A Hybrid Graphical Model for Robust Feature Extraction from Video

A. Taylan Cemgil Wojciech Zajdel Ben J. A. Kröse
Intelligent Autonomous Systems, University of Amsterdam
{cemgil, wzajdel, krose}@science.uva.nl

Abstract

We consider a visual scene analysis scenario where objects (e.g. people, cars) pass through the viewing field of a static camera and need to be detected and segmented from the background. For this purpose, we introduce a hybrid dynamic Bayesian network and derive an Expectation propagation (EP) algorithm for robust estimation of object shapes and appearance statistics. We demonstrate the viability of the approximation on an object detection task from real videos, where objects' smooth shapes are segmented from the background. The model is readily extendible to multi-object multi-camera scenarios and can be coupled in a transparent and consistent way with a hierarchical model for object identification under uncertainty.

1. Introduction

Foreground/background classification of pixels is a crucial preprocessing step in many computer vision problems, such as object tracking and identification where one wishes to suppress background pixels to collect reliable statistics about object features of interest. In many applications, the key assumption is that the visual sensors and the background image are static and objects of interest are in motion (e.g. people walking in a room full of furniture). Transient deviations from "steady-state" pixels values are detected and associated with foreground objects. Unfortunately, in reality, a background scene is almost never entirely static due to illumination changes, shadows, reflections, fog, wind or camera jitter [11]. All these factors render simple heuristics such as thresholding pixel differences useless.

From a statistical viewpoint, background estimation can be viewed as a novelty detection problem, e.g. [3]. Either implicitly or explicitly, probabilistic approaches attempt to estimate a process for the background pixels and try to detect "surprises". However, mainly due to computational considerations, many methods (e.g. [3, 10]) ignore spatial dependencies among neighboring pixels and only take temporal correlations into account. More recently, spatial correlations among pixels are implicitly taken into account by using a fixed basis transform [11] or are explicitly modeled using Markov random fields [13, 9].

In this paper we describe a probabilistic model that takes the spatial and temporal correlations into account in a flexible way, without much additional computational cost. We describe our model as a hybrid dynamic Bayesian network (with latent discrete child and continuous parent nodes [7]). The model is generative in nature, and reflects our a-priori assumptions about the visual scene – such as smooth object/shadow regions and a quasi-static background process. Exact computation of required quantities is intractable and we derive an approximate Expectation Propagation (EP) [8] algorithm that works essentially in linear time in the number of pixels, regardless of the correlation structure among them. An attractive feature of the model is that it allows for learning important image regions or object shapes easily[6]. Moreover, background estimation can be readily coupled in a transparent and consistent way to subsequent processing.

2. Model

We will denote each RGB pixel of a video stream as $y_{k,t}$ where $k = (i, j)$ with i and j corresponding to the vertical and horizontal spatial indices and t denoting the time frame. To simplify notation we will treat k as a linear index and let $k = 1 \dots K$ with K being the number of pixels per frame. We will also use the boldface notation $\mathbf{y}_t$ to denote all the pixels at t'th frame.

Each pixel has a binary indicator $r_{k,t} \in \{\text{back} = -1, \text{fore} = 1\}$, that indicates whether $y_{k,t}$ is associated with background or foreground object. For modeling shadows, we introduce a similar and independent binary indicator, $c_{k,t}$ with domain $\{\text{no-shadow} = -1, \text{shadow} = 1\}$. Note, that this enumeration leaves a possibility that a foreground object can be under a shadow as well. We will denote the collection of indicators in the t'th frame as $\mathbf{r}_t$ and $\mathbf{c}_t$ and refer to them as *masks*.

Visual scenes that we are interested in contain smooth shadow and object regions. Therefore, the corresponding masks, when viewed as a binary images, will exhibit long

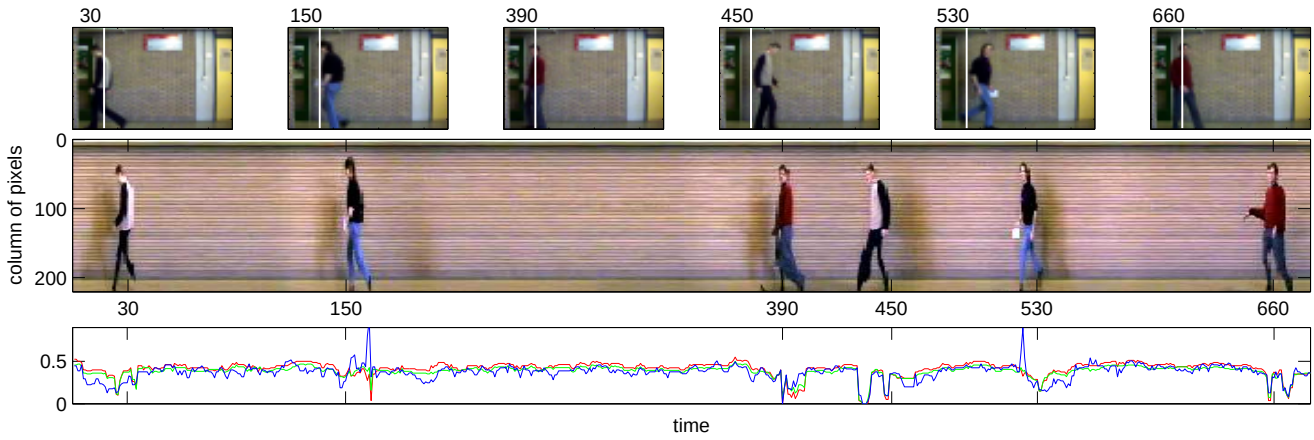

Figure 1. (Top), A few frames from an office scene. (Middle) A vertical slice through time, taken along the line shown on the frames. (Bottom) Variations in RGB values of a single pixel (at row 200). Occlusions and shadows can be clearly seen as immediate changes in intensity. Fluctuations in a single color channel illustrate the difficulty of detection when pixels are processed independently.

range correlations. A popular approach for modeling such correlations is by a Markov Random Field (MRF), where a positive coupling is assumed between adjacent pixels. Whilst a powerful and compact model, inference and especially learning in MRF's tend to be computationally expensive.

In this paper, we will investigate an alternative approach for introducing long range correlations on the binary masks. Our approach is based on the simple observation that signs of a collection of correlated Gaussian variables are also correlated. For an illustration of this idea, see Figure 2. We define a linear dynamical system (a Kalman filter model) on a collection of latent variables $x_{k,t}$, from which we obtain binary indicators $r_{k,t}$ by thresholding. Analogous models were proposed for unsupervised learning of static binary patterns [6], visualization [12] and for classification [14]. To our knowledge, however, these ideas were not employed in a dynamical visual scene analysis context.

2.1. Prior on masks

Linear dynamical systems are widely used state space models for continuous time series. In this model, data is assumed to be generated independently from a latent Markov process that obeys a single linear regime

$$\mathbf{s}_0 \sim \mathcal{N}(m, P) \qquad \mathbf{s}_t \sim \mathcal{N}(A\mathbf{s}_{t-1}, Q) \qquad \mathbf{x}_t \sim \mathcal{N}(W\mathbf{s}_t, R).$$

Here, m and P are prior mean and covariance, Q and R are diagonal covariance matrices and A and W are transition and observation matrices that describe the linear mappings between $\mathbf{s}_t$, $\mathbf{s}_{t-1}$ and $\mathbf{x}_t$. By integrating over $\mathbf{s}_{1:T}$, it is easy to see that this model induces on $\mathbf{x}_{1:T}$ a Gaussian distribution with a constrained (but in general full) covariance matrix. A tractable extension to this model is useful for modeling piecewise linear regimes with occasional regime switches [2]:

$$\begin{aligned} o_t &\sim p(o) \\ \mathbf{s}_t &\sim [o_t = \text{off}]\mathcal{N}(\mathbf{s}_{t-1}, Q) + [o_t = \text{on}]\mathcal{N}(m, P). \end{aligned}$$

Here, o_t is a binary variable that indicates a regime switch and [text] denotes an indicator, that evaluates to 1 (or 0) whenever the proposition "text" is true (or false). When $o_t = \text{on}$, we switch to a new regime by "reinitializing" the state vector from the prior.

To convert this model for real valued data into one for binary data, we use a clipping mechanism analogous as described in [6]. Under this model, quantization of the corresponding hidden variables yields binary masks $\mathbf{r}$ and $\mathbf{c}$:

$$\begin{aligned} o_t &= \begin{cases} \text{on} & r_{k,t-1} = \text{back } \forall k = 1 \dots M \text{or } t = 1 \\ \text{off} & \text{otherwise} \end{cases} \\ \mathbf{s}^r_t &\sim [o_t = \text{off}]\mathcal{N}(A^r\mathbf{s}^r_{t-1}, Q^r) + [o_t = \text{on}]\mathcal{N}(m^r, P^r) \\ \mathbf{x}^r_t &\sim \mathcal{N}(W^r\mathbf{s}^r_t, R^r) \\ r_{k,t} &= \text{sgn}(x^r_{k,t}). \end{aligned}$$

Appropriately chosen W^r and A^r yield typical $x_{k,t}$ that are smooth functions of k and t, hence their signs also alternate slowly. We use this property to model smooth object masks. When an object leaves the scene in the previous time frame, i.e. when all indicators switch to $r_{k,t-1} = \text{back}$ for $k = 1 \dots M$, we trigger the onset indicator by setting $o_t = \text{onset}$. For shadow masks, we use a simpler model without regime switching:

$$\begin{aligned} \mathbf{s}^c_0 &\sim \mathcal{N}(m^c, P^c) & \mathbf{s}^c_t &\sim \mathcal{N}(A^c\mathbf{s}^c_{t-1}, Q^c) \\ \mathbf{x}^c_t &\sim \mathcal{N}(W^c\mathbf{s}^c_t, R^c) & c_{k,t} &= \text{sgn}(x^c_{k,t}). \end{aligned}$$

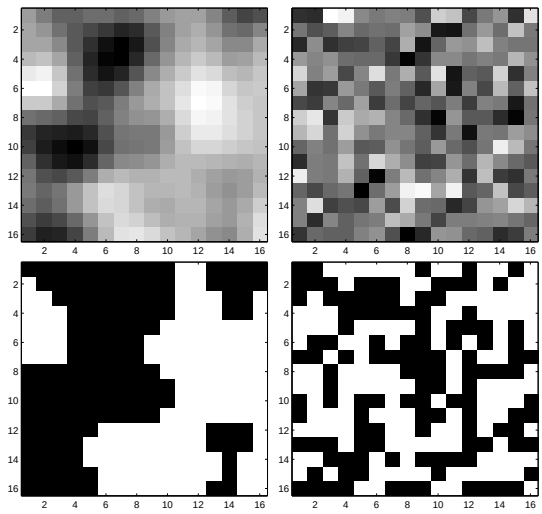

Figure 2. Random draws from the prior for different correlation values. Top row, $x_{1:K}$, bottom row, $r_{1:K}$, arranged as a 16×16 grid. In this example, the covariance matrix is parametrized as $K = \{\kappa_{k,k'}\}$ where $\kappa_{k,k'} = \exp(-\frac{1}{\theta^2}\|k - k'\|^2)$ with $\|\cdot\|$ denotes the Euclidean norm. Left column and right column correspond to samples with $\theta = 2$ and $\theta = 0.5$, respectively. If the correlation coeffi cient is chosen to be large for adjacent indices k and k', the corresponding random variables x_k and $x_{k'}$ will have a-priori close values. Hence, typical $x_{1:K}$ will be smooth functions (of k) and consequently their sign will also alternate smoothly. The main idea of this paper is to induce a spatio-temporal correlation structure via an embedded linear dynamical system.

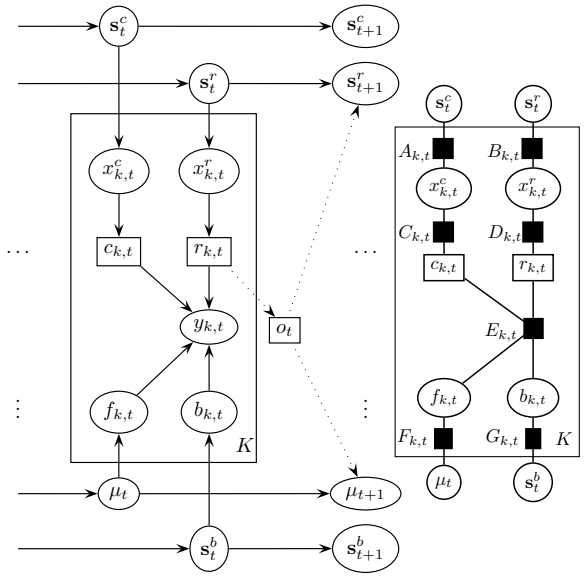

Figure 3. (Left) Graphical Model of a single time frame. The rectangles are plates that denote K repetitions of nodes inside. Square and oval nodes correspond to discrete and continuous variables, respectively. Variables f, b, y are vectors denoting the RGB components. Dotted arcs depict the regime switch mechanism that is triggered when an object leaves the scene. (Right) Loopy Factor graph of a single slice used in EP iterations. Forward messages to the next time slice (not shown) are passed only once. Filled rectangles (A, B, C, D, E, F, G) represent factors.

2.2. Generative model for pixel values

The pixels of the background image are assumed to be quasi-static. However, due to camera jitter and long term deviation in the illumination conditions, the pixel values are not exactly constant (See Fig.1). The background pixel values, denoted as $b_{k,t}$, are assumed to be generated from the following linear dynamical system:

$$\mathbf{s}_0^b \sim \mathcal{N}(m^b, P^b) \quad \mathbf{s}_t^b \sim \mathcal{N}(A^b \mathbf{s}_{t-1}^b, Q^b) \quad \mathbf{b}_t \sim \mathcal{N}(W^b \mathbf{s}_t^b, R^b).$$

The rationale behind this model is simple: Suppose that W^b is equal to a "snapshot" of the background image at time 0 and $A^b = 1$. In this case $\mathbf{s}_t^b$ is a scalar. We can interpret it as a global intensity variable that undergoes a random walk with transition noise variance Q^b. In general, we can represent background with multiple basis vectors arranged as columns of W^b. The expansion coefficients $\mathbf{s}^b$ will correspond to a low dimensional representation. The variables $(A, Q, W, R)^b$ can be learned offline to reflect the expected variation in the background image. Similar models for appearance have been considered in [4].

The foreground objects may have a variety of colors and textures depending upon the observed scene (e.g. highway, office). Hence, the prior distribution should be selected on a case by case basis, according to the features which one believes are important. We let $f_{k,t} \sim p(f|\mu_t)$ where μ_t is a parameter vector. When an object leaves the scene, as indicated by o_t, the parameter vector is reinitialized

$$\mu^{\text{new}} \sim p(\mu) \qquad \mu_t = [o_t = \text{off}]\mu_{t-1} + [o_t = \text{on}]\mu^{\text{new}},$$

where $p(\mu)$ is a suitable prior distribution. One choice is to describe object's appearance with a linear Gaussian model (e.g. factor analysis model [4]) and let $\mu_t = W_t^f$, where W^f are basis vectors representing the object's appearance. Due to the limited scope of this paper, we assume that the foreground pixels $f_{k,t}$ are drawn independently from an uniform distribution.

Pixel values and the masks render the observed image as

$$\begin{aligned} \rho_{k,t} &= [c_{k,t} = \text{no-shadow}] + [c_{k,t} = \text{shadow}]\rho \\ y_{k,t} &= \rho_{k,t}\left([r_{k,t} = \text{fore}]\, f_{k,t} + [r_{k,t} = \text{back}]b_{k,t}\right). \end{aligned} \tag{1}$$

Here, $0 < \rho < 1$ is a fixed parameter modeling the amount of illumination drop due to shadows. The graphical model is shown in Figure 3. A typical draw from the model is depicted in Figure 4

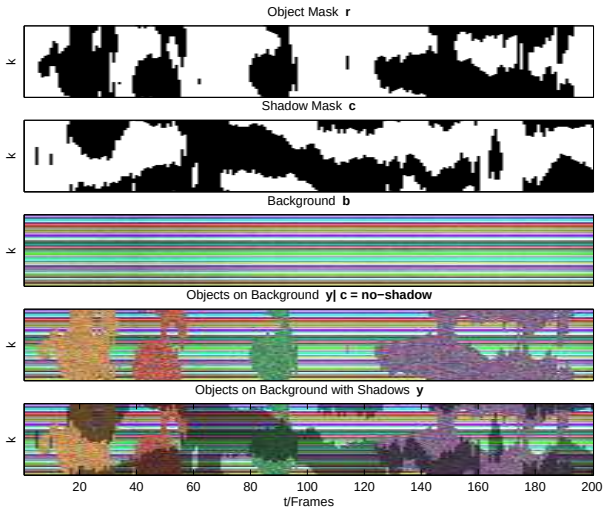

Figure 4. A random draw from the model. Note the spatial and temporal correlations in the generated masks.

3. Inference

We are interested in various marginals of the smoothed density $p(\mu_{1:T}, \mathbf{r}_{1:T}, \mathbf{s}^r_{1:T}|\mathbf{y}_{1:T})$ or the filtering density $p(\mu_t, \mathbf{r}_t, \mathbf{s}^r_t|\mathbf{y}_{1:t})$ (for online operation). In either case, the latent binary variables c, r render the model an intractable hybrid graphical model[7] where desired marginals can be computed only approximately.

To select a suitable inference method, consider the model structure: given the masks $\mathbf{r}$, $\mathbf{c}$, we could integrate over the background pixel process analytically (since it is a KFM). In principle, we could sample from the masks, however, unless the continuous latent parents $\mathbf{s}^c$ and $\mathbf{s}^r$ are low dimensional, the prior probabilities of masks can not be computed easily. One needs to sample from $\mathbf{s}^c$ and $\mathbf{s}^r$ and this renders Rao-Blackwellized particle filtering or Gibbs sampling computationally expensive [1]. Alternatively, one could approximate the prior by a mean-field approximation. However, due to the hard clipping mechanism, mean field with factorized Gaussians as the approximating family would break down (since Kullback-Leibler divergence between any Gaussian and a clipped Gaussian becomes ∞) and we need to relax clipping with a sigmoidal soft threshold or use a more exotic approximating family.

3.1. Expectation Propagation

Here, we investigate an alternative deterministic approximation method, based on Expectation propagation [8]. EP is an iterative message-passing algorithm and generalizes Loopy Belief Propagation (LBP) [15], in that it is directly applicable to arbitrary hybrid graphical models, including the model we have introduced. One can view EP as a local message passing algorithm on a factor graph [5] where messages are passed between factors –potentials representing local dependencies, and beliefs – approximate marginal potentials. Unlike multinomial or Gaussian models, where all marginals stay within a closed family, in hybrid models, beliefs computed from mixed messages may have complicated forms (e.g. mixtures, clipped Gaussians). In such cases, EP replaces a belief with a potential from an approximating exponential family, usually in KL sense, i.e. by matching moments. The quality of the approximation depends, how well these "summary statistics" represent the exact beliefs.

The algorithm can be summarized as the following fixed-point iteration between messages $m_{F\to\xi}$ and beliefs $q(\xi)$:

$$m_{F\to\xi}(\xi)^{\text{new}} := \frac{Zq^{\text{new}}(\xi)}{q(\xi)} m_{F\to\xi}(\xi) \quad (2)$$

$$q^{\text{new}}(\xi) = \text{argmin}_q \text{KL}(\tilde{q}|q) \quad (3)$$

$$\tilde{q}(\xi) := \frac{1}{Z}\sum_{F\setminus\xi}\psi(F)\prod_{\xi'\in F}\frac{q(\xi')}{m_{F\to\xi'}(\xi')} \quad (4)$$

Here ξ is an index that runs over variables, F is an index set that runs over factors, $\psi(F)$ is a local factor potential, $F\setminus\xi$ denotes variables adjacent to F except for ξ, and Z is a normalization constant [8].

Implementation We implemented each time-slice of our model with a simple factor graph as shown in Fig.3. This structure corresponds to a fully factorized approximation to the joint posterior. We use Gaussians and multinomials as natural approximating families for the continuous and discrete variables. The message update equations are derived as particular cases of the general scheme in Eq.4. The factor potentials $A_{k,t}, B_{k,t}, F_{k,t}, G_{k,t}$ are Gaussian, therefore $q^{\text{new}} = \tilde{q}$ is a Gaussian that can be computed analogously as in the well-known linear Gaussian models (see [5]). Below, we solve Eqs. 3–4 for the non-linear factors $D_{k,t}$ and $E_{k,t}$ (the equations for $C_{k,t}$ are identical to that of $D_{k,t}$).

For simplicity, we omit the time/pixel indicies $\{k,t\}$. Consider factor D and the adjacent variables $\{x,r\}$. The clipping mechanism translates to the simple local potential $\psi(x,r) = [rx > 0]$. Let $\frac{q(r)}{m_{F\to r}(r)} = Z_{\text{r}}[w(-1), w(1)]$ be a multinomial potential, $w(-1)+w(1)=1$ and $\frac{q(x)}{m_{F\to x}(x)} = Z_{\text{x}}\mathcal{N}(x|\mu,v)$ be Gaussian. To update belief $q(r)$ we substitute to Eq. 4 with $r\in\{-1,1\}$ and integrate x

$$\tilde{q}(r) = \frac{1}{Z}w(r)\int_x [xr>0]\mathcal{N}(x;\mu,v) = \frac{1}{Z}w(r)\lambda(r),$$

where $Z = Z_{\text{r}}Z_{\text{x}}\sum_r w(r)\lambda(r)$ and the normalization constant of a clipped Gaussian is $\lambda(r) = (1+r)/2 - (r/2)\,\text{erfc}(\mu/\sqrt{2v})$. This update corresponds to a "prior"

on indicator $r \in \{-1, 1\}$. The distribution $\tilde{q}$ is already multinomial, so we can directly substitute $q^{\text{new}}(r) = \tilde{q}(r)$.

Now, we consider the same factor D and update belief $q(x)$. Substituting to Eq. 4 with $r \in \{-1, 1\}$ gives

$$\tilde{q}(x) = \frac{1}{Z} \sum_r [xr > 0] w(r) \mathcal{N}(x|\mu, v).$$

Note that $\tilde{q}$ is unimodal. Minimization of Eq. 3 corresponds to setting the moments of the new belief $q^{\text{new}}(x) = \mathcal{N}(x|\mu_{\text{x}}, v_{\text{x}})$ equal to the moments $\tilde{q}(x)$

$$\mu_{\text{x}} = \sum_r w(r) \langle x|r \rangle \qquad v_{\text{x}} = \sum_r w(r) \langle x^2|r \rangle - \mu_{\text{x}}^2,$$

where the moments of a normalized clipped Gaussian density $[rx > 0]\mathcal{N}(x|\mu, v)/\lambda(r)$ are given by

$$\langle x|r \rangle = \mu + r\eta/\lambda(r) \qquad \langle x^2|r \rangle = \mu^2 + v + r\mu\eta/\lambda(r),$$

and $\eta = \exp(-\mu^2/2v)\sqrt{v/2\pi}$.

Next, consider factor E^ν and the adjacent variables $\{r, c, f^\nu, b^\nu\}$, where $\nu \in \{\text{R}, \text{G}, \text{B}\}$ denotes the color channel, and f^ν and b^ν are scalars denoting channel values for the current pixel. The equations for all channels are identical, so we omit the index ν. Let the current belief/message ratios be $\frac{q(r)}{m_{F\to r}(r)} = Z_{\text{r}}[w_{\text{r}}(1), w_{\text{r}}(-1)]$, $\frac{q(c)}{m_{F\to c}(c)} = Z_{\text{c}}[w_{\text{c}}(1), w_{\text{c}}(-1)]$ for the binary variables r and c. Let $\frac{q(f)}{m_{F\to f}(f)} = Z_{\text{f}}\mathcal{N}(f|\mu_{\text{f}}, v_{\text{f}})$ and $\frac{q(b)}{m_{F\to b}(b)} = Z_{\text{b}}\mathcal{N}(b|\mu_{\text{b}}, v_{\text{b}})$ denote Gaussian potentials, and $\delta()$ the Dirac-delta function. The factor function follows from Eq. 1

$$\begin{aligned}\psi(E) = &\left[\binom{r}{c} = \binom{1}{-1}\right] \delta(y - b) + \left[\binom{r}{c} = \binom{1}{1}\right] \delta(\rho y - b) \\ &+ \left[\binom{r}{c} = \binom{-1}{-1}\right] \delta(y - f) + \left[\binom{r}{c} = \binom{-1}{1}\right] \delta(\rho y - f),\end{aligned}$$

where y is a scalar denoting the observed pixel value in the current color channel, and $\left[\binom{a_1}{a_2} = \binom{b_1}{b_2}\right] = 1$(or $= 0$) when $a_1 = b_1$ and $a_2 = b_2$ (otherwise).

We start with updating belief $q(r)$ by substituting to Eq. 4 and integrating variables f, b, c

$$\tilde{q}(r) = \frac{1}{Z} w_{\text{r}}(r) \sum_c w_{\text{c}}(c) \mathcal{N}(y|\rho(c)\mu_{\xi(r)}, v_{\xi(r)})$$

with $\rho(1) = \rho$, $\rho(-1) = 1$, and ξ is an indicator function: $\xi(1) = \text{b}$, $\xi(-1) = \text{f}$. Similar results hold for $\tilde{q}(c)$. For discrete indicators, $\tilde{q}$ is multinomial, thus $q^{\text{new}} = \tilde{q}$.

Now consider updating $q(b)$. Substituting to Eq. 4 and integrating r, c, f gives $\tilde{q}(b)$ in the form of a mixture of two delta functions and a Gaussian density

$$\tilde{q}(b) = \alpha_1 \delta(b - \rho y) + \alpha_2 \delta(b - y) + \alpha_3 \mathcal{N}(b|\mu_{\text{b}}, v_{\text{b}}),$$

where

$$\alpha_1 = \frac{w_{\text{r}}(1) w_{\text{c}}(1)}{Z} \gamma(y/\rho), \qquad \alpha_2 = \frac{w_{\text{r}}(1) w_{\text{c}}(-1)}{Z} \gamma(y)$$
$$\alpha_3 = \frac{w_{\text{r}}(-1)}{Z} \sum_c \gamma(y/\rho(c)), \quad \gamma(y) = \mathcal{N}(y|\mu_{\text{f}}, v_{\text{f}}).$$

The new Gaussian belief $q^{\text{new}}(b) = \mathcal{N}(b|\mu, v)$ follows from moment matching between $q^{\text{new}}(b)$ and $\tilde{q}(b)$

$$\mu = \langle b \rangle_{\tilde{q}} = \alpha_1 \rho y + \alpha_2 y + \alpha_3 \mu_{\text{b}}$$
$$v = \langle b^2 \rangle_{\tilde{q}} - \langle b \rangle_{\tilde{q}} = \alpha_1 (\rho y)^2 + \alpha_2 y^2 + \alpha_3 (v_{\text{b}} + \mu_{\text{b}}^2) - \mu.$$

Analogous derivation holds for updating the belief $q(f)$.

Schedule We have used the following order of message and belief updates for a fixed time-slice t

loop until EP convergence
  for $k = 1 \dots K$
    $A_{k,t} \to x^c_{k,t}$, $B_{k,t} \to x^r_{k,t}$, $C_{k,t} \to c_{k,t}$, $D_{k,t} \to r_{k,t}$
    $F_{k,t} \to f_{k,t}$, $G_{k,t} \to b_{k,t}$
    $E_{k,t} \to c_{k,t}$, $E_{k,t} \to r_{k,t}$,$E_{k,t} \to f_{k,t}$, $E_{k,t} \to b_{k,t}$
    $A_{k,t} \to s^c_t$, $B_{k,t} \to s^r_t$, $F_{k,t} \to \mu_t$, $G_{k,t} \to s^b_t$

Here, $F \to \xi$ denotes simultaneous update of the message $m_{F\to\xi}$ and belief $q(\xi)$ according to the Eqs. 3–4. We refer to the inner loop as a single EP iteration. After convergence we move on to the next time-slice.

4. Experiments

In the following experiments we study the convergence rate of the EP-based inference in our model, and compare our model with two other popular techniques for recovering smooth masks: Markov Random Fields and mathematical morphology.

EP approximation Due to the iterative nature of EP, the convergence rate is a key factor for online implementation of our model. In this experiment we demonstrate the convergence and accuracy of EP approximation to the prior $p(\mathbf{r})$ (the source of intractability). Inference in this submodel is equivalent to calculating the integral of the Gaussian distribution $\int d\mathbf{s}^r p(\mathbf{x}^r|\mathbf{s}^r) p(\mathbf{s}^r)$ in one of the 2^K orthants specified by the mask configuration (see Fig.5, left). In one dimension, this integral can be evaluated easily as shown in the previous section; however in higher dimensions we need to reside to approximations. Fortunately, we can ensure by construction that the Gaussian is positive definite, hence the clipped Gaussian will be unimodal and we expect a factorized EP approximation to converge. In the middle panel of Fig.5 we show results of an experiment where we have computed the likelihood of all configurations of $\mathbf{r}$ for $K = 8$, ranked them and compared to importance sampling. We observe in this and similar experiments that EP converges in 2-3 iterations and is indeed very accurate. Following this observation, in the subsequent tests we run two EP iterations per frame.

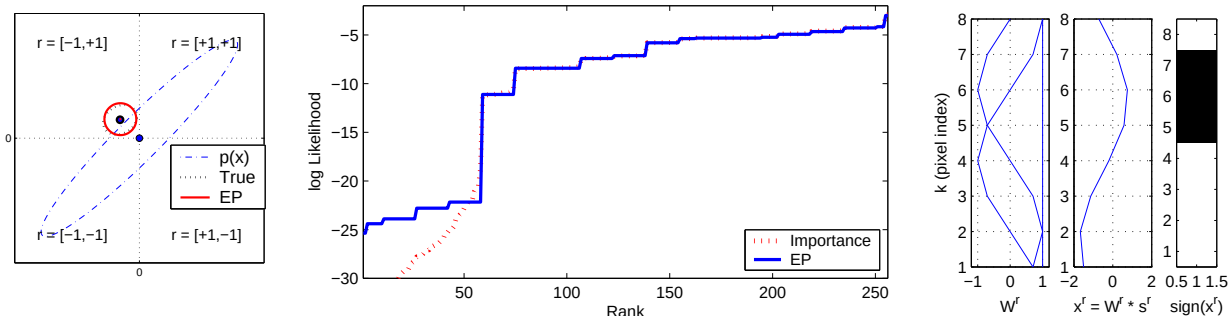

Figure 5. (Left) $K = 2$, a Gaussian clipped at $r = [-1, 1]$ and EP approximation to its true moments. (Middle) $K = 8$. Comparison of EP with importance sampling. We use a truncated Fourier basis $W_k^r = [1, \sin(\omega k), \cos(\omega k)]$ with $\omega = 2\pi/K$ with $R^r = 0.1^2 I, m^r = 0$. Note that many configurations have the same probability since $m^r = 0$. For importance sampling we have drawn 10^7 samples independently from $\mathbf{s}^r$ and integrated over $\mathbf{x}^r$ to compute $p(\mathbf{r})$. (Right) $K = 8$ pixels arranged in a column; three basis vectors W_k^r; an example mask obtained from coefficients $\mathbf{s}^r = [-0.4, -0.2, -1.2]^\top$.

Comparison with MRF To illustrate the performance on real data, we use the video sequence shown in Fig. 1. In this experiment we split video frames into individual vertical scanlines, which are processed independently. For each scanline (a column of 240 pixels) we define mask prior, similarly as in the right panel of Fig 5. In this way, we demonstrate the smooth changes of inferred masks over time.

We have trained the parameters of the background process $(A, Q, W, R, m, P)^b$ using an EM algorithm from frames of the empty scene. Mask parameters are set to $A^r = \gamma I$, $Q^r = (1-\gamma)I$ with the fudge factor $\gamma = 0.95$. W^r is taken as a truncated Fourier basis, the same as in the previous experiment. Other parameters are $(m^r, P^r) = ([1.5, 0.5, 0.5]^\top, 0.2I)$. Parameters for shadow masks are identical. In principle these can also be learned from presegmented data.

We also consider a similar model, where the prior correlations on masks are enforced with an MRF that couples neighboring pixels. Fig. 8 (Top) shows the object masks recovered by the filtering density $p(\mathbf{r}_t|\mathbf{y}_{1:t})$ in our model, and Fig. 8 (Bottom) in the MRF case. The differences between the two approaches become more clear in Fig. 9, where we inspect the filtering density on the "change-point" variable $p(o_t|\mathbf{y}_{1:t})$. This variable can be interpreted as an object-presence indicator. Objects can be automatically detected whenever $p(o_t|\mathbf{y}_{1:t})$ changes to a value grater than a suitable threshold τ. The advantage of our model is evident in Fig. 6, where we show the averaged false-alarm rates as a function of τ, using two videos recorded in an office-like environment (each of approx. 1500 frames).

Comparison with morphological smoothing Our model can be also used for smoothing of masks that were computed by an independent segmentation method. Typically, such post-precessing involves successive application

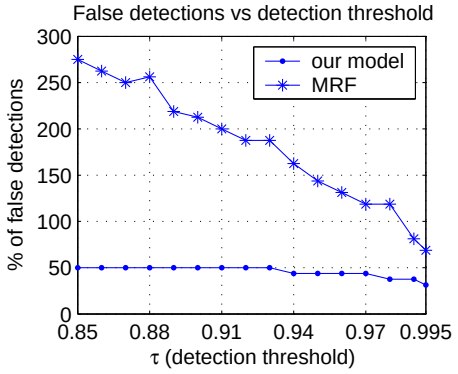

Figure 6. False-alarm rates for object detection as a function of detection threshold.

of mathematical morphology operations: erosion and dilation. Under our model, to infer underlying objects' shapes from a noisy mask $\hat{\mathbf{r}}_t$, we first find (with EP) the expected low-dimensional shape coefficients $\langle \mathbf{s}_t^r | \hat{\mathbf{r}}_t \rangle = \boldsymbol{\mu}_s$. The smooth shape (mask) $\mathbf{r}_t$ is reconstructed by clipping the linear projection of the coefficients $\boldsymbol{\mu}_s$; $\mathbf{r}_t = \text{sgn}(W^r \boldsymbol{\mu}_s)$.

In this experiment we split video frames into rectangular patches of $(N_y \times N_x)$ pixels and process each patch separately. The model parameters for each patch are identical. The basis vectors are linear functions of pixel coordinates; $W_{i,j}^r = [1, k_i + k_j, k_i - k_j]$, were k_i (respectively, k_j) are N_y (N_x) uniformly spaced points from $[-1, +1]$ (see Fig. 7). The other mask parameters are set to $A^r = \gamma I$, $Q^r = (1-\gamma)I$, $\gamma = 0.99$, $(m^r, P^r) = ([0, 0, 0]^\top, I)$. Fig. 10 we presents selected frames from the video sequence, where we compare original masks to the shapes recovered by our model and those recovered by mathematical morphology. The complete video is available at http://staff.science.uva.nl/∼wzajdel/demos/cvpr05.mpg.

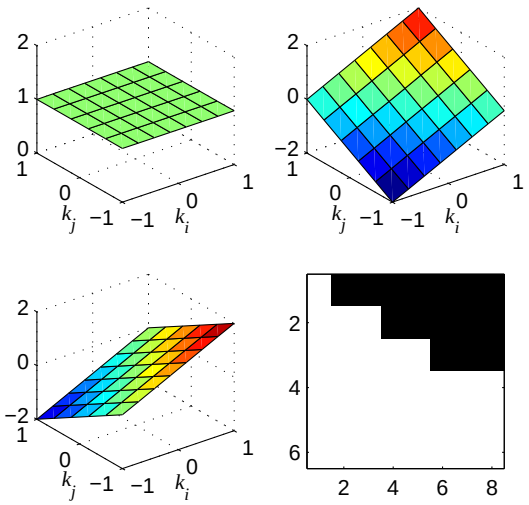

Figure 7. Basis vectors for modeling masks in two-dimensional patches. (Top Left) $W^r_{i,j} = 1$. (Top Right) $W^r_{i,j} = k_i + k_j$ (Bottom Left) $W^r = k_i - k_j$, were k_i (respectively, k_j) are N_y (N_x) uniformly spaced points from $[-1, +1]$. (Bottom Right) An example of a mask for $N_y = 6, N_x = 8$ generated from these vectors, with coeffi cient $\mathbf{s}^r = [0.52, 0.21, -0.92]^\top$.

5. Discussion and Conclusion

We have described a model for robust object detection, where background estimation need not be viewed as an independent preprocessing step. Automatic identity estimation based on the objects' features requires additional machinery; which can be readily coupled in a transparent way to the presented model. This allows us properly characterize any uncertainty in the foreground detection, which arguably is important for robust object identification.

An attractive feature of the model is that any image subregion with an arbitrary shape (e.g. full frames, scanlines or randomly scattered patches) may be processed, since given $\mathbf{s}$ variables, the spatial ordering of pixels is irrelevant. Additionally, the dimensionality of $\mathbf{s}$ variables can be tuned to trade off correlations and computation time (which scales as $O(|S|^2TK)$, where $|S| = \max_{\xi\in\{r,c,b\}} \dim \mathbf{s}^\xi_t$).

The model, as it now stands, assumes that at most one object is observed at a given time frame. This assumption holds for small image regions, but for larger images the model has to be extended, e.g. by considering multiple layers [4], each with a single object and a corresponding subgraph. Whilst the local message passing mechanism offers a computationally tractable solution, it remains to be seen, whether such an extension to the framework described here can be implemented in practice.

Acknowledgments

We thank Zoran Zivkovic for assistance with experiments. The work described in this paper was supported by the Dutch Technology Foundation STW and Dutch Ministry of Economic Affairs under the grant no ANN.5312.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1–2):5–43, 2003.

[2] P. Fearnhead. Exact and effi cient bayesian inference for multiple changepoint problems. Technical report, Dept. of Math. and Stat., Lancaster University, 2003.

[3] N. Friedman and S. Russell. Image segmentation in video sequences. A probabilistic approach. In *Uncertainty in Artifi cial Intelligence*, 1997.

[4] N. Joijc and B. J. Frey. Learning flexible sprites in video layers. In *IEEE Computer Vision and Pattern Recognition*, pages 199–206, 2001.

[5] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[6] D. D. Lee and H. Sompolinsky. Learning a continuous hidden variable model for binary data. In *Advances in Neural Information Processing Systems*, 1998.

[7] U. Lerner and R. Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *Uncertainty in Artifi cial Intelligence*, pages 310–318, 2001.

[8] T. Minka. *Expectation Propagation for approximate Bayesian inference*. PhD thesis, MIT, 2001.

[9] J. Rittscher, J. Kato, S. Joga, and A. Blake. A probabilistic background model for tracking. In *Proc. of European Conf. Computer Vision*, 2000.

[10] C. Stauffer and W. E. Grimson. Adaptive background mixture modelling for realtime tracking. In *IEEE Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[11] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. In *Proc. of European Conf. on Computer Vision*, pages 307–323, 2000.

[12] M. E. Tipping. Probabilistic visualisation of high-dimensional binary data. In *Advances in Neural Information Processing Systems*, 1998.

[13] D. Wang, T. Feng, H.-Y. Shum, and S. Ma. A novel probability model for background maintenance and subtraction. In *Int. Conf. on Vision Interface*, pages 109–117, 2002.

[14] C. K. I. Williams and D. Barber. Bayesian classifi cation with Gaussian processes. *IEEE Transactions on Pattern Ananlysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[15] J. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Int. Joint Conf. on Artifi cial Intelligence*, 2001. Distinguished Papers Track.

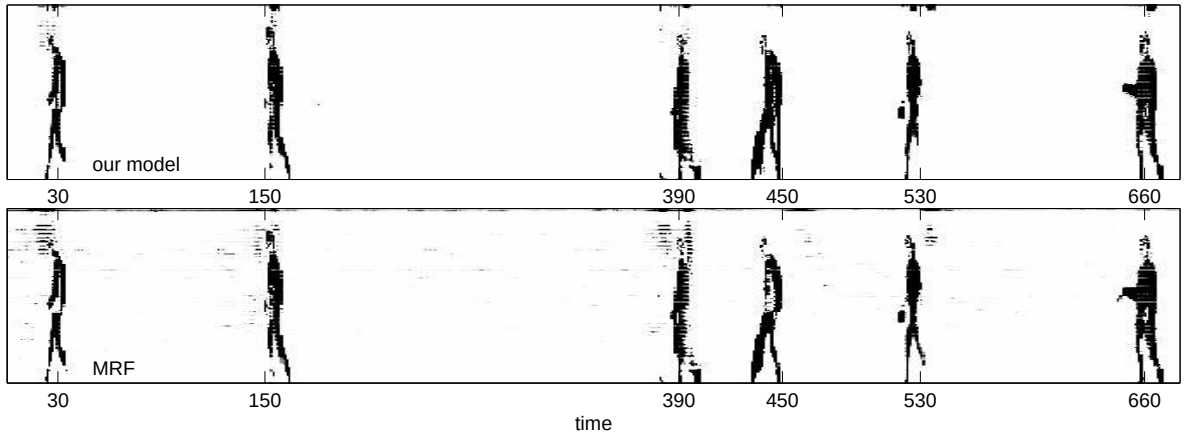

Figure 8. Filtered estimates of object masks $p(\mathbf{r}_t|\mathbf{y}_{1:t})$. (Top) Clipped Gaussian prior (our model). (Bottom) Markov Random Filed prior, with coupling strength 0.7. Compare with Fig. 1

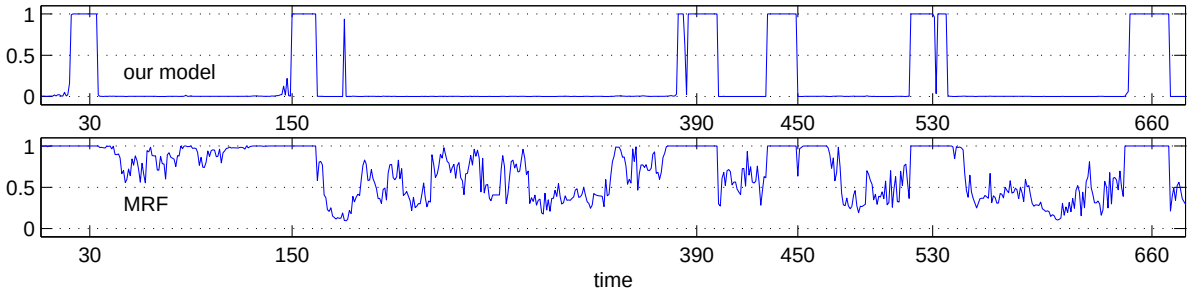

Figure 9. Filtering distribution on the "change-point" variable $p(o_t|\mathbf{y}_{1:t})$. (Top) Our model. (Bottom) MRF prior.

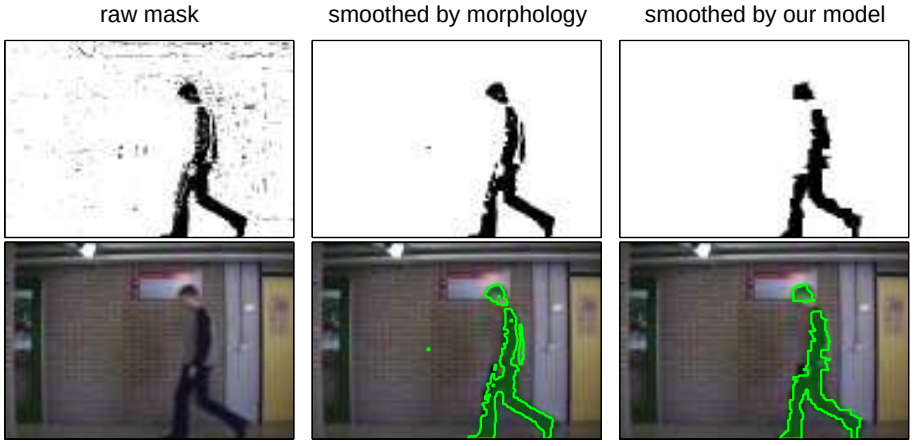

Figure 10. (Left) Segmentation of independent pixels with [10]. (Middle) Object shape recovered by morphology (1 erosion, 2 dilations, 1 erosion). (Right) Object shape recovered by our model, when applied independently to rectangular patches ($N_y = 6, N_x = 8$) within the frame.