Learning Attentional Policies for Tracking and Recognition in Video with Deep Networks

Loris Bazzani, Nando de Freitas LORIS.BAZZANI@UNIVR.IT, NANDO@CS.UBC.CA
University of British Columbia, Canada

Hugo Larochelle LAROCHEH@CS.TORONTO.EDU
University of Toronto, Canada

Vittorio Murino VITTORIO.MURINO@IIT.IT
Istituto Italiano di Tecnologia, Italy

Jo-Anne Ting JATING@CS.UBC.CA
University of British Columbia, Canada

Abstract

We propose a novel attentional model for simultaneous object tracking and recognition that is driven by gaze data. Motivated by theories of the human perceptual system, the model consists of two interacting pathways: ventral and dorsal. The ventral pathway models object appearance and classification using deep (factored)-restricted Boltzmann machines. At each point in time, the observations consist of retinal images, with decaying resolution toward the periphery of the gaze. The dorsal pathway models the location, orientation, scale and speed of the attended object. The posterior distribution of these states is estimated with particle filtering. Deeper in the dorsal pathway, we encounter an attentional mechanism that learns to control gazes so as to minimize tracking uncertainty. The approach is modular (with each module easily replaceable with more sophisticated algorithms), straightforward to implement, practically efficient, and works well in simple video sequences.

1. Introduction

Humans track and recognize objects effortlessly and efficiently, exploiting attentional mechanisms (Rensink, 2000; Colombo, 2001) to cope with a vast stream of data. In this paper, we use the human visual system as inspiration to build a model for simultaneous object tracking and recognition from gaze data, as shown in Figure 1. The proposed model also addresses the problem of gaze planning (i.e., where to look in order to achieve some goal, such as minimizing position or speed uncertainty).

The model consists of two interacting modules, ventral and dorsal, which are also known as the *what* and *where* modules respectively. The dorsal pathway is in charge of state estimation and control. At the lowest level of the dorsal pathway, a particle filter (Doucet et al., 2001) is used to estimate the states of the object under consideration, including location, orientation, speed and scale. We make no attempt to implement such states with neural architectures, but it seems clear that they could be encoded with grid cells (McNaughton et al., 2006) and retinotopic maps as in V1 and the superior colliculus (Rosa, 2002; Girard & Berthoz, 2005). At the higher level of the dorsal pathway, a policy governing where to gaze next is learned with an online hedging algorithm (Auer et al., 1998). This policy learning step could be easily improved using other bandit approaches and Bayesian optimization (Brochu et al., 2009; Cesa-Bianchi & Lugosi, 2006; Chaudhuri et al., 2009). The dorsal attentional mechanism is responsible for controlling saccades and, to a significant extent, smooth pursuit (Colombo, 2001).

The ventral pathway consists of a two hidden layer deep network. The second layer corresponds to a multi-fixation RBM (Larochelle & Hinton, 2010), as

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

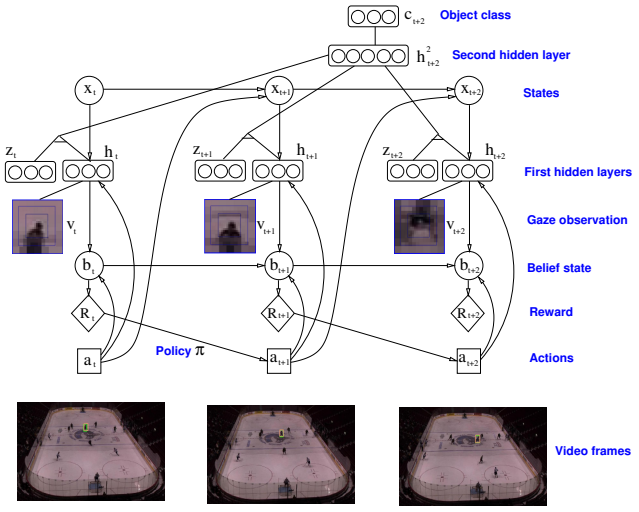

Figure 1. From a sequence of gazes $(\mathbf{v}_t, \mathbf{v}_{t+1}, \mathbf{v}_{t+2})$*, the model infers the hidden features* $\mathbf{h}$ *for each gaze (that is, the activation intensity of each receptive field), the hidden features for the fusion of the sequence of gazes and the object class* $\mathbf{c}$*. Only one time step of classification is kept in the figure for clarity.* $\mathbf{z}_t$ *indicates the relative position of the gaze in a template. The location, size, speed and orientation of the gaze patch are encoded in the state* $\mathbf{x}_t$*. The actions* $\mathbf{a}_t$ *follow a randomized policy* π_t *that depends on the cumulative reward* R_{t-1}*. This particular reward is a function of the belief state* $\mathbf{b}_t = p(\mathbf{x}_t|\mathbf{a}_{1:t}, \mathbf{h}_{1:t})$*, also known as the filtering distribution. Unlike typical commonly used partially observed Markov decision models (POMDPs), the reward is a function of the beliefs. In this sense, the problem is closer to one of sequential experimental design. With more layers in the ventral* $\mathbf{v} - \mathbf{h} - \mathbf{h}^2 - \mathbf{c}$ *pathway, other rewards and policies could be designed to implement higher-level attentional strategies.*

shown in Figure 1. It accumulates information from the first hidden layers at consecutive time steps. For the first layers, we use (factored)-restricted Boltzmann machines (RBMs) (Hinton & Salakhutdinov, 2006; Ranzato & Hinton, 2010; Welling et al., 2005), but autoencoders (Vincent et al., 2008), sparse coding (Olshausen & Field, 1996; Kavukcuoglu et al., 2009), two-layer ICA (Köster & Hyvärinen, 2007) and convolutional architectures (Lee et al., 2009) could also be adopted in this module. At present, we pre-train these appearance models.

The proposed system can be motivated from different perspectives. First, starting with (Isard & Blake, 1996), many particle filters have been proposed for image tracking, but these typically use simple observation models such as B-splines (Isard & Blake, 1996) and color templates (Okuma et al., 2004). RBMs are more expressive models of shape, and hence, we conjecture that they will play a useful role where simple appearance models fail. Second, from a deep learning computational perspective, this work allows us to tackle large images and video. The use of fixations synchronized with information about the state (e.g. location and scale) of such fixations, eliminates the need to look at the entire image or video. Third, the system is invariant to image transformations encoded in the state, such as location, scale and orientation. Fourth, from a dynamic sensor network perspective, this paper presents a very simple, but efficient, novel way of deciding how to gather measurements dynamically. Lastly, in the context of psychology, the proposed model realizes to some extent the functional architecture for dynamic scene representation of (Rensink, 2000). The rate at which different attentional mechanisms develop in newborns (including alertness, saccades and smooth pursuit, attention to object features and high-level task driven attention) guided the design of the proposed approach and was a great source of inspiration (Colombo, 2001).

Recently, a dynamic RBM state-space model was proposed in (Taylor et al., 2010). Both the implementation and intention behind that proposal are different from the approach discussed here. To the best of our knowledge, the approach presented here is the first successful attempt to combine dynamic state estimation from gazes with online policy learning for gaze adaptation, using deep belief network models of appearance. Many other dual-pathway architectures have been proposed in computational neuroscience, including (Olshausen et al., 1993; Postma et al., 1997), but we believe ours has the advantage that it is very simple, *modular* (with each module easily replaceable), suitable for large datasets and easy to extend.

2. Model specification

We describe the dorsal pathway in Sections 2.1 (state-space representation), 2.2 (reward function and control policy) and 2.3 (observation model for the state-space model). The corresponding state estimation and control algorithms are presented in Section 3. The ventral pathway is described briefly in Section 2.3.

2.1. State-space model

The standard approach to image tracking is based on the formulation of Markovian, nonlinear, non-Gaussian state-space models, which are solved with approximate Bayesian filtering techniques. In this setting, the unobserved signal (object's position, velocity, scale, orientation or discrete set of operation) is denoted $\{\mathbf{x}_t \in \mathcal{X}; t \in \mathbb{N}\}$. This signal has initial distribution $p(\mathbf{x}_0)$ and transition equation $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{a}_{t-1})$.

Here $\mathbf{a}_t \in \mathcal{A}$ denotes an action, at time t, defined on a discrete state space of size $|\mathcal{A}| = K$. The observations $\{\mathbf{h}_t \in \mathcal{H}; t \in \mathbb{N}^*\}$, are assumed to be conditionally independent given the state process $\{\mathbf{x}_t; t \in \mathbb{N}\}$. (Note that from the state space model perspective the observations are the hidden units of the first layer of the ventral path model.) In summary, the state-space model is described by the following distributions:

$$\begin{aligned} p(\mathbf{x}_0) & \\ p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{a}_{t-1}) & \quad \text{for } t \geq 1 \\ p(\mathbf{h}_t|\mathbf{x}_t, \mathbf{a}_t) & \quad \text{for } t \geq 1, \end{aligned}$$

where $\mathbf{x}_{0:t} \triangleq \{\mathbf{x}_0, ..., \mathbf{x}_t\}$ and $\mathbf{h}_{1:t} \triangleq \{\mathbf{h}_1, ..., \mathbf{h}_t\}$ represent the states and the observations up to time t, respectively. For the transition model, we will adopt a classical autoregressive process. The appearance model $p(\mathbf{h}_t|\mathbf{x}_t, \mathbf{a}_t)$ is slightly more involved and will be discussed in Section 2.3.

Our aim is to estimate recursively in time the *posterior distribution* $p(\mathbf{x}_{0:t}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t})$ and its associated features, including the marginal distribution $\mathbf{b}_t \triangleq p(\mathbf{x}_t|\mathbf{h}_{1:t}, \mathbf{a}_{1:t})$ — known as the *filtering distribution* or *belief state*. This distribution satisfies the following recurrence:

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

Except for standard distributions (*e.g.* Gaussian or discrete), this recurrence is intractable.

2.2. Reward function and policy

To complete the specification of the model, we need to introduce the policy $\pi(\cdot)$ and an instantaneous reward function $r_t(\cdot)$. The reward can be any desired behavior for the system, such as minimizing posterior uncertainty or achieving a more abstract goal. We focus on gathering observations so as to minimize the uncertainty in the estimate of the filtering distribution: $r_t(\mathbf{b}_t) \triangleq u[\widetilde{p}(\mathbf{x}_t|\mathbf{h}_{1:t}, \mathbf{a}_{1:t})]$. More specifically, as discussed later, this reward will correspond to the variance of the importance weights of the particle filter approximation $\widetilde{p}(\mathbf{x}_t|\mathbf{h}_{1:t}, \mathbf{a}_{1:t})$ of the belief state. In our current implementation, each action is a different gaze location. The objective is to choose where to look so as to minimize the uncertainty about the belief state.

We also need to introduce the cumulative reward of the control algorithm for each action:

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

The actions are distributed according to the following stochastic policy:

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

where $\eta > 0$ is a parameter. We have defined the policy in this way as it enables us to borrow decision making algorithms from the online learning framework (Auer et al., 1998) with very little effort. Here, we will adopt the *hedge algorithm* for full information games described in (Auer et al., 1998). Although this algorithm works well and has vanishing regret, as mentioned in the introduction, one could adopt other bandit techniques (Cesa-Bianchi & Lugosi, 2006; Chaudhuri et al., 2009) or Bayesian optimization (Brochu et al., 2009) to extend the policy to continuous action spaces and treat imperfect information games (only one gaze allowed at each time step).

2.3. Appearance model

We use (factored)-RBMs to model the appearance of objects and perform object classification using the gazes chosen by the control module. These undirected probabilistic graphical models are governed by a Boltzmann distribution over the gaze data $\mathbf{v}_t$ and the hidden features $\mathbf{h}_t \in \{0, 1\}^{n_h}$. We assume that the receptive fields $\mathbf{w}$, also known as RBM weights or filters, have been trained beforehand. We also assume that readers are familiar with these models and, if otherwise, refer them to (Ranzato & Hinton, 2010; Swersky et al., 2010).

In image tracking, the observation model is often defined in terms of the distance of the observations with respect to a template τ,

$$p(\mathbf{h}_t|\mathbf{x}_t, \mathbf{a}_t) \propto e^{-d(\mathbf{h}(\mathbf{x}_t, \mathbf{a}_t), \tau)},$$

where $d(\cdot, \cdot)$ denotes a distance metric and τ an object template (for example, a color histogram or spline). In this model, the observation $\mathbf{h}(\mathbf{x}_t, \mathbf{a}_t)$ is a function of the current state hypothesis and the selected action. The problem with this approach is eliciting a good template. Often color histograms or splines are insufficient. For this reason, we will construct the templates with (factored)-RBMs as follows. First, optical flow is used to detect new object candidates entering the visual scene. Second, we assign a template to the detected object candidate, which consists of several gazes covering the field of motion, as shown in Figure 2 for $K = 9$ gazes. The same figure also shows the typical foveated observations (higher resolution in the center and lower in the periphery of the gaze) and the receptive fields for these observations learned beforehand with an RBM. The control algorithm will be

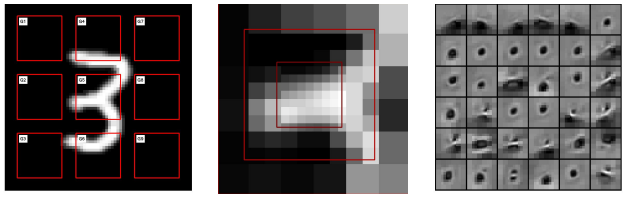

Figure 2. (a) Template with 9 gazes initialized automatically when motion is detected. (b) Foveal observation corresponding to gaze G5 in the template. (c) The most active RBM filters for this observation.

used to learn which of the gazes in the template are more fruitful. That is, each action will correspond to the selection of one of these gaze options. Finally, we define the likelihood of each observation directly in terms of the distance of the hidden units of the RBM $\mathbf{h}(\mathbf{x}_t, \mathbf{a}_t, \mathbf{v}_t)$ to the hidden units of each template region $\mathbf{h}(\mathbf{x}_1, \mathbf{a}_t = k, \mathbf{v}_1)$, $k = 1 : K$, initialized in the first frame. That is,

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

The above template is static, but conceivably one could adapt it over time.

The appearance module also performs object recognition, classifying a sequence of gaze instances selected with the gaze policy. We implement a multi-fixation model very similar to the one proposed in (Larochelle & Hinton, 2010), where the binary variables $\mathbf{z}_t$ (see Figure 1) are introduced to encode the relative gaze location, $\mathbf{a}_t$ in the present implementation, in a factored-RBM. We refer the reader to this citation for detailed information on this part of the model. We experimented with a single fixation module, but found the multi-fixation module of (Larochelle & Hinton, 2010) to increase classification accuracy. To improve the estimate the class variable $\mathbf{c}_t$ over time, we accumulate the classification decisions at each time step.

Note that the process of pursuit (tracking) is essential to classification. As the target is tracked, the algorithm fixates at random locations near the target's location estimate. The size and orientation of these fixations also depends on the corresponding state estimates. Note that we don't fixate exactly at the target location estimate as this would provide only one distinct fixation over several time steps if the tracking policy has converged to a specific gaze. It should also be pointed out that instead of using random fixations, one could again use the control strategy proposed in this paper to decide where to look with respect to the track estimate so as to reduce classification uncertainty. We leave the implementation of this extra attentional mechanism for future work.

1. *Initialization*, $t = 0$.
   - For $i = 1, \ldots, N$, $\mathbf{x}_0^{(i)} \sim p\left(\mathbf{x}_0\right)$, set $t = 1$, and initialize the template.
   - Set $R_0(\mathbf{a}_0 = k) = 0$ for $k = 1, \ldots, K$, where K is the number of gazes in the template
2. *Importance sampling step*
   - For $i = 1, \ldots, N$, $\widetilde{\mathbf{x}}_t^{(i)} \sim q_t\left(d\mathbf{x}_t^{(i)} \middle| \widetilde{\mathbf{x}}_{0:t-1}^{(i)}, \mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right)$ and set $\widetilde{\mathbf{x}}_{0:t}^{(i)} = \left(\mathbf{x}_{0:t-1}^{(i)}, \widetilde{\mathbf{x}}_t^{(i)}\right)$.
   - For $i = 1, \ldots, N$, $k = 1, \ldots, K$, , evaluate the importance weights

$$\widetilde{w}_t^{(i),k} \propto \frac{p\left(\mathbf{h}_t|\widetilde{\mathbf{x}}_t^{(i)}, \mathbf{a}_t = k\right) p\left(\widetilde{\mathbf{x}}_t^{(i)}|\widetilde{\mathbf{x}}_{0:t-1}^{(i)}, \mathbf{a}_{t-1}\right)}{q_t\left(\widetilde{\mathbf{x}}_t^{(i)} \middle| \widetilde{\mathbf{x}}_{0:t-1}^{(i)}, \mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right)}.$$

   - Normalise the importance weights $w_t^{(i),k} = \frac{\widetilde{w}_t^{(i),k}}{\sum_{j=1}^N \widetilde{w}_t^{(j),k}}$
3. *Gaze control step*
   - Update the policy

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

   - Receive rewards $r_{t,k} = \sum_{i=1}^N \left(w_t^{(i),k}\right)^2$ for $k = 1, \ldots, K$
   - Set $R_t(\mathbf{a}_t = k) = R_{t-1}(\mathbf{a}_t = k) + r_{t,k}$ for $k = 1, \ldots, K$
   - Sample action $k^\star$ according to the policy $\pi_t(\cdot)$.
   - Set $w_t^{(i)} = w_t^{(i),k^\star}$ for $i = 1, \ldots, N$
4. *Selection step*
   - Resample with replacement N particles $\left(\mathbf{x}_{0:t}^{(i)}; i = 1, \ldots, N\right)$ from the set $\left(\widetilde{\mathbf{x}}_{0:t}^{(i)}; i = 1, \ldots, N\right)$ according to the normalized importance weights $w_t^{(i)}$.
   - Set $t \leftarrow t + 1$ and go to step 2.

Figure 3. Particle filtering algorithm with gaze control.

3. Algorithm

Since the belief state cannot be computed analytically, we will adopt particle filtering to approximate it. The algorithm is shown in Figure 3. We refer readers to (Doucet et al., 2001) for a more in depth treatment of these sequential Monte Carlo methods. Assume that at time $t-1$ we have $N \gg 1$ particles (samples) $\{\mathbf{x}_{0:t-1}^{(i)}\}_{i=1}^N$ distributed according to $p\left(d\mathbf{x}_{0:t-1}|\mathbf{h}_{1:t-1}, \mathbf{a}_{1:t-1}\right)$. We can approximate this belief state with the following empirical distribution $\widehat{p}\left(d\mathbf{x}_{0:t-1}|\mathbf{h}_{1:t-1}, \mathbf{a}_{1:t-1}\right) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_{0:t-1}^{(i)}}\left(d\mathbf{x}_{0:t-1}\right)$. Particle filters combine sequential importance sampling with a selection scheme designed to obtain N new particles $\{\mathbf{x}_{0:t}^{(i)}\}_{i=1}^N$ distributed approximately according to $p\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right)$.

3.1. Importance sampling step

The joint distributions $p\left(d\mathbf{x}_{0:t-1}|\mathbf{h}_{1:t-1},\mathbf{a}_{1:t-1}\right)$ and $p\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$ are of different dimension. We first modify and extend the current paths $\mathbf{x}_{0:t-1}^{(i)}$ to obtain new paths $\widetilde{\mathbf{x}}_{0:t}^{(i)}$ using a proposal kernel $q_t\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{x}_{0:t-1},\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$. As our goal is to design a sequential procedure, we set $q_t\left(d\widetilde{\mathbf{x}}_{0:t}|\,\mathbf{x}_{0:t-1},\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right) = \delta_{\mathbf{x}_{0:t-1}}\left(d\widetilde{\mathbf{x}}_{0:t-1}\right) q_t\left(d\widetilde{\mathbf{x}}_t|\,\widetilde{\mathbf{x}}_{0:t-1},\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$, that is $\widetilde{\mathbf{x}}_{0:t} = (\mathbf{x}_{0:t-1},\widetilde{\mathbf{x}}_t)$. The aim of this kernel is to obtain new paths whose distribution $q_t\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right) = p\left(d\widetilde{\mathbf{x}}_{0:t-1}|\mathbf{h}_{1:t-1},\mathbf{a}_{1:t-1}\right) q_t\left(d\widetilde{\mathbf{x}}_t|\,\widetilde{\mathbf{x}}_{0:t-1},\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$ is as "close" as possible to $p\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$. Since we cannot choose $q_t\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right) = p\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$ because this is the quantity we are trying to approximate in the first place, it is necessary to weight the new particles so as to obtain consistent estimates. We perform this "correction" with importance sampling, using the weights:

$$\widetilde{w}_t = \widetilde{w}_{t-1}\frac{p\left(\mathbf{h}_t|\widetilde{\mathbf{x}}_t,\mathbf{a}_t\right)p\left(d\widetilde{\mathbf{x}}_t|\widetilde{\mathbf{x}}_{0:t-1},\mathbf{a}_{t-1}\right)}{q_t\left(d\widetilde{\mathbf{x}}_t|\,\widetilde{\mathbf{x}}_{0:t-1},\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)}.$$

The choice of the transition prior as proposal distribution is by far the most common one. In this case, the importance weights reduce to the expression for the likelihood. However, it is possible to construct better proposal distributions, which make use of more recent observations, using object detectors (Okuma et al., 2004), saliency maps (Itti et al., 1998), optical flow, and approximate filtering methods, as in the unscented particle filter. One could also easily incorporate strategies to manage data association and other tracking related issues. After normalizing the weights, $w_t^{(i)} = \frac{\widetilde{w}_t^{(i)}}{\sum_{j=1}^N \widetilde{w}_t^{(j)}}$, we obtain the following estimate of the filtering distribution:

$$\widetilde{p}\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right) = \sum_{i=1}^{N} w_t^{(i)}\delta_{\widetilde{\mathbf{x}}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right).$$

3.2. Gaze control step

We treat the problem of choosing a gaze (from the template) with a portfolio allocation algorithm called Hedge (Freund & Schapire, 1997; Auer et al., 1998). Hedge is an algorithm that, at each time step t, updates the policy $\pi_t(\mathbf{a}_t|R_{t-1})$ for each allowed action (see (Auer et al., 1998)). It then selects an action $k^\star$ according to this policy as shown in Figure 3. The immediate reward is defined as the variance of the normalized importance weights. This choice is motivated by the fact that the (factored)-RBM observation model for the state-space representation, $p\left(\mathbf{h}|\mathbf{x},\mathbf{a}\right)$, is very peaked.

Note that we assumed a perfect information game. However, it is possible to only observe one of the gazes at each time using the EXP3 algorithm from (Auer et al., 1998) or Bayesian optimization techniques (Brochu et al., 2009).

3.3. Selection step

The aim of the selection is to obtain an "unweighted" approximate empirical distribution $\widehat{p}\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$ of the weighted measure $\widetilde{p}\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t},\mathbf{a}_{1:t}\right)$. The basic idea is to discard the samples with small weights and multiply those with large weights. The introduction of this key step led to the first operational SMC method; see (Doucet et al., 2001) for details of implementation of this black-box routine.

4. Experiments

In this section, three experiments are carried out to evaluate quantitatively and qualitatively the proposed approach. The first experiment provides comparisons to other control policies on a synthetic dataset. The second experiment, on a similar synthetic dataset, demonstrates how the approach can handle large variations in scale, occlusion and multiple targets. The final experiment is a demonstration of tracking and classification performance on several real videos. For the synthetic digit videos, we trained the first-layer RBMs on the foveated images, while for the real videos we trained factored-RBMs on foveated natural image patches (Ranzato & Hinton, 2010).

The first experiment uses 10 video sequences (one for each digit) built from the MNIST dataset. Each sequence contains a moving digit and static digits in the background (to create distractions). The objective is to track and recognize the moving digit; see Figure 4. The gaze template had $K = 9$ gaze positions, chosen so that gaze G5 was at the center. The location of the template was initialized with optical flow.

We compare the policy learning algorithm against algorithms with deterministic and random policies. The deterministic policy chooses each gaze in sequence and in a particular pre-specified order, whereas the random policy selects a gaze uniformly at random. We adopted the Bhattacharyya distance in the specification of the observation model. A multi-fixation RBM was trained to map the first layer hidden units of three time consecutive time steps into a second hidden layer, and trained a logistic regressor to further map to the 10 digit classes. We used the transition prior as proposal for the particle filter.

Tables 1 and 2 report the comparison results. Track-

Table 1. Tracking error (in pixels) on several video sequences using different policies for gaze selection.

	0	1	2	3	4	5	6	7	8	9	Avg.
Learned policy	**1.2** (1.2)	**3.0** (2.0)	**2.9** (1.0)	**2.2** (0.7)	**1.0** (1.9)	**1.8** (1.9)	**3.8** (1.0)	**3.8** (1.5)	**1.5** (1.7)	**3.8** (2.8)	**2.5** (1.6)
Deterministic policy	18.2 (29.6)	536.9 (395.6)	104.4 (69.7)	2.9 (2.2)	201.3 (113.4)	4.6 (4.0)	5.6 (3.1)	64.4 (45.3)	142.0 (198.8)	144.6 (157.7)	122.5 (101.9)
Random policy	41.5 (54.0)	410.7 (329.4)	3.2 (2.0)	3.3 (2.4)	42.8 (60.9)	6.5 (9.6)	5.7 (3.2)	80.7 (48.6)	38.9 (50.6)	225.2 (241.6)	85.9 (80.2)

Table 2. Classification accuracy on several video sequences using different policies for gaze selection.

	0	1	2	3	4	5	6	7	8	9	Avg.
Learned policy	95.62%	**100.00**%	**99.66**%	99.33%	**99.66**%	**100.00**%	**100.00**%	**98.32**%	**97.98**%	**89.56**%	**98.01**%
Deterministic policy	**99.33**%	**100.00**%	98.99%	94.95%	5.39%	98.32%	0.00%	29.63%	52.19%	0.00%	57.88%
Random policy	98.32%	**100.00**%	96.30%	**99.66**%	29.97%	96.30%	89.56%	22.90%	12.79%	13.80%	65.96%

ing accuracy was measured in terms of the mean and standard deviation (in brackets) over time of the distance between the target ground truth and the estimate; measured in pixels. The analysis highlights that the error of the learned policy is always below the error of the other policies. In most of the experiments, the tracker fails when an occlusion occurs for the deterministic and the random policies, while the learned policy is successful. This is very clear in the videos at: `http://www.youtube.com/user/anonymousTrack`

The loss of track for the simple policies is mirrored by the high variance results in Table 1 (experiments 0, 1, 4, and so on). The average mean and standard deviations (last column of Table 1) make it clear that the proposed strategy for learning a gaze policy can be of enormous benefit. The improvements in tracking performance are mirrored by improvements in classification performance (Table 2).

Figure 4 provides further anecdotal evidence for the policy learning algorithm. The top sequence shows the target and the particle filter estimate of its location over time. The middle sequence illustrates how the policy changes over time. In particular, it demonstrates that hedge can effectively learn where to look in order to improve tracking performance (we chose this simple example as in this case it is obvious that the center of the eight (G5) is the most reliable gaze action). The classification results over time are shown in the third row.

The second experiment addresses a similar video sequence, but tracking multiple targets. The image scale of each target changes significantly over time, so the algorithm has to be invariant with respect to these scale transformations. In this case, we used a mixture proposal distribution consisting of motion detectors and the transition prior. We also tested a saliency proposal but found it to be less effective than the motion detectors for this dataset. Figure 5 (top) shows some of the video frames and tracks. The videos allow one to better appreciate the performance of the multi-target tracking algorithm in the presence of occlusions. Tracking and classification results for the real videos are shown in Figure 5 and the accompanying videos.

5. Conclusions and future work

We have proposed a decision-theoretic probabilistic graphical model for joint classification, tracking and planning. The experiments demonstrate the significant potential of this approach. There are many routes for further exploration. In this work we pre-trained the (factored)-RBMs. However, existing particle filtering and stochastic optimization algorithms could be used to train the RBMs online. Following the same methodology, we should also be able to adapt and improve the target templates and proposal distributions over time. This is essential to extend the results to long video sequences where the object undergoes significant transformations.

Deployment to more complex video sequences will require more careful and thoughtful design of the proposal distributions, transition distributions, control algorithms, continuous template models, data-association and motion analysis modules. Fortunately, many of the solutions to these problems have already been engineered in the computer vision, tracking and online learning communities. Admittedly, much work remains to be done.

Saliency maps are ubiquitous in visual attention studies. Here, we simply used standard saliency tools and motion flow in the construction of the proposal distributions for particle filtering. There might be better ways to exploit the saliency maps, as neurophysiological experiments seem to suggest (Gottlieb et al., 1998).

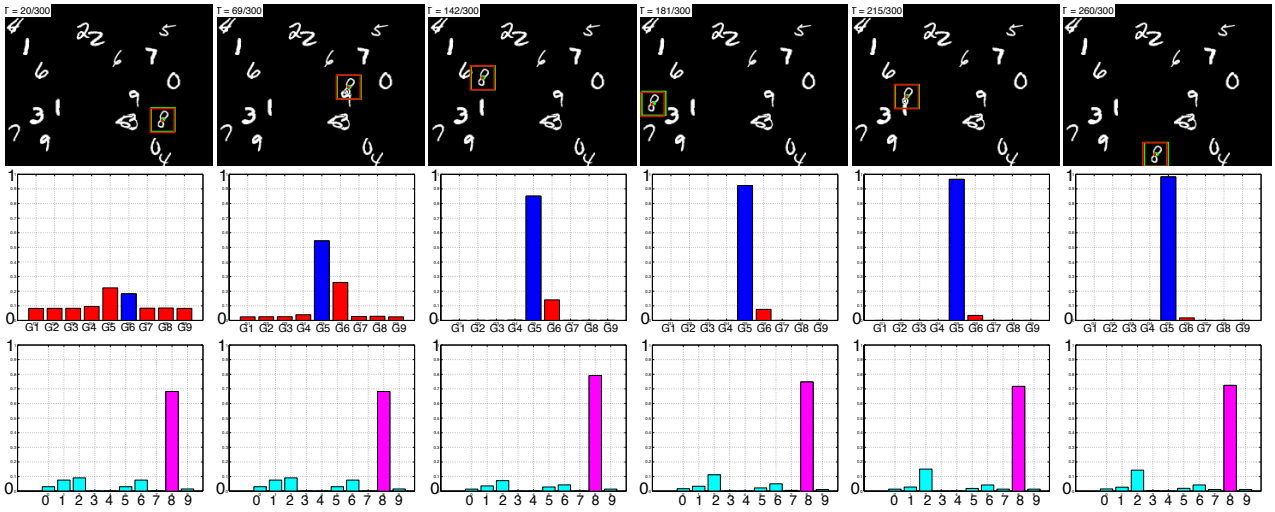

Figure 4. Tracking and classification accuracy results with the learned policy. First row: position of the target and estimate over time. Second row: policy distribution over the 9 gazes; hedge clearly converges to the most reasonable policy. Third row: cumulative class distribution for recognition.

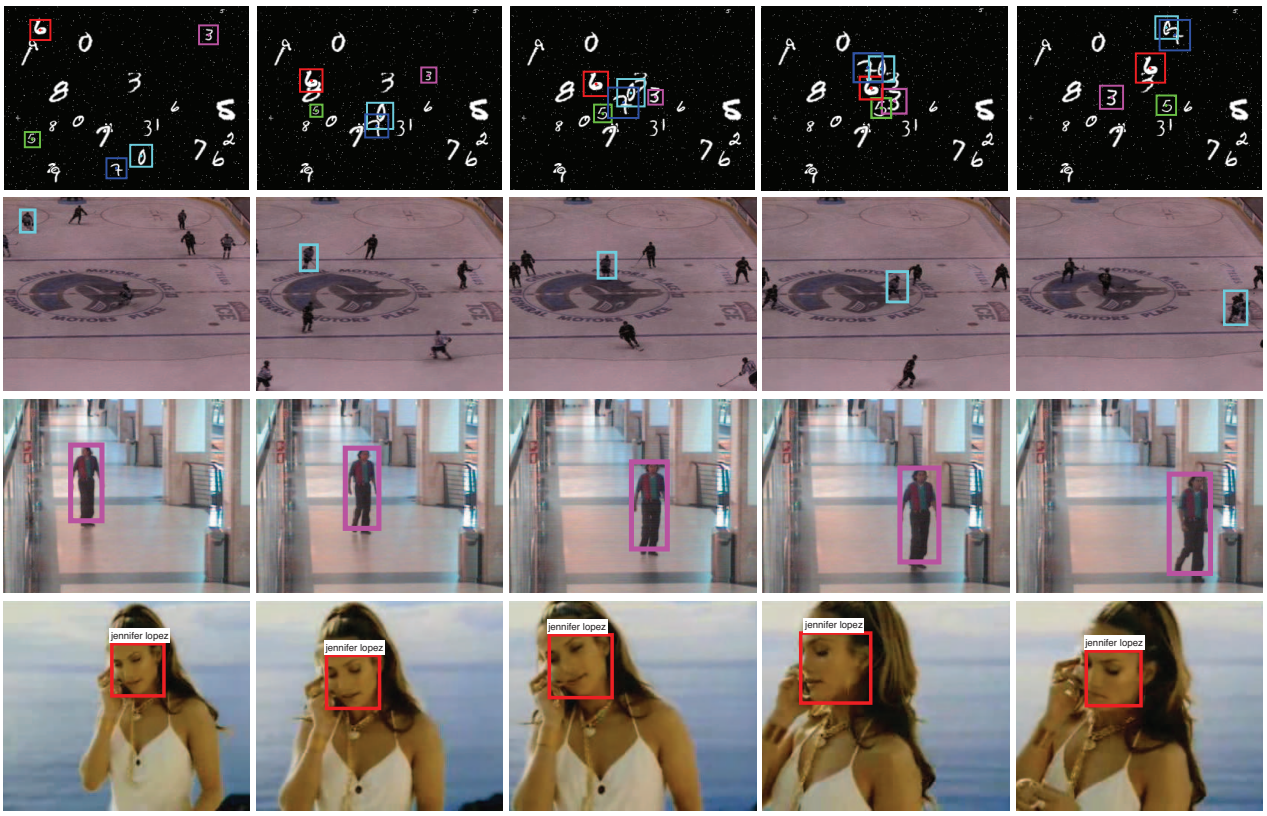

Figure 5. (Top) Multi-target tracking with occlusions and changes in scale on a synthetic video. (Middle and bottom) Tracking in real video sequences.

One of the most interesting avenues for future work is the construction of more abstract attentional strategies. In this work, we focused on attending to regions of the visual field, but clearly one could attend to subsets of receptive fields or objects in the deep appearance model.

Acknowledgments

We thank Ben Marlin, Kenji Okuma, Marc'Aurelio Ranzato and Kevin Swersky. This work was supported by CIFAR's NCAP program and NSERC.

References

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. Gambling in a rigged casino: the adversarial multi-armed bandit problem. Technical Report NC2-TR-1998-025, 1998.

Brochu, Eric, Cora, Vlad M, and de Freitas, Nando. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-023, CS dept, UBC, 2009.

Cesa-Bianchi, Nicolo and Lugosi, Gabor. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

Chaudhuri, Kamalika, Freund, Yoav, and Hsu, Daniel. A parameter-free hedging algorithm. In *NIPS*, 2009.

Colombo, John. The development of visual attention in infancy. *Annual Review of Psychology*, pp. 337–367, 2001.

Doucet, A, de Freitas, N, and Gordon, N. Introduction to sequential Monte Carlo methods. In Doucet, A, de Freitas, N, and Gordon, N J (eds.), *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55: 119–139, 1997.

Girard, B. and Berthoz, A. From brainstem to cortex: Computational models of saccade generation circuitry. *Progress in Neurobiology*, 77(4):215 – 251, 2005.

Gottlieb, Jacqueline P., Kusunoki, Makoto, and Goldberg, Michael E. The representation of visual salience in monkey parietal cortex. *Nature*, 391:481–484, 1998.

Hinton, GE and Salakhutdinov, RR. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Isard, M and Blake, A. Contour tracking by stochastic propagation of conditional density. In *European Computer Vision Conference*, pp. 343–356, 1996.

Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254 –1259, 1998.

Kavukcuoglu, K., Ranzato, M.A., Fergus, R., and Le-Cun, Yann. Learning invariant features through topographic filter maps. In *Computer Vision and Pattern Recognition*, pp. 1605–1612, 2009.

Köster, Urs and Hyvärinen, Aapo. A two-layer ICA-like model estimated by score matching. In *International Conference of Artificial Neural Networks*, pp. 798–807, 2007.

Larochelle, Hugo and Hinton, Geoffrey. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Neural Information Processing Systems*, 2010.

Lee, H., Grosse, R., Ranganath, R., and Ng, A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, 2009.

McNaughton, Bruce L., Battaglia, Francesco P., Jensen, Ole, Moser, Edvard I., and Moser, May-Britt. Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

Okuma, Kenji, Taleghani, Ali, de Freitas, Nando, and Lowe, David G. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Olshausen, Bruno A., Anderson, Charles H., and Essen, David C. Van. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13:4700–4719, 1993.

Postma, Eric O., van den Herik, H. Jaap, and Hudson, Patrick T. W. SCAN: A scalable model of attentional selection. *Neural Networks*, 10(6):993 – 1015, 1997.

Ranzato, M.A. and Hinton, G.E. Modeling pixel means and covariances using factorized third-order Boltzmann machines. In *Computer Vision and Pattern Recognition*, pp. 2551–2558, 2010.

Rensink, Ronald A. The dynamic representation of scenes. *Visual Cognition*, pp. 17–42, 2000.

Rosa, M.G.P. Visual maps in the adult primate cerebral cortex: Some implications for brain development and evolution. *Brazilian Journal of Medical and Biological Research*, 35:1485 – 1498, 2002.

Swersky, K., Chen, Bo, Marlin, B., and de Freitas, N. A tutorial on stochastic approximation algorithms for training restricted Boltzmann machines and deep belief nets. In *ITA Workshop*, pp. 1–10, 2010.

Taylor, G.W., Sigal, L., Fleet, D.J., and Hinton, G.E. Dynamical binary latent variable models for 3D human pose tracking. In *Computer Vision and Pattern Recognition*, pp. 631–638, 2010.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103, 2008.

Welling, M., Rosen-Zvi, M., and Hinton, G. Exponential family harmoniums with an application to information retrieval. *Neural Information Processing Systems*, 17: 1481–1488, 2005.