Where to look?

Nando de Freitas

CIFAR 2009

Two parts to this talk

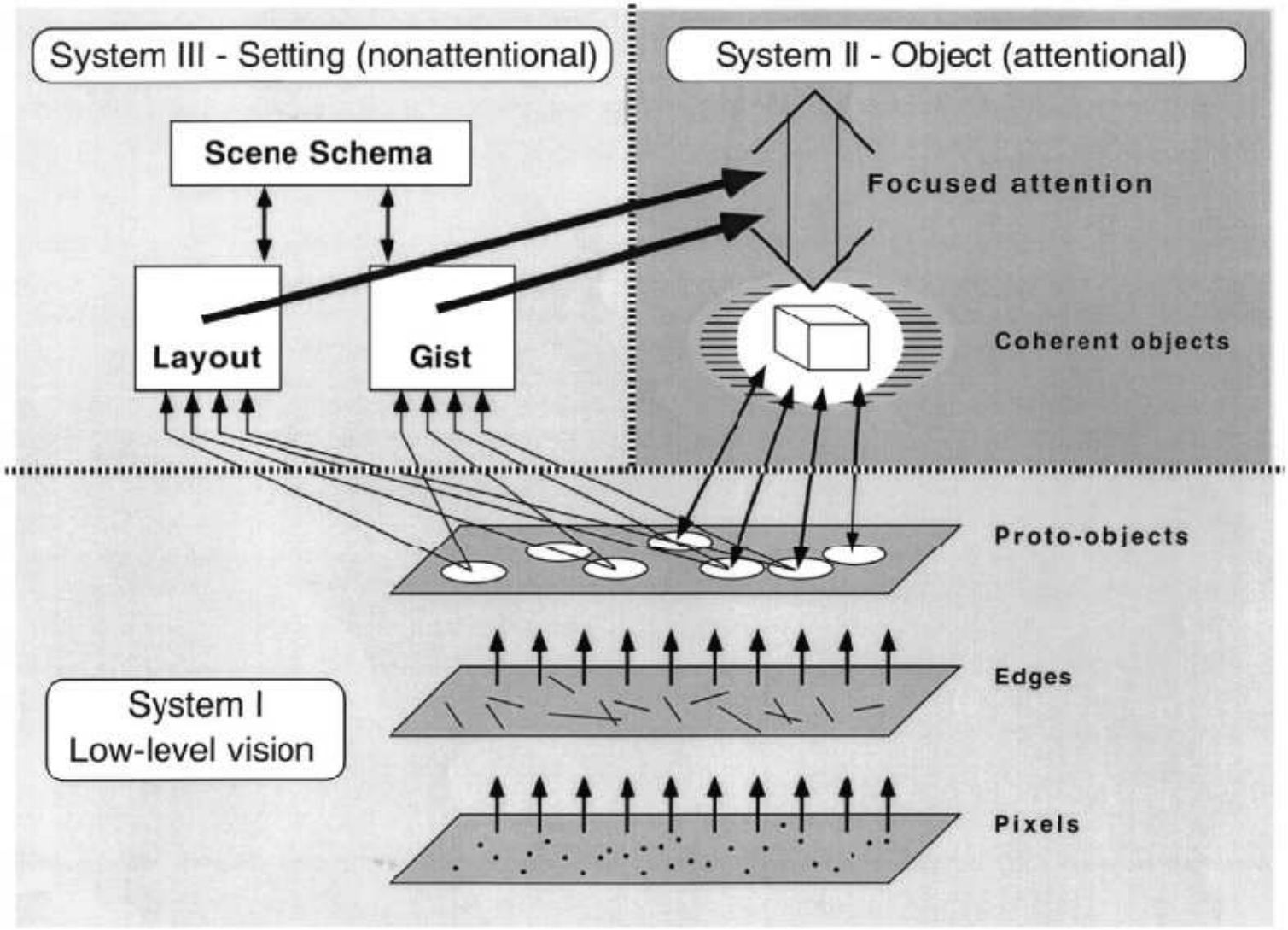
(i) Sequential optimal control approach

Julia Vogel

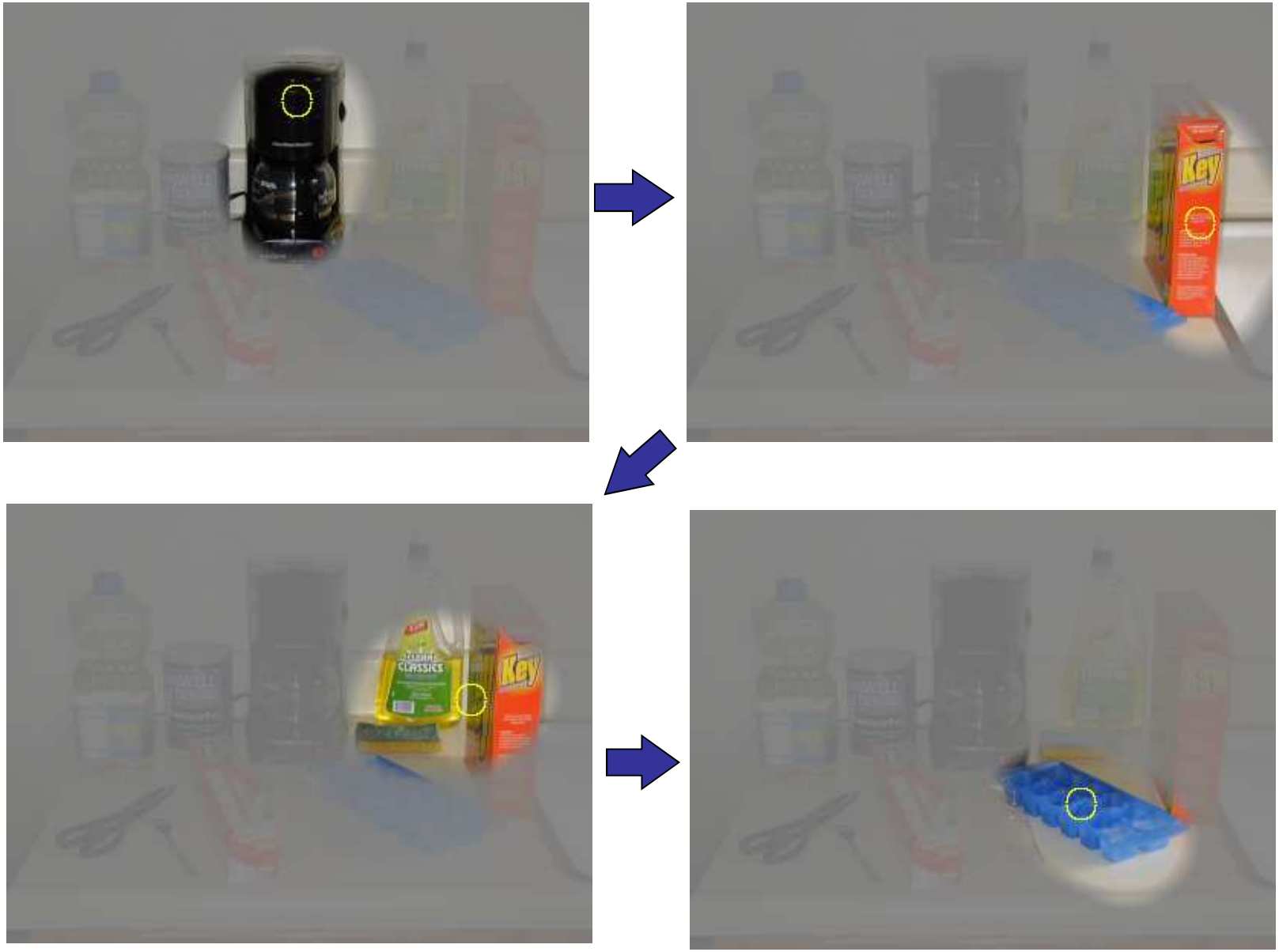
(ii) A hierarchical-temporal-memory-convolutional-Boltzmann-Machine (HTM-CRBM) approach

Bo Chen

Previous work: Dynamic scene representation



[Rensink, Visual Cognition, 2000]

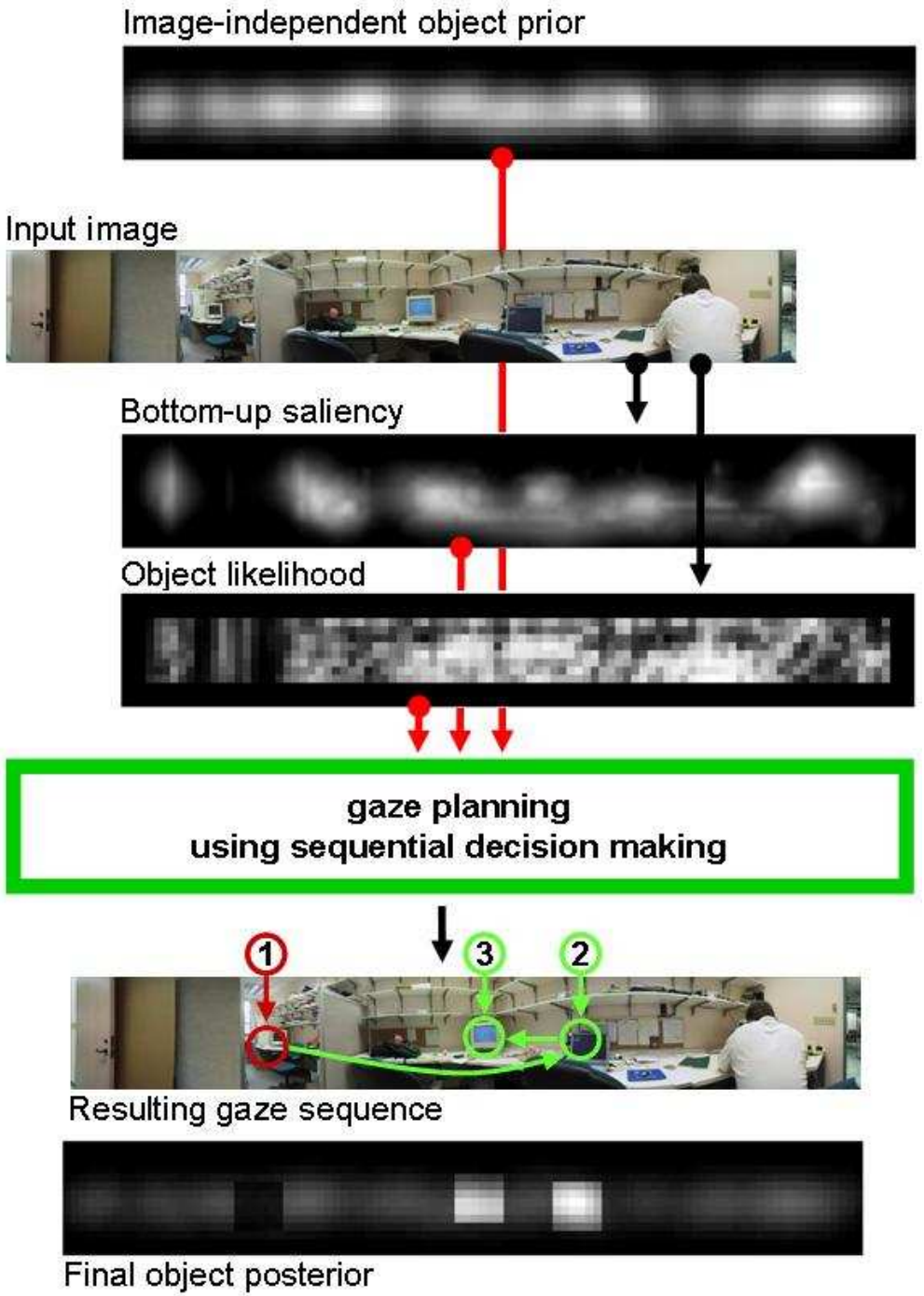


[Dirk Walther, 2006]

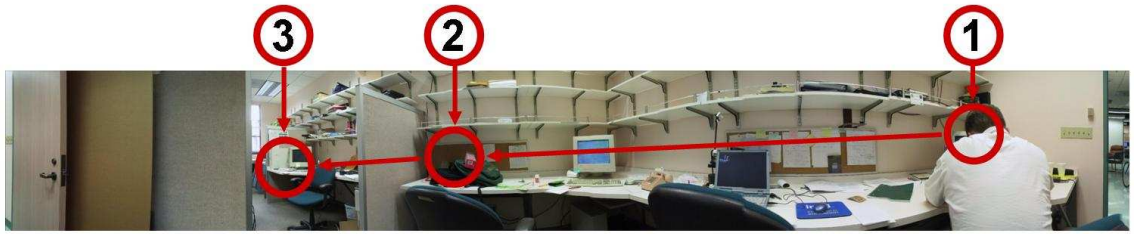
Use a POMDP that integrates:

- context priors
- bottom-up saliency
- target saliency

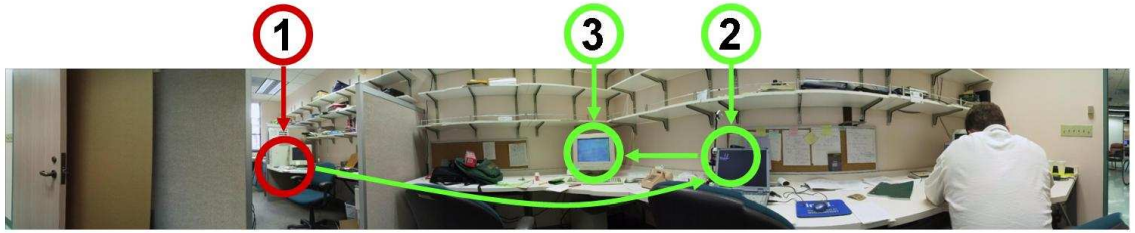
to select gaze a sequence that minimizes uncertainty in the posterior distribution over the object's location.



Algorithm



Gaze sequence when using only bottom-up saliency



Gaze sequence after gaze planning with bottom-up and top-down information

[Itti and Koch,2000]
[Walther, SaliencyToolbox 1.0, 2006]

- For each gaze order:
  - For each Monte Carlo simulation:
    - For each planning step:
      - Sample likelihood model based on computational gist
      - Use Bayes rule to compute the posterior distribution numerically
  - Approximate the expected cost (discounted posterior information increase)
- Choose gaze order with minimum expected cost

Likelihood object model

- Object likelihood $P(y_t|x_t)$: probability that area around particular location x_t is indicative of target.
- Main idea: Use very fast, crude estimate of object likelihood similar to humans catching the “gist” of the scene very quickly.
- We use Torralba’s gist, trained on monitors.

Actual object detector

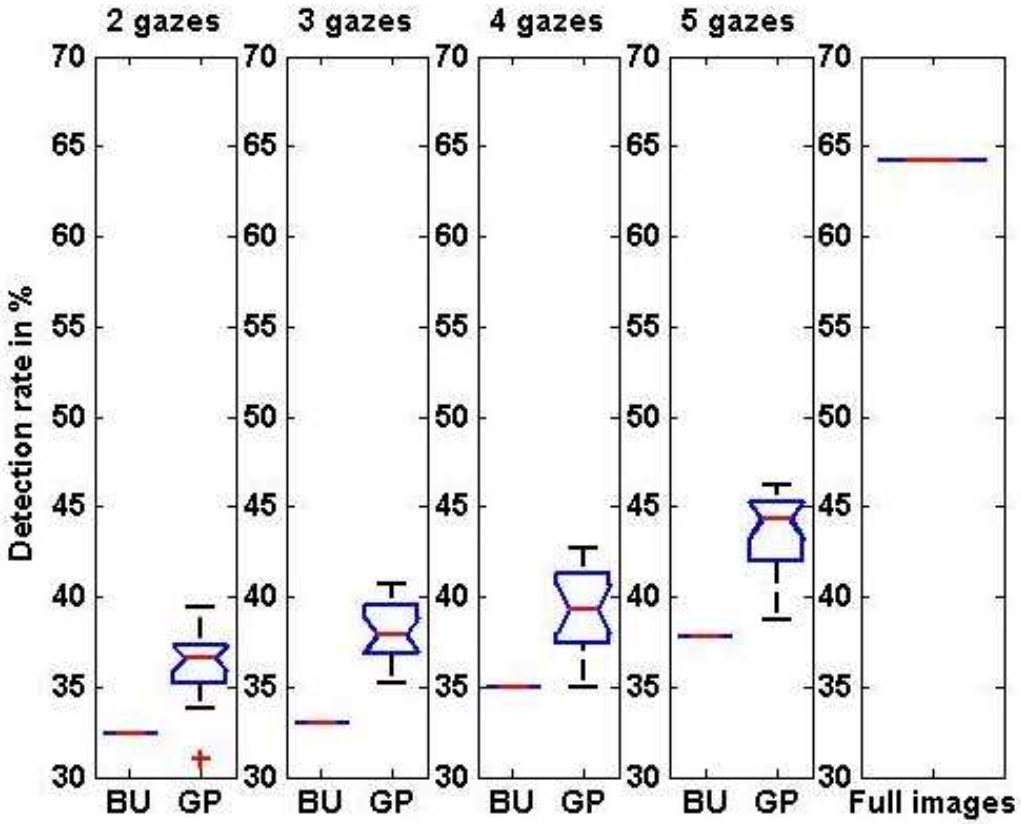
- Boosted detector of [Torralba, Murphy, Freeman, PAMI 2007]
- Analyzes window of 200x200 pixels around gaze location
- Also used for full image analysis during experiments

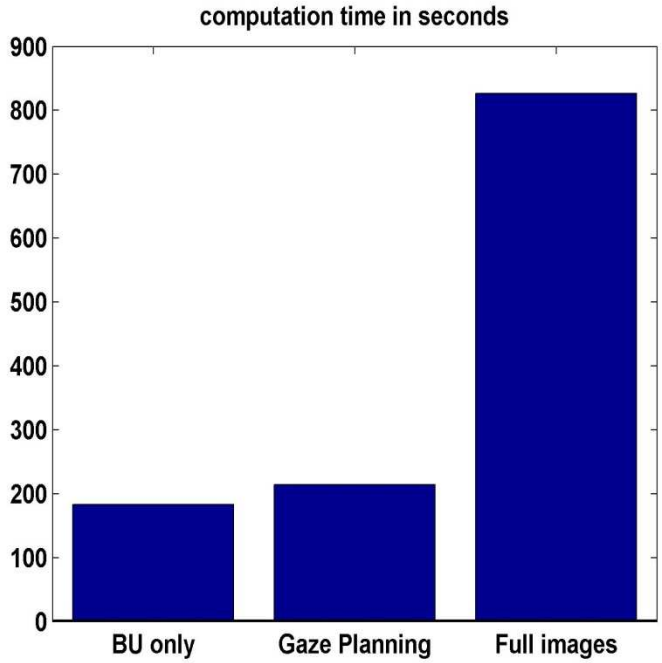
scale 1

scale 2

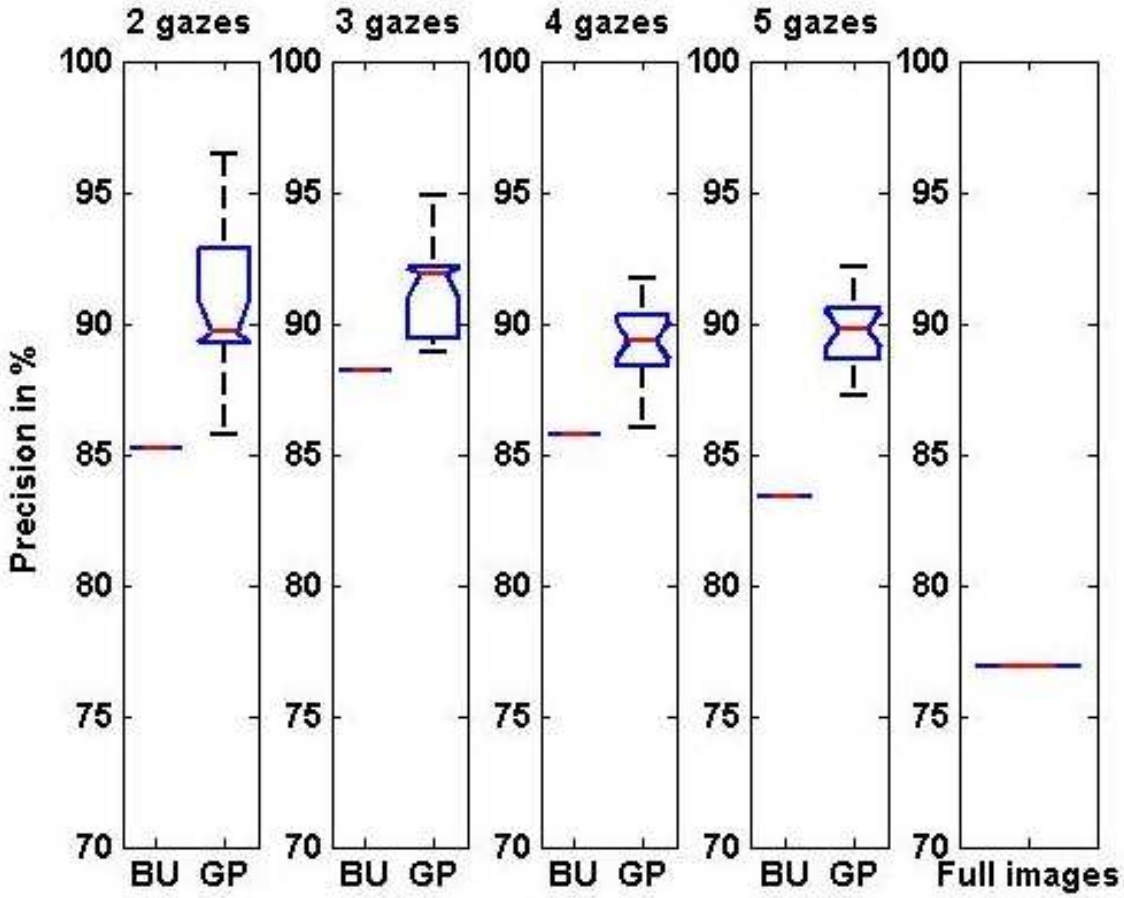
scale 3

Results: Detection rate





Results: Precision



Overt and covert attention

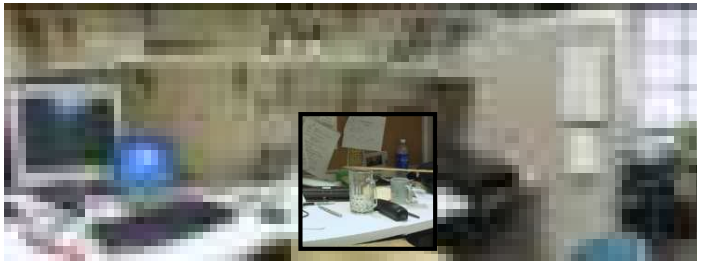
[Itti and Koch]

We need to hallucinate more

[Frank Ferrie]

Intuition: Learn office model

Given the first gaze, propagate information up the office model and down to hallucinate the rest of the image. Repeat as necessary.

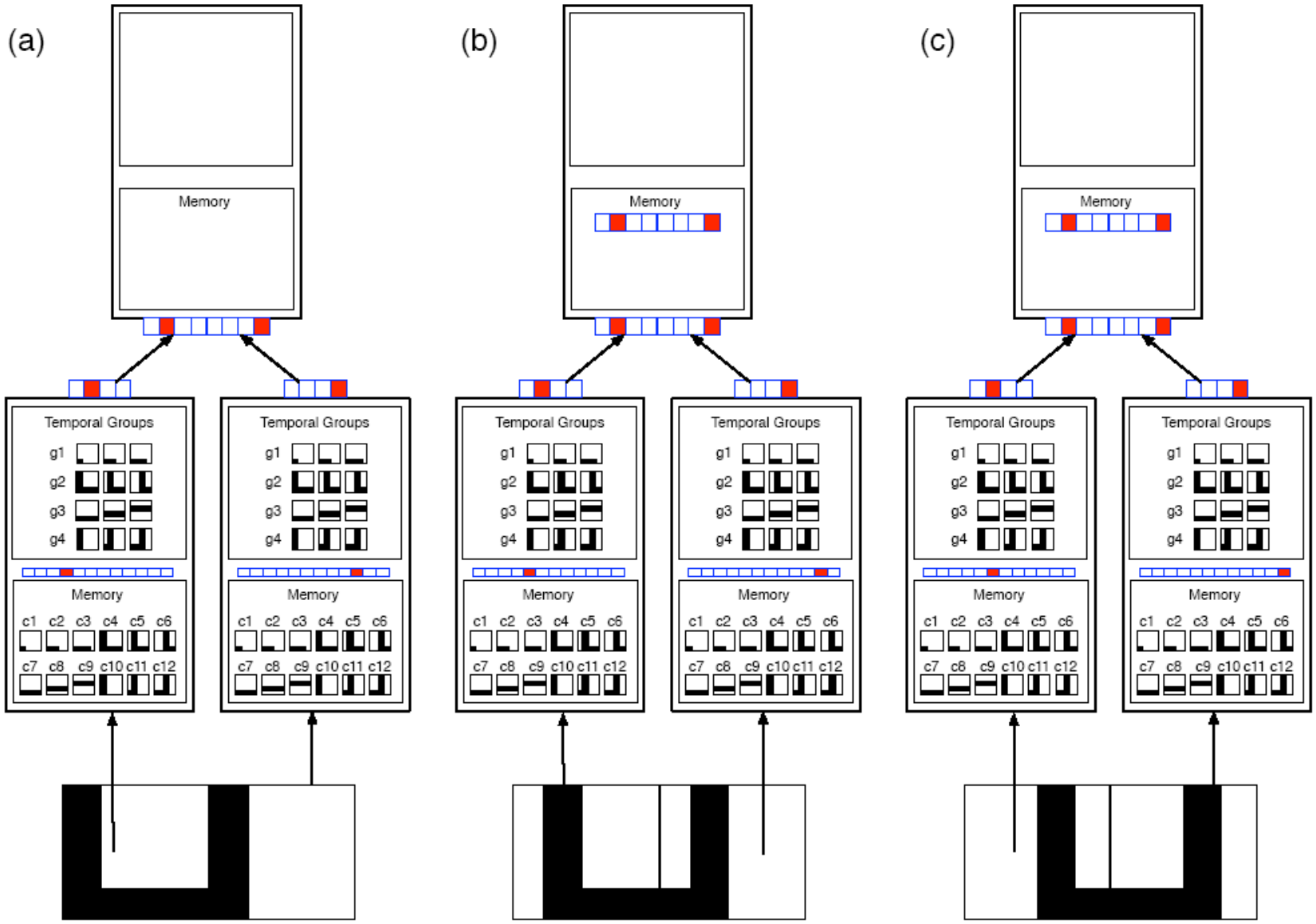


Where's the mouse?
How many computers in the office?
Is this a desk?

Some desiderata for architecture

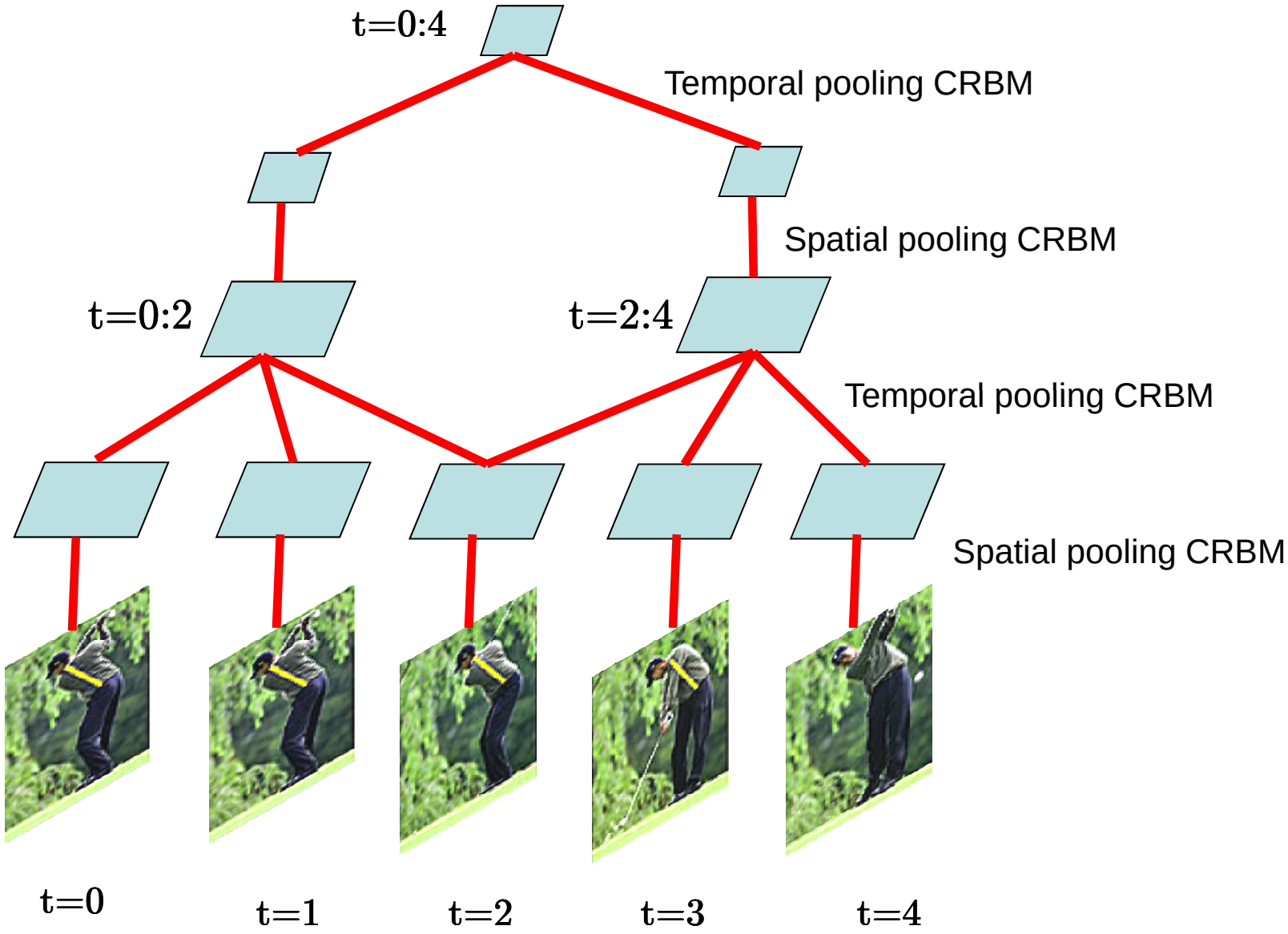
1. Must be able to **hallucinate**
2. Must model **space-time** signals
3. Must use common **units** of computation
4. Must be memory efficient – **hierarchical**
5. Must be **invariant** to signal transformations
6. Must have a **distributed** representation
7. Must explain Ron Rensink's analysis

Hierarchical Temporal Memory (HTM)

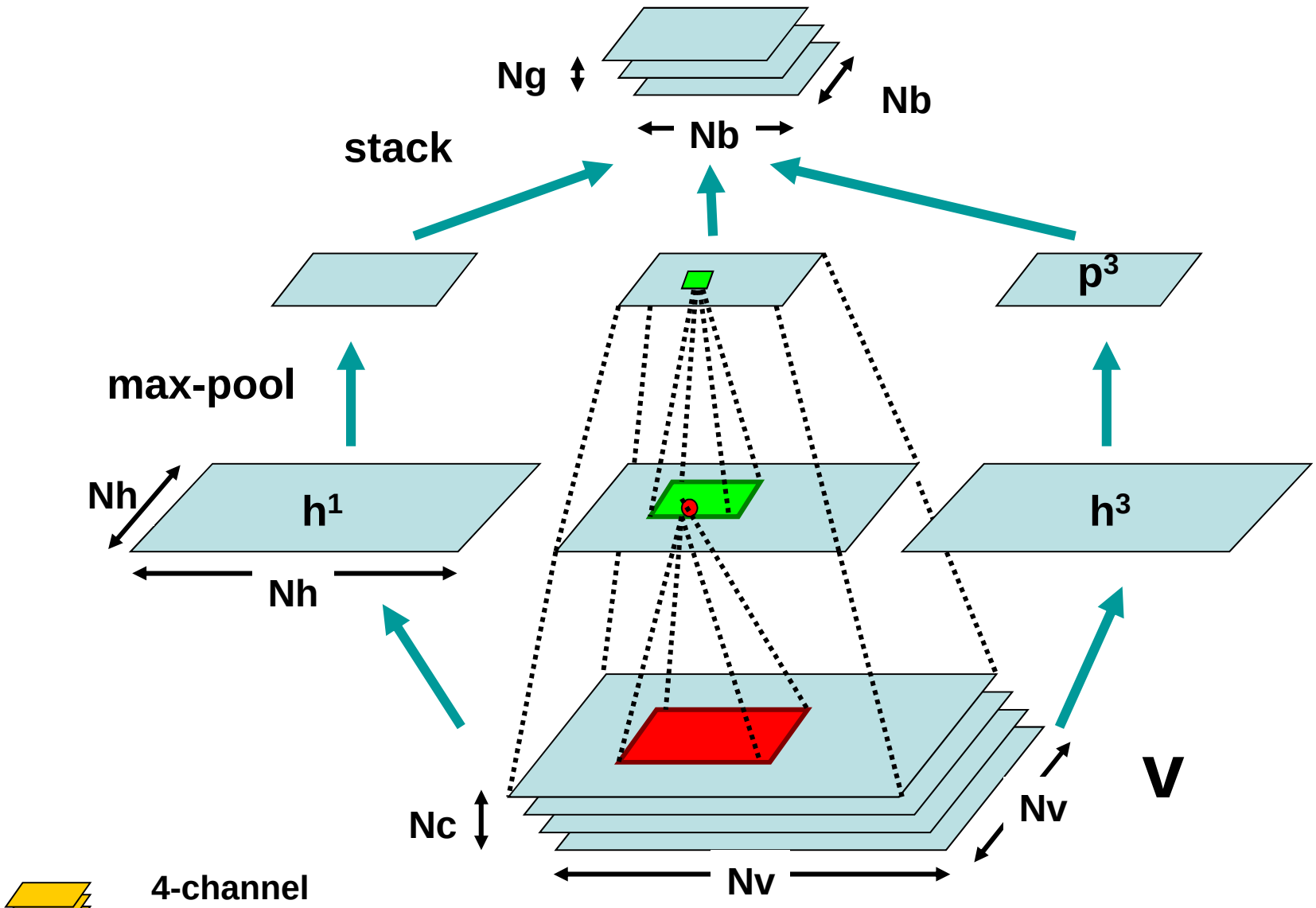


[Jeff Hawkins; Dileep George, 2008]

HTM-CRBM architecture

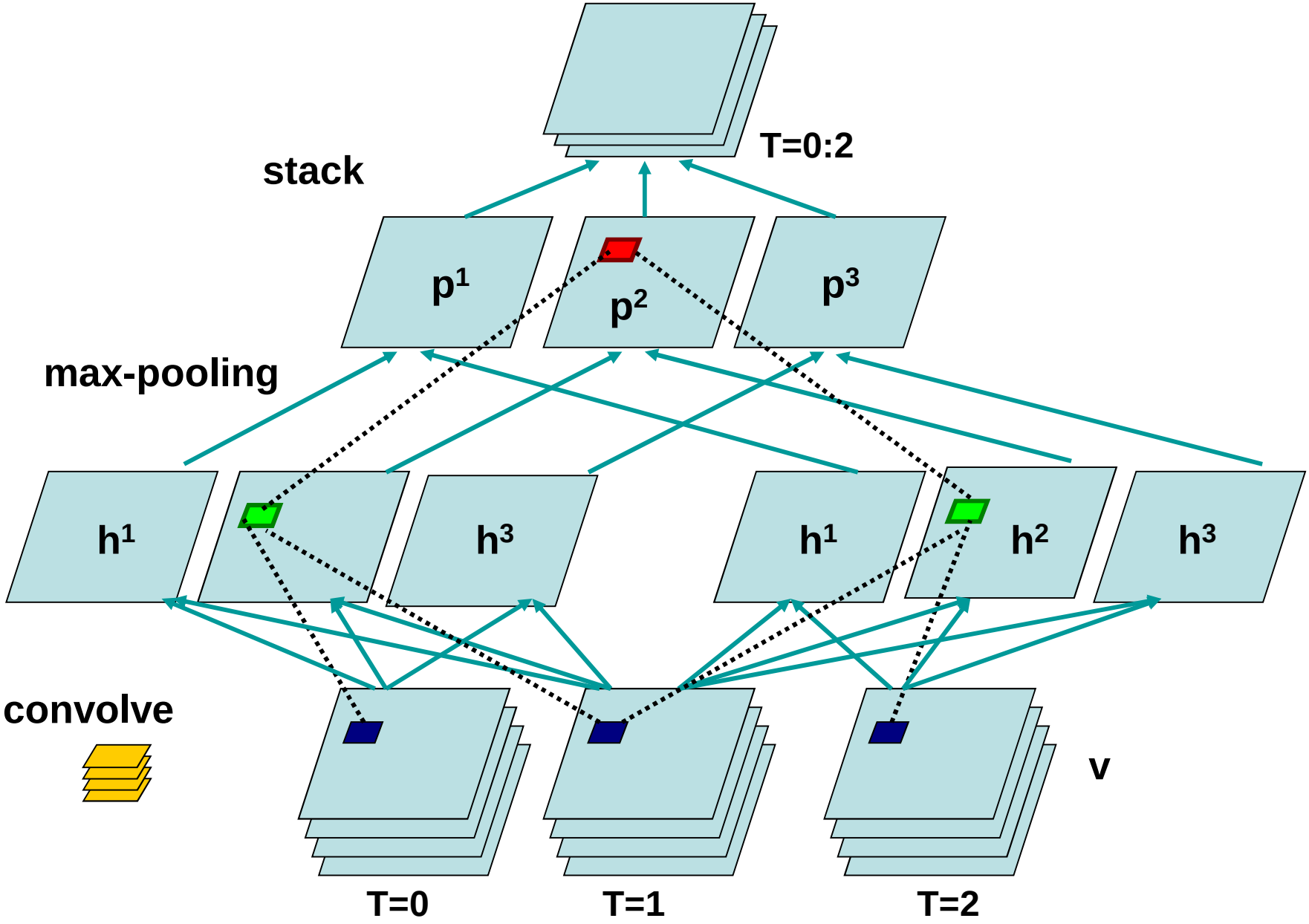


Spatial pooling CRBM

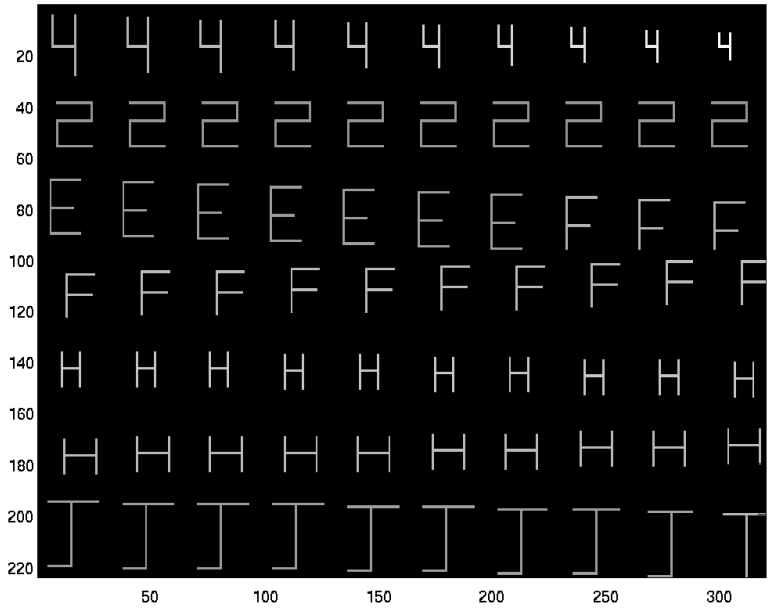


In this image, V, the input has $\mathrm{Nc}=4$ channels

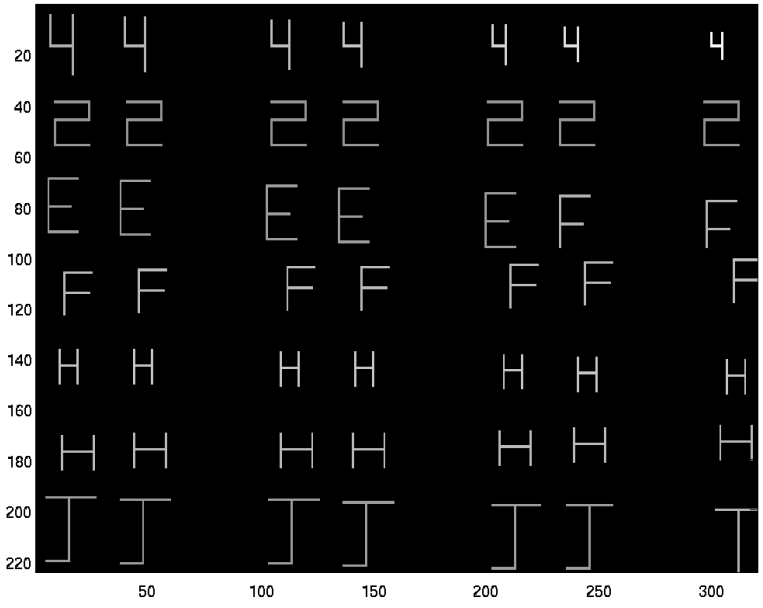
Temporal pooling CRBM



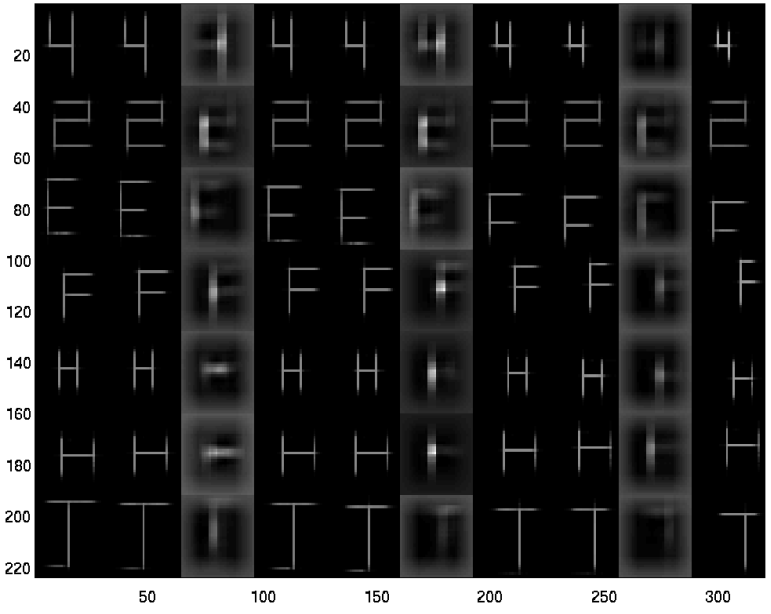
Training sequences



Sequences with missing data



Reconstructions

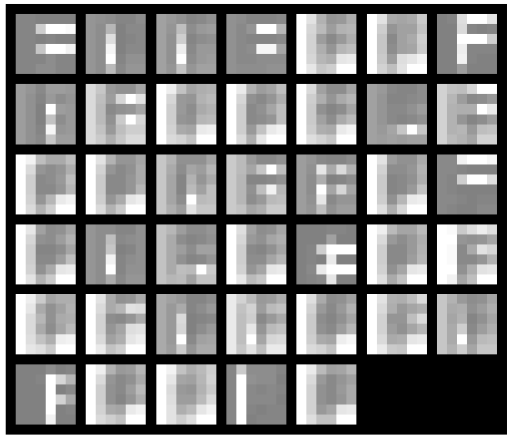


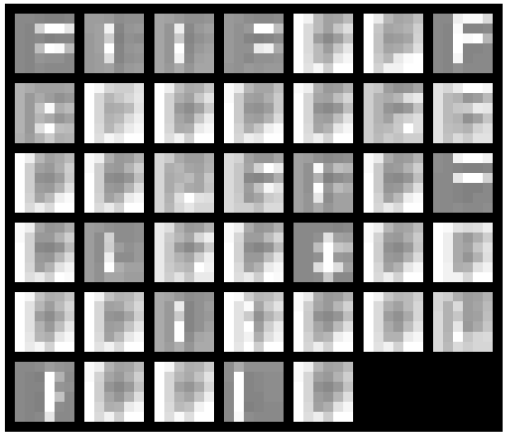
Mean field values of activations for a specific image in a sequence

Bottom up only

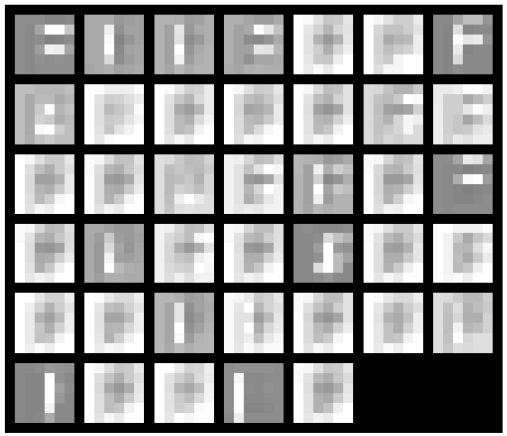
Bottom up top down

Original image





Missing image



Thank you

Many questions remain:

Optimal sequence for recognizing a specific something?

Best strategy for adding the “where pathway” / actions?

Agreement with what Ron says humans do?

Technical issues: scaling (computation & data), hyper-parameters, structure, and more.