the A/V condition, different from the A1st conditions, where parallel and serial yield different results. Moreover, if the vision system used a parallel configuration in the A1st condition, then the slopes in that condition would be the same as in the A/V. However, Spivey et al. (2001) found a steeper slope in the A1st condition suggesting that the sys-

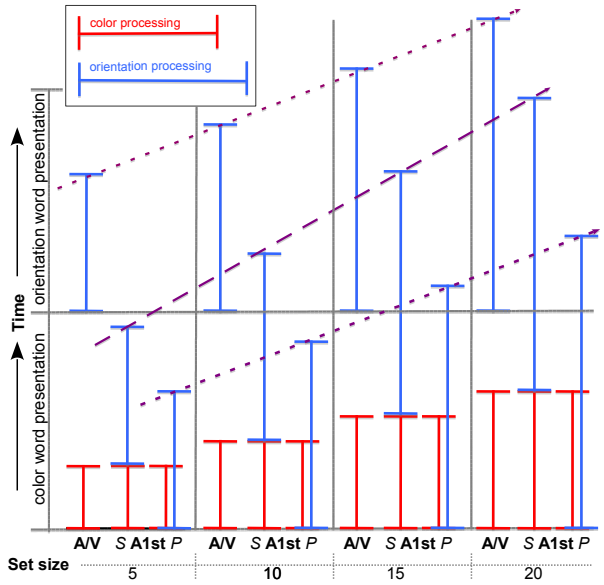

Figure 6: The influence on fixed color word duration on the overall RTs in the A/V compared to the A1st serial and parallel conditions (see text for details).

tem was configured serially. While it is possible that processing is configured in parallel in the A/V condition, this seems implausible given that the parallelism has no effect in that condition and that the system was not configured in a parallel fashion in the A1st condition where the parallelism could have been exploited. Additional evidence comes from our model simulations where serial (incremental) configurations have the best fit to the human data. Notice that the above argument does not rely on the incremental/non-incremental distinction, hence the argument holds for both incremental and non-incremental configurations.

Conclusion

We introduced a real-time embodied computational model that was used to investigate different processing configurations of a cognitive system that can perform conjunctive visual searches based on spoken natural language cues. We replicated the empirical findings from Experiment 1 in (Spivey et al., 2001) that show a significant difference between audio-visual concurrent instruction compared to audio-first instruction, which has been hypothesized to be due to a difference in processing configuration of the visual system in the two conditions. Based on our modeling results, we conclude that the same processing configuration is responsible for both conditions and that the differences in the experimental data in the two conditions are fully explained by the way the experimental manipulations impose processing constraints on the system. Future work will extend the model to the different published variations of the experiment in an attempt to show that the same underlying processing configuration can explain the results in all of these conditions.

Acknowledgments

This work was in part supported by NSF grant IIS-1111323 and ONR grants #N00014-11-1-0289 and #N00014-14-1-0149 to the first author. Thanks to Mike Spivey and his group for providing the stimuli and for helpful discussions.

References

Brick, T., & Scheutz, M. (2007). Incremental natural language processing for HRI. In *Proceedings of the second acm ieee international conference on human-robot interaction* (pp. 263–270). Washington D.C..

Cantrell, R., Scheutz, M., Schermerhorn, P., & Wu, X. (2010). Robust spoken instruction understanding for HRI. In *Proceedings of the 2010 human-robot interaction conference* (p. 275-282).

Chiu, E., & Spivey, M. (2011). Linguistic mediation of visual search: Effects of speech timing and display. In *European perspectives on cognitive science.*

Chiu, E., & Spivey, M. (2012). The role of preview and incremental delivery on visual search. In *Proceedings of the 34th annual conference of the cognitive science society.*

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409–436.

Hamker, F. (2004). A dynamic model of how feature cues guide spatial attention. In *Vision research* (Vol. 44, pp. 501–521).

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. In *Nature reviews, neuroscience* (pp. 194–203).

Krause, E., Cantrell, R., Potapova, E., Zillich, M., & Scheutz, M. (2013). Incrementally biasing visual search using natural language input. In *Proceedings of aamas* (pp. 31–38).

Reali, F., Spivey, M., Tyler, M., & Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Perception and Psychophysics*, *68*(6), 959–974.

Scheutz, M., Eberhard, K., & Andronache, V. (2004). A real-time robotic model of human reference resolution using visual constraints. *Connection Science*, *16*(3), 145–167.

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411–423.

Spivey, M., Tyler, M., Eberhard, K., & Tanenhaus, M. (2001). Linguistically mediated visual search. *Psychological Science*, *12*, 282–286.

Veale, R., Briggs, G., & Scheutz, M. (2013). Linking cognitive tokens to biological signals: Dialogue context improves neural speech recognizer performance. In *Proceedings of the 35th annual conference of the cognitive science society.*

Wolfe, J. (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 University Press.