Towards Context-Sensitive Visual Attention

Nick Hawes and Jeremy Wyatt
School of Computer Science,
University of Birmingham, Birmingham, UK
n.a.hawes@cs.bham.ac.uk

Abstract

In this paper we present a discussion of information processing context and how we believe a visual attention system should be influenced by contextual information. We support this argument with a proof-of-concept design and implementation of a context-sensitive extension to the Itti & Koch model of visual attention as part of an architecture for a cognitive system. Our model demonstrates improved performance in terms of both fixations and processing time on visual search tasks compared to the non-extended model.

1 Introduction

Tsotsos argues that cognitive systems do not solve general vision tasks, instead presenting the case for visual attention as a form of guidance that reduces complexity [11]. In this paper we take this approach one step further. Just as cognitive systems do not solve general vision tasks, one might argue that they do not solve general (i.e. context free) *attention* tasks; invariably, attention is paid to an object by a cognitive system in order to perform a particular task.

We believe that a cognitive system can take advantage of information from the task it is engaged in to reduce the cost of its processing. Just as low-level visual attention exploits characteristics of visual scenes (e.g. intensity boundaries) to locate objects, a cognitive system should be able exploit the characteristics of its current task to reduce the cost of processing the visual input related to its task. This argument is not driven by a desire to replicate the performance of biological systems, although it is in part inspired by a human's ability to find relevant objects in a cluttered scene by focusing on particular visual characteristics of the object. Instead, we are driven by the need to do task-driven processing of complex scenes in real-time in a cognitive system.

In the rest of this paper we discuss our approach to information-processing context and what it means for a process to be sensitive to this. We then go on to discuss the design of a context-sensitive extension to the Itti & Koch visual attention model [4], and its implementation as part of a cognitive system. The performance of our attention model is then analysed, and directly compared to the Itti & Koch model. The performance of our attention model does not challenge the state-of-the-art of such systems, and accordingly we do not consider its design to be the main contribution of this paper. Instead, the model should be taken as an exploratory design and implementation to support the presented discussion of context-sensitivity and the further integration of visual systems into broader cognitive systems.

2 Context Sensitivity

Before examining our proposed model of context-sensitive visual attention, it is necessary to clarify our interpretation of both *context*, and what it means to be sensitive to it. We can initially consider the term context as being short for *task context*, i.e. what an agent is currently doing, and all the information it has that is related to this. From this we could generalise to a notion of *information processing context*, where the context is defined as the information that the agent is currently processing (e.g. its current goals, inputs, outputs, the contents of short-term memory etc.). Within this wide selection of information, some information may accurately capture the current task context (e.g. an agent's goals), whilst other information may be closely related to this (e.g. a plan based on a current goal), and still other information may be more distantly, or not at all, related (e.g. the charge remaining in the agent's power source). Ideally any discussion of specific information processing contexts should be presented with reference to an architecture for a cognitive system. In this instance we are still designing our architecture from various parts [7, 3], so such reference would be premature.

We define a mechanism as being *context-sensitive* if it is able to inform and alter its processing using information from the current information context. For example, if a vision architecture is composed of several separate object-specific recognition processes that usually all run simultaneously, it could be considered to be context-sensitive if was

able to run only a specific process when the context dictated that the associated object was being searched for, thereby saving processing resources. Given our previous, wide, definition of context, it will be necessary for a context-sensitive process to be selective about the information it uses as a basis for its decisions. Although this may suggest that a narrower definition of context should be used, it is more general to assume that a context-sensitive mechanism has access to any relevant information.

Given the above definition of context-sensitivity, we can now discuss why we want to make visual attention sensitive to context. The intended use for visual attention in artificial vision systems is to quickly and cheaply suggest a series of regions in an input image that may contain objects. These regions are selected based on their inherent *saliency*, which is calculated from the input image based on a given algorithm (e.g. [4, 12]). Once a region of the image has been attended to, it can be further examined by other visual processes (e.g. edge detection, object recognition etc.). Visual attention therefore reduces the processing demands of the entire visual system by restricting the use of more processing intensive algorithms to the regions of an image which are suspected to contain an object[1].

This account of attention in an artificial vision system assumes, amongst other things, that all of the regions of the image that might contain an object are of equal importance to the cognitive system doing the processing. Only once objects have been extracted from attended regions can their importance be determined by subsequent (non-visual) processes. If the visual attention mechanism was sensitive to context, then it should be able to select regions of the scene that are not only likely to contain objects, but to contain objects that are somehow relevant to the current context. For example, if the cognitive system was searching for an apple in a bowl of fruit, a context-sensitive visual attention system should return regions containing objects that could feasibly contain apples before returning other regions (e.g. ones that could contain bananas). If one of these early regions did contain an apple, then the need to process and evaluate other, irrelevant, regions would be avoided.

As the previous paragraph implies, making visual attention context-sensitive starts to give limited semantics to the list of regions selected by the attention system. Candidate regions are selected in order of the potential relevance of the object that they might contain. This could lead to a semantic variant of the *pop-out* phenomena used regularly within the visual attention literature (e.g. [2]). For example, if the agent is holding a hammer, then regions of the scene that contain objects that could be nails may become more salient, and therefore get selected earlier for processing.

[1]This and subsequent discussions within the paper assume that all vision is object based. This is not a view we actually subscribe to, but it is a useful simplifying assumption to make when discussing visual attention.

A further argument for adding context sensitivity to a visual attention system is to increase the integration of artificial vision systems into whole cognitive systems. Most visual attention work is done in isolation (although there are exceptions, e.g. [1]), but this ignores the purpose of searching for an object: to do something with it. In our cognitive systems we wish to avoid the division often placed between vision and further cognitive operations. By allowing attention to be influenced by other cognitive subsystems (e.g. planning, linguistic interactions etc.) we may be able to reduce the processing demands on both sides of this imaginary divide.

Context-sensitivity is more commonly referred to as *top-down attention* or *top-down saliency* [9, 2]. This phrase is used to imply that external processes can produce particular behaviours in the visual attention system. As the approach we take in this paper is more akin to the modulation of one process by other processes, possibly in a non-direct manner (i.e. the information used to modulate the attention system may not be generated specifically for that purpose), we prefer the term "context-sensitive" to "top-down", although our work could also be described using the latter term.

3 Scenario and Context

The work reported on in this paper is part of the ongoing CoSy project [2]. Part of the work of CoSy is concerned with the *PlayMate scenario*, in which a robot and a human interact with a tabletop of objects to perform various tasks[3]. The visual attention system presented here was designed with PlayMate-like tasks in mind, and therefore the rest of this paper will concern itself with just this context. This focus does not rule out applying the work to other scenarios, just that it has not been designed and evaluated with other scenarios in mind. By way of contrast, the work in [2] applies a similar technique to a mobile robot, much like the one considered by CoSy's Explorer scenario.

The PlayMate scenario gives a number of sources for information that could be considered to contribute to the information-processing context. Linguistic interactions with the human can provide various cues as to what the relevant objects on the table are, as can the PlayMate's internal representations of goals and plans. We can also consider the interactions in the past. For example, if the human has recently spoken about one type of object, then moves on to discuss another type of object, the previous object could retain some relevance to the current context (perhaps mediated by some kind of decay mechanism).

How all this heterogeneous information can be harnessed in a way that is useful to other processes within a cognitive

[2]For more information on CoSy see http://www.cognitivesystems.org.

[3]More information on the PlayMate scenario is available at http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html

system, including visual attention, is an open question. One of the CoSy group's early developments has been a multimodal dialogue architecture that uses ontologies to mediate between modalities [6, 7]. This has demonstrated the power of allowing separate processes to share knowledge via ontologies. We believe that one way to capture contextual knowledge in a non-intrusive way is to monitor the ontological entities accessed by various processes, and use the current set of active ontological entities to represent the context.

3.1 Experimental System

The visual attention system described in the rest of this paper has been implemented in C++ as part of an early prototype of a PlayMate-like cognitive system. The system is composed of the aforementioned multimodal dialogue system integrated with a SIFT-based vision system [8]. The system's basic abilities involve learning names for objects in its world and answering questions about the spatial arrangement of scenes constructed with collections of these objects [6]. The system can be asked questions such as "where is the Coke can?" and "what colour is the Pepsi can?". It answers with phrases such as "near the Pepsi can" and "blue"[4].

Given that the tasks that can be performed by our experimental system are relatively limited at the moment, our access to contextual information is also similarly limited. The visual routines that underlie the behaviour of the system involve locating the object being discussed, and any other objects necessary to describe its position. As such, we are faced with the visual search problem commonly tackled by other visual attention researchers.

Unfortunately, the only pieces of contextual information that we can reliably use with our early system are the names of the objects being discussed. This reduces the previously discussed notion of context from being something that could encompass various types of information, both abstract and concrete, to something that can be uniquely associated with the process of visual search for a particular object. Whilst this currently prevents us from expanding upon our more general notion of context, it allows us to focus more directly on the visual attention aspects of the problem.

4 Approach and Algorithm

To create a context-sensitive visual attention system, we first need to start with a basic visual attention system (as we still require its standard abilities). For this we selected the model developed by Itti & Koch [4]. We refer to this as the basic or standard model for the remainder of the paper, using the terms to describe the relation between this model and our extension to it (not in terms of the wider research field). Although there are other models (e.g. [12]), Itti & Koch's model is perhaps the most extensively documented, and is based on a modular design that lends itself well to the type of modifications we have performed.

The details of the Itti & Koch model have been explained in great detail elsewhere (e.g. [5, 4]), so we will only provide a brief summary of the relevant mechanisms. The model starts with an original image from which intensity, colour and orientation maps are extracted in a Gaussian pyramid of spatial scales. These maps are then passed through a process of centre-surround differencing and normalisation to produce a set of *feature maps* for intensity, colour and orientation. These maps contain information on within-feature contrasts at different spatial scales. Each set of feature maps is then combined across scales and normalised to produce a *conspicuity map* for each feature. Finally these conspicuity maps are combined linearly into a single *saliency map* which is used in visual search to determine the next location to be attended.

When various objects are placed in front of an implementation of the Itti & Koch model, the most reliable object-dependent changes in its internal processing occur within the feature and conspicuity maps based on the colour and intensity channels of the original image. Specific objects reliably generate signatures in the colour and intensity maps regardless of other image features. Whilst orientation information is useful for extracting an object from a cluttered background, it is not possible to reliably associate a particular object with a particular set of values across the orientation maps when the object could appear in any pose in any part of the image (e.g. standing upright or lying on its side).

Motivated by the previous observation we decided to base our design on learning weights for the colour and intensity maps in this model. This produces a model that has a number of strengths and weaknesses. It's main strength is that the information it uses is already available, so it is a computationally inexpensive addition. It's main weakness comes from the lack of discriminative power available when using only colour and intensity information. Given this, we decided to use this model as an exploratory study in context-sensitivity, which will allow us to investigate whether making extra information available to visual attention has benefits for a cognitive system. Further discussion of the system is presented in Section 6.

4.1 Allowing Attentional Biases

As mentioned previously, we are restricting context to a search for a particular object. To allow the attention system to be effected by the context, we must modify it to learn biases for each object it must search for.

[4] Movies at http://www.cs.bham.ac.uk/research/robotics/movies.php.

The process of generating the final saliency map in the standard Itti & Koch model involves many linear combinations of maps of various features. Although alterations to any of these combinations could have an effect on the final map, the combination of the conspicuity maps (for colour, intensity and orientation) into the saliency map presents the largest opportunity to affect the final outcome of the system. Therefore this is where we alter the standard model. First, rather than combine the red-green and blue-yellow conspicuity maps into a single colour conspicuity map, we directly combine them into the final saliency map. Second, we introduce weights into the final linear combination to allow the effects of the red-green, blue-yellow and intensity conspicuity maps to have a varying presence in the saliency map. It is these weights which allow the basic visual attention model to be context-sensitive.

4.2 Learning Appearances and Biases

Given the limited contextual information in our experimental set-up, we currently only learn attentional weights for individual objects (rather than more general contexts). Learning is triggered by a human using a mouse to indicate the region of the screen containing the object, and then saying to the robot "this is a" followed by a description of the object, e.g. "Coke can" or "Pepsi can". When the language subsystem has parsed the utterance and determined its meaning, a learning event is triggered in the vision subsystem. This causes the appearance of the object to be learnt by the vision system, and also causes the visual attention system to learn the attention weights for the object. This is done in the following way. First, the mean value of each of the conspicuity maps is calculated. Then the mean value of the object region is calculated for each conspicuity map. The basic weight for each conspicuity maps is then calculated by subtracting the region mean from the map mean. This means that a map's weight is based on how prominent the given region is on it. If the weight is large and positive then the region excites the map, if it is large and negative then it inhibits the map. If it is small then it has only a small effect on the map regardless of sign. Once the basic weights are calculated, they are normalised to be all positive and sum to 1. This is done to remove explicit inhibition, as we found that negatively weighting maps causes them to interact badly with the other maps by overriding their saliency information during the linear combination thus reducing the effectiveness of the standard visual attention model. We don't remove inhibition completely though, as the map with the lowest negative weighting before normalisation is normalised to zero, so its associated map is not active in the final saliency map. It was found that this method of learning weights produced the most reliable results during search. The reasons for this need to be fully explored. In the implementation, once the weights have been calculated for an object they are stored in a database indexed by a label generated by the language subsystem.

Figure 1. Saliency maps from the standard and context-sensitive models.

The effects of contextual weighting can be seen in Figure 1. This figure shows two saliency maps from the our experimental setup. The top map was generated by our implementation of the standard Itti & Koch model. The bottom map was generated by our context-sensitive model when given the task of finding the Pepsi can (the leftmost can).

4.3 Applying Attentional Biases

In the experimental system, a visual search is triggered by the user asking "can you see the" followed by a description of the object. Again the language subsystem interprets the meaning of the utterance and extracts the name of the object contained in the query. This name is then passed to the vision subsystem along with a command to initiate a search. The boolean result of this search is then used to generate an utterance in response to the query.

In the experimental setup with the context-sensitive model, the visual search process is very similar to the visual search process used in other visual attention research. When search is initiated, the conspicuity maps in the attention model are weighted using the values previously learnt for the object being searched for, and then the saliency map is generated. This map is used to determine most salient region of the input image, which is then passed to the SIFT system. This system extracts any objects contained within the region. The labels of these objects are then compared to the label used to initiate search. If the object is found then the search process terminates. If the object is not found

Can	Intensity	Red-Green	Blue-Yellow
Pepsi	0	0.04	0.96
Sprite	0.3	0	0.7
Lucozade	0.31	0.69	0
Diet Coke	0.17	0.83	0
Coke	0.31	0.69	0
Sunmagic	0.2	0	0.8

Table 1. Conspicuity map weights for the evaluation objects.

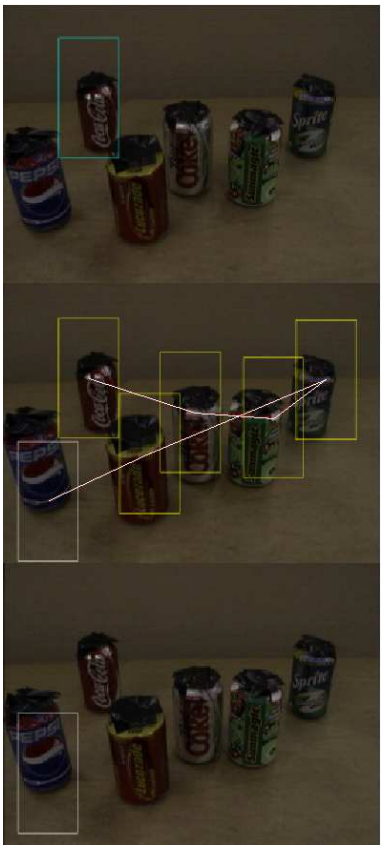

Figure 2. Three images showing the two attention models in action.

in the salient region, then this region is inhibited on the saliency map and another region is selected. Search terminates when either the object is found, or when the entire image has been considered by the search process.

5 Evaluation

To evaluate our context-sensitive visual attention model we used simple object search tasks to compare three different search models. The objects we used for the search tasks were soft drinks cans. These were selected because their various colourings offered different responses to both the basic attention model and our context-sensitive extension to it, their richly textured surfaces enabled the SIFT vision system to recognise them reliably, and their regular size allowed us to assume a fixed size region of interest for both learning and recognition. We selected six cans for our experiments. Table 1 shows the weights each can produced in the context-sensitive attention model.

For evaluation purposes, our implementation supports search without attention (the whole image is processed at once by the SIFT system), and search with visual attention without weighting (roughly equivalent to the standard Itti & Koch model), as well as search with context-sensitive visual attention. In the following results we denote these three approaches as *FI* (full image), *AT* (attention), and *CS* (context-sensitive) respectively. Search without attention is supported to provide empirical support for the use of visual attention in these types of tasks. Data from real integrated systems using attention and object recognition for search tasks is rarely seen in the attention literature, but it is included here because it should be of interest to researchers designing cognitive systems. Figure 2 shows the behaviour of the experimental system using the two attention models. The top image shows a scene and its most salient region as calculated using the basic model. The middle image shows the search path this map provides for the Pepsi can (considering regions of decreasing saliency until the can is found). The bottom image shows the search path for the Pepsi can when the context-sensitive model is used.

We measured the performance of the system using two related metrics: the average number of fixations per search, and the average search time in seconds. The number of fixations records the number of times a region of the image is selected for further processing. This is always one for the full image, as only one search is ever required. The search time records the number of seconds taken for the search process to find the target object. Fixations are useful for a high-level view of the system's performance. Search time provides a more direct measure of the performance of the system, and allows a direct comparison to be made between the full image and attentional approaches. In these experiments we did not consider the case in which the object was not present, although this is discussed in Section 6.

All the experiments were run with 640x480 video input from a B21r robot. The processes that form the cognitive architecture were distributed across 3 machines: a Pentium 3 based Linux PC on board the B21r was used for capturing the video input, a dual Opteron workstation was used for most of the internals of the architecture, and a Pentium M laptop ran the user interface and speech recognition. Region of interest size was fixed at 100x200. All experimental data in the rest of the paper is averaged over ten runs.

	Fixations			*Seconds*		
Can	FI	AT	CS	FI	AT	CS
Coke	1	1	1	2.00	0.60	0.61
Sprite	1	2	1	2.02	0.75	0.61
Coke	1	2	2	2.17	0.96	0.97
Lucozade	1	1	1	2.17	0.67	0.67

Table 2. Average search lengths for two dissimilar (above), and similar (below) objects.

To demonstrate the strengths and weaknesses of our approach Table 2 presents the results from experiments in which the system was asked to find one can in a scene containing only two cans. In the first of these experiments the cans were chosen because their context-sensitive weightings are quite different. In the second the cans were chosen because their weightings are very similar. The weightings can be seen in Table 1. In the first experiment the standard (i.e. non-weighted) attention model always finds the Coke can in the first fixation, followed by the Sprite can in the second. The application of the context-sensitive extension alters this behaviour in a positive way. The weightings cause the region containing the target object to be the most salient region in the image when the object is searched for. This in turn causes this region to be searched first, allowing the object to be found with a single fixation.

In the second experiment presented in Table 2 the standard model always finds the Lucozade can followed by the Coke can. In this case, the context-sensitive extension is ineffectual because of the similarity between the weightings of the two cans. This similarity means that the search order remains unchanged, and is based more on the behaviour of the basic model than the context-sensitive extension.

This demonstrates the basic behaviour of the system: either the target object becomes more salient based on the contextual weightings, or there is some *interference* between the weightings of the target object and the other objects in the scene, causing these other objects to be fixated upon before the target object (as in the case for the context-sensitive search for the Coke can in the second of the previously discussed experiments).

To investigate how the system behaves with more complex scenes, we produced twenty random arrangements of the six cans from Table 1. An example scene can be seen in Figure 2. For each scene we measured the performance of the system when asked to find each of the cans ten times using each of the possible search approaches. Example results of this for three different scenes are presented in Figures 3 to 5. The results of all of the experiments are summarised in the following paragraphs.

The general pattern across the experiments is that the Coke and Diet Coke cans are found early in the search by

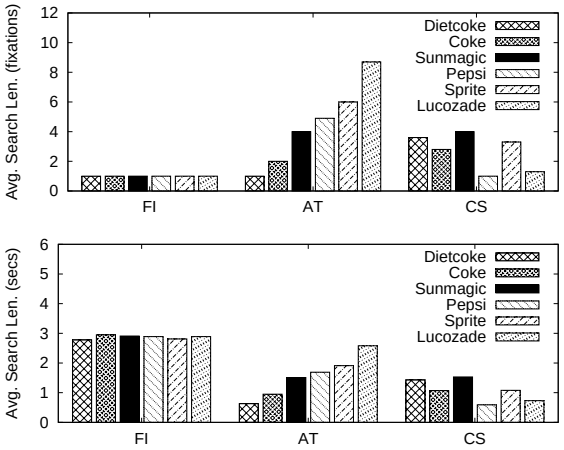

Figure 3. Results from scene 1

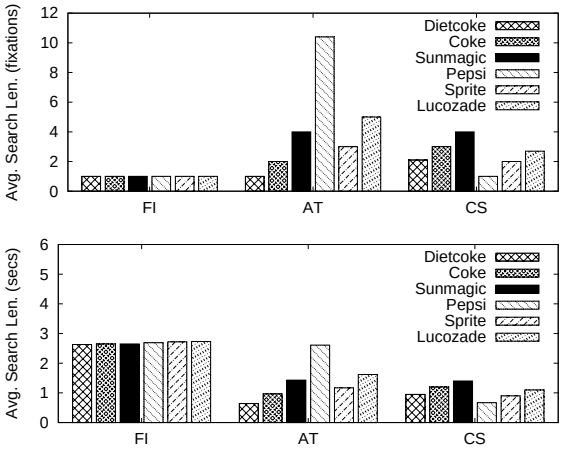

Figure 4. Results from scene 2

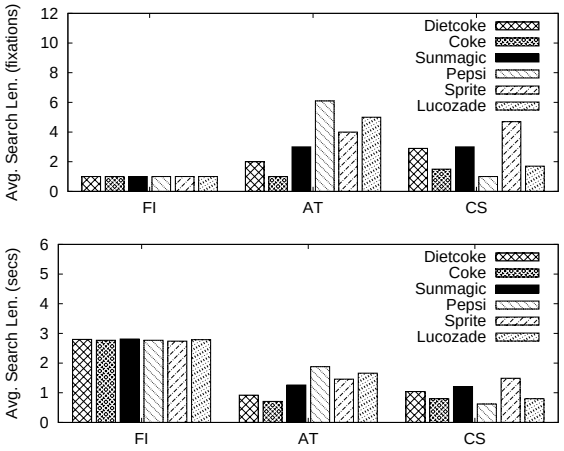

Figure 5. Results from scene 3

the basic attention model. The other cans are then found in an order that is fairly consistent across experiments on the same scene, but that varies across scenes. When the context-sensitive model is applied, it tends to slightly increase the length of the search for the objects usually found earlier by the standard model, whilst more significantly reducing the length of the search for the objects that it usually finds later. We believe this is because the objects that are found easily by the standard model stimulate more of its conspicuity maps, and therefore generate a stronger response in the final saliency map. The method we use to generate weights in our model tends to greatly discount the map that contributes least to the overall saliency of the object. This means that when the weightings are applied, objects that are salient in a single map, or in a pair of maps, have their saliency increased by a significant amount more than those that are salient across all of the maps. Objects that are salient in all three maps effectively have their saliency reduced because the weighting ignores one dimension of their saliency. This is illustrated by the results from the searches for the Pepsi can, which has a very strong weight on the Blue-Yellow conspicuity map. Using the standard model it is usually found after the first half of the objects have already been found. When the context-sensitive extension is used it is usually found on one of the first fixations.

This highlights a problem with our approach. If all the conspicuity maps are weighted, even by quite different values, the final context-sensitive saliency map varies very little from the map generated without weights. If only one or two of the conspicuity maps are weighted, then not all of the salient features of the target object are emphasised by the context-sensitive extension. A direct effect of this is that other objects may become more salient than the target object if their features are emphasised by the incorrect weightings. A more detailed weighting model, such as the one used in [2], would alleviate this problem to some degree, but any similar extension of the Itti & Koch model faces two related problems when dealing with non-toy environments. First, there is no guarantee that the learnt weightings for one object are unique to that object. Second, due to the relatively small number of distinguishing dimensions within the model, there is a significant chance of sub-sets of weightings being common across objects, and therefore the search for any one of these objects will consider all similarly weighted objects with an increased probability.

Despite these problems, the application of contextual information to the task of visual attention yielded positive effects in general. Across all of the experiments on all of the random scenes our context-sensitive attention model reduced the average time taken to find a particular object by 60% (2.75s to 1.1s) in a real-time comparison with the full image approach, and by 23% (1.43s to 1.1s) in a real-time comparison with our implementation of the Itti & Koch model. In terms of fixations our model takes 32% less fixations on average (4.03 fixations to 2.73 fixations) to find an object than our implementation of the Itti & Koch model.

6 Discussion

The first topic that it is necessary to discuss is the appropriateness of using colour and intensity to distinguish between different contexts. The results from our experiments demonstrate that colour and intensity information can distinguish between some objects, but this is quite different to characterising all visual input for a particular information processing context. It seems very unlikely that our current approach will be able to generate the kind of semantic pop-out described in Section 2, unless all of the associated visual entities shared a similar colour and intensity profile (which seems unlikely for non-trivial cases). One possible approach would be to determine the weights for the context by combining the weights for individual entities, but if the entities were sufficiently visually varied then this would just reproduce the behaviour of the standard visual attention model as all maps would be present to some degree.

Frintrop et al. [2] present a descendant of the Itti & Koch model that is modified in a similar way to the model presented here. They allow all of the maps generated by the attention model to be weighted in association with a particular object. Although this endows their model with increased discriminative power, it is still only more discriminative within the confines of colour, intensity and orientation. As our results show, this is useful for single object recognition (which is what Frintrop et al. use it for), but it would not allow us to model contexts involving multiple visually-distinct objects. Frintrop et al. weight both the initial feature maps (green, blue etc.) and the combined conspicuity maps. It seems that response in one is closely related to response in the other, so the weighting of both may not be required. That said, weighting the initial feature maps rather than the conspicuity maps appears to give a finer degree of within-feature discrimination, so in a future version of our model we will experiment with this approach.

The other key difference between the work presented here and that of Frintrop et al. is that they generate two overall maps: the standard saliency map, and a top-down map generated from the weighted maps. This is also the case for a number of other systems that use external information to weight a saliency map [9, 10]. This separation allows them to preserve the results of the standard saliency model, whilst also using results tailored to the particular search. This approach also allows them to have separate inhibitory and excitory object specific maps, although as they are combined into a single map before being combined with the basic saliency data this separation does not appear to be critical to the model. Although we eschew explicit inhibition

and choose to use relative excitation instead, this approach does not seem significantly different from the combination of separate inhibitory and excitory maps.

An interesting observation that has arisen from this work is that just as (visual) attention should be discussed with reference to the tasks it is being used for, it should also be discussed with reference to the architecture it is situated within. Without this, the effects of attention on the whole system cannot be evaluated. The SIFT system used in our experimental system can process both small regions of interest, and whole images, to recognise objects. As our results show, processing a whole image is inefficient when searching for a single object, but this is not the case when multiple objects are involved. If the system is asked to list all of the visible objects, then the effects of the task on the whole system become more apparent (this is similar to searching for an object which is not present). The cost of tackling this problem is made up of the cost of generating the saliency map, plus the cost of running the recogniser on regions of interest covering the whole image. The cost of tackling the problem with a single pass is the just cost of processing the whole image with the recogniser. Even if the performance of running the SIFT system on the whole image is the same as running it on number of small regions covering the same area (which is not the case) then there would always be the unnecessary overhead of generating the saliency map when the whole image must be searched. Although this problem could be overcome with the use of heuristics (e.g. search N regions of interest, where N could be heuristically determined), it does highlight how the efficacy of attentional systems are directly related to the performance of the systems they are providing information for.

7 Conclusions & Future Work

In this paper we presented a discussion of information processing context and how we believe a visual attention system should be influenced by contextual information. We presented an exploratory design and implementation of a context-sensitive extension to the Itti & Koch model of visual attention which demonstrated improved performance on visual search tasks when compared to the basic model. We also discussed a number of flaws with our approach, including its lack of discriminative power (which is also the case with similar models). In the future we intend to perform more direct comparisons with state-of-the-art attention systems (e.g. [2, 10]), and investigate other methods for allowing context to influence visual attention.

Acknowledgements

The research reported on in this paper was supported by the EU FP6 IST Cognitive Systems Integrated project Cognitive Systems for Cognitive Assistants “CoSy” FP6-004250-IP. The authors would also like acknowledge contributions made by other project members to both the theoretical and practical work presented here. In particular the work of colleagues at DFKI on the linguistic architecture and colleagues at the University of Ljubljana and TU Darmstadt on parts of the vision system.

References

[1] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *IJCAI ’99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146–1153, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[2] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Proceedings of the Annual Meeting of the German Association for Pattern Recognition (DAGM ’05)*, Wien, Austria, August 2005.

[3] N. Hawes, A. Sloman, and J. Wyatt. Requirements & designs: Asking scientifi c questions about architectures. In *Proceedings of AISB ’06: Adaptation in Artificial and Biological Systems*, volume 2, pages 52–55, April 2006.

[4] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res*, 40(10-12):1489–1506, 2000.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[6] G.-J. M. Kruijff, J. Kelleher, G. Berginc, and A. Leonardis. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI’06)*, Salt Lake City, UT, March 2006.

[7] G.-J. M. Kruijff, J. D. Kelleher, and N. Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In E. Andre, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, editors, *Perception and Interactive Technologies (PIT 2006)*. Springer Verlag, 2006.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999.

[9] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Res*, 45(2):205–231, January 2005.

[10] A. Torralba. Contextual influences on saliency. In G. R. Laurent Itti and J. K. Tsotsos, editors, *Neurobiology of Attention*. Elsevier, 2005.

[11] J. K. Tsotsos. Computational foundations for attentive processes. In G. R. Laurent Itti and J. K. Tsotsos, editors, *Neurobiology of Attention*. Elsevier, 2005.

[12] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nufb. Modeling visual attention via selective tuning. *Artif. Intell.*, 78(1-2):507–545, 1995.