Cognitive Processes in Contextual Cueing

Christian Balkenius (christian.balkenius@lucs.lu.se)
Lund University Cognitive Science
Kungshuset, Lundagård, SE-222 22 Lund, Sweden

Abstract

A computational model of learning in visual attention is described, and it is shown how it can integrate information over time to form a representation of the visual context. This visual context is subsequently used to guide the focus of attention in a visual search task. Computer simulations shows that the model can learn to use contextual cues when they are available. The performance using contextual cues is compared to a situation where they are not used. The result shows that contextual cueing is only useful when the visual scene is sufficiently complex to warrant the extra time needed to establish context identity.

Introduction

To understand a visual scene, it is necessary to integrate information from many attentional fixations. This type of process can be investigated by the contextual cueing paradigm (Chun and Jiang, 1998, Chun, 2000). In this type of experiment, subjects are shown different images with a number of distracters and a single target stimulus. The spatial locations of the distracters together predict the location of the target stimulus and by using this information; the reaction time in the visual search task will be reduced.

The contextual cueing paradigm is important because it may give insights on how we integrate and use information from many sources. It may also give ideas for technical solution in the area of visual scene analysis.

Context can tell us where in the image objects are (Biederman, 1982) or suggest their identity (Palmer, 1975). In this way, the visual context can be used both for faster visual search and to disambiguate the scene.

Much is known about how the context participates in different learning situations and it would be interesting to see whether it has an analogous role in the control of visual attention. Balkenius (2000) proposed a computational model that aims at characterizing the different learning processes in visual attention. The model included modules for context recognition and target prediction, but the context was used only to control habituation and extinction and not as a cue in itself. The aim of the present work is to extend that model with the ability to use the context as a cue to predict a target location.

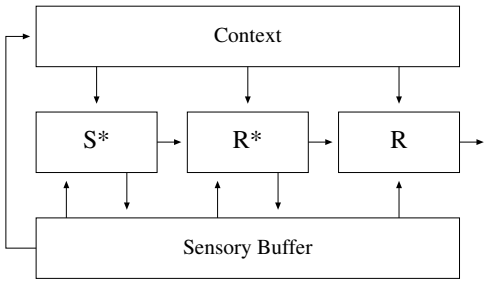

Figure 1: The original model described in Balkenius (2000). *S**: stimulus evaluation; *R**: response evaluation; *R*: fixed responses.

The Model

The new model is an extension of the computational model described in Balkenius (2000). The goal of the model is to explain how different learning processes contribute to the control of visual attention. There are five main components in the model (Fig. 1). A *sensory buffer* codes the visual input in different ways. This coding ranges from the detection of oriented contrasts to object identity. The sensory buffer also codes the visual input using a spatial code that allows attention to be directed toward the location of a stimulus. A *fixed response system* R reacts to the location coding in the sensory buffer to produce overt orienting reactions.

The shift of attention is controlled by two basic learning systems. The *stimulus evaluation system* S* assigns a value to each stimulus based on its reward history and the *response evaluation system* R* assigns values to stimulus–response pairs. Together, the S* and R* systems implement an actor–critic (or two-process) architecture for learning (Sutton & Barto, 1998, Mowrer, 1960/1973). The learning process in S* is classical conditioning and in R* it is instrumental conditioning.

In addition, the three modules, S*, R* and R, are influenced by the context system that codes the current visual context or situation (Balkenius & Morén, 2000). Previously, we have primarily investigated the inhibitory role of the context (Balkenius, 2000, Balkenius & Morén, 2000, Morén, 2002). Here, the model is extended to allow excitatory influences of the context on the response evaluation system R* (Fig. 2).

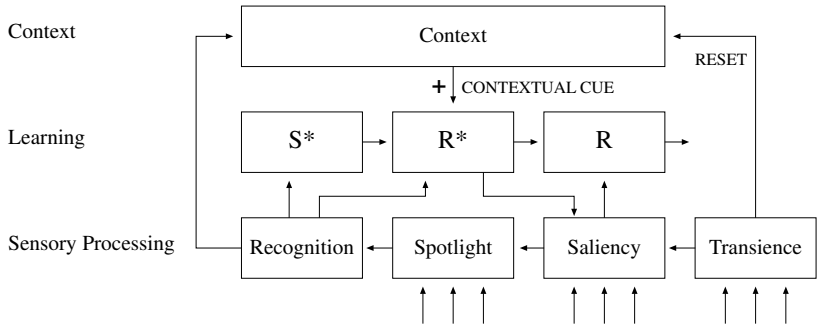

Figure 2: The modified model that allows contextual cueing from the context system to the response learning system. The sensory buffer is here divided into four components. *Transience*: Detection of image change; *Saliency*: The saliency map; *Spotlight*: Selection of input to recognition stage; *Recognition*: Detection of stimulus type. *S**, *R** and *R* as in Fig. 1

The Saliency Map

The model has recently been extended to process real visual input and not only abstract stimuli. To make this possible, the model has been extended with a more elaborate pre-attentive processing stage (Fig. 2). This processing stage was inspired by the model by Itti, Koch and Niebur (1998). The pre-attentive processing is applied to every location in the input image and consists of a filter bank that detects oriented lightness and color contrast as well as image motion. The output from all filters are integrated at each image location to give an initial salience at each part of the image. This salience estimation is used to calculate the probability of an attentional shift to that region. The salience over all image locations is thus used as a probability density function.

When an image location has been selected, a search is made for the local maximum on the salience map closest to the selected location. This location is subsequently used as target for the next attentional shift. An inhibition of return mechanism prevents the most salient image region from being selected repeatedly (Klein, 2000). As a result, a sequence of attentional shift will be produced over the image. This results in a passive, bottom-up mechanism for scanning a visual scene.

Sensory Processing

The output of the saliency map controls an attentional spotlight that moves over the image and selects a region for further processing (Fig. 2). In the simulated model, only the direction of the spotlight is controlled by the saliency map while its size is held constant. The stimulus in the spotlight of attention is sent to a recognition module and classified as belonging to one of the two categories target or distracter, In the next step, it is assigned a value by the stimulus evaluation module S*. This value is high for the target stimulus and low for distracter stimuli. The value of the stimulus in the current focus of attention also drives the learning in the target prediction system R*. In the simulations described below, there was no learning in S* itself (but see Morén, 2003).

Context System

The context system integrates information over several attentional shifts to form a code that remains constant in a certain visual situation although the attention (or gaze) moves back and fourth over the image. The creation of the context code proceedes in two steps (Balkenius & Morén, 2000). The first step consists of a binding between the stimulus and location code for each attentional fixation. These are bound together to form a compressed code that represents the conjunction between the stimulus and the attended location. The second step constructs a context code by a second binding process which constructs a compressed code for the set of all active bindings between a stimulus and a location.

The output from the context system is used for top-down control of the other modules in the model. An important aspect of the model is that while the bottom-up processing is controlled by the current focus of attention, the top-down control depends on a larger set of stimuli coded by the context system.

The context module also needs to know when a new attentional fixation is a part of the current context and when it signals a new context. In the previous version of the model, an explicit memory was used to generate prediction that could be matched to the actual input. In this way, the model would recognize when the context has changed and when to build a new context code.

In the current version, this mechanism is replaced by a module that looks at transient visual activation (Fig. 2). The amount of transient activity is calculated as the difference between two successive images. This can be seen as a simple form of mismatch between two successive inputs. The module is a part of the pre-attentive processing in the model. A sufficiently large transient activation will trigger an orienting reaction that will reset the context code and allow a new one to be built. This mechanism is sufficient for the situations investigated here. In a full model, both the explicit matching and the reactions to transient activation should be included.

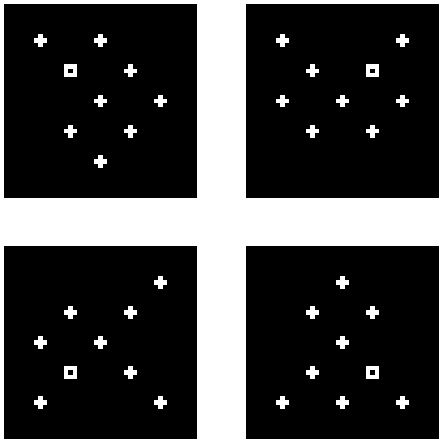

Figure 3: The four visual contexts used in the simulations.

Target Prediction

To predict the locations where the target will appear in the image, a response evaluation system R* is included in the model. It associates the current visual stimulus at the focus of attention and the context with the expected target location. This learning process is a form of instrumental learning and is controlled by the value-output from S*. A high value will function as reinforcement for the target predictor. When the model attends to a target stimulus, it will learn to predict its location in the current context.

The target predictor is implemented as a simple reinforcement learning algorithm (Sutton and Barto, 1998). The likelihood of finding a target at each location is calculated as a linear function of the current context coded as a vector. The associations between contexts and target locations are stored in a matrix which is updated each time the target is attended. Higher-order reinforcement was not included in the current model although it was included in the earlier model (Balkenius, 2000).

Simulations

The model was tested in computer simulations with four stimulus arrays as visual input (Fig. 3). Each stimulus array contains a single target feature (a square) and several distracter features (plus signs). The pre-attentive visual system was set up to initially assign a relatively smaller salience to the target feature than to the distracters to limit the amount of initial attention assigned to the target.

The distracters in each image were placed in such a way that that no individual distracter would identify the context and predict the target location. However, the spatial relations of several distracters would identify the context and could thus be used to predict the current target location. The model would thus have to pick up the correct distacter–context relation before it could use the context to guide the search for the target. There were four possible target locations.

Five different simulations were run with the model. In each simulation, the four stimulus arrays were presented in pseudo-random order to avoid any serial position effects.

In simulation A, the complete model was run to measure its performance on the four images in Fig. 3. Each stimulus array was presented for 50 simulated time steps (or ticks) followed by a blank screen for 50 time steps. After the presentation of each stimulus array, the time before the target stimulus was attended was measured. This was considered the reaction time. Since a trial did not finish until the target was attended, the accuracy of the model was always 100%, but the reaction time would go down if the model learned to predict the target locations.

Simulation B was identical to A except that the target prediction system R* was disconnected from R. The result of this manipulation was that the behavior of the model was controlled by top-down signals only since there was no way for the top-down signals to reach the output stage of the mode. This simulation was intended as a control group to show the effect of contextual cueing.

In simulation C, the effect of the context recognition was investigated. The current context was explicitly given as an input to the target predictor, thus bypassing the context system. This allowed R* to immediately generate target predictions without having to wait for the context system to identify the context. A comparison between target localization speed in simulation A and C indicates the time required by the context system to identify the context before it can be used by the target prediction system.

In simulation D, a single explicit context was given as input to the target prediction module for all four stimulus arrays. Thus, all contexts would appear identical to R*. The goal of this simulation was to investigate how much knowledge of the exact context (as in simulation A and C) decreases the reaction time compared to knowing only the four possible spatial locations. In this simulation, R* was expected to learn to predict that the target would be in one of the four possible locations although it could not predict which one.

Finally, in simulation E, the target prediction system received both the actual context as input and an additional constant context. This allowed the target prediction system to use both of the strategies of simulation A and D.

Results

The results of the five simulations are shown in Fig. 4 and Fig. 5. The reaction time in simulation A clearly decreases which implies that the model is learning the distracter–context and context–target relation. This contrasts with simulation B where no improvement in reaction time can be seen over time. When the context was explicitly given to the model in simulation C, the performance improved considerably. This was also the case

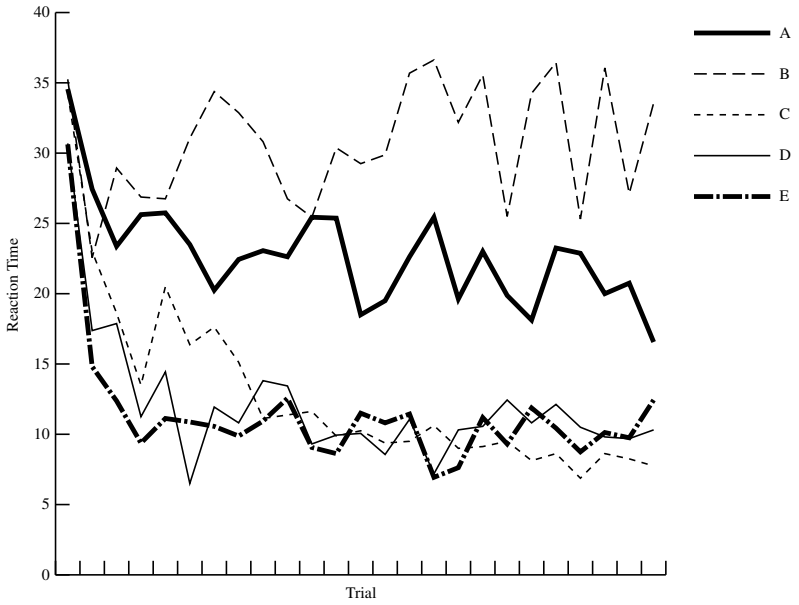

Figure 4: The development of the reaction during training in simulation A, B, C, D and E.

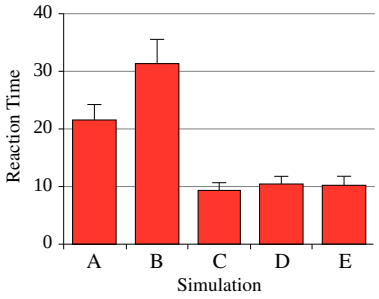

Figure 5: Mean reaction time after learning for each simulation.

in simulation D and E. In the last three simulations, the reaction time decreased much faster than in A.

The mean reaction times in the different simulations after learning was completed are shown in Fig. 5 together with their standard deviations. The difference in reaction times between contextual cueing in A (21.47 ticks) and no target prediction in B (31.25 ticks) is clearly seen. It is also clear that the difference between receiving a perfect context code as in C (9.28 ticks) and a single context code for all stimulus arrays as in D (10.38 ticks) is very small.

Discussion

The model can learn to predict a target location that depends on the visual context. When the contextual cueing had been learned, the reaction time decreased with 30% (simulation A) compared to when only bottom-up information was used (simulation B). However, the additional information given by the context can only be used once the model has decided on the identity of the present context and this takes some time. Comparing the final reaction time in simulation A and C shows that the average reaction time cost required to identify the context is 12.19 ticks which is 46% of the total reaction time in A. This implies that contextual cueing is only useful when the target localization task is sufficiently hard to warrant this extra cost. It also shows that if the context can be inferred by other means, the reaction time can be much reduced.

Interestingly, for the stimulus arrays used, knowing the four possible target locations is almost as good as knowing the exact location, and when the cost of identifying the context is taken into account, it outperforms the contextual cueing mechanism. This can be seen by comparing the reaction times in simulation C, D and E. This result is consistent with the experiment by Chun (2000), where contextual cueing had a significant, but small effect on target localization. The exact result obviously depends on the stimulus material used.

In Balkenius (2000), the different parts of the original computational model were associated with different functional brain areas. It was suggested that the contextual system resides in the hippocampus. This implies that patients with a damaged hippocampus should interfere with contextual cueing. This is the observation made by Chun and Phelps (1999). Evidence was also described that supports that the prefrontal cortex mediates the inhibitory contextual influence on visual attention. The excitatory influence on attention described here probably involves a direct influence of the hippocampus on motor systems.

The simulations make a number of qualitative predictions that could be tested experimentally. First, it sug-

gests that the reaction time in contextual cueing increases with the number of distracters in the scene as well as with the number of visual features that need to be attended before the context can be identified. This prediction is based on the assumption that attentional shifts are necessary to categorize the visual context.

Second, the model also suggests that the importance of contextual cueing increases when the target stimulus is less salient. By manipulating the relative salience of the distracters and targets, it should be possible to control how much contextual cueing contributes to a reduced reaction time.

Finally, the simulations predict that the gain from contextual cueing should increase when the number of target locations increases. In this case, it is no longer possible to learn a small number of probable target locations as in simulations D and E.

The model described here is an extension of the model described in Balkenius (2000), but lacks many parts of the original model. For example, the model simulated here did not have the ability to habituate to irrelevant distracters. Higher order conditioning was also excluded. This prevented the model from learning where in the visual array to look for contextual information. In the future, these mechanisms will be added again to investigate how they interact with the contextual cueing mechanism.

Acknowledgments

The code for the simulations presented in this paper are available as a part of the IKAROS project at http://www.lucs.lu.se/IKAROS/. I would like to thank Hideki Kozima for hosting me at CRL in Japan during the work with this article.

References

Balkenius, C. (2000). Attention, Habituation and Conditioning: Toward a Computational Model. *Cognitive Science Quarterly*, *1* (2), 171–214.

Balkenius, C., & Morén, J. (2000). A Computational Model of Context Processing. In J.-A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, & S. W. Wilson (Eds.), *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behavior* (pp. 256–265). Cambridge, MA: The MIT Press.

Biederman, I. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognit. Psychol. 14*, 143-177

Chun, M.M. and Jiang, Y. (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognit. Psychol. 36*, 28-71

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Science*, *4* (5), 170–178.

Chun, M.M. and Phelps, E.A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nat. Neurosci*, *2*, 844-847

Itti, L., Koch, C. and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans Patt Anal Mach Intell*, *20* (11), 1254–9.

Morén, J. (2002). *Emotion and Learning: A Computational Model of the Amygdala*, Lund University Cognitive Studies, 93.

Mowrer, O. H. (1960/1973). *Learning theory and behavior*. New York: Wiley.

Palmer, S.E. (1975). The effects of contextual scenes on the identification of objects. *Mem. Cognit. 3*, 519-526.

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

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Science*, *4* (4), 138–147.