Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2009-046
CBCL-279
October 2, 2009

Attentive processing improves object recognition

Sharat Chikkerur, Thomas Serre, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Attentive processing improves object recognition

Sharat Chikkerur, Thomas Serre and Tomaso Poggio
Center for Biological and Computational Learning, MIT

Abstract

The human visual system can recognize several thousand object categories irrespective of their position and size. This combination of selectivity and invariance is built up gradually across several stages of visual processing. However, the recognition of multiple objects in cluttered visual scenes presents a difficult problem for human as well as machine vision systems. The human visual system has evolved to perform two stages of visual processing: a pre-attentive parallel processing stage, in which the entire visual field is processed at once and a slow serial attentive processing stage, in which a region of interest in an input image is selected for "specialized" analysis by an attentional spotlight. We argue that this strategy evolved to overcome the limitation of purely feed forward processing in the presence of clutter and crowding. Using a Bayesian model of attention along with a hierarchical model of feed forward recognition on a data set of real world images, we show that this two stage attentive processing can improve recognition in cluttered and crowded conditions.

1 Introduction

Object recognition is an important function of the visual system and the brain has areas that specialize in this task. The capability of the human visual system supersedes computer vision systems in many respects. In particular, (i) The visual system can recognize several thousand categories of objects [Biederman, 1987]. (ii) Objects can be recognized despite changes in position, size, viewpoint and illumination [Wallis and Rolls, 1997, Biederman, 1992, Riesenhuber and Poggio, 1999b]. (iii) Objects can be located and recognized under clutter and in presence of other distracting objects [Serre et al., 2005]. While the mechanism for how the brain recognizes complex objects is well understood, how the visual system handles translation and clutter invariance is still an open question. In this work, we attempt to shed light on the latter two issues.

1.1 Background

In the brain, visual processing proceeds among two concurrent, yet interconnected streams. The ventral ("what") stream comprises a set of hierarchically organized regions (V1-IT) that process identity of objects ("what") and the dorsal ("where"') stream processes location and motion information [Ungerleider and Haxby, 1994]. Studies examining the tuning properties of these neurons have shown that the complexity of the preferred stimulus increases as we go further along the ventral stream: from simple oriented bars in area V1 [Hubel and Wiesel, 1959], curvature in V2 [Hegde and Van Essen, 2000, Ito and Komatsu, 2004], simple shapes in V4 [Pasupathy and Connor, 2001, Desimone and Schein, 1987, Gallant et al., 1996] and finally to object selective cell in area IT [Tanaka et al., 1991, Tanaka, 1996]. This gradual increase in selectivity provides a convincing explanation to how the brain recognizes complex objects, but cannot explain how it can do so in a location invariant manner.

In their seminal study, Hubel and Wiesel [1959] observed the existence of simple and complex cells in the striate cortex. Simple cells were found to be sensitive to both position and orientation of the stimulus while complex cells exhibited the same tuning with greater invariance to position. It was suggested that a complex cell achieves translation invariance by pooling responses of simple cells. Several neurally possible operations for pooling have thus far been proposed. It was suggested that the complex cell pools the energy (squared and rectified response from simple cells) providing translation invariance [Carandini et al., 2005]. Alternatively Riesenhuber and Poggio [1999b] suggested *max* pooling is more consistent with physiology (verified through subsequent studies [Lampl et al., 2004]). Another advantage of max-pooling is its robustness to clutter [Riesenhuber and Poggio, 1999a, Riesenhuber, 2005]. As long as the response of the distractors are lower than that of the preferred stimulus, max-pooling ensures that the distractors are ignored. Foldiak [Foldiak, 1991] proposed a computational model using the trace rule that explained how max pooling might develop in neurons.

Oram and Perret suggested that pooling mechanism between simple and complex cell may be extended to higher regions of the ventral stream to achieve invariant recognition. Based on Oram and Perret's conceptual proposals, Serre et al. [Serre et al., 2005] showed that a computational model of the ventral that gradually builds up specificity through composition and invariance through max-pooling can account for responses of V4 [Cadieu et al., 2007] and IT neurons [Hung et al., 2005] in the ventral stream. Since then, several computational models [Serre et al., 2007a, Ranzato et al., 2007, Mutch and Lowe, 2006, Wersing and Korner, 2003] have emulated the hierarchical feed forwardorganization of the ventral stream to achieve near human performance in object categorization on isolated objects.

Behavioral studies have shown that humans perform well under rapid object categorization task even when the presentation times are as small as 150ms [Fabre-Thorpe et al., 2001]. Under these conditions, visual processing is assumed to be purely feed-forward with no time for attentive processing. However, follow up studies showed that human performance degrades under clutter. This has been observed in physiological studies in area V4 [Desimone, 1998] and IT [Zoccolan et al., 2007] as well as human psychophysics [Serre et al., 2007b]. Computational models of object recognition [Walther and Koch, 2007] also exhibit this behavior. Thus, susceptibility to clutter seem to be symptomatic of feed-forward architectures. It is surprising then, that the human visual system can deal with visual complexity of the real work despite limitation presented by its feed-forward architecture. In this work, we propose that the apparent limitation of feed forward architecture is overcome with the help of visual attention, the ability to focus('attend') to behaviorally relevant stimuli while ignoring clutter.

1.2 Prior related work

The interaction between object recognition and attention is not well understood. Several conceptual proposals have been made thus far. Anderson and Essen [1987] suggested the existence of *shifter* or routing circuits that provided the higher regions in the ventral stream with an object centric co-ordinate frame that allowed invariant recognition. They suggested that the control signals for these dynamic circuits may be derived in a stimulus driven bottom-up fashion or through top-down volitional control. This model accounts for both invariant recognition and the role of attention. However, the neural correlations of this mechanism have remained elusive. Rensink [2000] proposed a triadic architecture where the image is processed in parallel to derive the gist of the scene. Also in parallel, local analysis is used to identify proto-objects without labeling them. In the second stage, attention is serially used to bind the proto-objects together into objects. Finally, local analysis is used to identify the attended object. Attention is serially shifted within the image. Navalpakkam and Itti proposed a fourth stage, where image driven bottom-up saliency is used to drive the shift of attention [Navalpakkam et al., 2005]. A similar proposal is made in scan path theory [Noton and Stark, 1971]. It was suggested that the image is first analyzed in parallel to obtain the gist of the scene. The scene gist and the behavioral task is then used to reason about the location of objects. Eye-movements are then made to the planned positions so that they can be analyzed at higher resolution through the fovea. Evidence for the scanpath theory comes from neural recording in areas FEF and LIP that are connected to eye-movements. It was found that the response of neurons in this area are proportional to the likelihood of target [Bisley and Goldberg, 2003]. The gist hypothesis is supported by evidence showing that humans can extract the gist of the scene in a very short time interval [Biederman et al., 1982]. Furthermore, the extraction of gist informations seems to proceed independently of object recognition [Schyns and Oliva, 1994]. Thus, experimental evidence suggests that visual perception seems to proceed in at least two stages: A pre-attentive parallel processing stage, in which the entire visual field is processed at once and a slow serial attentive processing stage, in which a region of interest in an input image is selected for "specialized" analysis by an attentional spotlight.

Previous studies have empirically shown that attention aids in recognition. Rutishauser et al. [2004] showed that restricting feature extraction and recognition to visually salient regions improves recognition. Miau and Itti [2001] showed that a cascade arrangement comprising bottom-up saliency and feed-forward recognition improves accuracy. In this approach bottom-up saliency is used to sequentially generated regions of interest (of fixed size), each of which is scrutinized using an object recognition system. Walther et al. [Walther and Koch, 2007] extended this approach to proto-objects obtained from bottom-up saliency map. Each segmented proto-object is analyzed using a feed-forward model of object recognition. They showed that this parallel-serial approach alleviates effects of crowding and clutter. However, these models assume that the attention is driven entirely by the visual scene in a bottom-up fashion. This ignores previous studies that have shown that scene context [Torralba, 2003] and the search task [Yarbus, 1967] influence the shift of attention (and eye-movements) exhibited by humans.

2 Our approach

In this work, we describe a two-stage approach to recognizing objects in clutter. Our work is based on a Bayesian model of visual attention that integrates top-down location and feature priors with image driven bottom-up information. It has been previously used to explain physiological effects of attention as well as predict human eye-movements [Chikkerur et al., 2009]. In this work, we integrate it with a feed-forward model of the ventral stream permitting object recognition under clutter.

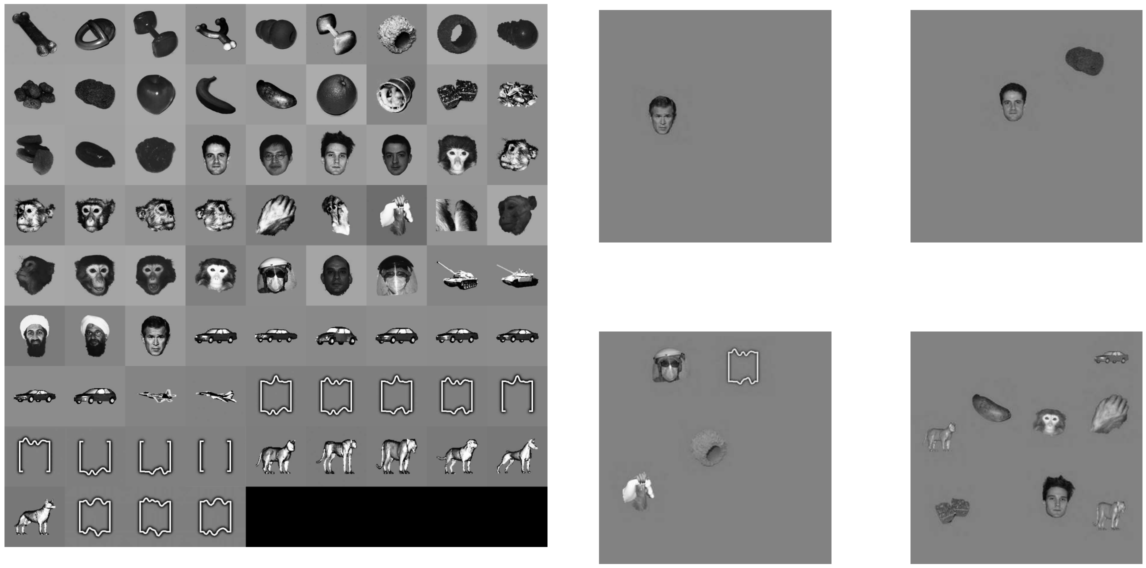
2.1 Model of attention

The Bayesian model of attention we use to generate the spotlight of attention is described in more detail elsewhere (see [Chikkerur et al., 2009] for details). Here, we briefly describe the model for the sake of completeness. Human and animal studies (see [Wolfe, 2007] for a recent review) have isolated at least three main components used to guide the deployment of an attentional spotlight: **(1)** Studies have shown that image-based *bottom-up* cues can capture attention, particularly during free viewing conditions. **(2)** Task dependence also plays a significant role in visual search [Wolfe, 2007]. Evidence for *top-down feature-based* attention comes from both imaging studies in humans [Kanwisher and Wojciulik, 2000] as well as monkey electro physiology studies [Maunsell and Treue, 2006]. **(3)** Structural associations between objects and their locations within a scene or *contextual cues*, have been shown to play a significant role in visual search and object recognition [Torralba, 2003]. How the visual system combines these cues and what the underlying neural circuits are, remain however largely unknown.

We use a Bayesian framework in which, bottom-up information serves as evidence and top-down spatial and feature cues serve as priors. The computational model is inspired by a Bayesian model of spatial attention proposed by Rao [Rao, 2005]. The main addition to the model is the inclusion of cortical feedback within ventral stream (providing feature-based attention) and areas from parietal cortex (providing spatial and context-guided attention). Feature priors bias the saliency map towards locations that share features with the target object while suppressing other location. Spatial priors can be used to bias the saliency map towards locations most likely to contain the object. The posterior probability of the location serves as as a task-dependent saliency map. When the location and feature priors are uniform, the system relies purely on visual bottom-up cues. Visual saliency is implicitly provided by divisive normalization.

2.2 Object recognition under clutter

In order to locate a target object within clutter, the model deploys feature-based attention in parallel at all locations via cascade of feedbacks from within ventral stream. This biases the saliency map in parietal cortex towards task-relevant locations (e. g., towards face targets if we are looking for faces or red targets if looking for red targets, etc). An attentional spotlight is created at the location with the highest saliency. Spatial attention is now used to enhance activation around the spotlight of attention and while suppressing it everywhere else. Feed forward processing now proceeds as if there was an isolated stimulus within the visual field. Thus, the overall effect of attentional processing is to isolate the target object while suppressing the clutter around it. In this study, we demonstrate the effectiveness of the approach on two sets of stimuli. Firstly, we show results on isolated stimuli where attention is purely feature-driven. Next, we apply the extended model on a animal vs. non-animal discrimination task. In natural images, other attentional cues such as saliency and context plays an important role. We use contextual information provided by the scene to bias the location and size where an animal is likely to be found. In both cases, we show that compared to purely feed-forward processing, a two-stage attention driven process improves recognition accuracy.



(a) Individual objects used to create the stimulus. (b) Example stimuli with one,two,four and eight objects.

Figure 1: Simple stimuli dataset

3 Simple stimuli

Hung et al. [2005] demonstrated neurons in area IT code for objects in translation and size invariant manner. A classier trained on responses of neurons can identify the category of the object over a wide range of position and sizes. They also demonstrated that a feed-forward computational model of the ventral stream can replicate invariance and recognition comparable to neurons in area IT. We used the same set of stimuli used in experiments by Hung et al.(see Fig. 1(b)). The dataset consists of 76 images belonging to 8 image categories (food, faces, monkey faces, hands, vehicles, lines and toys). In order to determine the behavior of the feed forward model under clutter, we generated images with 1,2,4 and 8 objects (see Fig. 1(b)) with object placed at random locations. The scale of the objects was chosen to be between 0.25x and 0.5x the image used in [Hung et al., 2005]. Because of the random placement, the images do not contain any contextual information and search has to proceed based on features (appearance) alone. We generated 50 images containing the target and 50 images where the target was absent. We repeated this process for each target category and cardinality thus generating 800×4 images in all.

3.1 Training

The feed-forward recognition system was trained with images of isolated objects. Additional training images were generated using translation and scaling of the original images. The recognition system uses ≈ 100 shape-based features trained using linear SVM classifier. A separate classifier is trained for each category (one vs. all).

In order to train the attention system, we selected a subset of of these features using a mutual information driven process [Fleuret, 2004]. The feature selection is done in order to minimize the computational complexity of the attention system. Wolfe's [Wolfe, 2007] guided search model provides a conceptual framework for many computational models for feature-based attention. In this model, several feature-maps computed in parallel are combined using a linear weighted combination. The weights enhance the features shared with the target object while suppressing those that are absent. Navalpakkam and Itti [Navalpakkam and Itti, 2006] used signal-to-noise ratio of each feature to bias the saliency map. In contrast, within our Bayesian framework, the preference for individual features is provided by the top-down prior. To train the attention model (determine $P(F^i|O)$, where F^i represents the i^{th} shape feature and O represents the object),

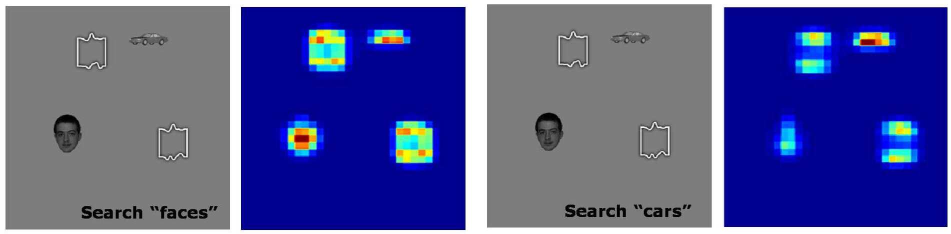


Figure 2: The figure illustrates the dependence of the search task on saliency computation: The feature-based attention biases the saliency map towards the search target. The top-down feature biases override the bottom-up evidence from the image.

we compute feature maps for each of the training image. The feature maps are discretized to maximize classification accuracy [Fleuret, 2004]. The feature F^i is said to be present if its detected at any location in the feature map. $P(F^i|O)$ is determined by simply counting the frequency of occurrence of each features within the training set.

3.2 Recognition

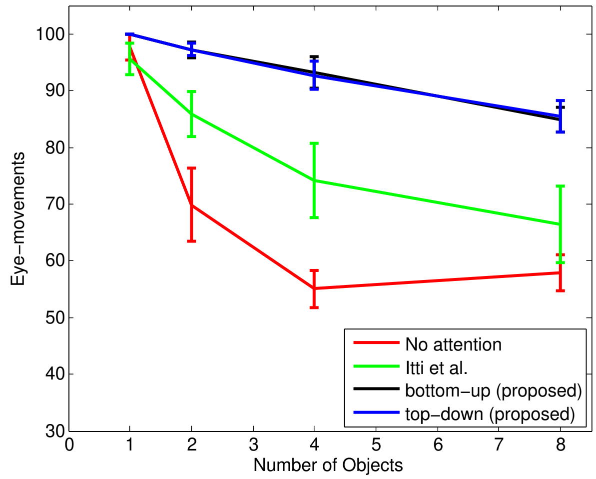
For recognition without attention, the entire image is processed using the feed-forward model. The target was said to be detected if the classifier output using the global feature was positive. In case of recognition with attention, the Bayesian model of attention is used to generate spotlights of attention. For each such spotlight, the region within the spotlight of attention is isolated and classified. If the target is not found, the current location and its surrounding are inhibited and attention is shifted to the next most salient location and detection is attempted again. The attentional shift is terminated when the saliency falls below a fixed threshold (10% of maximum) or when the target object is detected. It is to be noted that the saliency maps are task dependent (see Fig. 2). In our case, we have eight such saliency maps corresponding to each target object. Furthermore it is to be noted that mechanisms for selecting the size of the attentional spotlight is not well known. Computational models of attention have relied up fixed size [Itti et al., 1998] or adapting it to the image whereby the attentional spotlight is extended to proto-object boundaries [Walther and Koch, 2007]. The matter or visual acuity further confounds this choice. In this work, we assume the attentional spotlight to be of a fixed size.

3.3 Results

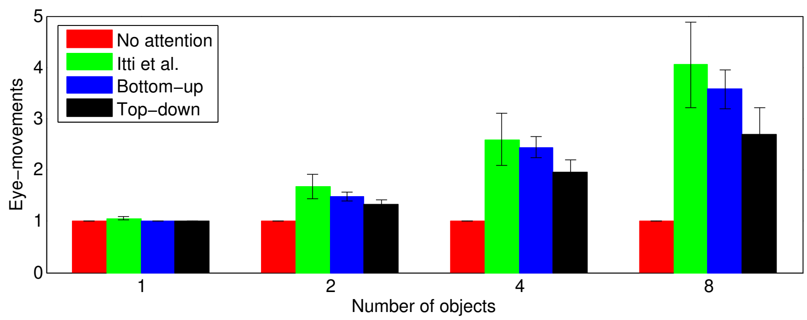
We use detection accuracy to quantify the detection performance. The simulation (see Fig. 3) shows that the detection performance degrades with the increasing the number of objects and approaches chance level at eight objects underscoring the effect of clutter. However, with attentional processing, the effect of clutter is less pronounced and recognition performance appears to be independent of the number of objects. The results show that the proposed top-down feature based attention performs better than a purely bottom-up attention in locating the objects of interest. Bottom-up attention using simple oriented features [Itti et al., 1998] performs the worst. More surprisingly, purely bottom-up attention using the shape-based feature performs just as well as the top-down attentional processing. This may be reconciled by examining the number of attentional shifts required to find the target object. Consider the case when the image has eight objects–one of which is the target. On an average, top-down feature based attention can locate the objects of interest in 2.6 shifts of attention. Whereas, the shape-based bottom-up approach requires 3.57 shifts of attention on an average. In this particular case, only feature-based priors are useful for the search task. In more complex real world images, the location and size of an object is strongly associated with the visual context and can be utilized to derive location priors.

4 Complex stimuli

The limit of feed-forward recognition in humans has been explored using a task based on discriminating images containing animals [Serre et al., 2007b]. The data-set consists of 600 images containing one or more animals and 600 distractor images



(a) Performance: Under purely feed forward conditions, the detection accuracy degrades when the number of objects within the image increases. However, recognition performance improves when attention is used.



(b) Search efficiency: The average number of attentional shifts required to find the target increases with the number of distractors. For a given number of distractors, top-down feature-based attention using shape features is the most efficient.

Figure 3: Performance measure for feature-based attention

comprising of natural and artificial objects (see Fig. 4). The studies showed that humans can perform this task well even when presentation time is as low as 50ms. In contrast to simple stimuli where the sample/set size is well defined [Wolfe, 2007], quantifying clutter in complex natural images is not straightforward although attempts have been made [Rosenholtz and Mansfield, 2005]. Serre et al. [2007b], segregated the image into four categories based on the amount of clutter and relative size of the animal with respect to background (The categories are ”head”,”near body”,”medium body” and ”far-body”). The presentation time was varied from 50ms to 80ms. It is assumed that 50ms is insufficient for attentive processing (based on the neural processing delay). Under these constrained conditions, the performance goes down with the amount of background clutter. When the mask time is increased to 80ms or removed altogether, the recognition performance in the cluttered category is found to improve. We argue that this improvement can may be attributed to attentive processing. To verify this hypothesis, we extended the feed-forward object recognition model [Serre et al., 2007a] to include attentional processing. Using the Bayesian model of attention described earlier, we combine top-down feature and spatial priors with evidence from the image to generate a saliency map is then used to generate spotlights of attention around the likely locations of animals. In the following we describe how the top-down feature and spatial priors are derived.

Figure 4: The animals vs. non-animals dataset used in [Serre et al., 2007b].The images in the dataset are divided into four categories with the depth of view and the amount of clutter increasing along the rows. The distractor non-animal images are matched for depth.

4.1 Training

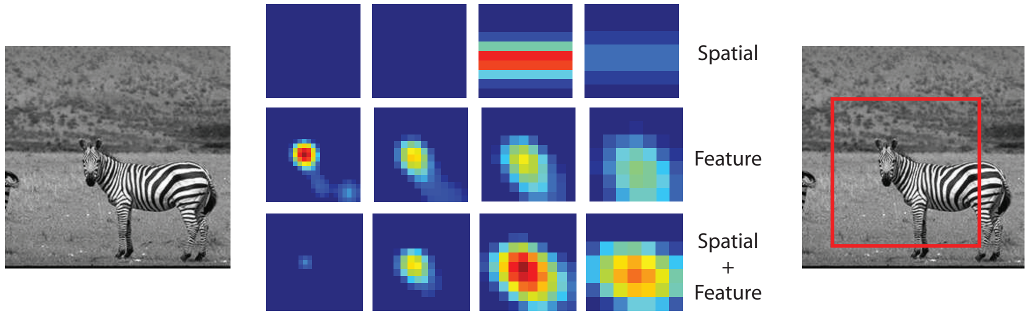
4.1.1 Feature-based attention

We use a dictionary of 64 shape-based features derived from the computational model of the ventral stream [Serre et al., 2007a]. These features were selected from an initial pool of several hundred features using mutual information based procedure outlined in [Fleuret, 2004]. The features are computed at four different scales corresponding to $0.85\times, 0.55\times, 0.36\times$ and $0.23\times$ the size of the image. Having multi-scale features allows us to determine the location as well as the size of the attentional spotlight. Using a set of 600 training images, the occurrence probability of these individual features ($P(F^i|O)$) within animal and non-animal images was learned.

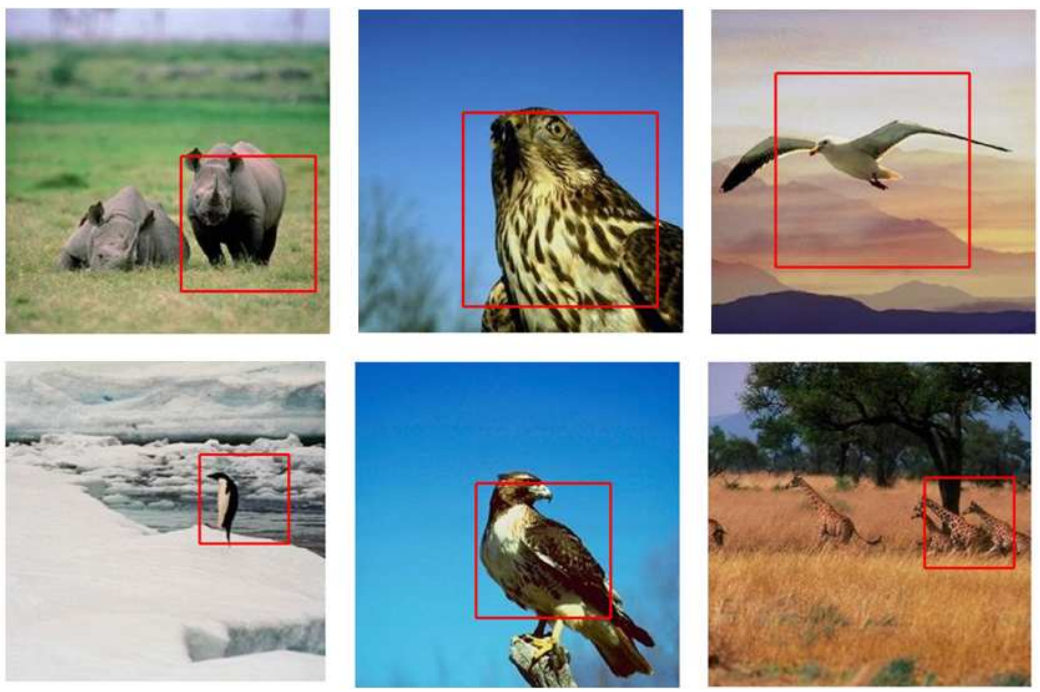
4.1.2 Spatial attention

Computer vision systems that are based on scanning [Bileschi, 2006, Dalal et al., 2006] do not make any assumption about the scene in which the objects are found. However, in the natural world, the scene and the objects within it share *contextual* relationships. Psychophysical experiments have shown that humans can reason about the location and size of objects in a scene within a very short time interval [Biederman et al., 1982]. This suggests that humans perceive the 'gist' or summary of scenes even before attending to the objects in them. Based on this observation, Torralba et al. [Torralba, 2003] proposed that a computational model of scene gist could similarly be used to reason about object probability and position prior to do object detection in computer vision applications. Previous approaches have used gist representations based on spatial distribution of oriented filter responses [Torralba, 2003, Itti et al., 2005]. In this work, we use biologically inspired shape descriptors [Serre et al., 2007a] to describe the 'shape' of the scene. The association between the image and the location (and scale) of animals was learned using a mixture of regressors [Murphy et al., 2003].

Given a vectorial scene-gist representation G, the probability of x-location, y-location and scale ($X = x, y, \sigma$) is given



(a) The figure illustrates the saliency map computation using spatial priors only,features only and a combination of both. The scale of the saliency map increases from left to right ($0.23\times$, $0.36\times$, $0.55\times$ and $0.85\times$ the size of the image).



(b) The figure illustrates the attentional windows computed using the Bayesian model on several images. Note that the size of the attention is not fixed. Instead it is determined by the scale corresponding to the most salient location.

Figure 5: Examples illustrating attentive processing on real-world natural images.

by

$$P(X|G) = \sum_k P(K|G)P(X|K,G) \tag{1}$$
$$P(X|K,G) \sim N(\mu_K + A_k^T G, \Sigma_K) \tag{2}$$
$$P(K|G) \sim \text{softmax}(K; W_k^T G) \tag{3}$$

K represents the canonical view, each of which imposes a different distribution on object location and size. A_k, μ_K represent the parameters of the individual regressors. The weights W_k and the softmax function provide a smooth transition between different constraints. $P(X|K)$ specifies the individual regressor for view K. To decrease the learning time and to avoid over fitting, we reduce the dimension of the the individual representations using PCA. We retain only the top 32

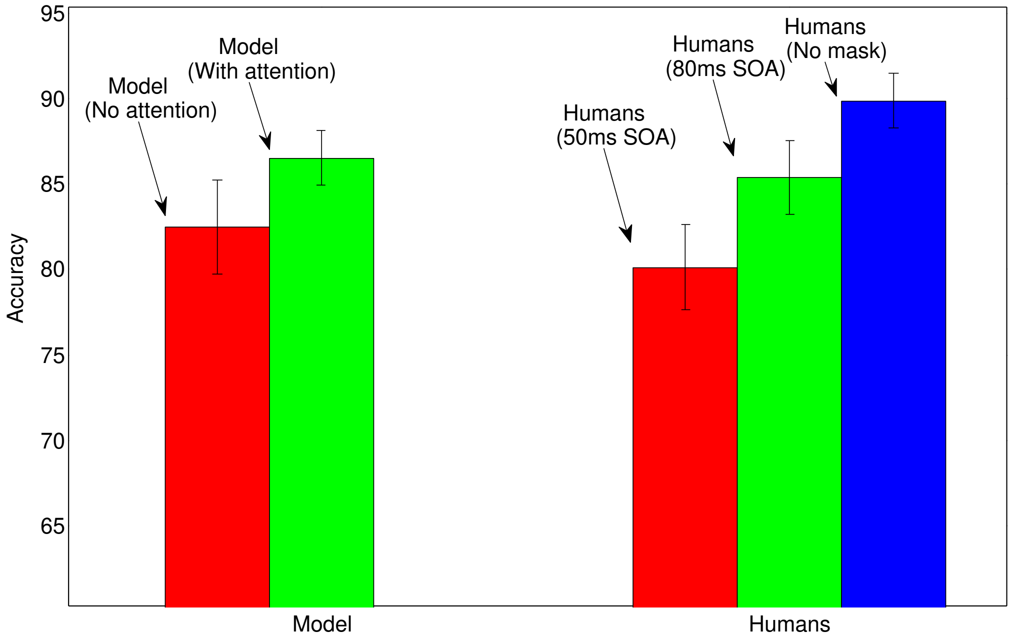


Figure 6: The performance of humans and the model on animals vs. non-animal task. The detection performance increases when processing with attention.

principal components in each case to provide a fair comparison. The number of regressors was fixed at $K = 5$ for all representations. Our informal study on the selection of K did not show any difference for higher values of K.

4.2 Recognition

Given a test image, a multi-scale saliency map is generated using the Bayesian model of attention (see Fig. 5). The most salient location and scale is used to generate a spotlight of attention (the size of the spotlight being determined by the scale). The region within the spotlight was isolated and then processed in a purely feed-forward fashion to finally decide if the image contains an animal or not. In our case, we restrict the processing to the first spotlight.

4.3 Results

The detection accuracy of the model increases from $82.5 \pm 2.75\%$ under purely feed forward condition to $86.5 \pm 1.6\%$ under attentive processing. On the other hand, human performance on the same data increases from $80.12 \pm 2.5\%$ when the presentation time is 50ms to $85.375 \pm 2.125\%$ when presentation time is increased to 80ms. When the mask is remove entirely, the performance improves further to $89.875 \pm 1.625\%$. Results suggest that the performance of the extended model using a single spotlight of attention increases in a manner similar to performance of human subjects who were given additional time to process the images. This suggests that attention may play an important role in recognition under clutter. However, the no-mask condition cannot be explained by attentive processing alone. The human visual system may rely upon higher level information that is not available to the model.

5 Conclusion

Although feed forward recognition system can achieve invariant recognition, they are susceptible to crowding and clutter. We argued that this problem can be overcome using attentive processing, a two stage process, where attention is used to isolate the object of interest followed by feed forward recognition. We demonstrated this approach using a bayesian model of attention followed by a hierarchical feed forward model of recognition. We evaluated the approach on two datasets of varying difficulty–a simple dataset comprising of isolated objects where attention can be driven to targets based on features alone [Hung et al., 2005] and data set comprising of animal and non-animal images where attention to the target object is driven context driven spatial cues in addition to feature information. In both cases, we showed that attentive processing improves recognition when compared to purely feed forward processing.

References

C. H. Anderson and D. C. V. Essen. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences*, 84(17):6297–6301, 1987.

I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological review*, 94(2):115–147, 1987.

I. Biederman. Human image understanding. In *Theory and Applications of Image Analysis: Selected Papers from the 7th Scandinavian Conference on Image Analysis*, page 3. World Scientific Pub Co Inc, 1992.

I. Biederman, RJ Mezzanotte, and JC Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143, 1982.

S. M. Bileschi. *StreetScenes: Towards scene understanding in still images*. PhD thesis, MIT, 2006.

J.W. Bisley and M.E. Goldberg. Neuronal activity in the lateral intraparietal area and spatial attention. *Science*, 299(5603): 81–86, 2003.

C. Cadieu, M. Kouh, A. Pasupathy, C.E. Connor, M. Riesenhuber, and T. Poggio. A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, 98(3):1733, 2007.

M. Carandini, J.B. Demb, V. Mante, D.J. Tolhurst, Y. Dan, B.A. Olshausen, J.L. Gallant, and N.C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

S. Chikkerur, T. Serre, Tan C., and T. Poggio. An integrated framework of visual attention using shape-based features. *CBCL, MIT Paper*, July 2009.

N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441, 2006.

R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society*, 1998.

R. Desimone and S. J. Schein. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835–868, 1987.

M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A Limit to the Speed of Processing in Ultra-Rapid Visual Categorization of Novel Natural Scenes. *Journal of Cognitive Neuroscience*, 13(2), 2001.

F. Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.

P. Foldiak. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology*, 76(4):2718–2739, 1996.

J. Hegde and D.C. Van Essen. Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, 20(5): 61–61, 2000.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J Physiol*, 148:574–91, 1959.

C.P. Hung, G. Kreiman, T. Poggio, and J.J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866, 2005.

M. Ito and H. Komatsu. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience*, 24(13):3313–3324, 2004.

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(11), 1998.

L. Itti, G. Rees, and J.K. Tsotsos. *Neurobiology of attention*. Academic Press, 2005.

N. Kanwisher and E. Wojciulik. Visual attention: insights from brain imaging. *Nature Reviews Neuroscience*, 1:91–100, 2000.

I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of neurophysiology*, 92(5):2704–2713, 2004.

J.H.R. Maunsell and S. Treue. Feature-based attention in visual cortex. *TRENDS in Neurosciences*, 29(6):317–322, 2006.

F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. In *Proc. IEEE Engineering in Medicine and Biology Society (EMBS), Istanbul, Turkey*, pages 789–792, Oct 2001.

K. Murphy, A. Torralba, and W.T. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. *Advances in Neural Information Processing Systems*, 16, 2003.

J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. In *Computer Vision and Pattern Recognition*, 2006.

V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Computer Vision and Pattern Recognition*, 2006.

V. Navalpakkam, MA Arbib, and L. Itti. Attention and scene understanding. *Neurobiology of Attention*, pages 197–203, 2005.

D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971.

A. Pasupathy and C.E. Connor. Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–2519, 2001.

M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition*, 2007.

R.P.N. Rao. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843–1848, 2005.

R.A. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40(10-12):1469–1487, 2000.

M. Riesenhuber. Object recognition in cortex: Neural mechanisms, and possible roles for attention. *Chapter in: Neurobiology of Attention, L. Itti, G. Rees, and J. Tsotsos (eds), Academic Press, Elsevier*, pages 279–287, 2005.

M. Riesenhuber and T. Poggio. Are Cortical Models Really Bound Review by the Binding Problem? *Neuron*, 24:87–93, 1999a.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature, Neuroscience*, 2:1019–1025, 1999b.

R. Rosenholtz and J. Mansfield. Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 761–770. ACM New York, NY, USA, 2005.

U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *Computer Vision and Pattern Recognition*, volume 2, 2004.

P.G. Schyns and A. Oliva. Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4): 195–200, 1994.

T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex, CBCL MIT paper, November 2005, 2005.

T. Serre, Wolf L., S. Bileschi, M. Reisenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007a.

T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424, 2007b.

K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139, 1996.

K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of neurophysiology*, 66(1):170–189, 1991.

A. Torralba. Modeling global scene factors in attention. *Journal of Optical Society of America*, 20(7):1407–1418, 2003.

L.G. Ungerleider and J.V. Haxby. 'What'and'where'in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165, 1994.

G. Wallis and E.T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2): 167–194, 1997.

D.B Walther and C. Koch. *Computational Neuroscience: Theoretical insights into brain function, Progress in Brain Research*, chapter Attention in Hierarchical Models of Object Recognition. 2007.

H. Wersing and E. Korner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15(7):1559–1588, 2003.

Jeremy M. Wolfe. Guided search 4.0: Current progress with a model of visual search. *Integrated Models of Cognitive System*, pages 99–119, 2007.

A. L. Yarbus. *Eye movements and vision*. Plenum press, 1967.

D. Zoccolan, M. Kouh, T. Poggio, and J.J. DiCarlo. Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. *Journal of Neuroscience*, 27(45):12292, 2007.