Attention and Scene Understanding

Vidhya Navalpakkam, Michael Arbib and Laurent Itti

University of Southern California, Los Angeles

Abstract

Abstract: *This paper presents a simplified, introductory view of how visual attention may contribute to and integrate within the broader framework of visual scene understanding. Several key components are identified which cooperate with attention during the analysis of complex dynamic visual inputs, namely rapid computation of scene gist and layout, localized object recognition and tracking at attended locations, working memory that holds a representation of currently relevant targets, and long-term memory of known world entities and their inter-relationships. Evidence from neurobiology and psychophysics is provided to support the proposed architecture.*

Key words: attention, scene understanding, bottom-up, top-down, gist, visual working memory, object recognition

1 Introduction

Primates, including humans, use focal visual attention and rapid eye movements to analyze complex visual inputs in real-time, in a manner that highly depends on current behavioral priorities and goals. A striking example of how a verbally-communicated task specification may dramatically affect attentional deployment and eye movements was provided by the pioneering experiments of Yarbus (1967). Using an eye-tracking device, Yarbus recorded the scanpaths of eye movements executed by human observers while analyzing a single photograph under various task instructions **(Fig. 1)**. Given the unique visual stimulus used in these experiments, the highly variable spatiotemporal characteristics of the measured eye movements for different task specifications exemplify the extent to which behavioral goals may affect eye movements and scene analysis.

Subsequent eye-tracking experiments during spoken sentence comprehension have further demonstrated how, often, the interplay between task demands and active vision is reciprocal. For instance, Tanenhaus *et al.* (1995) tracked eye movements while subjects received ambiguous verbal instructions about manipulating objects lying on a table in front of them. They demonstrated not only that tasks influenced eye movements, but also that visual context influenced spoken word recognition and mediated syntactic processing, when the ambiguous verbal instructions could be resolved through visual analysis of the objects present in the scene.

Building computational architectures that can replicate the interplay between task demands specified at a symbolic level (e.g., through verbally-delivered instructions) and scene contents captured by an array of photoreceptors is a challenging task. We review several key components and achitectures that have attacked this problem, and explore in particular the involvment of focal visual attention during goal-oriented scene understanding.

2 Basic Components of Scene Understanding

A recent overview of a computational architecture for visual processing in the primate brain was provided by Rensink (2000), and is used as a starting point for the present analysis. In Rensink's triadic architecture, low-level visual features are computed in parallel over the entire visual field, up to a level of complexity termed proto-objects (an intermediary between simple features such as edges and corners, and sophisticated object representations). One branch of subsequent processing is concerned with the computation of the so-called setting, which includes a fairly crude semantic analysis of the nature of the scene (its "gist", see [Gist of a Scene]), e.g., whether it is a busy city street, a kitchen or a beach; and its coarse spatial layout). In the other branch, attention selects a small spatial portion of the visual inputs and transiently binds the volatile proto-objects into coherent representations of attended objects. Attended objects are then processed in further details, yielding the recognition of their identity and attributes. **Fig. 2** builds upon and extends this purely visual architecture.

As the present review focuses on computational modeling of the interaction between cognitively-represented task demands and this simplified view of visual processing, we refer the reader to several articles in this volume which explore the putative components of the triadic architecture in further details (e.g., [Contextual influences on saliency], [Contextual guidance of visual attention], [Models of bottom-up attention and saliency]).

Visual attention has been often compared to a virtual spotlight through which our brain sees the world [Neurophysiological Correlates of the Attentional Spotlight], and shifts of attention have been classified into several types based on whether or not they involve eye movements (overt vs. covert), and are guided primarily by scene features or volition (bottom-up vs. top-down) (for review, see (Itti & Koch, 2001), [Models of bottom-up attention and saliency]). The first explicit biologically plausible computational architecture for controlling bottom-up attention was proposed by Koch and Ullman (1985) (also see (Didday & Arbib, 1975)). In their model, several feature maps (such as color, orientation, intensity) are computed in parallel across the visual field (Treisman & Gelade, 1980), and combined into a single salience map. Then, a selection process sequentially deploys attention to locations in decreasing order of their salience. We here assume a similar architecture for the attentional branch of visual processing and explore how it can be enhanced by modeling the influence of task on attention. The choice of features which may guide attention bottom-up has been extensively studied in the visual search literature [Guidance of Visual Search by Preattentive Information].

At the early stages of visual processing, task modulates neural activity by enhancing the responses of neurons tuned to the location and features of a stimulus [Attentional Modulation of Apparent Stimulus Contrast], [Biasing competition in human visual cortex], [Non-sensory signals in early visual cortex], [Attention Improves Discrimination Psychophysics] and many others in this book. In addition, psychophysics experiments have shown that knowledge of the target contributes to an amplification of its salience, e.g., white vertical lines become more salient if we are looking for them (Blaser *et al.*, 1999). A recent study even shows that better knowledge of the target leads to faster search, e.g., seeing an exact picture of the target is better than seeing a picture of the same semantic type or category as the target (Kenner & Wolfe, 2003). These studies demonstrate the effects of biasing for features of the target. Other experiments (e.g., (Treisman & Gelade, 1980)) have shown that searching for feature conjunctions (e.g., color×orientation conjunction search: find a red-vertical item among red-horizontal and green-vertical items) are slower than “pop-out” (e.g., find a green item among red items). These observations impose constraints on the possible biasing mechanisms and eliminate the possibility of generating new composite features on the fly (as a combination of simple features).

A popular model to account for top-down feature biasing and visual search behavior is *Guided Search* (Wolfe, 1994). This model has the same basic architecture as proposed by Koch and Ullman (1985), but, in addition, achieves feature-based biasing by weighing feature maps in a top-down manner. For example, with the task of detecting a red bar, the red-sensitive feature map gains more weight, hence making the red bar more salient. One question that remains is how the optimally set the relative feature weights such as to maximize the detectability of a set of behaviorally relevant targets among clutter (Navalpakkam & Itti, 2004).

Given a saliency map, several models have been proposed to select the next attended location, including various forms of winner-take-all (maximum-selector) algorithms (see [The Selective Tuning Model of Attention], [Models of bottom-up attention and saliency], [Probabilistic Models of Attention based on Iconic Representations and Predictive Coding], [The FeatureGate Model of Visual Selection]).

Having selected the focus of attention, it is important to recognize the entity at that scene location. Many recognition models have been proposed that can be classified based on factors including the choice of basic primitives (e.g., Gabor jets, geometric primitives like geons, image patches or blobs, and view-tuned units), the process of matching (e.g., self organizing dynamic link matching, probabilistic matching), and other factors (for review, see [Object recognition in cortex: Neural mechanisms, and possible roles for attention]).

Recognition is followed by the problem of memorization of visual information. A popular theory, the *object file theory of trans-saccadic memory* (Irwin, 1992), posits that when attention is directed to an object, the visual features and location information are bound into an object file (Kahneman & Treisman, 1984) that is maintained in visual short term memory across saccades. Psychophysics experiments have further shown that up to three or four object files may be retained in memory (Irwin, 1992). Studies investigating the neural substrates of working memory in primates and humans suggest that the frontal and extrastriate cortices may both be functionally and anatomically separated into a "what" memory for storing the visual features of the stimuli, and a "where" memory for storing spatial information (Wilson *et al.*, 1993).

To memorize the location of objects, we here extend the earlier hypothesis of a salience map (Koch & Ullman, 1985) to propose a two-dimensional topographic *task-relevance map* (TRM) that encodes the task-relevance of scene entities. Our motivation for maintaining various maps stems from biological evidence. Single-unit recordings in the visual system of the macaque indicate the existence of a number of distinct maps of the visual environment that appear to encode the salience and/or the behavioral significance of targets. Such maps have been found in the superior colliculus, the inferior and lateral subdivisions of the pulvinar, the frontal-eye fields and areas within the intraparietal sulcus [Dissociation of Selection from Saccade Programming], [Prefrontal Selection and Control of Covert and Overt Orienting]. Since these neurons are found in different parts of the brain that specialize in different functions, we hypothesize that they may encode different types of salience: the posterior parietal cortex may encode a visual salience map, while the pre-frontal cortex may encode a top-down task-relevance map, and the final eye movements may be generated by integrating information across the visual salience map and task-relevance map to form an attention guidance map possibly stored in the superior colliculus.

Our analysis so far has focused on the attentional pathway. As shown in figure 2, non-attentional pathways also play an important role; in particular, rapid identification of the gist (semantic category) of a scene is very useful in determining scene context, and is known to guide eye movements [Gist of a Scene]. It is computed rapidly within the first 150ms of scene onset, and the neural correlate of this computation is still unknown [Visual Saliency and Spike Timing in the Ventral Visual Pathway]. Recently, Torralba [Contextual influences on saliency] proposed holistic representation of the scene based on spatial envelope properties (such as openness, naturalness etc.) that bypasses the analysis of component objects and represents the scene as a single identity. This approach formalizes the gist as a vector of contextual features. By processing several annotated scenes, these authors learned the relationship between the scene context and categories of objects that can occur, including object properties such as locations, size or scale, and used it to focus attention on likely target locations. This provides a good starting point for modeling the role of gist in guiding attention. Since the gist is computed rapidly, it can serve as an initial guide to attention. But subsequently, our proposed TRM that is continuously updated may serve as a better guide. For instance, in dynamic scenes such as traffic scenes where the environment is continuously changing and the targets such as cars and pedestrians are moving around, the gist may remain unchanged and hence, it may not be so useful, except as an initial guide.

The use of gist in guiding attention to likely target locations motivates knowledge-based approaches to modeling eye movements, in contrast to image-based approaches. One such famous approach is the scanpath theory which proposes that attention is mostly guided in a top-down manner based on an internal model of the scene [Scanpath Theory, Attention and Image Processing Algorithms for Predicting Human Eye Fixations]. Computer vision models have employed a similar approach to recognize objects. For example, Rybak *et al.* [Attention-Guided Recognition Based on "What" and "Where" Representations: A Behavioral Model] recognize objects by explicitly replaying a sequence of eye movements and matching the expected features at each fixation with the image features. (also see [A Model of Attention and Recognition by Information Maximization], [Probabilistic Models of Attention based on Iconic Representations and Predictive Coding], [Attention Architectures for Machine Vision and Mobile Robots].

To summarize, we have motivated the basic architectural components which we believe are crucial for scene understanding. Ours is certainly not the first attempt to address this problem. For example, one of the finest examples of real-time scene analysis systems is *The Visual Translator* (VITRA) (Herzog

& Wazinski, 1994), a computer vision system that generates real-time verbal commentaries while watching a televised soccer game. Their low-level visual system recognizes and tracks all visible objects from an overhead (bird's eye) camera view, and creates a geometric representation of the perceived scene (the 22 players, the field and the goal locations). This intermediate representation is then analyzed by series of Bayesian belief networks which evaluate spatial relations, recognize interesting motion events, and incrementally recognize plans and intentions. The model includes an abstract, non-visual notion of salience which characterizes each recognized event on the basis of recency, frequency, complexity, importance for the game, and other factors. The system finally generates a verbal commentary, which typically starts as soon as the beginning of an event has been recognized but may be interjected if highly salient events occur before the current sentence has been completed. While this system delivers very impressive results in the specific application domain considered, due to its computational complexity it is restricted to one highly structured environment and one specific task, and cannot be extended to a general scene understanding model. Indeed, unlike humans who selectively perceive the relevant objects in the scene, VITRA attends to and continuously monitors *all* objects and attempts to simultaneously recognize *all* known actions. The approach proposed here differs from VITRA not only in that there is nothing in our model that commits it to a specific environment or task. In addition, we only memorize those objects and events that we expect to be relevant to the task at hand, thus saving enormously on computation complexity.

3 Discussion: Summary of a Putative Functional Architecture

In this section, we present a summary of an architecture which can be understood in four phases (figure 3). Partial implementation and testing of this architecture has been developed elsewhere (Navalpakkam & Itti, 2004), so that we here mainly focus on reviewing the componants and their interplay during active goal-oriented scene understanding.

3.1 Phase 1: Eyes Closed

In the first phase known as the "eyes closed" phase, the symbolic working memory (WM) is initialized by the user with a task definition in the form of keywords and their relevance (any number greater than baseline 1.0). Given the relevant keywords in symbolic WM, volitional effects such as "look at the center of the scene" could be achieved by allowing the symbolic WM to bias the TRM so that the center of the scene becomes relevant and everything else is irrelevant (but our current implementation has not explored this yet). For more complex tasks such as “who is doing what to whom,” the symbolic WM requires prior knowledge and hence, seeks the aid of the symbolic long-term memory (LTM). For example, to find what the man in the scene is eating, prior knowledge about eating being a mouth and hand-related action, and being related to food items helps us guide attention towards mouth or hand and determine the food item. Using such prior knowledge, the symbolic WM parses the task and determines the task-relevant targets and how they are related to each other. Our implementation (Navalpakkam & Itti, 2004) explores this mechanism using a simple hand-coded symbolic knowledge base to describe long-term knowledge about objects, actors and actions. Next, it determines the current most task-relevant target as the desired target. To detect the desired target in the scene, the visual WM retrieves the learned visual representation of the target from the visual LTM and biases the low-level visual system with the target’s features.

3.2 Phase 2: Computing

In the second phase known as the “computing” phase, the eyes are open and the visual system receives the input scene. The low-level visual system that is biased by the target’s features computes the biased salience map. Apart from such feature-based attention, spatial attention may be used to focus on likely target locations, e.g., gist and layout may be used to bias the TRM to focus on relevant locations (but this is not implemented yet). Since we are interested in attending to locations that are salient and relevant, the biased salience and task-relevance maps are combined by taking a pointwise product to form the attention-guidance map (AGM). To select the focus of attention, we deploy a Winner-take-all competition that chooses the most active location in the AGM. It is important to note that there is no intelligence in this selection and all the intelligence of the model lies in the WM.

3.3 Phase 3: Attending

In the third phase known as the “attending” phase, the low-level features or prototype objects are bound into a spatio-temporal structure or a mid-level representation called the “coherence field” (a mid-level representation of an object formed by grouping prototype objects into a spatio-temporal structure that exhibits coherence in space and time. Spatial coherence implies that the prototype objects at the different locations belong to the same object, and temporal coherence implies that different representations across time refer to the same object (Rensink, 2000); in our implementation, this step simply extracts a vector of visual features at the attended location). The object recognition module determines the identity of the entity at the currently attended location, and the symbolic WM estimates the task-relevance of the recognized entity (Navalpakkam & Itti, 2004).

3.4 Phase 4: Updating

In the final phase known as the "updating" phase, the WM updates its state (e.g., records that it has found the man's hand). It updates the TRM by recording the relevance of the currently attended location. The estimated relevance may influence attention in several ways. For instance, it may affect the duration of fixation (not implemented). If the relevance of the entity is less than the baseline 1.0, it is marked as irrelevant in the TRM, and hence will be ignored by preventing future fixations on it (e.g., a chair is irrelevant when we are trying to find what the man is eating. Hence, if we see a chair, we ignore it). If it is somewhat relevant (e.g., man's eyes), it may be used to guide attention to a more relevant target by means of directed attention shifts (e.g., look down to find the man's mouth or hand; not implemented). Also if it is relevant (e.g., man's hand), a detailed representation of the scene entity may be created for further scrutiny (e.g., a spatio-temporal structure for tracking the hand; not implemented). The WM also inhibits the current focus of attention from continuously demanding attention (inhibition of return in SM). Then, the symbolic WM determines the next most task-relevant target, retrieves the target's learned visual representation from visual LTM, and uses it to bias the low-level visual system.

This completes one iteration. The computing, attending and updating phases repeat until the task is complete. Upon completion, the TRM shows all task-relevant locations and the symbolic WM contains all task-relevant targets.

4 Conclusion

We have presented a brief overview of how some of the crucial components of primate vision may interact during active goal-oriented scene understanding. From the rather partial and sketchy figure proposed here, it is clear that much systems-level, integrative research will be required to further piece together more focused and localized studies of the neural subsystems involved.

References

Blaser, E., Sperling, G., & Lu, Z. L. 1999. Measuring the amplification of attention. *Proc. Natl. Acad. Sci. USA*, **96**(20), 11681–11686.

Didday, R. L., & Arbib, M. A. 1975. Eye Movements and Visual Perception: A “Two Visual System” Model. *Int. J. Man-Machine Studies*, **7**, 547–569.

Herzog, G., & Wazinski, P. 1994. VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. *Artificial Intelligence Review*, **8**(2-3), 175–187.

Irwin, D. E. 1992. Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 307–317.

Itti, L., & Koch, C. 2001. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

Kahneman, D, & Treisman, A. 1984. Changing views of attention and automaticity. *Pages 29–61 of:* Parasuraman, D, Davies, R, & Beatty, J (eds), *Varieties of attention*. New York, NY: Academic.

Kenner, N., & Wolfe, J. M. 2003. An exact picture of your target guides visual search better than any other representation [abstract]. *Journal of Vision*, **3**(9), 230a.

Koch, C, & Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**(4), 219–27.

Navalpakkam, V., & Itti, L. 2004. Modeling the influence of task on attention. *Vision Research, in press*.

Rensink, R. A. 2000. The dynamic representation of scenes. *Visual Cognition*, **7**, 17–42.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, **268**(5217), 1632–1634.

Treisman, A. M., & Gelade, G. 1980. A feature-integration theory of attention. *Cognit. Psychol.*, **12**(1), 97–136.

Wilson, F.A., O Scalaidhe, S.P., & Goldman-Rakic, P.S. 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, **260**, 1955–1958.

Wolfe, J. M. 1994. Guided search 2.0: a revised model of visual search. *Psyonomic Bulletin and Review*, **1(2)**, 202–238.

Yarbus, A. 1967. *Eye Movements and Vision*. Plenum Press, New York.

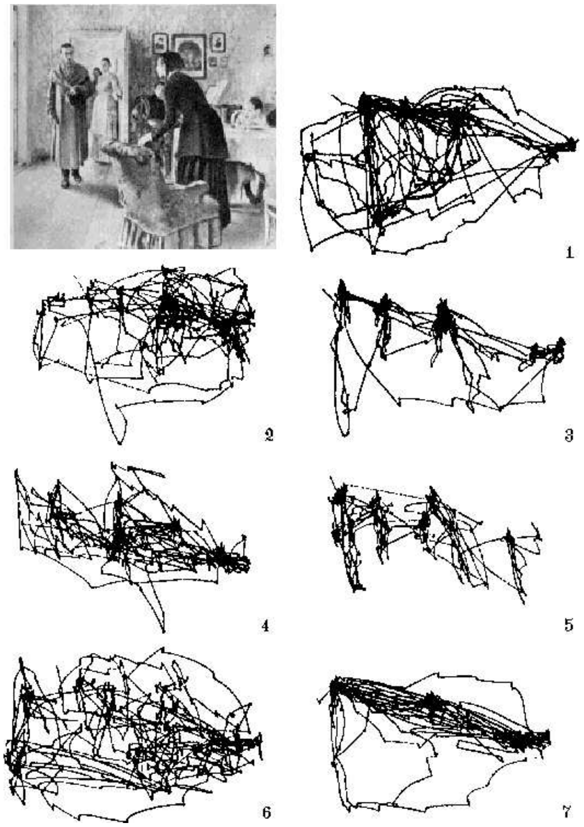

Fig. 1. Stimulus (top-left) and corresponding eye movement traced recorded from human observers under seven task specifications: 1) free examination of the picture, 2) estimate the material circumstances of the family in the picture, 3) give the ages of the people, 4) surmise what the family had been doing before the arrival of the "unexpected visitor," 5) remember the clothes worn by the people, 6) remember the position of the people and objects in the room, and 7) estimate how long the "unexpected visitor" had been away from the family. (from Yarbus, 1967).

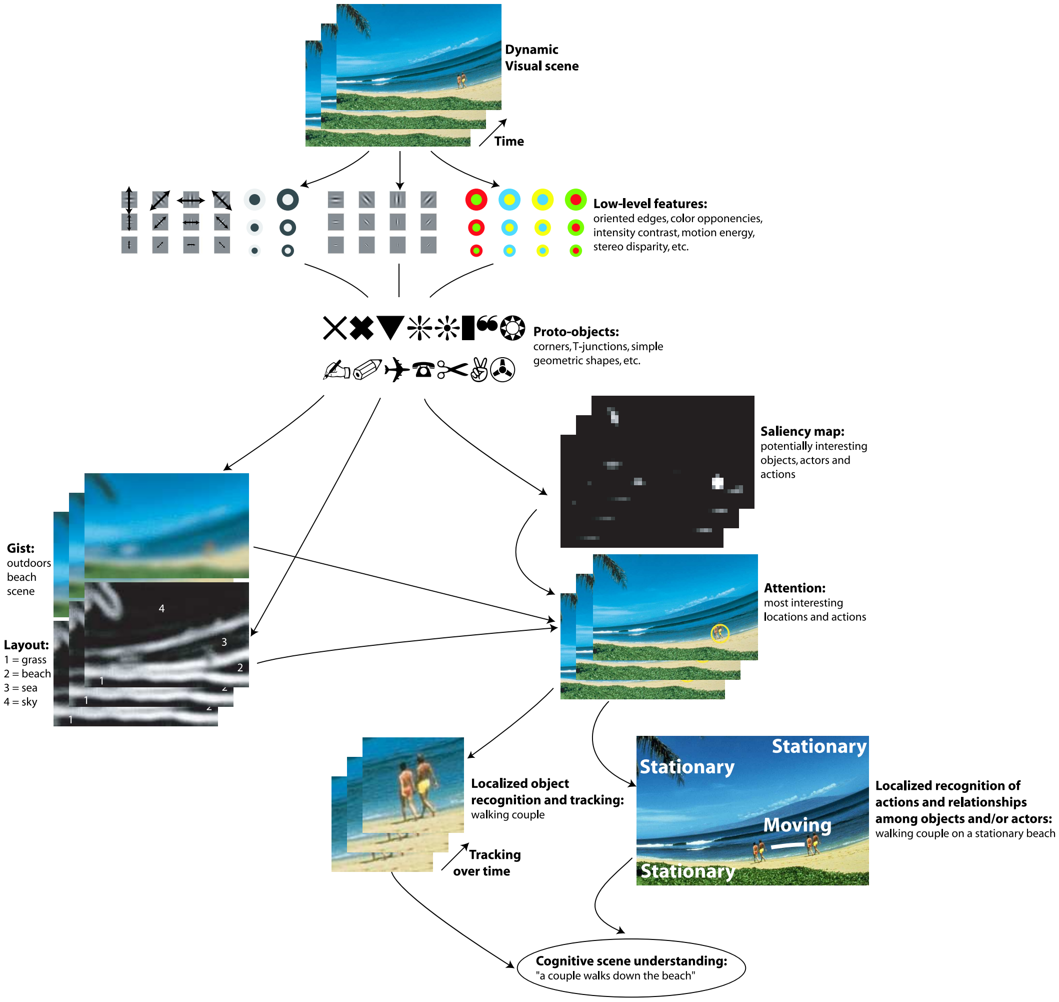
Dynamic
Visual scene
Time
Low-level features:
oriented edges, color opponencies, intensity contrast, motion energy, stereo disparity, etc.
Proto-objects:
corners, T-junctions, simple geometric shapes, etc.
Saliency map:
potentially interesting objects, actors and actions
Gist:
outdoors beach scene
Layout:
1 = grass
2 = beach
3 = sea
4 = sky
4
3
2
1
Attention:
most interesting locations and actions
Localized object recognition and tracking:
walking couple
Tracking over time
Stationary
Stationary
Moving
Stationary
Localized recognition of actions and relationships among objects and/or actors:
walking couple on a stationary beach
Cognitive scene understanding:
"a couple walks down the beach"

Fig. 2. The beginning of an architecture for complex dynamic visual scene understanding, starting from the visual end. This diagram augments the triadic architecture proposed by Rensink (2000), which identified three key components of visual processing: pre-attentive processing up to a proto-object representation (top), identification of the setting (scene gist and layout; left), and attentional vision including detailed object recognition within the spatially circumscribed focus of attention (right). In this figure, we have extended Rensink's architecture to include a saliency map to guide attention bottom-up towards salient image locations, and action recognition in dynamic scenes. The following figure explores in more details the interplay between the visual processing components depicted here and symbolic decriptions of tasks and visual objects of current behavioral interest.

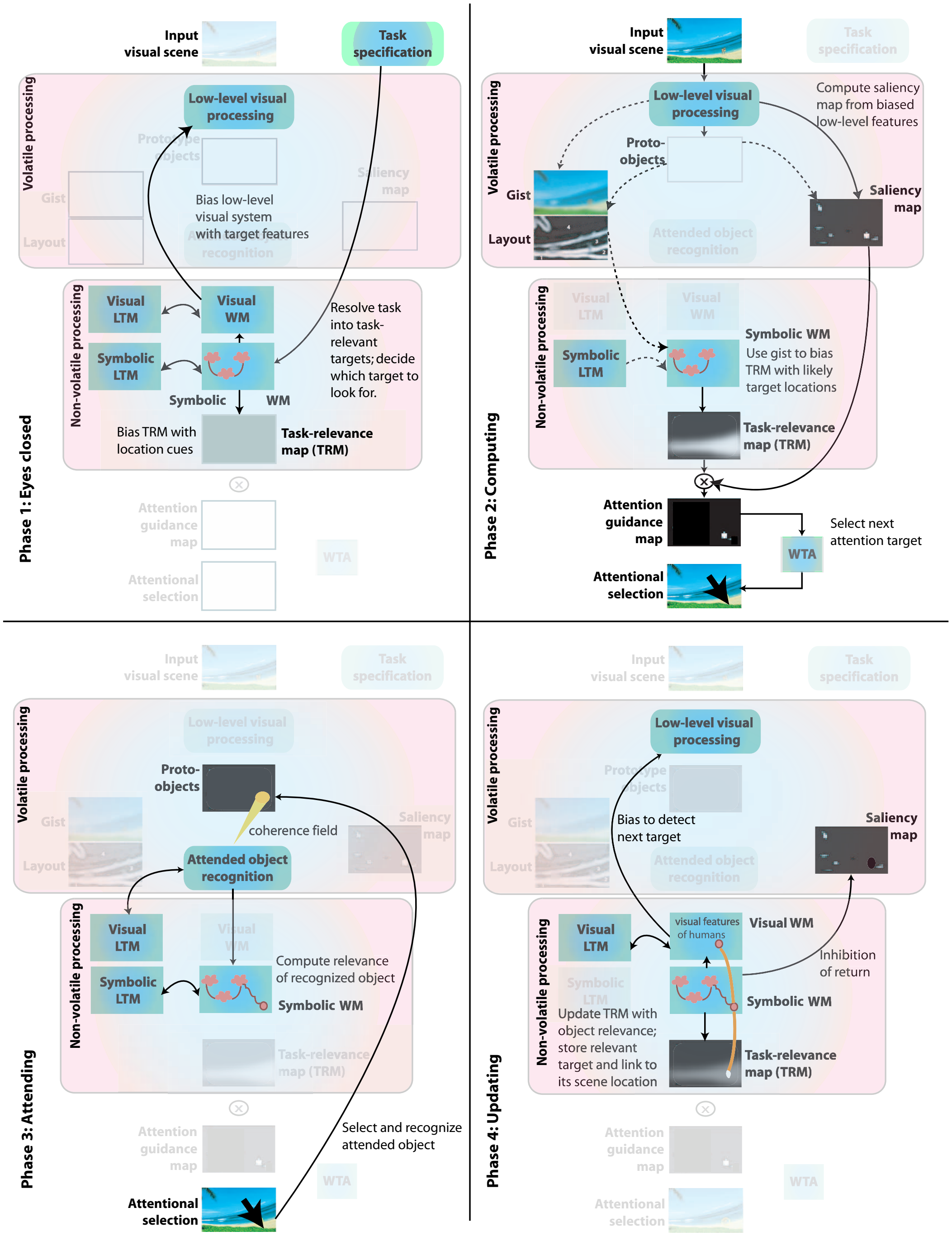

Fig. 3. Phase 1 (top left): Eyes closed, Phase 2 (top right): Computing, Phase 3 (bottom left): Attending, Phase 4 (bottom right): Updating. Please refer to section 3.2 for details about each phase. All four panels represent the same model; however, to enable easy comparison of the different phases, we have highlighted the components that are active in each phase and faded those that are inactive. Dashed lines indicate parts that have not been implemented yet. Following Rensink's (2000) terminology, volatile processing stages refer to those which are under constant flux and regenerate as the input changes.