BEVERLY

A Robot That Discovered What Caregivers Look Like

2008 Neuromorphic Engineering Workshop

Javier R. Movellan
Ian Fasel
Nick Butko
Temporal Dynamics of Learning Center
UCSD

Outline

- Timing and Social Contingency in Infants and Robots.
- Automatic Discovery of Object Categories.
- The Beverly Project.

Timing and Social Contingency

Timing and Social Contingency

Movellan & Watson (1987) Perception of Directional Attention, SRCD

Movellan & Watson (2002) The Development of Gaze Following as a Bayesian Systems Identification Problem. ICDL.

Movellan & Watson (1987) Perception of Directional Attention, SRCD

Movellan & Watson (2002) The Development of Gaze Following as a Bayesian Systems Identification Problem. ICDL.

Infant Robot Interaction

- After 3 vocalizations, 20 seconds in the experiment baby shows clear signs of having detected responsiveness.
- Turn taking: vocalizations follows by about 6 seconds of silence.
- Is what this child was doing a good idea given the statsitics of social interaction?

Not a trivial problem.

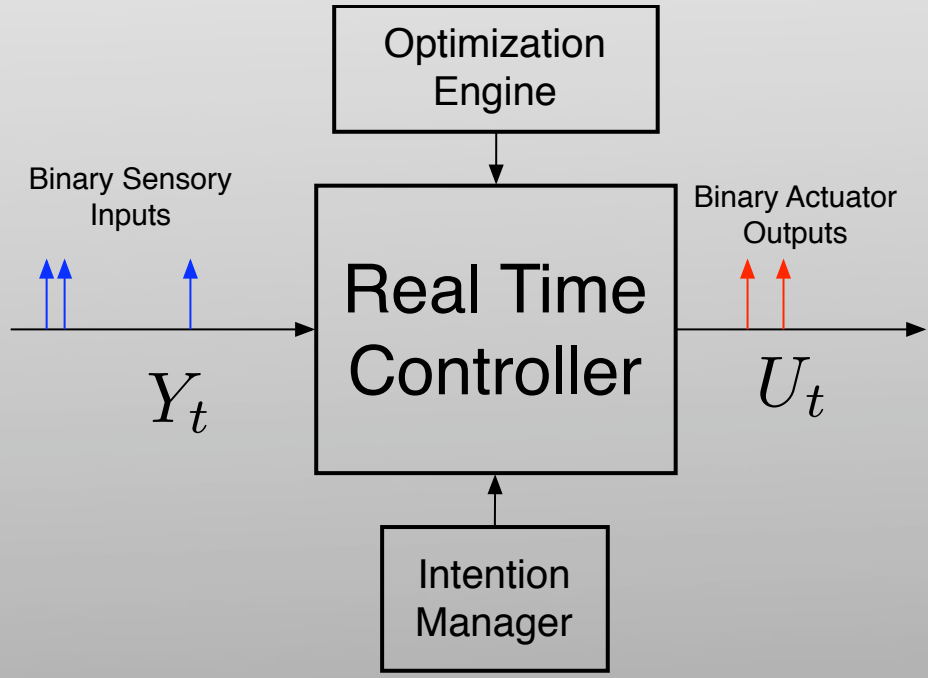
- People vary in degree of responsiveness.
- Significant time delays and uncertainty in the distribution of time delays.
- Significant background noise which varies from person to person and from situation to situation.
- Particularly difficult when working with simpler perceptual systems, like in robots.

A Barebones "Baby"

Binary Sensor: Sound energy crosses threshold.

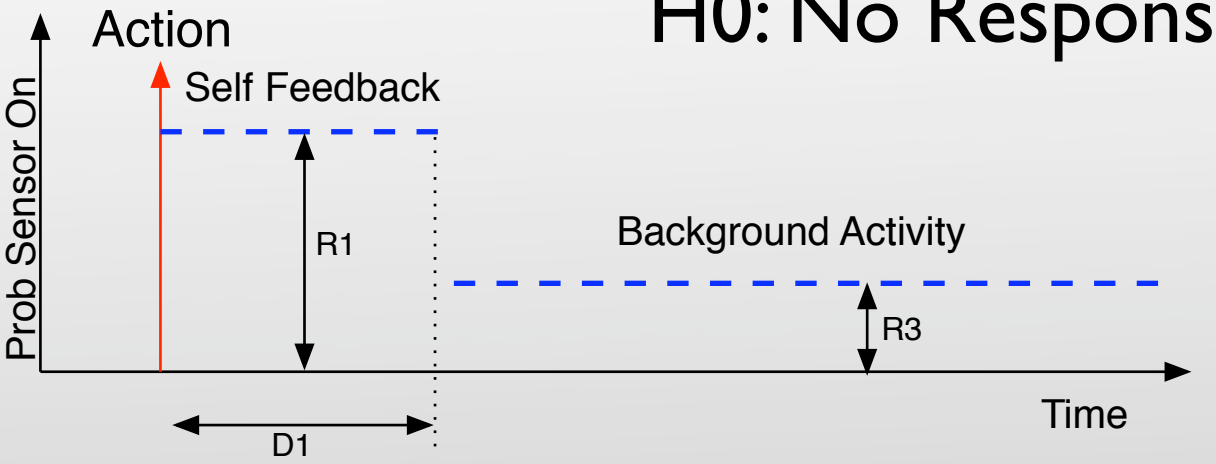
Binary Actuator: Vocalization/Silence

Continuous Operation: 30 Hz.

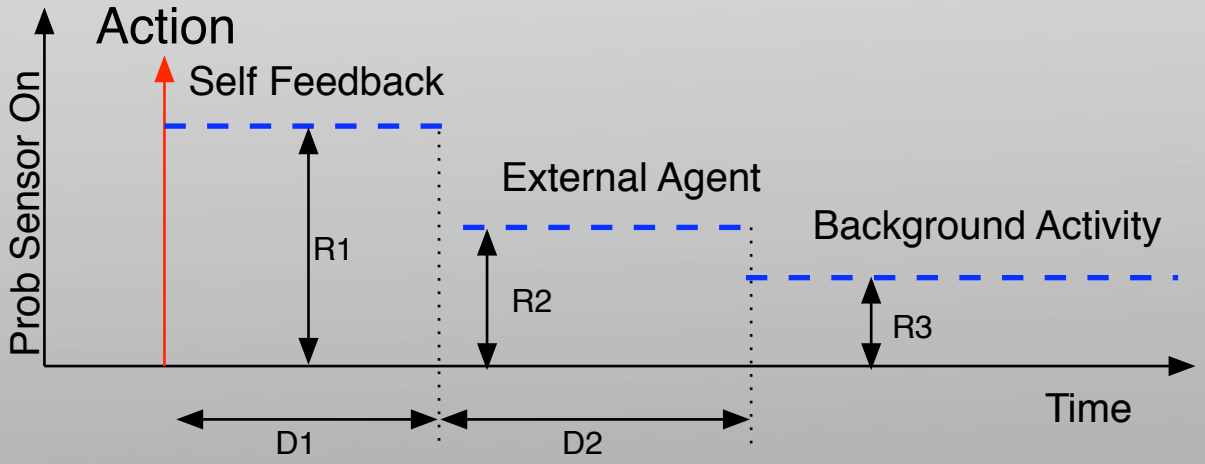


Timing: A signature of humanness.

H0: No Responsive Human



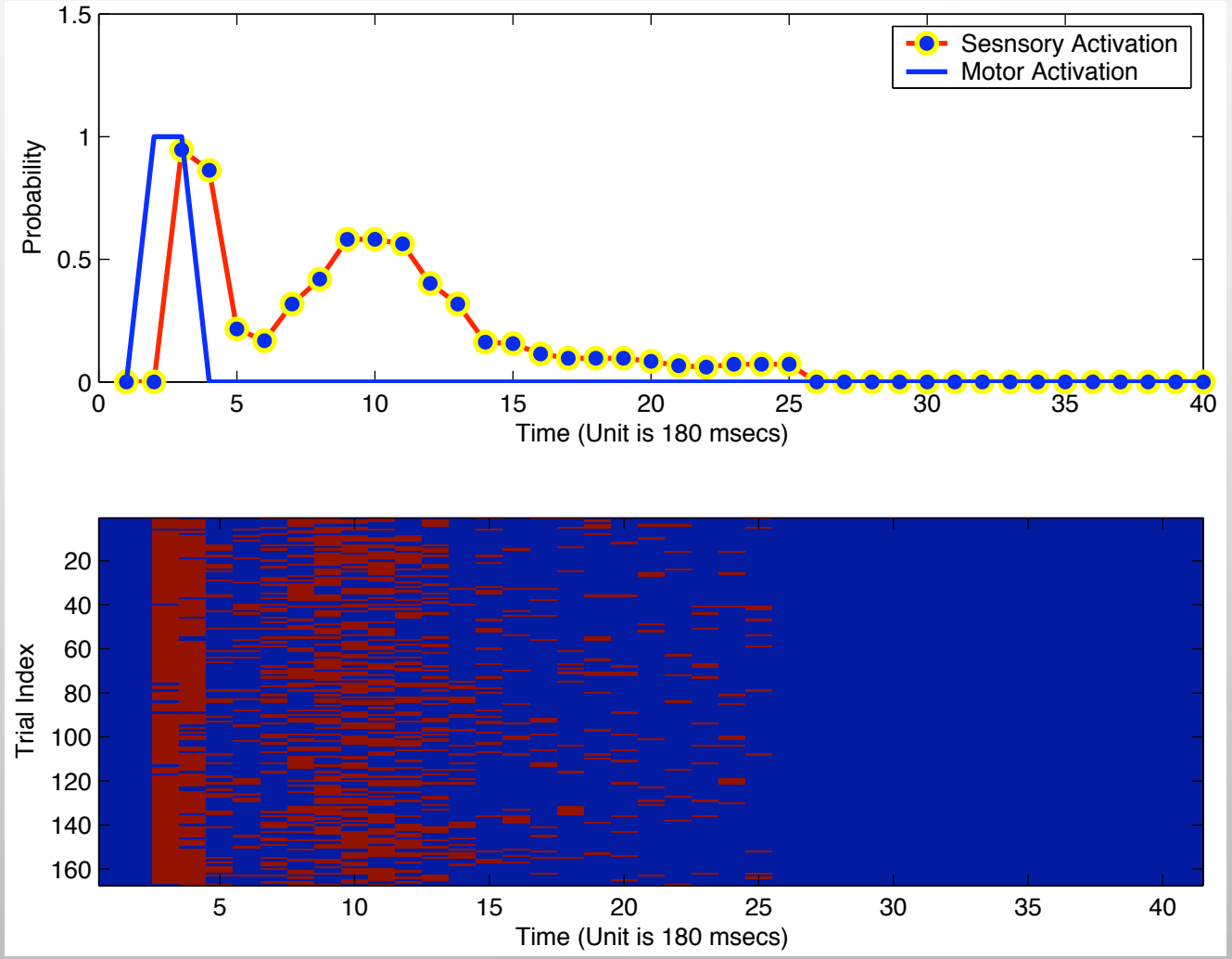
H1: Responsive Human



Infomax Model of Social Timing

- Hidden Variable of Interest: Which of two contingency clusters are we in?
- Unknown parameters: Current Background Noise. Specifics Responsiveness of the human in front of us.
- Fixed parameters: Self-feedback delay. Response time distribution for humans.
- Data: History of sensor and actuator activity.
- Utility: Information gained about the hidden variable (reduction in entropy of posterior distribution of cluster condition given observed sensory motor sequence). Epistemic Value.
- Controller Maps past history into moment by moment decisions to vocalize or to stay quiet.

Statistics of Social Timing



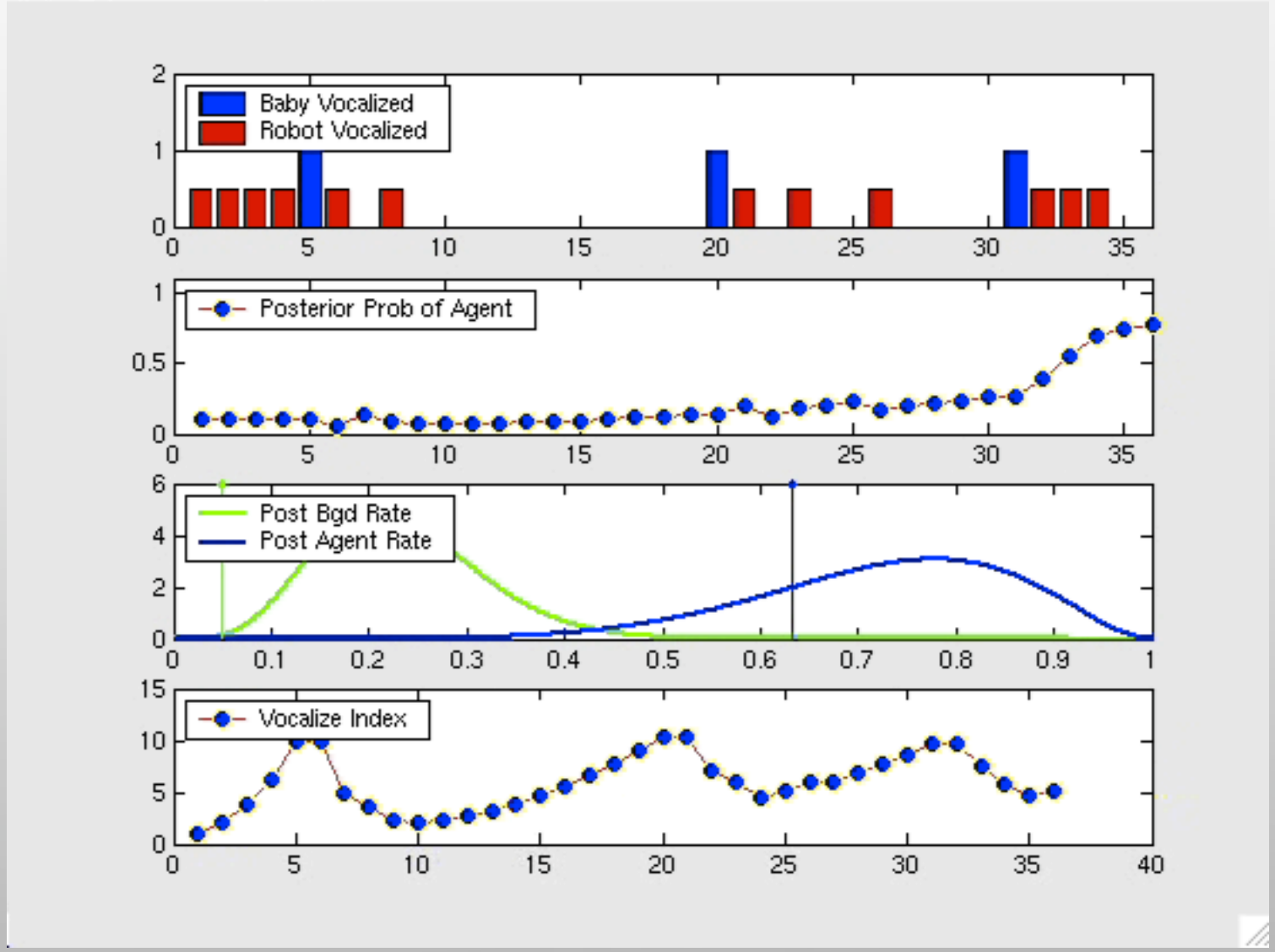
Finding Optimal Infomax Controller

- POMDP with unknown parameters. Exact Solution to Bellman Equation can be found using Dynamic Programing (Movellan, 2004). The result is a function that maps sufficient stats of data history into moment to moment decisions to vocalize or stay quiet (30 decisions per second).
- Very good approximate solution can be learned using *information gain as a reinforcement signal* (Butko, Fasel, Movellan, 2007).

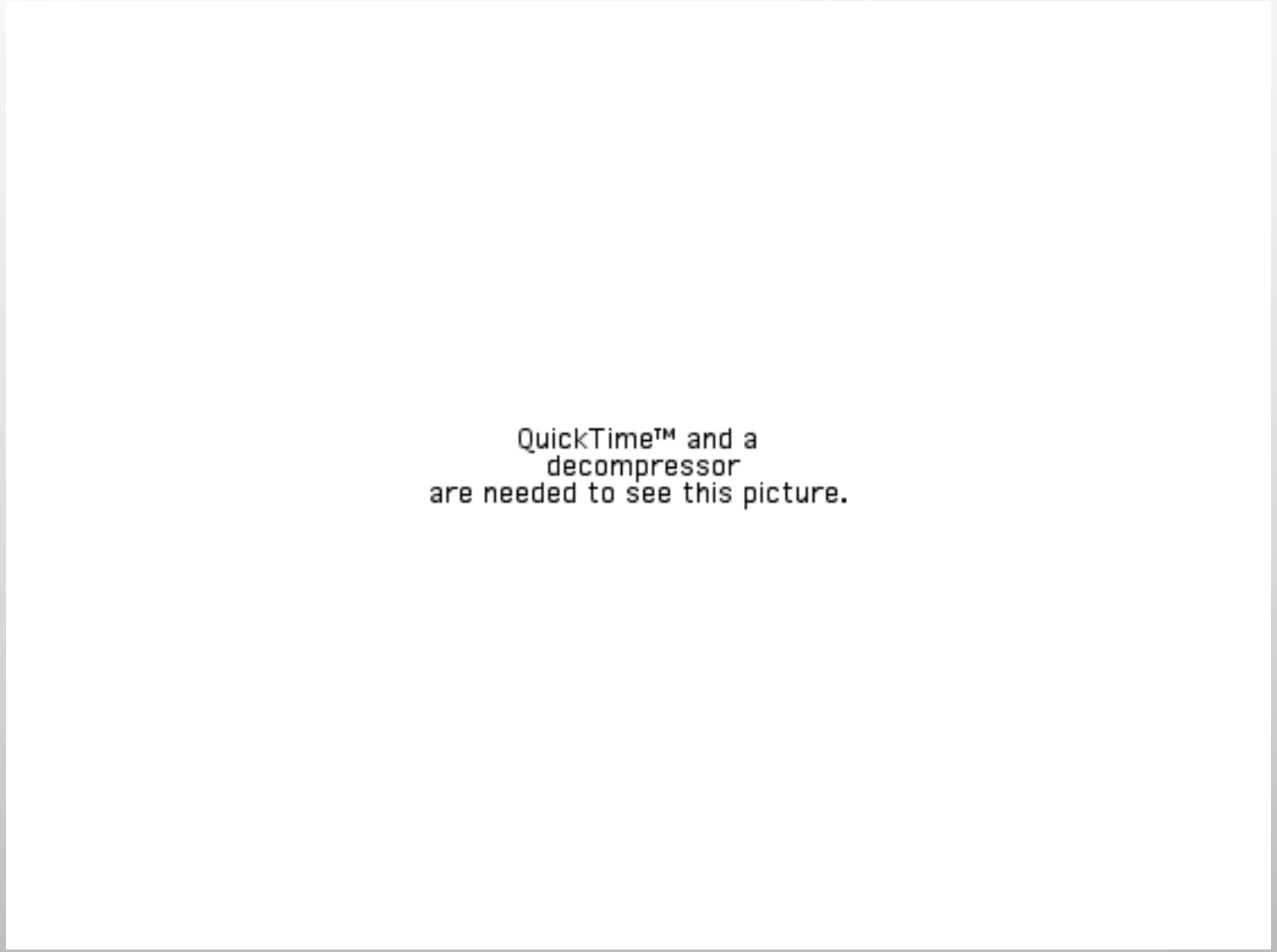
Optimal Controller Matched Qualitatively the Behavior of 9 Month Infants

- Turn-taking behavior, i.e., vocalizations followed by pauses (about 6 seconds long) as if waiting for a response.
- Turns are dynamic, responsive to the changing world.
- In low noise conditions discovers in two or three trials and about 20 seconds whether or not a human is present.

Simulation of Baby Experiment



Simulation of Baby Experiment



Optimal Controller Matched Qualitatively the Behavior of 9 Month Infants

- By nine months infants have developed methods to detect social contingency as quickly and accurately as it can be possibly done, given the statistics of social interaction.
- Note the model explores, but exploration is just part of the optimal control function that maps states to moment to moment actions.

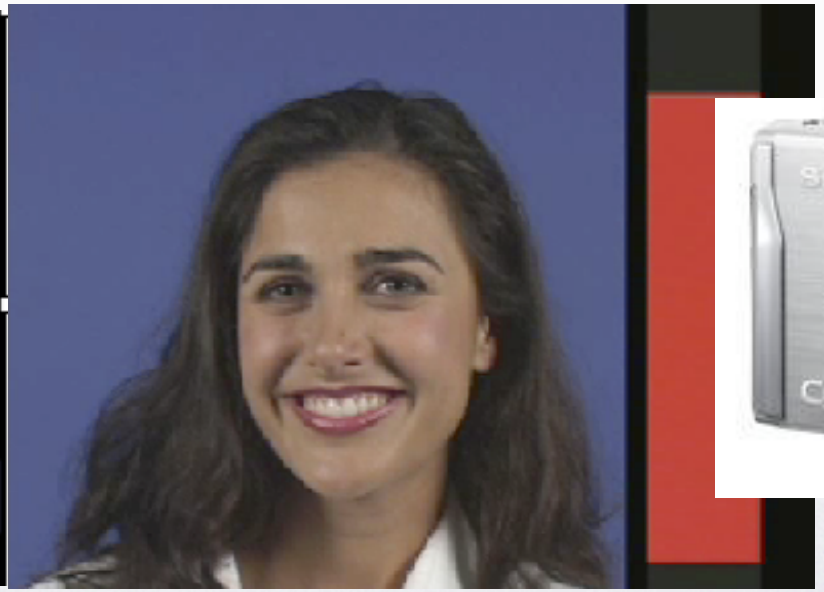
Infomax Control Demo

Automatic Discovery of Object Categories

Example applications

- For comercial level face detection and smile recognition we are using datasets of 100,000 images from the web to train our systems. These images have been labeled by hand one by one to get 20x20 patch with face and facial expression.

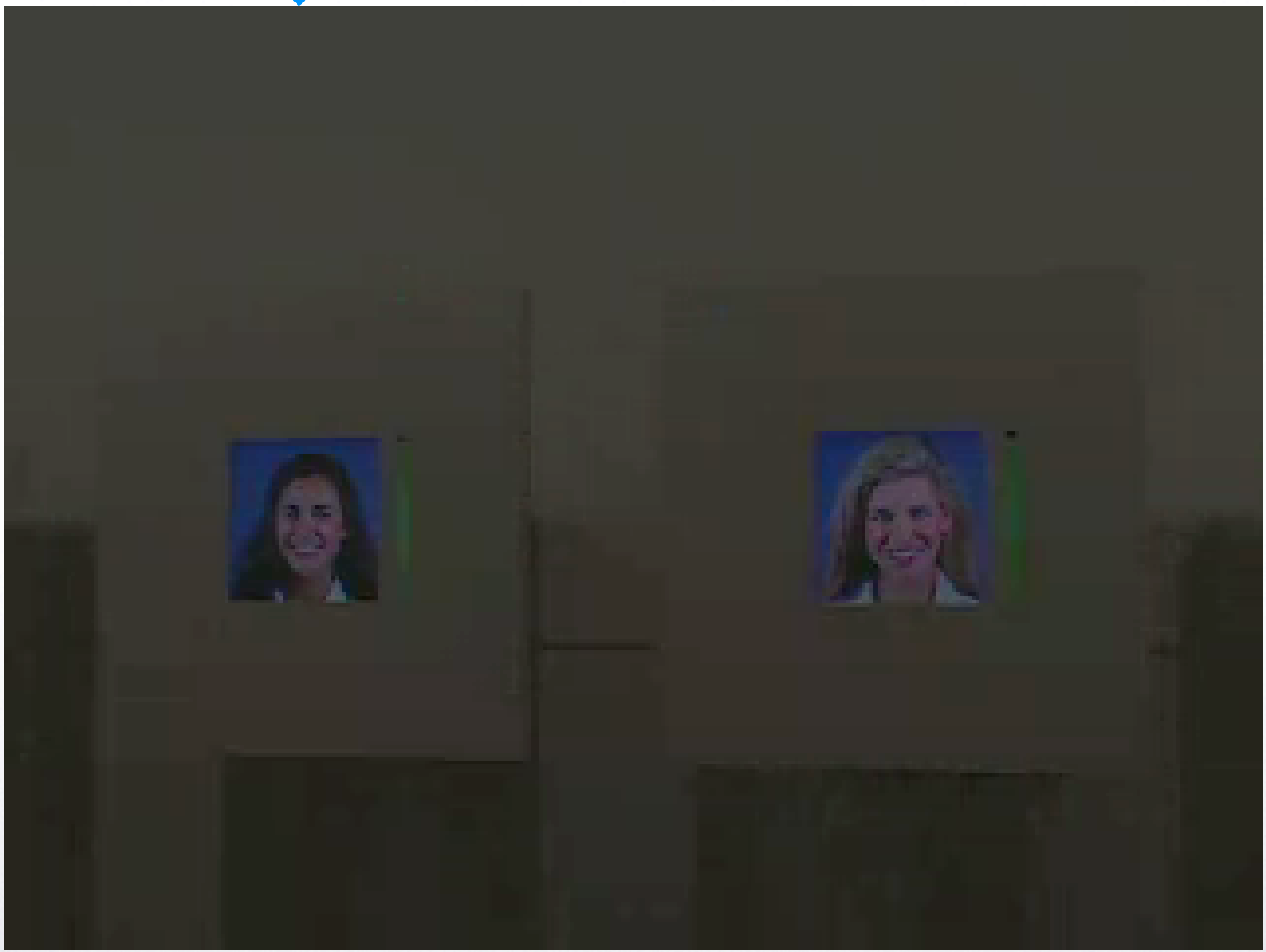
2002



2003

2008

Cheese Project



Christian Moeller
the Williamson Gallery, Art Center Pasadena, 2003

Supervised learning of object categories

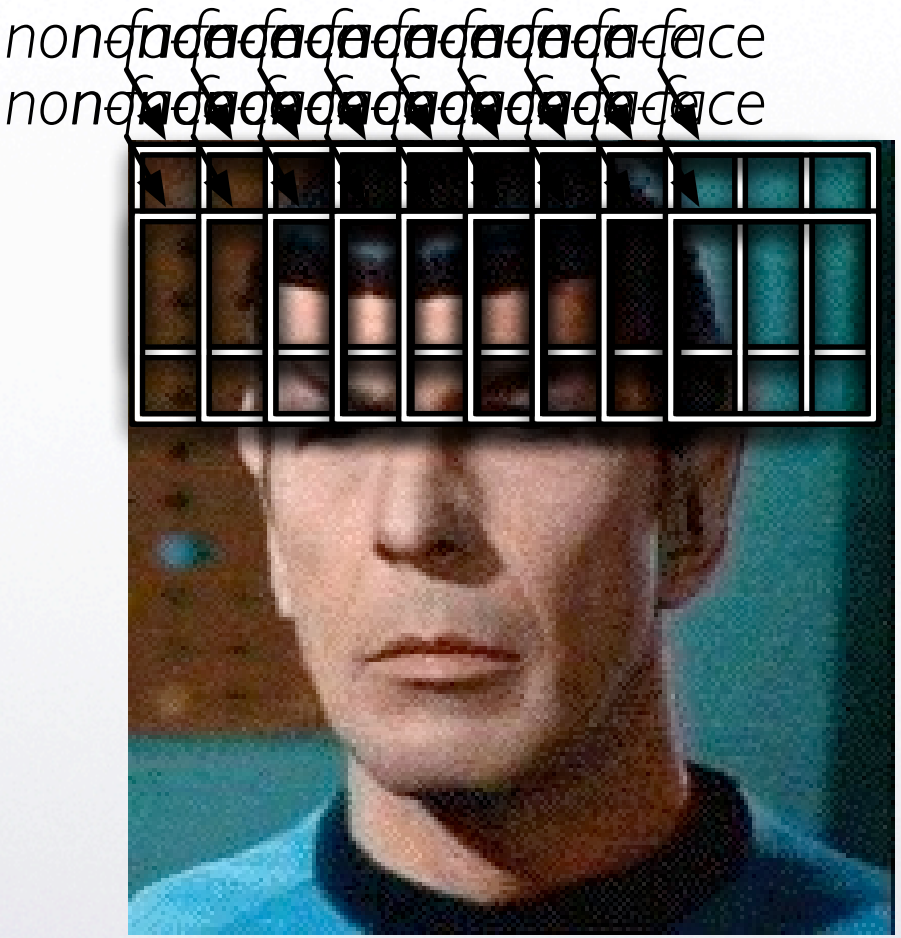
"Face"

"Non-Face"

- Train binary classifier on positive and negative examples of *segmented, scaled, and labeled* image patches.

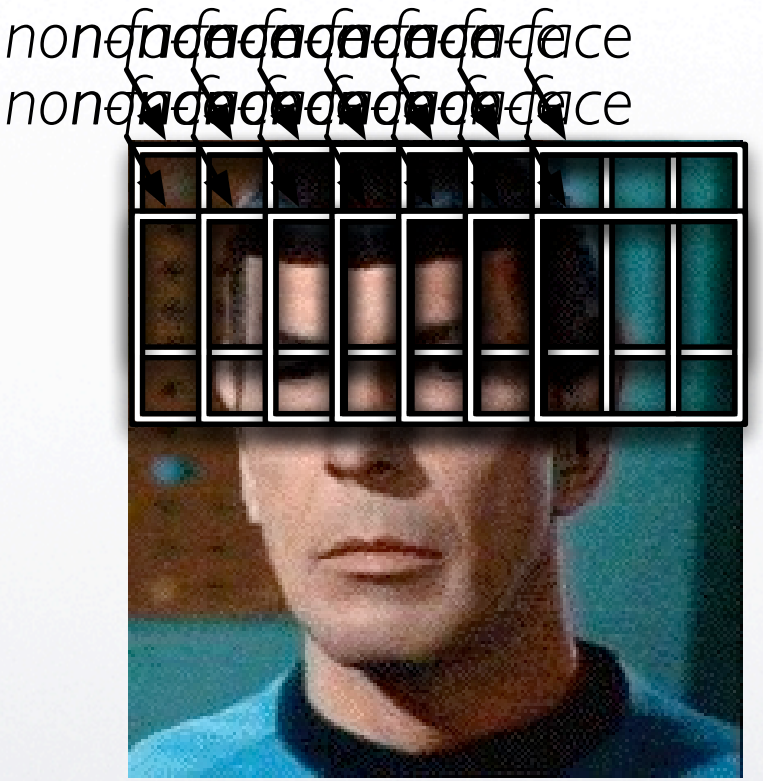
2.1. Object Detection

For all rectangles of a fixed size, label *face* vs. *nonface*:



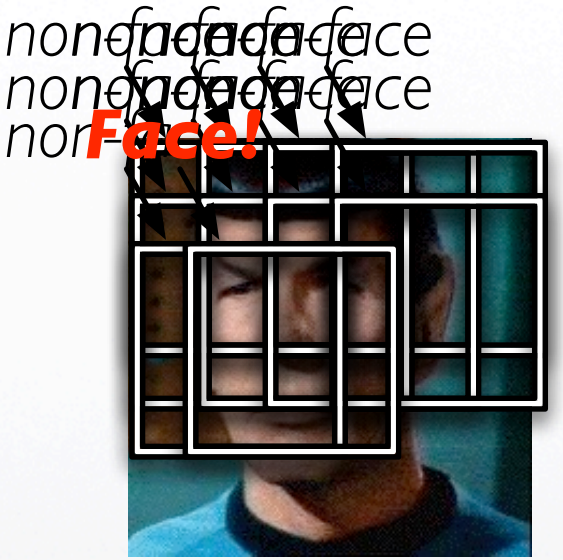
2.1. Object-based segmentation

For all rectangles of a fixed size, label *face* vs. *nonface*:



2.1. Object-based segmentation

For all rectangles of a fixed size, label *face* vs. *nonface*:



Over 400000 patches in a 640 x 480 image

- Could classify all patches in parallel
- Must be fast and accurate

2.1 Learning from unsegmented Images

- All we know is whether the object of interest is or is not in the image. We don't know where it is.
- We need to discover how the object category of interest looks like and detect it in new images.

Computational Analysis (Marr's Style)

Images are collections of pixels that satisfy the following principles:

1. **Common Cause**: Pixels rendered by different objects are conditionally independent of each other.
2. **Opacity:** Each pixel is rendered by a single object.
3. **Shift and Scale Invariance:** The appearance of objects is independent of their location on the image plane and their scale on the image plane.

Simplifying assumptions to maximize speed (not critical to the theory but critical to run in real time on current computers)

1. **Shape**: Objects render square image regions.
2. **Boltzmann**: Likelihood ratios of image patches have a product of experts Boltzmann distribution.
3. We will focus on *2 category problem* (object vs background). The theory works for multi category problems.

Learning

Let

$$G \stackrel{\text{def}}{=} \nabla L(\lambda)(z_k) \stackrel{\text{def}}{=} \left.\frac{\partial L(F_T + \epsilon\delta(z_k, \cdot))}{\partial \epsilon}\right|_{\epsilon=0}$$

$$= \left.\frac{\partial}{\partial \epsilon}\left(\sum_{i=1}^{m} \log \sum_{j=1}^{\kappa} u_{ij} e^{\epsilon\delta(z_k, z_j)} - m \log \sum_{l=1}^{\kappa} v_l e^{\epsilon\delta(z_k, z_l)}\right)\right|_{\epsilon=0}$$

$$= \sum_{i=1}^{m} \frac{u_{ik}}{\sum_{j=1}^{\kappa} u_{ij}} - m\frac{v_k}{\sum_{l=1}^{\kappa} v_l}$$

where

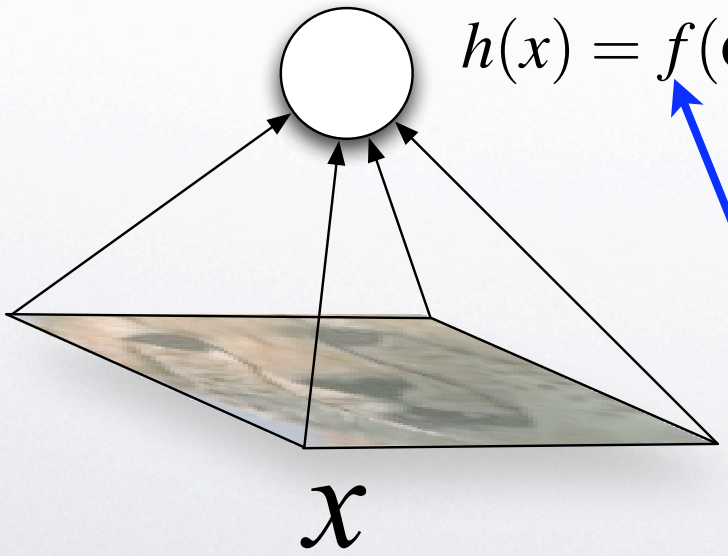
$$u_{ij} \stackrel{\text{def}}{=} \sum_{k=1}^{n_s} e^{F_T(x_{ik})}\delta(x_{ik}, z_j)$$

$$v_k \stackrel{\text{def}}{=} \frac{\beta}{n}\sum_{l=1}^{n} e^{F_T(\tilde{x}_l)}\delta(\tilde{x}_l, z_k) + \frac{1-\beta}{md}\sum_{i=1}^{m}\sum_{\{j:a_j \in s\}} e^{F_T(x_{ij})}\delta(x_{ij}, z_k)$$

Goal is to find h(x) that minimizes

A kernel from a very large pool of preselected candidates

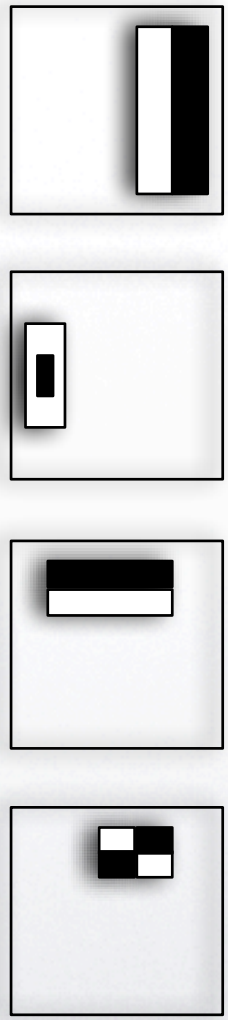
$$h(x) = f(\phi(x))$$



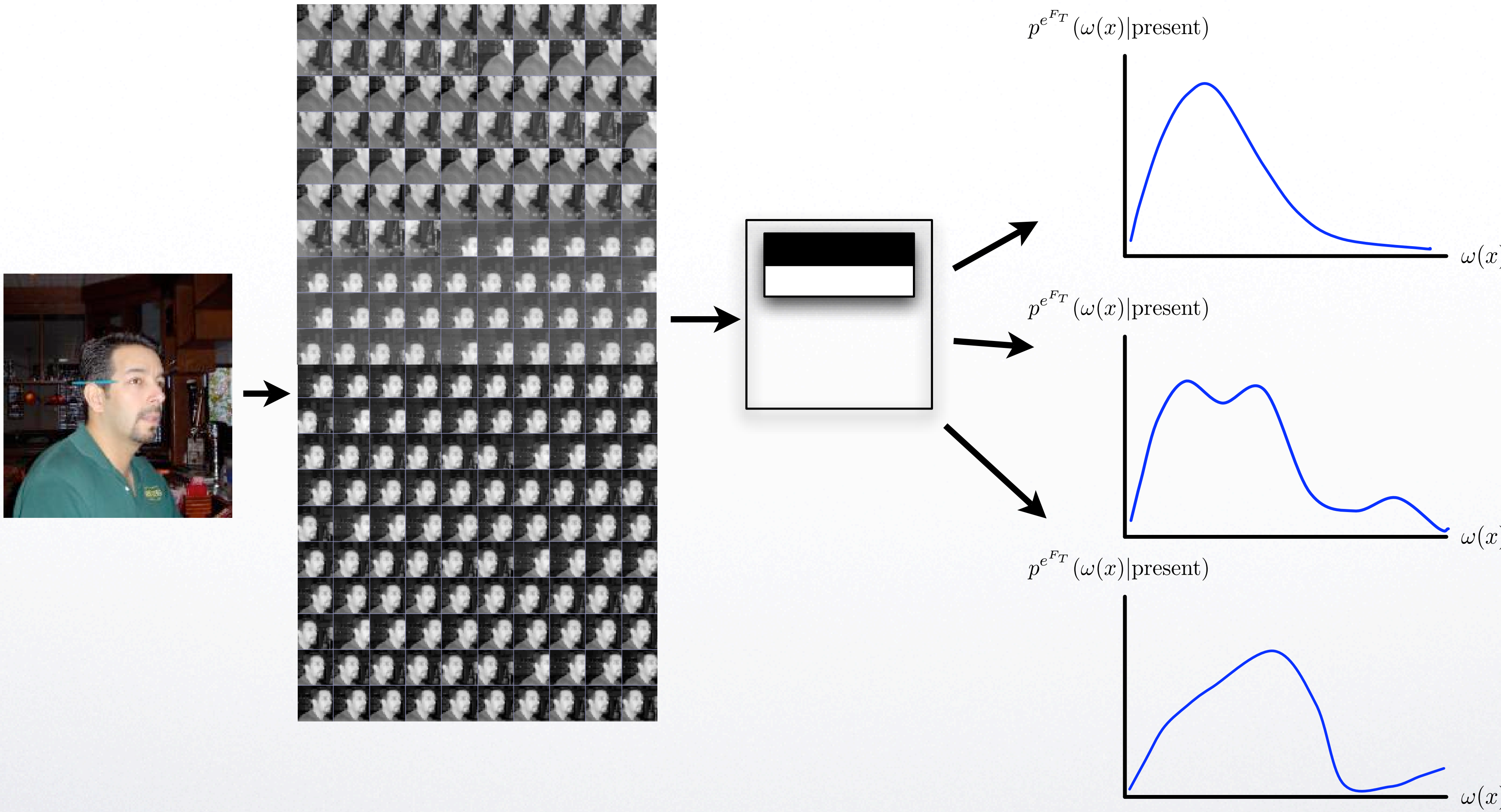
x

A smooth non-linear function optimized to max likelihood given the kernel

Box Kernels

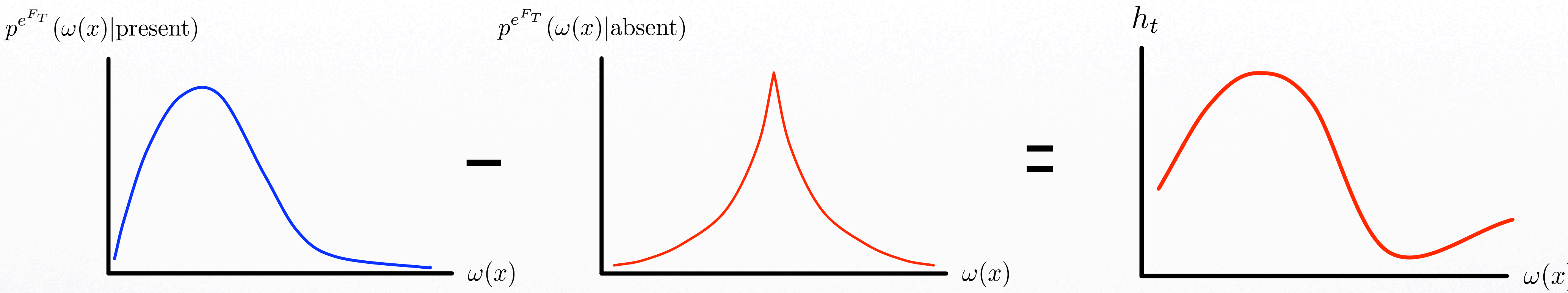


Learning



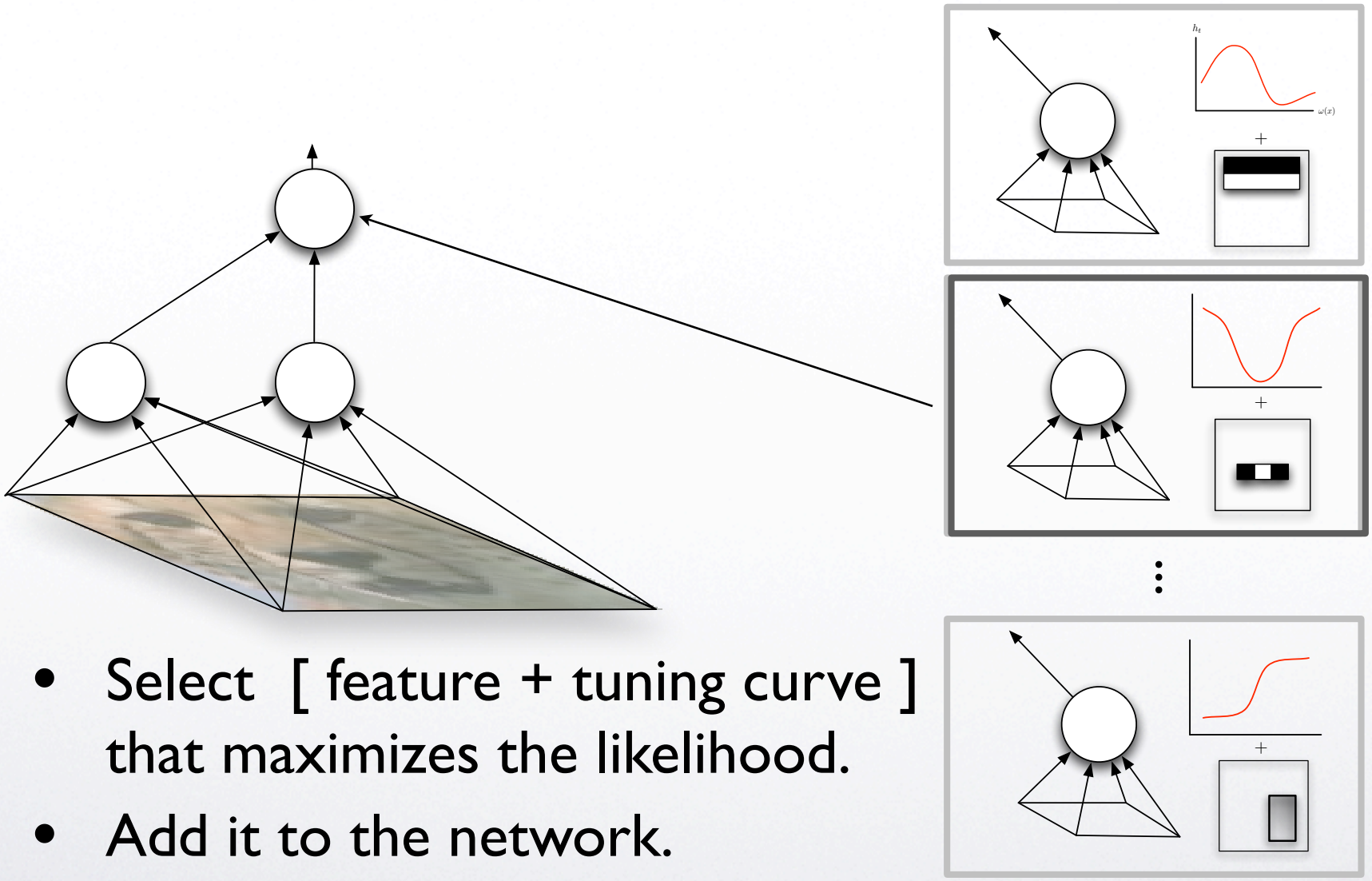
- Crop image patches at all scales
- Compute weighted density estimates of feature output

Learning



- The difference between these is the new transfer function.
- Repeat this process for the sample of background patches

Choosing Features.

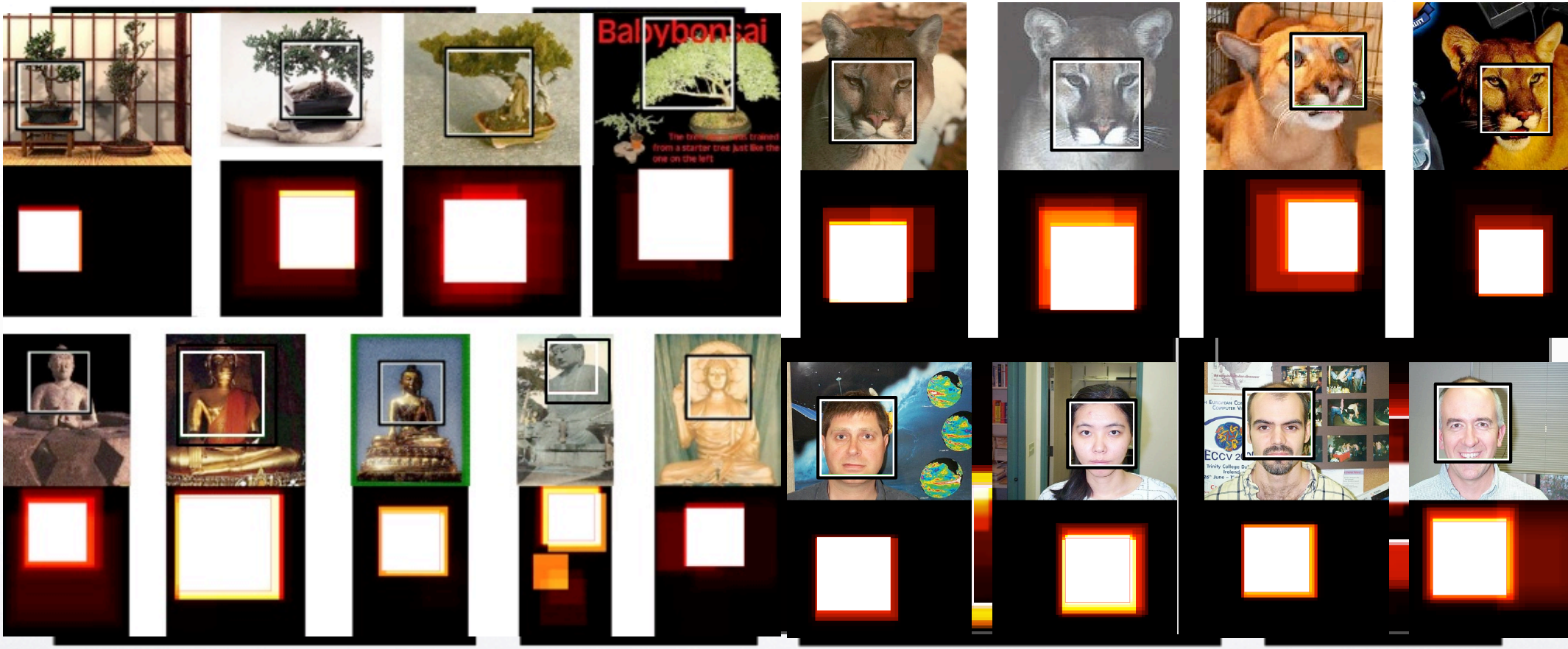


- Select [feature + tuning curve] that maximizes the likelihood.
- Add it to the network.

Inference

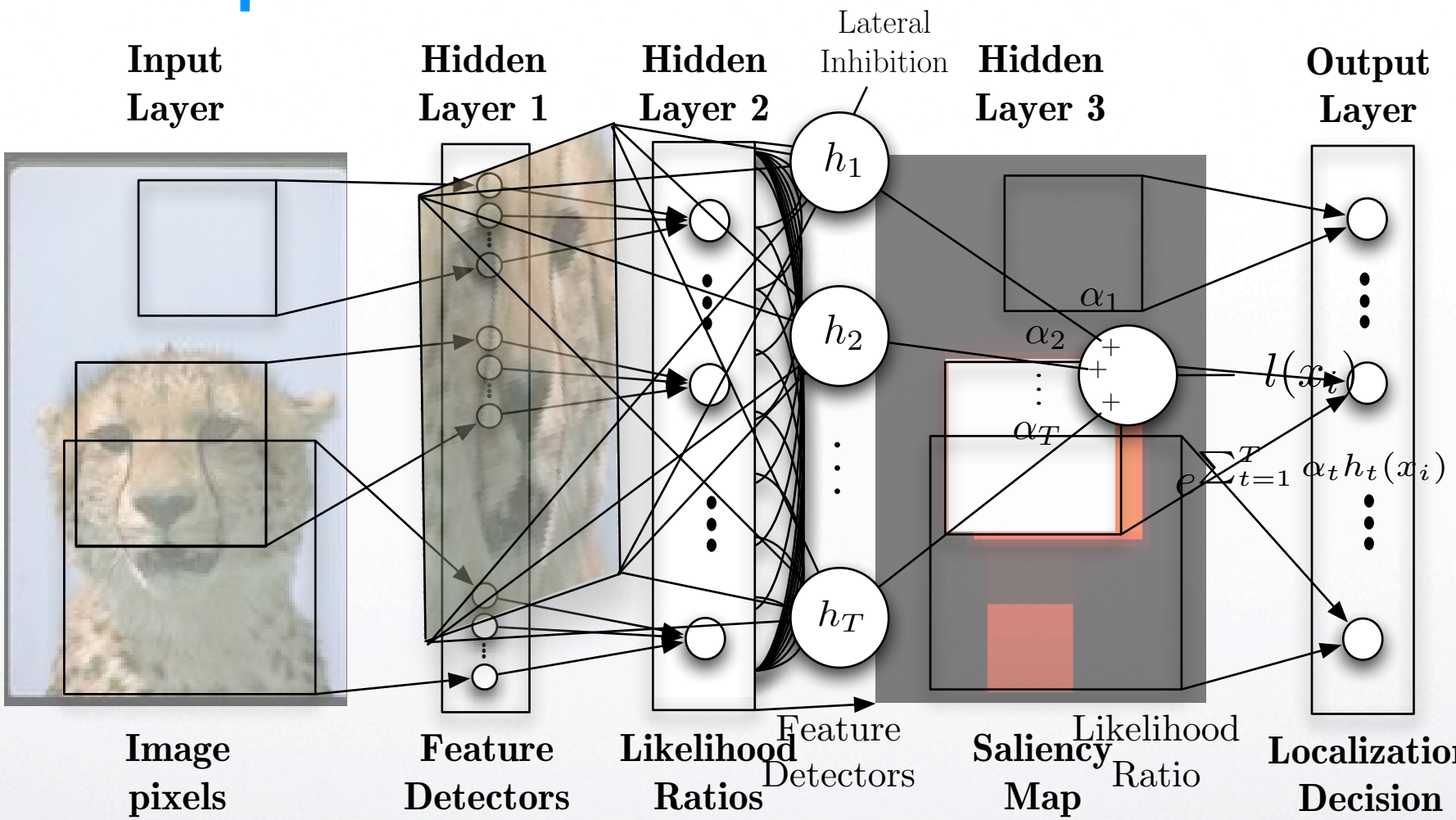
- After the network parameters are learned we can present the resulting network with a new image and request whether it contains the object of interest.
- We can in fact ask for the probability that a pixel contains the object of interest. This results on a segmental image field (*Segmental Boltzman Fields*)

Inference: Caltech 101



- Average two-alternative forced choice performance: 92.7%
- About 26 positive training images per category, 200 negatives

5.4. Implementation Level



- Implemented as convolutional neural net, 4,000,000 hidden units
- *Bayesian approach helps us understand the structure of the network.*

APA Monitor March 2007

BEVERLY

Nick Butko, Ian Fasel, Javier Movellan

Testing John Watson's Hypothesis:

- Infants identify caregivers by detecting contingencies between actions and sensors caused by social agents.

J. S. Watson (1972) Smiling cooing and ``the game", Merrill-Palmer Quarterly, 18:323.

BEVERLY

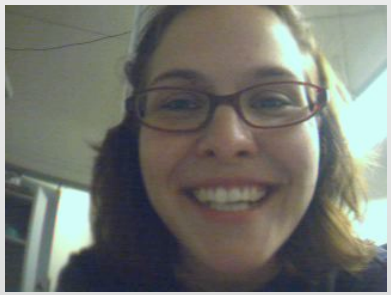
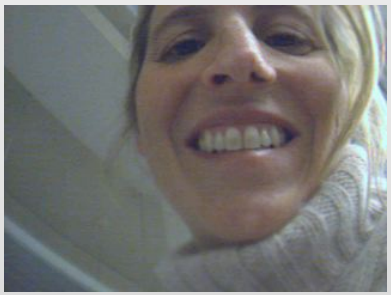
- Social Contingeny Detector: Can tell based on social contingency whether people are there. But it does not know how people look like.
- Segmental Boltzmann Fields: Can discover the appearance of objects from example images that contain or do not contain the object somewhere (Fasel & Movellan, 2006).

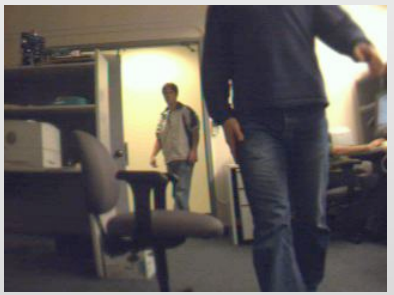
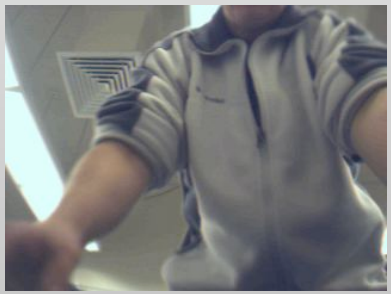
No Contingency Picture
Contingency Picture

Contingency Detector

Contingent Images

Non Contingent Images

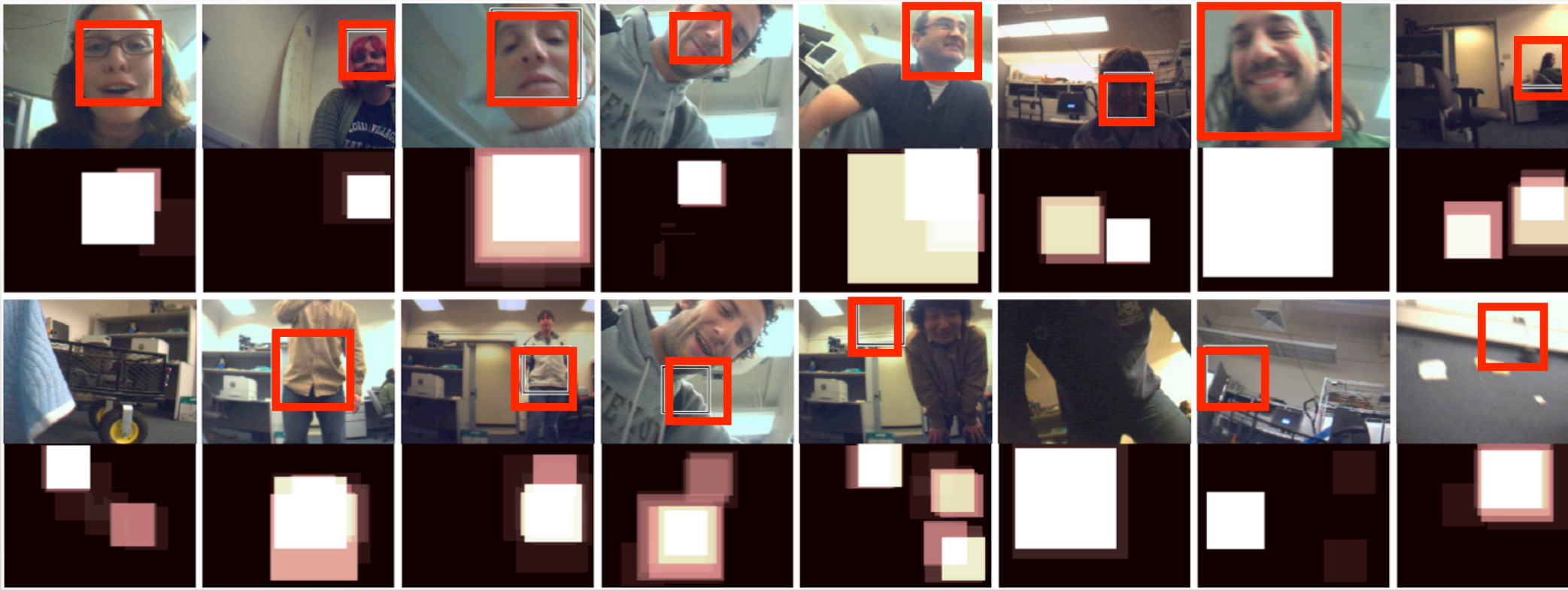




Segmental Boltzmann Fields Algorithm

- 88 minutes of continuous interaction with 9 humans instructed to "make baby robot excited"

- Will data collected with this simple contingency signal be sufficient to learn about what people look like?
  - Very noisy: e.g., "Contingent" images did not contain a face 26% of the time.
- How much interaction will be required to learn a good face detector? Months? Weeks? Days?

After 6 minutes of life:

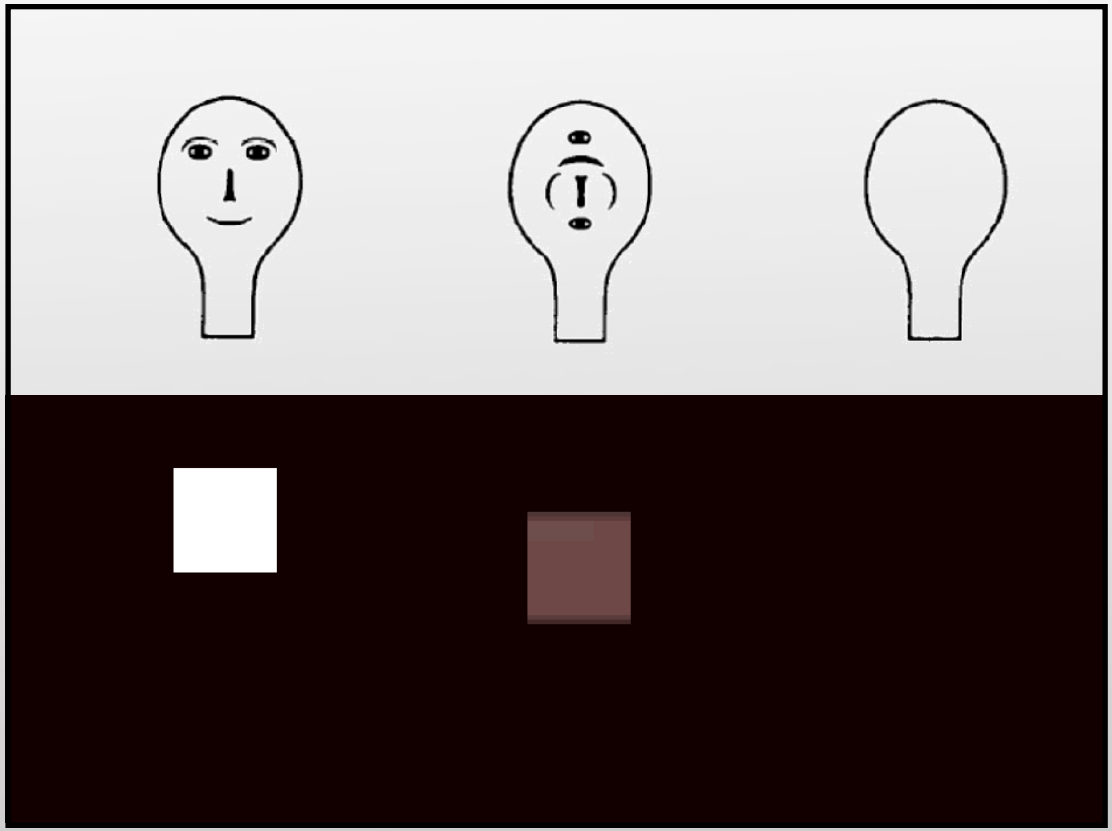


86.2% face detection (generalization images within dataset)

92.3% *person* detection (2 alternative forced choice task)

Butko, Fasel & Movellan (2006) *International Conference on Development and Learning.*

Preference for sketch faces



- 40-minutes-old infants preferentially track faces

Morton & Johnson(1991) CONSPEC and CONLEARN:A two process theory of infant face recognition, Psychological Review, 98, 2, 164-181.

Goren & Wu (1975) Visual following and pattern discrimination of face-like stimuli by newborn infants. Pedicatrics, 56, 544-549.

Lessons:

- John Watson's hypothesis is computationally plausible: Faces may become special because they explain the contingencies we experience in daily life.
- There is enough information in natural images so that the process of becoming special can occur within minutes of interaction with the world. Poverty of the stimulus argument may not be a good one in this case.
- Simple neural network implemented what a Bayesian analyst could see as abstract knowledge of how images are formed. Tuning was done via simple gradient desce
- Stochastic Optimal Control provides a useful formalism to understand learning as an active real time information gathering process.