A Unified Theory of Exogenous and Endogenous Attentional Control

Michael C. Mozer

Institute of Cognitive Science and
Department of Computer Science
University of Colorado, Boulder

Matthew Wilder

Department of Computer Science
University of Colorado, Boulder

David Baldwin

Department of Computer Science
Indiana University

Visual Search

Find the 20p coin

Find a coin that isn't round.

Are there more heads or tails?

How many gold coins are there?

Are all the coins British?

Are any coins the wrong size?

Visual Search

How is the visual system dynamically reconfigured to perform a remarkable variety of arbitrary tasks?

Attentional control

The ability to flexibly modulate attentional selection and visual perception based on task demands

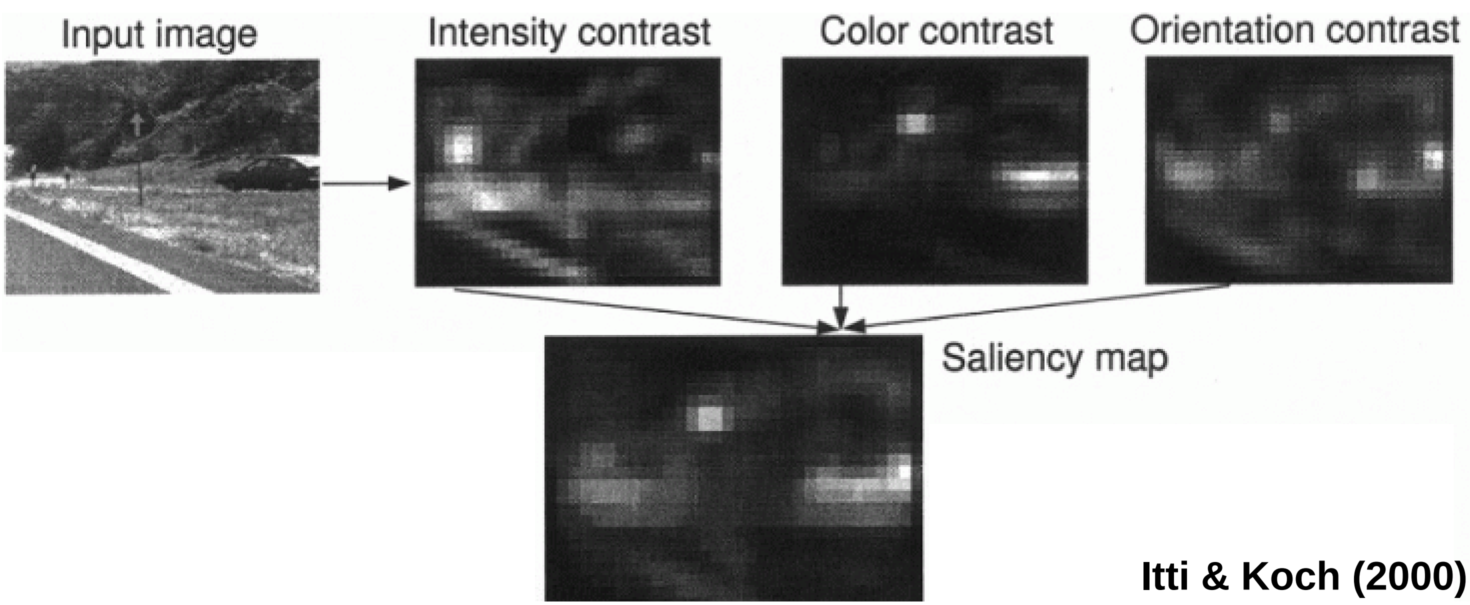
Attentional Control Strategies

Attentional Control Strategies

Exogenous

Attention guided to distinctive, locally contrasting visual features such as color, luminance, and texture discontinuities, and abrupt onsets.

e.g., Averbach & Coriell (1961); Posner & Cohen (1984); Itti & Koch (2000); Koch & Ullman (1985)



Itti & Koch (2000)

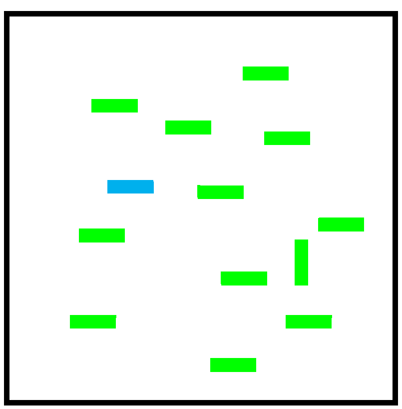
Attentional Control Strategies

Exogenous

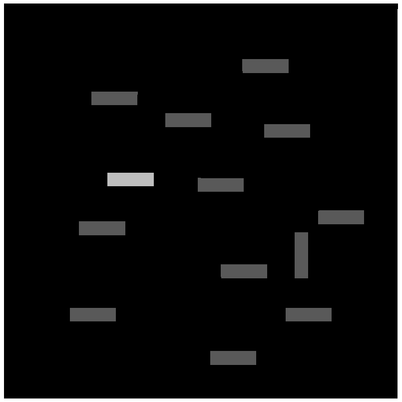
Feature-Based Endogenous

Attention guided to task-relevant features.

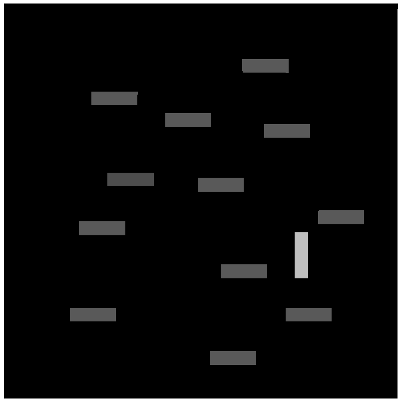
e.g., Baldwin & Mozer (2006); Mozer (1991); Navalpakkam & Itti (2005); Wolfe (1994)



find
blue



find
vertical



Attentional Control Strategies

Exogenous

Feature-Based Endogenous

Scene-Based Endogenous

Attention guided to regions of interest based on task and global scene gist.

e.g., Neider & Zelinsky (2006); Torralba, Oliva, Castelhano, & Henderson (2006)

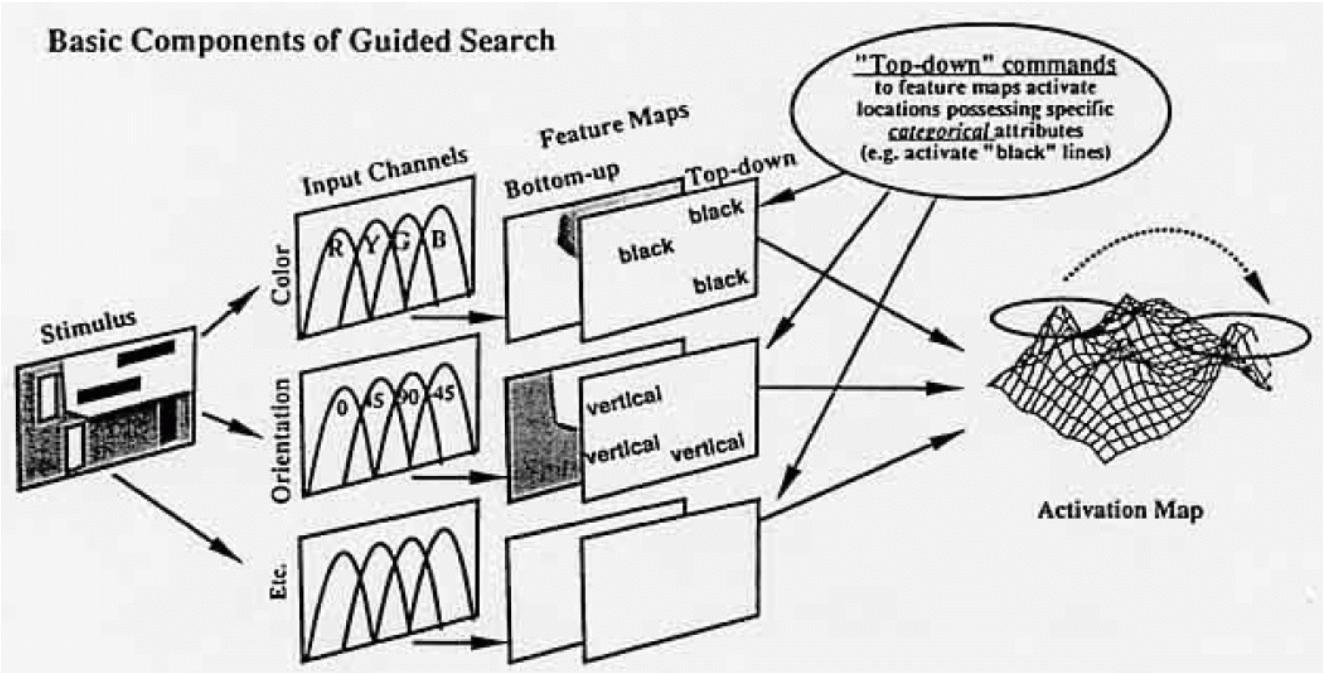
Torralba et al. (2006)

Theories of Attentional Control

If strategies are distinct, more than one might be applied in any situation.

Control processes need to arbitrate or integrate across strategies.

E.g., Wolfe (1994)



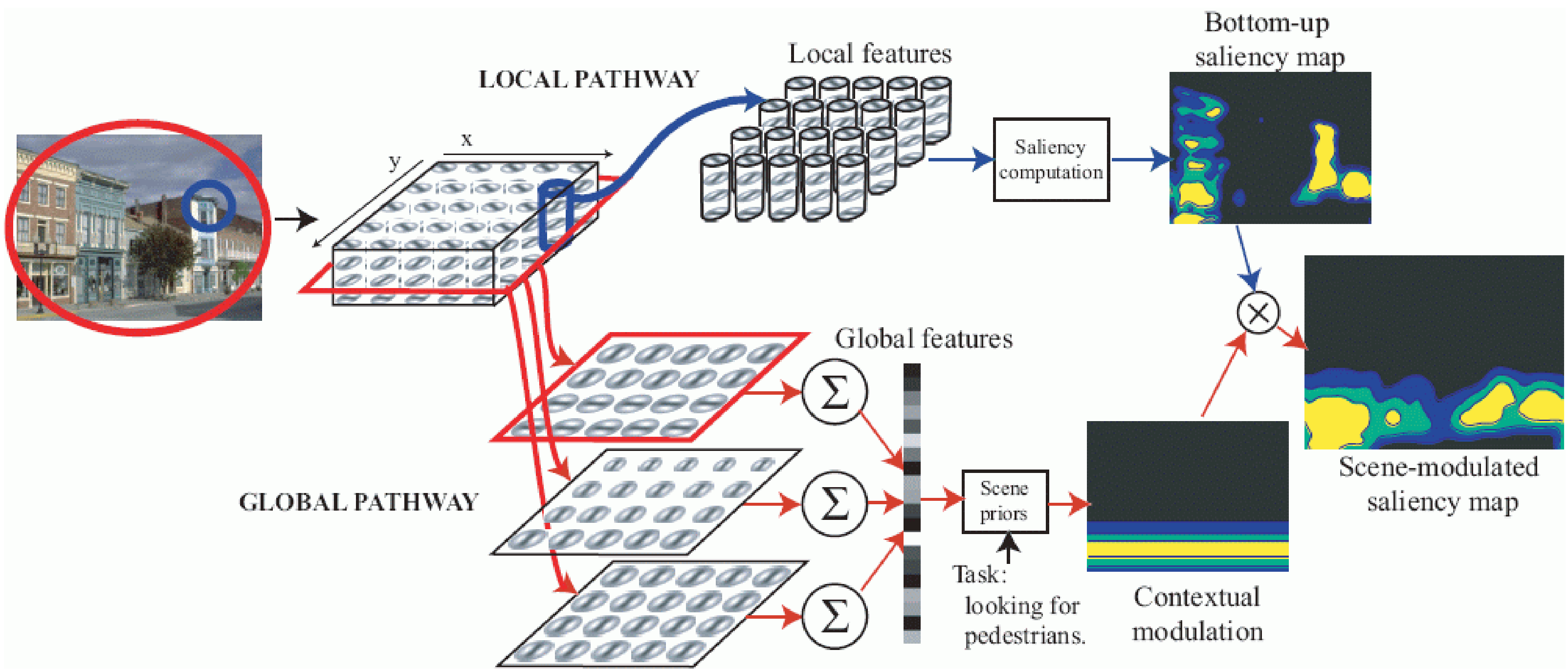
Theories of Attentional Control

If strategies are distinct, more than one might be applied in any situation.

Control processes need to arbitrate or integrate across strategies.

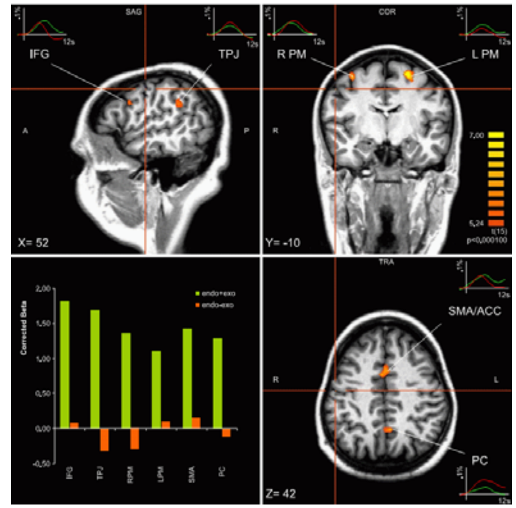
E.g., Wolfe (1994)

E.g., Torralba et al. (2006)



No Evidence for Distinct Mechanisms

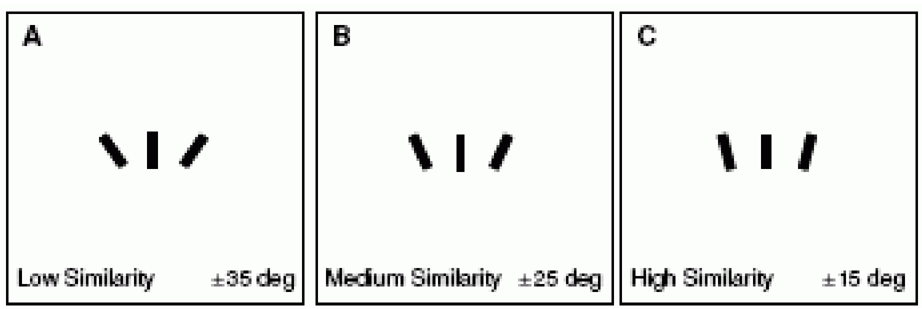
Neuroimaging suggests that endogenous and exogenous control do *not* involve distinct neural systems (e.g., Rosen et al., 1999; Peelen et al. 2004)



Peelen, Heslenfeld, & Theeuwes (2004)

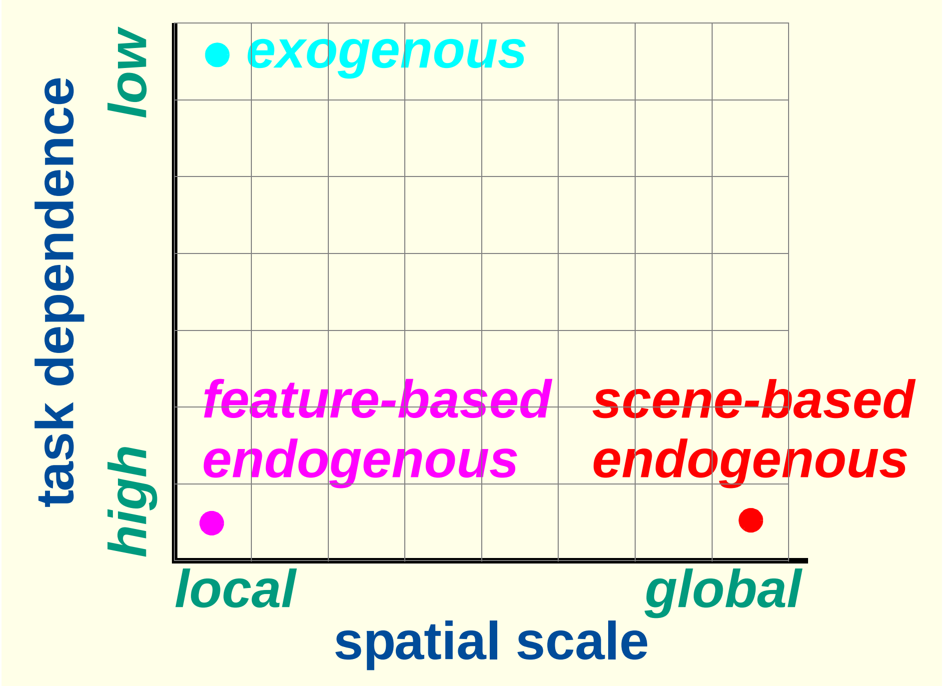
Behavioral data suggests a trade off among control strategies.

Increasing task difficulty via target/nontarget similarity decreases impact of an irrelevant singleton in brightness (Proulx & Egeth, 2006; Theeuwes, 2004).



A Unified Theory

Instead of conceiving of these strategies as three distinct and unrelated mechanisms, we characterize them as points in a *control space*.



Weak hypothesis

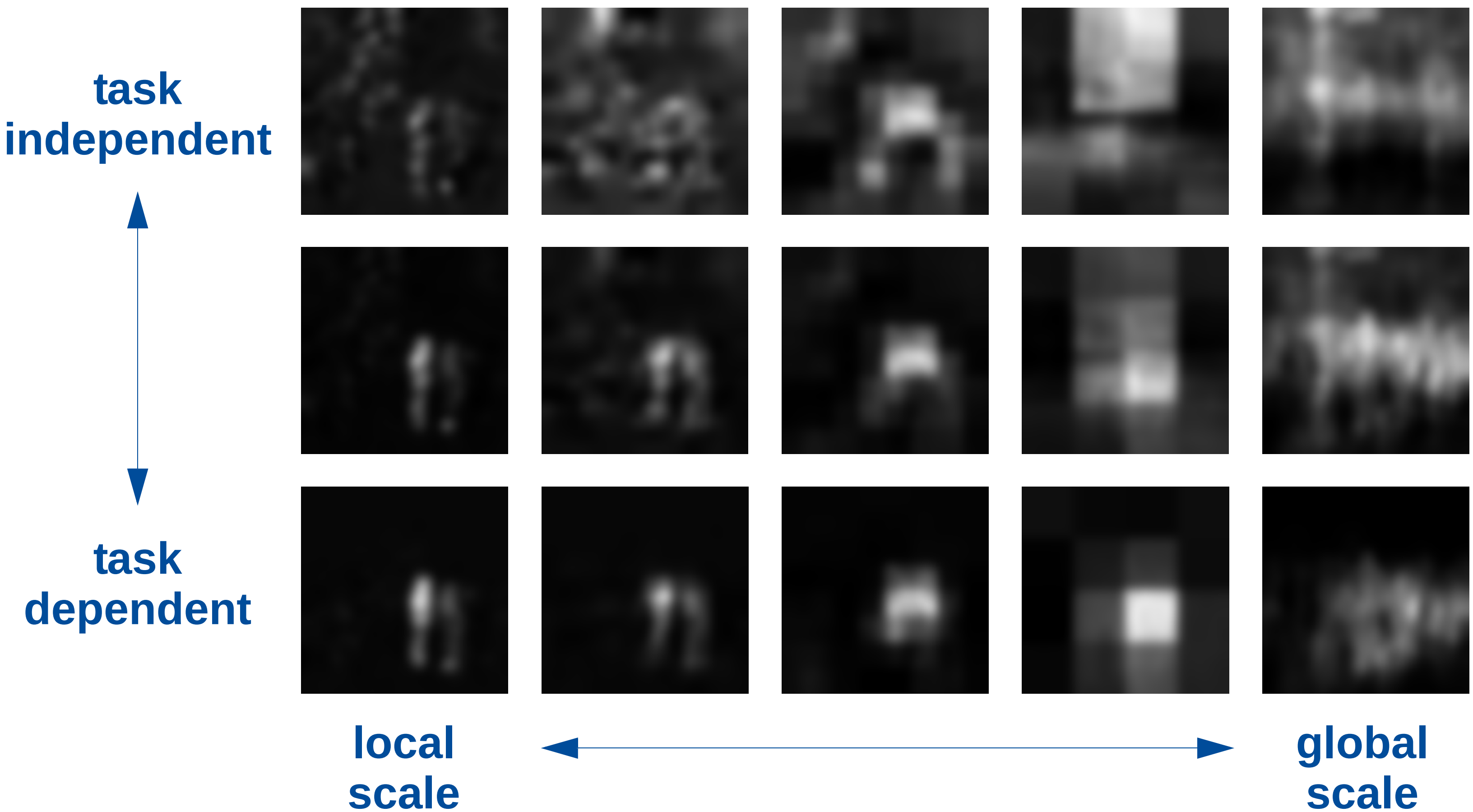
Control space offers a unified view and insights into the relationships among strategies.

Strong hypothesis

Attentional control at a particular instant for a particular task is defined by a single point in the space.

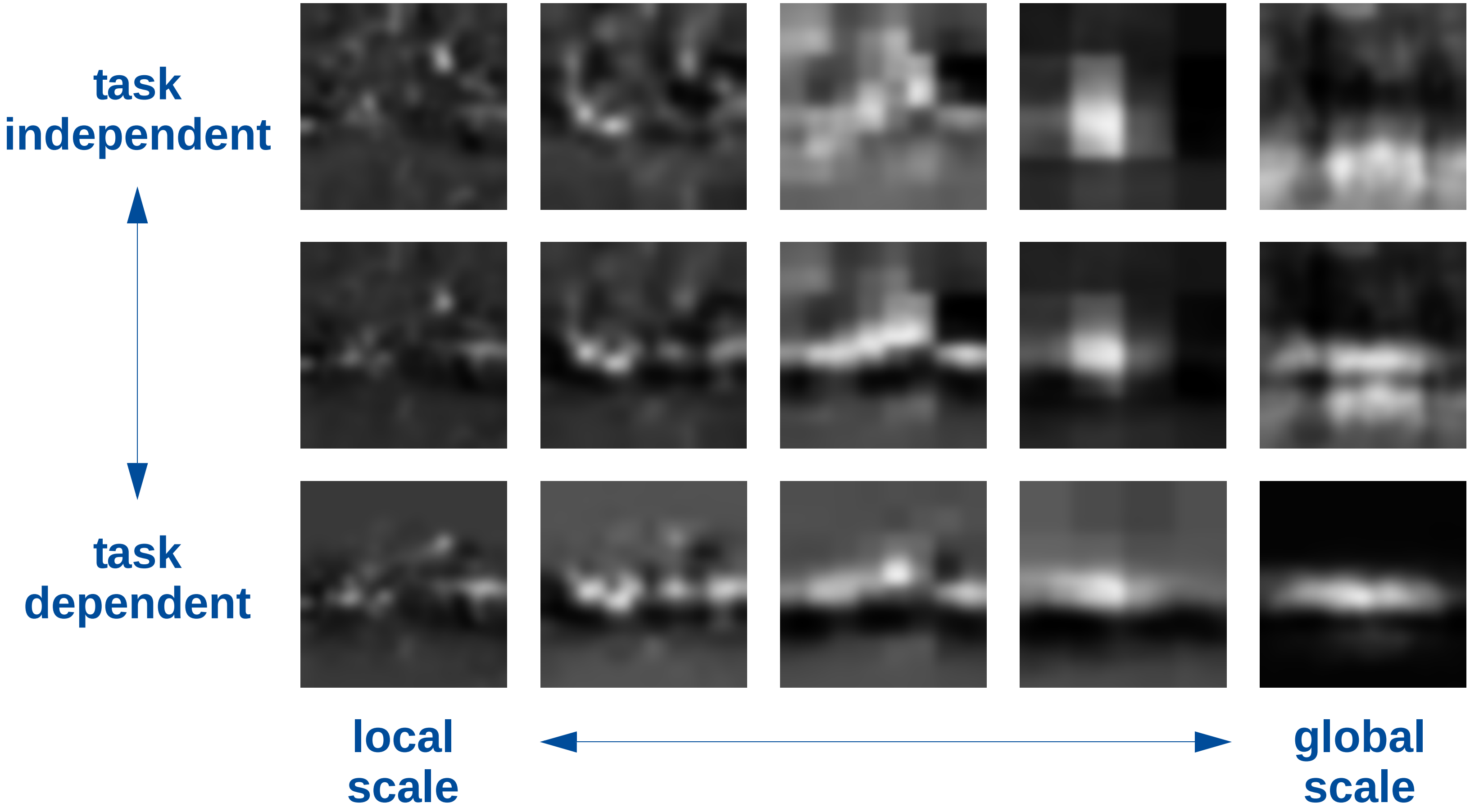
Example: Saliency Maps Over Control Space

Task: search for person



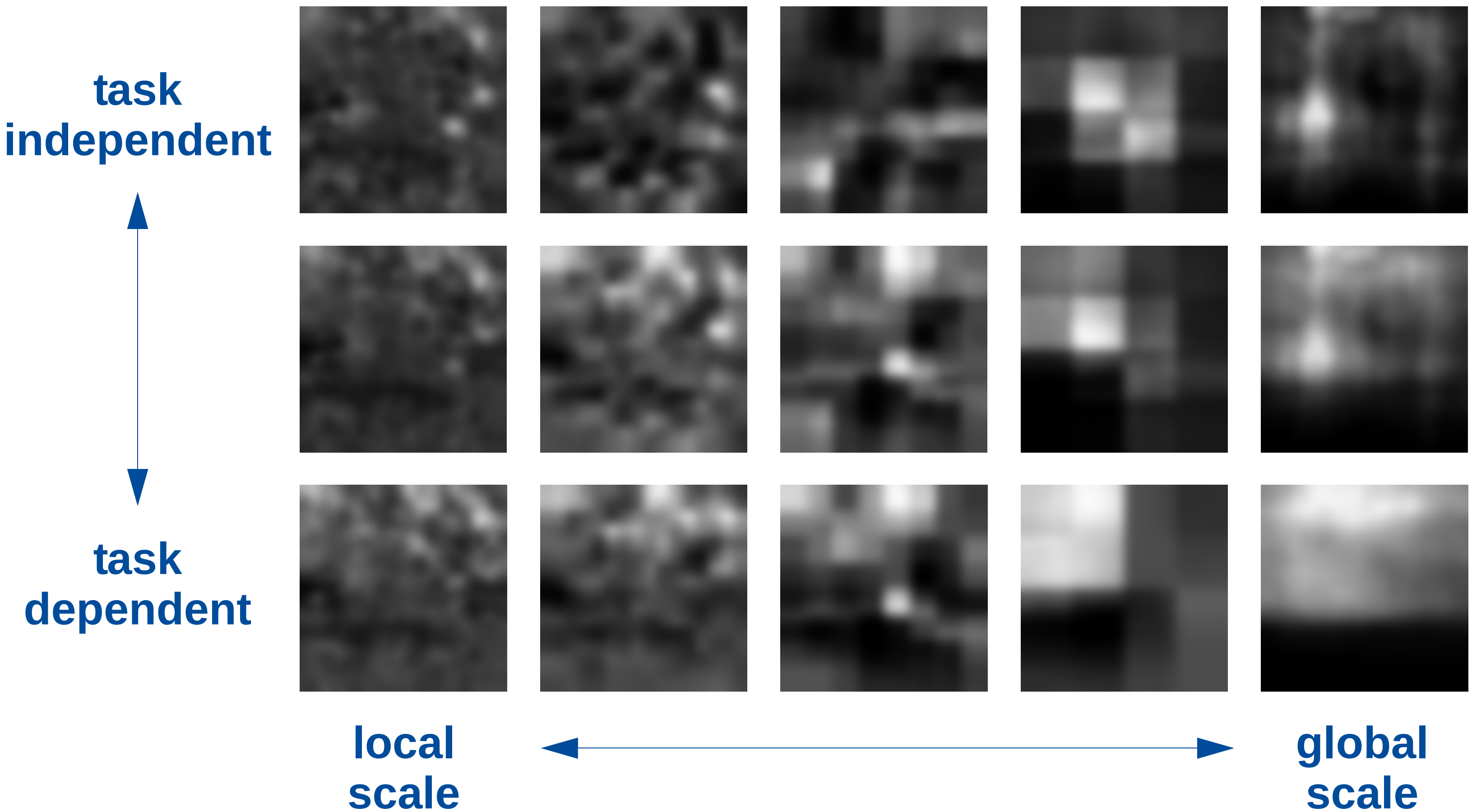
Example: Saliency Maps Over Control Space

Task: search for car



Example: Saliency Maps Over Control Space

Task: search for building



Our Framework

Input

images of real-world scenes and stimulus displays

Output

saliency map

Given current goals, model determines control parameters

- spatial scale
- task dependence
- object models to incorporate

Given control parameters, model configures processing pathway.

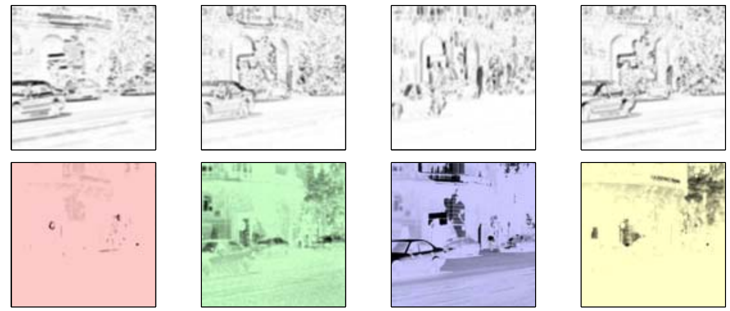
Processing Pathway: Generalizing Earlier Models

stage	Navalpakkam & Itti (2005); Wolfe (1994)	Torralba et al. (2006)
parallel feature detection with broad, overlapping tuning curves	color, orientation, luminance	color; orientation at multiple spatial scales
contrast enhancement via center-surround differencing	yes	yes, sort of, via cross-dimensional normalization
dimensionality reduction	no	yes
associative network to compute saliency	linear	mostly linear with a Gaussian squashing function

Processing Pathway: Preprocessing Image

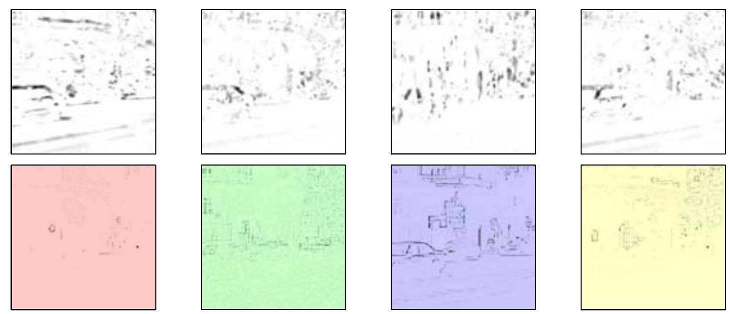
Feature extraction

local Gabor, RGBY filtering



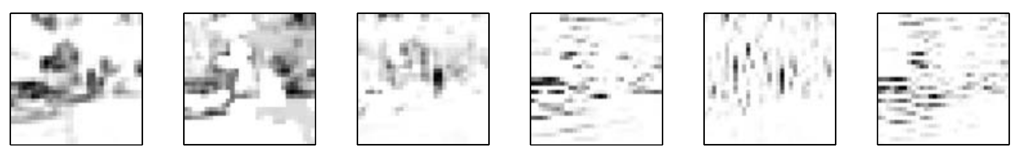
Contrast enhancement

center-surround differencing



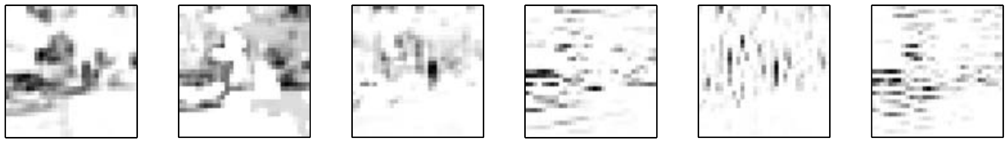
Dimensionality reduction

subsampling, PCA



Processing Pathway: Saliency Network

Preprocessed Representation

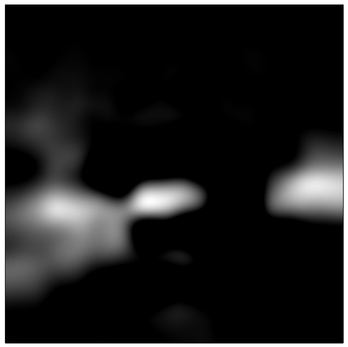


Association

rank-limited *linear* transform

Saliency map

linear summation across patches



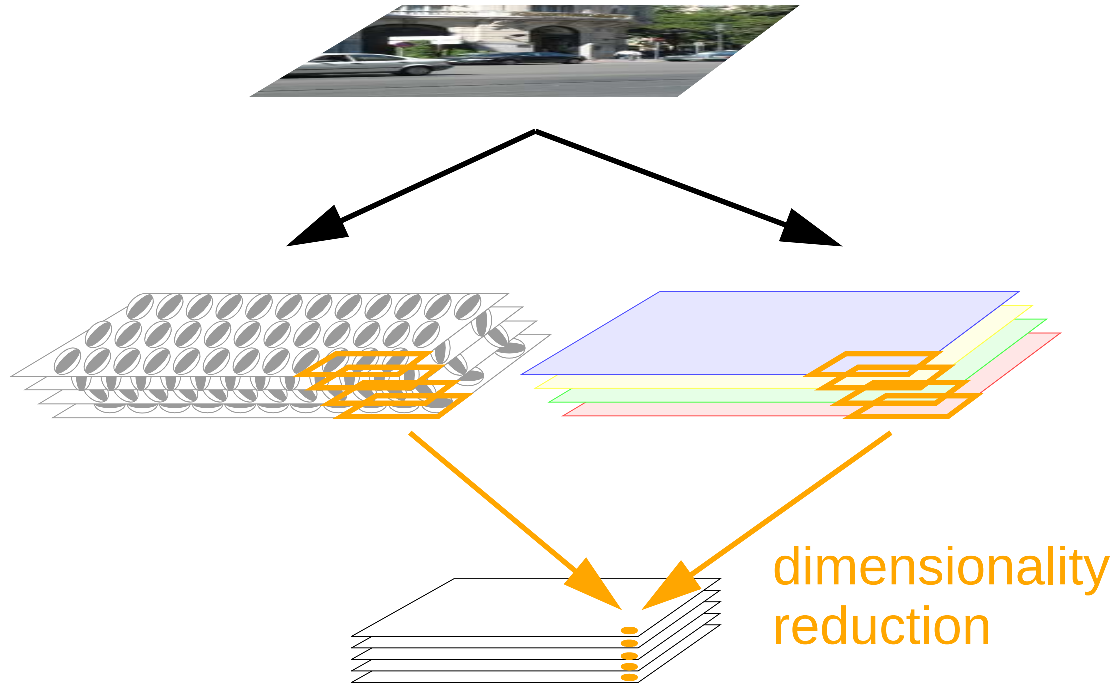
Implementation

Preprocessing

Dimensionality reduction via PCA on image patches

Trained with large natural image corpus

Location invariant



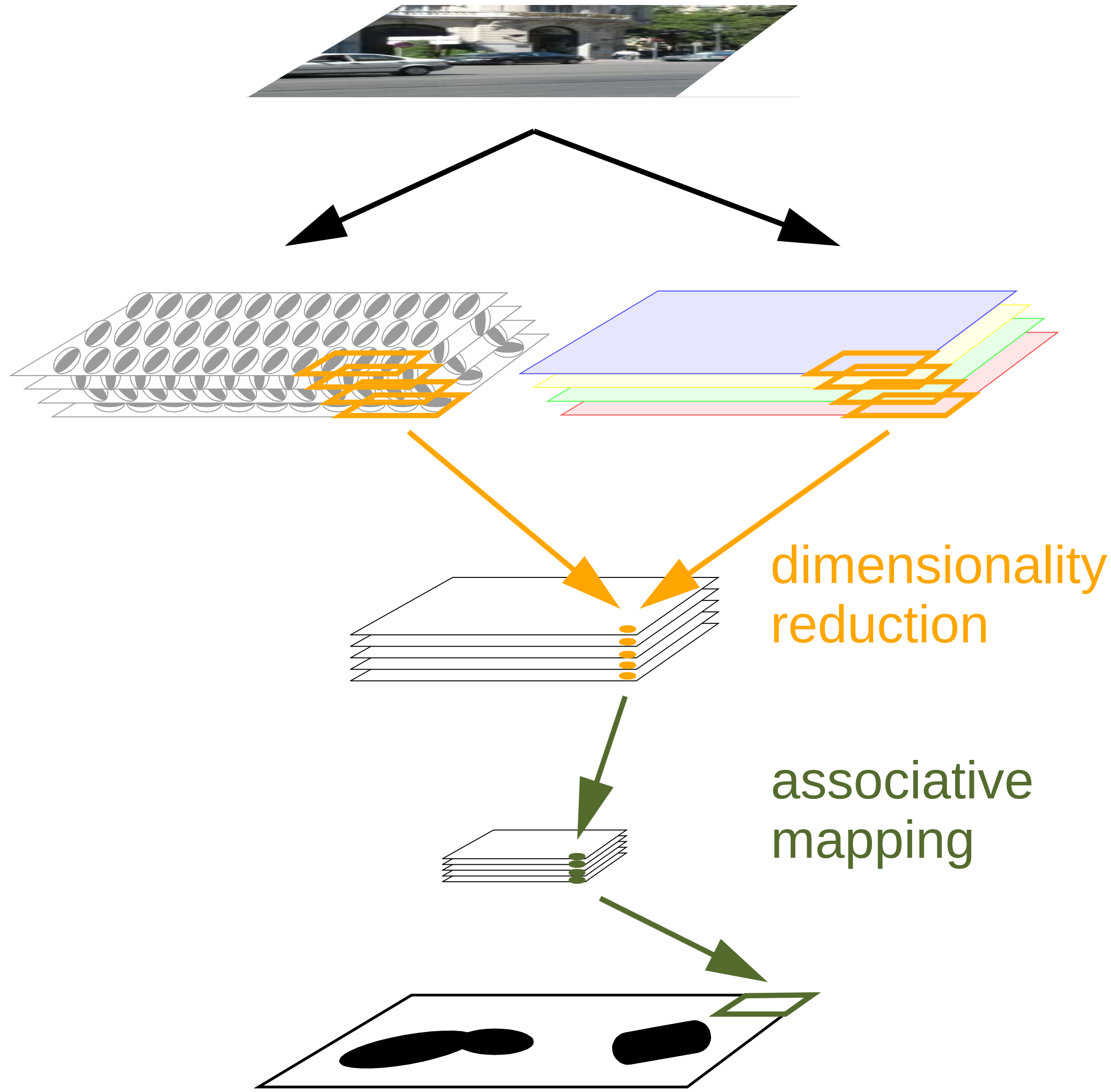
Implementation

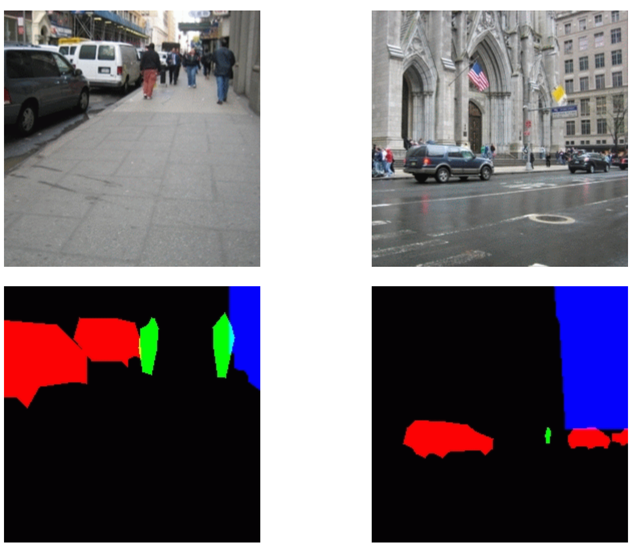
Preprocessing

Task dependent learning

Patches processed along parallel channels with separate learned connection strengths for each channel, task, and spatial scale.

Task: search for target object (car, person, building, lamp, tree, road, window, sign)





from LabelMe data base (Torralba and collaborators)

Implementation

Preprocessing

Task dependent learning

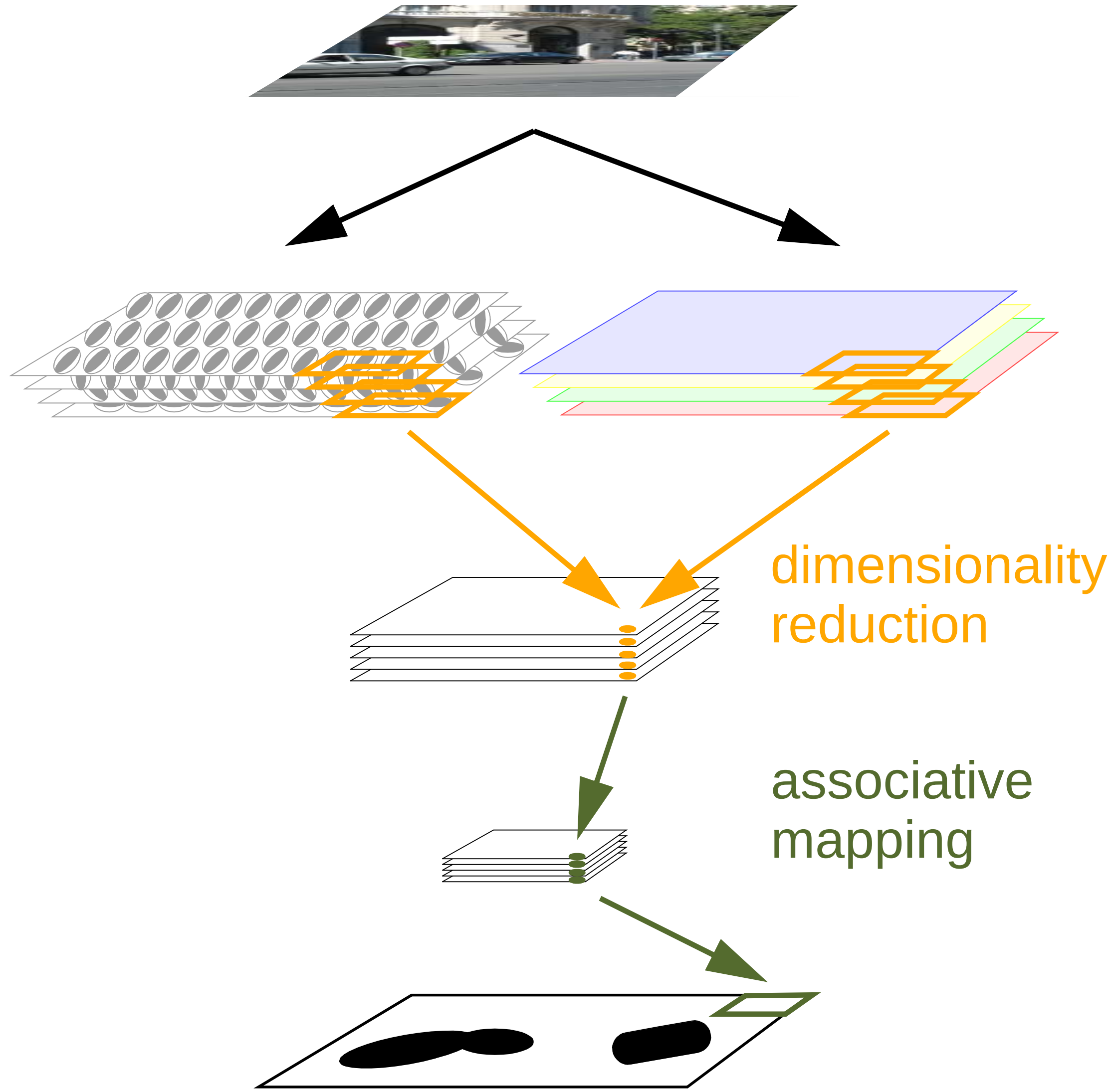
Control space

Spatial scale

- Diameter of overlapping receptive fields varied from 3% to 100% of image

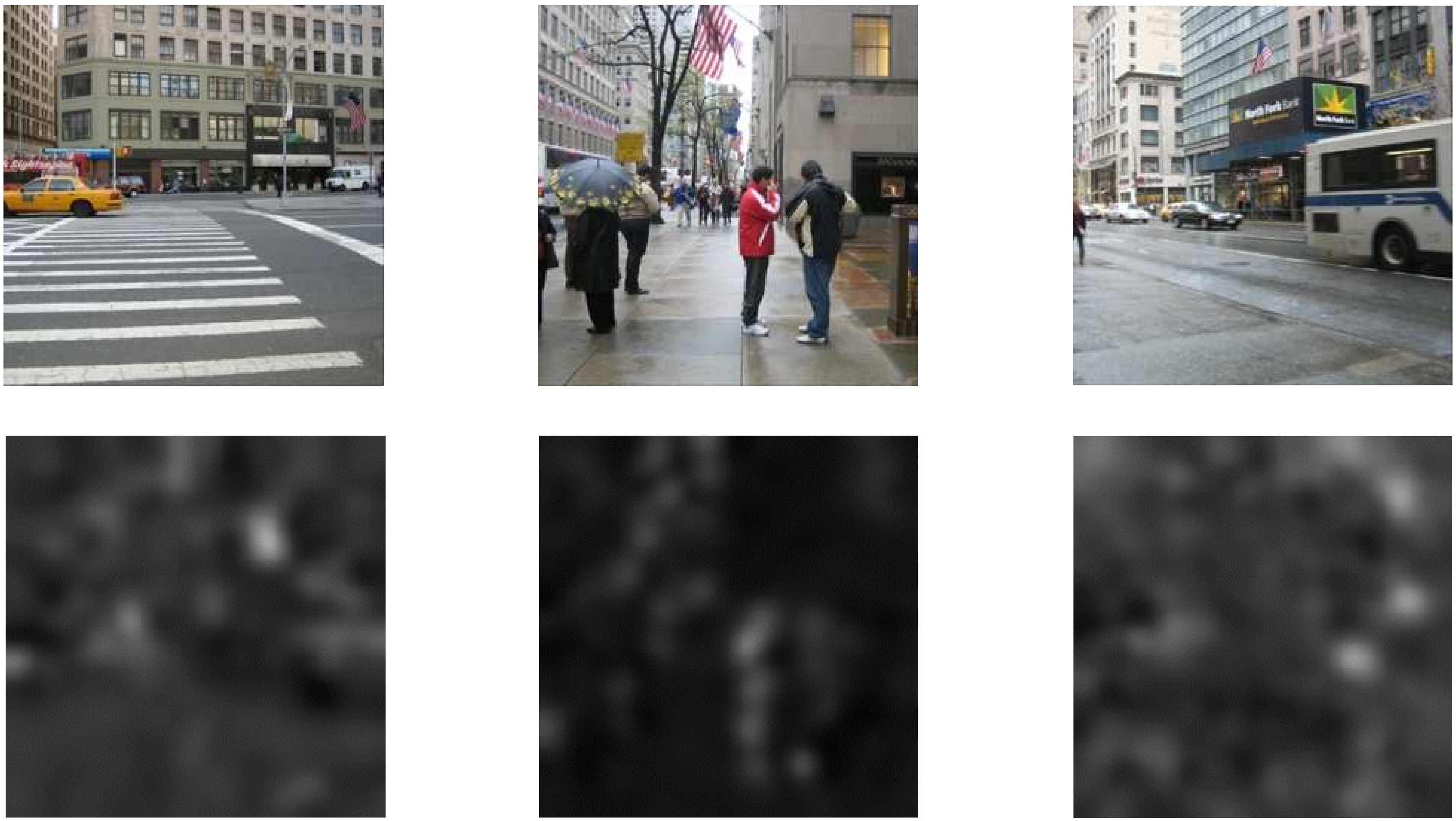
Task dependence

- Task-independent pathway is *average* of task-specific pathways.
- Intermediate task dependence via interpolation



Results: Exogenous Control

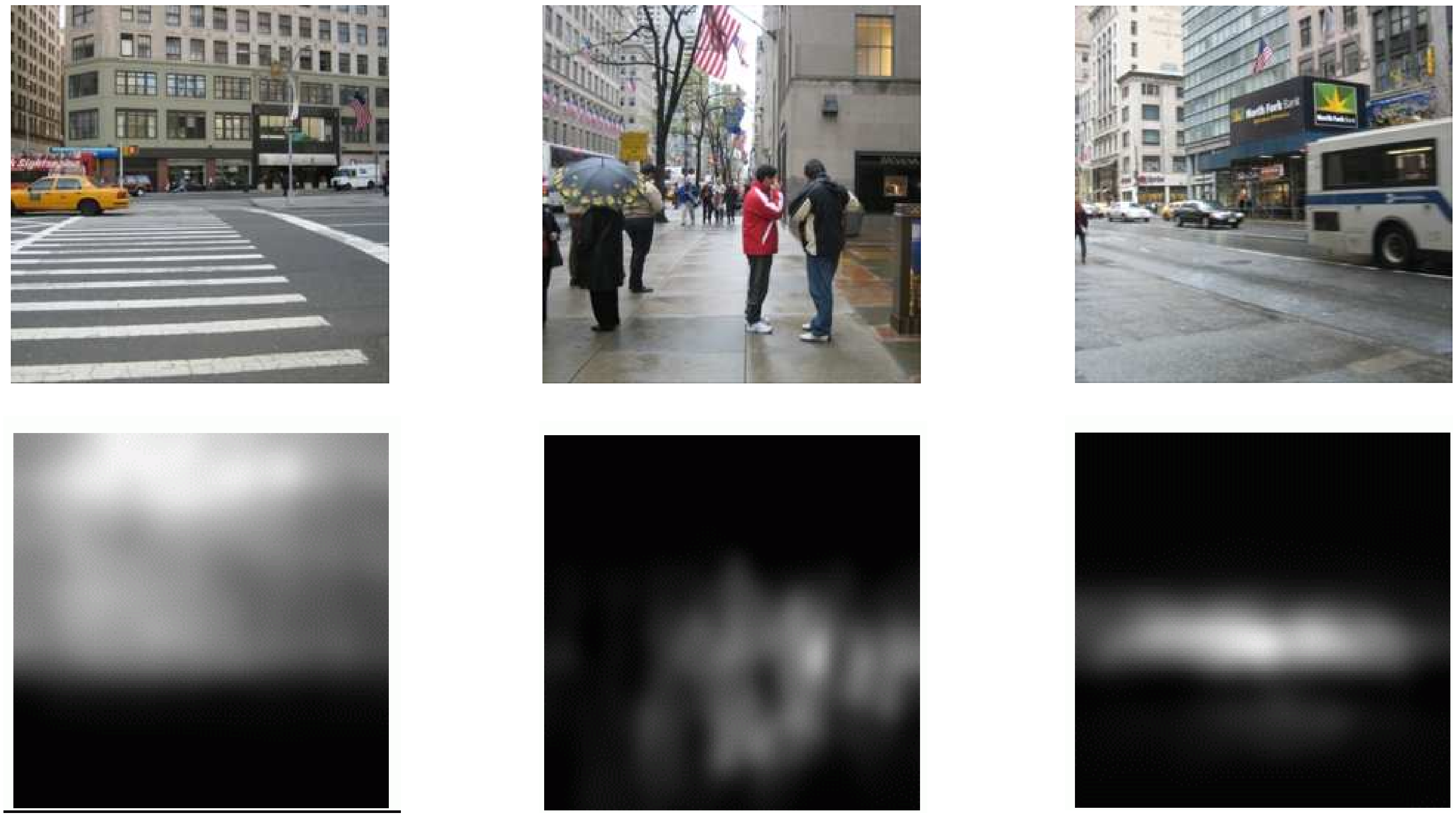
Fine scale, task independent pathway



Need larger data base;
Need to evaluate on Bruce & Tsotsos eye movement data set

Results: Contextual Guidance

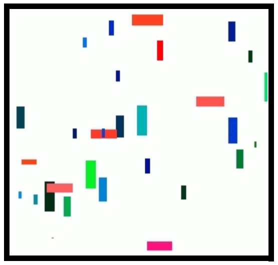
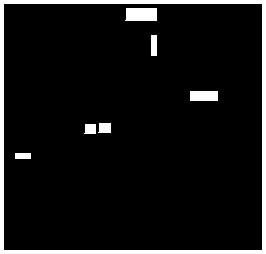
Coarse scale, task dependent pathway



Model produces results qualitatively similar to Torralba et al.

Results: Simple Feature Search

Train task-specific model for each feature

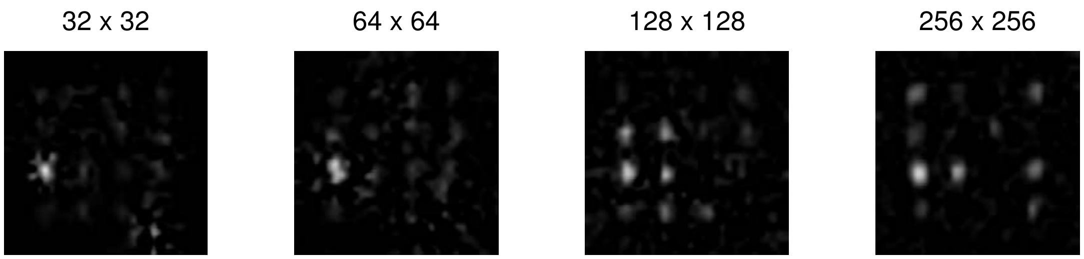
 

Results: Simple Feature Search

Train task-specific model for each feature

Evaluate saliency on test displays of varying size

RT ~ –log(proportion of saliency on target)

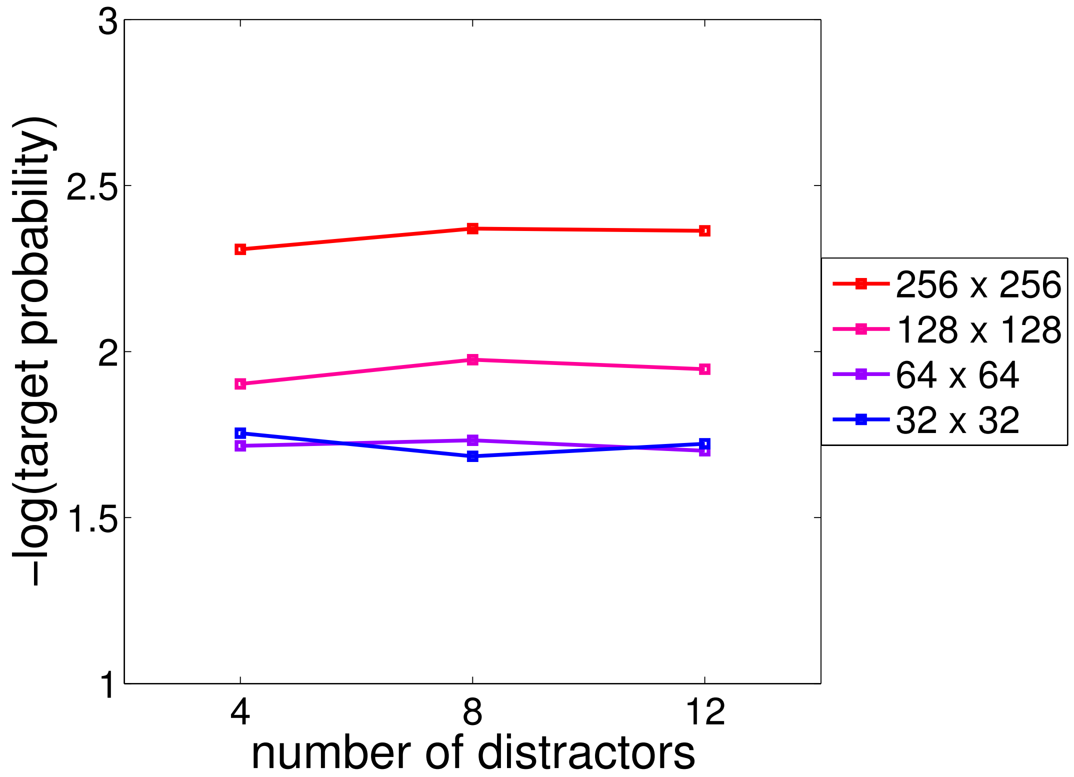


Results: Simple Feature Search

Train task-specific model for each feature

Evaluate saliency on test displays of varying size

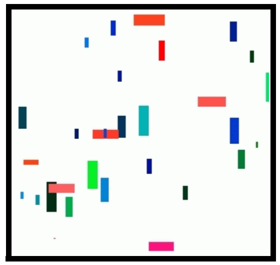
RT ~ –log(proportion of saliency on target)



Results: Conjunction Search

Train task-specific model for conjunction (red vertical)

Or train single features and combine models (red+vertical)



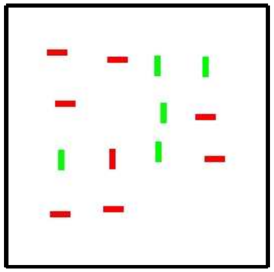
Results: Conjunction Search

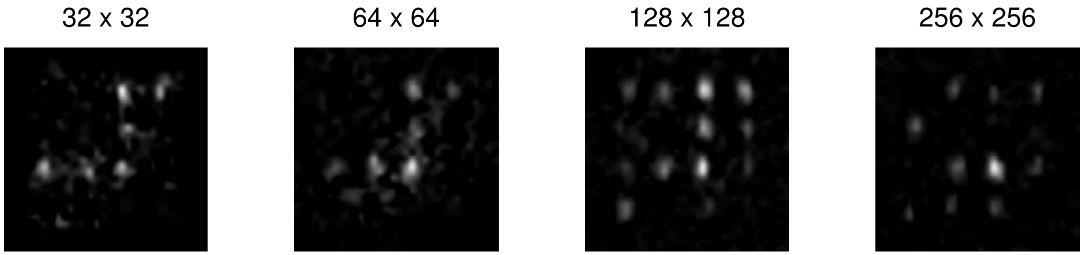
Train task-specific model for conjunction (red vertical)

Or train single features and combine models (red+vertical)

Evaluate saliency on test displays of varying size

RT ~ –log(proportion of saliency on target)





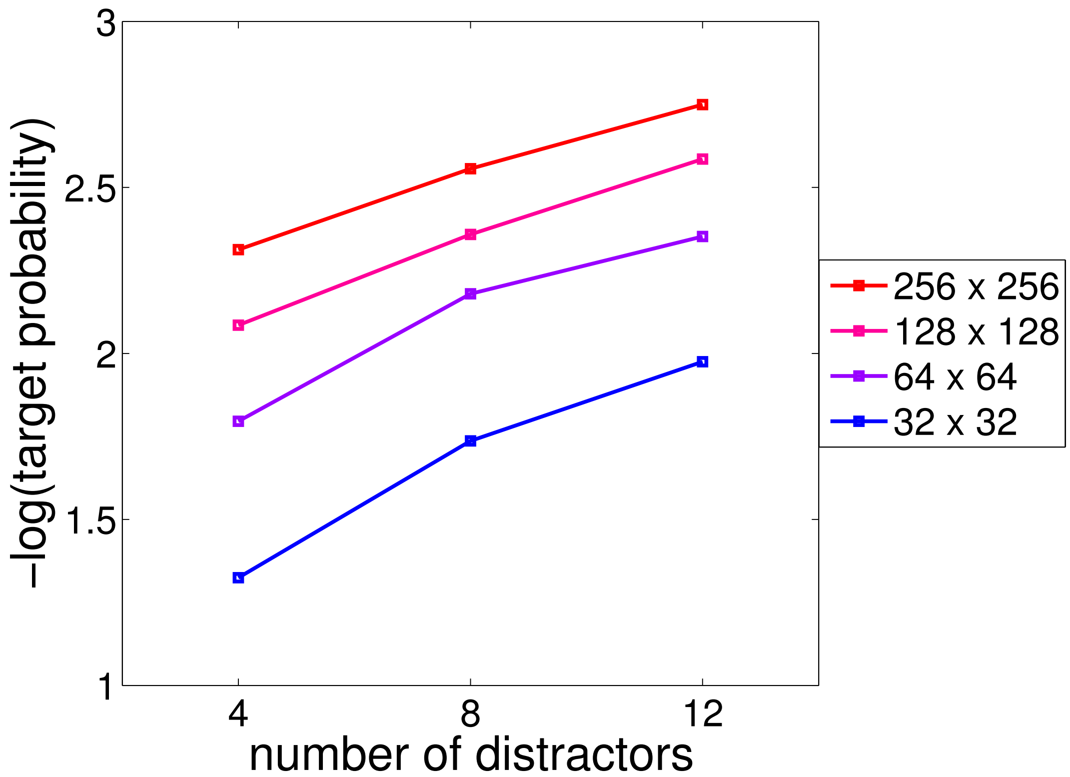
Results: Conjunction Search

Train task-specific model for conjunction (red vertical)

Or train single features and combine models (red+vertical)

Evaluate saliency on test displays of varying size

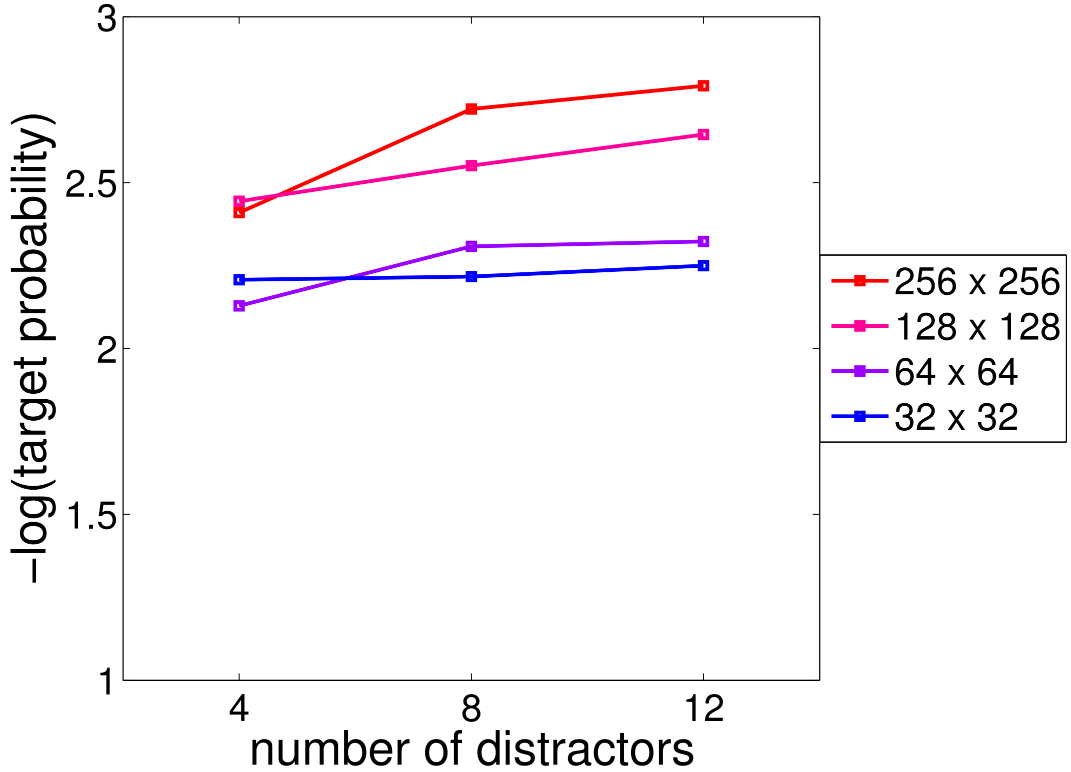
RT ~ –log(proportion of saliency on target)



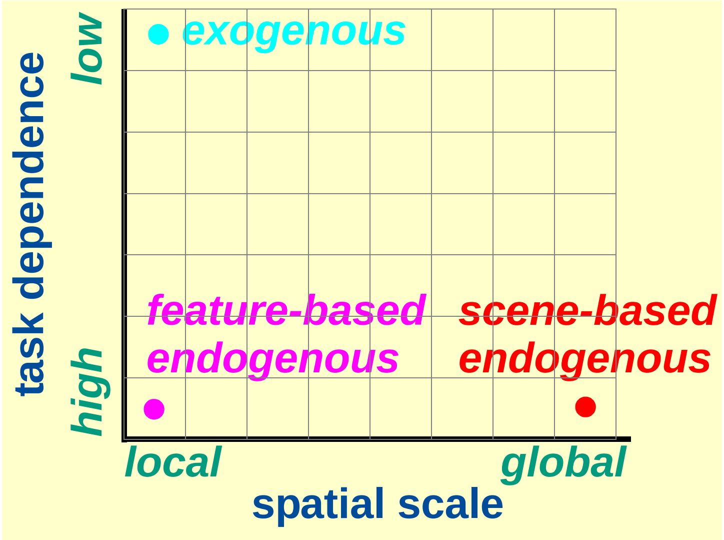
Results: Pop Out

Train single features and combine models

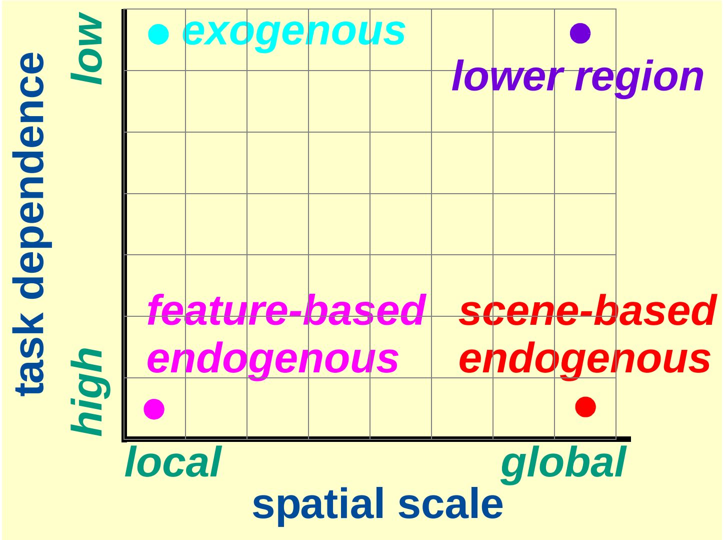
red + green + horizontal + vertical



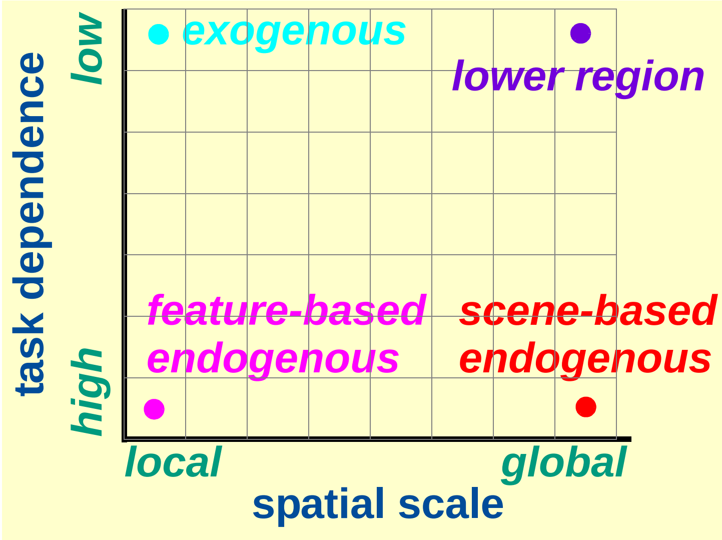
Other Phenomena



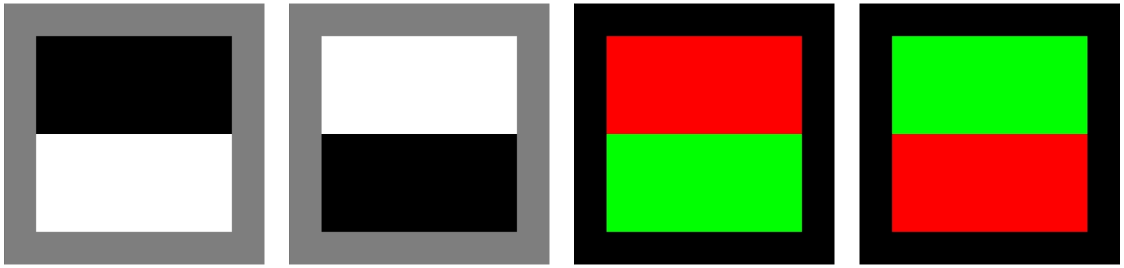
Other Phenomena



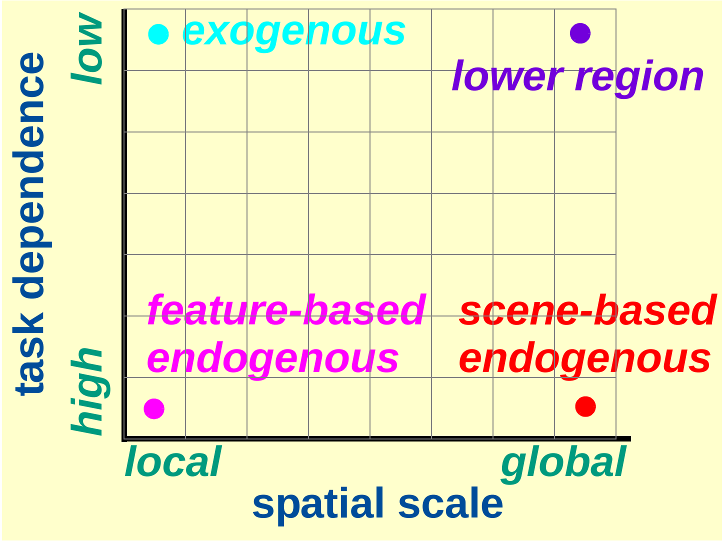
Other Phenomena



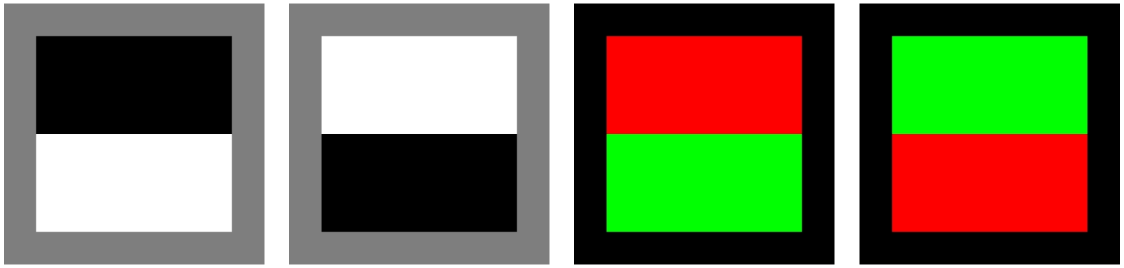
Vecera et al. (2002) found that in the absence of other cues, subjects preferred lower region of visual field as figure.

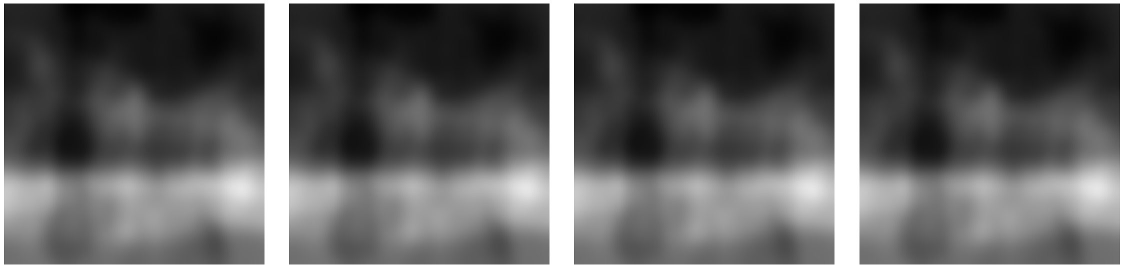


Other Phenomena

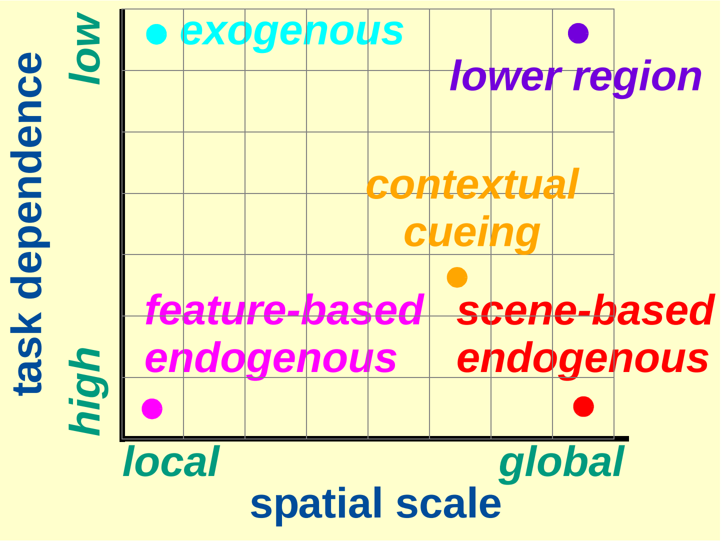


Vecera et al. (2002) found that in the absence of other cues, subjects preferred lower region of visual field as figure.



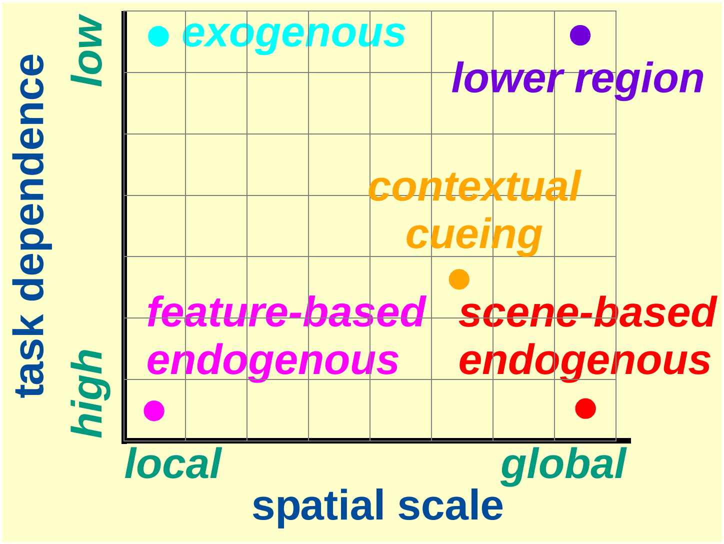


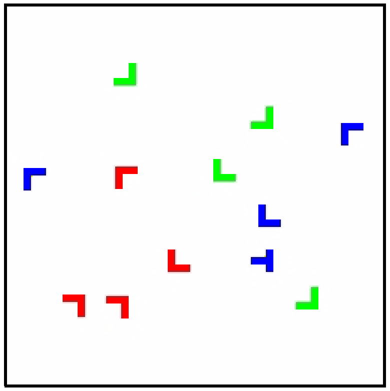
Other Phenomena



Other Phenomena

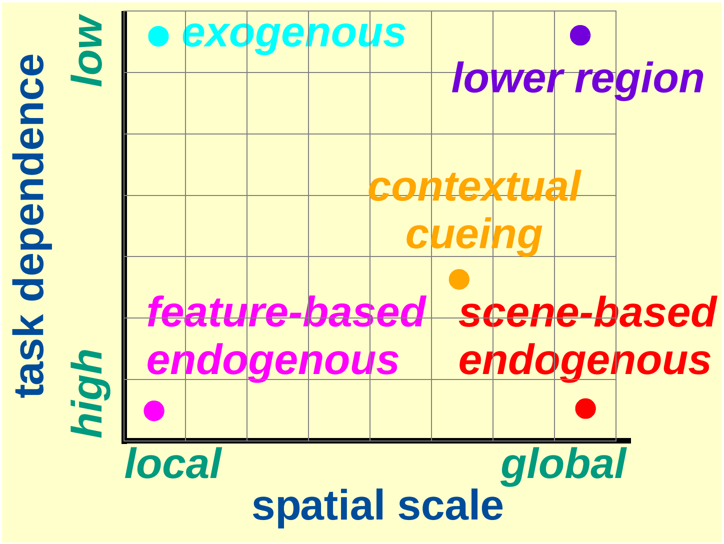
Chun and Jiang (1998) found that repeating configurations in a visual search task led to (60 ms) speed up.

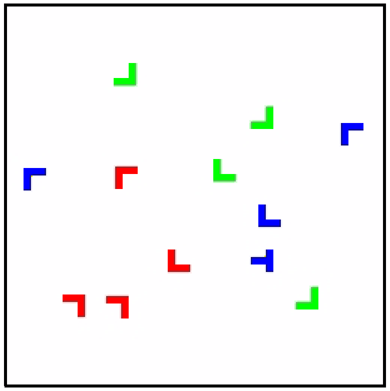


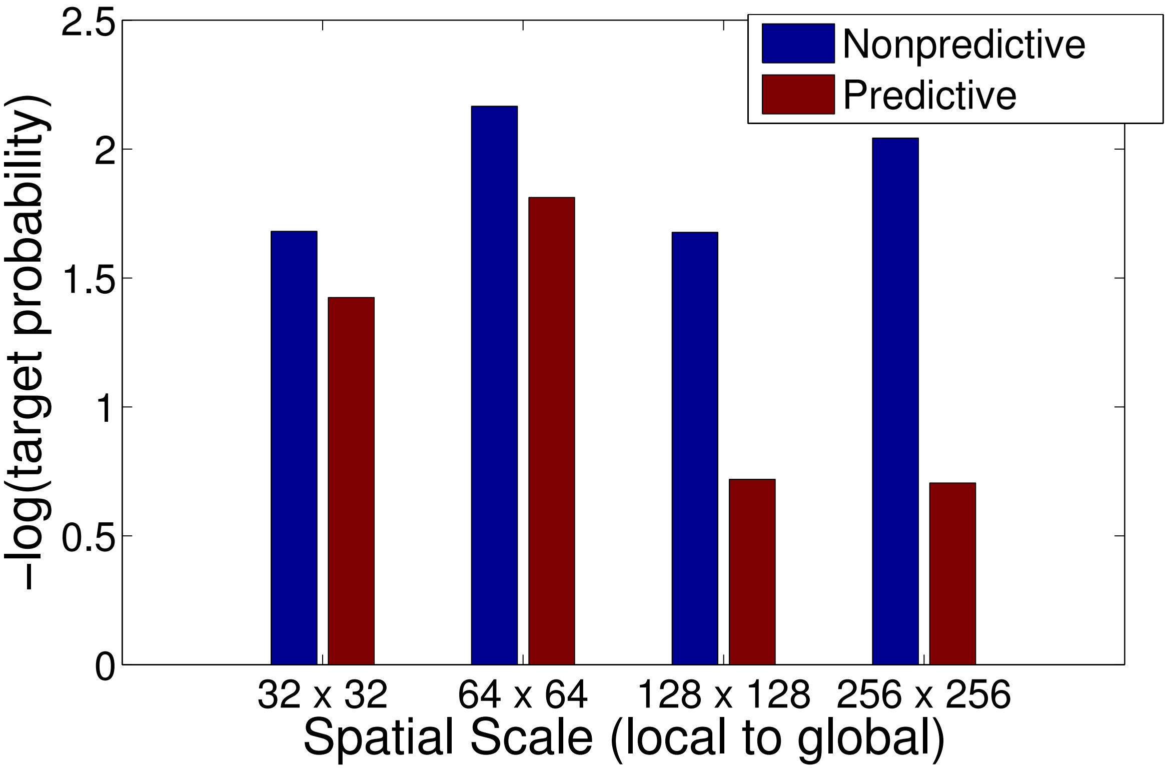


Other Phenomena

Chun and Jiang (1998) found that repeating configurations in a visual search task led to (60 ms) speed up.







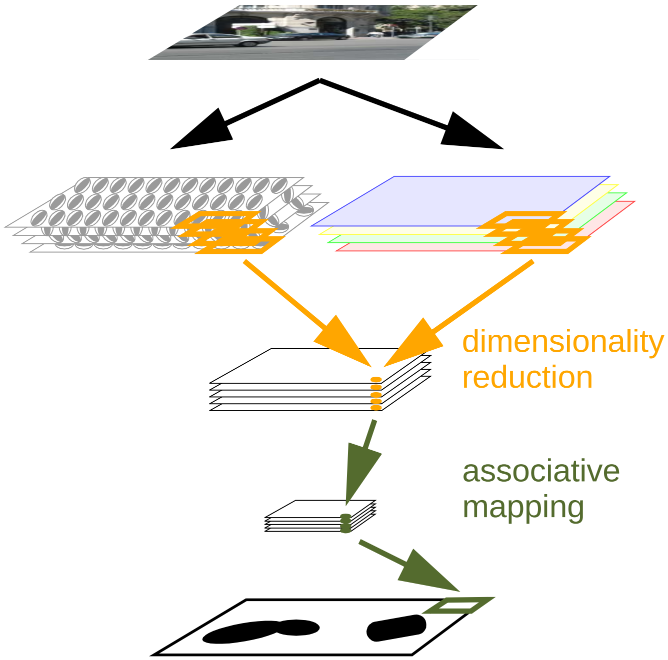
What Have We Built?

What Have We Built?

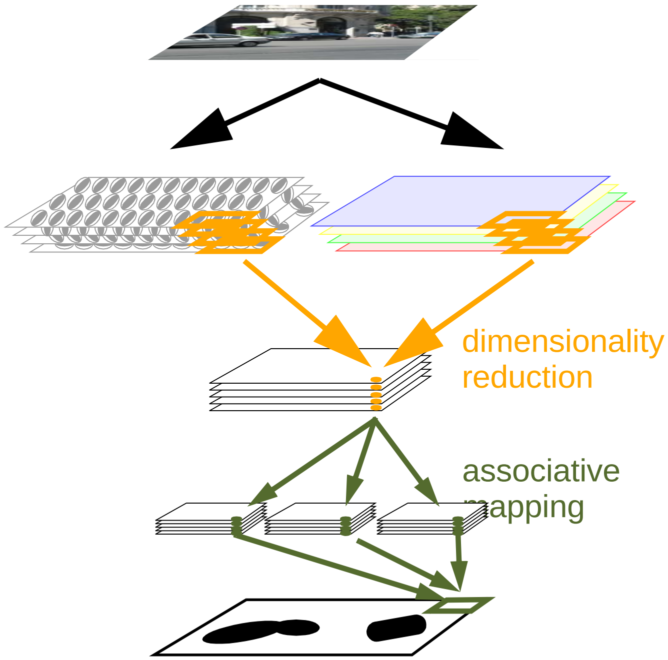
Model performs a crude sort of object recognition.

Estimates P(target$_x$ | features$_x$)

Accuracy limited by dimensionality reduction and linearity of associative network



What Have We Built?



Model performs a crude sort of object recognition.

Estimates P(target$_x$ | features$_x$)

Accuracy limited by dimensionality reduction and linearity of associative network

Linearity has benefits!

Tasks can be combined simply by gating in associative units.

red+vertical
car+bus+bike+train
exogenous control = inclusion of all tasks

No local optima -> gradient descent learning can be incremental and ongoing

Mapping Model to the Brain

If attentional salience computation is related to object recognition, maybe salience is what arises when we do a “quick and dirty” mapping, e.g., V1->IT and other projections where we skip layers.

And feedback from higher layers in posterior cortex to lower layers can serve to gate activity by saliency.

Feedback from higher layers in frontal areas serves to specify which pools of hidden units to gate out or in.

Summary so far

1. **Presented perspective on attentional control that attempts to integrate theoretical ideas from existing models and provide a unified framework for considering a range of phenomena.**

2. **Attention is not a primitive, prewired mechanism, but is intricately tied to task experience and object knowledge.**

   I'm late joining the game: SAIM and Itti models also posit strong links between object knowledge and attention.

   Models suggest different roles of cortical feedback

3. **Efficient attentional control requires learning about environment in which task is performed.**

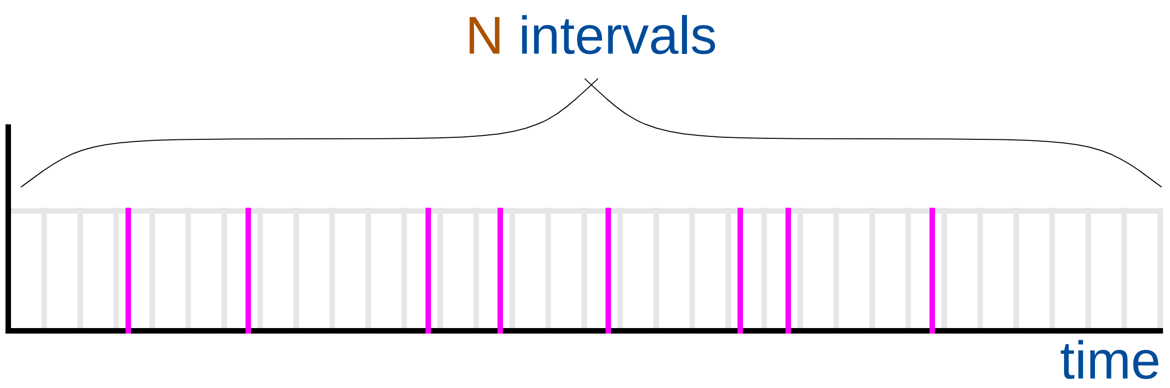
   Take this one step further: Learning about environment *is* attentional control.

Experience-Guided Search

Experience-Guided Search

Assumes visual features are represented by rate-coded spiking neurons

Simon says, “wrong!”



F_{xi}: count of the number of spikes observed for feature i at location x

$\mathbf{F}_{x}$: spike counts for all features at location x

$$\textbf{Saliency} \equiv P(T_x \mid \mathbf{F}_x, \rho)$$

task statistics

feature spike counts at location x

target at location x

If features are conditionally independent (wrong!),

$$P(T_x \mid \mathbf{F}_x, \rho) = \frac{P(T_x)\prod_i P(F_{xi} \mid T_x, \rho)}{\sum_{t=0}^{1} P(T_x = t)\prod_i P(F_{xi} \mid T_x = t, \rho)}$$

Saliency $\equiv P(T_x \mid \mathbf{F}_x, \rho)$

task statistics

feature spike counts at location x

target at location x

Bayes rule under assumption of independent features:

$$P(T_x \mid \mathbf{F}_x, \rho) = \frac{P(T_x)\prod_i P(F_{xi} \mid T_x, \rho)}{\sum_{t=0}^{1} P(T_x = t)\prod_i P(F_{xi} \mid T_x = t, \rho)}$$

Saliency $\equiv P(T_x | \mathbf{F}_x, \rho)$

task statistics

feature spike counts at location x

target at location x

Bayes rule under assumption of independent features:

$$P(T_x | \mathbf{F}_x, \rho) = \frac{P(T_x)\prod_i P(F_{xi} | T_x, \rho)}{\sum_{t=0}^{1} P(T_x = t)\prod_i P(F_{xi} | T_x = t, \rho)}$$

$\mathrm{Binomial}(\rho_{it}, N)$

mean spiking rate of feature i

for target (t=1) or distractor (t=0)

$$\textbf{Saliency} \equiv P(T_x | \mathbf{F}_x, \rho)$$

task statistics

feature spike counts at location x

target at location x

Bayes rule under assumption of independent features:

$$P(T_x | \mathbf{F}_x, \rho) = \frac{P(T_x) \prod_i P(F_{xi} | T_x, \rho)}{\sum_{t=0}^{1} P(T_x = t) \prod_i P(F_{xi} | T_x = t, \rho)}$$

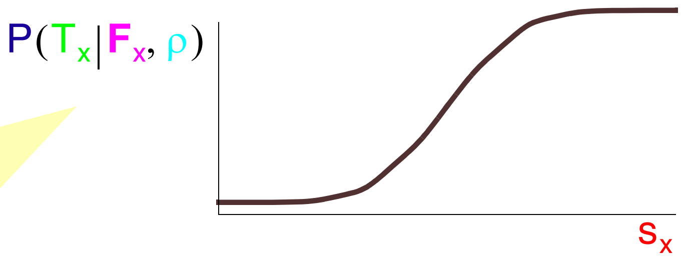
$$\cancel{\text{Binomial}(\rho_{it}, N)}$$

$$\text{Gaussian}(N\rho_{it}, N\rho_{it}(1 - \rho_{it}))$$

mean spiking rate of feature i

for target (t=1) or distractor (t=0)

$$P(T_x | F_x, \rho) = \frac{1}{1 + e^{-(c_1 + c_2 s_x)}}$$



$$P(T_x|F_x, \rho) = \frac{1}{1 + e^{-(c_1 + c_2 s_x)}}$$

Because attentional priority depends on relative saliency, we can substitute s_x for $P(T_x|F_x, \rho)$.

$$P(T_x | \mathbf{F}_x, \rho) = \frac{1}{1 + e^{-(c_1 + c_2 s_x)}}$$

spike rate of feature
i in location x

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})} (\tilde{f}_{xi} - \rho_{it})^2$$

activation
~~spike rate~~ of feature
i in location x

$$P(T_x|\mathbf{F}_x, \rho) = \frac{1}{1 + e^{-(c_1 + c_2 s_x)}}$$

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})}(\tilde{f}_{xi} - \rho_{it})^2$$

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})}(\tilde{f}_{xi} - \rho_{it})^2$$

$$= c_0 + \sum_i c_{i1}\tilde{f}_{xi} + c_{i2}\tilde{f}_{xi}^2$$

Experience-Guided Search

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})}(\tilde{f}_{xi} - \rho_{it})^2$$

$$= c_0 + \sum_i c_{i1}\tilde{f}_{xi} + c_{i2}\tilde{f}_{xi}^2$$

Guided Search

$$s_x = \sum_i c_{i1}\tilde{f}_{xi}$$

Differences Between EGS and GS

1. EGS includes terms quadratic in $\tilde{f}_{xi}$
2. GS determines constants via heuristics or optimization; in EGS, constants follow directly from task environment
3. GS retards model via noise, limits on gains; EGS doesn't.

Experience-Guided Search

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})}(\tilde{f}_{xi} - \rho_{it})^2$$

$$= c_0 + \sum_i c_{i1}\tilde{f}_{xi} + c_{i2}\tilde{f}_{xi}^2$$

Guided Search

$$s_x = \sum_i c_{i1}\tilde{f}_{xi}$$

Two Further Claims

1. **Bias that all features are considered relevant in the absence of experience**

   Achieved by treating ρ as a Beta random variable with imaginary-count prior

   $E[\rho_{i0}] < E[\rho_{i1}]$

2. **Environment is nonstationary**

   With probability λ, environment and/or task can change.

From these two claims, we have a total of 3 free parameters in the model.

Qualitative performance does not depend on parameters as long as $\lambda > 0$ and $E[\rho_{i0}] < E[\rho_{i1}]$

What It Boils Down To

- **Generate stimulus sequence corresponding to experiment.**
- **On each trial, perform feature extraction on display.**
- **Compute saliency at each location x**

$$s_x = \sum_i \sum_{t=0}^{1} \frac{1-2t}{\rho_{it}(1-\rho_{it})}(\tilde{f}_{xi} - \rho_{it})^2$$

- **Response time ~ saliency rank of target**
- **Update statistics of targets and distractors**

$$\alpha_{it} \leftarrow \lambda\alpha_{it}^0 + (1-\lambda)\left(\alpha_{it} + \sum_{x \in \chi_t} \tilde{f}_{xi}\right)$$

$$\beta_{it} \leftarrow \lambda\beta_{it}^0 + (1-\lambda)\left(\beta_{it} + \sum_{x \in \chi_t} 1 - \tilde{f}_{xi}\right)$$

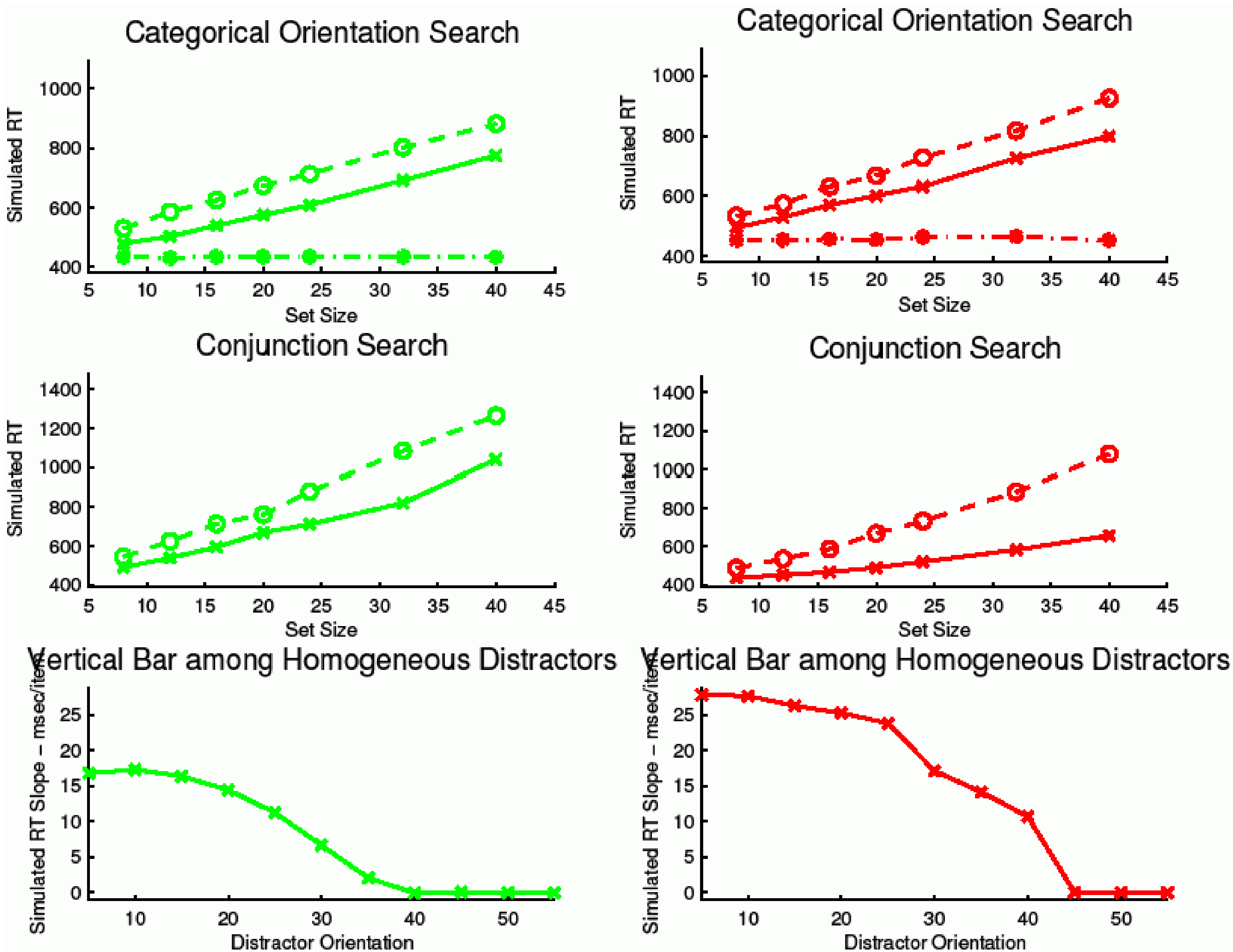
where $\rho_{it} = \dfrac{\alpha_{it}}{(\alpha_{it} + \beta_{it})}$

NOTE TO MIKE:

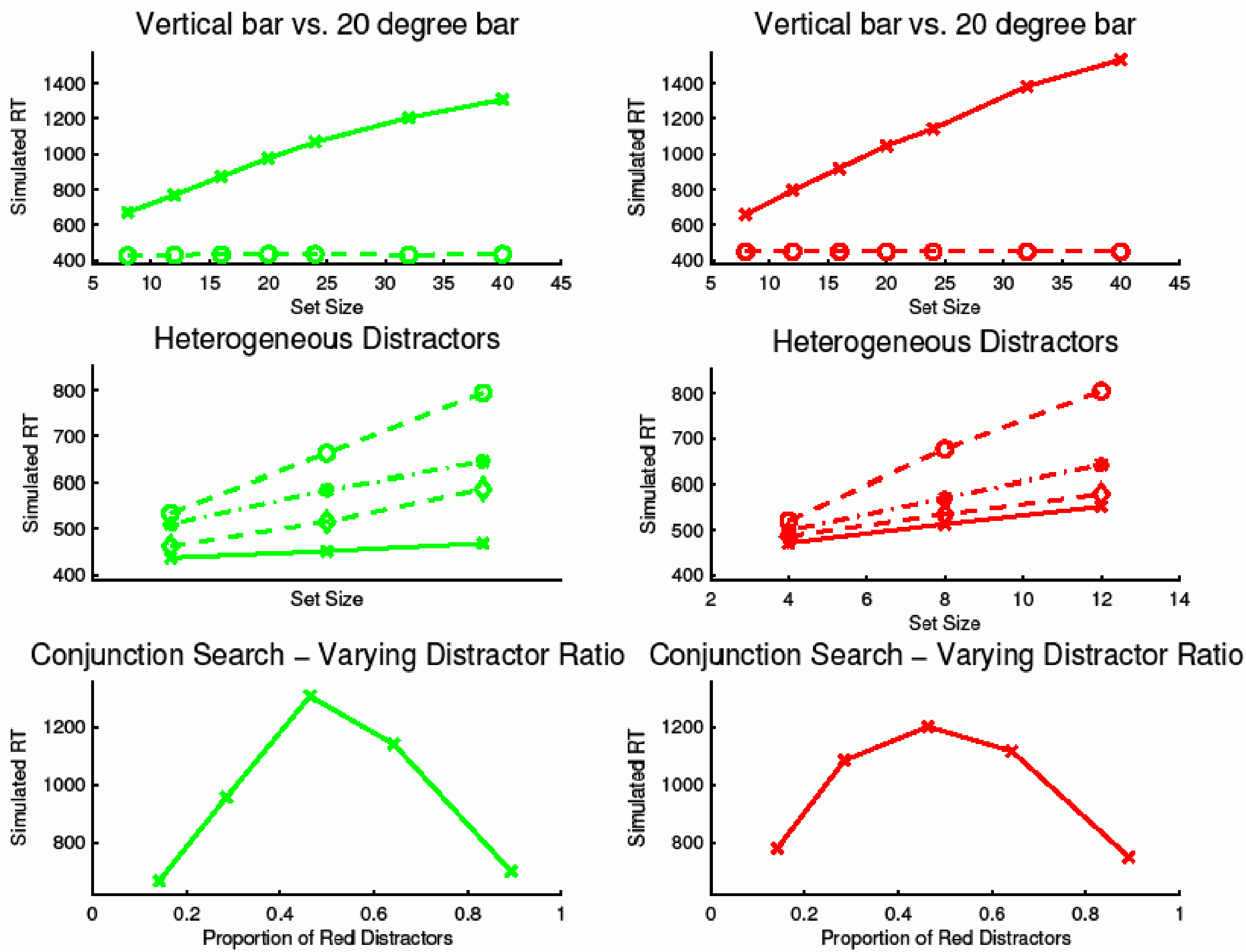
show examples of rho distribution changing over time

present generative model: binomial is an assumption

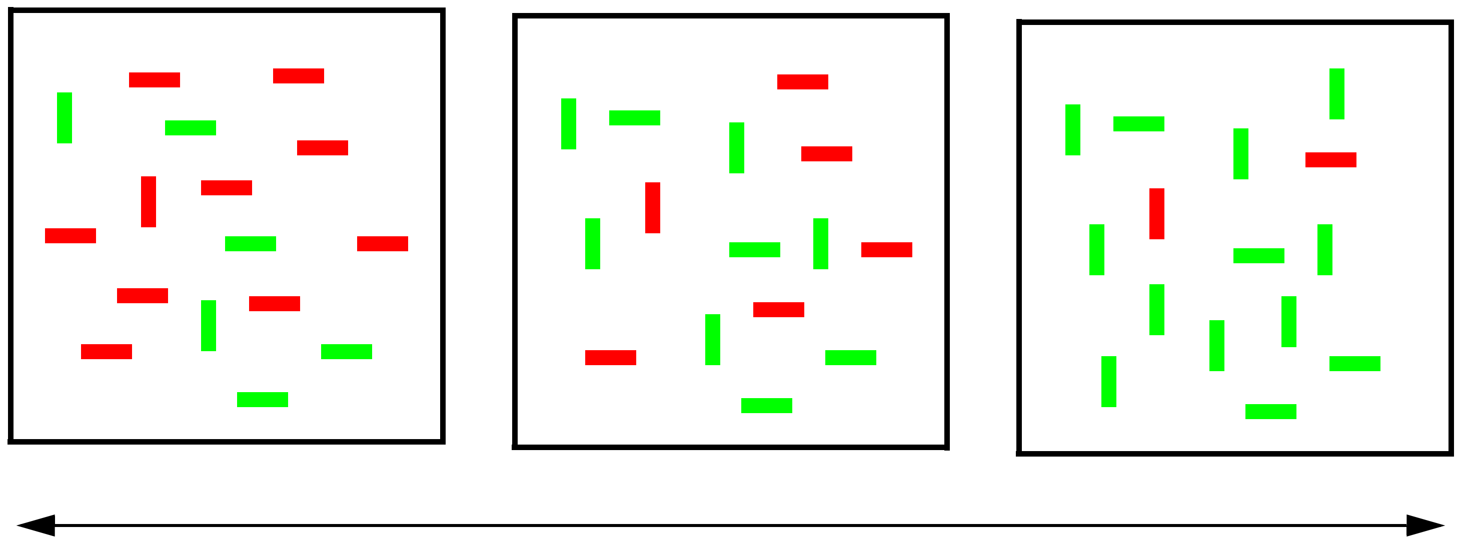
Simulation Results



Simulation Results



Simulation Results: Varying Distractor Proportion



mostly red distractors | equal number red and vertical | mostly vertical distractors

Usual story

Two stage filtering process

Our story

When statistics of the environment make one feature a more reliable cue, it is weighed more heavily.

Summary

Theories of attentional control invoke specialized mechanisms

- rule-based heuristics
- conflict monitoring and error detection
- optimization of performance

Experience-Guided Search model pushes the idea that attentional control arises directly from statistical inference on the task environment in which an individual is operating.

But so far we focused on adaptation to the ongoing stream of experience and trial-to-trial *changes* in control.

Adaptation is one thing, but the *big* question is how we translate instructions to action, i.e., how control is *initiated*.

Instruction Following

The two models we presented offer stories about how a task description can lead to an initial configuration of model.

Integrated control-space model

task -> look up of object models

Experience-guided search

task -> specification of priors in feature values