# On the Optimality of Spatial Attention for Object Detection

Jonathan Harel and Christof Koch

California Institute of Technology, Pasadena, CA 91125

**Abstract.** Studies on visual attention traditionally focus on its physiological and psychophysical nature [16, 18, 19], or its algorithmic applications [1, 9, 21]. We here develop a simple, formal mathematical model of the advantage of spatial attention for object detection, in which spatial attention is defined as processing a subset of the visual input, and detection is an abstraction with certain failure characteristics. We demonstrate that it is suboptimal to process the entire visual input given prior information about target locations, which in practice is almost always available in a video setting due to tracking, motion, or saliency. This argues for an attentional strategy independent of computational savings: no matter how much computational power is available, it is in principle better to dedicate it preferentially to selected portions of the scene. This suggests, anecdotally, a form of environmental pressure for the evolution of foveated photoreceptor densities in the retina. It also offers a general justification for the use of spatial attention in machine vision.

## 1 Introduction

Most animals with visual systems have evolved the peculiar trait of processing subsets of the visual input at higher bandwidth (faster reaction times, lower error rates, higher SNR). This strategy is known as focal or spatial attention and is closely linked to sensory (receptor distribution in the retina) and motor (eye movements) factors. Motivated by such wide-spread attentional processing, many machine vision scientists have developed computational models of visual attention, with some treating it broadly as a hierarchical narrowing of possibilities [1, 2, 8, 9, 17]. Several studies have demonstrated experimental paradigms in which various such attentional schemes are combined with recognition/detection algorithms, and have documented the resulting computational savings and/or improved accuracy [4–7, 20, 21].

Here, we seek to describe a general justification for spatial attention in the context of an object detection goal (detecting targets in images wherever they occur). We take an abstract approach to this phenomenon, in which both the attentional and detection mechanisms are independent of the conclusions. Similar frameworks have been proposed by other authors [3, 10]. The most common justification for attentional processing, in particular in visual psychology, is the computational saving that accrue if processing is restricted to a subset of the image. For machine vision scientists, in an age of ever decreasing computational costs of digital processors, and for biologists in general, the question is whether there are other justifications for the *spatial spotlight of attention*. We will address this in three steps which form the core substance of this paper:

1. (Section 2) We demonstrate that object detection accuracy can be improved using attentional selection in a motivating machine vision experiment.

2. (Section 3) We model *a generalized form* of this system and demonstrate that accuracy is optimal with attentional selection if prior information about target locations is not or cannot be used to bias detector output.

3. (Section 4) We then demonstrate that, even if priors are used optimally, if there is a fixed computational resource which can be concentrated or diluted over locations in the visual scene, with corresponding modulations in accuracy, that it is optimal to process only the most likely target locations. We show how the optimal extent of this spatial attention depends on the environment, quantified as a specific tolerance for false positives and negatives.

## 2 Motivating Example

### 2.1 Experiment

An important problem in machine vision is the detection of objects from broad categories in cluttered scenes, in which a target may only take up a small fraction of the available pixels. We built a system to solve an instance of this "object detection" problem: detecting cars and pedestrians wherever they occurred in a fully annotated video of 4428 frames, captured at 15fps at VGA (640x480) resolution.

Training images (47,459 total, of which 4,957 are positive examples) were gathered from [11] and [12]. The object detection system worked in two steps for each frame independently:

1. A saliency heat map [9] for the frame (consisting of color, orientation, intensity, motion, and flicker channels) was computed and subsequently serialized into an ordered list of "fixation" locations (points) using a choose-maximum/inhibit-its-surround iterative loop. A rectangular image crop ("window") around each fixation location was selected using a crude flooding-based segmentation algorithm.

2. The first $F \in \{1, 3, 5, 7, 9\}$ fixation windows were then processed using a detection module (one for cars and one for pedestrians), which in turn decided if each window contained its target object type or not. The detection modules based their classification decision on the output of an SVM, with input vectors having components proportional to the multiplicity of certain quantized SIFT [14] features over an image subregion, with subregions forming a pyramid over the input image – this method has proven quite robust on standard benchmarks [13].

### 2.2 Results

We quantified the performance by recording four quantities for each choice of $F$ windows per frame: (1) True Positive Count (TPC) – the number of windows, pooled over the entire video[1], in which a detection corresponded to a true object at that location. (2) False Positive Count (FPC) – windows labeled as a target where there was actually not one, and using the False Negative Count, FNC (number of targets undetected), (3) precision = TPC/(TPC+FPC) – fraction of detections which were actually target objects, and (4) recall = TPC/(TPC+FNC) – fraction of target objects which were detected.

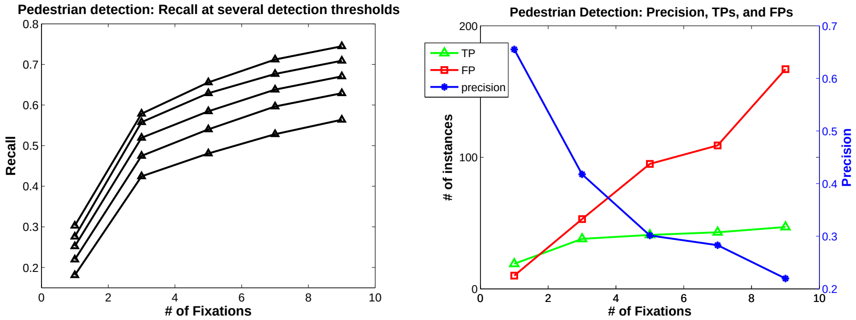

Fig. 1. Result of running detector over entire video. As the number of windows processed per frame increases, recall rate increases (left), while precision rate decreases (right). Left: curves for different settings of the SVM detection threshold.

The results for pedestrian detection are shown in Fig. 1. Results on cars were qualitatively equivalent.

Each data point in Fig. 1 corresponds to results over the pooled video frames, but at each frame the number of windows processed is not the same: we parameterize over this window count along the x-axis. All plots in this paper use this underlying *attention-parameterizing* scheme, in which processing one window corresponds to maximally focused attention, and processing them all corresponds to maximally blurred attention. The results in Fig. 1 indicate that, in our experiment, the recall rate increases as more windows are processed per frame, whereas the precision rate falls off. Therefore, in this case, it is reasonable to process just a few windows per frame, i.e., implement an attentional focus, in order to balance performance, independent of computational savings.

This can be understood by considering that lower-saliency windows are a priori unlikely to contain a target, and so their continued processing yields a false positive count that accumulates at nearly the false positive rate of the detector. The true positive count, on the other hand, saturates at a small number proportional to the number of targets in the scene. These two trends yield a decreasing precision ratio. This is seen more directly in Fig. 2 below, where we plot the average number of pedestrians contained in the first $F$ fixation windows of a frame, noting that the incremental increase (slope) per added window is decreasing. We will see in the next section how the behavior observed here is sensitive to incorporating priors into detection decisions.

[1] results shown are for 20% of the frames uniformally sampled from the video

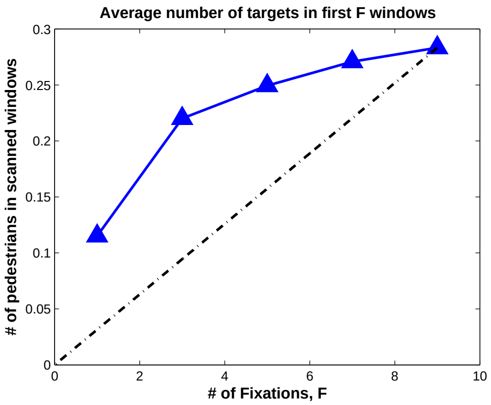

Fig. 2. The average number of pedestrians contained in the first $F$ windows. The dotted line connects the origin to the maximum point on the curve, showing what we would observe if pedestrians were equally likely to occur at each fixation. But since targets are more likely to occur in early fixations, the slope decreases.

# 3 A simple mathematical model of spatial attention for object detection

In this section, we model *a generalized form* of the system in the experiment above, and explore its behavior and underlying assumptions.

## 3.1 Preliminaries

We suppose henceforth that our world consists of images/frames streaming into our system, that we form window sets over these images, *somehow* sort these windows in a negligibly cheap way (e.g., according to fixation order from a saliency map, or due to an object tracking algorithm), and then run an object detection module (e.g., a pedestrian detector) over only the first $w$ of these windows on each frame, according to sorted order, where $w \in \{1, 2, ..., N\}$. We will refer to the processing of only the first $w$ windows as *spatial attention*, and the smallness of $w$ as the *extent of spatial attention.*[2]

We will model the behavior of a detection system as a function of $w$. Define[3]

$$\begin{aligned} T(w) &\triangleq \#\text{ targets in first } w \text{ windows} \\ FPC(w) &\triangleq \#\text{ false positives in first } w \text{ windows (incorrect detections)} \\ TPC(w) &\triangleq \#\text{ true positives in first } w \text{ windows (correct detections)} \\ FNC(w) &\triangleq \#\text{ false negatives (in entire image after processing } w \text{ windows)} \\ TNC(w) &\triangleq \#\text{ true negatives (in entire image after processing } w \text{ windows)} \end{aligned}$$

[2] see Appendix for table of parameters

[3] $C$ is for count, as in FalsePositiveCount = FPC

These counts determine the performance of the detection system, and so we will calculate their expected values, averaged over many frames. To do this, we define the following: For a single frame/image, let $T_i$ be the binary random variable indicating whether there is in truth a target at window $i$, with 1 corresponding to presence. Let $D_i$ be the binary random variable indicating the result of the detection on window $i$, with 1 indicating a detection. Then:

$$\begin{aligned}
E[T(w)] &= \sum_{i=1}^{w} E[T_i] = \sum_{i=1}^{w} p_i, \text{ where } p_i \triangleq \Pr\{T_i = 1\} \\
E[FPC[w]] &= \sum_{i=1}^{w} E[FP_i] \text{ where } FP_i = \begin{cases} 1 \text{ if } D_i = 1 \text{ and } T_i = 0 \\ 0 \qquad \text{otherwise} \end{cases} \\
&= \sum_{i=1}^{w} p(D_i = 1 | T_i = 0) \cdot (1 - p_i) = fpr \cdot (w - E[T(w)]) \\
E[TPC(w)] &= \sum_{i=1}^{w} E[TP_i], \text{ where } TP_i = \begin{cases} 1 \text{ if } D_i = 1 \text{ and } T_i = 1 \\ 0 \qquad \text{otherwise} \end{cases} \\
&= \sum_{i=1}^{w} p(D_i = 1 | T_i = 0) \cdot p_i = tpr \cdot E[T(w)]
\end{aligned}$$

Where the false and true positive rates, $fpr \triangleq p(D_i = 1 | T_i = 0)\ \forall i$, and $tpr \triangleq p(D_i = 1 | T_i = 1)\ \forall i$, are taken to be properties of the detector. Similarly,

$$E[FNC(w)] = n - E[TPC(w)], \text{ where } n \triangleq E\left[\sum_{i=1}^{N} T_i\right] = \sum_{i=1}^{N} p_i = E[T(N)]$$

$$\text{Since } \sum_{i=1}^{N} T_i = TPC(w) + FNC(w) = \#\text{ of windows with a target in image}$$

And

$E[TNC(w)] = (N - n) - E[FPC(w)]$, because:

$$N - \sum_{i=1}^{N} T_i = FPC(w) + TNC(w) = \#\text{ of windows without a target in image}$$

### 3.2 Decreasing precision underlies utility of spatial attention

We shall now use the quantities defined above to model the precision and recall trends demonstrated in the motivating example. But, first we must make a modeling assumption: suppose that $p_i$ is decreasing in $i$ such that:

$$E[T(w)] = n \frac{1 - \exp(-w/k)}{1 - \exp(-N/k)} \qquad (1)$$

which has a similar form to that in Fig. 2. Note that this yields $E[T(0)] = 0$, and $E[T(N)] = n$, as above, where $n$ represents the average number of target-containing windows in a frame. Below we plot this profile for several settings of $k$, with $n = 2$ and $N = 1000$ (more nearly continuous/graded than the motivating experiment for smoothness):

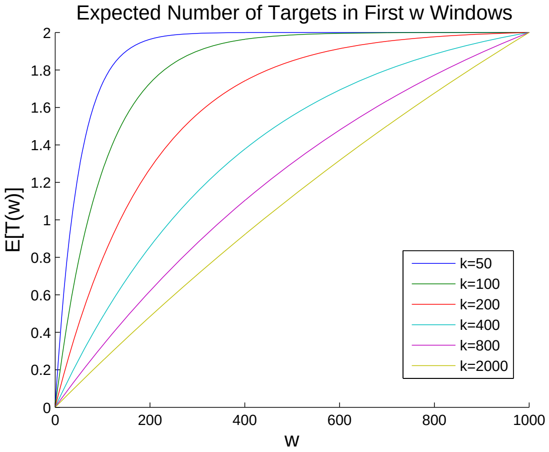

Fig. 3. A model of the average number of targets in highest $w$ priority windows.

Larger values of $k$ correspond to $E[T(w)]$ profiles which are closer to linear. Linearly increasing $E[T(w)]$ corresponds to constant $p_i$ so that $\sum_{i=1}^{w} p_i$ increases an equal amount for each increment of $w$. Concave down profiles above the line corresponding to decreasing $p_i$ profiles, in which the incremental contribution to $E[T(w)]$ from $\sum_{i=1}^{w} p_i$ is higher for low $w$. Such decreasing $p_i$ represent an ordering of windows where early windows are more likely to contain targets than later windows. In practice, one can almost always arrange such an ordering since targets are likely to remain in similar locations from frame to frame, be salient, or move, or be a certain color, etc.. Here, we are not concerned with how this ordering is carried out, but assume that it is.

Let subscript-$M$ denote a particular count accumulated over $M$ frames. As the number of frames $M$ grows,

$$\lim_{M\to\infty} T_M(w) = \lim_{M\to\infty} \sum_{image=1}^{M} T_{image}(w) = M \cdot E[T(w)]$$

by the Central Limit Theorem, where $T_{image}(w)$ is the number of targets in *image*. Using similar notation, the precision after $M$ images have been processed approaches:

$$\lim_{M\to\infty} prec_M(w) = \lim_{M\to\infty} \frac{TPC_M(w)}{TPC_M(w) + FPC_M(w)}$$

$$= \frac{M \cdot E[TPC(w)]}{M \cdot E[TPC(w)] + M \cdot E[FPC_M(w)]} = \frac{E[TPC(w)]}{E[TPC(w)] + E[FPC_M(w)]}$$

Equivalently, the recall approaches

$$\lim_{M\to\infty} rec_M(w) = \frac{E[TPC(w)]}{E[TPC(w)] + E[FNC_M(w)]}.$$

Define $prec(w) \triangleq \lim_{M\to\infty} prec_M(w)$, and $rec(w) \triangleq \lim_{M\to\infty} rec_M(w)$.

Using the model equation (1), and the equilibrium precision and recall definitions, we see that we can qualitatively reproduce the experimental results observed in Fig. 1:

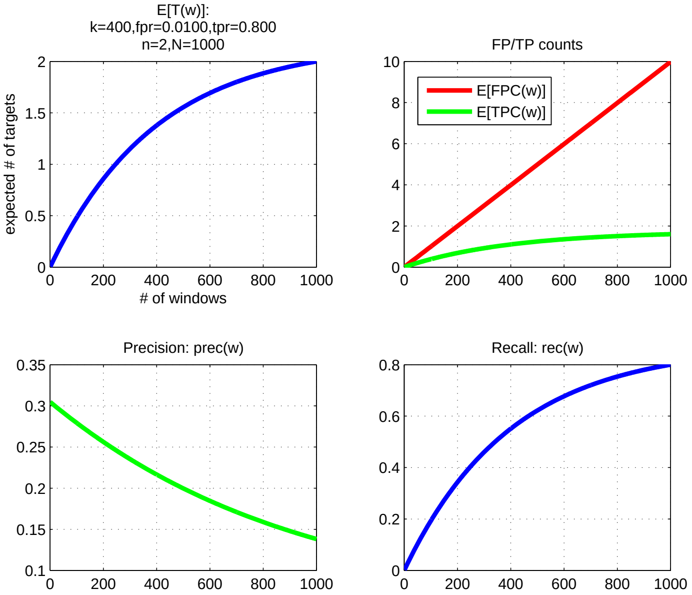

Fig. 4. Equilibrium precision and recall rates using a model $E[T(w)]$

Simulation results suggest that this decreasing precision, increasing recall holds under a wide variety of concave profiles $E[T(w)]$ (including all parameterized in (1)), and detector rates properties $(tpr, fpr)$. A few degenerate cases will flatten the precision curve: a linear $E[T(w)]$ and/or a zero false positive rate, i.e., zero ability to order windows, and a perfect detector, respectively. Otherwise, recall and precision pull performance in opposite directions over the range of $w$, and optimal performance will be somewhere in the middle depending on the exact parameters and objective function, e.g., area under ROC or precision-recall curve. Therefore, it is in this context best to process only the windows most likely to contain a target in each frame, i.e., implement a form of spatial attention.

**$tpr, fpr$ fixed $\forall i$ means having little faith in, or no ability to calculate, one's prior belief.** This model is realistic if one does not have faith in, or ability to calculate, one's prior belief: i.e., the order of windows is known, but not specifically $P(T_i = 1)$. Formally, in a Bayesian setting, one would assume that there is a pre-decision detector output $D_{ic} \in \boldsymbol{\theta}$ with constant known densities $p(D_{ic}|T_i)$. Then,

$$tpr = P(D_i = 1|T_i = 1) = \Pr(D_{ic} \in \theta^+|T_i = 1), \tag{2}$$

where $\theta^+$ is the largest set such that

$$LLR = \frac{p(D_{ic}|T_i = 1)P(T_i = 1)}{p(D_{ic}|T_i = 0)P(T_i = 0)} > 1 \; \forall D_{ic} \in \theta^+ \tag{3}$$

Very notably, the definition in (2) yields a *tpr* which is *not the same for all* $i$ (as modeled previously), and in particular, which depends on the prior $P(T_i = 1) = p_i$. Similarly,

$$fpr = P(D_i = 1|T_i = 0) = \Pr(D_{ic} \in \theta^+|T_i = 0),$$

also depends on $p_i$. Only if one assumes that $P(T_i = 1) = P(T_i = 0)$, then (3) is the same for all $i$, and so is (2). Having constant *tpr* and *fpr* $\forall i$ is also equivalent to evaluating the likelihood ratio as:

$$LLR = \left(\frac{p(D_{ic}|T_i = 1)}{p(D_{ic}|T_i = 0)}\right)^{\gamma} \frac{P(T_i = 1)}{P(T_i = 0)}$$

in the limit as $\gamma \to \infty$, or putting little faith into the prior distribution. This is somewhat reasonable given the motivating experimental example in section 2. The output of the detector is somehow much more reliable than whether a location was salient in determining the presence of a target, and the connection between saliency and probability of a target $P(T_i = 1)$ may be changing or incalculable.

Importantly, if a prior distribution is available explicitly, then the false positive counts $FPC(w)$ saturate at high values of $w$ which are unlikely to contain a target, and the utility of not running the detector on some windows is eliminated, although it still saves compute cycles.

## 4 Distributing a fixed computational resource

In the previous section, we assume that it makes sense to process a varying number of windows with the same underlying detector for each window. A more realistic assumption about systems in general is that they have a fixed computational resource, and that it can be and should be fully used to transform input data into meaningful detector outputs.

Now, suppose the same underlying two-step model as before: frames of images stream in to our system, we somehow cheaply generate an ordered window set on each of these, and select a number $w$ of the highest-priority windows, each of which will pass through a detector.

Here, we impose an additional assumption: that the more detection computations are made (equivalently, the more detector instances there are to run in parallel), the weaker each individual detection computation/detector must be, in accordance with the conservation of computational resource. Below, we derive a simple detector degradation curve, and then use it to characterize the relationship between the risk priorities of a system (tolerance for false positives/negatives) and its optimal extent of spatial attention, viz., how many windows down the list it should analyze.

### 4.1 More detectors, weaker detectors

We assume that a detector $DT$ is an abstraction which provides us with information about a target. For simplicity, suppose that it informs us about a particular real-valued target property $x$, like its automobility or pedestrianality. Then the information provided by detector $DT$ is:

$$I_{DT} \triangleq H_0 - H_{DT} \triangleq H(P_0(x)) - H(P_{DT}(x))$$

where $P_{DT}(x)$ is the density function over $x$ output by the detector, and $H_0 = H(P_0(x))$ is the entropy in $x$ before detection, where $P_0(x)$ is the prior distribution over $x$.

It seems intuitively clear that given fixed resources, one can get more information out of an aggregate of cheap detectors than out fewer more expensive detectors. One way to quantify this is by assuming that the fixed computational resource is the number of compute "neurons" $R$, and that these neurons can be allocated to understanding/detecting in just one window, or divided up into $s$ sets of $R/s$ neurons, each of which will process a different window/spatial location. There are biological data suggesting that neurons from primary sensory cortices to MTL [15] fire to one concept/category out of a set, i.e. that the number of concepts encodable with $n$ neurons is roughly proportional to $n$, and so the information $n$ neurons carry is proportional to $\log(n)$. Thus, a good model for how much information each of $s$ detectors provides is $\log\left(\frac{R}{s}\right)$, where $\log(R)$ is some constant amount of information provided if the entire computational resource were allocated to one detector.

Let $DT_1$ denote the singleton detector comprised of using the entire computational resource $R$, and $DT_s$ denote one of the $s$ detectors using only $R/s$ "neuronal" computational units. Then,

$$I_{DT_1} = H_0 - H_{DT_1} = \log(R), \text{ and } I_{DT_s} = H_0 - H_{DT_s} = \log(R/s) \Longleftrightarrow$$
$$H_{DTs} - H_{DT_1} = \log(R) - \log(R/s) = \log(s),$$

that is, that the output of each of $s$ detectors has $\log(s)$ bits more uncertainty in it than the singleton detector.

### 4.2 FPC, TPC, FNC, and TNC for this system

We will assume this time that the detector is Bayes optimal, i.e. that it incorporates the prior information into its decision threshold. For simplicity, and with some loss of generality, assume that the output probability density on $x$ of the detectors is Gaussian around means +1 and −1 corresponding to target present and absent, resp., with standard deviation $\sigma_{DT}$. Then, since the differential entropy of a Gaussian is $\log(\sigma\sqrt{2\pi e})$, a distribution which is $\log(s)$ bits more entropic than the normal with $\sigma_{DT_1}$ has standard deviation $s \cdot \sigma_{DT_1}$, where $\sigma_{DT_1}$ characterizes the output density over $x$ of the detector which uses the entire computational resource. Therefore, since we assume we process $w$ windows, we will employ detectors with output distributions having $\sigma = w \cdot \sigma_{DT_1}$.

The expected false positive count of our system, if it examines $w$ windows is, from section 3.1:

$$\begin{aligned} E[FPC(w)] &= \sum_{i=1}^{w} p(D_i = 1|T_i = 0)p(T_i = 0) \\ &= \sum_{i=1}^{w} fpr_i \cdot p(T_i = 0) \end{aligned} \tag{4}$$

To calculate $fpr_i$, we examine the likelihood ratio at window $i$, corresponding to the prior $p_i$ :

$$\begin{aligned} LLR_i &= \frac{p(D_i|T_i = 1)}{p(D_i|T_i = 0)} \frac{p_i}{1 - pi} \\ &= \frac{\exp(-(D_i - 1)^2/2\sigma^2)}{\exp(-(D_i + 1)^2/2\sigma^2)} \cdot \frac{p_i}{1 - pi} \\ &= \exp(2D_i/\sigma^2) \cdot \frac{p_i}{1 - pi} \end{aligned}$$

$D_i = 1$ when $LLR_i > 1 \Longrightarrow$

$$\begin{aligned} \exp(2D_i/\sigma^2) > \frac{1 - p_i}{p_i} \Longleftrightarrow 2D_i/\sigma^2 > \log\left(\frac{1 - p_i}{p_i}\right) \Longleftrightarrow \\ D_i > \frac{\sigma^2}{2} \log\left(\frac{1 - p_i}{p_i}\right) \end{aligned}$$

Thus,

$$\begin{aligned} fpr_i &= p\left( D_i > \frac{\sigma^2}{2} \log\left(\frac{1 - p_i}{p_i}\right) \middle| T_i = 0 \right) \\ &= Q\left( \frac{\frac{\sigma^2}{2} \log\left(\frac{1-p_i}{p_i}\right) + 1}{\sigma} \right) \\ &= Q\left( \frac{\sigma}{2} \log\left(\frac{1 - p_i}{p_i}\right) + \frac{1}{\sigma} \right) \end{aligned} \tag{5}$$

where $Q(\cdot)$ is the complementary cumulative distribution function the standard normal. Substituting (5) into (4) gives:

$$E[FPC(w)] = \sum_{i=1}^{w} Q\left(\frac{\sigma}{2}\log\left(\frac{1-p_i}{p_i}\right) + \frac{1}{\sigma}\right)(1-p_i). \tag{6}$$

Similarly,

$$E[TPC(w)] = \sum_{i=1}^{w} Q\left(\frac{\sigma}{2}\log\left(\frac{1-p_i}{p_i}\right) - \frac{1}{\sigma}\right)p_i \tag{7}$$

and the other two are dependent on these as usual: $E[TNC(w)] = (N-n) - E[FPC(w)]$, and $E[FNC(w)] = n - E[TPC(w)]$.

### 4.3 Optimal distributions of the computational resource

Equations (6)-(7) are difficult to analyze as a function of $w$ analytically, so we investigate their implications numerically. To begin, we use a model from equation (1), with $n = 3$ expected targets per total frame, $N = 100$ windows, prior profile parameter $k = 20$, and $\sigma_{DT_1} = 2/N$. The results are shown below:

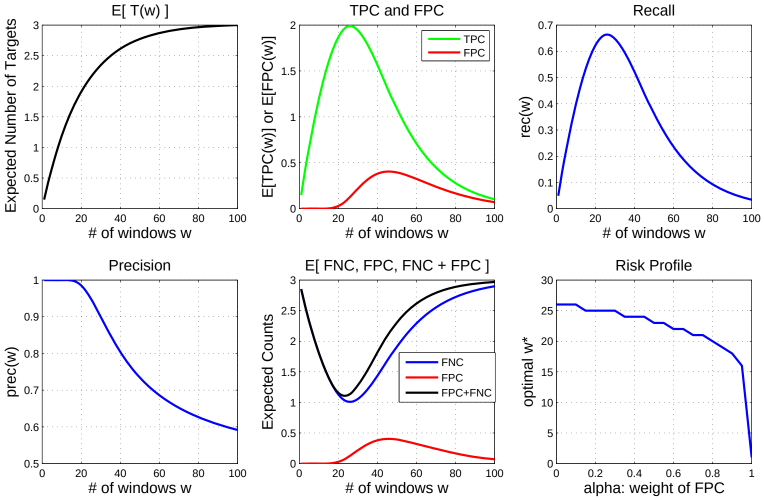

Fig. 5. Performance of an object detection system with fixed computional resource.

We observe the increasing recall, decreasing precision trend for low $w$ values, now *even with perfect knowledge of the prior*. This suggests that, at least for this setting of parameters, resources are best concentrated among just a few windows. The most striking feature of these plots, for example of the expected true positive count shown in green, is that there is an optimum around 20 or so windows. This corresponds to where the aggregate information of the thresholded detectors is peaked – beyond that, the detectors are spread too thinly and become less useful. Note that this is in contrast to the aggregate information of the pre-threshold real-valued detection outputs, which increases monotonically as $w \log(R/w)$.

It is interesting to understand not only that subselecting visual regions is beneficial for performance, but how the exact level of spatial attention depends on other factors. We now introduce the notion of a "Risk Profile":

$$w^*(\alpha) = \arg\min_w \left\{\alpha E[FPC(w)] + (1-\alpha)E[FNC(w)]\right\}.$$

That is, suppose a system has penalty function which depends on the false positives and false negatives. Both should be small, but how the two compare might depend on the environment: a prey may care a lot more about false negatives than a predator, e.g.. For a given false positive weight $\alpha$, the optimal $w^*$ corresponds to number of windows among which the fixed computational resource should be distributed in order to minimize penalty. We find (see Fig. 6), that an increasing emphasis on false negatives (low $\alpha$), leads to a more thinly distributed attentional resource being optimal. Thus, in light of this simple analysis, it makes sense that an animal with severe false negative penalties, such as a grazer with wolves on the horizon, may have evolved to spread out its sensory-cortical hardware over a larger spatial region – and indeed grazers have an elongated visual streak rather than a small fovea.

The general features of the plots shown in Fig. 5 hold over a wide range of parameters. We summarize the numerical findings by showing the risk profiles for a few such parameter ranges:

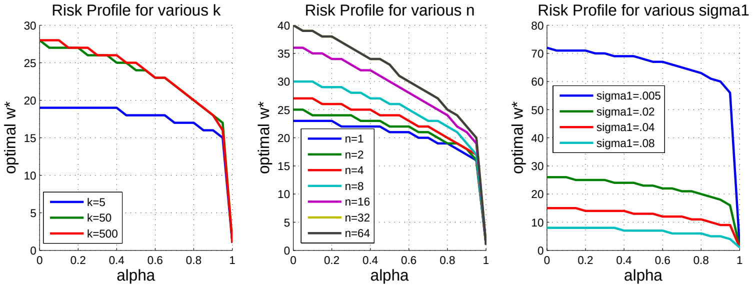

Fig. 6. The optimal number of windows out of 100 to process, for increasing $\alpha$, the importance of avoiding false positives relative to false negatives. sigma1$\equiv \sigma_{DT_1}$

The important feature of all these plots is that the optimal number of windows $w$ over which to distribute computation in order to minimize the penalty function is always less than $N = 100$, and that the risk profiles increase to the left, with increasing false negative count importance, for a wide range of parameterized conditions.

## 5 Conclusions

We have demonstrated, first in experiment and then using a simple numerical model, the critical importance of attentional selection for increased accuracy in a detection task. We find that processing scene portions which are a priori unlikely to contain a target can hurt performance if this prior information is not utilized to bias detection decisions. However, if the computational resources available for detection are fixed and must be distributed somehow to various scene portions, with a corresponding dilution in accuracy, it is best to concentrate them on scene portions which are a priori likely to contain a target, even if prior information biases detector outputs optimally. Note that this argues for an attentional strategy independent of computational savings – no matter how great the computational resource, it is best focused attentionally. We also show how a system which prioritizes false negatives high relative to false positives benefits from a blurred focus of attention, which may anecdotally suggest an evolutionary pressure for the variety in photoreceptor distributions in the retinae of various species. In conclusion, we provide a novel framework within which to understand the utility of spatial attention, not just as an efficiency heuristic, but as fundamental to object detection performance.

## 6 Acknowledgements

We wish to thank DARPA for its generous support of a research program for the development of a biologically modeled object recognition system, and our close collaborators on that program, Sharat Chikkerur at MIT, and Rob Peters at USC.

## 7 Appendix

Table of parameters:

| | |
|---|---|
| $N$ | # of windows available to process in a frame |
| $w$ | # of windows processed in a frame |
| $n$ | average # of target-containing windows in a frame |
| $k$ | poverty of prior information $\Rightarrow$lower $k$, better a priori sorting of windows |
| $\sigma_{DT_1}$ | standard deviation of detector output, if only one detector is used |

## References

1. J.K. Tsotsos, S.M. Culhane, W.Y. Kei Wai, Y. Lai, N. Davis, "Modeling visual attention via selective tuning", *Artificial Intelligence*, 1995
2. Y. Amit, D. Geman, "A Computational Model for Visual Selection", *Neural Computation*, 1999

3. A. Yu, P. Dayan, "Inference, Attention, and Decision in a Bayesian Neural Architecture", *Proc. Neural Information Processing Systems (NIPS)*, 2004
4. J. Bonaiuto, L. Itti, "Combining attention and recognition for rapid scene analysis", *Proc. IEEE-CVPR Workshop on Attention and Performance in Computer Vision (WAPCV 2005)*, 2005.
5. U. Rutishauser, D. Walther, C. Koch, P. Perona, "Is attention useful for object recognition?", *Proc International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004
6. F. Miau, C.S. Papageorgiou, L. Itti, "Neuromorphic algorithms for computer vision and attention", *Proceedings of Annual International Symposium on Optical Science and Technology (SPIE)*, 2001
7. F. Moosmann, D. Larlus, F. Jurie, "Learning Saliency Maps for Object Categorization", *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*, 2006
8. C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", *Hum. Neurobiol.*, 1985
9. L. Itti, C. Koch, "Computational modeling of visual attention", *Nature Reviews Neuroscience*, 2001
10. Y. Ye, J.K. Tsotos, "Where to Look Next in 3D Object Search", *Proc. of Internat. Symp. on Comp. Vis.*, 1995
11. http://cbcl.mit.edu/software-datasets/streetscenes/
12. http://labelme.csail.mit.edu/
13. S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006
14. D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, 2004
15. S. Waydo, A. Kraskov, R. Quian Quiroga, I. Fried, C. Koch, "Sparse Representation in the Human Medial Temporal Lobe", *Journal of Neuroscience*, 2006
16. A. Treisman, "How the deployment of attention determines what we see." *Visual Cognition*, 2006
17. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2001
18. H.E. Pashler *The Psychology of Attention.* MIT Press: Cambridge, 1998)
19. J. Braun, C. Koch, J.L. Davis, editors, *Visual Attention and Cortical Circuits.* MIT Press: Cambridge, 2001
20. D. Walther, C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, 2006
21. S. Mitri, S. Frintrop, K. Pervolz, H. Surmann, A. Nuchter, "Robust Object Detection at Regions of Interest with an Application in Ball Recognition", *Proc. of International Conference on Robotics and Automation (ICRA)*, 2005