The Interplay of Bottom-Up and Top-Down Mechanisms in Visual Guidance during Object Naming

Moreno I. Coco, George L. Malcolm and Frank Keller
School of Informatics
University of Edinburgh

A long standing controversy in visual cognition research regards the mechanisms and factors guiding the allocation of visual attention. There are two main theoretical views, which support either a bottom-up (stimulus-based, Itti and Koch 2000) or a top-down (object-based, Nuthmann and Henderson 2010) approach. The bottom-up approach assumes that visual attention is guided by low-level features of a scene (i.e., saliency). Saliency models can successfully predict fixation location for tasks which do not set up any explicit target object, i.e. the goals are underspecified. However, when a target is explicitly assigned, i.e. in visual search, saliency seems to be overridden by guidance based on top-down information about scene objects (e.g., Yang and Zelinsky 2009). The only model of visual attention that we are aware of, which includes contextual information is Torralba et al. [2006]. Such model uses a contextually modulated saliency map, and is able to predict fixation location better than saliency alone. Crucially, however, saliency information is still a key component of the model. So, it seems that both context and saliency have to be utilized in order to achieve reasonably good modelling performance.

However, on the experimental side, there is contradicting evidence about the role of saliency and context; especially when tested across different tasks. On one hand, saliency seems to elicit early fixation on targets in memorization or free-viewing [Underwood and Foulsham, 2006], but not across search tasks (see Underwood 2009 for a review). On the other hand, objects out-of-context seem to trigger a similar effect but in tasks involving naming and/or object recognition [Mackworth and Morandi, 1967].

Here, we present an eye-tracking object naming study that tries to clarify these theoretical issues by testing the interplay between saliency and context, while providing experimental data for further modelling work.

In a 2x2 design, participants are asked to name 5 most important objects in photo-realistic scenes. The saliency (Salient; Non-Salient) and contextual fit (In-Context; Out-Of-Context) of a target object are manipulated. General eye-movement measures (e.g., first pass) show effects of saliency on the latency and duration of first fixation, whereby a salient object is fixated earlier and for longer time than a non-salient object. Out-of-context objects are fixated later than salient objects, and are looked at overall less during the trial compared to in-context objects. Importantly, we also find the initiation time (i.e., the time to program the first eye-movement) to be significantly longer for out-of-context objects. Moreover, by exploring the time-course of fixations before and after naming, we confirm that out-of-context objects are looked at less before naming than salient objects, which are instead looked at more especially if contextually appropriate. Our results contrast with Underwood et al. [2007], where out-of-context objects are found to be processed earlier than in-context objects, and challenges accounts stating that saliency plays no role during goal-oriented tasks (e.g. Yang and Zelinsky 2009)

On the theoretical side, our results suggest that during gisting, visual attention is guided by contextual expectations; thus, objects violating the gist will take longer to be recognized. Instead, low-level information attracts visual attention to a salient target shortly after scene onset, facilitates its naming and maintains an effect on looks over the course of the whole trial, i.e., the effect of saliency does not seem to decay. On the computational side, our results suggest a model of visual attention based on both saliency and contextual information. In particular, saliency is used to modulate fixation location and give weight to the probability of naming a certain object; importantly such an effect does not linearly decay over time but it is rather related to the phase of the task, e.g., inspection vs naming. Contextual expectations are used to modulate the saliency predictions based on the semantic relation between the objects and the scene; thus, objects contextually appropriate are more likely to be named earlier and looked at for longer time.
References


