Zoya Gavrilov

Based on:
“A Computational Perspective on Visual Attention” by John K. Tsotsos

VISUAL ATTENTION

Talk Layout

- PART I
  - The foundations of attention research
- PART II
  - Selective Tuning: a computational model of attention

Based on the book: "A Computational Perspective on Visual Attention" by John K. Tsotsos, 2011

PART I: FOUNDATIONS
Talk Layout

- What is attention and why do we need it?
- The complexity of visual processing tasks
- What approximations can we make to the general vision problem?
  - Problems encountered along the way
- Why do we need attention? (revisited)
  - How attention be used to provide approximations and remedy problems
- Computational and biological (attentive) mechanisms for information reduction

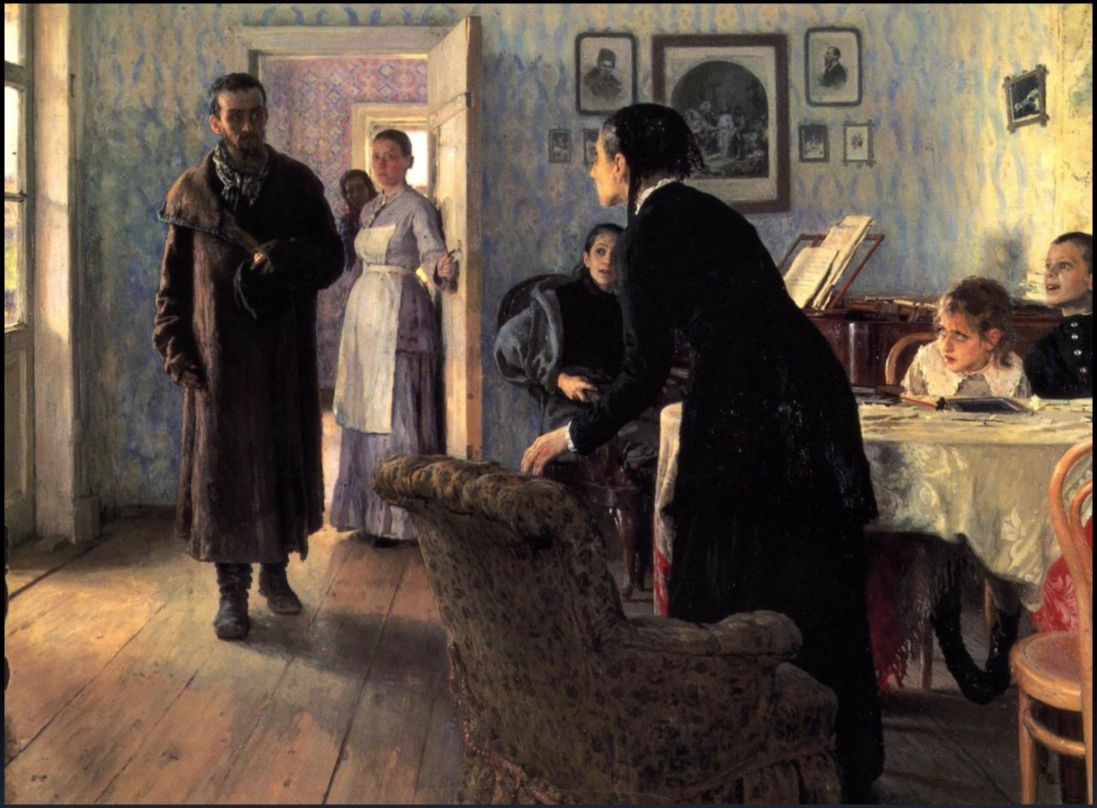
"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought."

- William James, 1890

"After many thousands of experiments, we know only marginally more about attention than about the interior of a black hole."

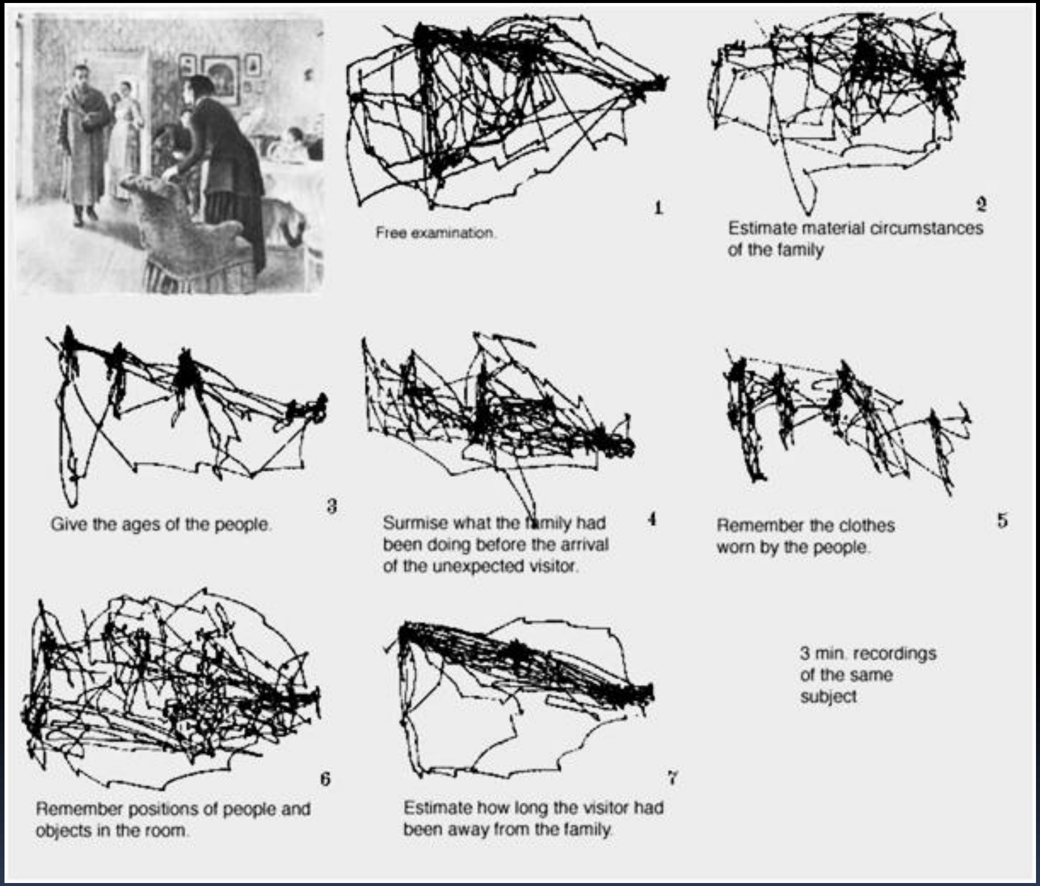
- James Sutherland, 1998

Alfred Yarbus, 1979



The Unexpected Visitor
Ilya Repin, 1888

Attentional mechanisms at work…


1
Free examination.
2
Estimate material circumstances of the family
3
Give the ages of the people.
4
Surmise what the family had been doing before the arrival of the unexpected visitor.
5
Remember the clothes worn by the people.
6
Remember positions of people and objects in the room.
7
Estimate how long the visitor had been away from the family.
3 min. recordings of the same subject

Scan Paths

Posner, 1980

Attention has 3 major functions:

(1) ALERTING - providing the ability to process high-priority signals

(2) ORIENTING – permitting orienting and overt foveation of a stimulus

(3) SEARCH – allowing search to detect targets in cluttered scenes

Overt attention: orienting the body, head, eyes to foveate a stimulus

Covert attention: attention to a stimulus in the visual field without eye movements (Helmholtz, 1896)

Saccade preparation biases visual detection at the location to which the saccade is prepared (Moore and Fallah, 2004)

- **Sequential Attention Model:** eye movements necessarily preceded by covert attentional fixations (Posner, 1980)
- **Oculomotor Readiness Hypothesis:** covert and overt attention as independent and co-occuring because they are driven by the same visual input (Klein, 1980)
- **Premotor Theory of Attention:** covert and overt attention as the result of activity of the motor system that prepares eye saccades (Rizzollati et al., 1987)
- **Current:** overt orienting is preceded by covert orienting; overt and covert orienting are exogenously activated by similar stimulus conditions; endogenous covert orienting of attention is not mediated by endogenously generated saccadic programming (Klein, 2004)

Elements of Attention

- Representational:
  - Working memory
  - Salience
  - Fixation history
- Control:
  - Recognition
  - Binding
  - Search
- Observed:
  - Movement
  - Time course
  - Task accuracy

Information Reduction

Elements of Attention

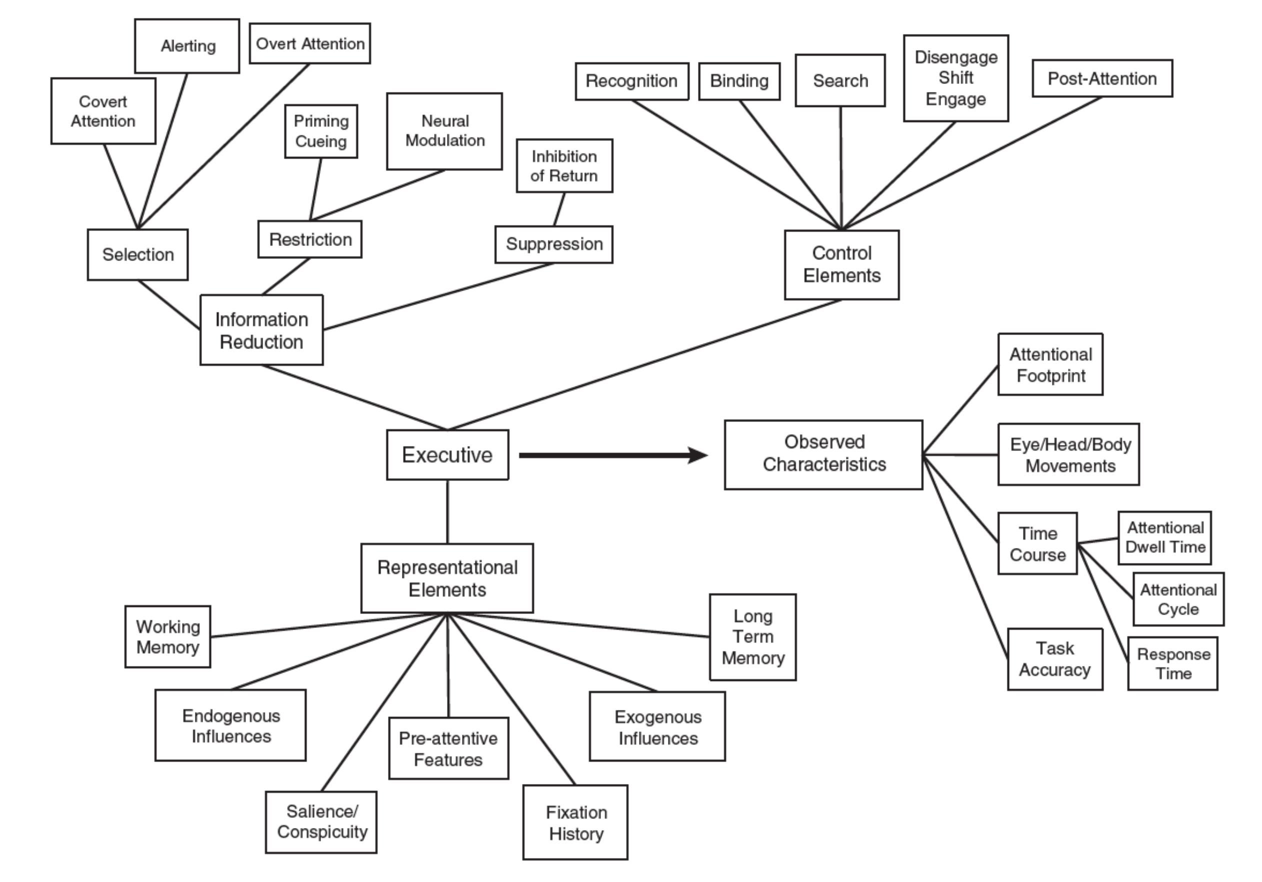
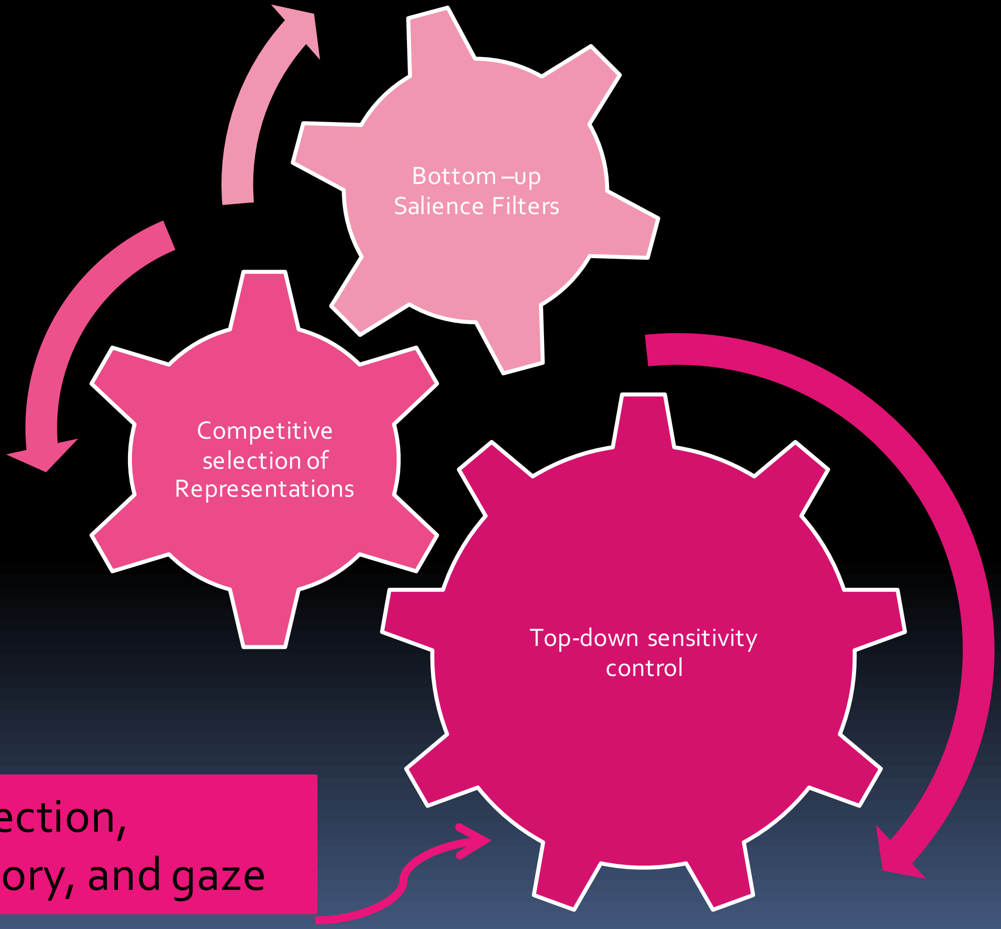


Image copyrighted: "A Computational Perspective on Visual Attention, John K. Tsotsos, The MIT Press, Cambridge MA, 2011"

Knudsen, 2007



Combines selection, working memory, and gaze

Attention for information reduction

- In A.I., attention as a search-limiting heuristic
  - Computational resources
- Adaptation to dynamic needs
- Span of attention -> bottleneck form
  - Perceptual load
  - Working memory

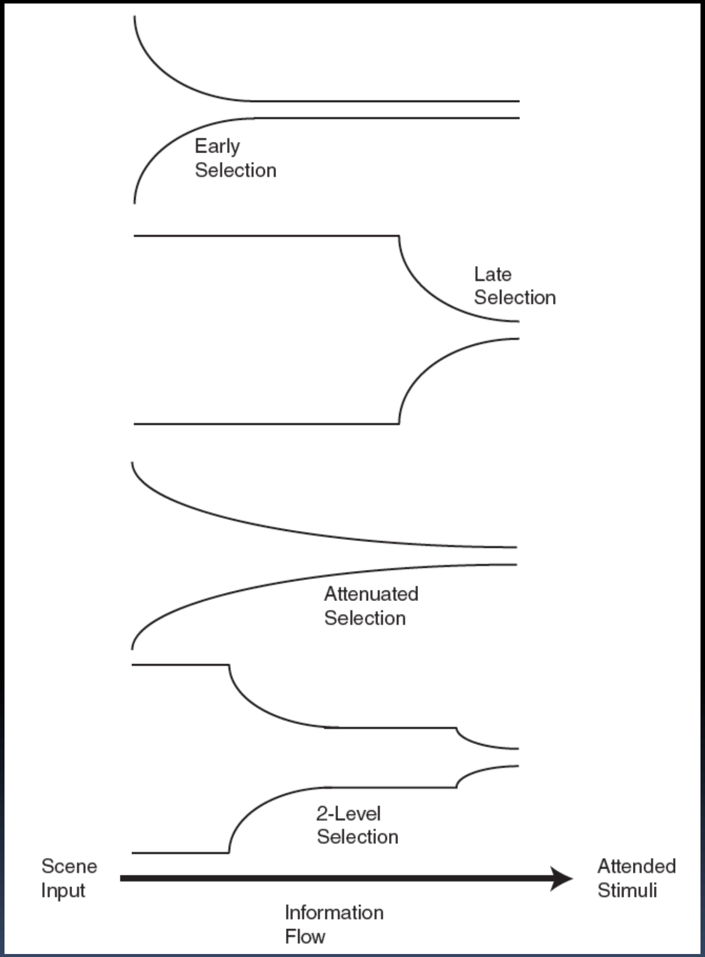


Image copyrighted: "A Computational Perspective on Visual Attention, John K. Tsotsos, The MIT Press, Cambridge MA, 2011"

"Attention is the process by which the brain controls and tunes information processing."

- John Tsotsos, 2011

Motivation

The hallmark of human vision is its generality!

Motivation

Worst-case complexity of unbounded visual search:

$$O(NP^2 2^{PM})$$

N = # object/event prototypes in one's knowledge base

P = # input locations/pixels

M = # measurements/features computed at each pixel

"As computers become faster, the sizes of P and M that can be handled grow, giving the appearance that the theory does not matter BUT this does not scale since feed-forward approaches remain exponential in the general case and outside the range of any improvements in computing power." (Tsotsos, 2011)

Improving time complexity

(1) Hierarchical Organization

(2) Pyramid Representation

(3) Spatiotemporal Localization

(4) Receptive Field Selectivity

(5) Feature Selectivity

(6) Model Space Selectivity

Hierarchical Organization

- Model space reduced from O(N) to O(lgN)
- Biederman's figure of 30,000 categories reduces to 15 levels of a binary decision tree

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$O(P^2 2^{PM} lgN)$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$O(NP^2 2^{PM})$

Pyramid Representation

- Layered representation, with decreasing spatial resolution
- Bidirectional connections between locations in adjacent layers
- Coarse-to-fine search
- Processing need only consider input from previous layer (supported by Hubel and Wiesel, 1965)
- Reduces P

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$$O(P^2 2^{PM} \lg N)$$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$$O(NP^2 2^{PM})$$

Spatiotemporal Localization

- Reduce number of possible receptive fields from $O(2^P)$ to $O(P^{1.5})$
- Assumption: contiguous receptive fields of all possible sizes centered at all locations in the image

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$O(P^{3.5}2^M \lg N)$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$O(NP^2 2^{PM})$

Receptive Field Selectivity

- Selection not only of location, but also of a local region or size
- Reduces P to P’

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$O(P'2^{M}lgN)$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$O(NP^{2}2^{PM})$

Feature Selectivity

- Reduce features to those actually present in the image or important to task at hand
- Reduces M to M’

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$O(P'M'\lg N)$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$O(NP^2 2^{PM})$

Model Space Selectivity

- Reduce model space to encompass relevant objects/events
- Reduces N to N’

(1) Hierarchical Organization
(2) Pyramid Representation
(3) Spatiotemporal Localization
(4) Receptive Field Selectivity
(5) Feature Selectivity
(6) Model Space Selectivity

$O(P'M'N')$

- N = # prototypes
- P = # pixels
- M = # features

compare with:

$O(NP^2 2^{PM})$

10 Problems with Pyramids

(1) Boundary Problem

- increasingly more of the periphery unrepresented at higher layers of the pyramid

(2) Blurring Problem

- large portion of output layer affected by single events at input layer

(3) Cross-talk Problem

- two separate visual events activate overlapping subpyramids (interference between events)

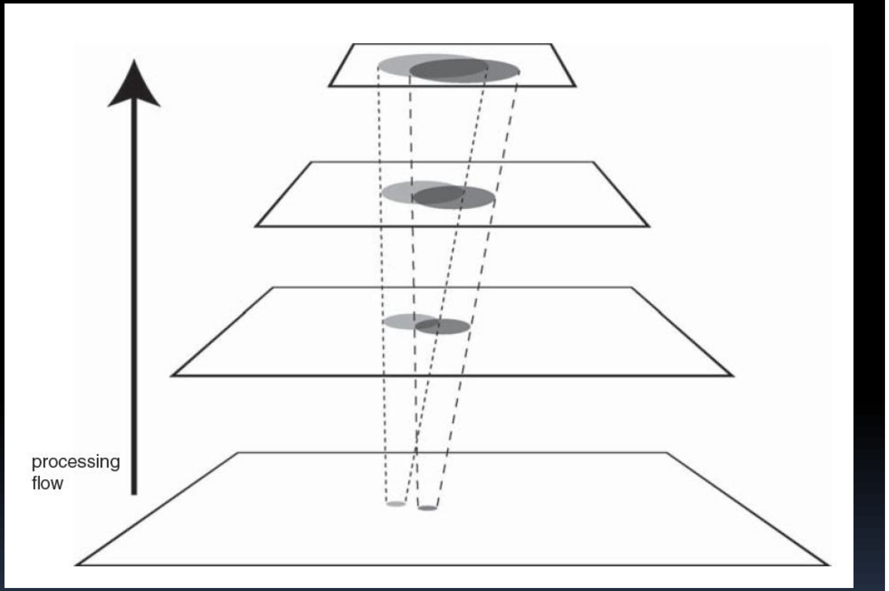
(4) Context Problem

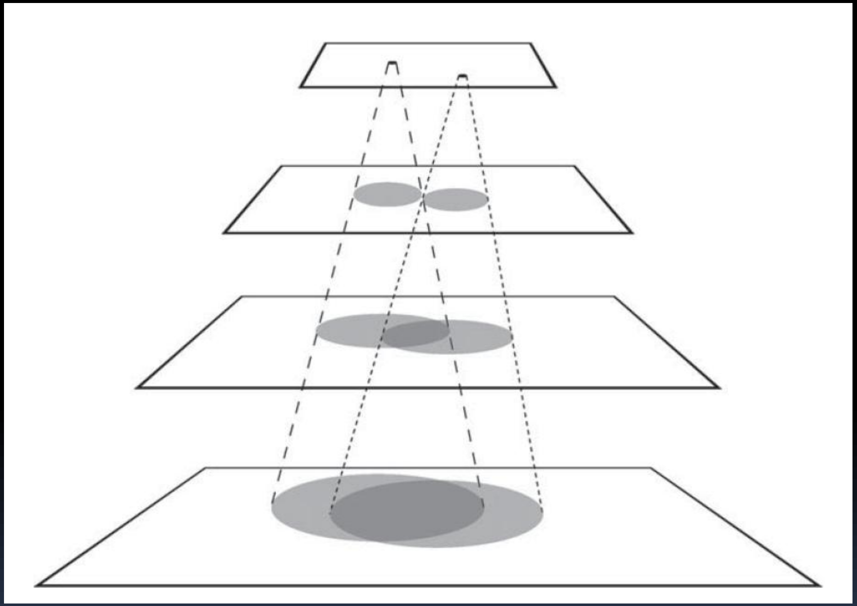
- neuron at higher level of the pyramid affected by an extended portion of the visual field (an event along with its context)

(5) Routing Problem

- exponential number of feedforward and feedback connections (which pathways represent the best interpretation of the stimulus?)

Problems with Pyramids





Images copyrighted: "A Computational Perspective on Visual Attention, John K. Tsotsos, The MIT Press, Cambridge MA, 2011"

Problems with Pyramids

(6) Multiple Foci Problem

- can we attend to more than one stimuli at once? What if there is interference in receptive fields?

(7) Convergent Recurrence Problem

- top-down signals overlap and converge in their receptive fields in lower layers of the pyramid

(8) Spatial Interpolation Problem

- inverse of sampling problem, by considering feedback direction

(9) Spatial Spread Problem

- inverse of context problem, by considering top-down information flow (divergence of signal)

(10) Lateral Spread Problem

- Due to lateral connectivity, signals may spread rapidly to neighboring neurons, obfuscating the signal

Why do we need attention?

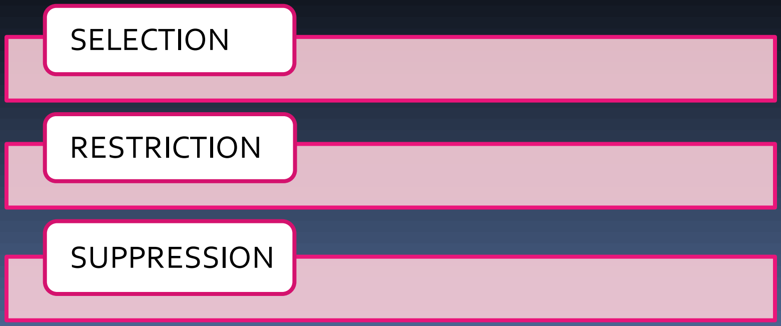
- to select region of interest
- to select features of interest
- to control information flow
- to solve context, blurring, cross-talk, boundary problems
- to solve order of selection/reselection for time-varying scenes
- to balance data-directed and task-directed processing

"Attention is a set of mechanisms that help tune and control the search processes inherent in perception and cognition."

- John Tsotsos, 2011

Information Reduction Mechanisms

- Selection:
  - Spatiotemporal region of interest
  - Features of interest
  - World. Task, object, event models
  - Gaze and viewpoint
  - Best interpretation of response



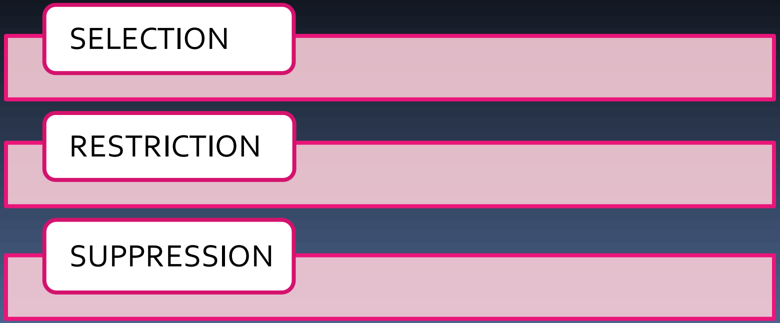
SELECTION

RESTRICTION

SUPPRESSION

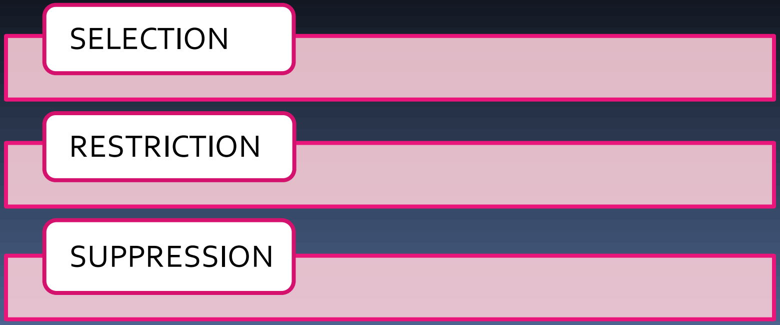
Information Reduction Mechanisms

- Restriction:
  - Task-relevant search-space pruning (priming)
  - Location cues
  - Fixation points
  - Search depth control during task satisfaction
  - Modulating neural tuning profiles



Information Reduction Mechanisms

- Suppression:
  - Inhibition of return
  - Spatial and feature surround inhibition
  - Suppression of task-irrelevant computations



Attentive Suppression

- Overall response of the neuron is a function of both signal (preferred stimulus within receptive field) and noise (everything else in receptive field)
- Inhibition-on-surround (e.g. Gaussian-shaped) to suppress noise

Attentive Processes that reduce complexity of visual processing

- Interest point operators
  - Good features can be located unambiguously in different views of a scene
  - Local maxima of directional variance measures

FIT (Feature Integration Theory):
feature maps are computed in parallel, while chosen points can be processed serially
BUT experimental data demonstrates no clear separation of strategies into serial or parallel

- Perceptual grouping
  - Limit possible combinations of positions/regions under consideration using cues of proximity, similarity, collinearity, symmetry, familiarity
  - Task-dependent choice of cues (and interaction of cues)
- Active vision
  - Data acquisition process depends on current state of data interpretation
  - E.g. salience-based fixation control
- Predictive methods
  - Domain and task knowledge to guide processing (planning)

Monkey visual cortex (Buschman and Miller, 2007)

- Top-down signals arise from frontal cortex
  - prefrontal neurons reflected target location during top-down attention
- Bottom-up signals arise from sensory cortex
  - parietal neurons signaled target location earlier during bottom-up attention
- Different modes of attention implied by synchrony at different frequencies
  - Low frequency for top-down attention
  - High frequency for bottom-up attention

Pyramid Vision (Burt, 1998)

3 Elements of Attention:

- Foveation to examine selected regions of the visual world at high resolution
- Tracking to stabilize the images of moving objects within the eye
- Interpretation (high-level) to anticipate where salient information will occur in a scene

Gain Modulation

Note: mechanism by which this occurs is still unclear!

- Change in the response amplitude of a neuron through the nonlinear combination of sensory, motor, and/or cognitive information
  - close to multiplicative interactions
- An input may affect the gain of the neuron to another input without changing the receptive field properties of the neuron (Salinas and Sejnowski, 2001)
- Gain fields: distributed, multi-modal representations

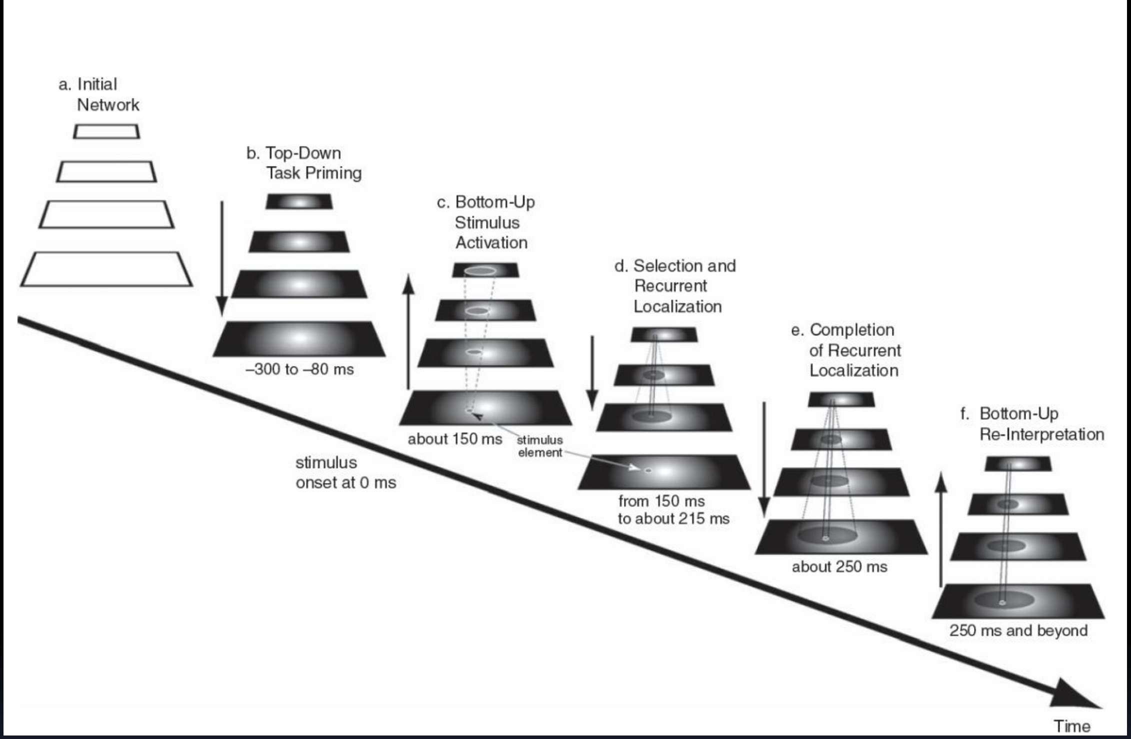
2 Models of Gain

Which model best represents attention function is still debated!

- **Contrast Gain Model:** attention modulates the effective contrast of stimuli at attended locations (multiplication of contrast necessary to reach a given level of response)
- **Response Gain Model**: attention causes neuronal response to be multiplied by a constant gain factor, resulting in increases in firing rate that grow larger with contrast

Task-dependency

- Attentional process selects the strongest-responding neurons over the task-relevant representation
- Categorization requires a single feed-forward pass through the visual system (~150 ms)
- Identification requires traversing the hierarchy downward, beginning with the category neuron and moving down through afferent neurons (~65 ms)
  - Extent of downward traversal is task-dependent
- Localization requires additional time for a complete top-down traversal (~250 ms)
  - Visual processing time difficult to disentangle from motor processing time (for pointing at stimulus)


a. Initial Network
b. Top-Down Task Priming
–300 to –80 ms
c. Bottom-Up Stimulus Activation
about 150 ms
stimulus element
stimulus onset at 0 ms
d. Selection and Recurrent Localization
from 150 ms to about 215 ms
e. Completion of Recurrent Localization
about 250 ms
f. Bottom-Up Re-Interpretation
250 ms and beyond
Time

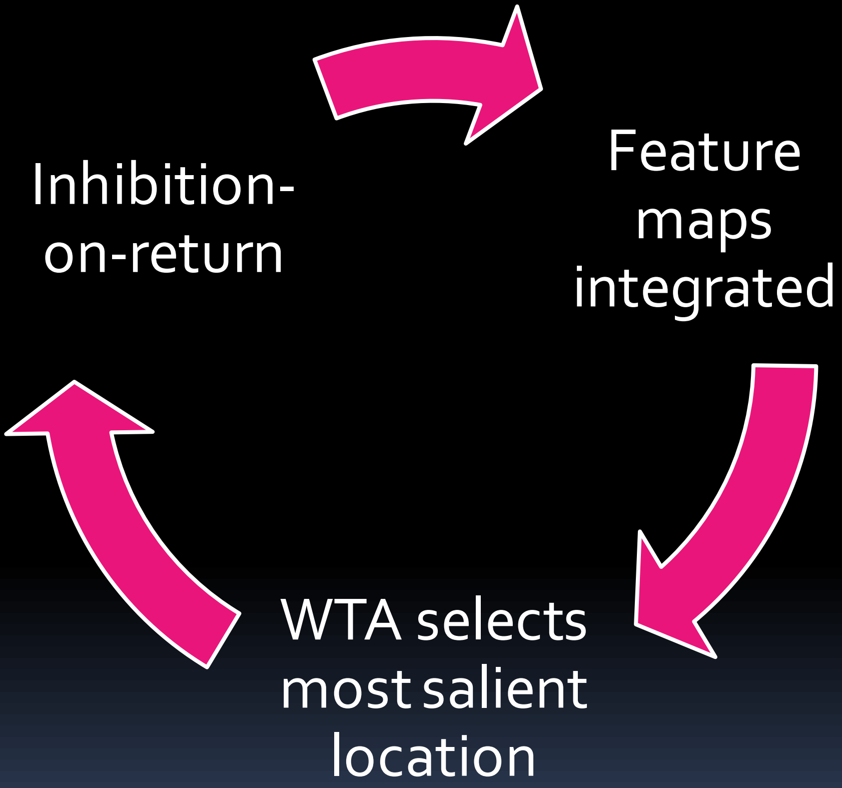
Computational models

- Selective Routing
  - Stimulus selection and transmission through visual cortex (how signals in the brain are transmitted to ensure correct perception)
- Temporal Tagging
  - Correlated neuronal activity (synchronized firing)
- Emergent Attention
  - Attention as a property of large assemblies of neurons involved in competitive interactions
  - Selection as a result of local dynamics and top-down biases
- Saliency Map

Saliency

- Features: color, orientation, curvature, texture, scale, vernier offset, size, spatial frequency, motion, shape, onset/offset, pictorial depth cues, stereoscopic depth
- Validation: individual neurons selective for oriented bars, binocular disparity, speed of translational motion, color opponency, etc. (even for particular faces)
- BUT how does binding occur?

Saliency Map (Koch and Ullman, 1985)



Inhibition-on-return

Feature maps integrated

WTA selects most salient location

Human cortex studies demonstrated that maxima of response that are found within a neural population correspond with the attended location.

Problem with current models

- Focus on the manifestations not the causes of attention
- Models are developed at different levels of abstraction, with differing functional components, and designed for specific purposes, making comparison of models difficult
- Attention should not be modeled as monolithic, but a set of interacting mechanisms
- Considerations of information flow complexity are necessary

PART II: A computational model of attention

John Tsotsos (first introduced in 1990)

SELECTIVE TUNING (ST)

Representation

- Pyramid: layers of neurons with decreasing spatial resolution and increasing abstraction
- Connectivity: feedback, feedforward, lateral connections (outgoing are diverging, incoming are converging)
- Decisions: made in a competitive manner; occur at the level where task-relevant information is computed
- Selection: WTA at a given layer; hierarchical strategy, recursively examining only afferent going to best responses (in a top-down manner)
- Refinement: final feed-forward pass to reinterpret stimulus within reduced context

TO BE CONTINUED . . .