Attention Driven Visual Processing for an Interactive Dialog Robot

Thomas Müller
Technische Universtiät München
Faculty of Informatics
Robotics and Embedded Systems
Boltzmannstr. 3
85748 Garching / München / Germany
muelleth@cs.tum.edu

Alois Knoll
Technische Universtiät München
Faculty of Informatics
Robotics and Embedded Systems
Boltzmannstr. 3
85748 Garching / München / Germany
knoll@cs.tum.edu

ABSTRACT

In this paper we propose an attention-based vision system for the JAST interactive dialog robot. The robotic vision system incorporates three submodules: object recognition, gesture recognition and self recognition.

The performance boost of our biologically inspired vision system is based on two assumptions: first, generally attention is attracted by regions of high intensity or hue gardients as well as scene dynamics (bottom-up attention attraction), and second, attentioninal focus can be directed by higher level modules, whether volitional or not, in an inhibitory or reinforcing way (top-down attention control). The system proposed in this paper is able to utilize these assumptions and organize its computational efforts accordingly.

Integrated into an efficient data management architecture, the vision system is capable of continuously publishing results to the cognitive layer of the robot and thus enables operations in realtime. Furthermore, the modular system structure and the asynchronous communication paradigm allows for efficient integration of additional modules, be it visual or any other sensory input data.

The main contribution of this work is the application of neuroscience findings and biologically plausible theories of attention based visual processing to a real-world robotic setup. Here, our experimental results show tremendous speed-ups using either the bottom-up attention attractors or the principle of top-down attention control as input data filters for further visual analysis, reaching the peak in a combination of the two.

Keywords

Computer Vision; Bioinspired Robot Vision; Active Perception and Vision; Realtime Visual Processing; Autonomous Robots; Gesture, Object and Self Recognition;

SAC'09 March 8-12, 2009, Honolulu, Hawaii, U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03 ...$5.00.

1. INTRODUCTION

Many traditional computer vision system, either for surveillance tasks with fixed cameras, robot vision on mobile platforms or any other kind of visual processing suffer from an enormous computational complexity. The main reason for this is, that these traditional systems utilize *complete analysis approaches*. The basic, biologically inspired idea of the authors (and some more in the scientific community) is to apply an attention-based type of filter on the great amount of visual input data and only perform further analysis on the tiny rest. This residual is what we call the regions of interest (ROIs).

There have been many approaches to computation of salient features in a static image, e.g. Reinagel and Zador [1] shows that high contrast regions seem to attract attention or Kadir and Brady [2] reports that salient regions can be computed using multiscale images. Gilles [3] on the other hand argues that local complexity can be a measure of saliency. Also, a learning approach for visual saliency models has been proposed recently by Keinzle *et al.* [4]. Following these ideas, one foundation of our approach is the claim, that fundamental attention attractors originating from sensory input can be either static salient features in a single frame or dynamics in the input data sequence (considering temporal properties).

Inspired by the idea of O'Regan and Noë [5], which they claim to be biologically plausible, we extend the saliency attention approach with the idea, that vision is a process of active and sometimes even volitional exploration of the environment. Thus, considering the second before mentioned assumption, i.e. based on the theory of *inhibition of return* – as shown to be plausible in human visual psychophysics by Posner and Cohen [6], we implement top-down cognitive feedback in the proposed system. Moreover, we will go one step further and integrate a possiblity not only for attention inhibition, but also for directed attention guidance. This reinforcement is triggered by cognitive processes – reasoning about relevant additional information to gain from a specific region (see Section 3). Although we do not claim to implement the entire framework of O'Regan's and Noë's theory – e.g. the *inattentional* or *change blindness*, we in deed show that a system utilizing the basic ideas performs considerably better than without.

Not contradicting, but complementing the work of e.g. Itti *et al.* [7, 8] or Walther *et al.* [9], we do not want to fo-

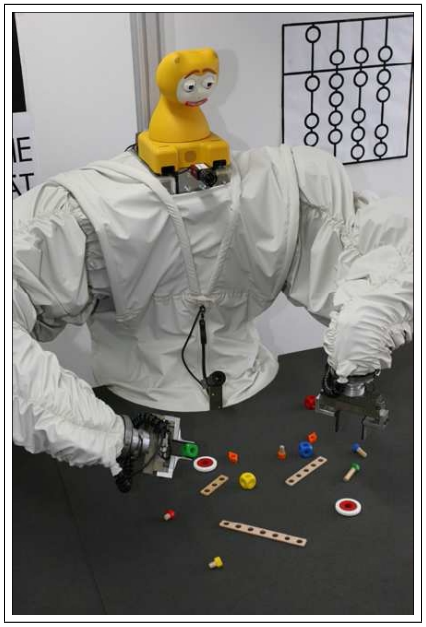

Figure 1: The JAST human-robot dialog system.

cus solely on building a biologically plausible visual systems, but our primary target is to apply the underlying ideas of such frameworks to a real-world robotic setup. We therefore avoid complex neural, connectionist or machine learning techniques where possible, giving preference to discrete algorithms. These fast and efficient algorithms allow for realtime performance and high accuracy for manipulation tasks on standard hardware.

The vision system presented in this paper is part of the "**J**oint-**A**ction **S**cience and **T**echnology" human-robot dialog system.

1.1 The JAST Robot

The overall goal of the JAST project is to investigate the cognitive and communicative aspects of jointly-acting agents, both human and artificial. The robot torso (Figure 1) being built as part of the project [10, 11] 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, 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.

Vision processing in the JAST system is performed on the output of a single top-view camera. It provides an image stream of 7.5 frames per second at a resolution of 1024x768 pixels. The output of the vision process is published to the multimodal fusion component [12], where it is used for disambiguating spoken input from the user. Moreover, combined hypotheses representing the users requests are produced and reasoning on the properties of the observed world parameters can be done.

1.2 Vision Architecture and Analysis Stage

The vision system presented here (Figure 2) applies an asynchronous communication mechanism (ACM). Therefore we can implement non-blocking behaviour and still guarantee the required frequency for result publishing, as publishing incomplete analysis results is tolerated. Derived from common standards, intermediate vision data is managed in limited-size priority-queues.

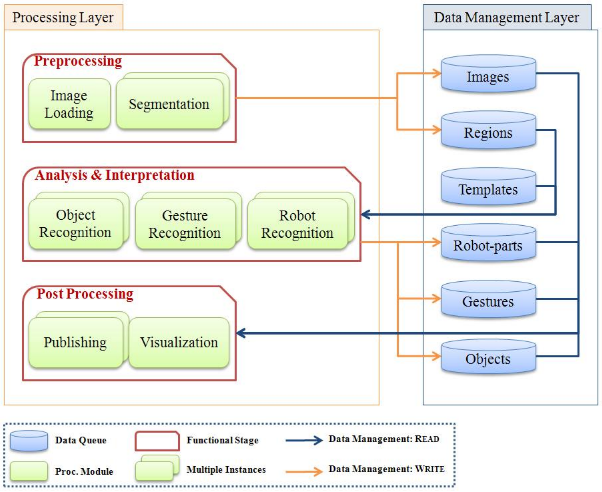

Figure 2: Architectural overview of the vision system.

Concerning parallelization techniques [13], the JAST vision system applies a wait-free [14] combination of *data-domain* and *function-domain* parallelization. As previously shown by Müller *et al.* [15], this combined approach performs very well in practice because anchor points for distributed computation can be designed to be independent concerning memory and workflow and the system avoids starvation according to Herlihy [16].

Object recognition within the scene is fairly straight forward once the regions of interest are identified. We apply the *OpenCV*-implementation of a template-matching algorithm on 20 different rotations of each template we consider to be relevant. The templates are generated from previously taken samples (a future version of the system will extract them online).

Since the OpenCV library offers a good choice of different similarity measures, e.g. cross-correlation (see below) and a couple of other patch-comparison methods – for example least-square-errors (LSE) or correlation-coefficients.

$$CC_{T,ROI} = \text{max}_{t_{\Delta x,\Delta y,\theta}} \left[\sum_{p\in T} \left(T(p)\cdot ROI(t(p))\right) \right] \qquad (1)$$

In Equation 1 the transform $t_{\Delta x,\Delta y,\theta}(p)$ denotes the projection of p into the region ROI according to the translation $\Delta x, \Delta y$ (automatized within OpenCV) and a discrete rotation value $\theta \in [0; 2\pi]$ applied by our algorithm.

Gesture recognition is implemented as a two step approach: first, specific invariants have to be extracted from a region, and second, the gesture has to be classified.

To classify an extracted set of invariants, we find the K nearest neighbors which are calculated based on the weighted distance of each training vector to the input invariant. The training vectors are created in advance and remain stable throughout the whole process. Next, we choose the K vectors from the training pool which have the shortest normalized (Euclidian) distance to the given invariants. A naïve Bayes probability for the invariants can then be computed for each available class of gestures (details in [17]). Figure 3 shows typical results of the recognition process.

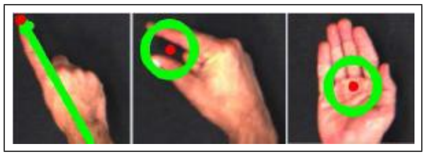

Figure 3: Typical results of the gesture recognition subsystem.

Self recognition or robot recognition is acomplished by a cognitive feedback algorithm. From robot sensors one can retrieve the current joint parameters of each joint of the robot's arms. This information is used to adjust a 3D model of the robot accordingly. Based on this information, in combination with link properties and the position of the torso, that are known from a priori, the system can compute the 3D cartesian position of each joint applying forward kinematics. This information is then used to identify corresponding regions in the input image.

2. ATTENTION ATTRACTORS

Before passing information to the *Analysis and Interpretation* stage described above (see Figure 2), we apply a novel early processing mechanism, which is the main topic of interest here. This section herein describes the effects directly triggered by the input data.

The goal of attention attraction subsequently described is to generate a saliency map. But we want to extend the idea of the saliency map to a more general map of visual attention and emphasize our technical perspective, so we introduce the term *attention condensation layer* here.

Both algorithms described in this section directly process sensory input data, so we call the emerging effects *bottom-up* attention attractors. Itti *et al.* [8] refer to these effects as "scene-dependent", on the contrary to "task-dependent" ones that originate from higher cognitive processing.

2.1 Static Saliency

Our approach for detecting salient local features in a single, thus static, input image relies on a comparison of intensity and hue [18]. A background model is used, which can be trained in advance. This model is represented as a 2D normalized joint histogram [19]. Creating the model is straight forward, as it is sufficient to only analyze one empty input frame, i.e. one that does not contain any objects, gestures or robot parts and can thus be automated.

In the saliency detection step the model is compared to hue-intensity distribution of image patches in the input image – e.g. by applying the *Bhattacharyya* distance. If the distance is greater than a certain threshold, the patch is considered to be non-background and worth analyzing it within the recognition stage (see Figure 4). Histograms of arbitrary patches can be efficiently obtained with the adaptive integral histogram approach described by Müller *et al.* [20].

2.2 Dynamic Saliency

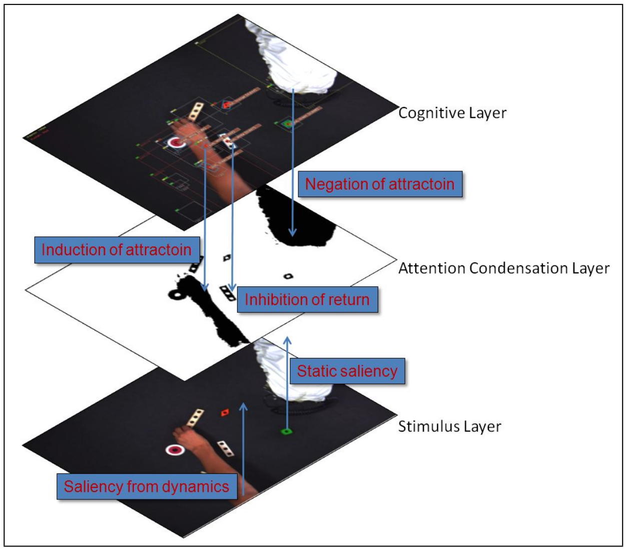

Figure 4: The visual layers for attention based compuation.

The extraction of saliency from dynamics we describe here, is an extension to the approach for detecting locally salient regions explained above. The basic idea is to create and evaluate a disparity map, combine it with the saliency map and observe its behavior for a number of frames.

Regions of particular interest are blobs of pixels moving in an uniform manner for a number of frames. On the basis of this observation, the system is able to infer regions containing high dynamics, that are then considerd to be worth analyzing. Technically speaking, we keep track of regions that show local salience, and as soon as they start moving in the scene we set a flag indicating dynamic saliency.

3. COGNITIVE FEEDBACK

As the second fundamental principle in our attention based robot vision system we assume, that the cognitive layer should be able to influence the amount of attention payed to a bottom-up attracted region by giving some sort of feedback. From a technical point of view this means projecting knowledge about a scene or about constraints in the world into the attention condensation layer. The active or sometimes even conscious projection of world knowledge can either cause inhibitory effects, attract attention, or increase the level of attention.

Neuroscientists (e.g. Li *et al.* [21]) often call these effects on the primary visual cortex of humans *top-down* effects, as they originate from higher levels of cognition, or *task-dependent* [8], i.e. as the high level task or plan influences lower level visual attention to specific regions. Their experiments show, that the same attention attractors have very different influence on the focus of attention and thus the activation of processing units under variations of the task to accomplish.

3.1 Inhibition of Return

"Inhibition of return" (e.g. by Posner *et al.* [6, 22]) constitutes the theoretical foundation of one of the algorithms used to control the focus of attention in a top-down manner.

Here we are talking about a situation where the system's attention attractors got activated and regions for analysis from the bottom-up view were identified. In this case the inhibition of return mechanism avoids re-analyzing regions that have been previously processed (Figure 4).

In order to achieve this, the system keeps track of any object, gesture or part of the robot visible in the scene. Many of these items are likely to appear at the same or very close position in consecutive frames. The level of attraction for a ROI $a(ROI, t)$ is in this case proportionally decreased with the number of sequential frames t it appears in. Here we propose two methods, either linear or non-linear degression.

$$a(ROI, t) = \begin{cases} \text{linear:} & \delta(ROI) - t \\ \text{non-linear:} & tanh(-\alpha t) + 1 \end{cases} \quad (2)$$

Within the linear degression, we have to specify a $\delta(ROI)$ which specifies the maximum number of consecutive attractions subject to the size of a region (larger regions need more time for analysis, therefore the inhibition of return affects big regions later), whereas within the non-linear case, we have to specify a factor α that determines the duration of the attraction and its strength subject to t. Usally $\alpha \simeq \delta(ROI)^{-1}$ is a good choice here.

Combined with a signum thresholding function T_ϵ, we can compute the time t corresponding to $T_\epsilon = 0$ after which no more processing is performed and no more ressources are used for the analysis of a region. In the linear case $\epsilon \geq 0$ is permitted, while $\epsilon > 0$ is a constraint in the non-linear case.

$$T_\epsilon(ROI, t) = \begin{cases} 1 & \text{if } \; a(ROI, t) \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3.2 Volitional Attention Control

Thinking about human attention control again, we find that we are able to control our focus of attention and direct it to a certain region, or generally to a subset of the perceived input data. This was shown to be plausible in Stroop's historical psychological experiments [23] and is still being researched on, e.g. by Cohen *et al.* [24].

There are basically two complementary options for this kind of attention control effects. First, if we suppose an object to be at a certain location, we are able to take a closer look, even when lacking the bottom-up stimulus. And second, even if there is bottom-up attraction, we can volitionally choose not to pay attention to this specific attractor and ignore the stimulus. The system proposed in this paper applies both control strategies by utilizing specific interfaces in the higher level cognition modules (indicated by top-down arrows in Figure 4).

Conscious Induction of Attraction on the one hand is an interface allowing for emphazising of specific or creation of new areas of interest in the attention condensation layer. Hence attention on existing regions is reinforced or regions of interest are generated artificially.

Negation of Attraction on the other hand, unlike the inhibition of return, is not automatically triggerd by the visual activation tracking mechanism, but rather originating from a higher level of cognition. Still, invoking the interface causes a similar effect: the computational effort put into the analysis of the designated region is repealed.

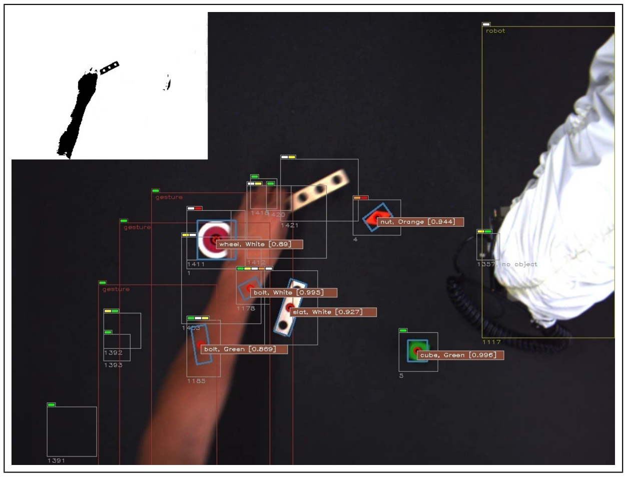

Figure 5: The figure shows a snapshot of an analyzed scene and the binarized attention condensation layer in the upper-left corner.

4. EXPERIMENTAL RESULTS AND CONCLUSION

Figure 5 exemplarily depicts an image of a dynamic input sequence and the analysis performed. In the figure some relevant information for high level cognitive / reasoning modules is annotated: a gesture is recognized and several objects are detected and their location, orientation and color is identified. The boxes around the identified items are not published, but indicate the corresponding ROIs extracted with the attention attraction algorithms.

The snapshot taken from a typical input video sequence[1] shows all of the effects described in the above sections:

Static saliency Region 1421, the slat in the subject's hand, is computed with this approach. Its intensity differs a lot from the background and so the previously unseen region is considered to be worth paying attention.

Dynamic saliency The red regions, likely to be gesture regions, are extracted using the moving blob approach. In the image, the last few blob positions are depicted. Also, regions 1391–1393 in the image originate from scene dynamics, but unfortunately they are false positives.

Inhibition of return Regions 4 and 5 containing an orange nut and a green cube are not reanalyzed, although static attention attractors were activated by the bottom-up mechanism. In this case the inhibition of return mechanism avoids the waste of resources on these specific regions.

Conscious induction The virtual regions 1185 and 1178, are projected by the cognitive layer. From the whole video sequence one can see, that these objects exist on the table, although they are invisible in the snapshot. The cognitive layer infers, that objects do not disappear so suddenly and thus virtual regions are generated and reanalyzed.

Negation of attraction The cognitive layer prohibits further analysis for the small region 1357 next to the robot, as it is part of the inferred position of the robot and thus does not contain relevant information. Although, by means of the region size it could possibly contain an object, the poperty "no object" is assigned.

One has to consider, that a performance analysis cannot

[1] http://www.youtube.com/watch?v=JupXjgdYzY4

be performed straight forward, because, as shown in [15], the system operates massively parallel in the function **and** the data domain. This means, lacks of computational ressources are compensated for with frame-drops. But still, we are able to measure according to a very basic metric: the time it takes, until each object in a scene is detected and analyzed.

We first analyze the system's performance without any attention driven improvements. We find, that evaluating a single frame even for a system only using object recognition (no gesture and no robot recognition) with respect to 16 possible template objects (a typical number for the JAST setup) and 20 rotations takes around 120 seconds for analyzing the 21 objects! But, when applying the bottom-up attention attraction mechanism of static saliency, the time needed for processing the whole scene already decreases to 6.39 seconds. Further on, considering a sequence of image frames containing the one from above, enabling the inhibition mechanism improves the performance to more than real-time, once all regions were analyzed. Obviously the results given here refer to real-world scenarios as being common in the JAST setup. Indeed, we can artificially set up arbitrarily complex scenes – e.g. filling up the entire table with overlapping objects results in a single ROI for the whole image and totally ruin the benefit of bottom-up saliency mechanisms.

In order to show the value of attraction on dynamics and conscious attention focussing, we consider the example of moving a hand or robot arm in the scene. First, dynamic saliency compensates for inhibition of return, so moving objects are reanalyzed although the stimulus itself might remain almost static, but spatial changes or distortions trigger the analysis. Finally, conscious mechanisms allow to compensate for unlogic attention attraction, such as unnatural region behaviour (sudden disappearances or appearances) or false positives due to errorneous saliency.

5. ACKNOWLEDGEMENT

This work was supported by the EU integrating project *Joint-Action Science and Technology* (FP6-003747-IP).

6. REFERENCES

[1] P. Reinagel and A. M. Zador, "Natural scene statistics at the center of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[2] T. Kadir and M. Brady, "Saliency, scale and image description," *IJCV*, 2000.

[3] S. Gilles, "Robust description and matching of images," Ph.D. dissertation, University of Oxford, 1998.

[4] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Proc. NIPS '06*, 2006.

[5] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioural and Brain Sciences*, vol. 24, pp. 939–1031, 2001.

[6] M. Posner and Y. Cohen, "Components of visual orienting," *Attention and Performance*, vol. 10, pp. 531–556, 1984.

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

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[9] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition – a gentle way," in *Biologically Motivated Computer Vision*, 2002, pp. 251–267.

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

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

[12] M. Giuliani and A. Knoll, "Integrating multimodal cues using grammar based models," in *Proc. ICMI*, 2007.

[13] D. E. Culler, J. P. Singh, and A. Gupta, *Parallel Computer Architecture: A Hardware/Software Approach.* Morgan Kaufmann Publishers, 1999.

[14] H. Sundell and P. Tsigas, "Fast and lock-free concurrent priority queues for multi-thread systems," in *Int. Parallel and Distributed Proc. Symp.*, 2003.

[15] T. Müller, P. Ziaie, and A. Knoll, "A wait-free realtime system for optimal distribution of vision tasks on multicore architectures," in *Proc. ICINCO '08*, 2008.

[16] M. Herlihy, "Wait-free synchronization," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 13, no. 1, pp. 124–149, 1991.

[17] P. Ziaie, T. Müller, M. E. Foster, and A. Knoll, "A naïve bayes classifier with distance weighting for hand-gesture recognition," in *Proc. Intl. CSI Computer Conference '08*, 2008.

[18] Y. Hu, D. Rajan, and L.-T. Chia, "Detection of visual attention regions in images using robust subspace analysis," *Journal of Visual Communication and Image Representation*, vol. 10, no. 3, pp. 199–216, April 2008.

[19] G. Pass and R. Zabih, "Comparing images using joint histograms," *Multimedia Systems*, vol. 7, no. 3, 1999.

[20] T. Müller, C. Lenz, S. Barner, and A. Knoll, "Accelerating integral histograms using an adaptive approach," in *Proc. ICISP'08*, ser. LNCS. Springer, July 2008, pp. 209–217.

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

[22] M. I. Posner, R. D. Rafal, L. S. Choate, and J. Vaughan, "Inhibition of return: Neural basis function," *Cognitive Neuropsychology*, vol. 2, no. 3, pp. 211–228, August 1985.

[23] J. R. Stroop, "Studies of interference in serial verbal reactions," *Journal of Experimental Pyschology*, vol. 18, pp. 643–662, 1935.

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