Evolution of Attention Mechanisms for Early Visual Processing

Thomas Müller and Alois Knoll

Robotics and Embedded Systems
Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany

ABSTRACT

Early visual processing as a method to speed up computations on visual input data has long been discussed in the computer vision community. The general target of a such approaches is to filter non-relevant information from the costly higher-level visual processing algorithms. By insertion of this additional filter layer the overall approach can be speeded up without actually changing the visual processing methodology.

Being inspired by the layered architecture of the human visual processing apparatus, several approaches for early visual processing have been recently proposed. Most promising in this field is the extraction of a saliency map to determine regions of current attention in the visual field. Such saliency can be computed in a bottom-up manner, i.e. the theory claims that static regions of attention emerge from a certain color footprint, and dynamic regions of attention emerge from connected blobs of textures moving in a uniform way in the visual field. Top-down saliency effects are either unconscious through inherent mechanisms like inhibition-of-return, i.e. within a period of time the attention level paid to a certain region automatically decreases if the properties of that region do not change, or volitional through cognitive feedback, e.g. if an object moves consistently in the visual field. These bottom-up and top-down saliency effects have been implemented and evaluated in a previous computer vision system for the project JAST.

In this paper an extension applying evolutionary processes is proposed. The prior vision system utilized multiple threads to analyze the regions of attention delivered from the early processing mechanism. Here, in addition, multiple saliency units are used to produce these regions of attention. All of these saliency units have different parameter-sets. The idea is to let the population of saliency units create regions of attention, then evaluate the results with cognitive feedback and finally apply the genetic mechanism: mutation and cloning of the best performers and extinction of the worst performers considering computation of regions of attention. A fitness function can be derived by evaluating, whether relevant objects are found in the regions created.

It can be seen from various experiments, that the approach significantly speeds up visual processing, especially regarding robust ealtime object recognition, compared to an approach not using saliency based preprocessing. Furthermore, the evolutionary algorithm improves the overall performance of the preprocessing system in terms of quality, as the system automatically and autonomously tunes the saliency parameters. The computational overhead produced by periodical clone/delete/mutation operations can be handled well within the realtime constraints of the experimental computer vision system. Nevertheless, limitations apply whenever the visual field does not contain any significant saliency information for some time, but the population still tries to tune the parameters - overfitting avoids generalization in this case and the evolutionary process may be reset by manual intervention.

Keywords: Early Visual Processing, Attention Mechanisms, Evolutionary Algorithm

Further author information: (Send correspondence to T. Müller)
E-mail: muelleth@cs.tum.edu, Telephone: +49 (0)89 289 25761

1. INTRODUCTION

Considering robot perception systems the scientific community investigates multiple approaches using diverse kinds of sensory devices. Still, thinking of the human, the visual sensors remain the most informative devices for enviroment perception. This fact, in the opinion of the authors is still under-represented in real world robotic applications. The reasons for this are manifold, first, information from visual sensors is often noisy, second, even if perfect "laboratory" data can be provided, the analysis of the input data is computationally expensive and thus time consuming. Furthermore, perfectly tuned vision systems are often inflexible and may not be reparametrized in an online manner.

Within this paper, a robot system for human-robot interaction is used as a demonstrator platform. Here, it is necessary to perform face detection and tracking quickly and efficiently, as the robot has to react on the users actions instantly. We therefore propose a novel approach for image preprocessing in order to speed up the expensive recognition system. Furthermore, the approach has to provide for an online reparametrization facility, as for example the lighting conditions, may change during an experiment. Also, we intended to develop a general approach, which is not inherently dependent on the task of face recognition itself, but rather general enough to cope with a variety of object recognition tasks.

In brief, the proposed solution comprises early visual processing mechanisms similar to the human visual apparatus: *static saliency* for detection of regions with a certain color footprint within a cluttered environment, *dynamic saliency* for identification of regions of non-uniform object movements in the visual field, and a strategy for back-projecting *cognitive feedback* in order to compute reinforcement / inhibition on the computed bottom-up effects (previous work elaborated on the implementation of these effects for the task of object recognition[1,2]). All of these attention mechanisms apply parameters and hence these parameters have to be tuned somehow. While in real biological systems these parameters are partly evolved over generations of individuals (phylogenetic training) or adjusted during lifetime (ontogenetic training), the proposed system strives to produce suitable parameters for certain tasks after some time induced by ground-truth feedback by evolution of the early processing units in an online manner. After some time, i.e., after the preprocessing stabilizes, the expensive ground-truth detection - which clearly is the most robust and accurate method - can be switched to a cyclic mode with large delays and the processing relies primarily on a detection by saliency and tracking closed loop, with only occasional ground-truth updates.

2. EXPERIMENTAL SETUP

The target setup for our approach is the human robot interaction system which operates as the main demonstrator platform for the JAST "**J**oint-**A**ction **S**cience and **T**echnology" project.

The overall goal of the JAST project is to investigate the cognitive and communicative aspects of jointly-acting agents, both human and artificial. The human-robot dialog system being built as part of the project[3,4] is designed as a platform to integrate the projects empirical findings on cognition and dialog with research on autonomous robots, by supporting symmetrical, multimodal human-robot collaboration on a joint construction task*.

The robot (Figure 1) consists of a pair of mechanical arms with grippers and an animatronic talking head. The input channels consist of speech recognition, object recognition, gesture recognition, face tracking, and robot sensors; the outputs include synthesized speech, emotional expressions, head motions, and robot actions. The user and the robot work together to assemble a wooden construction toy on a common work area, coordinating their actions through speech, gestures, and facial motions.

*http://www.youtube.com/watch?v=yXZXnAQ15LI

Figure 1. The JAST human-robot dialog system.

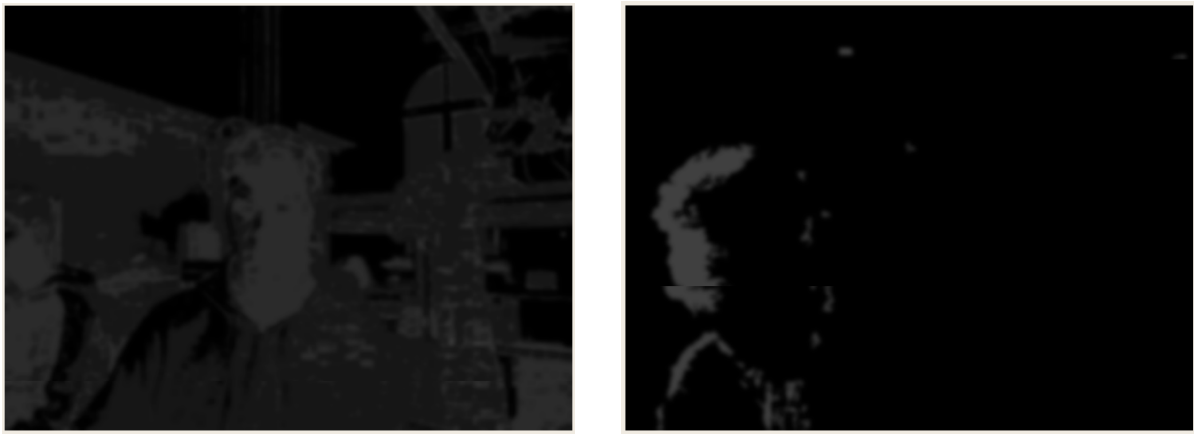
3. DESIGN OF SALIENCY UNITS

The robot system uses a camera mounted inside the nose of the robot's head in order to keep the head turned towards the human interaction partner(s). The recognition process has to be fast enough to make interactions intuitive for the cooperating human. While the ground-truth recognition process itself is based on cascaded Haar classifiers[5] which are provided by the open source computer vision library *OpenCV*,[6] this approach is computationally expensive. The solution, i.e., utilizing fast early processing to constrain detection only on relevant regions of the visual input, is based on two principles of attention:

- **Bottom-Up**
  Attention attraction in a bottom-up, scene-dependent[7] manner is defined as an effect originating directly from the visual input sensors. Two effects can be distinguished: attending regions that seem significantly different from the what is thought to be background (regions with a color which is of certain interest); or, regions, i.e. a patch of input pixels, that move in a uniform way within the visual field.

- **Top-Down**
  Top-down[8] or task-dependent[9] attention effects consider feedback of higher levels of cognition. This comprises projection of regions into the visual field that are anticipated to be interesting from task-aware cognitive layers (reinforcing feedback); or, on the contrary, not of interest for the moment (inhibiting feedback).

The saliency units of the proposed system implement both bottom-up and top-down effects to a certain degree. The following methodology is integrated into each of the saliency units.



(a) Saliency from a color footprint. (b) Saliency from scene dynamics.

Figure 2. Bottom-up attention effects.

3.1 Bottom-Up Attention Attractors

Considering bottom-up effects, each unit implements a thresholding approach[10] for static saliency. As colored input images in the HSV color space are used, this method can be parametrized with six values. Evaluating salient regions here means extracting regions by checking each pixel of the input image. As long as a pixel p satisfies the following condition, it is considered belonging to a salient region

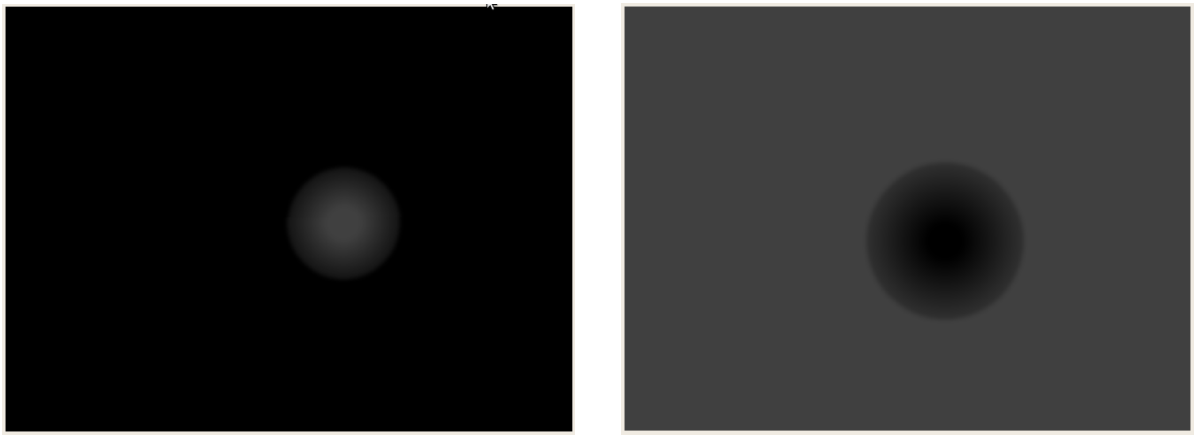
$$\forall c \in h, s, v : t_c - \delta_c \leq p_c \leq t_c + \delta_c \tag{1}$$

where t_c is the threshold value for a certain channel and δ_c the tolerance value for this channel. Salient pixels are colored brighter, while others are colored darker. To suppress noise the result image is smoothed using a Gaussian kernel after applying several dilation / erosion operations. The result is a grayscale image of saliency (with three different levels of gray from the sum of values from each color channel).

Considering dynamic moving blobs of pixels in the input image, i.e., the second bottom-up attention attractor, a mixture-of-gaussian (MOG) approach is used. The mixture of gaussians is a widely used approach for background-subtraction.[11,12] It approximates the background (and its dynamics) using several gaussian distribution models. This property is exploited, as we consider regions moving atypically in the visual field to attract attention. The MOG models static background and repetitive movements very well, but it seizes to model unexpected motions, which is precisely what we are trying to find regarding saliency from dynamics. Parameters of this method are amongst others window-size and forget-time. To avoid overfitting, we choose the window size of the MOG to be rather small, leaving the forget-time adjustable by default. The result of the application is a binary image (see Figure 2).

3.2 Saliency from Top-Down Feedback

Top-down feedback in the proposed system comprises projection of the output from a camshift tracker into the preprocessing layer. With respect to the experimental application this tracker tries to follow faces in the image. camshift, "Continuously Adaptive Mean Shift",[13] is an extension of the well-known mean-shift algorithm.[14] It is widely applied to visual tracking using color features. Once initialized (with a region of interest) the mean direction of movement is calculated from one frame to the next. It is able to model both motions in x- and y-directions and scaling, as well as changes of the target orientation. The output of the tracker is a 2D box with orientation information. This information is back-projected into the saliency layer by simply drawing decreasing values in an inside-out fashion from



(a) Attention attracting cognitive feedback. (b) Inhibitory mask.

Figure 3. Top-down attention effects.

the center of the box outwards. The result of this again is a grayscale image with dark background and bright circles (see Figure 3).

The same output, seen in a temporal context, can be used to model inhibition of return (IOR). Inhibition of return states an effect discovered by Posner and Cohen.[15,16,17] It essentially claims, that regions, that have been attended in the close past (between 500 and 3000 milliseconds), receive an inhibitory feedback signal. In our scenario, tracked regions from within this period in the past, i.e., pixels values of former regions of interest, are decreased by a certain value. The result of this process is a negated image, it assignes bright values to any pixel, and darker values within the inhibited regions (see Figure 3).

The parameters for top-down effects are the initial (center) brightness value of the circular tracker output for attention attraction, and the darkness value of past tracker outputs (again the center value) of circular inhibited regions.

3.3 Combination of Saliency Effects

The above methods are all computed seperately for each cycle of preprocessing. The weighted sum of the output images (grayscale or binary) states the final saliency mask. An example is shown in Figure 4. In the current version of the system hard-coded, i.e. fixed equal values, for the combination weights are used. Certainly it would be interesting to investigate, whether it is possible to adjust these weight in the evolutionary process described below (see also discussion).

4. EVOLUTION OF SALIENCY UNITS

The exciting part of this paper comes next. As stated in the respective paragraphs of the last section, each of the saliency units cyclically utilizes a set of parameters to compute the saliency mask. These parameters are initialized randomly within all individuals of the population (the set of saliency units). The parameters are then adjusted online (or when the training process is started – this is just the choice of the user). In general the ground-truth data for has to be provided somehow to the evolution unit. An "evolution unit" receives inputs from the saliency units (the saliency masks), compares their computational result with the ground-truth data and assignes a fitness value to the unit. According to the general system of evolutionary algorithms, the units performing best are cloned. Their parameter

Figure 4. Combined saliency mask.

sets P are then mutated according to a learning rate α[†]

$$\forall p \in P : p' = p + \alpha \left(rand(p)) - 0.5|range(p)| \right), \tag{2}$$

where $range(p)$ is the range of values the parameter can take, e.g., $p \in [0, 255]$ for color values, and $rand(p)$ returns a random value within the range of a parameter. Finally, the units with low fitness values are deleted and replaced by the mutated clones of fitter ones. The evolution unit cyclically performs this operation of cloning, mutation and extinction of units.

Informally, we specify the fitness of a saliency unit as follows: regarding the computed grayscale saliency mask for a certain task,

- the higher values of the pixels within the ground-truth area, and
- the lower the pixel-values outside the ground-truth area are,

the fitter the unit is considered to be. Thus, for a single frame, the fitness value *inside* the ground-truth area can be written as

$$f_i = \frac{1}{255} m_i \quad \text{with} \quad m_i = \frac{1}{|x_1 - x_0||y_1 - y_0|} \sum_{x,y \in a} \text{img}(x, y), \tag{3}$$

where m_i is the mean pixel value in the area $a = [(x_0, y_0), (x_1, y_1)]$, while the fitness value *outside* the ground-truth area is

$$f_o = \frac{255 - m_o}{255} \quad \text{with} \quad m_o = \frac{1}{width \cdot height - |x_1 - x_0||y_1 - y_0|} \sum_{x,y \notin a} \text{img}(x, y). \tag{4}$$

The average fitness f of a saliency unit over a set of sample frames S is then

$$f = \frac{1}{|S|} \sum_{s \in S} (\alpha f_i^s + \beta f_o^s), \tag{5}$$

where α and β are the weights for inside/outside fitness values with $\alpha + \beta = 1$ and $\alpha, \beta \in [0, 1]$[‡].

[†]The face tracking application shows good results for $\alpha \sim 0.05$.

[‡]The system shows good results for the facetracking application for $\alpha = 0.75$ and $\beta = 0.25$.

5. RESULTS AND DISUCSSION

As mentioned before, the goal of our experimental application is fast face detection and recognition for a human robot interaction scenario. The ground-truth data for our experimental setup is obtained from a robust face-detector using cascaded Haar classifiers[5] (see left-top in Figure 5).

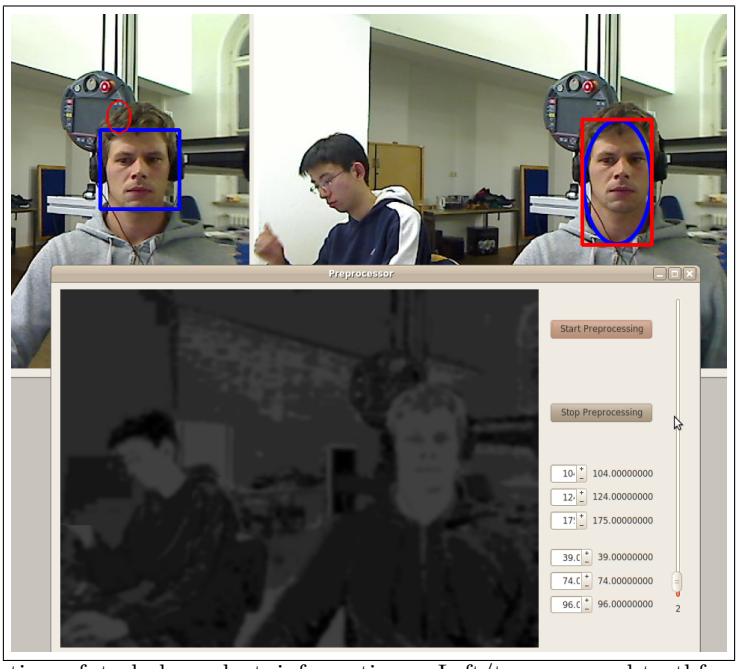


Figure 5. Backprojection of task-dependent information. Left/top: ground-truthface-detector, right/top: camshift tracker result, bottom: mask of "fittest"saliency unit (here only static bottom-up mask).

The detected faces, together with the output of the saliency units, are then passed to the evolution unit as input data. The evolutionary algorithm only uses cloning / mutation of the fittest, currently without applying crossover strategies. The following results correspond to running the system in online-training mode, which means, the *Viola-Jones* face-detector is running continuously along with the adaption of preprocessing units. The tracker for cognitive feedback is (re-)initialized by the face-detector cyclically, so it is based on ground-truth data.

We see that, running the system for several minutes, the evolved saliency mask adapts to the task of face detection (see Figure 6) without manual intervention. The figure shows the output of the saliency unit with the best fitness in the population. The population has a size of fifty individuals, which is a compromise (in fact the upper limit) regarding computational resources and evolutionary performance.

From the results of several runs of about five minutes each, within the target application of optimization for face detection and tracking, we find that the system seems to adjust the inhibitory value towards insignificance. Furthermore, as the background is rather stable, also the effect of dynamic saliency seems not very important – the parameters seem random by inspection of the result set P. This is not quite what we expected, as originally, we hoped to find greater impact of these effects on the final result. But in general, this finding could be correspondent to the specification of the task (in particular to the ground-truth data used), not to the design of the evolutionary system itself.

On the other hand, the static color footprint seems to evolve quickly towards a stable skin-color detector. The results for the color footprint parameters are comparable for all runs, where the only difference was the time the system needed to evolve these parameters. Also, the tracker feedback seems

(a) Some input frame.

(b) Randomly initialized.

(c) After one minute.

(d) After two minutes.

(e) Stable result (insignificant improvement).

(f) Combined with top-down feedback.

Figure 6. Evolution of a visual saliency mask (only fittest unit at a time is shown). (b) to (e) only show the bottom-up effect – the user interface provides buttons to toggle display of each of the effects.

a useful indicator of saliency, as even though the tracker is not very stable, the weight (center value of the bright circle) was evolved to a value above 200 (in a range of $[0, 255]$) for all experimental runs.

Switching from training to productive phase, i.e., suspending all but the fittest saliency unit, and increasing the delay for ground-truth updates to 30 seconds, we see, that a simple blob detector (region-growing and connected components with a minimum region size) effectively performs the same task as ground-truth detection for (re-)initialization of the tracker in much less time. Continuously running the ground-truth detector on our hardware we experience a performance of $\sim 3\,\mathrm{Hz}$, while using the saliency unit results in a performance of $\sim 20\,\mathrm{Hz}$.[§]

6. CONCLUSION

In this paper we have demonstrated a novel system for efficient design of bio-inspired early processing units within a visual perception system for a robotic target setup (face detection and tracking). Alongside, the paper shows, how these early processing units can be evoloved towards autonomous parametrization of a saliency mask. This mask can be applied to speed-up the face recognition system of the robot.

The results show, that the system indeed evoloves a saliency mask highlighting regions in the input image stream which are likely to contain the desired target object, a face. Furthermore, we believe the system is in general able to tune towards any kind of object for recognition if sufficient feedback is supplied from higher levels of cognitive architecture. We also strongly believe, although the results do not sufficiently prove this, that inhibitory feedback as well as bottom-up attention attraction from scene dynamics are crucial for a general robotic perception system.

[§]The experimental hardware comprises a Intel(R) Core2 Quad with 2.5 GHz and 3 GB RAM.

As the preliminary results presented in this paper seem promising to us, in the future, we will try to extend the approach and investigate on the following topics. (1) online adaptation of the weighted sum of saliency: the weights for combination of the attention effects are currently hard-coded and each set to 0.25. In the future it would be interesting to find, if the proposed system is able to adjust these weights within the evolutionary process. (2) cross-over: the evolutionary process by now only applies cloning / mutation and deletion of individuals. Certainly it would be interesting, whether the performance improves, if cross-over strategies are implemented in the process of parameter mutation.

REFERENCES

[1] Müller, T. and Knoll, A., "Attention Driven Visual Processing for an Interactive Dialog Robot," in [*Proceedings of the 24th ACM Symposium on Applied Computing (SAC'09)*], (March 2009).

[2] Müller, T. and Knoll, A., "Humanoid Early Visual Processing Using Attention Mechanisms," in [*Proceedings of Cognitive Humanoid Vision Workshop at IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'08)*], (December 2008).

[3] Foster, M. E., By, T., Rickert, M., and Knoll, A., "Humanrobot dialogue for joint construction tasks," in [*Proc. ICMI*], (2007).

[4] Rickert, M., Foster, M. E., Giuliani, M., By, T., Panin, G., and Knoll, A., "Integrating language, vision, and action for human robot dialog systems," in [*Proc. ICMI*], (2007).

[5] Viola, P. and Jones, M., "Robust real-time object detection," in [*IEEE ICCV Workshop on Statistical and Computational Theories of Vision*], (2001).

[6] OpenCV, "Open Source Computer Vision Library (OpenCV)," tech. rep., Intel Corporation (2010).

[7] Itti, L. and Koch, C., "Computational modelling of visual attention," *Nature Reviews Neuroscience* **2**, 194–203 (March 2001).

[8] Li, W., Piëch, V., and Gilbert, C. D., "Perceptual learning and top-down influences in primary visual cortex," *Nature Neuroscience* **7**, 651–657 (May 2004).

[9] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence* **20**, 1254–1259 (November 1998).

[10] Sezgin, M. and Sankur, B., "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging* **13**, 146–165 (Jan. 2004).

[11] Lipton, A., Fujiyoshi, H., and Patil, R., "Moving target classification and tracking from real-time video," in [*Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on*], 8 –14 (Oct. 1998).

[12] Stauffer, C. and Grimson, W., "Learning patterns of activity using real-time tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**, 747 –757 (Aug. 2000).

[13] Bradski, G. R., "Computer vision face tracking for use in a perceptual user interface," *Intel technology Journal* **Q2** (1998).

[14] Comaniciu, D., Ramesh, V., and Meer, P., "Real-time tracking of non-rigid objects using mean shift," in [*Computer Vision and Pattern Recognition*], (2000).

[15] Posner, M. and Cohen, Y., "Components of visual orienting," *Attention and Performance* **10**, 531–556 (1984).

[16] Cohen, J. D., Aston-Jones, G., and Gilzenrat, M. S., "A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration," in [*Cognitive Neuroscience of Attention*], Posner, M. I., ed., ch. Cognitive Models of Attention, 71–90, Guilford Publications (2004).

[17] Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J., "Inhibition of return: Neural basis function," *Cognitive Neuropsychology* **2**, 211–228 (August 1985).