A Goal Oriented Attention Guidance Model

Vidhya Navalpakkam and Laurent Itti

Departments of Computer Science, Psychology and Neuroscience Graduate Program
University of Southern California - Los Angeles, CA 90089
navalpak@usc.edu, itti@usc.edu

Abstract. Previous experiments have shown that human attention is influenced by high level task demands. In this paper, we propose an architecture to estimate the task-relevance of attended locations in a scene. We maintain a task graph and compute relevance of fixations using an ontology that contains a description of real world entities and their relationships. Our model guides attention according to a topographic attention guidance map that encodes the bottom-up salience and task-relevance of all locations in the scene. We have demonstrated that our model detects entities that are salient and relevant to the task even on natural cluttered scenes and arbitrary tasks.

1 Introduction

The classic experiment of Yarbus illustrates how human attention varies with the nature of the task [17]. In the absence of task specification, visual attention seems to be guided to a large extent by bottom-up (or image-based) processes that determine the salience of objects in the scene [11, 6]. Given a task specification, top-down (or volitional) processes set in and guide attention to the relevant objects in the scene [4, 1]. In normal human vision, a combination of bottom-up and top-down influences attract our attention towards salient and relevant scene elements. While the bottom-up guidance of attention has been extensively studied and successfully modelled [13, 16, 14, 8, 5, 6], little success has been met with understanding the complex top-down processing in biologically-plausible computational terms.

In this paper, our focus is to extract all objects in the scene that are relevant to a given task. To accomplish this, we attempt to solve partially the bigger and more general problem of modelling the influence of high-level task demands on the spatiotemporal deployment of focal visual attention in humans. Our starting point is our biological model of the saliency-based guidance of attention based on bottom-up cues [8, 5, 6]. At the core of this model is a two-dimensional topographic saliency map [9], which receives input from feature detectors tuned to color, orientation, intensity contrasts and explicitly encodes the visual salience of every location in the visual field. It biases attention towards focussing on the currently most salient location. We propose to extend the notion of saliency map by hypothesizing the existence of a topographic task-relevance map, which explicitly encodes the relevance of every visual location to the current task. In the proposed model, regions in the task-relevance map are activated top-down, corresponding to objects that have been attended to and recognized as being relevant. The final guidance of attention is derived from the activity in a further explicit topographic

map, the attention guidance map, which is the pointwise product of the saliency and task-relevance maps. Thus, at each instant, the model fixates on the most salient and relevant location in the attention guidance map.

Our model accepts a question such as "who is doing what to whom" and returns all entities in the scene that are relevant to the question. To focus our work, we have not for the moment attacked the problem of parsing natural-language questions. Rather, our model currently accepts task specification as a collection of object, subject and action keywords. Thus, our model can be seen as a question answering agent.

2 Related Work

Attention and identification have been extensively studied in the past. A unique behavioral approach to attention is found in [12] where the authors model perception and cognition as behavioral processes. They guide attention using internal models that store the sequence of eye movements and expected image features at each fixation. Their main thrust is towards object recognition and how attention is modulated in the process of object recognition. In contrast, we model human attention at a level higher than object recognition.

The Visual Translator (VITRA) [3] is a fine example of a real time system that interprets scenes and generates a natural language description of the scene. Their low level visual system recognises and tracks all visible objects and creates a geometric representation of the perceived scene. This intermediate representation is then analysed during high level scene analysis to evaluate spatial relations, recognise interesting motion events, and incrementally recognise plans and intentions. In contrast to VITRA, we track only those objects and events that we expect to be relevant to our task, thus saving enormously on computation complexity. The drawback of the VITRA project is its complexity that prevents it from being extended to a general attention model. Unlike humans that selectively perceive the relevant objects in the scene, VITRA attends to *all* objects and reports only relevant ones.

A good neural network model for covert visual attention has been proposed by Van der Laar [15]. Their model learns to focus attention on important features depending on the task. First, it extracts the feature maps from the sensory retinal input and creates a priority map with the help of an attentional network that gives the top down bias. Then, it performs a self terminating search of the saliency map in a manner similar to our salience model [6]. However, this system limits the nature of its tasks to pyschophysical search tasks that primarily involve bottom-up processes and are already fulfilled by our salience model successfully [7] (using databases of sample traffic signs, soda cans or emergency triangles, we have shown how batch training of the saliency model through adjustment of relative feature weights improves search times for those specific objects).

In [10], the authors propose a real time computer vision and machine learning system to model and recognize human behaviors. They combine top-down with bottom-up information in a closed feedback loop, using a statistical bayesian approach. However, this system focusses on detecting and classifying human interactions over an extended period of time and thus is limited in the nature of human behavior that it deals with. It

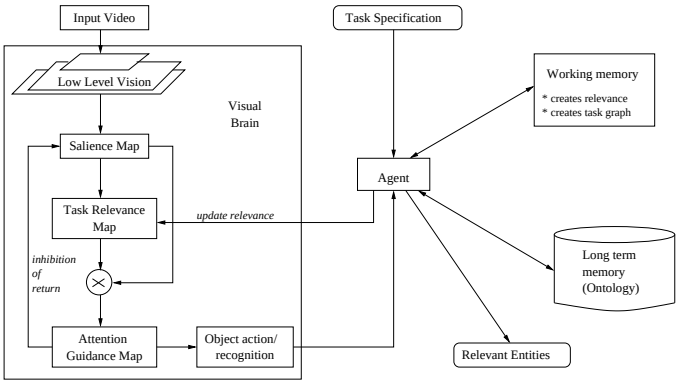

Fig. 1. An overview of our architecture

lacks the concept of a task/goal and hence does not attempt to model any goal oriented behavior.

3 Architecture

Our attention model consists of 4 main components: the visual brain, the working memory, the long term memory (LTM) and the agent. The visual brain maintains three maps, namely the salience map, task-relevance map and attention guidance map. The salience map (SM) is the input scene calibrated with salience at each point. Task-Relevance Map (TRM) is the input scene calibrated with the relevance at each point. Attention Guidance Map (AGM) is computed as the product of SM and TRM. The working memory (WM) creates and maintains the task graph that contains all entities that are expected to be relevant to the task. In order to compute relevance, the WM seeks the help of the long term memory that contains knowledge about the various real-world and abtract entities and their relationships. The role of the agent is to simply relay information between the visual brain and the WM; WM and the LTM. As such, its behavior is fairly prototyped, hence the agent should not be confused with a homunculus.

The schema of our model is as shown in figure 1. The visual brain receives the input video and extracts all low level features. To achieve bottom-up attention processing, we use the salience model previously mentioned, yielding the SM [8, 5, 6]. Then, the visual brain computes the AGM and chooses the most significant point as the current fixation. Each fixation is on a scene segment that is approximately the size of the attended object [2]. The object and action recognition module is invoked to determine the identity of the fixation. Currently, we do not yet have a generic object recognition module; it is done by a human operator. The agent, upon receiving the object identity from the visual brain, sends it to the WM. The WM in turn communicates with the LTM (via the agent) and determines the relevance of the current fixation. The estimated relevance of the current fixation is used to update the TRM . The current fixation is inhibited from returning in the SM. This is done to prevent the model from fixating on the same point continuously. The visual brain computes the new AGM and determines the next fixation. This process

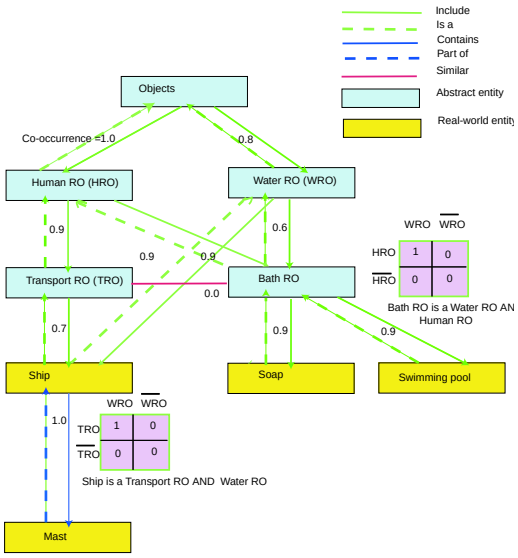

Fig. 2. A sample object ontology is shown. The relations include *is a*, *includes*, *part of*, *contains*, *similar*, *related*. While the first five relations appear as edges within a given ontology, the *related* relation appears as edges that connect the three different ontologies. The relations *contains* and *part of* are complementary to each other as in *Ship contains Mast*, *Mast is part of Ship*. Similarly, *is a* and *includes* are complementary. The co-occurrence measure is shown on each edge and the conjunctions, disjunctions are shown using the truth tables.

runs in a loop until the video is exhausted. Upon termination, the TRM is examined to find all the relevant entities in the scene.

The following subsections describe the important new components of our model in detail. The basic saliency mechanism has been described elsewhere [8, 5, 6].

3.1 LTM

The LTM acts as the knowledge base. It contains the entities and their relationships. Thus, for technical purposes, we refer to it as ontology from now on. As stated earlier, our model accepts task specification in the form of object, subject and action keywords. Accordingly, we have the object, subject and action ontology. In our current implementation, our ontology focusses primarily on human-related objects and actions. Each ontology is represented as a graph with entities as vertices and their relationships as edges. Our entities include real-world concepts as well as abstract ones. We maintain extra information on each edge, namely the granularity and the co-occurrence. Granularity of an edge ($g(u, v)$ where (u, v) is an edge) is a static quantity that is uniquely determined by the nature of the relation. The need for this information is illustrated with an example. While looking for the hand, fingers are considered more relevant than man because $g(hand, fingers) > g(hand, man)$. Co-occurrence of an edge ($c(u, v)$) refers to the probability of joint occurrence of the entities connected by the given edge. We illustrate the need for this information with another example. While looking for the hand, we consider pen to be more relevant than leaf because $c(hand, pen) > c(hand, leaf)$.

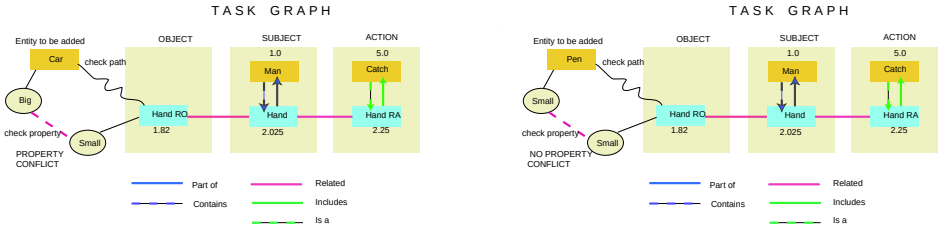

Fig. 3. To estimate the relevance of an entity, we check the existance of a path from entity to the task graph and check for property conflicts. While looking for a hand related object that is small and holdable, a big object like car is considered irrelevant; whereas a small object like pen is considered relevant.

Each entity in the ontology maintains a list of properties apart from the list of all its neighbours. These properties may also serve as cues to the object recognition module. To represent conjunctions and disjunctions or other complicated relationships, we maintain truth tables that store probabilities of various combinations of parent entities. An example is shown in figure 3.

3.2 WM

The WM estimates the relevance of a fixation to the given task. This is done in two steps. The WM checks if there exists a path from the fixation entity to the entities in the task graph. If yes, the nature of the path tells us how the fixation is related to the current task graph. If no such path exists, we declare that the current fixation is irrelevant to the task. This relevance check can be implemented using a breadth first search algorithm. The simplicity of this approach serves the dual purpose of reducing computation complexity (order of number of edges in task graph) and still keeping the method effective. In the case of object task graph, we perform an extra check to ensure that the properties of the current fixation are consistent with the object task graph (see figure 3). This can be implemented using a simple depth first search and hence, the computation complexity is still in the order of the number of edges in task graph which is acceptable.

Once a fixation is determined to be relevant, its exact relevance needs to be computed. This is a function of the nature of relations that connect the fixation entity to the task graph. It is also a function of the relevance of neighbours of the fixation entity that are present in the task graph. More precisely, we are guided by the following rules: the mutual influence on relevance between any two entities u and v decreases as a function of their distance (modelled by a $decay\_factor$ that lies between 0 and 1). The influence depends directly on the nature of the edge (u, v) that is in turn determined by the granularity ($g(u, v)$) and co-occurrence measures ($c(u, v)$). Thus we arrive at the following formula for computing relevance (R).

$$R_v = \max_{u:\ (u,v)\ is\ an\ edge} (R_u * g(u, v) * c(u, v) * decay\_factor) \quad (1)$$

Fig. 4. In the figure, the first column shows the original scene, followed by the TRM (locations relevant to the task) and finally, the attentional trajectory. The shapes represent fixations where each fixation is on a scene segment that is approximately the size of the object. The human operator recognized fixations as car, building, road or sky. When asked to find the cars in the scene, the model displayed results as shown in the first row. When asked to find the buildings in the scene, the model's results were as shown in the second row.

The relevance of a fixation depends on the entities present in the task graph. Hence, an important phase is the creation of the intial task graph. The initial task graph consists of the task keywords. For instance, given a task specification such as "what is John catching"; we have "John" as the subject keyword and "catch" as the action keyword. After adding these keywords to the task graph, we further expand the task graph through the "is a " relations. Our new task graph contains "John is a man", "catch is a hand related action". As a general rule, upon addition of a new entity into the task graph, we expand it to related entities. Here, we expand the initial task graph to "hand related action is related to hand and hand related object". Thus even before the first fixation, we have an idea about what entities are expected to be relevant. Once the initial task graph is formed, the model fixates and the WM finds the relevance of the new fixation based on the techniques discussed above. Upon addition of every entity into the task graph, its relevance is propagated to its neighbours.

4 Results

We tested our model on arbitrary scenes including natural cluttered scenes. To verify the model, we ran it on several images asking different questions on the same image and the same question on different images. On the same scene, our model showed different entities to be relevant based on the task specification. Two such examples are illustrated here. On a city scene, we asked the model to find the cars. Without any prior knowledge of a city scene, our model picked the relevant portions of the scene. On the same scene,

Fig. 5. In the figure, the first column is the original image, followed by the TRM after five attentional shifts and the final TRM after twenty attentional shifts. When asked to find the faces of people in the scene, the model displayed results as shown in the first row. When asked to determine what the people were eating, the model's results were as shown in the second row. The human operator recognized fixations as some human body part (face, leg, hand etc) or objects such as bottle, chandelier, plate, window, shelf, wall, chair, table.

when the model was asked to find the buildings, it attended to all the salient features in the buildings and determined the roads and cars to be irrelevant (see figure 4). On a natural cluttered scene, we asked the model to determine the faces of people in the scene and find what they were eating. As expected, the model showed that the relevance of entities in the scene varied with the nature of the task. For the first task, the model looked for human faces and consequently, it marked human body parts as relevant and other objects as irrelevant. While in the second task, the model looked for hand related objects near the human faces and hands to determine what the people were eating (see figure 5). Thus, even in arbitrary cluttered scenes, our model picks up the entities relevant to the current task.

5 Discussion and Outlook

Our broader goal is to model how internal scene representations are influenced by current behavioral goals. As a first step, we estimate the task-relevance of attended locations. We maintain a task graph in working memory and compute relevance of fixations using an ontology that contains a description of worldly entities and their relationships. At each instant, our model guides attention based on the salience and relevance of entities in the scene. At this infant stage, most of the basic components of our proposed architecture are in place and our model can run on arbitray scenes and detect entities in the scene that are relevant to arbitrary tasks.

Our approach directly contrasts with previous models (see section 2) that scan the entire scene, track all objects and events and subsequently analyze the scene to finally determine the task-relevance of various objects. Our aim is to prune the search space, thereby performing as few object identifications and attentional shifts while trying to analyse the scene. Towards this end, our salience model serves as a first filtration phase where we filter out all non salient locations in the scene. As a second phase of filtration, we attempt to further prune the search space by determining which of these salient locations is relevant to the current task. Thus, our approach is to perform minimal attentional shifts and to incrementally build up knowledge of the scene in a progressive manner.

At this preliminary stage, the model has several limitations. It cannot yet make directed attentional shifts, nor does it support instantiation. In future, we plan to expand the ontology to include more real-world entities and model complex facts. We also plan to allow instantiation such as "John is an instance of a man"; where each instance is unique and may differ from each other. Including directed attentional shifts into our model would require that spatial relations also be included in our ontology (e.g., look up if searching for a face but found a foot) and would allow for more sophisticated top-down attentional control. Knowledge of such spatial relationships will also help us prune the search space by filtering out most irrelevant scene elements (e.g., while looking for John, if we see Mary's face, we can also mark Mary's hands, legs etc are irrelevant provided we know the spatial relationships). Several models already mentioned provide an excellent starting point for this extension of our model [12]. Finally, there is great opportunity within our new framework for the implementation of more sophisticated rules for determining the next shift of attention based on task and evidence accumulated so far. This in turn will allow us to compare the behavior of our model against human observers and to obtain a greater understanding of how task demands influence scene analysis in the human brain.

6 Acknowledgements

We would like to thank all ilab members for their help and suggestions. This research is supported by the Zumberge Faculty Innovation Research Fund.

References

1. M Corbetta, J M Kincade, J M Ollinger, M P McAvoy, and G L Shulman. Voluntary orienting is dissociated from target detection in human posterior parietal cortex [published erratum appears in nat neurosci 2000 may;3(5):521]. *Nature Neuroscience*, 3(3):292–297, Mar 2000.
2. D Walther, L Itti, M Reisenhuber, T Poggio, C Koch. Attentional Selection for Object Recognition - a Gentle Way. *BMCV2002*, in press.
3. Gerd Herzog and Peter Wazinski. VIsual TRAnslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2-3):175–187, 1994.
4. J B Hopfinger, M H Buonocore, and G R Mangun. The neural mechanisms of top-down attentional control. *Nature Neuroscience*, 3(3):284–291, Mar 2000.
5. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

6. L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.
7. L. Itti and C. Koch. Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Journal of Electronic Imaging*, 10(1):161-169, Jan 2001.
8. 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.
9. C Koch and S Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–27, 1985.
10. Nuria M. Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.
11. D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, Jan 2002.
12. I. A. Rybak, V. I. Gusakova, A.V. Golovan, L. N. Podladchikova, and N. A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Research*, 38:2387–2400, 1998.
13. A M Treisman and G Gelade. A feature-integration theory of attention. *Cognitive Pyschology*, 12(1):97–136, Jan 1980.
14. J K Tsotsos. Computation, pet images, and attention. *Behavioral and Brain Sciences*, 18(2):372, 1995.
15. van de P. Laar, T. Heskes, and S. Gielen. Task-dependent learning of attention. *Neural Networks*, 10(6):981–992, 1997.
16. J M Wolfe. Visual search in continuous, naturalistic stimuli. *Vision Research*, 34(9):1187–95, May 1994.
17. A Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.