A NEURAL MODEL COMBINING ATTENTIONAL ORIENTING TO OBJECT RECOGNITION: PRELIMINARY EXPLORATIONS ON THE INTERPLAY BETWEEN WHERE AND WHAT

Florence Miau and Laurent Itti
Department of Computer Science, University of Southern California, Los Angeles, CA 90089-2520

Abstract: We propose a model of primate vision that integrates both an attentional orienting (“where”) pathway and an object recognition (“what”) pathway. The fast visual attention front-end rapidly selects the few most conspicuous image locations, and the slower object recognition back-end identifies objects at the selected locations. The model is applied to classical visual search tasks, consisting of finding a specific target among an array of distracting visual patterns (e.g., a circle among many squares). The encouraging results obtained, in which substantial speedup is achieved by the combined attention-recognition model while maintaining good recognition performance compared to an exhaustive search, suggest that the biologically-inspired architecture proposed represents an efficient solution to the difficult problem of rapid scene analysis. Keywords: Visual attention, object recognition, scene analysis, bottom-up, top-down.

I. INTRODUCTION

When we search for specific objects in visual scenes, we do not exhaustively scan the entire scene in an orderly manner, nor do we attempt to recognize objects at every spatial location. The selection of locations to be analyzed is determined by our focal visual attention system, which rapidly focuses on “interesting” image regions under two types of guiding influences: Bottom-up or image-derived cues, and top-down or task-derived cues [3], [2]. The interaction between bottom-up and top-down cues allows primates to analyze complex visual scenes in a small fraction of the time which would be required to exhaustively scan the entire scene. In the primate brain, anatomical and functional separation between localization and identification of objects is observed: cortical areas along the dorsal stream of visual processing (including posterior parietal cortex) are concerned with spatially directing attention towards conspicuous image locations (“where”), while areas along the ventral stream (including inferior temporal cortex) are concerned with localized identification of attended objects (“what”) [11].

We present a neuromorphic model combining both the where and what processing streams. The model is built upon the synergy between two biological models: a rapid, massively parallel model for the bottom-up control of visual attention extracts the few most salient locations from an incoming visual scene, and a computationally expensive model of object recognition identifies objects at attended locations. In the following sections, we describe the combined attention-recognition model, and present results obtained with simple search arrays (find a circle among rectangles).

II. VISION ALGORITHMS USED IN THIS STUDY

The present work studies the integration of two very different types of vision algorithms: First, a fast bottom-up attention front-end eliminates as many obviously irrelevant locations as possible, while maintaining a low rate of target misses. At each selected location, an object recognition back-end attempts to identify attended objects, while minimizing both target misses and false alarms, at the cost of much greater computational complexity. In this section, both components are presented, followed by a description of their integration into a complete target detection system.

A. Bottom-up Visual Attention Front-End

The front-end is derived from our saliency-based, bottom-up model for the control of covert visual attention in primates [5], [2], [4]. It employs a biologically-plausible architecture to rapidly select the few most conspicuous locations in any given scene. The input image is decomposed into seven channels (intensity contrast, red/green and blue/yellow double color opponencies, and 0°, 45°, 90° and 135° orientation contrast) at six spatial scales,
# Report Documentation Page

<table>
<thead>
<tr>
<th>Report Date</th>
<th>N/A</th>
<th>Dates Covered (from... to)</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 Oct 2001</td>
<td>N/A</td>
<td>-</td>
<td>N/A</td>
</tr>
</tbody>
</table>

## Title and Subtitle
A Neural Model Combining Attentional Orienting to Object Recognition: Preliminary Explorations on the Interplay Between Where and What

## Author(s)

## Performing Organization Name(s) and Address(es)
Department of Computer Science University of Southern California Los Angeles, CA 90089-2520

## Sponsoring/Monitoring Agency Name(s) and Address(es)
US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500

## Distribution/Availability Statement
Approved for public release, distribution unlimited

## Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

## Abstract

## Subject Terms

## Report Classification
Unclassified

## Classification of Abstract
Unclassified

## Number of Pages
4
yielding a total of 42 feature maps. The six intensity maps, twelve color maps, and 24 orientation maps are combined into the three summary maps shown in Fig. 1. After iterative spatial competition for saliency within each map [2], [4], which allows only a sparse number of locations to remain active, all maps are combined into the unique saliency map [7], that is, a scalar topographic neuronal map which encodes for stimulus saliency irrespectively of the visual modality by which a given stimulus is salient. The saliency map is scanned by the focus of attention in order of decreasing saliency, through the interaction between a winner-take-all neural network (which selects the most salient location at any given time) and an inhibition-or-return mechanism (which transiently suppresses the currently attended location from the saliency map) [5], [6]. In Fig. 1, the system selected the two isolated pedestrians as being the two most salient locations in the image, mostly because of their strong responses in the orientation channel.

The system performs remarkably well at picking out salient targets from cluttered environments. Experimental results include the reproduction by the model of human behavior in classical visual search tasks (pop-out versus conjunctive search, and search asymmetries; see [10], [2]): a demonstration of very strong robustness of the saliency computation with respect to image noise [5]; and the automatic detection of traffic signs, pedestrians, military vehicles and other salient objects in natural environments [2], [4] (see http://iLab.usc.edu). What is remarkable in these results is not only the wide range of applications, all using color images in which high amounts of noise, clutter, variations in illumination conditions, shadows and occlusions were always present. Even more interesting is that the same model is able, with no tuning or modification, to detect salient traffic signs in roadside images taken from a low-resolution camera (512 x 384) on a vehicle, pedestrians in urban settings, various salient objects in indoor scenes, military vehicles in very large (6144 x 4096) and highly cluttered color images of rural scenery, or salient ads in screen grabs of web pages.

B. Object recognition back-end

Some hallmarks of the human visual system are invariance to image transformations (for example, we can recognize a specific face among many, while being rather tolerant to changes in viewpoint, scale, illumination and expression), robustness to clutter (target objects can be detected in complex real world scenes), and speed. A recent model of object recognition in visual cortex, HMAX from MIT [9], has suggested computational mechanisms key to biological object recognition that differ substantially from current computer vision approaches. The model is consistent with a number of physiological and psychophysical data and leads to a comprehensive perspective on the biological basis of object identification and classification.

The model (Fig. 2) is based on a hierarchical feedforward architecture that starts out from units tuned to bar-like stimuli with small receptive fields, similar to “simple cells” in primary visual cortex. At the top of the hierarchy are view-tuned units, as found in inferotemporal cortex, one of the highest areas in the ventral visual stream. These view-tuned units respond specifically to a view of a complex object while showing tolerance to changes in stimulus scale and position. Feature complexity and transformation invariance are built up through a succession of feature complexity-increasing ’C’ layers. While feature complexity is increased by a template match operation, HMAX uses a new pooling mechanism, a max operation, to increase invariance in the C layers by pooling over units tuned to the same feature but at different scales and positions [9].

C. Integrating Attention and Object Recognition

In our first attempt to combine target detection and identification into a single computational framework, here we simply crop a 128 x 128 sub-image from the input scene, at each attended location, and feed that sub-image to the recognition module. The recognition back-end is trained to differentiate between several possible objects.

III. Results

We present simple results using synthetic search array stimuli [10]: a single circular shape (target) is embedded in an array of randomly rotated and jittered rectangular shapes (distractors). The circle is not particularly salient
among the rectangles (because it contains no unique low-
level feature, such as a different color or a unique ori-
tentation), and does not immediately “pop-out” from the
surrounding rectangles. Thus, the attentional front-end
successively selects various shapes in the array, and the
recognition back-end is used to determine when the se-
lected shape is the target circle.

A. Learning and Simulation Results

In order to recognize various shapes, we build one view-
tuned unit in HMAX for each shape to be recognized.
Each such unit receives weighted inputs from 256 C2
cells (Fig. 2). The corresponding 256 synaptic weights
are trained using two sample images (one containing the
target circle and another containing the distractor rect-
gle). An additive learning rule is iteratively applied,
which increases the weights associated to C2 cells re-
sponding strongly to the target, and decreases the weights
associated to the C2 cells responding strongly to the dis-
tractor. For each C2 cell $i$:

$$w_i \leftarrow w_i + (1 \pm \eta C_{2i}) \quad \text{[+ if target, - if distractor]}$$

After each update, the vector of weights $w$ is nor-
malized to unit norm. The update is iterated with
$\eta \in \{0.0, 0.2, 0.3, 0.4\}$, over 9 possible numbers of repetitions
between 240 and 1,200. The output of the view-tuned cell
is finally obtained by taking the difference between the
sum of the C2 responses with weights all equal to 1/256
and the sum of the C2 responses with trained weights; a
positive response means that the target was found.

For five different pairs of circles and rectangles ($128 \times
128$), we train and evaluate the performance of our system
using arrays made of the shapes used for training. Arrays
contain one target (circle) among many randomly jittered
and oriented distractors (rectangles). Speckle color noise
is added to each array with uniform density of 0.1. For
each target/distractor pair, we use six array sizes ($2 \times 2,$
$3 \times 3, 4 \times 4, 5 \times 5, 6 \times 6, 7 \times 7$) and ten randomized
instantiations of the arrays for each size.

Depending on the pair, we obtain overall good discrim-
ination, as shown in the receiver operating characteristic
(ROC) curves of Fig. 3. Pair No. 2 yields best discrimina-
tion, because of the strong orientation difference be-
tween both shapes. Hence, it is possible to obtain excel-
ent recognition performance with HMAX if the patterns
to be discriminated reasonably differ. The worst results
are obtained for pair 4, in which target and distractors
are very similar in size, position and overall shape.

On our computer systems (800MHz Pentium-III Linux
machines), each $128 \times 128$ HMAX evaluation is achieved
in 6.77s. For a $1280 \times 1280$ array, attempting recognition
at every location on a grid with $64 \times 64$ pixel spacing (to
ensure sufficient overlap between adjacent locations to be
analyzed) would hence require $\approx 2.977h \approx 50$min. The
computation of the saliency map, however, only takes 32s
for the same $1280 \times 1280$ array, which is equivalent to the
time required for $\approx 4.75$ recognitions. Thus, our imple-
mentation results indicate that if the attention front-end
can eliminate just five recognitions of the 441 done in
an exhaustive search, the combined model will perform
faster than the exhaustive search.

B. Generalization

One important test to evaluate the usability of our
combined model in practice regards the robustness of the
recognition process with respect to the exact coordinates
of the focus of attention. Indeed, attention is directed to
objects based on the activity in the saliency map, which
has a coarse scale compared to the original image (16
times smaller horizontally and vertically). Thus, it can-
not be guaranteed that attention will always exactly be
centered onto the object, but variations of $\pm 16$ pixels in
the coordinates of the focus of attention can reasonably
be expected, due to noise or sampling artifacts.

We train the HMAX model using pair No. 2, and eval-
uate how the system can recognize translated and ro-
tated versions of the images in pair No. 2. Translations
of -70 to 70 pixels are systematically applied to the rect-
angular shape (Fig. 4), as well as rotations of 0 to 180°
(Fig. 5), in a manner similar to that previously described
with HMAX [9]. Overall, we observe good invariance
of recognition with respect to translation and rotation. This
result indicates that our combined approach, in which the
exact positioning of the focus of attention onto an object
to be recognized may be subject to small variations, is
overall robust thanks to the small-scale translation and rotation invariance built into the HMAX model. This systematic analysis directly consolidates the very good ROC results obtained with search arrays (Fig. 3).

IV. DISCUSSION AND CONCLUSION

Our results indicate that much computation time can be gained by the use of a fast visual attention front-end to guide a more expensive object recognition back-end. This approach substantially differs from traditional target detection and recognition approaches, for example using template matching. Here indeed, the main characteristic of our bottom-up attention model is its nonspecificity, which allows it to attend to a wide range of salient objects (traffic signs, pedestrians, military vehicles, etc.) under a wide variety of imaging conditions (indoors, outdoors, in color or grayscale, and in low or high resolution), in a manner similar to what humans do when exploring new scenes [12]. Although more efficient dedicated computer vision algorithms exist for, e.g., the detection of traffic signs [1], those cannot pick up tanks hidden in an aerial view of a forest (in [1], for example, includes segmentation of specific colors, and a specific algorithm to detect triangular signs).

A challenge for future research will be to extend HMAX to natural color images. Preliminary testing with grayscale images containing pedestrians suggested that the current HMAX has difficulty discriminating between pedestrians and the wide variety of other salient locations selected by the attention model. In order to allow HMAX to perform as well as the best available computer vision techniques [8] when background clutter is present, we will need to both increase the number of low-level features and the number of layers in the HMAX architecture.

In conclusion, we have demonstrated a full implementation of a complete target detection system inspired from biology. The success of this approach shows that using a relatively crude and non-specific attention focusing system, based on a small set of simple early vision features, can substantially accelerate a search for targets. Our results hence suggest that even sophisticated pattern recognition algorithms can benefit from a spatial reduction in their search space based on simple image cues.

Acknowledgments: This work was supported by startup funds to L.I. from the USC School of Engineering. We thank Max Riesenhuber and Tomaso Poggio from the MIT AI lab for advice and the HMAX code, and Christof Koch from Caltech for discussions.

REFERENCES