CHAPTER

105

Attention Architectures for Machine Vision and Mobile Robots

Lucas Paletta, Erich Rome and Hilary Buxton

ABSTRACT

Computer vision systems that are applied for image understanding in real-world environments require the capability to focus operations on task relevant events in an ongoing input stream of visual information. Attentive systems must indirectly provide solutions to characteristic challenges in real-world processing, such as the complexity in input imagery and uncertainty in the acquired information. We address successful methodologies on saliency and feature selection, describe attentive systems with respect to object and scene recognition, and review saccadic interpretation under decision processes. In robotic systems, we understand attention embedded in the context of optimizing sensorimotor behavior and multisensor-based active perception. We present an overview on system architectures that play a crucial role in attentive robots, with emphasis on multimodal information fusion and humanoid robots.

I. INTRODUCTION

Vision systems with the task of operating in real-world environments must tackle challenges that arise from the specific conditions and the uncertainty in the visual information of the captured images. Selective attention is necessary to focus on information that is relevant to a current task and mandatory to make a choice for the processing of appropriate data sources in response to a specific system state in space and time. In machine vision, the development of enabling technologies such as video surveillance systems, miniaturized mobile sensors, and ambient intelligence systems involves the real-time analysis of enormous quantities of data. Knowledge has to be applied concerning what needs to be attended to and when and what to do in a meaningful sequence, in correspondence with visual feedback. Methods on attention and control are mandatory to render computer vision systems more robust. In mobile robots, the embodied actuators may affect perception in an even greater sense, introducing mobility and thereby deciding the observer's viewpoint, linking physical presence with specificity in its sensoric experience. The following sections review attention architectures in machine vision and robotic systems, with an emphasis on research results presented at the International Workshop on Attention and Performance in Computer Vision 2003 (Paletta et al., 2003).

II. ATTENTIVE COMPUTER VISION SYSTEMS

Attention architectures in machine vision are introduced from the view of bottom-up processing in terms of saliency operators. In analogy to top-down paths in human perception (Braun et al., 2001), task-dependent modulation of feature extraction is a relevant issue, in particular regarding the task of object search in real-world scenes. Contextual modulation of processing enables us to take advantage of context cues in order to focus processing on most promising, such as salient image information. In addition, saccadic integration operates on the process chain of contextually related local interpretations of a global object or scene information. Finally, in dynamic vision, the extraction of visual motion plays an outstanding role in providing essential cues and attracting attention.

A. Saliency from Feature Selection

Saliency of a local image area must be defined on the basis of the specific visual information and, accordingly, on the basis of an appropriate feature detector. Attentive processing in computer vision initially used saliency-based models, in which the strength of the response of feature detectors determined candidate locations by matching (Clark and Ferrier, 1988). In a model of purely bottom-up information processing, the Culhane-Tsotsos feature detector (Culhane and Tsotsos, 1992) builds a hierarchy of representations relying on the assumption that input cell values directly reflect how salient a specific location is. A saliency operator based on information measures with respect to spatial locations and scales of objects in an image is provided by Jagersand (1995). It results from the expected information gain from Kullback contrasts between successive resolution lengths.

More elaborated models of attentive stimulus-driven search were proposed by Itti et al. (1998), combining first multiscale image features into a single topographical saliency map. Competition among neurons in this map give rise to a single winning location that corresponds to the next attended target (Sec. IIB). The underlying model is based on the Feature Integration Theory from Treisman and Gelade (1980). A referring implementation that is invariant to similarity transformation is described by Lionelle and Draper (2003). The saliency of local image regions has more recently become relevant in object recognition and wide-base line stereo (Fraundorfer and Bischof, 2003), taking into account the three closely interrelated aspects of saliency, scale, and content. The detector is translation, rotation, and scale invariant.

In current machine attention, bottom-up selection plays an important role in providing early cues in a multistage competetive scheme of attention processing (Navalpakkam and Itti, 2002). Backer and Mertsching (2003) introduced a cascaded computation by selecting a small number of discrete items in a preattentive phase analyzing symmetry, eccentricity, color contrast, and depth, and then applied smiattentive processes of tracking and information accumulation until a single cue of interest could be more efficiently selected.

B. Object and Scene Recognition

Top-down processing has been emphasized in computer vision tasks related to object search. The main focus of research is on how to integrate bottom-up and top-down information to attain an efficient decision on where to focus attention to within the input image. A critical task in computer vision is the recognition of entities of interest in the visual appearance (i.e., object recognition). In real-world environments, object recognition must cope with a high number of degrees of freedom (pose, scale, illumination, etc.), and uncertainty in the visual information plays a major role. Attention supports object search by constraining the dimension of the search space by limiting the visual information to a restricted set of hypotheses bounded to regions.

One central thesis is that attention acts to optimize the search procedure inherent in a solution to vision (Tsotsos et al., 1995). The accordingly proposed selective tuning model emphasized the role of top-down processing in attention mechanisms. Analogously, task-based attention (de Laar et al., 1997) operates in top-down information paths, adjusting a processing layer of intermediate features in a goal-driven manner.

An influential model in attention modeling is to feed bottom-up information into a competitive processing layer, where feature responses are competing according to a winner-takes-all (WTA) strategy for priority in interpretation (Itti et al., 1998; Navalpakkam and Itti, 2002). Ramstrøm and Christensen (2002) proposed a distributed control layer inspired by market principles in which bottom-up feature responses are competing according to a game theoretical model—the competetive equilibrium—with top-down information.

Visual attention in terms of an interactive process has been considered by Lee et al. (2003), in which attentional behavior should be biased with respect to particular objects, spatial locations, and time. A spiking neural network, for example, can bias selective behavior in such a way that it speeds up (facilitates) or slows down (interferes) with the processing of a given visual stimulus, gradually relating "where" and "when" access of information in the network. The network can then manipulate the amount of bottom-up and top-down influence on a search task to investigate the dynamic and modulatory aspects of selective attention.

A face detection task demonstrates the new functionality (Fig. 105.1), allocating the focus of attention to possible target locations. Skin color, facial features, and ellipse-like shape determine a bottom-up map that is correlated with cued features. The net input $\mathrm{net}_j(t)$ of a spiking neural network node j at time t is,

$$\mathrm{net}_j(t) = \alpha \sum_{i=1}^{n} w_{ij}(t) x_i^B + \beta \sum_{i=1,r=1}^{n,m} u_{irj}(t) \left(x_i^{\mathrm{B}}\right)^2 x_r^{\mathrm{T}} \qquad (1)$$

where B and T stand for the bottom-up and top-down inputs, n and m are the dimensions of the bottom-up and top-down inputs, and w and u are the bottom-up

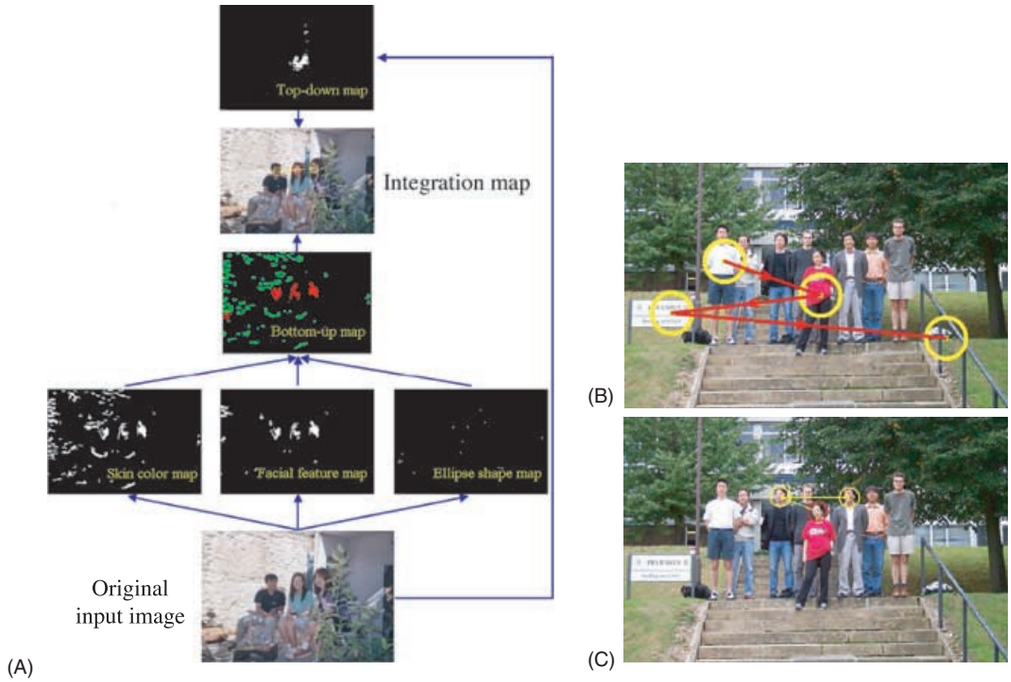

FIGURE 105.1 Example of interactive processing with natural images containing faces (Lee et al., 2003). The model (A) allocates focus of attention to possible target locations. The more task-relevant the target location with respect to the cue, the more likely the location is selected early in the attentional trajectory. Results from bottom-up WTA attention (B) are contrasted to scanpaths from graded interactive top-down attention processing (C).

and multiplicative weights, respectively. The $(x_i^B)^2$ term reinforces correlation between the bottom-up and top-down information stream, depending on whether they are consistent or not. The impact of the top-down information is illustrated by the different behaviors in Fig. 105.1B and C, showing the refined selection of targets not only due to generic visual features (B) but also from the maximization of information between extracted and object-cued features (C).

These computational models of selective attention in the visual search task support the spotlighting of a sequence of regions of interest. Applications are in the detection of objects of interest and scene recognition that is simply based on the occurrence of objects within. Any geometrical or probabilistic relation between serially focused information is not investigated here. However, context analysis would be mandatory for any more extended analysis of the scene or part-based object recognition.

C. Contextual Cueing and Attention Strategies

In computer vision, we face the challenging task of detecting objects of interest in outdoor environments. Changing illumination, different weather conditions, and noise in the imaging process are the most important issues that require a truly robust detection system. Recent work on real-time video interpretation therefore considers attentional mechanisms (Navalpakkam and Itti, 2002) and cascaded systems (Viola and Jones, 2001) to coarsely analyze the complete video frame in a first step, reject irrelevant hypotheses, and iteratively apply increasingly complex classifiers with appropriate level of detail (Sec. IIA). Attention from context priming (Torralba and Sinha, 2001; Ogris and Paletta, 2003) makes sense out of globally defined environmental features to set priors on object-related observable variables to obtain spatial pointers to regions of interest and therefore significantly improves the quality of service in real-time interpretation. Cascaded processing serves as fundamental methodology in sequential saccadic interpretation, giving rise to improved analysis from any updating evidence.

1. Context Priming

Investigations on the binding between scene recognition and object localization made in experimental psychology have produced clear evidence that highly local features play an important role into facilitating

detection from predictive schemes (Hollingworth and Henderson, 2002). In particular, the visual system infers knowledge about stimuli occurring in certain locations, leading to expectancies regarding the most probable target in the different locations (location-specific target expectancies). Related work on scene recognition concerns the contextual mapping from objects to objects (Rimey, 1993) and from global scene features to object hypotheses (Torralba and Sinha, 2001) with respect to a static environment.

The extraction of scene landmarks has recently been applied for priming tasks in attentive object detection (Ogris and Paletta, 2003). Landmarks can be detected more robustly and rapidly than arbitrary objects. The use of landmark configurations inherently includes local spatial context and thus becomes more locally discriminative and predictive than using global features (Fig. 105.2A). Landmark configurations are then mapped to visual object events in local proximity. Defining the mapping in terms of a probabilistic estimation of direction β,

$$p(\mathbf{x}_i | l_j, o_k, \beta(l_j, o_k)) \propto N_{l_i}(\mu_\beta, \sigma_\beta) \quad (2)$$

where image pixel $\mathbf{x}_i$ provides confidence p for being on a line to landmark l_i, and $\lambda_{i,t} \in l_i$ is an associated landmark appearance sampled at time t. This confidence is exponentially decreasing—according to a Gaussian $N(\cdot)$—with increasing angle β from the reference line μ_β (from landmark l_i to object o_k). The approach integrates, then, those landmarks l_j that have been consecutively visited in an observation sequence and been selected as estimators for the next object location, by

$$p(\mathbf{x}_i | l_1, \cdots | l_j, \cdots, o_k, \beta(l_j, o_k)) = \prod_{j=1}^{J} p(\mathbf{x}_i | l_j, o_k, \beta(l_j, o_k)) \quad (3)$$

Figure 105.2B illustrates the result of a recursive estimation of attention from predictions of individual landmarks. Experimental results on frames of a video scene demonstrate an average prediction error of ≈3° in the attention direction and a hit rate of ≈94% (Ogris and Paletta, 2003).

2. *Saccadic Information Integration*

The framework of active and purposive vision laid the conceptual basis to understand computer vision tasks from the perspective of acquiring and optimizing sequential decisions in goal-driven tasks. In this context, Bandera et al. (1996) introduced reinforcement learning to improve the performance of foveal visual attention for the task of model-based object recognition. Recent attempts to model Markovian decision processes for automatic gaze control in face and scene perception (Henderson et al., 2001) and attentive viewpoint control (Paletta and Pinz, 2000) demonstrate the potential for automatic saccade and viewpoint control for the interpretation of objects and scenes.

III. ATTENTION IN ROBOTIC SYSTEMS

The understanding of how to design intelligent robotic systems has seen a paradigm shift during the last decade. Robot architectures are no longer control driven by symbolic artificial intelligence (sense-plan-act) and do not rely on universal perceptual reconstructions of the environment by purposive sensor-driven control (sense-act) and interpretation schemes. Instead, with reference to situatedness, embodiment, and context-relatedness of task performance, visual perception has been relocated as a functional basis for behavior-based control (Arkin, 1998). Attention has been playing an increasingly important role in providing solutions to the control of a. growing stream of sensory input. Mobile robots are often 'thrown' into a complex environment where they have to apply their knowledge and find out about what

(a) (b)

FIGURE 105.2 Spatial attention from local context (Ogris and Paletta, 2003). (A) Landmark configurations are discriminated locally to provide (B) attentive pointers to nearby object locations.

needs to be attended to and when and what to do in correspondence with visual feedback.

A. Autonomous Multisensor Robots

Common tasks in the control of autonomous mobile robots are collision avoidance, navigation, and tracking and manipulation of objects (Coelho et al., 2001). In order to execute these tasks correctly, the robot needs to detect objects and free space in its environment quickly and reliably, emphasizing the role of machine attention to find points of interest.

Multisensor data provide new challenges on selective attention to autonomous and mobile robot systems. In particular, humanoid robots are involved in a multitude of sensory inputs, such as from image feature, motion, and audio signal streams (Vijayakumar et al., 2001). Vijayakumar et al. (2001) applied a WTA network for saliency computation according to Itti et al.'s (1998) model on a 30 degrees of freedom (DOF) humanoid with pan and tilt peripheral and foveal vision, demonstrating motion and person detection in video-rate attention control.

Objects usually have range discontinuities at their borders that can help to detect them. Models comprising depth as a feature typically use stereo vision to compute it, which is computationally expensive, and only a fraction of the image pixels contribute to the computed 3D point clouds. Frintrop et al. (2003) describe attention from a multimodal 3D laser scanner. The attention system has a similar structure as the Neuromorphic Vision Toolkit (Itti et al., 1998), but uses only intensity and orientation as features. Depth information is intensity coded, and the system is capable of simultaneously processing both remission and depth value images, generating a single saliency map for both. Figure 105.3 illustrates how focus of attention from reflectance and range images contributes to classification and detection using fusion of the individual sensor maps.

Cognitive aspects in robot attention were implemented by Dickinson et al. (1997). Here, an active object recognition strategy combined the use of an attention mechanism for viewpoint selection to disambiguate recovered object features. The attention mechanism consisted of a probabilistic search through a hierarchy of predicted feature observations. Paletta and Rome (2000) describe a robotic system that uses task-specific knowledge in order to direct attention to certain regions of interest. Object detection is performed by an appearance-based classification of the regions of interest (ROIs). Because the domain objects (sewer pipe inlets) are often very similar, the system learns an information-fusion approach to disambiguate objects. The classification is performed iteratively with images taken from different viewpoints,

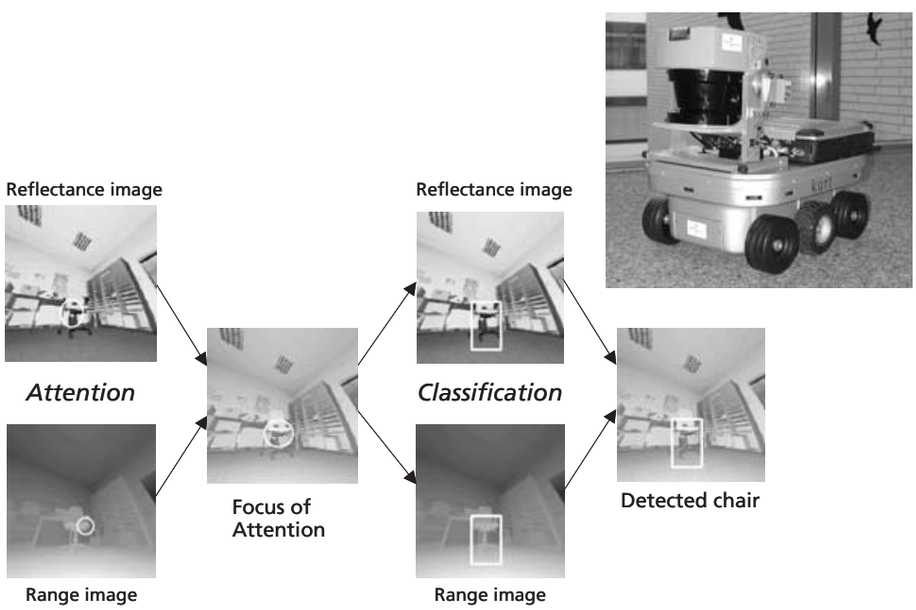

FIGURE 105.3 The custom 3D range finder mounted on top of the mobile robot KURT2 (upper right). Office scene imaged with the 3D scanner in range- and remission-value modes. The system, described in (Frintrop et al., 2003, 2004), performs attention-based ROI selection and successive classification within the ROI.

until the confidence is sufficient for making a decision. The robot KISMET (Breazeal and Scasselatti, 1999) integrated perception, attention, drives, emotions, behavior arbitration, and expressive acts in order to interact socially with humans. The attention system was based on Wolfe's model (Wolfe, 2000) and integrated perceptions with habituation effects and influences from the robot's motivational and behavioral state to create a context-dependent attention activation map. Finally, joint attention is highly relevant in the supervised learning stages of human development, specifically for the capability to interact with the environment. Nagai et al. (2003) demonstrated the finding of and attending to a salient object in the robot's view and a sensorimotor coordination when the visual attention succeeds. Based on these mechanisms, the robot learns the sensorimotor coordination when the robot can watch the salient object by shifting its gaze direction from the caregiver's face to the object.

B. Visuomotor Attention

Oculomotor control is not just important for spatial but also for sequential attention. Because eye movements must be executed in a sequential manner, it is crucial to focus visual attention at the right time on the right targets so that subsequent information processes, in particular motor planning and execution, receive relevant information sufficiently fast to update ongoing processes. From this viewpoint, oculomotor control may, for example, be a crucial constraint on how movement of other body parts are planned. Miyashita et al. (1996) showed that anticipatory saccades in sequential procedural learning in monkeys are tightly coupled to the limb-motor system. Similend, Shibata et al. (2001) developed a biomimetic gaze stabilization that is mused for attentional mechanisms in a humanoid robot (Vijayakumar et al., 2001).

Actually, robot visuomotor attention is in its early stages. Coelho et al. (2001) demonstrated the importance of learning of visual features that the observer has to attend to for more efficient grasping. They constructed constellations of visual features to predict relative hand-object postures that lead reliably to haptic utility.

IV. CONCLUSION

Attention is an ongoing research topic in computer vision that gains increasingly in importance under the guidance of cognitive vision system research. The classic application fields, such as video analysis, surveillance, and robotics, and not the only ones to depend on attention methodologies. Emerging technologies, such as mobile vision, wearable computing, and service robotics, are by nature limited in computational resources but should operate in everyday's complex contexts and work environments. These upcoming challenges to both computational and semantic complexity provide the motivation to reinforce research on attentional mechanisms in machine vision and mobile robot perception.

References

Arkin, R. C. (1998). "Behavior-Based Robotics." MIT Press, Cambridge, MA.

Backer, G., and Mertsching, B. (2003). Two selection stages provide efficient object-based attentional control for dynamic vision. *In* "Proceedings of the Workshop on Attention and Performance in Computer Vision," pp. 9–16. Joanneum Research, Graz, Austria.

Bandera, C., Vice, F. J., Bravo, J. M., Harmon, M. E., and Baird, L. C. (1996). Residual Q-learning applied to visual attention. *In* "Proceedings of the International Conference on Machine Learning," (Lorenza Saitta Ed.) pp. 20–27. Morgan Kaufmann Publishers, San Francisco, CA: Bari, Italy.

Braun, J., Koch, C., Lee, D. K., and Itti, L. (2001). Perceptual consequences of multilevel selection. *In* "Visual Attention and Cortical Circuits" (J. Braun, C. Koch, and J. L. Davis, Eds.), pp. 215–242. MIT Press, Cambridge, MA.

Breazeal, S., and Scasselatti, B. (1999). A context-dependent attention system for a social robot. *In* "Proceedings of the International Joint Conference on Artificial Intelligence," (Thomas Dean, Ed.) Vol. 2, pp. 1146–1153. Morgan Kaufmann, San Francisco, CA: Stockholm, Sweden.

Clark, J., and Ferrier, N. (1988). Modal control of an attentive vision system. *In* "Proceedings of the International Conference on Computer Vision", pp. 514–523. Tampo, Florida.

Coelho, J., Piater, J., and Grupen, R. (2001). Visual perceptual categories for reaching and grasping with a humanoid robot. *Robotics Autonomous Syst*. **37**, 195–219. Santa Maighenta Ligure, Italy, Springer.

Culhane, S., and Tsotsos, J. (1992). An attentional prototype for early vision. *In* "Proceedings of the European Conference on Computer Vision," pp. 551–560. Sauta Maighenta Ligure, Italy, Springer.

de Laar, P. V., Heskes, T., and Gielen, C. (1997). Task-dependent learning of attention. *Neural Networks* **10**, 981–992.

Dickinson, S. J., Christensen, H. I., Tsotsos, J. K., and Olofsson, G. (1997). Active object recognition integrating attention and viewpoint control. *Comput. Vis. Image Understanding*, **67**, 239–260.

Fraundorfer, F., and Bischof, H. (2003). Utilizing saliency operators for image matching. *In* "Proceedings of the Workshop on Attention and Performance in Computer Vision", pp. 17–24. Joanneum Research, Graz, Austria.

Frintrop, S., Nüchter, A., and Surmann, H. (2004). Visual attention for object recognition in spatial 3D data. *In* "Proceedings of the 2nd Workshop on Attention and Performance in Computational Vision," pp. 75–82. Joanneum Research, Graz, Austria.

Frintrop, S., Rome, E., Nüchter, A., and Surmann, H. (2003). An attentive, multi-modal laser "eye". *In* "Proceedings of the International Conference on Computer Vision Systems," pp. 202–211. Joanneum Research, Graz, Austria.

Henderson, J. M., Falk, R., Minut, S., Dyer, F. C., and Mahadevan, S. (2001). Gaze control for face learning and recognition by humans and machines. *In* "From Fragments to Objects: Segmentation and Grouping in Vision" (T. Shipley and P. Kellman, Eds.), pp. 463–481. Elsevier Science, Oxford.

Hollingworth, A., and Henderson, J. (2002). Accurate visual memory for previously attended objects in natural scenes. *J. Expe. Psychol. Hum. Perception Performance* **28**, 113–136.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis Machine Intell.* **20**, 1254–1259.

Jagersand, M. (1995). Saliency maps and attention selection scale and spatial coordinates: An information theoretic approach. *In* "Proceedings of the International Conference on Computer Vision," pp. 195–202. Boston, MA.

Lee, K., Buxton, H., and Feng, J. (2003). Selective attention for cue-guided search using a spiking neural network. *In* "Proceedings of the Workshop on Attention and Performance in Computer Vision," pp. 55–63. Graz, Austria.

Lionelle, A., and Draper, B. (2003). Evaluation of selective attention under similarity transforms. In "Proceedings of the Workshop on Attention and Performance in Computer Vision," pp. 31–38. Graz, Austria.

Miyashita, K., Kato, R., Miyauchi, S., and Hikosaka, O. (1996). Anticipatory saccades in sequential procedural learning monkeys. *J. Neurophysiol*. **76**, 1361–1365.

Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). Joint attention emerges through bootstrap learning. *In* "Proceedings of the International Conference on Intelligent Robotic Systems," pp. 168–173. Las Reges, Nevarda.

Navalpakkam, V., and Itti, L. (2002). A goal oriented attention guidance model. *In* "Proceedings of the International Workshop on Biologically Motivated Computer Vision," Heinrila H. Bulthoff, Seong-Whau Lee, Tomah A. Poggio, Clairliau Wallranean (Eds.) pp. 453–461. Tübingen, Germany, Springer.

Ogris, G., and Paletta, L. (2003). Contextual cueing for spatial attention in object detection from video. *In* "Proceedings of the Workshop on Attention and Performance in Computer Vision," pp. 64–72. Joanneum Research, Graz, Austria.

Paletta, L., Humphreys, G., and Fisher, R. (Eds.) (2003). "Proceedings of the International Workshop on Attention and Performance in Computer Vision." Joanneum Research, Graz, Austria.

Paletta, L., and Pinz, A. (2000). Active object recognition by view integration and reinforcement learning. *Robotics Autonomous Syst.* **31**, 71–86.

Paletta, L., and Rome, F. (2000). Learning fusion strategies for visual object detection. *In* "Proceedings of the International Conference on Intelligent Robots and Systems," pp. 1446–1452. Takamatsu, Japan.

Ramstrøm, O., and Christensen, H. (2002). Visual attention using game theory. *In* "Proceedings of the International Workshop on Biologically Motivated Computer Vision," Heinrila H. Bulthoff, Seong-Whau Lee, Tomah A. Poggio, Clairliau Wallranean (Eds.) pp. 462–471.

Rimey, R. D. (1993). Control of selective perception using bayes nets and decision theory. Technical Report TR468. Computer Science Department, University of Rochester.

Shibata, T., Vijayakumar, S., Conradt, J., and Schaal, S. (2001). Biomimetic oculomotor control. *Adaptive Behav*. **9**, 189–207.

Torralba, A., and Sinha, P. (2001). Statistical context priming for object detection. *In* "Proceedings of the International Conference on Computer Vision," pp. 763–770. Vaucourer, Canada. IEEE.

Treisman, A., and Gelade, G. (1980). A feature integration theory of attention. *Cogn. Psychol*. **12**, 97–136.

Tsotsos, J. K., Culhane, S. M., Wai, W, Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modelling visual attention via selective tuning. *Artificial Intell*. **78**, 507–545.

Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt visual attention for a humanoid robot. *In* "Proceedings of the International Conference on Intelligent Robots and Systems," Maui, Hi. Vol. 4, pp. 2332–2337.

Viola, P., and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In* "Proceedings of the Conference on Computer Vision and Pattern Recognition," Kauai, HI. pp. 511–518. IEEE.

Wolfe, J. M. (2000). Visual attention. *In* "Seeing" (K. K. DeValois, Ed.), pp. 335–386. Academic Press, San Diego, CA.