On the Optimality of Spatial Attention for Object Detection

Jonathan Harel[1]
Christof Koch[2]

[1]Ph.D. student in Electrical Engineering
[2]Prof. of Computation & Neural Systems, Biology
California Institute of Technology

Presentation on May 12, 2008 at WAPCV 2008 in Santorini, Greece
Full paper available from publications link at http://www.klab.caltech.edu/~harel/

wapcv 08

My motivation for this project

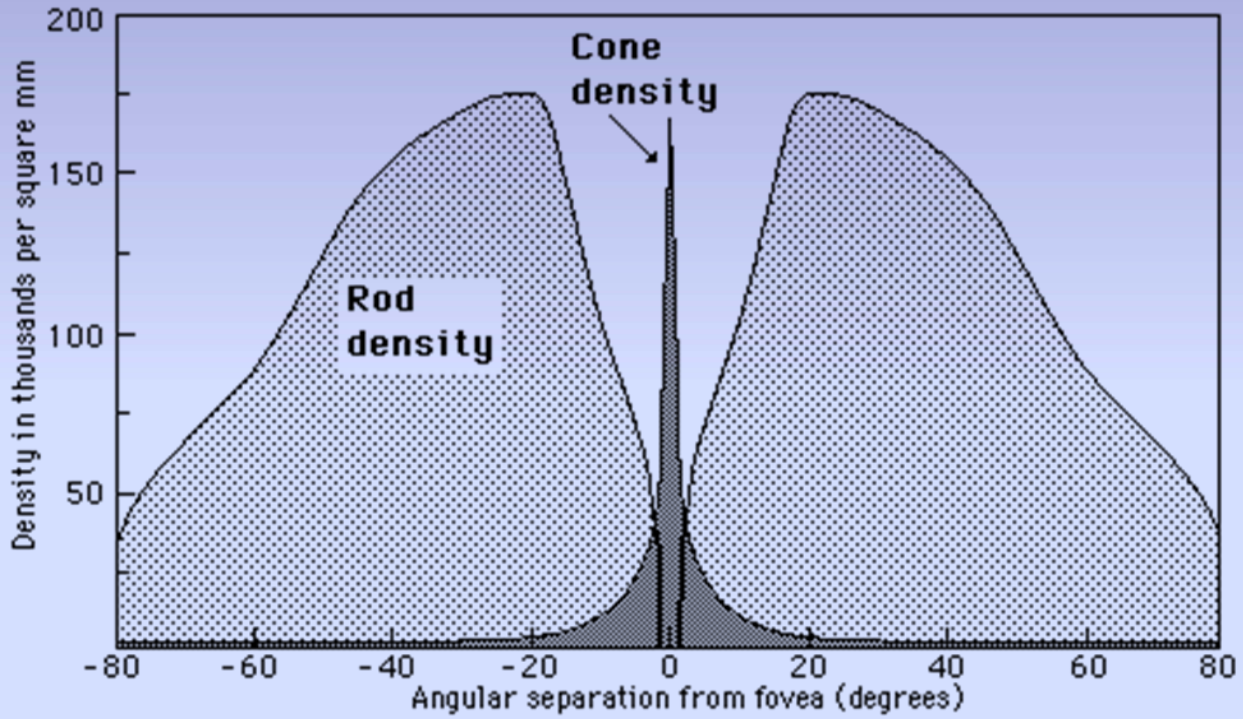
- Attempt to understand the benefits of attention for machine vision in a general setting

- Attempt to understand the principles behind its emergence in biology

We need eye-movements (overt attention) for basic tasks

population 

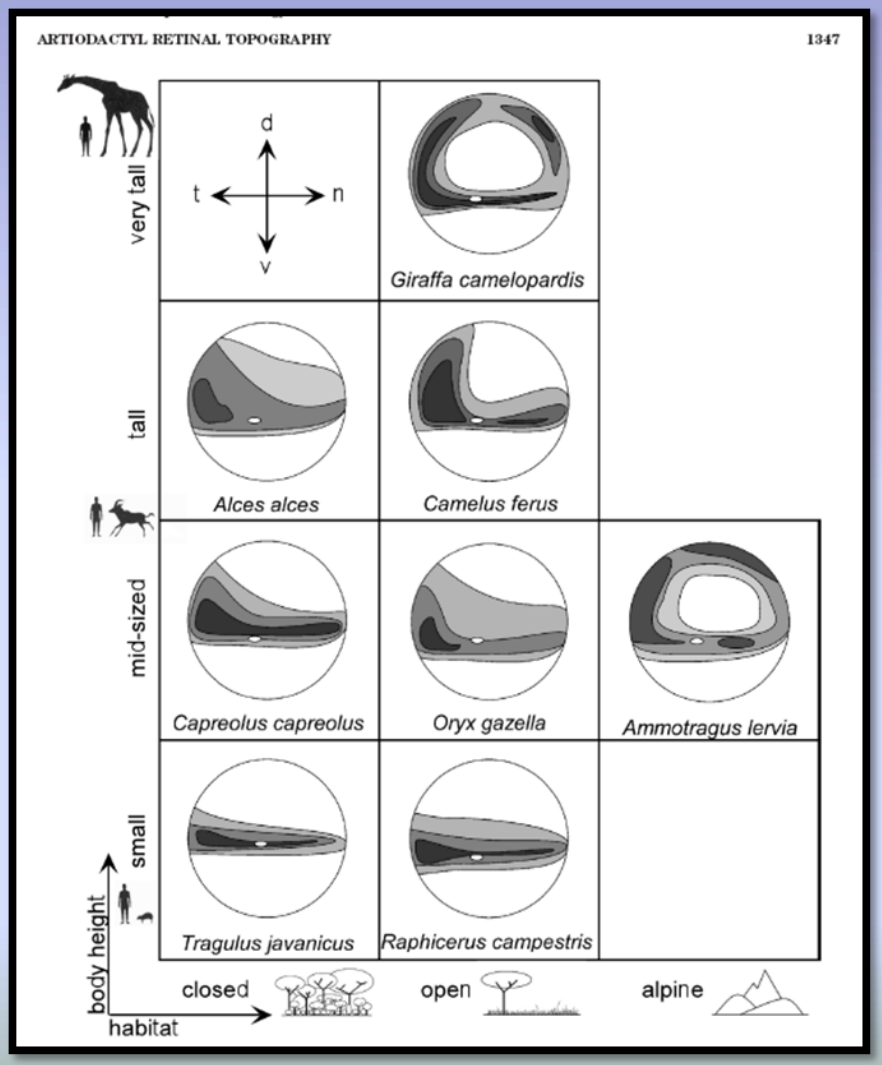
Photoreceptors in the human retina



http://hyperphysics.phy-astr.gsu.edu/hbase/vision/rodcone.html

Photoreceptors in other animals: visual streaks and beyond

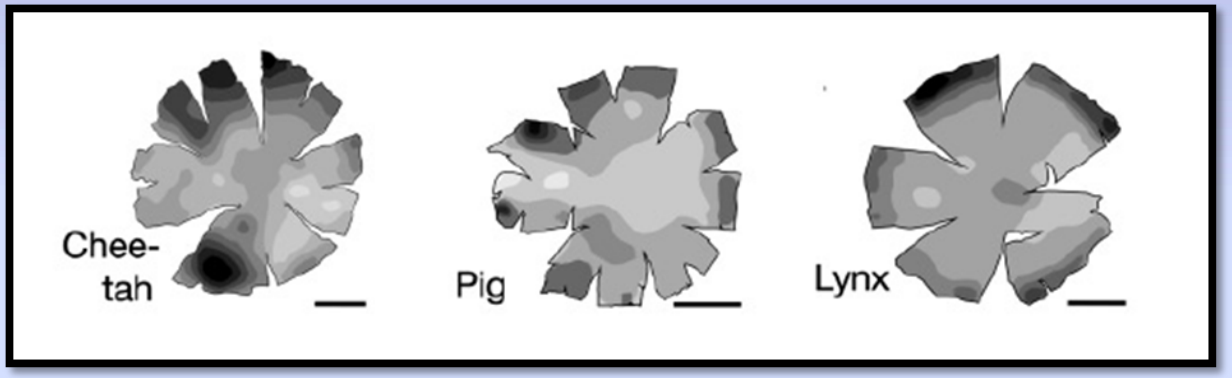
M-Cone Distributions in Artiodactyls (even hoofed creatures)



Schiviz et al., "Retinal Cone Topography of Artiodactyl Mammals: Influence of Body Height and Habitat", *J. Comparative Neurology*, 2008.

Further retinal structure:
chromatic vision concentrated outside *area centralis*

Percentage of cone population which is S-type in flattened retinae
(white-black range is 0-30%)

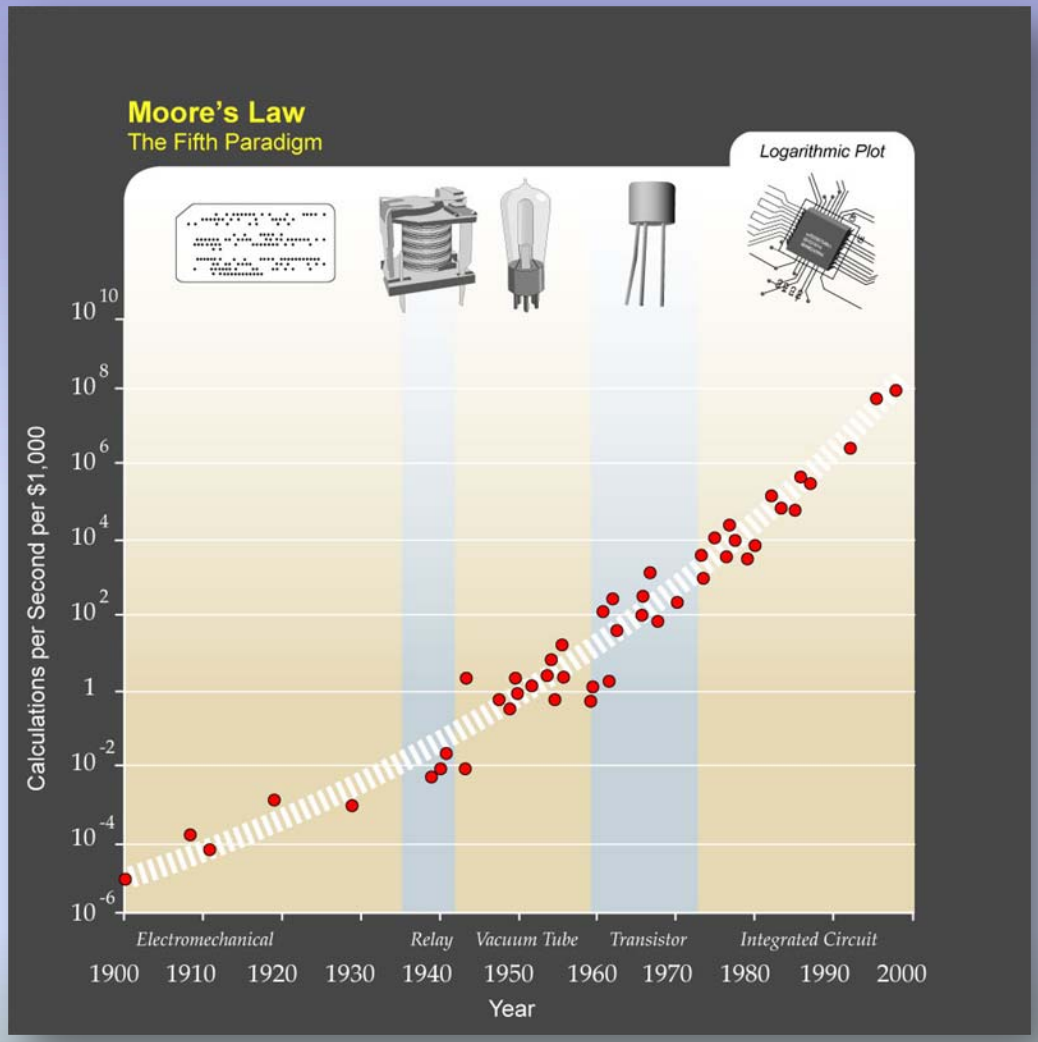


Anhelt et al., “Independent variation of retinal S and M cone photoreceptor topographies: A survey of four families of mammals”, *Visual Neuroscience*, 2006.

Okay, zoologist.
a common concern

- If attention is a way to focus processing on a subset of the input, why should machine vision scientists care?
- We can just throw more hardware at the problem.

Availability of computational power



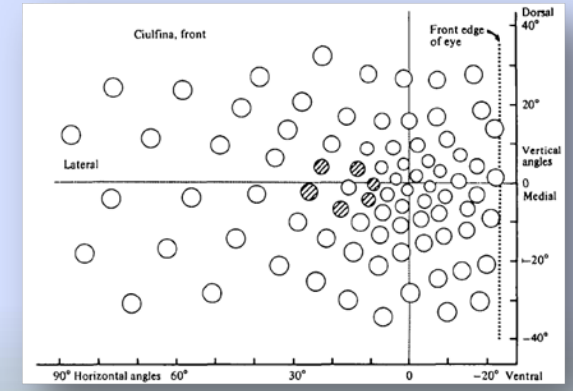
http://en.wikipedia.org/wiki/Moore's_law

Lesson from biology:

sensory focusing naturally selected *across computational scales*

praying mantis: ~10^5 neurons

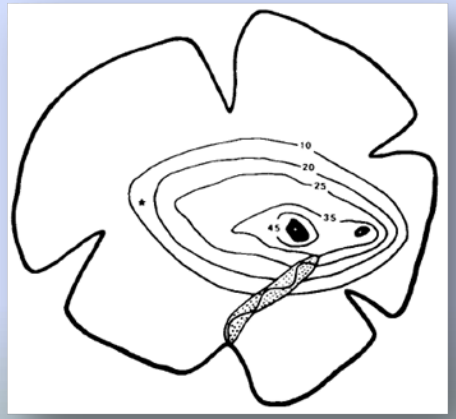
number of ommatidia (5/circle)



G. Horridge, 1978

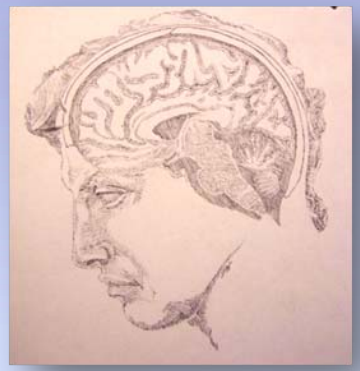
Chilean eagle

ganglion density map (10^3/mm^2)

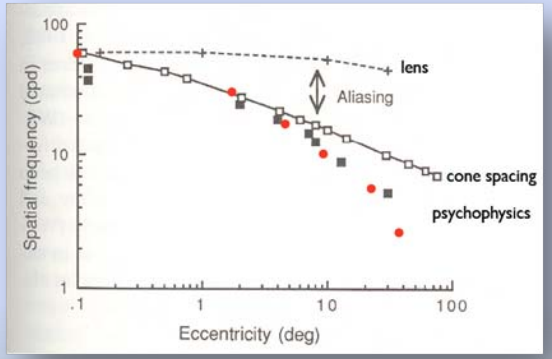


O. Inzunza, 1991

human : ~10^{11} neurons



visual acuity



P. Perona Lecture Slides, CNS 186, Caltech

A possible explanation

- “The streak, or any other distribution, is rather seen as minimising redundancy only for a subset of the image … [not] significant in the animal’s lifestyle.”

A. Hughes, Letter to the Editors of *Vision Research,* 1981.

Uneven retinal sampling = compression code?

- *"*relative entropy … is the maximum compression possible ... One minus the relative entropy is the *redundancy*."
  - C. Shannon, *The Mathematical Theory of Communication,* 1948.

the information *relevant to the animal* is represented more efficiently by uneven sampling

Why should we compress some visual information?

- Higher spatial acuity in some areas means potentially better decision making on subsets of the input which are important behaviorally

- *No matter how much hardware you have*, it's probably better to focus it (as naturally selected in *Animalia*), because it gives you a better system without additional mechanical cost

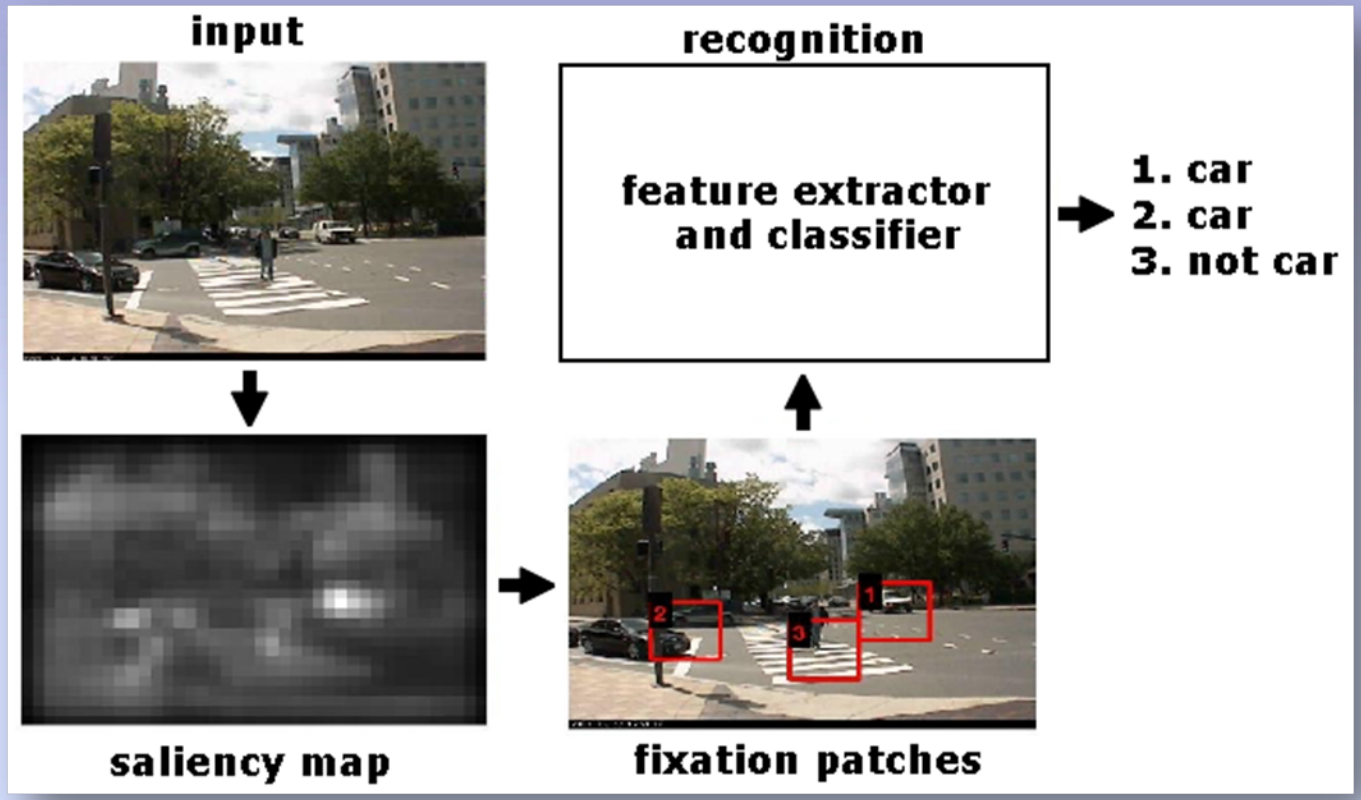
- Can we quantify this?

Unpacking this in a restricted case:
the advantage of spatial attention for object detection

Disclaimer: attention

- Definition: Preferential allocation of more processing resources to some sensory input (e.g., more processors, more firing, more neurons, etc.).

Begin with a toy experiment



Results

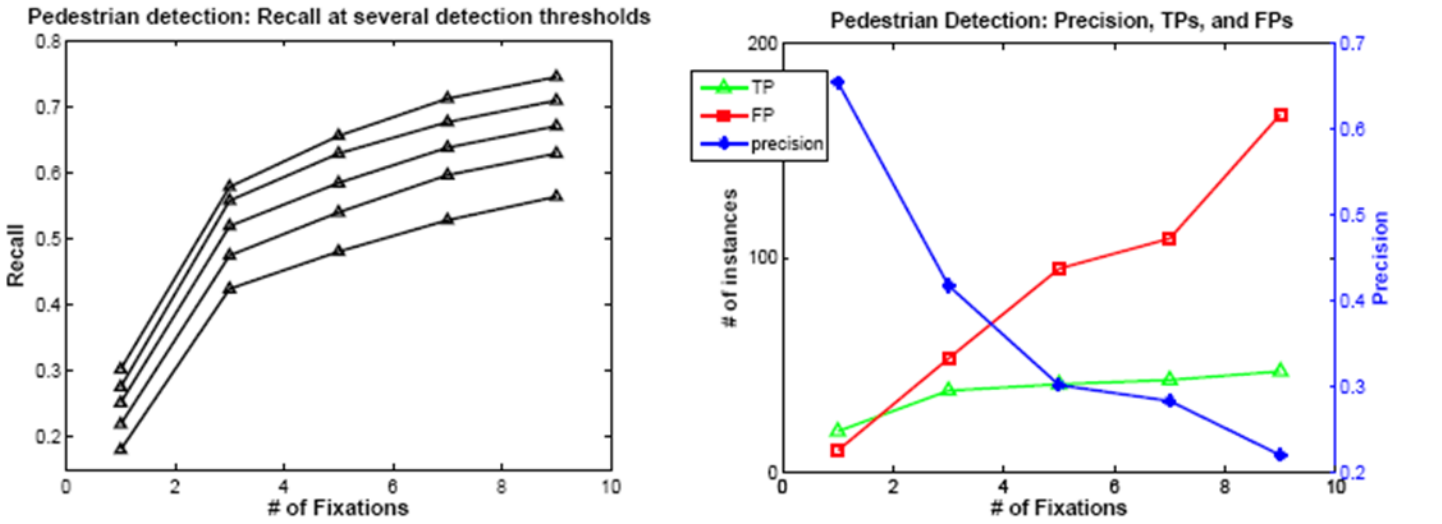


Fig. 1. Result of running detector over entire video. As the number of windows processed per frame increases, recall rate increases (left), while precision rate decreases (right). Left: curves for different settings of the SVM detection threshold.

Recall = fraction of targets detected
Precision = fraction of detections which are targets

As *more* "fixation windows" are processed, *precision decreases* since the false positive rate remains constant, while windows are increasingly unlikely to be targets. Argues for processing only most important areas of the scene.

The catch

- So you make more mistakes if you process unlikely target locations in the exact same fashion as you process likely target locations.

- But – what if you bias detectors, i.e. reflect this reduced likelihood of being a target via *modulation by the prior*. Then what?

- Yes, in that case it's better to process the *whole scene.*

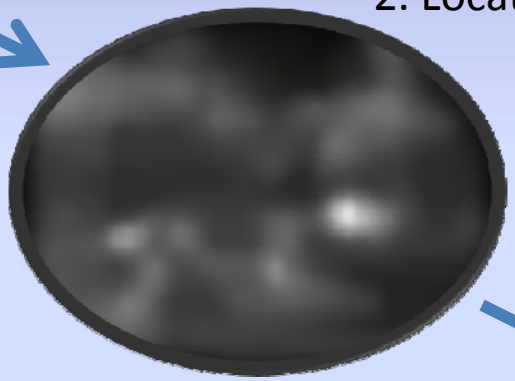
But what if you can make *more accurate detections* when you make fewer of them?

- This just reflects a conversation of computational resources: you can either make more, less accurate detections, or fewer, more accurate ones.
- Then, it again becomes better to process the most likely locations.
- How? It gets messy.

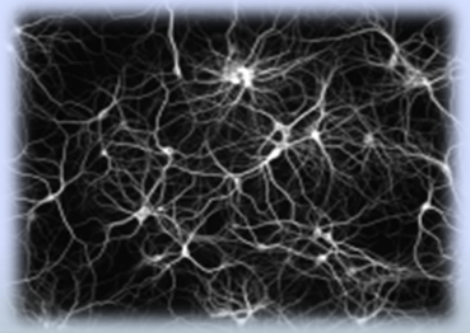
Begin with a simple model

1. The visual world is observed.

2. Locations in the scene are prioritized.



3. Some number of locations is processed by object detectors



4. The behavior of the system depends on the reliability of these detections.

Note

- We will not talk about
  - How image locations are prioritized (including whether the prioritization is due to bottom-up or top-down effects)
  - How visual processing is carried out exactly

- We will only examine logical conclusions of such a model.

Some modeling assumptions

1. On statistics of detector output
2. On how “information content” of a detector is measured
3. On how this “information content” degrades with number of detectors employed
4. Nature of location prioritization

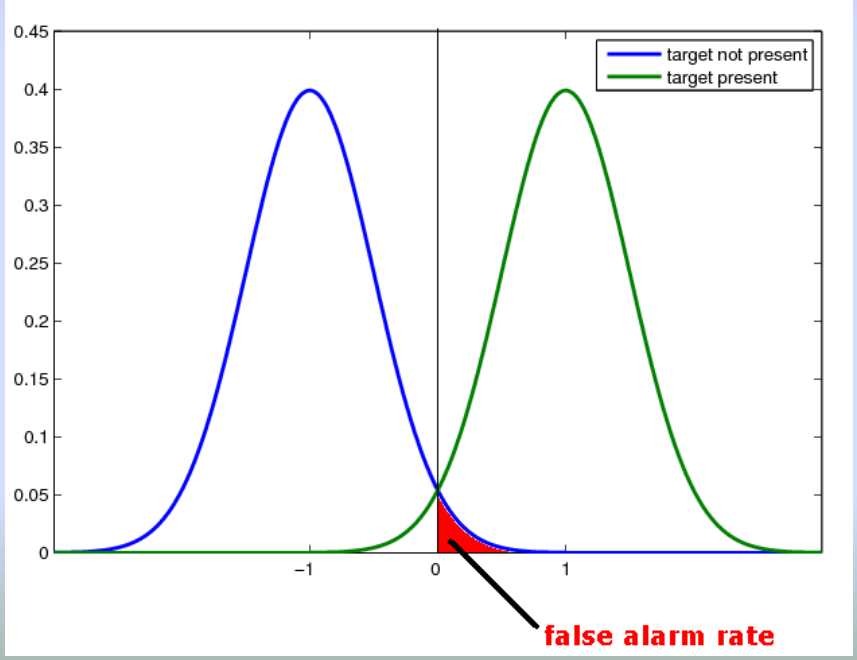
1. Statistics of detector output

In order to incorporate prior probabilities,

The detector must provide *some object* D such that the *conditional* p(D|target {present,absent}) can be modulated by the *prior* p(target {present,absent}).

Assume: D is a real value, and that the conditionals are normal.

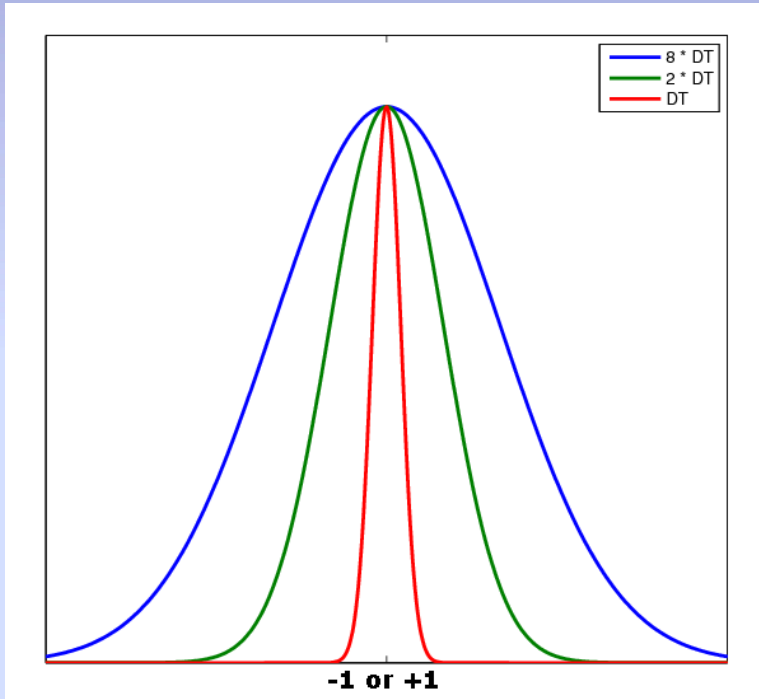
p(D|target {present,absent})



Detection threshold shown for uniform prior belief.

2. How “information content” of a detector is measured

p(D|target {present,absent}) modulated by σ

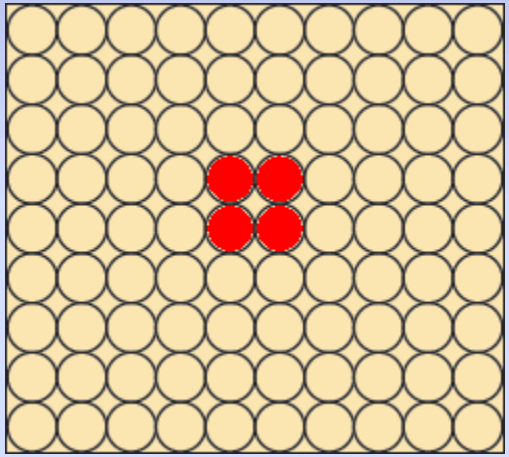


$$H(\sigma) = \log(\sigma\sqrt{2\pi e})$$
$$I(\sigma) = H_0 - H(\sigma)$$

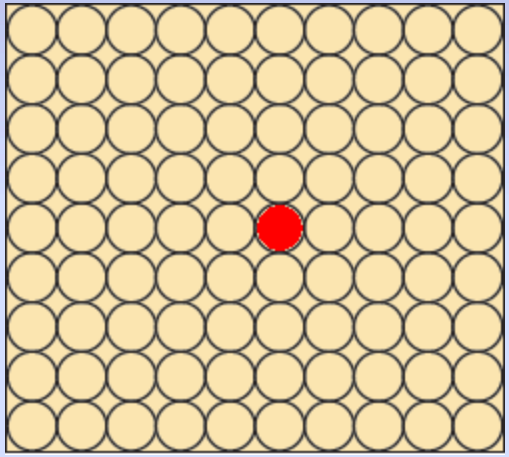
Larger σ => less informative detector

3. How σ increases with number of detectors employed

- Let's model our computational resource as having a fixed number of computational "nodes".
- How should performance degrade if we use only some of them to compute?



using 4 computational nodes per detector => can have many detectors



using 1 computational node per detector => can have even more, weaker detectors

Assume: Logarithmic information content loss

- Suppose each node can encode one possible state in an ensemble. Then R nodes can encode log(R) bits, and R/s nodes can encode log(R/s)=log(R) – log(s) bits.

- This is a ‘sparse’ coding scheme (evidence from V1 to Hippocampus)

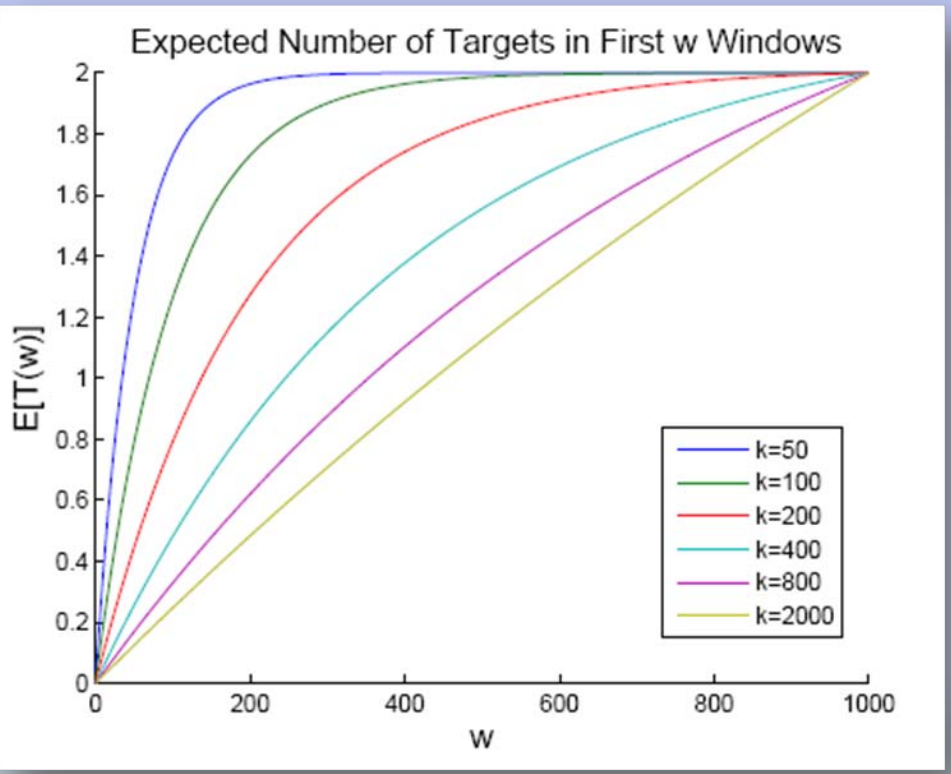
How this translates to increasing σ

- If we use all R resource nodes to perform a detection, let's say we have a detector which has output Gaussian with entropy H_{DT_1} corresponding to σ_{DT_1}
- Then if we use just R/s nodes, our output should be log(s) bits more entropic, meaning $\sigma_{DT_s} = s \cdot \sigma_{DT_1}$

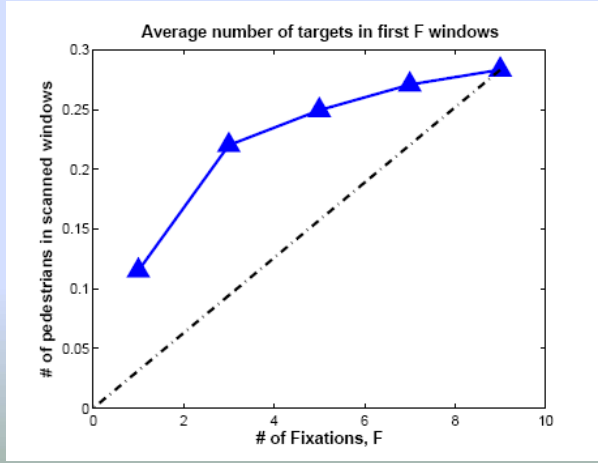
4. Nature of location prioritization

Assume: a scene contains n targets on average, and N locations, ordered according to priority. If we process the first w of them, we expect to have E[T(w)] targets on average, where:

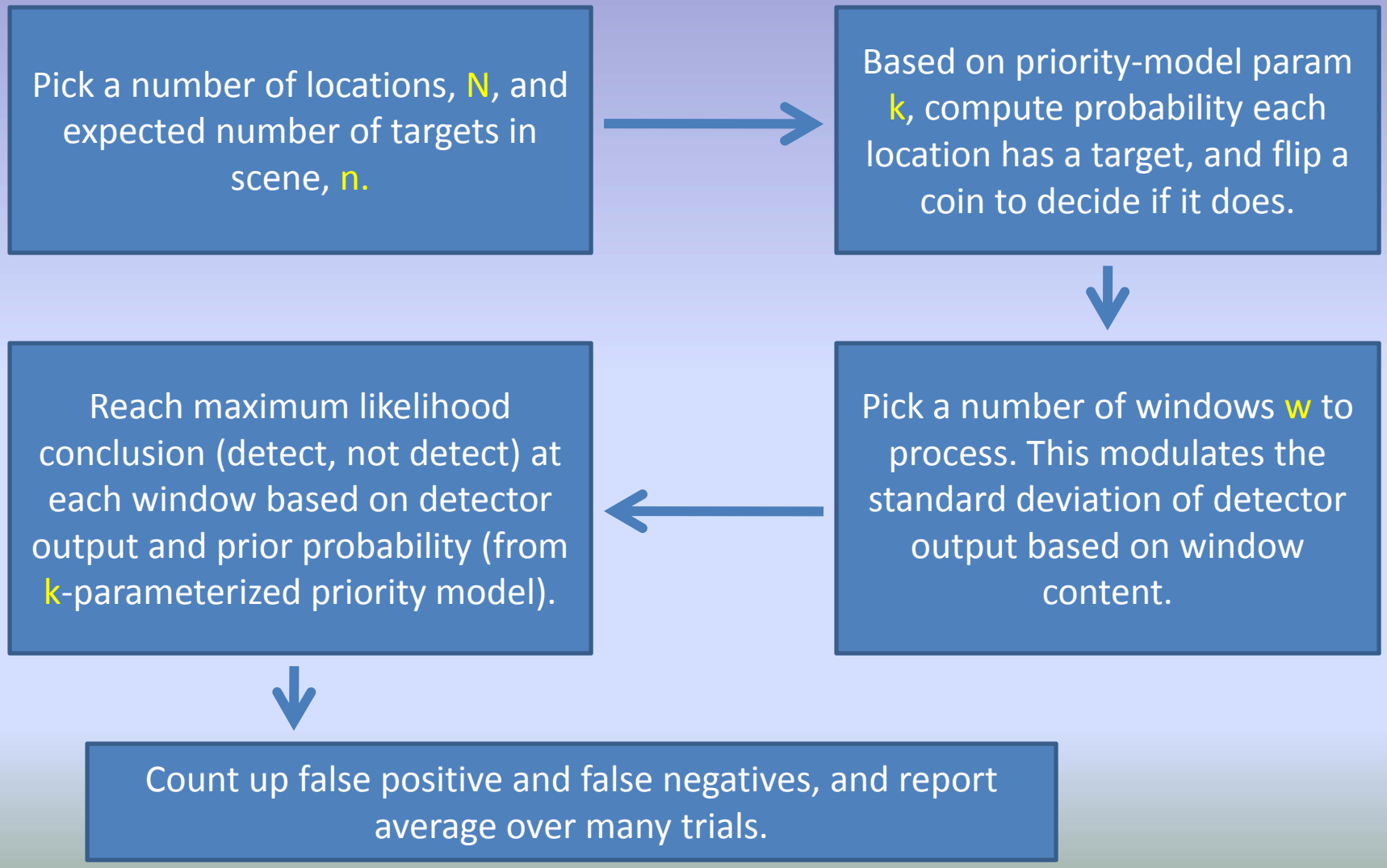
$$E[T(w)] = n\frac{1-\exp(-w/k)}{1-\exp(-N/k)}$$



From experiment:



Simulate model, get results



Simulation results: best to process only most important locations: w* < N=100

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

w	# of windows processed in a frame
n	average # of target-containing windows in a frame
k	poverty of prior information ⇒lower k, better a priori sorting of windows
σ_{DT_1}	standard deviation of detector output, if only one detector is used

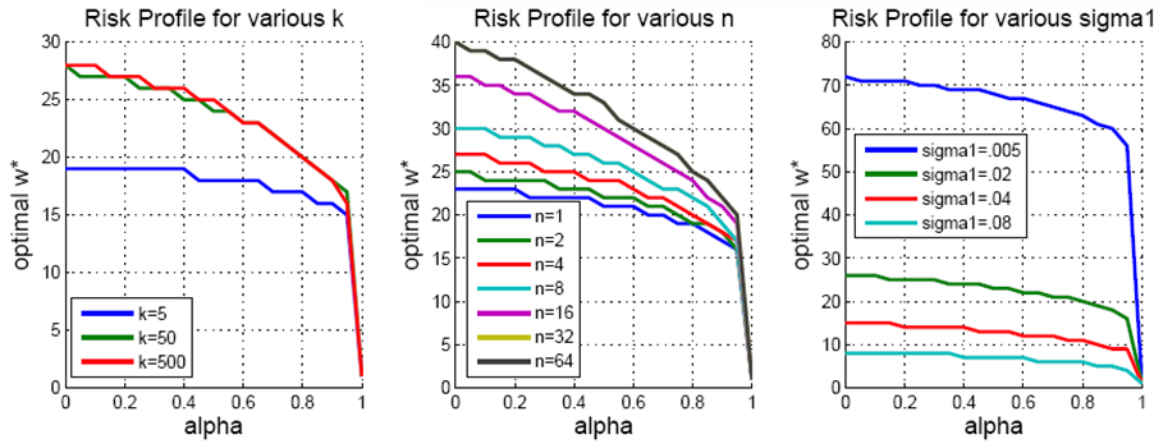


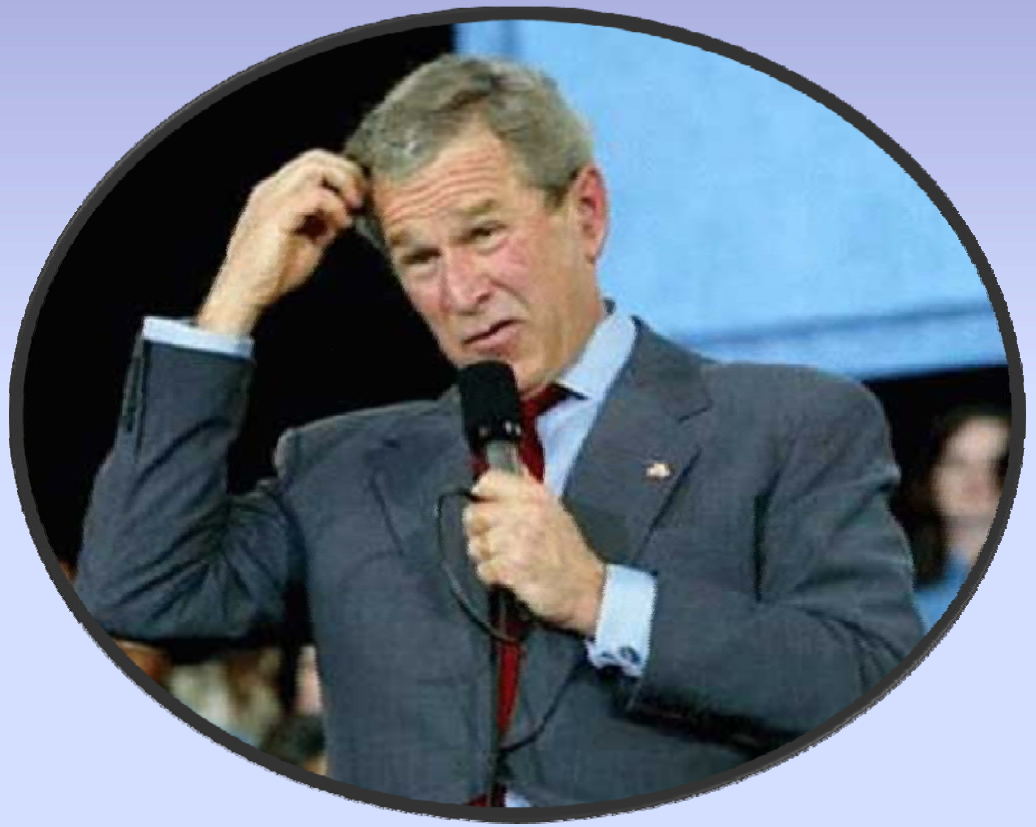
Fig. 6. The optimal number of windows out of 100 to process, for increasing α, the importance of avoiding false positives relative to false negatives. sigma1$\equiv \sigma_{DT_1}$

w* mostly governed by where sum information in binary detectors begins to fall off.

Conclusions

- We see that in the animal kingdom, selective, overt visual attention has emerged in species of all brain sizes.
- We present a model which can choose between processing all locations or only the most important given a fixed resource.
- We conclude that, in animal or machine, selective spatial attention to subsets of the visual input is always preferable to the lack thereof.

Thanks for listening! Questions?



Full paper available from publications link at http://www.klab.caltech.edu/~harel/