



Loris Bazzani^{1,2}, Nando de Freitas², Jo-Anne Ting²

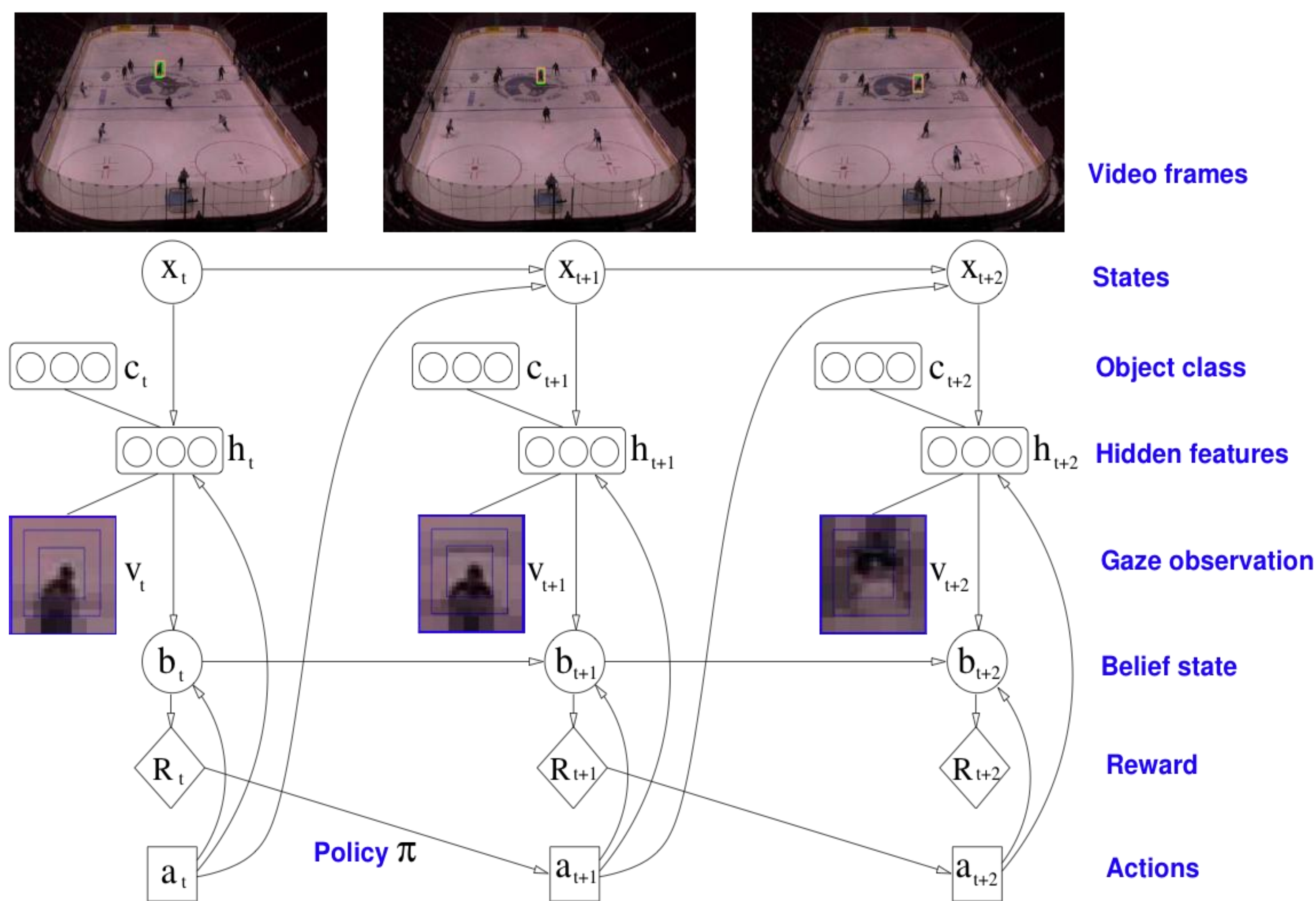
¹ Computer Science, University of Verona, Verona, Italy

² Computer Science, University of British Columbia, Vancouver, Canada

Goals

- Ensuring that deep models can cope with vast video streams
 - Tracking and recognition driven by gaze data
 - Learning a policy to decide where to look at
- Ensuring that deep models are invariant with respect to scale, orientation, location, etc.
 - The **ventral pathway** models object appearance and object label using deep (factored)-restricted Boltzmann machine
 - The **dorsal pathway** models the location, orientation, scale and speed of the attended object
- Improving appearance models for image tracking
- Ensuring that the system is easy to implement and modular

The Proposed Model



- The distribution of the states is estimated with particle filtering
- An attentional mechanism learns to control gazes so as to maximize different objectives

State Space Model & Tracking

- Markovian, nonlinear, non-Gaussian state-space models

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{a}_{t-1}) \text{ for } t \geq 1$$

$$p(\mathbf{h}_t | \mathbf{x}_t, \mathbf{a}_t) \text{ for } t \geq 1$$

- The filtering distribution, aka belief state, is

$$b_t \triangleq p(\mathbf{x}_t | \mathbf{h}_{1:t}, \mathbf{a}_{1:t}) \propto p(\mathbf{h}_t | \mathbf{x}_t, \mathbf{a}_t) \int p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{a}_{t-1}) p(d\mathbf{x}_{t-1} | \mathbf{h}_{1:t-1}, \mathbf{a}_{1:t-1})$$

- The integral is approximated by particle filtering

Policy

- Here, we select actions so as to minimize the uncertainty in the estimate of the filtering distribution, but the reward can be any desired behavior: e.g. maximizing classification accuracy or achieving more abstract goals. The **cumulative reward** of the control algorithm for each action is

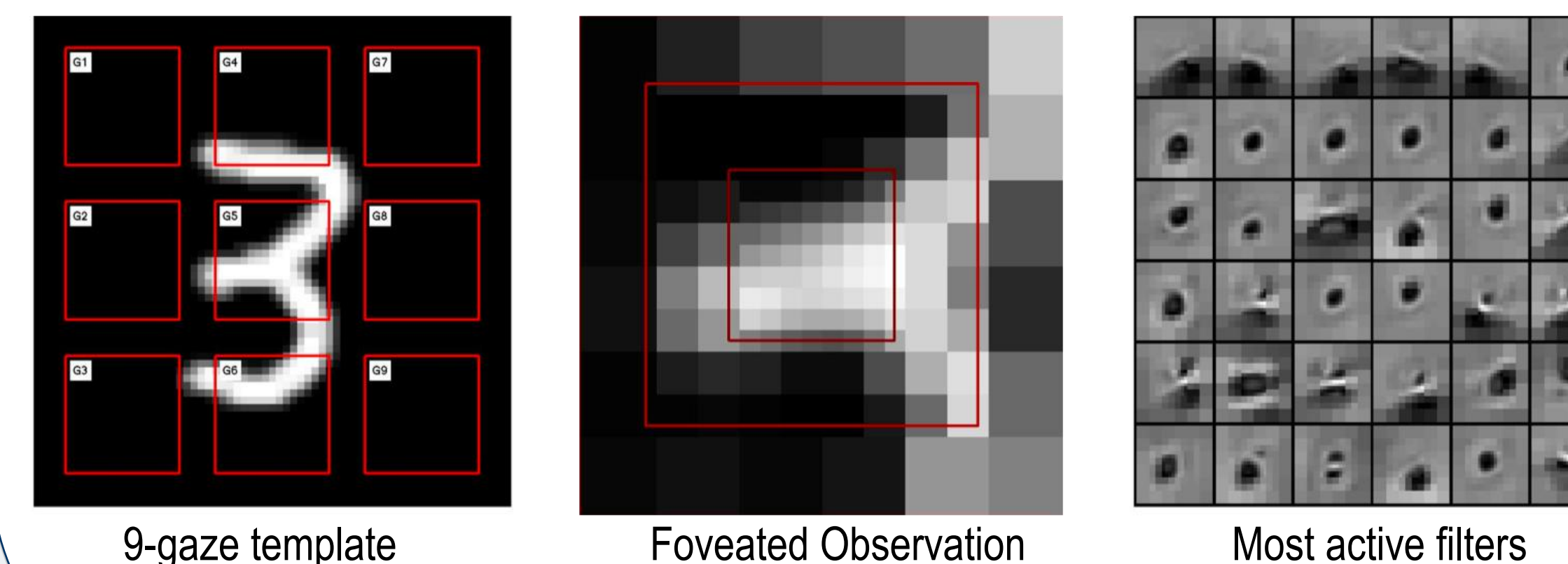
$$R_T(\mathbf{a}_T = k) = \sum_{t=1}^T r_t(p(\mathbf{x}_t | \mathbf{h}_{1:t}, \mathbf{a}_t = k, \mathbf{a}_{1:t-1}))$$

- We use the **hedge algorithm** of Freund and Schapire to learn a stochastic policy:

$$\pi_t(\mathbf{a}_t = k | R_{t-1}) = \frac{\exp(\eta R_{t-1}(\mathbf{a}_t = k))}{\sum_{j=1}^K \exp(\eta R_{t-1}(\mathbf{a}_t = j))}$$

Appearance Model

- We use **(factored)-RBMs** to model the appearance of objects and perform object classification using the gazes chosen by the control module



Appearance Model

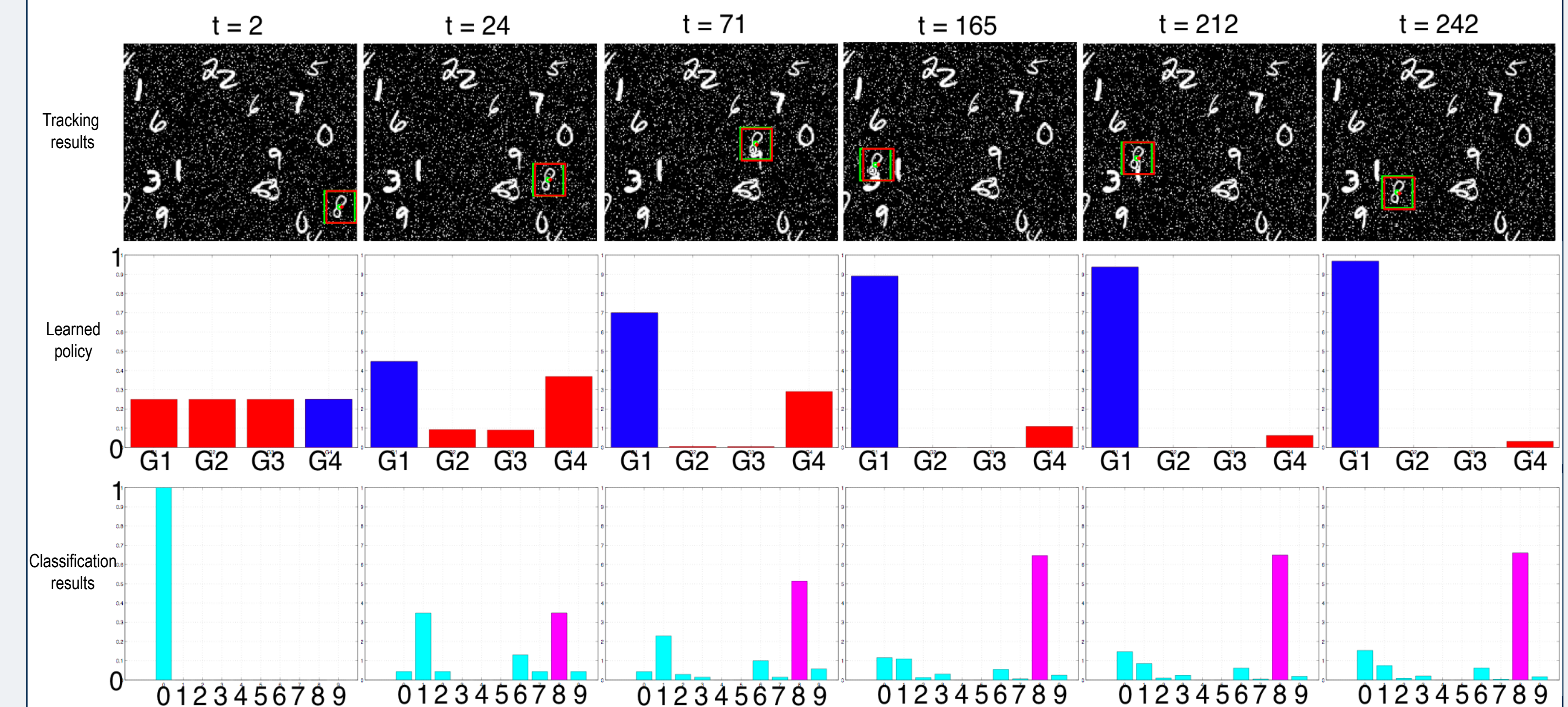
- In image tracking, the observation model is defined in terms of the distance of the observations with respect to a learned RBM template:

$$p(\mathbf{h}_t | \mathbf{x}_t, \mathbf{a}_t = k) \propto e^{-d(\mathbf{h}(\mathbf{x}_t, \mathbf{a}_t = k, \mathbf{v}_t), \mathbf{h}(\mathbf{x}_1, \mathbf{a}_t = k, \mathbf{v}_1))}$$

- For object recognition, we train a classifier on the latent representations of gaze instances and accumulate class decisions over time.

Experiments

- Tracking, online policy learning and recognition on a synthetic example:



- Tracking results on a multi-target, multi-scale synthetic example and a single-target multi-scale real example:

