MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Learning concise models of human activity from ambient video via a structure-inducing M-step estimator

Matthew Brand

TR-97-25 November 1997

Abstract

We introduce a method for structure discovery in data and use it to learn a normative theory about the behavior of the visual world from coarse image representations. The theory takes the form of a concise probabilistic automaton—specifically, a continuous-output hidden Markov model (HMM)—but the induction method applies generally to any conditional probability model. The learning algorithm introduces and exploits an entropic prior for fast, simultaneous estimation of model structure and parameters. Although not motivated as such, the prior and its maximum *a posteriori* (MAP) estimator can be understood as an exact formulation of minimum description length (MDL) for Bayesian point estimation; we present an exact solution for the MAP estimator which thus folds MDL into the M-step of expectation-maximization (EM) algorithms. Consequently there is no speculative or wasted computation as in search-based MDL approaches. In contrast to conventionally trained HMMs, entropically trained models are so concise and highly structured that they are interpretable, and can be automatically converted into a flowchart and/or a map of characteristic activities (motion patterns) in the field of view. In this paper we examine the model formed by the system from roughly a half-hour of video of office activity, then demonstrate its ability to detect unusual behavior.

Publication History:–

1. First printing, TR-97-25, November 1997

Learning concise models of human activity from ambient video via a structure-inducing M-step estimator

Matthew Brand

18nov97

Abstract

We introduce a method for structure discovery in data and use it to learn a normative theory about the behavior of the visual world from coarse image representations. The theory takes the form of a concise probabilistic automaton—specifically, a continuous-output hidden Markov model (HMM)—but the induction method applies generally to any conditional probability model. The learning algorithm introduces and exploits an entropic prior for fast, simultaneous estimation of model structure and parameters. Although not motivated as such, the prior and its maximum *a posteriori* (MAP) estimator can be understood as an exact formulation of minimum description length (MDL) for Bayesian point estimation; we present an exact solution for the MAP estimator which thus folds MDL into the M-step of expectation-maximization (EM) algorithms. Consequently there is no speculative or wasted computation as in search-based MDL approaches. In contrast to conventionally trained HMMs, entropically trained models are so concise and highly structured that they are interpretable, and can be automatically converted into a flowchart and/or a map of characteristic activities (motion patterns) in the field of view. In this paper we examine the model formed by the system from roughly a half-hour of video of office activity, then demonstrate its ability to detect unusual behavior.

1 Introduction

How visual entities behave is often more important than how they appear; from flies to humans there are many examples of natural vision systems computing the former without computing much of the latter. Behavioral descriptions support many key inferences; in some cases they provide a better basis for recognition than appearance. Here we consider learning a model of human behavior from medium- to long-term ambient video. We shall take a pattern-discovery approach; by behavior, we mean nothing more than spatio-temporal patterns in the motion, pose, and position of the observed person. Desiderata for such a model include: It should partition the visual data stream into coherent activities; it should allow the detection of anomalous behaviors; and it should be computationally lightweight. We find we can meet these criteria

with a new algorithm that induces low-entropy probabilistic automata from time-series of coarse image representations. These models are concise and interpretable, which strongly contrasts with probabilistic models typically obtained from polynomial-time algorithms (e.g., expectation-maximization),

The key to this result is a new M-step estimator for expectation-maximization (EM) algorithms which effects simultaneous structure and parameter learning in conditional probability models. Applied to hidden Markov models (HMM), the algorithm finds a concise representation of the hidden structure of a signal by trimming uninformative edges from the state transition graph and/or removing entire states. The basis of the new M-step is an entropic prior on parameter values and a solution for the maximum *a posteriori* (MAP) estimator. As we show below, the MAP estimate minimizes both the entropy of the model and its cross-entropy with the data's sufficient statistics, whereas maximum likelihood (ML) methods only minimize the latter. Recursive estimation tends to extinguish uninformative parameters, which can then be trimmed from the model without loss of posterior probability. By recursively simplifying a randomly initialized model we can induce the structure of relations between hidden variables. We call this whole process *entropic estimation*.

Entropic estimation sparsifies the conditional probability table, yielding a concise and computationally lightweight model. In practice, surviving states tend to be highly correlated with meaningful partitions of the data, while surviving transitions provide a nearly minimal perplexity model of the signal dynamics.

2 Related work

Vision: There is wide interest in learning normative models of activity from vision, but the literature on learning over time spans greater than a few seconds is sparse. Brand has compiled a normative model of animate/inanimate object interactions into a finite-state machine, then learned the visual correlates of this model via couplings of HMMs, producing parses of video on the scale of minutes [2]. Hogg et al. have shown how to learn characteristic motion maps for pedestrian plazas, which are themselves representations of non-parametric distributions over collections of pedestrian trajectories on the scale of hours [4, 6].

HMMs: The literature of structure-learning in HMMs is, to date, based entirely on generate-and-test algorithms. These algorithms work by selecting a single state to be merged [11] or split [12, 5], then retraining the model to see if any advantage has been gained. Though these efforts use a variety of heuristic techniques and priors (including MDL) to avoid failures, much of the computation is squandered and reported run-times range from hours to days. Here we develop an EM structure-learning algorithm that converges in seconds.

MDL: The entropic prior is simply related to the lower bound of the coding length of the model (see eqn. 2), and thus entropic estimation is closely connected to minimum description length methods and stochastic complexity (SC) [14, 8, 9]. Of recent attempts to find or approximate MDL estimators, two stand out: Recently, Yamanishi developed a general Monte Carlo approximation to computing the SC—one of the first that is not heavily biased by the user's choice of encoding scheme [16]. Vovk

has derived a computable MDL estimator for single-parameter probability models over discrete sample spaces [13]. Here we present an exact estimator for multi-parameter models for continuous sample spaces. Our formulation provides a unified Bayesian framework for two issues that are often treated separately in the MDL literature: 1) estimating the number of parameters, and 2) estimating their values.

3 An entropic prior

In entropic estimation we want to move parameter values as far as possible from their initial random values. Parameters at chance add virtually no information to the model, and are therefore wasted degrees of freedom. In contrast, parameters near the extrema $\{0, 1\}$ are informative because they impose strong constraints on the class of signals accepted by the model. In Bayesian terms, we desire a prior that asserts that parameters that do not reduce uncertainty are improbable. We can capture this intuition in a surprisingly simple form:

$$P_e(\theta_i) \propto \theta_i^{\theta_i} \tag{1}$$

In a complete model of N conditional probabilities $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_N\}$ we write

$$P_e(\boldsymbol{\theta}) \propto \boldsymbol{\theta}^{\boldsymbol{\theta}} = \prod_i \theta_i^{\theta_i} = \exp\left[\sum_i \theta_i \log \theta_i\right] = e^{-H(\boldsymbol{\theta})} \tag{2}$$

whence we can see that the prior measures how free the model is from ambiguity. The bolded convex curve in figure 1 shows how this prior is averse to chance values. Combining $P_e(\cdot)$ with the multinomial yields the biased entropic prior:

$$P_e(\boldsymbol{\theta}|\boldsymbol{\omega}) \propto \left[\prod_i \theta_i^{\omega_i}\right] \frac{P_e(\boldsymbol{\theta})}{P(\boldsymbol{\omega})} \propto \prod_i \theta_i^{\theta_i + \omega_i} \tag{3}$$

where ω_i is a bias for event type i.

This prior is obviously conjugate to the multinomial, so we may also consider $\boldsymbol{\omega}$ to be evidence, in which case the posterior takes the same form as eqn. 3. As figure 1 shows, with scant evidence this distribution skews to stronger odds, but with increasing evidence it converges to "fair" odds for $\boldsymbol{\omega}$, and is thus consistent. Note that this is the opposite behavior that one obtains from a Dirichlet prior, which skews to weaker odds when data is scarce.

3.1 MAP estimator

To obtain MAP estimates we set the derivative of log-likelihood to zero, using Lagrange multipliers to ensure $\sum_i \theta_i = 1$,

$$0 \quad = \quad \frac{\partial}{\partial \theta_i}\left(\log \prod_i \theta_i^{\omega_i + \theta_i} + \lambda \sum_i \theta_i\right) \tag{4}$$

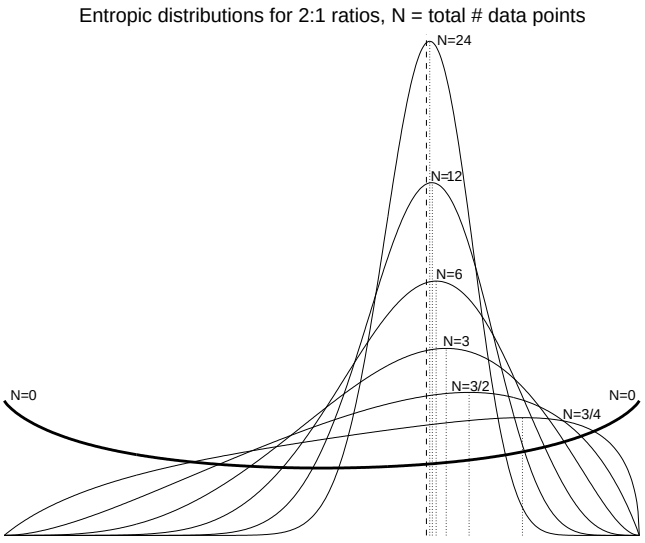

Figure 1: Entropic distributions for binomial parameter $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ s.t. $\theta_1 + \theta_2 = 1$, where evidence $\boldsymbol{\omega} = \{\omega_1, \omega_2\}$ has been generated in the ratio $\omega_1 = 2\omega_2$ and $N = \omega_1 + \omega_2$ is the total observed evidence for each distribution (plotted curve). The bolded convex curve $\exp(-H(\boldsymbol{\theta}))$ shows how extremal values are preferred in the absence of evidence ($N = 0$). Dotted verticals show the MAP estimates, extremal at $N = 0$ but converging to the ML estimate (dashed line at $\theta_1 = 2/3$) as $N \to \infty$.

$$
\begin{aligned}
&= \sum_i \frac{\partial}{\partial \theta_i} (\omega_i + \theta_i) \log \theta_i + \lambda \sum_i \frac{\partial}{\partial \theta_i} \theta_i \\
&= 1 + \frac{\omega_i}{\theta_i} + \log \theta_i + \lambda \qquad (5)
\end{aligned}
$$

We obtain θ_i by working backward from the Lambert W function, a multivalued inverse function satisfying $W(x) e^{W(x)} = x$. Taking logarithms and setting $y = \log x$,

$$
\begin{aligned}
0 &= -W(x) - \log W(x) + \log x \qquad (6) \\
&= -W(e^y) - \log W(e^y) + y \\
&= \frac{-1}{1/W(e^y)} + \log 1/W(e^y) + \log z + y - \log z \\
&= \frac{-z}{z/W(e^y)} + \log z/W(e^y) + y - \log z \qquad (7)
\end{aligned}
$$

Setting $\theta_i = z/W(e^y)$, $y = 1 + \lambda + \log z$, and $z = -\omega_i$, eqn. 7 simplifies to eqn. 5, implying that

$$
\hat{\theta}_i = \frac{-\omega_i}{W(-\omega_i e^{1+\lambda})} \qquad (8)
$$

Equations 5 and 8 yield a fast iterative procedure for the entropic MAP estimate: Calculate $\boldsymbol{\theta}$ given λ, normalize $\boldsymbol{\theta}$, calculate λ given $\boldsymbol{\theta}$, repeat. This typically converges in 2-4 iterations. Solutions lie in the W_{-1} branch of Lambert's function. (In practice, some additional algebra is needed to handle intermediate values for which $W(-e^{-x})$ has no real branch or where machine precision is exhausted.)

3.2 Interpretation

Entropy: Some manipulation of the negative log-posterior (which is minimized) allows us to understand the MAP estimate in terms of entropy:

$$
\begin{aligned}
-\log \prod_i \theta_i^{\theta_i+\omega_i} &= -\sum_i (\theta_i+\omega_i)\log\theta_i \qquad (9)\\
&= -\sum_i (\theta_i \log\theta_i + \omega_i\log\theta_i - \omega_i\log\omega_i + \omega_i\log\omega_i)\\
&= -\sum_i \theta_i\log\theta_i + \sum_i \omega_i \log\frac{\omega_i}{\theta_i} - \sum_i \omega_i\log\omega_i\\
&= H(\boldsymbol{\theta}) + D(\boldsymbol{\omega}||\boldsymbol{\theta}) + H(\boldsymbol{\omega}) \qquad (10)
\end{aligned}
$$

When $H(\boldsymbol{\omega})$ is fixed, the MAP estimate minimizes the sum of the parameter entropy $H(\boldsymbol{\theta})$ and the cross-entropy $D(\boldsymbol{\omega}||\boldsymbol{\theta})$ between the parameters $\boldsymbol{\theta}$and the data's sufficient statistics $\boldsymbol{\omega}$. Equivalently, it minimizes coding length. As we will see in §4.1, we can also simplify the structure of the model, such that the data's sufficient statistics change and $H(\boldsymbol{\omega})$ declines as well.

An extended discussion of the meaning of the multinomial MAP estimator, its use in other probabilistic models, and its relation to problems in graph theory can be found in [3]. We have also derived minimum-entropy MAP estimators for covariance, mean, and weight parameters; a forthcoming paper describes applications to mixtures-of-Gaussians, radial basis functions, neural networks, and other popular models.

4 Use in entropic HMM training

In entropic estimation of HMM transition probabilities, we follow the conventional E-step, calculating the probability mass for each transition to be used as evidence $\boldsymbol{\omega}$:

$$
\gamma_{j,i} \quad = \quad \sum_t^{T-1} \alpha_j(t)\, P_{i|j}\, p(y_{t+1}|s_i)\, \beta_i(t+1) \qquad (11)
$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are obtained from forward-backward analysis and follow the notation of Rabiner [7]. $P_{i|j}$ is the current estimate of the probability that state i will follow state j. For the M-step, we calculate new estimates $\{\hat{P}_{i|j}\}_i = \boldsymbol{\theta}$ by applying the MAP estimator in §3.1 to each $\boldsymbol{\omega} = \{\gamma_{j,i}\}_i$. That is, $\boldsymbol{\omega}$ is a vector of the evidence for each kind of transition out of a single state; from this evidence the MAP estimator calculates probabilities $\boldsymbol{\theta}$. (In Baum-Welch reestimation, the maximum-likelihood estimator simply sets $\hat{P}_{i|j} = \gamma_{j,i} / \sum_i \gamma_{j,i}$.)

In recursive estimation (e.g., EM), the entropic estimator drives weakly supported parameters toward zero, concentrating evidence on surviving parameters until their estimates converge to near the ML estimate, at which point the algorithm terminates.

4.1 Model trimming

In [3] we show that HMM parameters remaining near zero can also be deleted with no loss of probability mass iff

$$P_{i|j} \quad \leq \quad \exp\left[-\sum_{t=1}^{T-1}\gamma_j(t)\right] \tag{12}$$

where $\gamma_j(t)$ is the probability of state j at time t. This is derived by balancing any loss in the likelihood with a gain in the prior. More generally, trimming is licensed for *any* probabilistic model with an entropic prior on θ_i when

$$\theta_i \frac{\partial}{\partial\theta_i} H(\boldsymbol{\theta}) \quad \geq \quad \theta \frac{\partial}{\partial\theta_i} \log P(\boldsymbol{D}\,|\boldsymbol{\theta}) \tag{13}$$

Trimming bumps the model out of a local probability maximum and allows further training in a lower-dimensional and possibly smoother parameter subspace. A similar test licenses state deletion. There is an interesting question as to how much state trimming is always desirable, since overfitting is generally due to excess parameters, not states. Perhaps this is why we find that entropic training naturally reserves some excess states for representing common subpaths in the transition graph; this is a form of compression that reduces the coding length and computational expense of the model. We call such states *gating* states as opposed to conventional *data-modeling* states because their output probabilities are near-zero almost everwhere and typically do not need to be computed. In the example developed below, state number 5 is a particularly good example of a gating state.

We have observed that entropic training has a number of interesting properties: (1) Smaller transition probabilities are driven toward zero, at which point the entire transition can be deleted from the model, reducing compute time and ambiguity. (2) Entropically trained HMMs tend to generalize better to held out test data than conventionally trained HMMs and also classify more accurately. (3) State output distributions tend to have slightly tighter covariances and states are more clearly identified with regions of the signal. (4) Entropically trained HMMs tend to attain the same low perplexity regardless of initial conditions, while the perplexity of conventionally trained HMMs is a function of their initial state count. Properties (1) and (2) can be proven; properties (3) and (4) have merely been observed in trials with datasets drawn from vision, genetics, handwriting, and speech, and will be demonstrated in the second half of this paper.

5 Learning a model of office activity

HMMs are the probabilistic model of choice for modeling signals from humans, principally because they are robust to variations in the timing and sequencing of signal struc-

tures, and they make optimal use of contextual information in time. They are essentially nondeterministic finite-state automata with probabilistic outputs. Although there is a strong conjecture that human signals such as language have recursive (context-free or -sensitive) structure, it is also widely admitted that the recursion depth is quite finite and therefore finite-state machines are probably adequate models for many tasks.

Entropically estimation can remove excess parameters and therefore show some resistance to overfitting and improved generalization; this we will demonstrate in figure 8. However, because of their facility for discovering concise structural models, we are more interested in using entropically trained HMMs to learn the structure of longer term behavior in visual domains such as traffic intersections, factory floors, animal colonies, etc. Office activity is a particularly good test because of the challenging range of time spans: Fast events such as answering the phone may take a few seconds while slow activities such as writing take hours. Here we demonstrate that much of this structure can be discovered via entropic estimation from lightweight, coarse visual tracking data.

5.1 Image representation

Continuous-output HMMs require a reasonably short observation vector which represents the content of each image. Different image representations will lead to models that emphasize different coherencies in the data. We experimented with two kinds of observation vectors: a "stripe" representation and a "blob" representation. Stripe data consists of mean location and extent of foreground pixels in a vertical or horizontal stripe across the image. Blob data consists of ellipse parameters fitting the single largest connected set of foreground pixels in the image.

In both cases foreground pixels are identified with reference to an acquired statistical model of the background texture and camera noise. The foreground consists of pixels that change substantially, ostensibly due to motion. These are modeled via multivariate gaussian distributions over color and location, and are re-estimated in each frame. Pixels are sorted into foreground or background by likelihood ratio; morphological dilation connects the foreground pixels using a seed from the previous frame [15]. For stripe data 5-10 stripes were used in each direction; the observation vector consisted of [*mean, extent,* Δ*mean,* Δ*extent*] for each stripe. For blob data a single bivariate gaussian (ellipse) was fitted to the foreground pixels; the observation vector consisted of [$mean_x$, $mean_y$, $\Delta mean_x$, $\Delta mean_y$, *mass,* Δ*mass, elongation, eccentricity*]. One of the goals of training is to tune model states to interesting regions in this signal; for the rest of the paper we will refer to this tuning interchangeably as "output distribution" and "receptive field."

Note that we are making some important simplifications of the task: We are interested in the behavior of a single person in a relatively stable environment, hence our choice of office work. Highly dynamic environments will "break" the vision front end. However, we did contend with variation in the form of moving shadows, natural light from the window, uncontrolled light from nearby cubicles, moved objects and furniture in the room, and pedestrians visible in the hallway. The image processing is robust to these variations in lighting insofar as they fit the learned background noise model. Similarly, motion in the background does not register if it is not visually connected to

the person being tracked; a second person entering the room is ignored unless there is a contact or occlusion with the person being tracked. Arguably, such visual events are highly informative and their inclusion in the feature vector is a good thing.

5.2 Training

Approximately 30 minutes of data were taken at 5Hz from an SGI IndyCam. Data was collected automatically and at random over several days by a program that started recording whenever someone entered the room after it had been empty 5+ minutes. Backgrounds were re-learned during these absences to accommodate changes in lighting and room configuration. After automatic deletion of blank frames (when the subject exits the room and field of view), roughly 21 minutes of training data remained.

Three sequences ranging from 1000 to 1900 frames in length were used for entropic training of 12, 16, 20, 25, and 30-state HMMs. States were initialized to tile the image with their receptive fields. Transition probablities were initialized to prefer motion to adjoining tiles; first-state probabilities were set to zero for non-edge states. It was found that variation in the initial receptive fields or state counts made little difference in the gross structure or performance of the final model. Training took six seconds on an SGI R10000 running Matlab.

As might be expected, models built on blob data have receptive fields tuned primarily to location, motion, and gross shape; models built on stripe data are more sensitive to body articulations (e.g., having one's arm out to write on a whiteboard or pick up a phone), and less attuned to gross shape and attitude. In both cases the results are similar; we will concentrate on the blob data since the results lend themselves to clearer visualizations.

5.3 Results

Entropic training yielded a substantially sparsified transition matrix (figure 2) which is easily converted into a human-readable representation of characteristic office activity. Figure 3 shows the corresponding state machine, or flowchart. The graph is automatically generated from the transition matrix; grouping of states into activities was done by adaptive clustering on a proximity matrix which combined Mahalonobis distance and transition probability between states. (Clustering is purely for improving the layout and readability of the graph, and has no algorithmic value—even without the clustering the graph is intelligible.) Labels were added by the author after training and clustering.

Some states deserve special explanation: State 5 is a gating state that does not model data but simplifies paths between other states; state 7 responds mainly to elongation and represents getting up and sitting down; state 10 represents staring at the screen; state 9 represents looking down to and up from the keyboard. Note that most of the transitions were trimmed, and many of the transitions are reversible, since many office activities have symmetric transitions, e.g., going to and from the whiteboard. Figure 4 shows how the states map onto regions in the field of view. Figure 5 show some frames from a non-training sequence to which specific states are strongly tuned. However, since several states respond mainly to characteristic velocities rather than to pose or position, these images are not entirely representative.

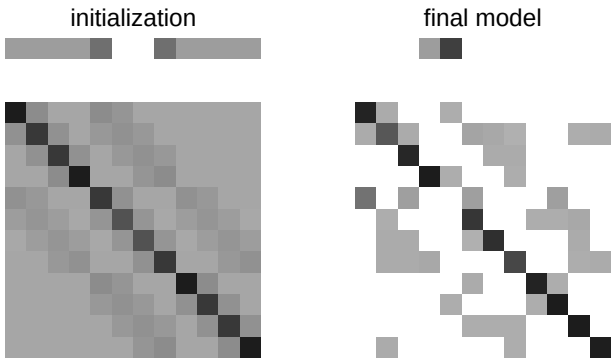

Figure 2: Transition matrix before and after entropic training. The top row indicates prior probabilities of each state; each subsequent row indicates the transition probabilities out of a state. Color key: $\circ = 0$; $\bullet \rightarrow 1$.

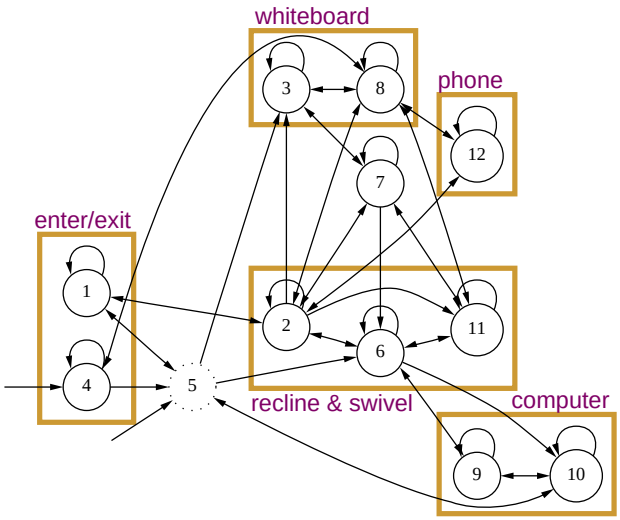

Figure 3: An activity graph generated by entropic training.

Even though 12 states is probably suboptimal, one state was reserved for gating rather than data-modeling. Entropic training with larger initial state counts resulted in even sparser models with similar qualitative structure (figure 6). Conventional training, of course, does not produce sparse or interpretable models (see figure 7).

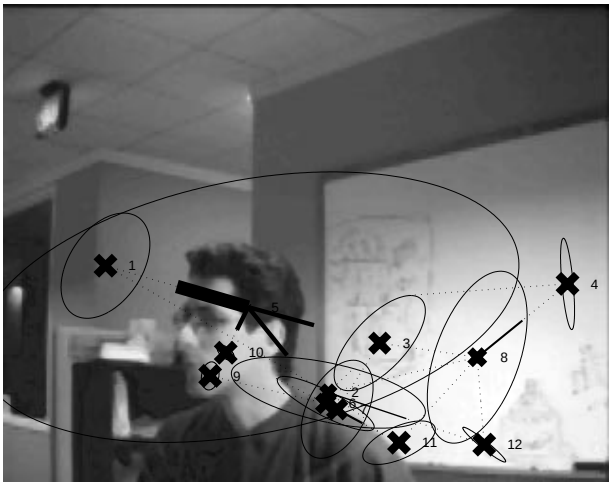

Figure 4: Locative receptive fields for the states in figure 3, projected onto a typical image. The $times$s and ellipses show per-state means and covariances for the centroid of the visual blob. (Ellipses *do not* show blob extent, shape, or velocity). Thickness of $times$s indicate dwell probabilities; thickness of arc-lines indicate transition probabilities.

5.4 Anomaly detection

To study the significance of the entropically trained parameters, we compared the ability of entropically trained and conventionally trained HMMs to detect anomalous data—in this domain, unusual behavior. Four data sets were used: (a) training data; (b) held out test data; (c) reversed held out test data; (d) data taken after the subject had consumed 4 cups of espresso. These data sets differ principally in the ordering, rhythm, and timing of actions, and therefore emphasize the discriminative power of the transition parameters. (The coffee sequence was originally part of the held-out data but it was found to have an unusually low probability; later, consulting the recording logs, we found out why.) There were three test conditions: (1) entropically estimated parameters; (2) conventionally estimated parameters; (3) transition parameters flattened to chance. Condition (3) tests whether the transitions or output parameters are responsible for the model's selectivity. Figure 8 shows that the entropic HMM has a smaller train-test divergence (e.g., better generalization) and was most successful in distinguishing abnormal behavior (backwards and jittery). The performance of the flattened model shows that little of that selectivity is due just to the output parameters.

This addresses a common criticism of continuous-output HMMs—that model selectivity is determined mainly by model structure, secondly by output distributions, and only lastly by transition probabilities, because they have the smallest dynamic range [1]. (Historically some users have found structure so selective that parameter values can be ignored, e.g., [10]). Entropic estimation makes transition parameters first-class citizens by expanding their dynamic range—to infinity, in fact, if one considers

Figure 5: Sample frames assigned high state-specific probabilities by the model.

trimmed parameters. Another advantage of blurring the distinction between parameter values and model structure is that as the parameter matrix sparsifies, credit assignment in learning becomes exponentially less diffuse and more effective [1].

6 Conclusion

We have shown how a computer can form concise theories of behavior from gigabyte-scale video streams using very coarse representations of change in the image. Given a half-hour of ambient video of office activity, the system generates a probabilistic model of normative activity, a readable flow-chart of work activities, and a map of significant

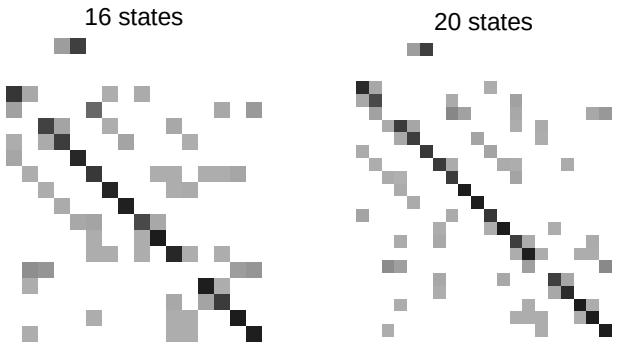

Figure 6: Entropic training with more states results in increasingly sparse models with similar structure.

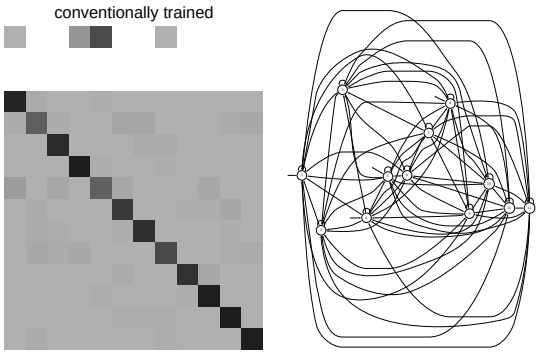

Figure 7: Conventional training fails to discover a structured model, though hints of the the entropic model are faintly visible. The equivalent flowchart is shown at right. Compare with figures 2 and 3

events and processes in the field of view. In contrast to what one gets from conventional estimation methods, the learned model is so concise and sparse that it is interpretable as a theory of the signal.

The key to this result is an algorithm that exploits an entropic prior to do simultaneous structure and parameter estimation for conditional probability models. The expectation-maximization algorithm simultaneously minimizes the entropy of the model and its cross-entropy with the data; can escape local probability maxima through model simplifications; is monotonic; and converges in seconds. As a result, the entire system can learn or monitor the behavior of its environment using less than one-fourth of a modern workstation's compute power.

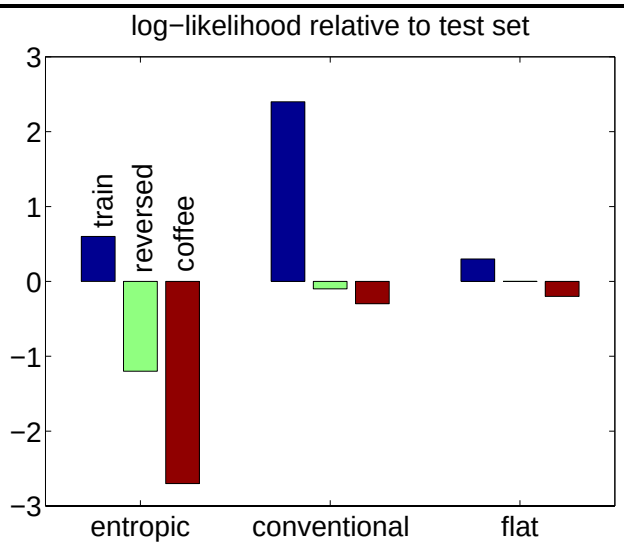

Figure 8: Model log-likelihoods normalized to sequence length. The entropic model is by far the most discriminative. Likelihoods are plotted relative to the test set because that is the *de facto* standard of normal activity.

Acknowledgements

Flowcharts were generated by the program `dot`, provided courtesy of AT&T Research.

References

[1] Yoshua Bengio and Paulo Frasconi. Diffusion of context and credit information in Markovian models. *Journal of AI Research*, 3:249–270, 1995.

[2] Matthew Brand. The "Inverse Hollywood Problem": From video to scripts and storyboards via causal analysis. In *Proceedings, Conference on Artificial Intelligence, AAAI*, 1997. Also available as MIT Media Lab Vision and Modeling TR #410.

[3] Matthew Brand. Structure discovery in hidden Markov models via an entropic prior and parameter extinction. Technical report, Mitsubishi Electric Research Labs, October 1997. Submitted to Neural Computation.

[4] J.H. Fernyhough, A.G. Cohn, and D.C. Hogg. Generation of semantic regions from image sequences. In *ECCV96*, pages II:475–484, 1996.

[5] Shiro Ikeda. Construction of phoneme models — Model search of hidden Markov models. In *International Workshop on Intelligent Signal Processing and Communication Systems*, Sendai, October 1993.

[6] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *IVC*, 14(8):609–615, August 1996.

[7] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[8] Jorma Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[9] Jorma Rissanen. *Stochastic Complexity and Statistical Inquiry*. World Scientific, 1989.

[10] H. Sakoe and C. Chiba. Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume ASSP-26, pages 43–49, February 1978.

[11] Andreas Stolcke and Stephen Omohundro. Best-first model merging for hidden Markov model induction. Technical Report TR-94-003, International Computer Science Institute, 1947 Center St., Berkeley, CA, 94704, USA, April 1994.

[12] Jun-Ichi Takami and Shigeki Sagayama. Automatic generation of the hidden Markov model by successive state splitting on the contextual domain and the temporal domain. Technical Report SP91-88, IEICE, December 1991.

[13] Volodya G. Vovk. Minimum description length estimators under the optimal coding scheme. In Paul Vitányi, editor, *Proceedings, Computational Learning Theory / Europe*, pages 237–251. Springer-Verlag, 1995.

[14] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computing Journal*, 11:185–195, 1968.

[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. In *SPIE Proceedings*, volume 2615, 1995.

[16] Keiji Yamanishi. A randomized approximation of the MDL for stochastic models with hidden variables. In *Proceedings, Computational Learning Theory*, pages 99–109, 1996.