Nonparametric Bayesian attentive video analysis

Giuseppe Boccignone
Natural Computation Lab–DIIIE
Universitá di Salerno, via Melillo 1, 84084 Fisciano, Italy

Abstract

We address the problem of object-based visual attention from a Bayesian standpoint. We contend with the issue of joint segmentation and saliency computation suitable to provide a sound basis for dealing with higher level information related to objects present in dynamic scene. To this end we propose a framework relying on nonparametric Bayesian techniques, namely variational inference on a mixture of Dirichlet processes.

1. Introduction

Visual attention not only restricts various types of visual processing to certain spatial areas of the visual field but also accounts for object-based information, so that attentional limitations are characterized in terms of the number of discrete objects which can be simultaneously processed [7]. The object-based nature of attention is readily apparent in dynamic situations, in which object tokens must be maintained over time, e.g., multiple object tracking [7].

Clearly, there may be a hierarchy of units of attention, from intra-objects parts to perceptual groups. Similarly, segmentation processes, that bundle parts of the visual field together as units, are likely to take place at all levels of visual processing. Further, units of segmentation processes may serve as the focus of attention, while the units of other segmentation processes may be in part the result of object-based attention [7, 4].

This multiplicity of levels raises a number of ontological and epistemological concerns on the definition of an object; in most cases one is free to consider almost anything as an object [7]. However, objecthood is more well defined at earlier levels of visual analysis [7], which are the ones we are mostly dealing with here. Indeed, many experimental results suggest that this packaging of the world into units may occur quite early, and even pre-attentively [7]. Some of these processes are early, using "quick and dirty" heuristics to obtain "proto-object" units for further processing, which meanwhile may serve as potential preliminary units of attention [4].

In this note, in the framework of Bayesian models of attention (e.g., [9],[5]) and taking into account issues raised by object-based theories [7], we propose a computational model to cope in a unified way with both segmentation and gaze control. Indeed the joint computation of segmentation and saliency is suitable to set a sound basis for exploiting higher level information related to objects present in dynamic scene, and eventually to set the focus of attention (FOA). To this end we address nonparametric Bayesian techniques, and in particular variational inference on Dirichlet process mixture (DPM) representation [3].

2. Bayesian object-based attention

The observed image sequence $\{\mathbf{x}_t\}_{t=1}^{T}$ is modeled as a spatio-temporal volume of contiguous frame features $\mathbf{x}_t = \{\mathbf{x}_{n,t}\}_{n=1}^{N}$, where n is a site index standing for the coordinates $\mathbf{r} = \{r_x, r_y\}$. In this representation every pixel is mapped to a $7D$ feature vector, $\mathbf{x}_{n,t} = \{\mathbf{x}_{n,t}^{col}, \mathbf{x}_{n,t}^{vel}, \mathbf{x}_{n,t}^{space}\}$, where $\mathbf{x}_{n,t}^{col}$ is a $3D$ vector of a suitable color space, $\mathbf{x}_{n,t}^{vel}$ is a $2D$ motion vector, e.g. velocities as derived from optical flow algorithms, and $\mathbf{x}_{n,t}^{space} = \mathbf{r}_t$ are the spatial coordinates in frame t.

In such representation, we may provide a minimal notion of (proto-)object existing at time t, say $\mathrm{O}_t = \{\mathbf{r}_t, \mathcal{A}_t, \mathcal{O}\}$, in terms of its position, appearance parameters, and label, respectively [9]. The set of labels $\mathcal{O}$ is discrete and can index any category of objects, but for the purposes of this study, objects are defined in terms of foreground moving regions embedded within a noisy background; also, in this study appearance $\mathcal{A}_t$ will simply represent the object bounding box (from which size and aspect ratio can be derived). Thus, the problem of deploying attention at a certain object $\hat{\mathrm{O}}_t$, given that features $\mathbf{x}_t$ are observed and prior knowledge is available, can be formulated in Bayesian terms as the problem of making a decision upon the posterior probability

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

to be inferred via Bayes' rule $p(\mathrm{O}_t|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|\mathrm{O}_t)p(\mathrm{O}_t)}{p(\mathbf{x}_t)}$. To this end, the numerator can be developed as the product of the followig terms: $p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{A}_t, \mathcal{O})$, representing the likelihood of observing specific object features; $p(\mathbf{r}_t|\mathcal{A}_t, \mathcal{O})$, the probability of focusing on position $\mathbf{r}_t$ conditioned on object appearance (e.g., the size, or aspect ratio); the prior joint probability $p(\mathcal{A}_t, \mathcal{O})$ of observing specific objects and appearance parameters. The latter is set to a constant, since here we are dealing with scenes in which the actual objects of interest are mostly represented by people engaged in different kinds of actions either individual or collective (walking, talking, etc.), occurring in different parts of the scene. Then, inference can be performed as:

$$p(\mathrm{O}_t|\mathbf{x}_t) \simeq \frac{p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{A}_t, \mathcal{O})p(\mathbf{r}_t, |\mathcal{A}_t, \mathcal{O})}{p(\mathbf{x}_t)}. \quad (1)$$

The term $p(\mathbf{x}_t)^{-1}$ can be thought of as low-level saliency information in the sense of Shannon [9], which biases object-dependent information as provided by the terms in the numerator. Thus, Eq. 1 straightforwardly shows how the FOA can be inferred by taking into account object-based cues, but weighted with spatiotemporal low-level cues (denominator).

From Eq. 1 the choice of gazing at a certain object $\hat{\mathrm{O}}_t$ at time t, can be taken as $\hat{\mathrm{O}}_t = \arg\max_{\mathrm{O}_t} p(\mathrm{O}_t|\mathbf{x}_t)$.

In order to compute the posterior $p(\mathrm{O}_t|\mathbf{x}_t)$, on the one hand we have to specify low-level information $p(\mathbf{x}_t)$; on the other hand, we have to take into account object-dependent information so to provide a suitable form for $p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{A}_t, \mathcal{O})$ and $p(\mathbf{r}_t, |\mathcal{A}_t, \mathcal{O})$. This latter problem involves some sort of preliminary segmentation of objects present in the scene, which in turn should be based on the selective organization of the observable (low-level) features. Indeed, this is but one example of the close relationship between attention and segmentation, which we discussed in the introductory session.

Here, we propose to model $p(\mathbf{x}_t)$ in a form which, beyond shaping bottom-up saliency, is meanwhile suitable to provide a preliminary segmentation of the scene. One such representation is a mixture of components.

A K component mixture model takes the general form $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\theta}_k)$ [2]. Each mixture component, or cluster, belongs to a parameterized family of probability densities $p(\mathbf{x}|\boldsymbol{\theta}_k)$, e.g., in the case of gaussian components $p(\mathbf{x}|\boldsymbol{\theta}_k) = N(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are the means and covariance matrices, respectively. In such representation, each data point $\mathbf{x}_n$ is generated by independently selecting one of K clusters according to the multinomial distribution and then sampling from the chosen cluster's data distribution. When learning mixtures from data, it is often useful to place an independent conjugate prior $\mathcal{G}_0$, with hyperparameters λ, on each cluster's parameters, so that the latter can be generated through the draw $\boldsymbol{\theta}_k \sim \mathcal{G}_0(\lambda)$, $k = 1, \ldots K$. Here, "$Y \sim \mathcal{S}$" means "Y has the distribution $\mathcal{S}$", Similarly, the mixture proportion for the classes $\boldsymbol{\pi} = \pi_k{}_{k=1}^{K}$ can be assigned a symmetric Dirichlet prior with concentration parameter α, so that $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha/K \ldots \alpha/K)$.

Then, each data point $\mathbf{x}_n$ can be generated through the following hierarchy of draws: 1. $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha/K \ldots \alpha/K)$; 2. $\boldsymbol{\theta}_k \sim \mathcal{G}_0$; 3. $z_n \sim \mathrm{Mult}(\boldsymbol{\pi})$; 4. $\mathbf{x}_n \sim p(\mathbf{x_n}|\boldsymbol{\theta}_{z_n})$. Here, the unobserved indicator variable $z_n = k \in \{1, \ldots K\}$ specifies the unique cluster associated with observation $\mathbf{x}_n$, and $\mathrm{Mult}(\boldsymbol{\pi})$ is a multinomial distribution on mixing proportions.

Finite mixture models assume the number of clusters K to be a fixed, known constant. In general, determining an appropriate mixture size is a difficult problem, which has motivated a wide range of model selection procedures [2]. However, by placing prior distributions on infinite mixtures, DPM models are promising candidates for clustering applications where the number of clusters is unknown a priori, and very important, they naturally yield a clustering effect [3].

3. Saliency and segmentation via DPM

The DPM model assumes that a Dirichlet process $\mathcal{DP}$, with scaling parameter α is used as a nonparametric prior to generate a distribution $\mathcal{G}$ from the base distribution $\mathcal{G}_0$ and can be obtained from the finite K mixture model described above by taking the limit $K \to \infty$. Precisely, the observed data are generated through the following hierarchy of draws: 1. $\mathcal{G} \sim \mathcal{DP}(\alpha, \mathcal{G}_0)$; 2. $\boldsymbol{\theta}_n \sim \mathcal{G}$; 3. $\mathbf{x}_n \sim p(\mathbf{x_n}|\boldsymbol{\theta}_n)$. Here $\boldsymbol{\theta}_n$ plays the role of $\boldsymbol{\theta}_{z_n}$.

A characterization which provides a constructive approach to the DPM generative process is the *stick breaking* procedure. Let $\boldsymbol{v} = \{v_k\}_{k=1}^{\infty}$, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k\}_{k=1}^{\infty}$ be two infinite collection of random variables such that $v_k \sim \mathcal{B}(1, \alpha)$, the Beta distribution, and $\boldsymbol{\theta}_k \sim \mathcal{G}_0$. Then, the mixing proportions π_k can be obtained by iteratively breaking a unit length stick into an infinite number of pieces, where the length of each successive piece is independently drawn from $\mathcal{B}(1, \alpha)$. Formally: $\pi_k(\boldsymbol{v}) = v_k \prod_{j=1}^{k-1}(1 - v_j)$. Then, $\mathcal{G}$ is discrete with support consisting of a countably infinite set of atoms drawn independently from $\mathcal{G}_0$: $\mathcal{G} = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{v})\delta(\boldsymbol{\theta}_k)$, where $\delta(\boldsymbol{\theta}_k)$ is the distribution concentrated at the single point $\boldsymbol{\theta}_k$.

In such setting, the DPM process to generate the observed data is the following: 1. $v_k \sim \mathcal{B}(1, \alpha)$; 2. $\boldsymbol{\theta}_k \sim \mathcal{G}_0(\lambda)$; 3. $z_n \sim \mathrm{Mult}(\boldsymbol{\pi}(\boldsymbol{v}))$; 4. $\mathbf{x}_n \sim p(\mathbf{x_n}|\boldsymbol{\theta}_{z_n})$.

In our case, the observable data $\mathbf{x}_n$ are drawn from

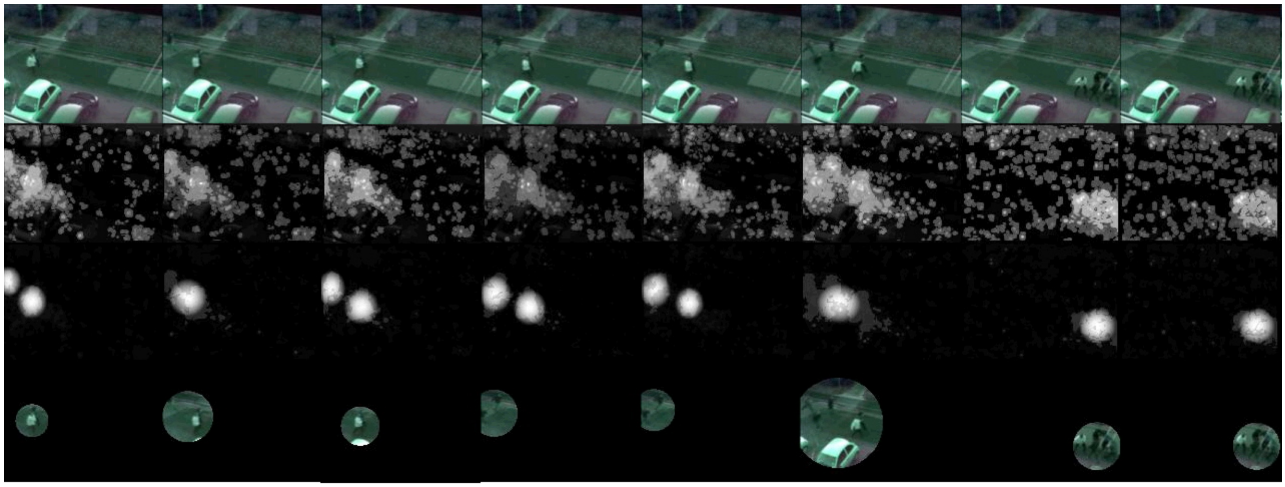

Figure 1. Example of results on a video from the data-set. Top row, from left to right: an excerpt of input frames; first 6 frames show a "run together" interaction, last 2 frames a "fight" interaction. Second row shows the corresponding grey-level coded saliency $p(\mathbf{x}_t)^{-1}$. Third row shows the log-posterior density. The bottom row presents the FOA chosen from the posterior.

an exponential family distribution (Gaussian), and the base distribution for the $\mathcal{DP}$ is the corresponding conjugate prior (i.e., the Gaussian distribution for $\boldsymbol{\theta}_k = \boldsymbol{\mu}_k$, and the Gaussian-Wishart for $\boldsymbol{\theta}_k = \boldsymbol{\Sigma}_k$ [2]).

A straightforward way of implementing the DPM generative process is by using MCMC methods (e.g., [6]). However, their computational cost is prohibitive for fields such as video processing and analysis. An appealing alternative are variational Bayes techniques [2]. In such framework we can rewrite the generative model in terms of the joint pdf as follows:

$$p(\boldsymbol{x}, \mathbf{z}, \boldsymbol{\xi}) = \prod_{n=1}^{N} p(\mathbf{x_n}|\boldsymbol{\theta}_{z_n}) p(z_n|\pi(\mathbf{v})) \prod_{k=1}^{\infty} p(\boldsymbol{\theta}_k|\lambda)\mathcal{B}(1,\alpha) \quad (2)$$

where $\boldsymbol{\xi} = \{\mathbf{v}, \boldsymbol{\Theta}\}$. Variational inference lower bounds the log marginal likelihood $\mathcal{L}(\boldsymbol{x})$ through the negative free energy [2] as:

$$\mathcal{L}(\boldsymbol{x}) \geq \sum_{\mathbf{z}} \int_{\boldsymbol{\xi}} Q(\mathbf{z}, \boldsymbol{\xi}) \log \frac{p(\boldsymbol{x}, \mathbf{z}, \boldsymbol{\xi})}{Q(\mathbf{z}, \boldsymbol{\xi})} \quad (3)$$

Mean field variational inference can be achieved by using the following factorized family of variational distributions:

$$Q(\mathbf{z}, \boldsymbol{\xi}) = \prod_{n=1}^{N} q(z_n; \boldsymbol{\gamma}_n) \prod_{k=1}^{K} q(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_{\theta_k}) q(v_k; \boldsymbol{\gamma}_{v_k}), \quad (4)$$

where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_n, \boldsymbol{\gamma}_{\theta_k}, \boldsymbol{\gamma}_{v_k}\}$ are the variational parameters of $q(z_n; \boldsymbol{\gamma}_n)$ $q(v_k; \boldsymbol{\gamma}_{v_k})$ and $q(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_{\theta_k})$, namely, the multinomial, Beta and exponential family distributions (again, Gaussian for $\boldsymbol{\theta}_k = \boldsymbol{\mu}_k$, and Gaussian-Wishart for $\boldsymbol{\theta}_k = \boldsymbol{\Sigma}_k$). It is worth noting that this is a *truncated stick-breaking* representation ($K < \infty$) only with respect to the variational distribution, while the model still is a full DPM [3].

By inserting Eq. 4 into Eq. 2, approximate inference is achieved by alternating optimization of the free energy over $Q(\mathbf{z})$ and $Q(\boldsymbol{\xi})$ [2]. In particular it has been shown [3] that a simple coordinate ascent algorithm can be set to maximize the lower bound, by iteratively updating the variational parameters $\boldsymbol{\gamma}$. Due to space limitations, we refer to [3] for detailed derivation of parameter update equations.

At convergence, the variational algorithm provides us with two kinds of informations: i) the approximated evidence $p(\mathbf{x}_t)$, which we will use to compute the denominator of Eq. 1, thus, *the low-level saliency*; ii) the *segmented regions* (clusters), where each point has been assigned a label z_n, together with cluster parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Note that the number of components is automatically determined.

In the chosen spatio-temporal representation, the most likely component (corresponding to the largest cluster) will be the one modeling background regions, with near-zero velocities; on the contrary, the regions of interest will be those with: 1) significant mean velocity $\tilde{\mu}_k^{vel}$; 2) large spatial support, with respect to other spatio-temporal blobs that have non null velocity but tiny space/time support, and mostly representing vari-

ations due to noise. Then, the proto-object feature likelihood of Eq. 1 is computed as

$$p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{A}_t, \mathcal{O}) = \frac{exp\{(\mathbf{x}_{n,t}^{vel} - \tilde{\mu}_k^{vel})^2\}}{\sum_n exp\{(\mathbf{x}_{n,t}^{vel} - \tilde{\mu}_k^{vel})^2\}}. \quad (5)$$

Denote $\mathcal{R} = \{R_1, \ldots, R_S\}$ the spatial regions with $\tilde{\mu}_k^{vel}$ velocity. In order to compute the probability of focusing on position $\mathbf{r}_t$ conditioned on object appearance $p(\mathbf{r}_t|\mathcal{A}_t, \mathcal{O})$, we simply determine the maximum spatial support of these regions, $M_R = \max\{|R_1|, \ldots, |R_S|\}$, and to each region $R_i$we assign an importance weight $w_i = |R_i|/M_R$. Then, for all regions in $\mathcal{R}$ we compute

$$p(\mathbf{r}_t|\mathcal{A}_t, \mathcal{O}) = w_i \mathcal{N}(\mathbf{r}_t; \mu_{R_i}, \Sigma_{R_i}), \mathbf{r}_t \in R_i, \quad (6)$$

where μ_{R_i} is the center of mass - the center of the FOA - and covariance Σ_{R_i} is a diagonal matrix, $\Sigma_{R_i}(1,1), \Sigma_{R_i}(2,2)$ encoding the width and the height of the region, determining the spread of the FOA.

4. Simulation

Simulation has been performed on the public *BEHAVE Interactions Test Case* [1], a data-set which comprises videos of people acting out various interactions, under varying illumination conditions and spurious reflections due to camera fixed behind a window. An illustrative example, which is representative of results achieved on such data-set, is provided in Fig. 1.

Each original RGB frame (640×480 pixels) of the sequence is represented in CIE-Lab color space, $\mathbf{x}_t^{col} = \{x_t^L, x_t^a, x_t^b\}$ and down-sampled via a 3-level Gaussian pyramid. Optical flow features $\mathbf{x}_{n,t}^{vel}$ have been estimated on the lowest level of the pyramid from the posterior probability $p(\mathbf{x}_t^{vel}|\frac{\partial x_t^L}{\partial \mathbf{r}}, \frac{\partial x_t^L}{\partial t})$ following [8].

For what concerns the variational DPM, the truncation level K was set to 15, and the concentration parameter α initialized to 1. Parameters inferred at time $t-1$ were used to initialize variational parameters on frame t. The variational algorithm finds the modes of the predictive distribution after $6-7$ iterations on the average, and the saliency distribution uses $3-4$ modes to represent the observed scene.

Differently from other works on attentive vision, the chosen FOA from the distribution has a time varying scale (due to $p(\mathbf{r}_t|\mathcal{A}_t, \mathcal{O})$), which is a function of the appearance of the segmented moving blobs. Fig. 1 bottom row, shows how the FOA switches from the first two "runners" to the two following behind, mostly due to velocity change; then, later on, it stabilizes on the center of a fighting group. It is also worth noting the noisiness of the low-level saliency map (second row) with respect to the posterior distribution map (third row).

The system is currently implemented in plain MATLAB code, with no specific optimizations. As regards actual performance, 90% of the execution time is spent for the variational DPM procedure, which takes an average elapsed time of 9 secs per frame, running on a 2 GHz Intel Core Duo processor, 2 GB RAM, under Mac OS X 10.4.11.

5. Final remarks

We have addressed the issue of joint segmentation and saliency computation in dynamic scenes, using a mixture of Dirichlet processes, as a sound basis for computational modeling of object-based visual attention. The idea of using mixture modeling for low-level saliency was first proposed in [9], but limited to classic finite mixtures, in the context of static images and without addressing the issue of segmentation. Very recently image segmentation in a nonparametric Bayesian setting has been proposed [6], but the cost of the MCMC sampling adopted for inference and learning is critical for dynamic scenes. In this respect, variational techniques have proven to be a viable tool for video analysis. Beyond the theoretical interest in modeling object-based visual attention and related problems [7, 4], preliminary results obtained show that the approach proposed here can be fruitful for most recent trends in applications like foveated video coding, active video-surveillance, interaction analysis.

References

[1] BEHAVE Interactions Test Case Scenarios. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, USA, 2006.

[3] D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2005.

[4] J. Driver, G. Davis, C. Russell, M. Turatto, and E. Freeman. Segmentation, attention and phenomenal visual objects. *Cognition*, 80(1-2):61–95, 2001.

[5] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems 19*, pages 1–8. MIT Press, 2006.

[6] P. Orbanz and J. Buhmann. Nonparametric Bayesian Image Segmentation. *Int. J. of Comp. Vis.*, 77(1–3):24–45, 2008.

[7] B. Scholl. Objects and attention: the state of the art. *Cognition*, 80(1-2):1–46, 2001.

[8] E. Simoncelli, E. Adelson, and D. Heeger. Probability Distributions of Optical Flow. In *Proc. CVPR*, pages 310–315. IEEE Computer Society, 1991.

[9] A. Torralba. Modeling global scene factors in attention. *J. Opt. Soc. Am. A*, 20(7):1407–1418, 2003.