Neuromorphic algorithms for computer vision and attention

Florence Miau[1], Constantine Papageorgiou[2] and Laurent Itti[1]

[1] Department of Computer Science, University of Southern California, Los Angeles, CA
[2] Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

ABSTRACT

We describe an integrated vision system which reliably detects persons in static color natural scenes, or other targets among distracting objects. The system is built upon the biologically-inspired synergy between two processing stages: A fast trainable visual attention front-end ("where"), which rapidly selects a restricted number of conspicuous image locations, and a computationally expensive object recognition back-end ("what"), which determines whether the selected locations are targets of interest. We experiment with two recognition back-ends: One uses a support vector machine algorithm and achieves highly reliable recognition of pedestrians in natural scenes, but is not particularly biologically plausible, while the other is directly inspired from the neurobiology of inferotemporal cortex, but is not yet as robust with natural images. Integrating the attention and recognition algorithms yields substantial speedup over exhaustive search, while preserving detection rate. The success of this approach demonstrates that using a biological attention-based strategy to guide an object recognition system may represent an efficient strategy for rapid scene analysis.

Keywords: Visual attention, object recognition, scene analysis, bottom-up, top-down

1. INTRODUCTION

When we search for specific objects in visual scenes, we usually do not exhaustively scan the entire scene in an ordered manner, and we do not attempt to recognize objects of interest at every spatial location. Rather, the study of visual search over the past two decades,[1–5] and in particular recent experiments measuring eye movements during search,[6] suggest that primates use non-sequential, yet non-random and, to some degree, reproducible[7,8] scanning strategies to rapidly explore a new scene. The selection of locations to be analyzed is determined by our focal visual attention system, which rapidly scans "interesting" image regions under two guiding influences: Bottom-up or image-derived cues, and top-down or task-derived cues.[9,5] The interaction between bottom-up and top-down cues allows primates to analyze and comprehend complex visual scenes in a small fraction of the time which would be required to exhaustively explore the entire scene. For example, in classical "pop-out" search tasks, an odd target among an array of identical distractors is effortlessly found in 200-300 ms,[1] while approximately as much time is necessary for *a single* voluntary shift of our attentional focus (covert shift) or of our eyes (overt shift).[10]

Remarkably, however, psychophysical,[1] electrophysiological[11] and modeling[2,12,9] evidence suggests that the bottom-up guidance of focal attention relies on very simple visual cues, such as local intensity, color and orientation contrasts. Hence, the remarkable performance of biological systems in target search tasks seems almost surprising: One would expect, from a computational and machine vision viewpoint, much more complex cues to be required, for example involving rotation- and scale-invariant template matching. For most machine vision problems, employing a very crude front-end to determine a small set of candidate locations to be analyzed in detail indeed seems dangerous, as one would expect such strategy to yield much higher miss rates than exhaustive search.

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 inferotemporal cortex) are concerned with localized identification of attended objects ("what").[13] In the present study, we investigate whether combining computer vision models of both the where and what processing streams may yield a faster and more powerful integrated system.

We approach the problem from two perspectives: First we evaluate the simple coupling between a biologically-inspired visual attention front-end and one of the best-performing object recognition back-ends, using support vector machine (SVM) classifier to identify pedestrians in natural scenes. This first combined model is evaluated on a database of outdoors color images, and yields a speedup greater than 3-fold with less than 5% degradation in

recognition performance, compared to an exhaustive search. Because in this first integrated system the attention and recognition components do not share any of the low-level image processing, we then conduct preliminary investigations with a more biologically-plausible model, in which the object recognition back-end is inspired from the properties of neurons in inferotemporal cortex. The performance of this combined system, which is not yet as robust as the SVM-based approach with natural scenes, is evaluated with simple synthetic stimuli (visual search arrays). We find that using the attention front-end yields very substantial speedup, as computing where to attend takes less time than attempting just 5 localized recognitions (out of 441 recognitions to be attempted on a 1280×1280 image) but typically cuts the number of locations to be examined by a factor 4 or more (depending on each specific stimulus image).

2. VISION ALGORITHMS USED IN THIS STUDY

The present work studies the integration of two very different types of computer 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, one of our two detailed back-ends performs object recognition, and attempts to minimize both target misses and false alarms, at the cost of much greater computational complexity. In this section, the three components are presented, followed by a description of their integration into a complete target detection system.

2.1. 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.[14,9] It employs a biologically-plausible architecture to rapidly select the most conspicuous locations in any given scene. The first stage of the model analyzes the input image in a massively parallel manner, by extracting simple biological vision features (intensity contrast, color double-opponencies, and orientation contrast) at multiple spatial scales. Within each of the resulting topographic feature maps, a spatial competition for salience is applied, which enhances locally unique responses while suppressing extended areas of uniform activity. After within-feature competition, the feature maps are combined into a single scalar "saliency map," which topographically encodes for local conspicuity independently of particular features dimensions. In the second stage of the model, the location of maximum activity in the saliency map, that is, the most conspicuous image location, is detected by a winner-take-all (WTA) neural network, and a fixed-size, circular "focus of attention" is directed towards it. In order to allow the WTA to subsequently detect the next most salient location, a localized inhibitory feedback mechanism transiently suppresses the currently attended location from the saliency map (hence implementing a so-called "inhibition of return" (IOR) mechanism[15]). The interaction between the WTA and IOR yields an attentional "scanpath", which explores the most active locations in the saliency map in order of decreasing saliency **(Fig. 1)**.

When attempting to detect pedestrians in natural scenes, we compared the baseline performance of the attention front-end (which does not know anything about pedestrians), to a trained version in which the weights by which various visual features contribute to the saliency map are learned from example images. We used a simple supervised learning technique to bias the model towards more strongly considering features which are characteristic of pedestrians. Training consists of determining a set of non-topographic, global weights applied to each feature map before addition to the saliency map. Feature weights are adjusted by comparing each map's response inside and outside manually outlined image regions containing targets.[16] This procedure promotes, through an increase in weights, the participation to the saliency map of those feature maps which show higher activity inside targets than outside. It may correspond in biological systems to a simple change in the gain associated with a given feature type, as has been demonstrated, for example, under volitional top-down attentional control.[17]

2.2. Two Alternative Object Recognition Back-Ends

The goal of the back-end is to determine whether or not a target of interest is present at a given image location. In this study, we use two alternative back-ends: First, the support vector machine (SVM)-based trainable object recognition system developed at MIT,[18,19] which has been shown to be among the best available algorithms for object recognition in cluttered natural scenes, although its relevance to biological systems is not evident. Using this very powerful back-end, the resulting attention-recognition system is applied to the detection of pedestrians in color outdoors scenes. Second, a biological object recognition model developed at MIT, HMAX model,[20] is also tested. Although this more recent model does not yet achieve recognition rates which can compete with the first back-end in the presence of noise and clutter, it is of particular interest because of its biological relevance and of the

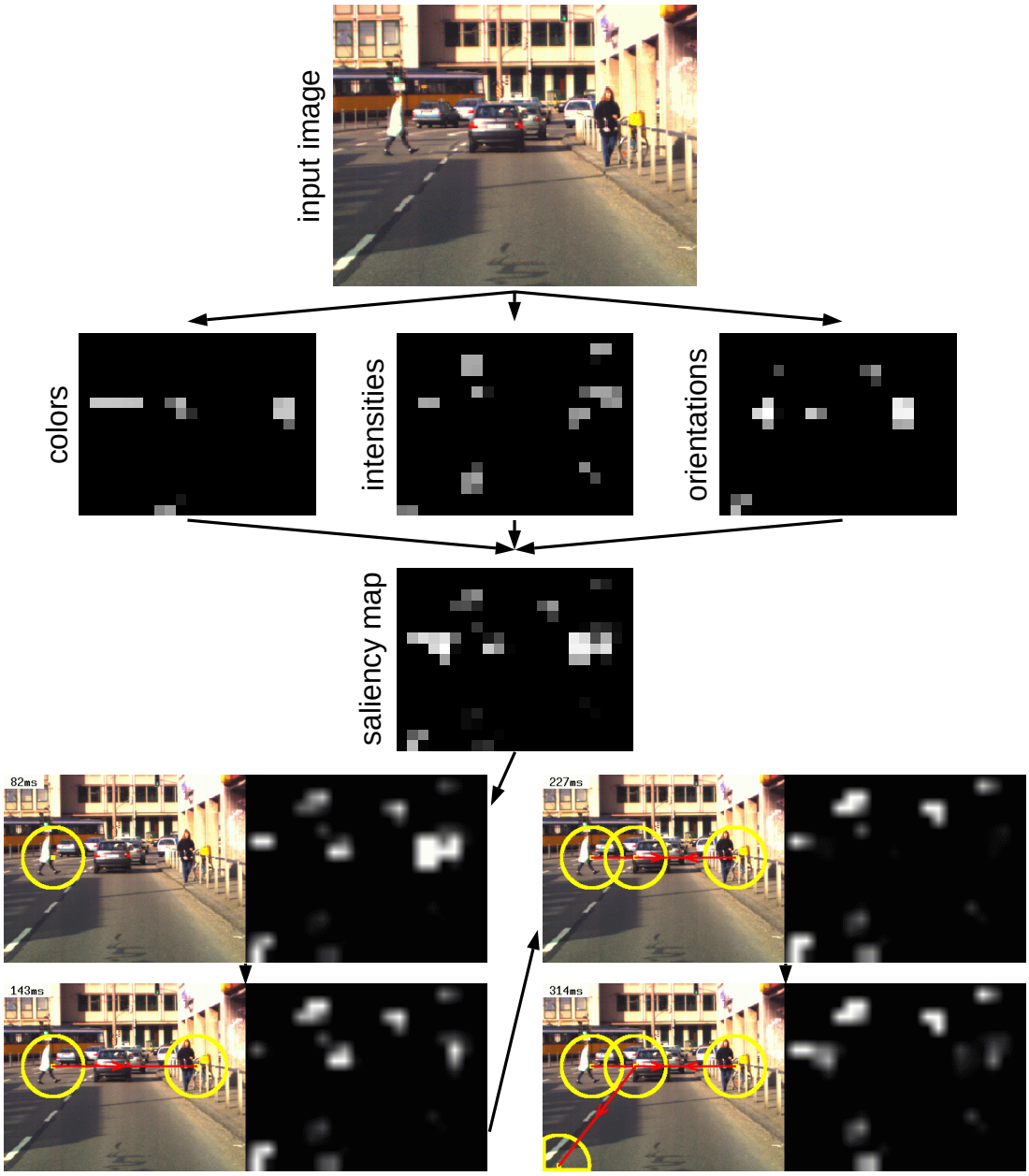


Figure 1. Overview of the bottom-up attentional selection front-end[9,14]. The input image is decomposed into seven channels (intensity contrast, red/green and blue/yellow double color opponencies, and 0°, 45°, 90° and 135° orientation contrast) at six spatial scales, yielding a total of 42 feature maps. Here the six intensity maps, twelve color maps, and 24 orientation maps are combined into the three summary maps shown. After iterative spatial competition for salience within each map, which allows only a sparse number of locations to remain active, all maps are combined into the unique saliency map. 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, as shown on the four bottom frames). In the example shown, 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.

similarity between its early processing stages and those of the attention model. Using this model, performance of the full system is evaluated using simpler stimuli, consisting of search arrays in which a single target circular shape is embedded among many distracting rectangular shapes.

2.2.1. The trainable SVM-based object recognition system

This model uses an object representation which maximizes inter-class variability while keeping intra-class variability low, based on multiscale intensity differences detected by Haar wavelets.[21] This representation yields more robust recognition than pixel-based, region-based, or traditional fine-scale edge-based representations when images present a high degree of variability in the shapes, textures, illumination conditions and color of the pedestrian patterns.

In the first stage of the model, a Haar wavelet transform is run over the input image and yields a set of coefficients at two spatial scales, which contains the responses of the wavelets over the entire image. Three types of wavelets are used, responding to vertical, horizontal, and diagonally oriented intensity differences. Color information is used by applying the Haar transform to each color channel separately, and by selecting, at each location, as the final wavelet coefficient the one from the three color channels which gives the maximal response.

Feature selection is derived from a training database consisting of 1,848 frontal and rear color views of pedestrians (924 originals, plus their mirror images), scaled to 128×64 pixels, and centered at the pedestrian. Using the Haar wavelet representation, both coarse-scale (32×32 pixels) and fine-scale (16×16 pixels) features are computed. At these scales of wavelets, there are 1,326 total features for the 128×64 pattern. The system hence uses a 1,326-element vector which describes these features for each pattern being processed.

Using the 1,848 pedestrian patterns and a set of 7,189 negative patterns gathered from images of outdoor scenes not containing pedestrians, a Support Vector Machine (SVM-based system) classifier is trained to differentiate between pedestrian and non-pedestrian patterns. Support Vector Machines is a training technique which, instead of minimizing the training error of a classifier, minimizes an upper bound on its generalization error. This technique has recently received much attention and has been applied to areas such as handwritten character recognition,[22] 3D object recognition,[23] text categorization,[24] and object detection.[18,19] Some appealing characteristics of SVM-based system rely in that, first, by choosing different kernel functions, one can implement various classifiers (e.g., polynomial classifiers, multilayer perceptrons, or radial basis functions), second, in that the only tunable parameter is a penalty term for misclassifications, and, third, in that the algorithm finds the separating decision surface which should provide the best out-of-sample performance. The SVM-based system decision surface is obtained by solving a quadratic programming problem; for more details on the algorithm, see refs.[25,26]

2.2.2. The HMAX biological object recognition model

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, the HMAX model from MIT,[20] 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, the HMAX model 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.[20]

2.3. Integration of detection and recognition Models

Using only the object recognition component of the system in a baseline approach, to detect targets in a new image we sequentially shift a detection window over the entire image. In addition, to achieve multi-scale detection in the case of pedestrians with the SVM-based back-end system, we incrementally resize the image and slide the detection

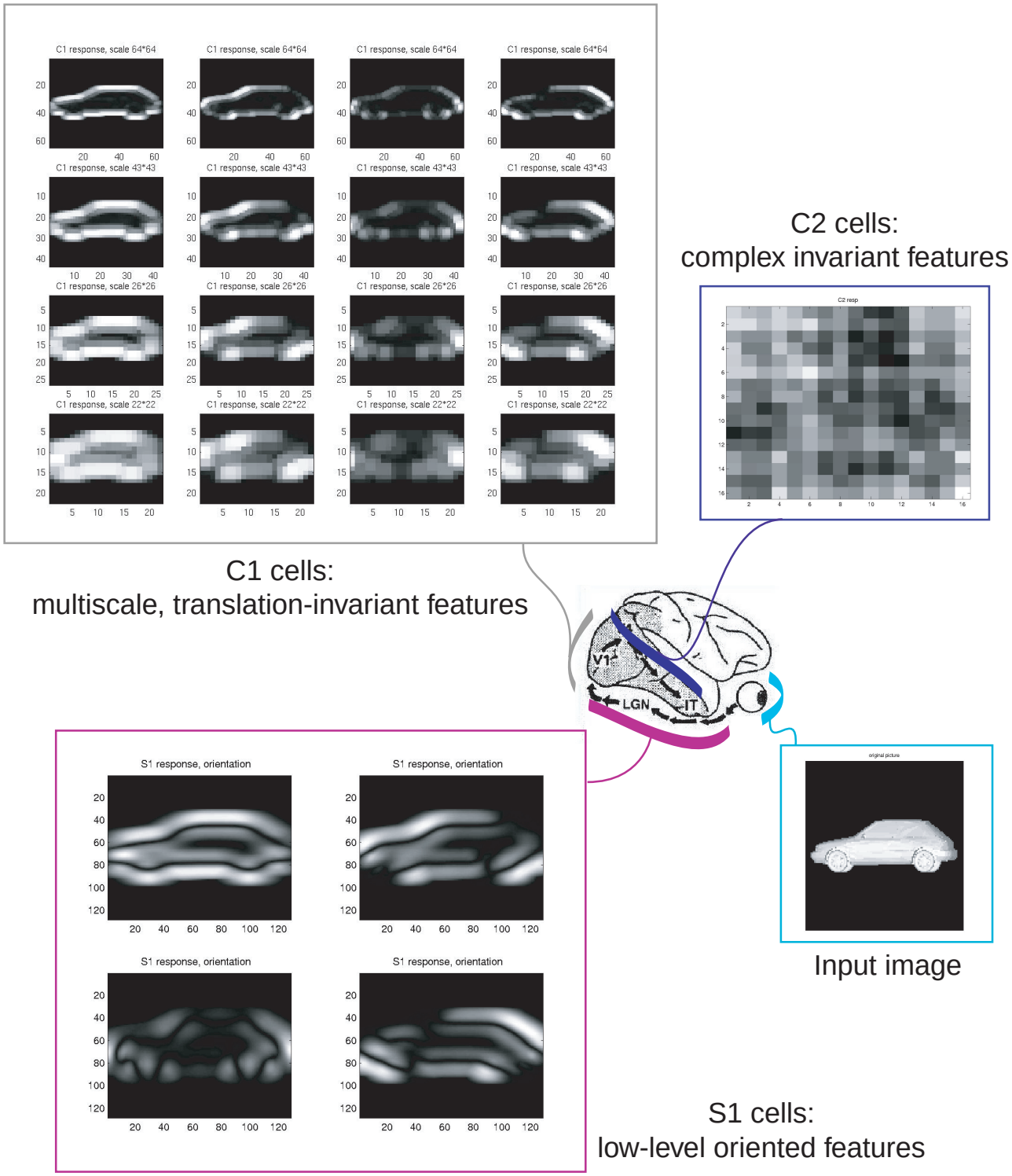


Figure 2. The HMAX model architecture for object recognition[20]

window over each resized image. In the Results Section, we will use this exhaustive approach as a baseline for comparison with the combined attention-recognition system.

		False Alarms	
Generic Model	First target	1.36 ± 2.35	(72 images)
	All targets	4.16 ± 3.70	(58 images)
Trained Model, Training Set	First target	1.27 ± 2.55	(36 images)
	All targets	5.30 ± 5.70	(30 images)
Trained Model, Test Set	First target	1.06 ± 2.15	(36 images)
	All targets	4.23 ± 4.01	(30 images)

Table 1. Number of false detections before targets were found, for the 72 images studied. The generic model, which detects salient locations irrespectively of their nature, visited an average of 1.36 locations before attending to a pedestrian (the first pedestrian found was hence at between the second and third attentional shifts on average). It made an average of 4.16 false detections before sequentially finding all pedestrians in 58 images, and failed to find all pedestrians in the remaining 14 images within the time allocated for the search (up to about 15-20 shifts). After training, we can see the number of shifts necessary to find a first target is reduced, but this is at the cost of increasing the time necessary to find all targets.

Using a cooperation of the front-end and back-end models, the biologically-inspired combined system proposed here searches for targets only in the restricted set of image locations provided by the attentional front-end. In the current implementation, the link between front- and back-end is extremely simple: The focus-of-attention front-end provides a list of candidate (x, y) locations to be analyzed in more detail by the back-end object recognition model.

3. RESULTS

3.1. Detection of Pedestrians Using Attention and SVM-based Recognition

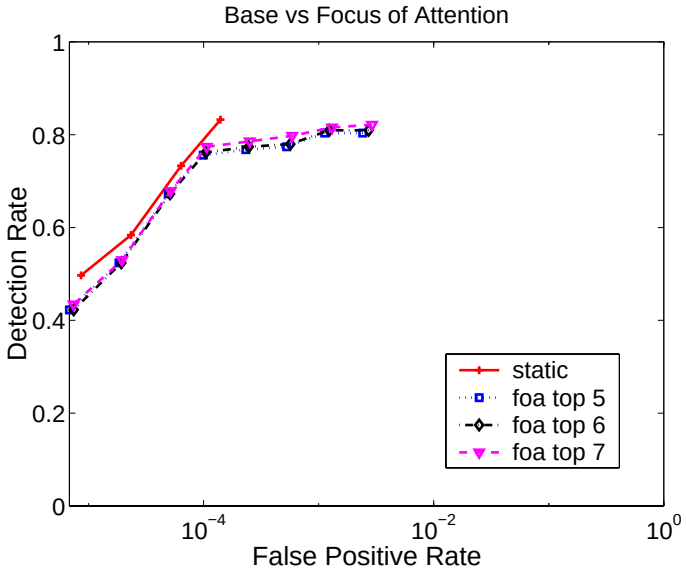
The system integrating the attention front-end to the SVM-based pedestrian detection back-end system was evaluated on a database of 72 color images at 384×256 resolution, and compared to the baseline system (in which an exhaustive search replaces the attention front-end). The images were acquired with a consumer-electronics camera around the MIT campus, and contain substantial amounts of imaging noise, large variations in illumination and shadows, important clutter, and numerous salient "distracting" objects such as buildings, vehicles, flags, advertising signs and trees. In addition, the pedestrians shown in these images vary in size, posture, viewing angle, illumination, colors and contrast. One to six persons were present in each image.

3.1.1. Focus-of-Attention Target Detection Performance

The generic (untrained) bottom-up attention front-end yielded remarkable target detection performance in most images, with on average the 2.36-th attended location being a pedestrian. Training, however, did not substantially improve target detection rate. Since we have previously reported more dramatic improvements using similar training,[16] we attribute the relatively poor improvement obtained here to the too high diversity of target features relatively to the number of features implemented.[16] A summary of the results in terms of the number of false alarms before targets were attended to is presented in **Table 1**. The predictions of the model for the 72 images studied can be interactively examined on our World-Wide-Web site at `http://iLab.usc.edu/bu/javaDemo/` (odd-numbered images are from the training set, while even-numbered ones are from the test set).

3.1.2. Integrated System Performance

We computed receiver-operating characteristic (ROC) curves which compare the integrated system to the baseline (exhaustive) system **(Fig. 3)**. The performance of the version using the attention front-end is slightly worse than for the baseline system. However, fewer locations are considered by the attention-based system, resulting in a significantly increase in processing speed. Indeed, processing time for one image using the baseline approach was approximately one hour on a 450MHz Linux system. As this time directly scales with the number of patterns considered, it was reduced by a factor 3.3 when only the 5 most salient locations were analyzed, i.e., less than 18 minutes per image **(Fig. 3)**. In counterpart, finding the five most salient locations using our front-end only required 30 seconds. The almost superimposed ROC curves for $N \in \{5, 6, 7\}$ indicate that considering more than the five most salient regions essentially does not impact performance.



Front-end	Patterns Processed
Base	100.00%
FOA (top 7)	38.97%
FOA (top 6)	35.14%
FOA (top 5)	29.22%

Figure 3. (left) ROC curves for pedestrian detection, comparing the baseline (exhaustive) system to the trained integrated system using the attentional front-end. Although the attention-based system only explores a small fraction of all image locations, detection rate is only slightly degraded (by less than 5% on average). Analyzing the 5, 6, or 7 most salient locations yields essentially identical results. **(right)** Percentage of total patterns processed by the baseline and integrated models. For the model using the focus-of-attention front-end (FOA), all locations in the analysis window around each attended location are reported.

3.2. Detection of a Circle Among Rectangles using Attention and HMAX model

The above results demonstrate that integrating attention and recognition may yield a very powerful yet computationally efficient target detection system. One aspect which can be further improved in the above system is to try and merge the early image processing stages of both models. To this end, we here present experiments with a second back-end, the HMAX model, whose early stages of processing are very similar to those of the attention front-end.

We present simple results using synthetic search array stimuli[1]: 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 orientation), 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 selected shape is the target circle.

3.3. Learning and Simulation Results

In order to recognize various shapes, we build one view-tuned unit in HMAX model 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 rectangle). An additive learning rule is iteratively applied, which increases the weights associated to C2 cells responding strongly to the target, and decreases the weights associated to the C2 cells responding strongly to the distractor. For each C2 cell i:

$$w_i \leftarrow w_i * (1 \pm \eta \mathrm{C2}_i) \qquad [+ \text{ if target, - if distractor}]$$

After each update, the vector of weights $\mathbf{w}$ is normalized to unit norm. The update is iterated with $\eta \in \{0.2, 0.3, 0.4\}$, 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.

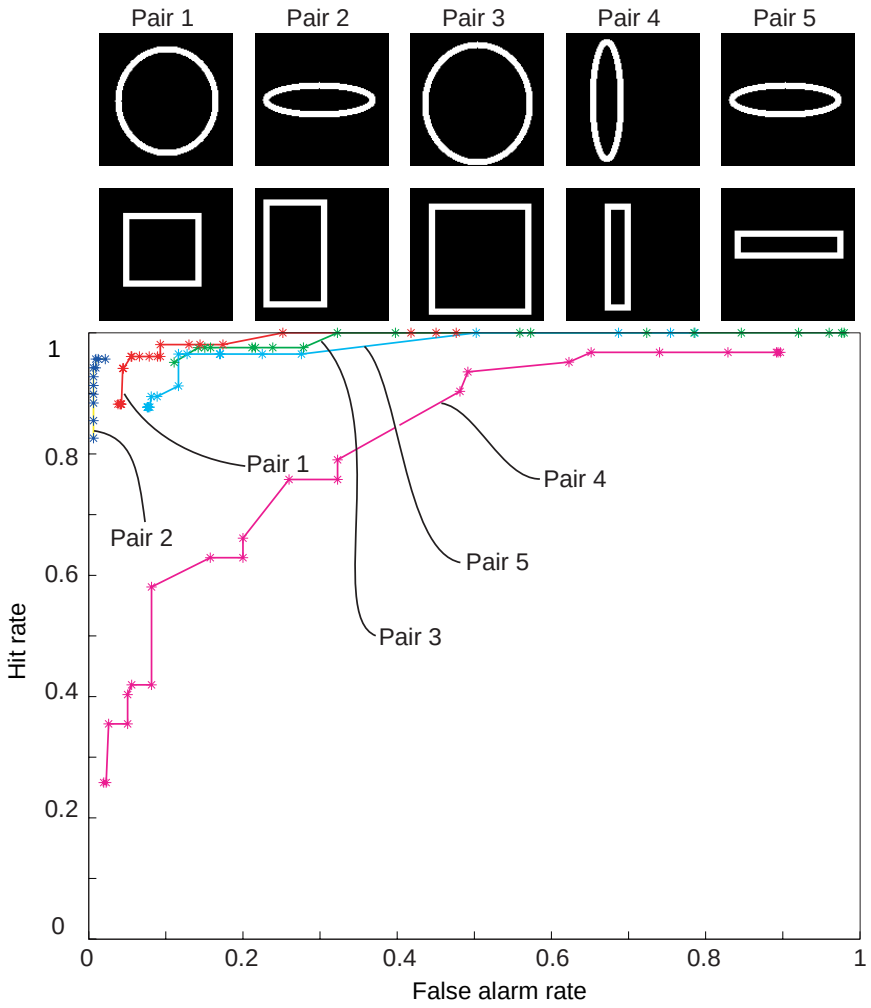


Figure 4. ROC curve for five different pairs of shapes.

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

Depending on the pair, we obtain variable discrimination results, as shown in the receiver operating characteristic (ROC) curves of **Fig. 4**. Pair No. 2 yielded yielding best discrimination, because of the strong orientation difference between both shapes. These results indicate that it is possible to obtain very good recognition performance if the patterns to be discriminated are well chosen.

On our computer systems (800MHz Pentium-III Linux machines), each 128×128 HMAX model evaluation is achieved in 6.75s. For a 1280×1280 array, attempting recognition at every location on a grid with 64×64 pixel spacing (to ensure sufficient overlap between adjacent locations to be analyzed) would hence require $\approx 2,977$s $\approx$ 50min. The computation of the saliency map, however, only takes 32s for the same 1280×1280 array, which is equivalent to the time required for ≈ 4.75 recognitions. Thus, our implementation results indicate that if the attention front-end can eliminate five recognitions of the 441 done in an exhaustive search, the combined model will perform faster than the exhaustive search.

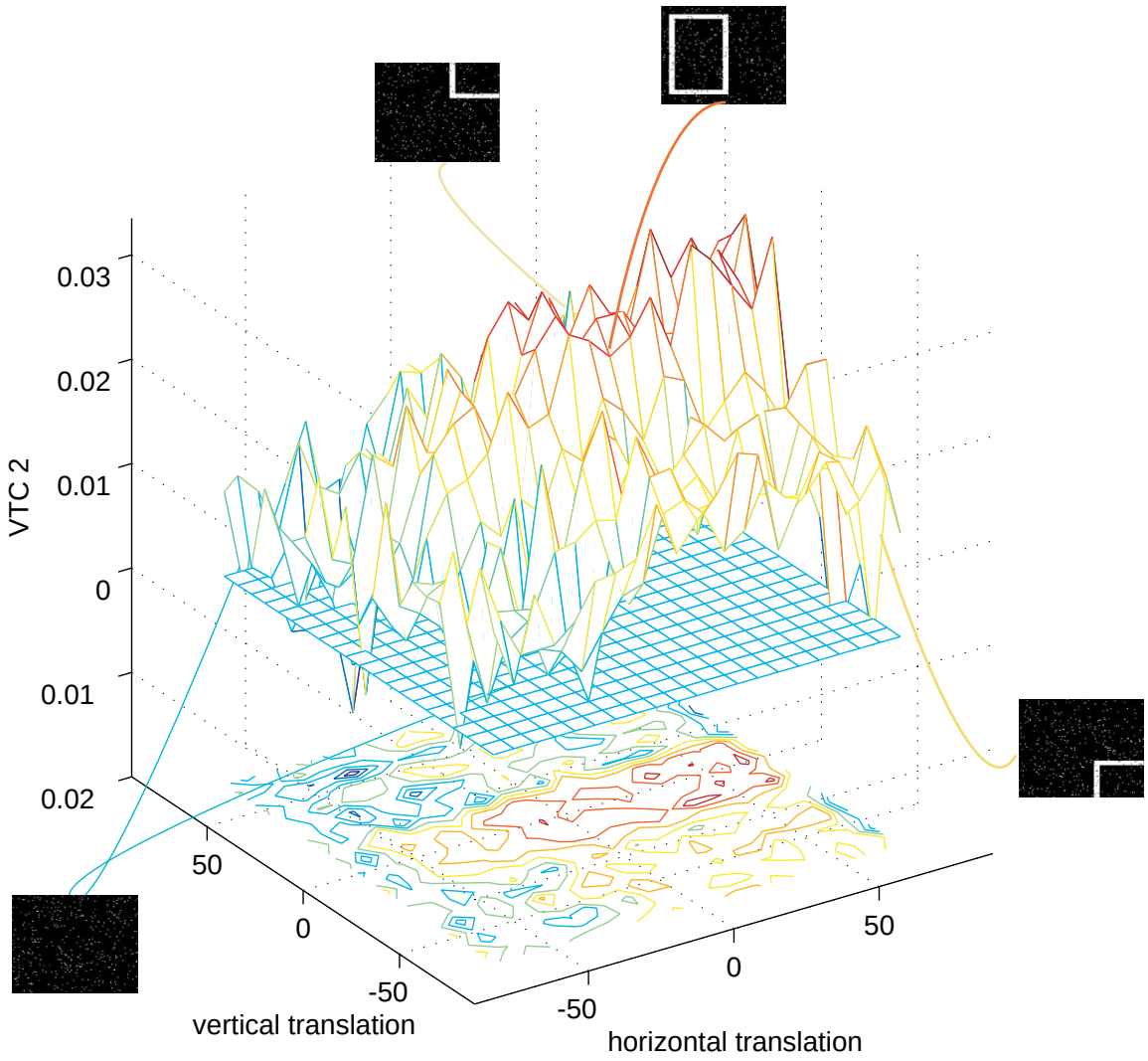


Figure 5. Responses of HMAX model when the input is translated. Recognition is robust (VTC2 response > 0) for almost the whole range of ± 70 pixels of horizontal and vertical translation studied.

3.4. 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 cannot be guaranteed that attention will always exactly be centered onto the object, but variations of ± 16 pixels in the coordinates of the focus of attention can reasonably be expected, due to noise or sampling artifacts.

We trained the HMAX model using pair No. 2, and evaluated how the system can recognize translated and rotated versions of the images in pair No. 2. Translations of -70 to 70 pixels were systematically applied to the rectangular shape **(Fig. 5)**, as well as rotations of 0 to 180° **(Fig. 6)**, in a manner similar to that previously described with HMAX model.[20] Overall, we observed 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

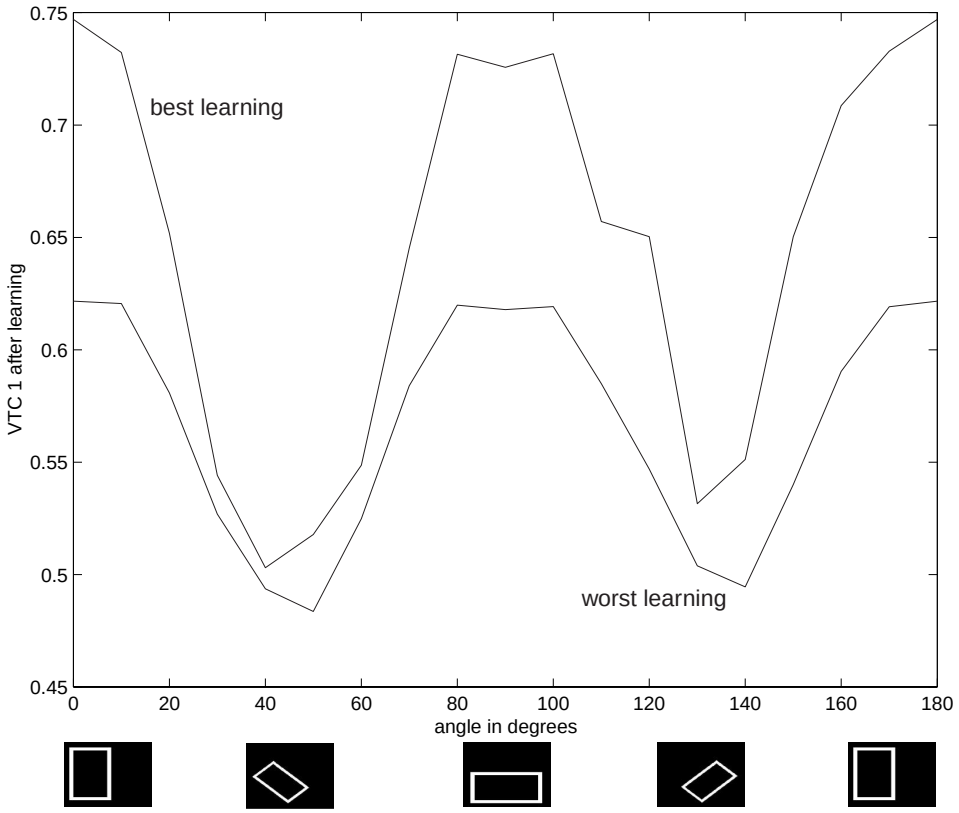


Figure 6. Dependence of the HMAX model response on stimulus rotation.

obtained with search arrays **(Fig. 4)**.

Finally, to evaluate robustness to variations in target shape, we trained the HMAX model using pair No. 2 and evaluated how the system could recognize the squares and circles from the other pairs **(Fig. 7)**. The results for pair No. 2 are identical to previously, and good results are also obtained for pair No. 5, for which the shapes are similar to those in pair No. 2. The results on the other pairs indicate that good hit rates can be obtained if one is to tolerate some false alarms.

3.5. Discussion

Through the integration of the attention and SVM-based models, we have demonstrated a complete artificial vision system, which integrates a rapid "where" attention-based orienting component, as well as a more computationally expensive "what" object recognition component. Integrating both components yielded a substantial decrease in processing time, by a factor three, at the cost of only slightly degrading target detection rate, by less than 5%.

Integrating saliency information to our object detection framework however remains imperfect in several ways: First, the center of each salient region is often not at the exact center of the target person (since one part of the person may be more salient than another), hence requiring that a relatively large window of possible target locations be examined around each attended location. We are currently investigating a refined version of our front-end, which would also provide an estimate of target size as well as more precise object location. Such improvement should allow us to reduce the search window around each attended location as well as to reduce the number of spatial scales examined by the object recognition back-end for each candidate target. The manner in which biological systems, which also appear to encode salience in relatively crude topographic maps,[11,27] solve this problem is not fully understood, though neuronal models have been proposed for translation- and scale-invariant object recognition.[28] Second, the lack of improvement in detection rate when using the 5, 6 or 7 most salient locations suggests that non-salient targets

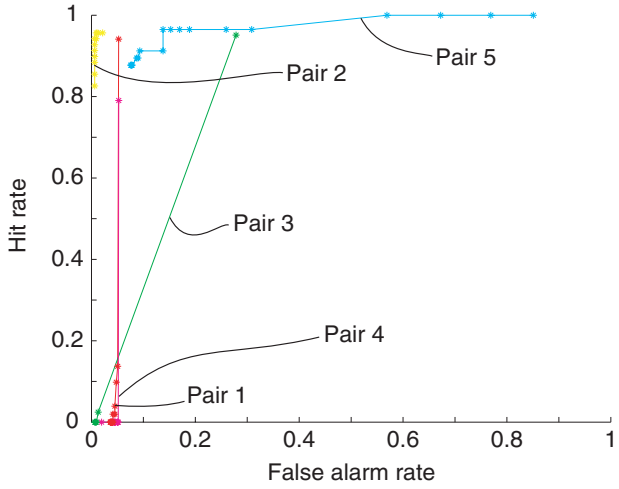


Figure 7. ROC curve for five different pairs of shapes after learning on pair No 2.

will not be visited by the attentional system unless a large number of attentional shifts is retained. Considering more attentional shifts however reduces the appeal of our approach in terms of decreasing computation time. One solution to this problem would be to integrate a larger number of features into the front-end, which is likely to render the supervised training more efficient.[16] A complementary approach, directly inspired from primate behavior, would be to integrate a top-down, volitional component to our purely bottom-up attention focusing mechanism, which could actively guide attention towards likely target locations as increasing amounts of information is gathered about the topography and organization of the scene.

Our results with the HMAX model indicate that fast and reliable target detection could be obtained on simple stimuli, using a fully biologically-inspired system. Although not yet as performant as the system using the SVM-based back-end, this approach appears very promising because it lends itself to further developments in which the early stages of visual processing may be shared by the attention and recognition components. A challenge for future research will be to extend HMAX model to natural color images. Preliminary testing with grayscale images containing pedestrians suggested that the current version of HMAX model has difficulty discriminating between pedestrians and the wide variety of other salient locations attended to by the attention model.

4. CONCLUSION

In conclusion, we have demonstrated two full, detailed implementations of a complete target detection system inspired from biological architectures. 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 research was supported by startup funds from the Charles Lee Powell Foundation and the USC School of Engineering, by the National Eye Institute, and by the National Science Foundation. We thank Maximilian Riesenhuber for help with the HMAX model, and Tomaso Poggio and Christof Koch for excellent advice and discussion.

REFERENCES

1. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit Psychol* **12**, pp. 97–136, Jan. 1980.
2. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol* **4**(4), pp. 219–27, 1985.

3. J. M. Wolfe, "Guided search 2.0: a revised model of visual search," *Psychonomic Bull Rev* **1**, pp. 202–238, 1994.
4. J. K. Tsotsos, "Computation, pet images, and attention," *Behavioral and Brain Sciences* **18**(2), p. 372, 1995.
5. L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience* **2**, pp. 194–203, Mar. 2001.
6. B. C. Motter and E. J. Belky, "The guidance of eye movements during active visual search," *Vision Res* **38**, pp. 1805–15, June 1998.
7. D. Noton and L. Stark, "Scanpaths in eye movements during pattern perception," *Science* **171**, pp. 308–11, Jan. 1971.
8. S. Egner, L. Itti, and C. R. Scheier, "Comparing attention models with different types of behavior data," in *Investigative Ophthalmology and Visual Science (Proc. ARVO 2000)*, vol. 41, p. S39, Mar. 2000.
9. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research* **40**, pp. 1489–1506, May 2000.
10. J. Wolfe, "Visual search: a review," in *Attention*, H. Pashler, ed., London, UK: University College London Press, 1996.
11. J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature* **391**, pp. 481–4, Jan. 1998.
12. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modeling visual-attention via selective tuning," *Artificial Intelligence* **78**(1-2), pp. 507–45, 1995.
13. L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of visual behavior*, D. G. Ingle, M. A. A. Goodale, and R. J. W. Mansfield, eds., pp. 549–586, MIT Press, Cambridge, MA, 1982.
14. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, pp. 1254–1259, Nov. 1998.
15. M. I. Posner and Y. Cohen, "Components of visual orienting," in *Attention and Performance X*, H. Bouma and D. G. Bouwhuis, eds., pp. 531–556, L. Erlbaum, Hilldale, NJ, 1984.
16. L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging* **10**, pp. 161–169, Jan. 2001.
17. S. Treue and J. C. Martinez Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature* **399**, pp. 575–579, June 1999.
18. C. Papageorgiou, T. Evgeniou, and T. Poggio, "A trainable pedestrian detection system," in *Intelligent Vehicles*, pp. 241–246, Oct. 1998.
19. C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision* **38**, pp. 15–33, 2000.
20. M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat Neurosci* **2**, pp. 1019–1025, Nov. 1999.
21. S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, pp. 674–93, July 1989.
22. B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifier," in *Proc. 5th ACM Workshop on Computational Learning Theory*, pp. 144–152, (Pittsburgh, PA), July 1992.
23. T. Vetter and V. Blanz, "Estimating coloured 3D face models from single images: An example based approach," in *Proceedings of the European Conference on Computer Vision ECCV'98*, (Freiburg, Germany), 1998.
24. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning*, 1998.
25. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
26. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**(2), pp. 121–167, 1998.
27. K. G. Thompson and J. D. Schall, "Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex," *Vision Res* **40**(10-12), pp. 1523–1538, 2000.
28. B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A multiscale dynamic routing circuit for forming size- and position-invariant object representations," *J Comput Neurosci* **2**, pp. 45–62, Mar. 1995.