Bayesian Analysis (2018) **13**, Number 1, pp. 85–113

Latent Marked Poisson Process with Applications to Object Segmentation

Sindhu Ghanta*, Jennifer G. Dy†, Donglin Niu‡, and Michael I. Jordan§

Abstract. In difficult object segmentation tasks, utilizing image information alone is not sufficient; incorporation of object shape prior models is necessary to obtain competitive segmentation performance. Most formulations that incorporate both shape and image information are in the form of energy functional optimization problems. This paper introduces a Bayesian latent marked Poisson process for segmenting multiple objects in an image. The model takes both shape and image feature/appearance into account—it generates object locations from a spatial Poisson process, then generates shape parameters from a shape prior model as the latent marks. Inferentially, this partitions the image: pixels inside objects are assumed to be generated from an object observation/appearance model and pixels outside objects come from a background model. The Poisson process provides (non-homogeneous) spatial priors for object locations and the marks allow the incorporation of shape priors. We develop a hybrid Gibbs sampler that addresses the variation in model order and nonconjugacy that arise in this setting and we present experimental results on synthetic images and two diverse domains in real images: cell segmentation in biological images and pedestrian and car detection in traffic images.

Keywords: spatial Poisson process, segmentation, Bayesian nonparametrics, object detection.

1 Introduction

Computer vision is a field that is deeply concerned with probabilistic inference. The problem of processing an image to uncover the physical causes of a light field is an inverse problem (Marr, 1982) and visual inference problems are naturally formulated within a Bayesian paradigm (Zhu et al., 1998). The spatial–temporal structure in images presents abundant opportunities for the exploitation of prior knowledge; indeed, models from spatial statistics have long provided a fruitful interaction between computer vision and Bayesian analysis. In particular, models based on Markov random fields (Geman and Geman, 1984) have been used to capture spatial dependencies and models based on the Poisson process (Gelfand et al., 2009) naturally express uncertainties regarding the numbers and locations of objects in images. The latter connection has also helped spur the development of Bayesian nonparametric models aimed at image processing and computer vision problems, with random partitions and random measures providing further realism regarding complex properties of images; examples include (Orbanz and

*Parallel Machines, sindhu.ghanta@gmail.com
†Northeastern University, jdy@ece.neu.edu
‡Northeastern University, donglin.niu@gmail.com
§University of California, Berkeley, jordan@cs.berkeley.edu

© 2018 International Society for Bayesian Analysis https://doi.org/10.1214/16-BA1035

Buhmann, 2008; Blei and Frazier, 2011; Ren et al., 2011) and (Sudderth and Jordan, 2009).

Missing from much of the existing Bayesian spatial statistics literature has been an explicit role for objects and their shapes. Shape models can be a powerful constraint in vision problems. Even if the focus is solely segmentation, methods that combine shape prior information with image feature information can improve segmentation performance compared to methods that utilize image feature information alone (Chan and Zhu, 2005; Cootes et al., 1995; Shotton et al., 2006). The incorporation of shape priors is popular in level-set methods and optimization-based formulations (Chan and Zhu, 2005; Cremers et al., 2007; Leventon et al., 2000; Huang and Metaxas, 2008). There have also been proposals that incorporate shape priors in graph-based segmentation algorithms (Kumar et al., 2005). These formulations have generally been designed to discover a single object in a visual scene (for example, segmentation of the heart in a computed tomography image). However, one may be interested in detecting multiple occurrences of similar objects or patterns in an image (for example, cells in an image) or in segmentation of multiple objects/patterns in an image that may overlap (Vese and Chan, 2002; Vu and Manjunath, 2008). Such multiple-object segmentation is the focus of the current paper.

We introduce a Bayesian latent marked Poisson process model for segmenting multiple objects/patterns in an image that takes into account both shape prior information and image feature information. We utilize a spatial Poisson process (Baddeley, 2007; Gelfand et al., 2009) as a nonparametric prior for the number of objects along with their locations. Each object has a corresponding shape generated from a shape prior model (Cootes et al., 1995). This provides a partitioning of the image. When locations and shapes of objects are determined, the pixels that are inside objects are assumed to be generated from an object image feature/appearance model and pixels outside objects are assumed to come from a background model.

Not only does a spatial Poisson process allow us to model the number of objects along with their locations, but it also provides a natural model for spatial context information. Often, one has domain knowledge about parts of an image scene that have a high/low probability of occurrence of an object. For example, in traffic surveillance images, cars will be found only on roads and their probability of occurrence in other parts of the image is very low. This information is captured naturally by the Poisson intensity parameter. Traditionally in computer vision, context is represented in the form of pairwise geometric relationship between different objects (He et al., 2004); here we are able to introduce a spatial context prior in the form of the non-homogeneous Poisson intensity parameter. Moreover, posterior inference for the non-homogeneous Poisson intensity provides a probability map exhibiting regions where high or low object concentration occur.

Our model is based on the formalism of marked Poisson processes (MPPs). In particular, the shape associated with each object location is treated as the mark associated with the location. Bayesian analysis of marked Poisson processes has been applied to a variety of domains (Xiao et al., 2015; Taddy, 2010; Rotondi and Varini, 2003). There exists an extensive literature on the use of general marked point processes for detection

of objects (Descombes and Zerubia, 2002; Baddeley and Lieshout, 1993); however, this literature has focused on simple shapes and synthetic images. Exceptions that focus on real-world images include (Lafarge et al., 2010; Rue and Hurn, 1999; Baddeley and Van Lieshout, 1992). These work differ from ours in that the marks in their case are *observed* random variables; in our case the shape marks are *latent* random variables. In addition, these methods discourage object overlap between objects, either using a hard-core process or a Strauss process. In contrast to these approaches, our model utilizes a marked Poisson process with no constraints on overlap between objects.

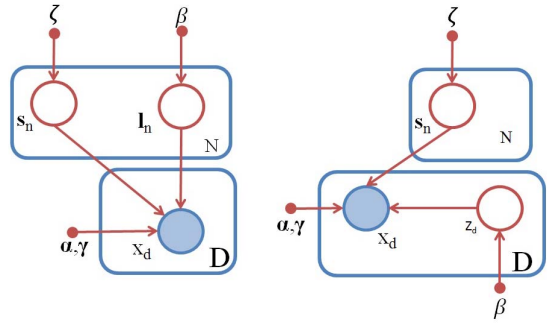
A challenge associated with utilizing a Poisson process for object detection is the changing model order as each random outcome can have different number of objects. Typically, a reversible-jump Markov chain Monte Carlo (RJMCMC) sampling strategy is used to address this problem (Ge and Collins, 2009; Zhou et al., 2014), but this strategy can be a computational bottleneck. Our work introduces a novel inference strategy based on a hybrid Gibbs sampler. We take advantage of the finite resolution (number of pixels or voxels) in computer vision problems. Instead of performing inference on the continuous location variable, we take advantage of the discretization to obtain a Gibbs sampler. Despite the discretization, our approach still computes posterior probabilities with respect to the underlying Poisson process.

The paper is organized as follows. In Section 2, we describe the basic specification of our model for generating and segmenting multiple objects/shapes of the same type. We describe the challenges associated with inference on the model and our approach to ameliorating these challenges in Sections 3 and 4. Then, in Section 5.1, we explain how to extend the model when we have multiple objects/shapes of different types. In Section 6, we discuss the results of experiments on one synthetic and two real-world applications: Cell image segmentation and traffic (car and pedestrian) detection is presented. These experiments exhibit the use of our model in unsupervised mode (for the cell data) and in supervised mode (for the car and pedestrian scenes). We present our conclusions in Section 7.

2 Model Specification

An image is represented by a data matrix $\mathbf{X} \in \mathbb{R}^{D_r \times D_c}$, where D_r and D_c are the number of rows and columns respectively (total number of pixels, $D = D_r \times D_c$). Each element in this matrix is a pixel, represented by a scalar in the case of a gray-scale image, a vector in the case of RGB color images and in general a feature vector capturing local image characteristics (e.g., wavelet, Fourier coefficients, co-occurrence or histogram features). To simplify notation, we assume two-dimensional images; however, the concepts here can be extended to three-dimensional and higher-dimensional images.

Our model takes the following form. We assume that N objects are generated at locations $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_N]$ as a draw from a Poisson process with Poisson intensity parameter $\beta(\tau)$, where τ represents the 2D image plane. Here $\mathbf{l}_n = [l_{n,r}, l_{n,c}]$ is a pair of coordinates. Note that N is random. A shape parameter, $\mathbf{S}$ for each object, with origin/center at $\mathbf{l_n}$, is generated from a shape prior distribution with hyper-parameter $\boldsymbol{\zeta}$.



(a) Original model (b) Modified model

Figure 1: A graphical representation of the proposed model.

The shape parameters, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]$ define the object contour and thus a partitioning of the image: Observation data (the feature representation of each pixel) inside and outside the contour are generated from foreground and background appearance models with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively. Let all the hyper-parameters be represented by $\boldsymbol{\theta} = [\beta(\tau), \boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}]$.

The overall probability model is given as follows:

$$p(\mathbf{L}, \mathbf{S}, \mathbf{X}|\boldsymbol{\theta}) \quad = \quad \Big[\underbrace{p(\mathbf{L}|\beta(\tau))}_{\text{Poisson process}} \cdot \underbrace{p(\mathbf{S}|\mathbf{L}, \boldsymbol{\zeta})}_{\text{Shape distribution}} \Big] \times \underbrace{p(\mathbf{X}|\mathbf{L}, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma})}_{\text{Appearance/Likelihood}} , \tag{1}$$

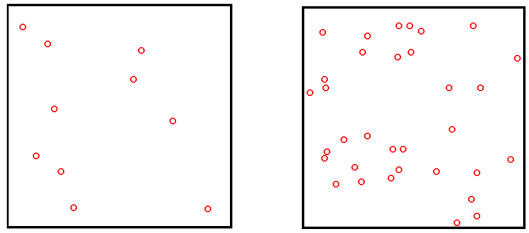
where the *appearance model*—the probability of the pixel data given the hyper-parameters—takes the following form:

$$p(\mathbf{X}|\mathbf{L}, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \quad = \quad \prod_{d=1}^{D} [p(x_d|\boldsymbol{\alpha})]^{I_d} [p(x_d|\boldsymbol{\gamma})]^{1-I_d}, \tag{2}$$

where I_d is a random variable that takes a value 1 when the pixel d belongs to an object and 0 when the pixel does not belong to an object (i.e., it belongs to the background). Note that this random variable is a deterministic function of $\mathbf{L}$ and $\mathbf{S}$. A graphical representation of the model is shown in Figure 1(a).

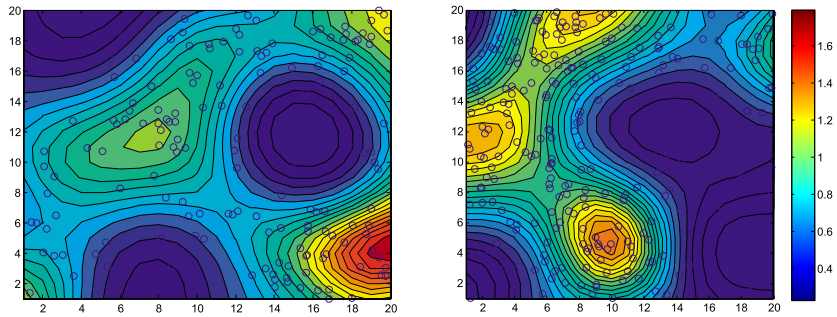
2.1 Spatial Poisson Process Prior

A sample from a spatial Poisson process consists of a random number of points at random locations on the 2D plane based on the underlying intensity function $\beta(\tau)$. When $\beta(\tau) = \beta$, the process is called *homogeneous*. When $\beta(\tau)$ varies with τ, we say it is *non-homogeneous*. A Poisson process is defined in general on a locally compact metric space $\mathcal{S}$ with intensity measure Λ (which is finite on every compact set and has no atoms) as a point process on $\mathcal{S}$ such that, for every compact set $\mathcal{B} \subset \mathcal{S}$, the count



(a) $\Lambda(\mathcal{B}) = 10$ (b) $\Lambda(\mathcal{B}) = 30$

Figure 2: A realization of a homogeneous Poisson process.



(a) (b)

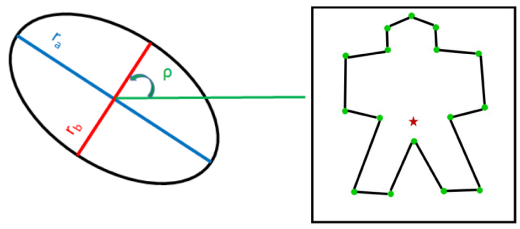
Figure 3: A realization of a non-homogeneous Poisson process.

$N(\mathcal{B})$ has a Poisson distribution with mean $\Lambda(\mathcal{B})$. If $\mathcal{B}_1, \ldots, \mathcal{B}_m$ are disjoint compact sets, then $N(\mathcal{B}_1), \ldots, N(\mathcal{B}_m)$ are independent (Baddeley, 2007). In a spatial setting, $\mathcal{S} = R^2$ and $\Lambda(\mathcal{B}) = \int_{\mathcal{B}} \beta(\tau) d\tau$.

An example of a draw from a homogeneous Poisson prior is shown in Figure 2. The number of objects that can be expected in the 2D plane in this case is given by the scalar Poisson intensity parameter. The distribution of these objects in the 2D plane is uniform. A draw from a non-homogeneous Poisson process is shown in Figure 3. The colormap of contours in Figure 3 represents the value of Poisson intensity at each point in the 2D plane. Red and blue colors indicate a high and low Poisson intensity value respectively. One should expect to see more objects in the region containing red color compared to the regions that contain blue color.

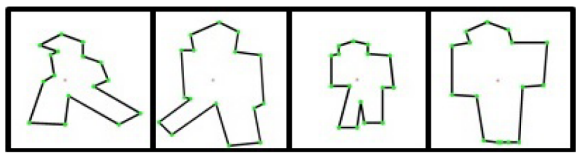
2.2 Shape Prior

In many vision problems, the objects of interest have distinct shape characteristics. Here, we wish to incorporate such knowledge as a shape prior. There are two possible scenarios for building a shape model: one is when the shape model has a known simple parametric form, and the other is when the shape model is complex.

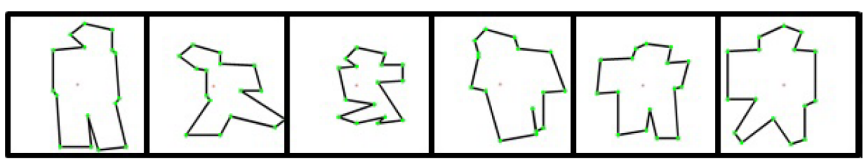


(a) (b)

Figure 4: (a) Shape parameters of an ellipse: r_a is the major axis, r_b is the minor axis, and ρ is the orientation. (b) Landmark points (green) for shape prior of pedestrians. Red star is the origin/center point.



(a)



(b)

Figure 5: (a) Shapes extracted from training images. (b) Synthetic images generated from the inferred shape prior.

Simple Known Parametric Model In a simple example of our framework, shape is described by a fixed collection of parameters, $\boldsymbol{\theta}$. For example, we may consider elliptical shapes (a common simplification used in modeling cells, 2D image slices of blood vessels and the human face). Ellipses can be described by three parameters: the major axis r_a, the minor axis r_b and the rotation angle ρ (which implies $\mathbf{s} = [r_a, r_b, \rho]$). One could assume uniform priors on these parameters (which are hyper-priors of the overall model) or any other appropriate distribution based on either domain knowledge or training data. We also assume that the major and minor axes are independent of the rotation angle. Figure 4(a) shows the parameters of an ellipse shape model.

Complex Shape Model There are several ways to build complex shape models. In this paper, we adapt a simple approach that utilizes a landmark distribution model for shape (Cootes et al., 1995). In this approach, a complex shape is parameterized by landmark points along the boundary of the object. Landmark points are representative points labeled by a domain expert or annotator who aims to capture points that are

significant for the application, such as the highest point of an object, curvature extrema, or points along a boundary. Figure 4(b) shows an example of a 16-point model of the boundary of a person. Each point on the object has a row and column element, giving rise to a 32-dimensional vector. Therefore, there are m parameters for a $m/2$ point model. Each shape, $\mathbf{s}$, is represented by an m-dimensional vector with elements comprising the row and column location of each landmark point with respect to an origin (the red star in Figure 4(b)).

A prior shape model is constructed as follows. Given a set of P training shape templates (see Figure 5(a)), the shapes are first aligned by adjusting the scale, translation and rotation that minimizes a weighted sum of squared distances between equivalent points on each shape template as described in (Cootes et al., 1995). Then, for each shape, $\mathbf{s}_i$, where $i = [1, ..., P]$, we calculate its deviation, $d\mathbf{s}_i$, from the mean, $\bar{\mathbf{s}}$:

$$d\mathbf{s}_i = \mathbf{s}_i - \bar{\mathbf{s}} \tag{3}$$

and the covariance matrix $\mathbf{C}$, using

$$\mathbf{C} = \frac{1}{P} \sum_{i=1}^{P} d\mathbf{s}_i d\mathbf{s}_i^T. \tag{4}$$

One can sample from a Gaussian distribution with this mean and covariance matrix to generate a synthetic shape. Figure 5(b) shows examples of synthetic images generated from the inferred shape prior. To take into account variation in scale and rotation, additional variables that represent scaling a and rotation angle ρ can be introduced and assumed to be generated from uniform or gamma distributions depending on the application. We also assume that the shape, scale and rotation angle are independent of each other. For notational convenience, we represent the set of all parameters representing a complex shape—$\mathbf{s}$, a and ρ—by $\mathbf{s}$. When the number of landmark points is large, principal component analysis (PCA) can be used for dimensionality reduction as described in Cootes et al. (1995). Note that we have chosen this simple landmark-based shape model as a concrete illustration; within our framework one can deploy more powerful and flexible shape models (see, e.g., (Bhattacharya and Dunson, 2010) which provides a Bayesian nonparametric model for shape).

2.3 Feature/Appearance Model

We assume that data (pixels), $\mathbf{x}_d$, inside a shape contour are generated from an object feature/appearance model, $p(\mathbf{x}_d|\boldsymbol{\alpha})$, and data outside any shape boundary are generated from a background model, $p(\mathbf{x}_d|\boldsymbol{\gamma})$. As indicated earlier, an observation pixel, $\mathbf{x}_d$, can be represented by its gray-scale value or color values. For textured objects, $\mathbf{x}_d$ would be represented by a feature vector based on image characteristics (e.g., wavelet, Fourier, co-occurrence, histogram features). One can utilize a multinomial distribution over pixel intensities or a truncated Gaussian distribution based on the features used for the application. Detailed examples of foreground/object and background models are provided in our experiments.

3 Leveraging the Finite Nature of Image Pixels

Finding the posterior distribution of the number of objects, their locations and corresponding shapes in a Bayesian framework for a continuous location parameter is challenging due to the need to compare different model orders. In addition, we must treat non-conjugate priors, which result from the fact that the likelihood depends on partitioning based on shape boundaries. A standard approach to providing posterior inference in such a setting involves employing Metropolis–Hastings sampling within the reversible jump Markov chain Monte Carlo (RJMCMC) framework. However, it is advantageous to use Gibbs sampling as it samples directly from the conditional posterior distribution, does not involve rejections and one does not have to worry about choosing an appropriate proposal distribution. However, the basic Gibbs sampling strategy does not handle trans-dimensional jumps and hence cannot be used directly for inference in our setting.

To make Gibbs sampling applicable to our setting, we take advantage of the fact that there are only a finite number of pixels in an image and that we only care about location information up to a pixel resolution. For example, it is enough for us to know that an object center occurs in an area $d\tau$ covered by a pixel rather than the exact continuous value of the location. We introduce a latent variable $Z = [z_1, \ldots, z_D]$, where each z_d is a discrete variable representing the number of object centers at a particular pixel location d. Note that the number of object centers that can occur at any given pixel is unbounded.

According to the definition of a spatial Poisson process, for disjoint sets $B_1, \ldots, B_N$, the number of objects lying in these sets, $N(B_1), \ldots, N(B_N)$, are independent random variables. In our case, the disjoint sets are individual pixels. This value depends only on the underlying Poisson intensity parameter in the area covered by the pixel. Let this value of Poisson intensity parameter at each pixel be given by $\beta_d = \int_{p-1}^{p} \int_{q-1}^{q} \beta(u, v) du dv$, where p and q are the indices of the pixel and u and v are continuous random variables that represent location in the 2D plane. z_d has a Poisson distribution with parameter β_d. The total number of objects N in the image has a Poisson distribution with parameter I, where I is the integral of the intensity parameter of the Poisson process for the entire image given by $I = \int_\tau \beta(\tau)$.

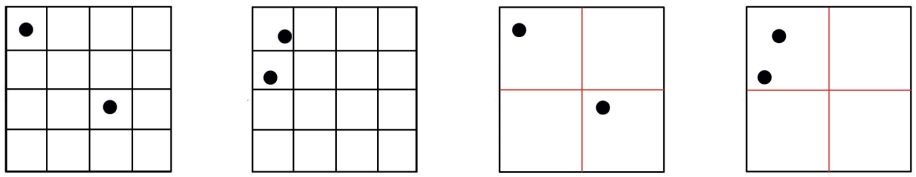
Given the number of objects N in the image, the joint distribution is a multinomial distribution:

$$p(Z|N) = \frac{p(Z)}{p(N|I)} = \frac{\prod_{d=1}^{D} \frac{\exp(-\beta_d)\beta_d^{z_d}}{z_d!}}{\frac{\exp(-I)I^N}{N!}} = \frac{N!}{z_1! \ldots, z_d!} \left(\frac{\beta_1}{I}\right)^{z_1} \cdots \left(\frac{\beta_D}{I}\right)^{z_D} \quad (5)$$

and our overall model takes the following form:

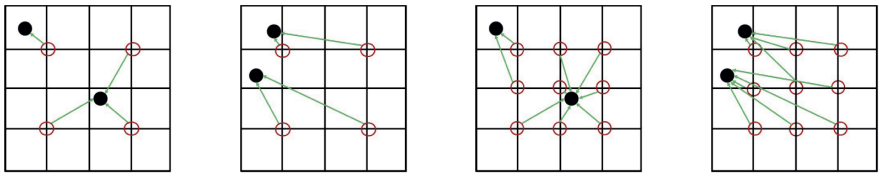
$$p(Z, \mathbf{S}, \mathbf{X}|\boldsymbol{\theta}) = p(Z|N)p(\mathbf{S}|\mathbf{L}, \boldsymbol{\zeta})p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \quad (6)$$

as depicted in Figure 1(b). The random number of objects in the image can be represented by $\sum_{d=1}^{D} z_d$.



(a) −6.158 (b) −6.158 (c) −3.386 (d) −4.079

Figure 6: (a) Location Set-1 and (b) Location Set-2 where the image is divided into 16 regions. (c) and (d) show the division of space into 4 regions. Value of Poisson log-likelihood for each configuration is reported below the corresponding figure.



(a) −5.688 (b) −8.816 (c) −8.67 (d) −13.03

Figure 7: (a) Location Set-1 (b) Location Set-2 with $M = 4$ and (c), (d) show the same set with $M = 9$. Value of Poisson log-likelihood for each configuration is reported below the corresponding figure.

4 Inference

Posterior inference for the proposed model involves calculating the posterior distribution over the variables $(Z, \mathbf{S})$ given the observation $\mathbf{X}$. We develop a Gibbs sampling framework to perform this inference. A naive application of Gibbs sampling would compute the following conditional probability:

$$p(z_d|Z_{-d}, \mathbf{S}, \boldsymbol{\theta}) \propto p(z_d|Z_{-d}, \beta(\tau))p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma}). \tag{7}$$

One of the properties of the spatial Poisson process is that the distribution of points in the 2D plane does not depend on the presence of other points; the occurrence of a point at any location depends only on the underlying Poisson intensity parameter. Sampling z_d in this way does not capture the complete spatial randomness property of the Poisson process for a given set of object locations. This happens due to discretization of the space. The random variable Z has a distribution which is a functional of the spatial Poisson process. However, it is not possible to invert this functional by considering the values of z_d independently.

As an illustration, consider Figures 6(a) and 6(b) as two sets of observations and the goal of calculating the likelihood of these observations as the outcome of a uniform Poisson process with prior $\int \lambda(\tau) = 2$. The number of points in both sets is the same. Intuitively, one expects the likelihood value to be low when the observations are clustered

and high when they are spread apart. When the image is discretized into 16 parts; both observations will have the same likelihood according to $log(\prod_{d=1}^{16} p(z_d|\lambda)) = -6.158$. If instead the image is discretized into 4 pixels/parts, the grid in Figure 6(c) (-3.386) has higher likelihood compared to Figure 6(d) (-4.079). The relative likelihood value depends on the number of parts the image has been divided into and $\prod_{d=1}^{D} p(z_d|\lambda)$ cannot capture this spatial randomness as every value of z_d is independent (Skellam, 1952). In our case, digital images are naturally discretized into pixels and introducing the latent variable z_d for the presence of an object center cannot distinguish clustered objects from uniformly distributed objects (assuming homogeneous Poisson intensity). A similar issue is encountered in maximum likelihood estimation of the Poisson intensity parameter where distance methods are proposed to ameliorate this problem (Skellam, 1952; Pollard, 1971).

Distance methods assume a grid of M points evenly distributed on the surface of the image (denoted by hollow red circles). The probability of finding the nearest object/observation point at a distance r_m from the mth grid point is given by $p(r_m) = (2\pi r_m \beta)\exp(-\pi r_m^2 \beta)$. Let the distances of nearest object occurrence from this grid of M points be denoted by $r_1, r_2, \ldots r_M$. For example, Figures 7(a) and 7(b) assume four grid points in the image. We will describe how the number M can be chosen and its associated trade-offs later in this section. Extending the likelihood equation for non-homogeneous case given in Pollard (1971), we derive the following equation for sample distances $r_1, r_2, \ldots r_M$:

$$p(R) = \prod_{m=1}^{M} \sum_{d=1}^{D} (2\pi r_m)(p_d \beta_d)\exp(-\pi r_m^2 \beta_d), \tag{8}$$

where $R = [r_1, \ldots, r_M]$ and p_d is the proportion of image a pixel region represents, which is equal to $1/D$. Note that this equation assumes that r_m are independent and we inherit the assumptions in Pollard (1971) that the region within the nearest neighboring point to a grid point has a uniform intensity β_d and the boundary problem is ignored. For the homogeneous case, this equation reduces to $p(R) = (2\pi\beta)^M \exp(-\pi\beta \sum_{m=1}^{M} r_m^2)(r_1 r_2 \ldots r_M)$.

The likelihood value in terms of the distances R given in (8) captures the spatial randomness for a given set of object locations and intensity value. For example, $p(R)$ in Figure 7(a) (-5.688) is higher compared to Figure 7(b) (-8.816). Even in a case where more number of grid points are chosen ($M = 9$), the likelihood of the configuration in Figure 7(c) (-8.67) is higher than the configuration of Figure 7(d) (-13.03). This is a desirable property to determine the configurations that are more likely to occur than others. Note that Skellam (1952); Pollard (1971) introduced distance methods for point estimation of the Poisson intensity parameter, $\beta(\tau)$, using a maximum likelihood criterion, given observed Poisson locations. We, on the other hand, are interested in exploring the possible object locations z_d and their likelihood given the intensity value. We view R as an auxiliary variable (as shown in (9)) that captures the spatial randomness of the objects in the image. Based on this augmentation, we derive a Gibbs sampler as follows:

$$p(z_d, R|Z_{-d}, \mathbf{S}, \boldsymbol{\theta}) \propto p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\theta}) p(z_d, R|Z_{-d}, \beta(\tau)). \tag{9}$$

Note that R is deterministic given Z. Given that we are interested only in z_d for the purposes of inference, we discard the values of R after they are used in the sampling process. We have:

$$p(z_d, R|Z_{-d}, \beta(\tau)) = p(R|Z_{-d}, \beta(\tau))p(z_d|Z_{-d}, R, \beta(\tau)). \tag{10}$$

Note that z_d can take discrete values in the range $[0, \infty)$; i.e., we do not place any restrictions on the number of object centers that can appear in one pixel.

For Gibbs sampling, all the variables except the one being sampled are assumed to be known and constant. This implies that one needs to sample the values of R with only one unknown value z_d. The possible values that R can take given Z_{-d} reduces to two cases (R_o and $R_\emptyset$, if $z_d = 0$ or $z_d > 0$ respectively):

$$p(R|Z_{-d}, \beta(\tau)) = \begin{cases} p(R_o), \textit{if } z_d = 0, \\ p(R_\emptyset), \textit{if } z_d > 0. \end{cases} \tag{11}$$

This value for both cases can be calculated from (8) as R is deterministic given the values of Z. All the cases where $z_d > 0$ give the same $R = R_\emptyset$ value as the distance between a fixed point and its nearest object remains the same irrespective of the number of objects present in that pixel. This leads to the following conditional probabilities:

$$\begin{aligned} p(z_d = k, R|Z_{-d}, \beta(\tau))_{k=0} &= (1/H)p(R_o)p(z_d = 0|Z_{-d}, R_o) \\ &= (1/H)p(R_o)e^{-\beta_d}, \\ p(z_d = k, R|Z_{-d}, \beta(\tau))_{k\neq 0} &= (1/H)p(R_\emptyset)p(z_d = k|Z_{-d}, R_\emptyset) \\ &= (1/H)p(R_\emptyset)(e^{-\beta_d}{\beta_d}^k/k!), \end{aligned} \tag{12}$$

where

$$\begin{aligned} H &= \sum_{k=0}^{\infty} p(R)p(z_d = k, R_k|Z_{-d}) \\ &= p(R_o)e^{-\beta_d} + p(R_\emptyset)\sum_{k=1}^{\infty} \frac{e^{-\beta_d}\beta_d^k}{k!} \\ &= p(R_o)e^{-\beta_d} + p(R_\emptyset)(1 - e^{-\beta_d}). \end{aligned} \tag{13}$$

To set the number of grid points M for calculating R, we notice that if M is too sparse compared to the number of objects in the image, most values of z_d will have no influence on the value of R and $p(z_d, R|Z_{-d}, \beta(\tau)) = e^{\beta_d}\beta^{z_d}/z_d!$; this ceases to capture the spatial randomness of the objects. On the other hand, having the grid points that are too dense will result in increased computational expense. As a default, we chose M to be ten pixels apart in the image.

Step 1 of Algorithm 1 provides the pseudocode for sampling z_d based on three cases: deletion of a point, remaining at the same state and addition of a point. This involves calculating the likelihood for all the possible cases followed by normalization giving a

Algorithm 1 Gibbs Sampling for a Marked Poisson Process.

Initial: Sample the variables $\{z_d\}_1^D$ for every pixel and $\{\mathbf{s}_n\}_1^N$ for every object, where number of objects is given by $N = \sum_1^D z_d$. Given a set of these values from the previous iteration, sample a new set as follows:

Step 1: $Z = Z^t$ and $\mathbf{S} = \mathbf{S^t}$. For $d = 1, \ldots, D$, draw a new sample for z_d from the following probabilities:

$$p(z_d = k, R|Z_{-d}, \mathbf{S}, \boldsymbol{\theta}) \propto \begin{cases} p(z_d = k, R|Z_{-d}, \beta(\tau))p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \textit{if } k \leq k^t \\ p(z_d = k, R|Z_{-d}, \beta(\tau)) \int p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma})p(\mathbf{s}_{N+1}|\boldsymbol{\zeta})ds, \\ \textit{if } k = k^t + 1, \end{cases}$$

if $k = k^t + 1$, sample a new parameter s_{N+1} from $p(s_{N+1}|\zeta)p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ and $N = N + 1$; if $k < k^t$, delete an object and reduce dimension of $\mathbf{s}$, $N = N - 1$.

Step 2: For $n = 1, \ldots, N$.

Sample the shape parameter of each object: $p(\mathbf{s}_n|Z, \mathbf{S}_{-n}, \boldsymbol{\theta}) \propto p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\theta})p(\mathbf{s}_n|\boldsymbol{\zeta})$.

Step 3: $t = t + 1$, $Z^t = Z$ and $\mathbf{S^t} = \mathbf{S}$.

REPEAT Steps 1, 2 and 3 until convergence.

Output: We assume convergence when first order statistics vary below a pre-set threshold ϵ: save and report the mean values of all object's location and shape.

discrete probability distribution. The final value of z_d is sampled from this distribution. This is similar to the Gibbs sampling framework for Dirichlet process mixtures.

The shape variables are assumed to be independent of each other. Sampling the shape variable is given as follows and constitutes Step 2 of Algorithm 1:

$$p(\mathbf{s}_n|\mathbf{S}_{-n}, \mathbf{L}, \boldsymbol{\zeta}) \propto p(\mathbf{X}|Z, \mathbf{S}, \boldsymbol{\theta})p(\mathbf{s}_n|\boldsymbol{\zeta}). \tag{14}$$

Because our prior is not conjugate to the observation likelihood, we use sampling–resampling to approximate a sample from the posterior (Smith and Gelfand, 1992). We sample a fixed number (300) of shapes from the prior distribution and then calculate the corresponding likelihood for each case. We then sample a shape from these candidate shapes by sampling from a discrete distribution (for each shape model) whose weights are the normalized scores of (14).

Implementation Details Initialization of $\{z_d\}_1^D$ in the algorithm can be done randomly. However, to speed-up the process, we initialize using a simple heuristic. For intensity based features, we apply a threshold on the infinity norm of the vector. The default threshold used in our experiments was just the midpoint between the maximum and minimum value. We then use the center of the connected components of this binary image to initialize the object centers. More details are provided in the experiments section. Inference using Gibbs sampling on this model is expensive due to the introduction of the latent variable Z, whose size is equal to the total number of pixels. For this reason, we use a hybrid Gibbs sampler that combines a single site and a blocked Gibbs sampler. Blocks of pixels in the image are sampled for the value of Z at first. Blocks that show very low likelihood (compared to the likelihood of presence of an object observed so far)

for the occurrence of an object in it are removed from further sampling. In other words, these are the pixels whose appearance is closer to the background than the foreground model. This allows us to get rid of obvious non-object center locations. After this step, the rest of the image is sampled for Z at every pixel location. The algorithm has been implemented in MATLAB and takes an average of 10 minutes per $[1024 \times 1344]$ image on a 12 GB RAM, 2.4 GHz computer. This limitation of high computational time can be overcome by parallelizing Gibbs sampling (Terenin et al., 2015; Gonzalez et al., 2011). Sampling of location and the corresponding shape parameters in different parts of the image can be distributed over several computer nodes. These nodes can communicate back with a master node for global parameter update at constant intervals.

4.1 Inference of the Non-Homogeneous Poisson Prior

Inference of the non-homogeneous Poisson intensity prior given a set of training examples is not straightforward. In this section we present the details regarding learning a non-homogeneous Poisson intensity prior parameter as suggested in Adams et al. (2009).

Let the set of observations in a region τ be denoted by $\{\mathbf{l}_n\}_{n=1}^N$. Given these observations, one needs to infer the underlying Poisson intensity parameter $\beta(\tau)$. One does not know the functional form of the intensity parameter a priori. We thus make use of a Cox process, where the Poisson intensity prior is drawn from another underlying stochastic process. In particular, we utilize a log Gaussian Cox process so as to respect the constraint that the intensity must be nonnegative. That is, we model the intensity as an exponential function of a random realization from a Gaussian process: $\beta(\tau) = \exp(g(\tau))$, where $g(\tau) : R^2 \to R$ is a random scalar function having a Gaussian process prior. This formulation has the following advantages: (a) there is no issue with edge effects, (b) the intensity within a bounded window, given a realization of the process can be predicted using Bayesian methods, (c) higher-order properties take a simple expression and theoretical properties can be easily derived, and (d) the model is amenable to interpretation.

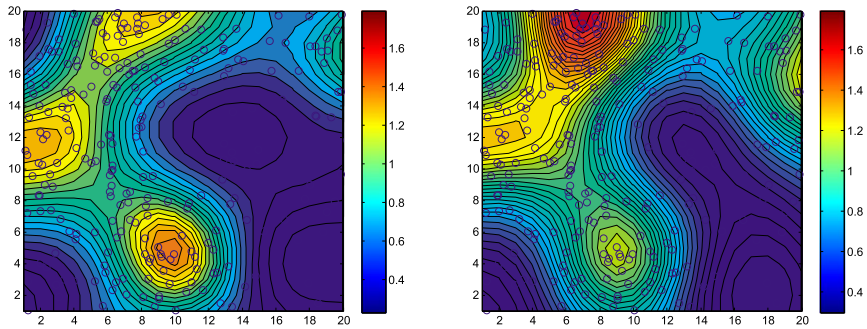
The likelihood for a given set of observations is given by

$$p(\mathbf{l}_{n=1}^N|\beta(\tau)) = \exp\{-\int_\tau d\mathbf{l}\beta(\mathbf{l})\} \prod_{n=1}^N \beta(\mathbf{l_n}), \tag{15}$$

where $\beta(\tau) = \exp(g(\tau))$. Inferring $\beta(\tau)$ given a set of observations requires inference of the underlying function g. The posterior distribution of g is given by

$$\exp(g)|\{\mathbf{l}_n\}_{n=1}^N \sim \frac{\mathcal{GP}(g)\exp\{-\int_\tau \exp(g)d\tau)\}\prod_n \exp g(\mathbf{l}_n)}{\int dg\mathcal{GP}(g)\exp\{-\int_\tau \exp(g)d\tau)\}\prod_n \exp g(\mathbf{l}_n)}.$$

We wish to use an MCMC algorithm to sample from this distribution. The posterior distribution over the Gaussian process g is unfortunately doubly-intractable (due to the presence of an integral over the Gaussian process in both the numerator and denominator of the Metropolis–Hastings acceptance probability) but the intractability can be



(a) Original: An example of a sample from a spatial Poisson process with underlying intensity indicated by color.

(b) Inferred: Poisson intensity functional indicated by contours inferred from observation of points from one sample.

Figure 8: An example of inference of Poisson intensity given a set of observations.

overcome by using an approach due to Adams et al. (2009); Moller et al. (1998, 2006), in which an additional layer of sampling of "fantasy events" is employed such that the integral in the numerator and denominator cancels. In our setting the fantasy events are the locations of new points $\{e\}$ that we generate apart from the given observations $\{\mathbf{l}_n\}_{n=1}^N$ (based on the current and proposed states of g). We use delayed rejection sampling (Green and Mira, 2001) to minimize rejections resulting in faster convergence. An example problem is shown in Figure 8(a).

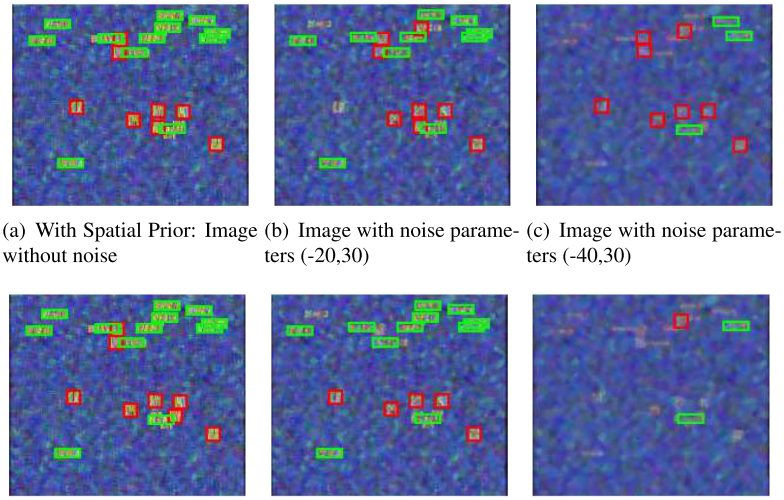
Figure 8(b) shows the inferred intensity given the observations. Ideally, the inferred contours in Figure 8(b) and the actual (ground-truth) contours of Figure 8(a) should be the same. We infer the contours (intensity parameter) given only one sample (one set of observations/points drawn from the Poisson process) and the results are very close. This learned probability map provides us with a spatial context prior indicating the high density and low density spatial regions of object occurrence. Given a set of training images with object location annotations, we infer the intensity parameter $\beta(\tau)$ and utilize this learned probability map as the spatial prior for our model.

5 Extensions to the Model

In this section, we present extensions of the proposed model to the case of multiple categories. We present two types of extensions, one involving multiple categories with the same appearance model and the other involving multiple categories with different appearance models.

5.1 Multiple Categories with the Same Appearance Model

We generally wish to be able to detect multiple categories in an image. For example, one might want to detect both cars and pedestrians from a traffic surveillance video.



(a) With Spatial Prior: Image without noise (b) Image with noise parameters (-20,30) (c) Image with noise parameters (-40,30)

(d) Without Spatial Prior: Image without noise (e) Image with noise parameters (-20,30) (f) Image with noise parameters (-40,30)

Figure 9: Results of inference on synthetic images. Figures (a), (b) and (c) (top) show results from the proposed model with a non-homogeneous spatial Poisson prior. Figures (d), (e) and (f) (bottom) show results from the proposed model without a spatial prior.

Assuming there are C categories of objects in the image, the joint likelihood is given by

$$p(\mathbf{L}, \mathbf{S}, \mathbf{X}|\boldsymbol{\theta}) \quad = \quad \prod_{c=1}^{C} \Big[p(\mathbf{L}_c|\beta_c(\tau) \cdot p(\mathbf{S}_c|\mathbf{L}_c, \boldsymbol{\zeta_c}) \Big] \times p(\mathbf{X}|\mathbf{L}, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \tag{16}$$

where, $\mathbf{L}_c = [\mathbf{l}_1, \ldots, \mathbf{l}_{N_c}]$ and $\mathbf{S}_c = [\mathbf{s}_1, \ldots, \mathbf{s}_{N_c}]$. In this formulation each category has its own Poisson and shape prior parameters. The number of objects in each category is represented by N_c. The appearance prior for all the categories is kept the same. Case study 2 of the experiments section presents an illustration of this scenario. We demonstrate the advantage of using a non-homogeneous Poisson intensity as a prior in a case with multiple categories in the following experiments on synthetic data.

Experiments on Synthetic Data

We test the performance of the proposed method on synthetic images that contain two categories of objects: rectangles and squares. We compare the results to those obtained from a model that is missing the spatial prior.

The goal is to be able to infer the object location along with its category, given the appearance and spatial prior. We add different levels of Gaussian noise to these images to make the detection challenging. Figure 9 shows the segmentation results of our model with incorporation of spatial context through a non-homogeneous Poisson prior as shown by the top figures (a), (b) and (c) for varying noise level and the

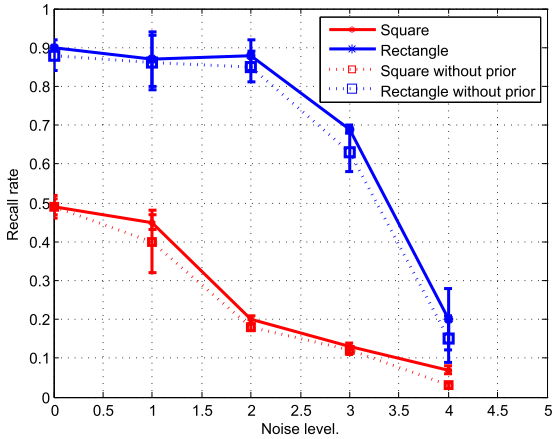


Figure 10: Results of our algorithm on synthetic images.

segmentation results without the spatial prior in the bottom figures (d), (e), and (f). In the absence of noise, notice that the appearance of objects in the image is different from background, thus both models can detect the objects easily. However, when noise is added, a non-homogeneous Poisson prior provides spatial context to help push the likelihood value towards the presence of an object. This proves to be useful in detection of objects under noisy conditions which is true in most real world scenarios.

Five images are generated from one sample of the spatial Poisson process with varying levels of noise. No noise is added to the first image and the rest of the images have a Gaussian noise with parameters mean and variance given by $[-10, 30]$, $[-20, 30]$, $[-30, 30]$ and $[-40, 30]$ respectively followed by smoothing using a median filter of size $[5, 5]$. 50 different images were generated from different spatial Poisson priors. Each image has four additional noisy images associated with it. The performance of the algorithms on a total of 250 synthetic images is reported in Figure 10. We plot the recall rate as a function of noise.

A constant number of iterations (20000) were used for both the models to make a fair comparison. Clearly, the model with a spatial Poisson prior outperforms the model without it. These results are intuitive as the model with a spatial prior will give a very low probability to a model with very few objects. This experiment was also performed on images where there were no objects and both models correctly detected no objects in every case.

5.2 Multiple Categories with Different Appearance Models

In this section, we address the case where each category has its own unique appearance model. Let the appearance parameters for each category be represented by $\boldsymbol{\alpha}_c$. The joint likelihood given by (16) becomes problematic when there is overlap between objects.

(a) (b) (c) (d)

Figure 11: Templates used for training shape and appearance of maple leaves.

One needs to decide which object is in front in order to assign it to the appropriate appearance model. Based on this decision, the indicator variable $I_{d,c}$ is equal to one for one of the object categories c or zero if it belongs to background.

We introduce a random variable $b_{c,n}$ that represents the probability of the nth object of category c to be in front. This random variable is assumed to be drawn from a uniform distribution $b \sim U[u_{min}, u_{max}]$, where $u_{min} = 0$ and $u_{max} = 1$. Let these parameters be denoted by ε and $\mathbf{b} = [b_{1:N_1,1}, \ldots, b_{1:N_C,C}]$. In the generative framework of the model, this parameter is drawn for each location of the object of each category and when there is an overlap of shape boundaries, the object with higher b value is assigned to be in front. The binary variable, $I_{d,c}$, is thus assigned according to (18). The modified joint likelihood is given by (17).

$$p(\mathbf{L_{1:C}}, \mathbf{S}_{1:C}, \mathbf{X}|\boldsymbol{\theta}) = \prod_{c=1}^{C} \Big[p(\mathbf{L}_c|\beta_c(\tau) \cdot p(\mathbf{S}_c|\mathbf{L}_c, \boldsymbol{\zeta_c}) p(\mathbf{b}|\varepsilon) \Big] \times p(\mathbf{X}|\mathbf{L}, \mathbf{S}, \boldsymbol{\alpha}_{1:C}, \boldsymbol{\gamma}), \quad (17)$$

$$I_{d,c} = \begin{cases} \delta(c_n), \textit{if } b_n > b_{n'}, \\ \delta(c_{n'}), \textit{if } b_{n'} > b_n. \end{cases} \quad (18)$$

Maple leaves data was collected to illustrate an example with multiple categories and different appearance. There are two types of leaves in the image based on the side on which it landed on the ground. Each has its own appearance model. The background consists of grass. In this case, overlap needs to be considered when detecting the leaves. The shape model was trained with 22 landmark points on 6 examples. The images used for training are shown in Figure 11. A uniform hyper-prior is chosen over the random variables ρ and a for variation in orientation and size of the leaves. Results on nine images are shown in Figure 12. Red and blue color on the boundaries of the leaves indicate different categories. The average segmentation accuracy of this experiment is 0.95 and the precision and recall rate for object detection is 1 and 0.94 respectively.

6 Experiments

In this section, we present two real-world case studies. The first case study involves segmentation of cells from fluorescence microscopy images. Inference is accomplished in

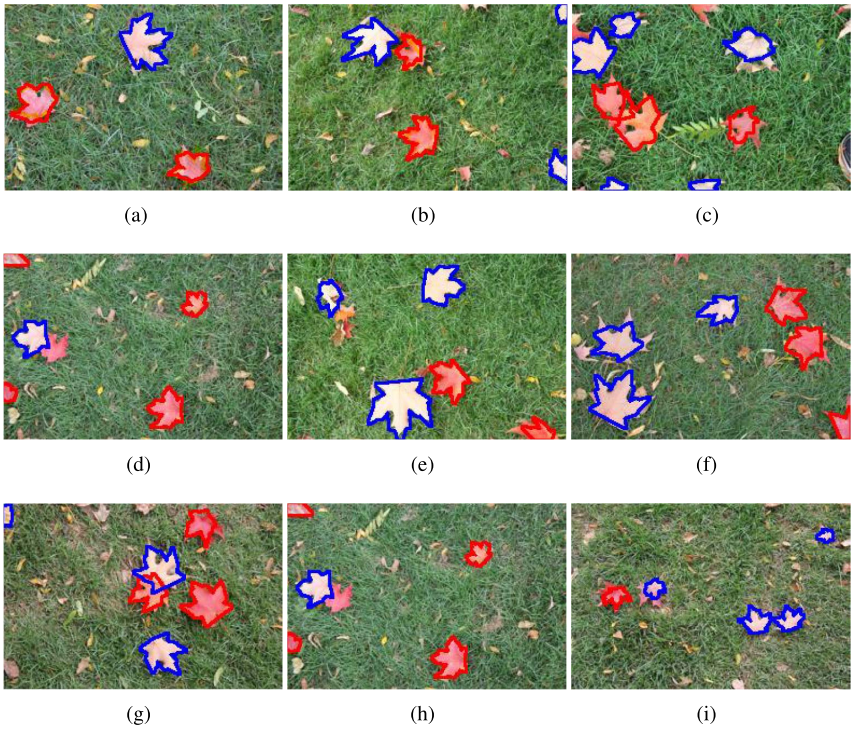
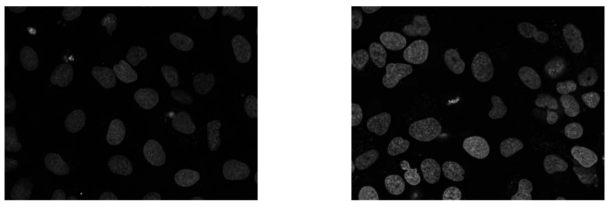


Figure 12: Results on a dataset consisting of maple leaves with different appearance (red and pink). Red and blue boundaries indicate different categories of leaves found by our method.

an unsupervised mode using a homogeneous spatial Poisson prior. The second application involves detection of cars and pedestrians (two categories) from traffic surveillance data. Inference in this case is accomplished in a supervised mode and a non-homogeneous Poisson is utilized.

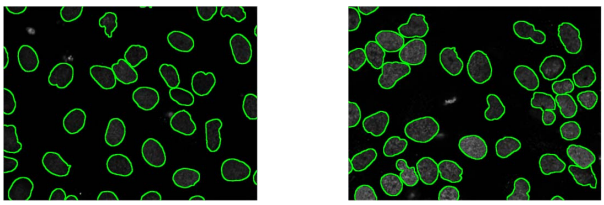
6.1 Case Study 1: Cell Image Segmentation (Unsupervised Mode)

The goal in this application is to segment and detect cell nuclei (Coelho et al., 2009) in fluorescence microscope images. The dataset contains two sets of images, 'gnf' and 'ic100', each set containing 50 images. The resolution of each image is $[1024 \times 1344]$ pixels. The nuclei of the cell show variation in appearance, shape and orientation. The shapes of the cell nuclei are very close to an ellipse. Therefore, variation in shape is encompassed in a flexible shape prior that assumes an ellipse with major axis (r_{ma}), minor axis (r_{mi}) and angle (r_a). Uniform priors are assumed because all possible values are equally probable; ($r_{mi}, r_{ma} \sim U[0, MAX]$) where $MAX = 300$ and angle ($r_a \sim U[-\pi, \pi]$). We assume a multinomial distribution over the intensity values (whose range is $[0, 255]$) for the appearance prior. Since this is an unsupervised setting, appearance



(a) (b)

Figure 13: Two example images from the cell dataset.



(a) (b)

Figure 14: Groundtruth annotations on two examples from the cell dataset.

parameters for foreground and background are determined adaptively for each test image. To do this, the intensity of the image is divided into two clusters using the k-means algorithm (Forgy, 1965; Macqueen, 1967). The resulting histograms for the two clusters are used to calculate the multinomial distribution parameters for the foreground (cell) and background appearance priors. Note that the cells are always brighter than the background. This domain knowledge is utilized to consider the cluster with a higher mean to belong to foreground. Initialization of the number of objects and their locations is done using a simple heuristic. A binary image is obtained by thresholding the original grayscale image. The value of this threshold is the maximum value in the cluster that belongs to the background appearance model. After this step, connected components are identified and their midpoints are used as the initial object locations.

Performance of the proposed algorithm is compared with other unsupervised algorithms. Bayesian nonparametric image segmentation (NBIS) (Orbanz and Buhmann, 2008) and graph-based algorithm (GB) (Felzenszwalb and Huttenlocher, 2004) can automatically determine the number of segments in an image. Spectral clustering (SC) (Ng et al., 2002) partitions the image into a constant number of regions specified by the user. This was set to 50 in our experiments. We also compare against level sets (LS) (Dufour et al., 2005) and the algorithm (MINS) proposed in (Lou et al., 2014), both of which are capable of inferring segmentations and learning the number of objects. Code for implementation of all the competing algorithms has been downloaded from the authors' websites. The competing algorithms are sensitive to parameter tuning; hence, the parameters are tuned on a validation set of two images chosen randomly.

Example images of cell data are shown in Figure 13 and Figure 14 displays the

Method	RandI	SA
NBIS	$.81 \pm .01$	$.53 \pm .11$
GB	$.81 \pm .008$	$.58 \pm .1$
SC	$.84 \pm .02$	$.62 \pm .11$
LS	$.78 \pm .01$	$.45 \pm .09$
MINS	$.77 \pm .008$	$.33 \pm .09$
Ours	$\mathbf{.86 \pm .006}$	$\mathbf{.69 \pm .03}$

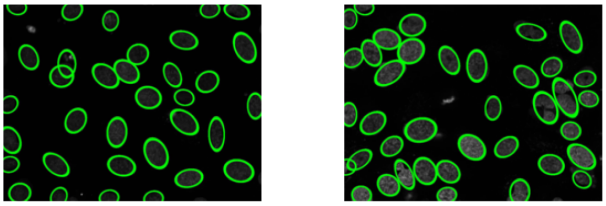
Table 1: Segmentation results for the cell data.

corresponding ground-truth annotations for these two example images. Figure 15 show the segmentation results provided by our proposed model on these same two images. Note that the results are reasonable and match the ground truth segmentation well. Figure 16 displays the results for GB, Figure 17 for MINS, Figure 18 for NBIS, Figure 19 for level sets, and Figure 20 for spectral clustering on these same two images. Note that the segmentation results on competing models tend to be noisier.

Table 6.1 reports the average segmentation results based on the Rand index (Rand, 1971) (RandI) and average segmentation accuracy (Everingham et al., 2010) (SA) on all the 100 cell images. The Rand index is the ratio of the number of pixels that have been classified correctly to the sum of correctly and incorrectly classified pixels (Rand, 1971). Segmentation accuracy is defined as the ratio of true positives to the sum of true positives, false positives and false negatives (Everingham et al., 2010). Higher values of Rand index and SA indicate better performance. Note that our proposed method resulted in the best segmentation performance compared to all other competing methods.

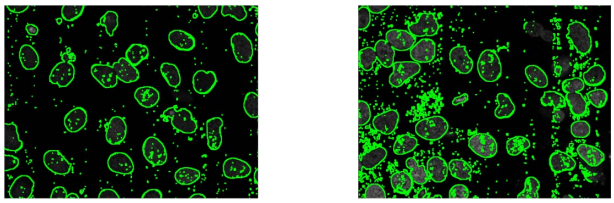
Because our algorithm can perform both segmentation and detection, we also report detection performance based on precision and recall. Precision is defined as the ratio of true positives with true positives and false positives. Recall is defined as the ratio of true positives with true positives and false negatives. We follow the rules specified in (Everingham et al., 2010) to calculate these values. The higher the value of these measures, the better. Table 6.1 reports the detection results in terms of average and standard deviation of precision and recall on the 50 cell images in each folder, 'gnf' and 'ic100'. Note that our algorithm has a better performance compared to other algorithms in the experiments. Graph-based methods have very noisy detections as can be seen from Figure 16, showing high false positives and recall values but very poor precision as shown in Table 6.1.

Apart from (LS) (Dufour et al., 2005) and (MINS) (Lou et al., 2014), all other algorithms used for comparison are primarily designed for segmentation of images. This implies that one needs to post-process the results to determine which segments belong to background and which ones belong to the foreground. This is accomplished using k-means. The intensity of the image is divided into two classes using the k-means algorithm and the resulting mean values are used to determine the clusters belonging to background and foreground. Clusters belonging to the category with low mean are assumed to be from background.



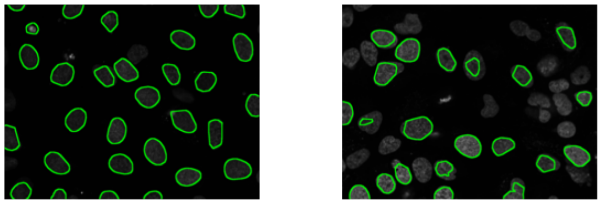
(a) (b)

Figure 15: Results of our algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.



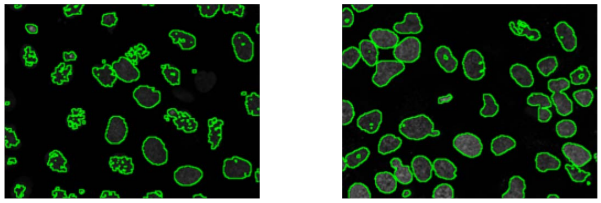
(a) (b)

Figure 16: Results of graph based (GC) algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.



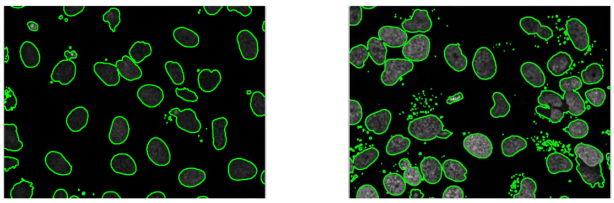
(a) (b)

Figure 17: Results of MINS algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.



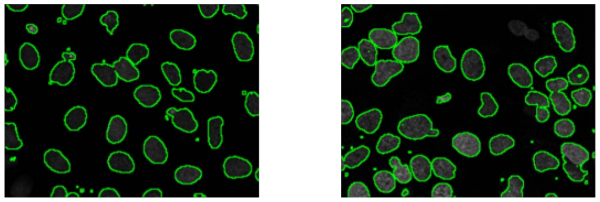
(a) (b)

Figure 18: Results of NBIS algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.



(a) (b)

Figure 19: Results of level sets algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.



(a) (b)

Figure 20: Results of spectral clustering algorithm on two example images from the cell dataset. Corresponding ground-truth annotations are shown in Figure 14.

Our algorithm is implemented in MATLAB and takes an average of 10 minutes per $[1024 \times 1344]$ image on a 12 GB RAM, 2.4 GHz computer. The NBIS and spectral clustering (SC) algorithms take 40 minutes and 14 minutes respectively. Graph-based algorithm (GB) and MINS take 61 and 25 seconds per image respectively. However, their segmentation and object detection performance is much worse than our algorithm. The level-set method takes 15 minutes. Parameters of mean and propagation weight were tuned for optimal performance. This was done once for each of the folders 'gnf' and 'ic100'. The maximum number of iterations allowed was set to 500.

6.2 Case Study 2: Traffic Surveillance Dataset (in Supervised Mode)

The traffic surveillance dataset (Wang et al., 2013) contains video of a static camera showing the traffic of pedestrians and vehicles. A spatial prior in this case is very useful to model the high probability of a car appearing on the road compared to any other part of the image. The pedestrians in the dataset also show a spatial pattern that can be captured with a non-homogeneous spatial Poisson prior, with higher probability on walking paths compared to other regions. Ground truth for training is provided by the dataset itself consisting of 350 images used to train the model and 100 test images which are used to test our model.

	Precision		Recall	
	gnf	ic100	gnf	ic100
NBIS	$.56 \pm .07$	$.21 \pm .03$	$.9 \pm .02$	$\mathbf{.83 \pm .017}$
GB	$.01 \pm .005$	$.01 \pm .002$	$.73 \pm .08$	$.56 \pm .12$
SC	$.29 \pm .016$	$.0816 \pm .006$	$\mathbf{.94 \pm .02}$	$.56 \pm .17$
LS	$.35 \pm .14$	$.12 \pm .01$	$.4 \pm .01$	$.17 \pm .03$
MINS	$.67 \pm .02$	$.13 \pm .01$	$.46 \pm .1$	$.28 \pm .17$
Ours	$\mathbf{.72 \pm .01}$	$\mathbf{.36 \pm .06}$	$.6 \pm .01$	$.3 \pm .03$

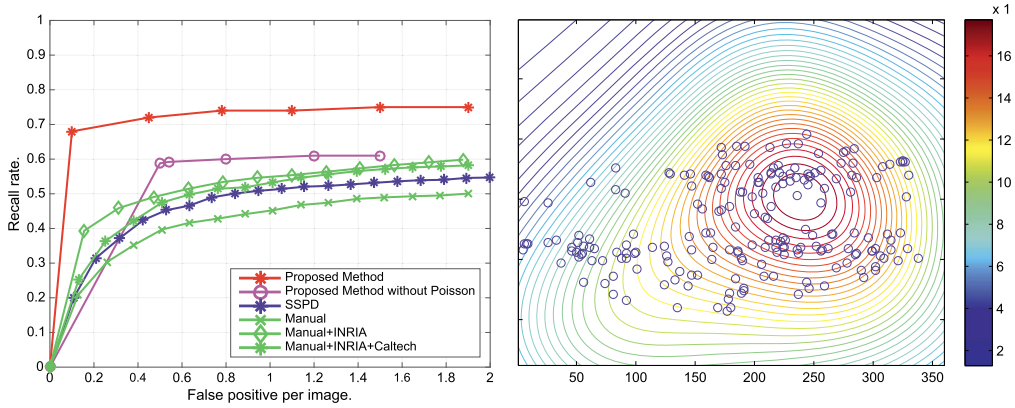
Table 2: Detection results for the cell data.

(a) (b)

Figure 21: Results on traffic images. (a) and (b) show example detection/segmentation results for pedestrians and cars.

Training: Shape and Appearance Prior A complex shape prior with 16 landmark points as shown in Section 2.2 is used to model the shape of pedestrians. An orientation parameter is not required in this case as the orientation of pedestrians in the frame does not change. Size of the object varies with location in the image; pedestrians who are farther from the camera appear smaller in size than those who are closer. Therefore, the size parameter a of the object is drawn from a uniform distribution $U[a_{min,l}, a_{max,l}]$ where, $a_{min,l} = a_{min} \times l_r, a_{max,l} = a_{max} \times l_r$, l_r is row location of the object in the image, and a_{min}, a_{max} are constants. A rectangular box whose length and width are determined by the outcome of a multinomial distribution with parameters $[\zeta_l, \zeta_w]$ is used for the shape prior on cars. The length and width of the rectangular box is discrete and therefore a multinomial distribution is chosen as the appropriate prior distribution.

The images are in color which implies that they have three channels R, G and B. We refer to the vector containing the R, G and B values at each location as the pixel intensity. These pixel intensities of objects from different categories (cars and pedestrians) show huge variation and modeling them directly is complex. Alternatively, one can model the difference in intensity value of each pixel from the observed mean background value; $e = |B_d - x_d|$, where B_d is the mean background intensity value



(a) (b)

Figure 22: Results on traffic images: (a) shows the plot of recall rate vs. false positive per image of our proposed method compared with other methods for pedestrians, and (b) shows the contour plot of the posterior Poisson intensity value inferred by the algorithm for pedestrians.

calculated from the training data. The probability of e is given as follows:

$$P(e|\phi_{c,1}, \phi_{c,2}) = \frac{\phi_{c,1}\exp(-e\phi_{c,1})}{(1-\exp(-\phi_{c,2}\phi_{c,1}))}, \quad 0 < e < \phi_{c,2}, \tag{19}$$

which is a right-truncated exponential distribution (truncated at $\phi_{c,2}$), where c represents the category which is foreground and background in our case. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2] = [\boldsymbol{\phi}_{f,1}, \boldsymbol{\phi}_{f,2}]$ and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2] = [\boldsymbol{\phi}_{b,1}, \boldsymbol{\phi}_{b,2}]$ denote appearance parameters for foreground and background respectively. The hyper-parameter values for the shape and appearance priors are calculated using maximum likelihood on the training set.

Testing Initialization of the object centers is done using a heuristic approach. A binary image is obtained by thresholding the infinity norm of the feature vector. This threshold is set to a default value (between the maximum and minimum value of the features). Midpoint of the connected components in the binary image are used as the object centers.

The dataset was designed for object detection (pedestrians) and does not contain segmentation annotations in the training set. We therefore evaluate the results only with respect to object detection accuracy. Results reported in (Wang et al., 2013) for different approaches are used for comparison. The approaches that we compare with are: a generic HOG+SVM detector trained on the CUHK dataset; on the INRIA (Dalal and Triggs, 2005) and CUHK dataset; on an additional dataset (Griffin et al., 2007); and a scene-specific pedestrian detector (Wang et al., 2013). We refer to the fourth approach that used adaptive context cues as SSPD (Wang et al., 2013). Segmentation examples

are shown in 21. Results are reported in Figure 22 which plots recall rate as a function of false positive per image, where we see that our algorithm outperforms competing models for detecting multiple pedestrians on this dataset.

Additional benefits of the model include the ability to generate synthetic shapes from the inferred shape prior (see Figure 5(b)) and the inference of posterior distribution of the spatial Poisson process that can show the high stress/traffic areas (i.e., the areas of high probability of occurrence of objects), depicted by the contour plot in Figure 22(b). This can be inferred and analyzed for different times of the day as well. Moreover, our model can detect and segment multiple object types (such as, cars and pedestrians) simultaneously.

7 Discussion

A novel probabilistic generative model for multiple object detection and segmentation is presented. This model is based on a latent marked Poisson process. The Poisson process has been pivotal in driving the research in Bayesian nonparametrics. Our proposed Bayesian latent marked Poisson process provides a natural structure to incorporate number, location, shape and feature/appearance information. The spatial information is integrated into the model using a non-homogeneous spatial Poisson process prior. Inference on this model is challenging due to changing model order which is the natural outcome of using a Poisson process. We provide a new formulation to perform inference using hybrid Gibbs sampling by taking advantage of the finite number of pixels in images. Possible extensions to the model are presented along with experimental results on two diverse real-world applications. The results show that this model can outperform competing algorithms for segmenting and detecting multiple objects.

References

Adams, R., Murray, I., and MacKay, D. (2009). "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 9–16. NY, USA: ACM. 97, 98

Baddeley, A. (2007). *Stochastic Geometry: Lectures given at the C.I.M.E. summer school held in Martina Franca, Italy, September 13–18, 2004*, chapter Spatial point processes and their applications, 1–75. Springer. MR2327290. doi: https://doi.org/10.1007/978-3-540-38175-4_1. 86, 89

Baddeley, A. J. and Van Lieshout, M. N. M. (1992). "Object recognition using Markov spatial processes." In *International Conference on Pattern Recognition*, 136–139. The Hague, The Netherlands. 87

Baddeley, A. J. and Van Lieshout, M. N. M (1993). "Stochastic geometry models in high-level vision." *Journal of Applied Statistics*, 20(5–6): 231–256. 87

Bhattacharya, A. and Dunson, D. B. (2010). "Nonparametric Bayesian density esti-

mation on manifolds with applications to planar shapes." *Biometrika*, 97: 851–865. MR2746156. doi: https://doi.org/10.1093/biomet/asq044. 91

Blei, D. M. and Frazier, P. I. (2011). "Distance dependent Chinese restaurant processes." *Journal of Machine Learning Research*, 12: 2461–2488. MR2834504. 85

Chan, T. and Zhu, W. (2005). "Level set based shape prior segmentation." In *IEEE Conference on Computer Vision and Pattern Recognition*, 1164–1170. San Diego, CA. 86

Coelho, L. P., Shariff, A., and Murphy, R. F. (2009). "Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms." In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 518–521. Boston, MA. 102

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). "Active shape models-their training and application." *Computer Vision and Image Understanding*, 61(1): 38–59. 86, 90, 91

Cremers, D., Rousson, M., and Deriche, R. (2007). "A Review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape." *International Journal of Computer Vision*, 72(2): 195–215. 86

Dalal, N. and Triggs, B. (2005). "Histograms of oriented gradients for human detection." In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 886–893. San Diego, CA. 108

Descombes, X. and Zerubia, J. (2002). "Marked point processes in image analysis." *IEEE Signal Processing Magazine*, 19(5): 77–84. 87

Dufour, A., Shinin, V., Tajbakhsh, S., Guillen-Aghion, N., Olivo-Marin, J.-C., and Zimmer, C. (2005). "Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces." *IEEE Transactions on Image Processing*, 14(9): 1396–1410. 103, 104

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). "The pascal visual object classes (VOC) challenge." *International Journal of Computer Vision*, 88(2): 303–338. 104

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). "Efficient graph-based image segmentation." *International Journal of Computer Vision*, 59(2): 167–181. 103

Forgy, E. (1965). "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications." *Biometrics*, 21: 768. 103

Ge, W. and Collins, R. T. (2009). "Marked point processes for crowd counting." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2913–2920. Miami, FL. 87

Gelfand, A. E., Le, S., and Carlin, B. (2009). "Analysis of marked point patterns with spatial and nonspatial covariate information." *Annals of Applied Statistics*, 3: 943–962. MR2750381. doi: https://doi.org/10.1214/09-AOAS240. 85, 86

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741. 85

Gonzalez, J., Low, Y., Gretton, A., and Guestrin, C. (2011). "Parallel Gibbs sampling: From colored fields to thin junction trees." In *Artificial Intelligence and Statistics*. Ft. Lauderdale, FL. 97

Green, P. J. and Mira, A. (2001). "Delayed rejection in reversible jump Metropolis–Hastings." *Biometrika*, 88(4): 1035–1053. MR1872218. doi: https://doi.org/10.1093/biomet/88.4.1035. 98

Griffin, G., Holub, A., and Perona, P. (2007). "Caltech-256 object category dataset." Technical Report 7694, California Institute of Technology. 108

He, X., Zemel, R., and Carreira-Perpinan, M. (2004). "Multiscale conditional random fields for image labeling." In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, II-695–II-702. Washington, DC, USA. 86

Huang, X. and Metaxas, D. N. (2008). "Metamorphs: Deformable shape and appearance models." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8): 1444–1459. 86

Kumar, M. P., Torr, P. H. S., and Zisserman, A. (2005). "OBJ CUT." In *IEEE Conference on Computer Vision and Pattern Recognition*, 18–25. San Diego, CA. 86

Lafarge, F., Gimel'farb, G., and Descombes, X. (2010). "Geometric feature extraction by a multi-marked point process." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9): 1597–1609. 87

Leventon, M. E., Grimson, W. E. L., and Faugeras, O. (2000). "Statistical shape influence in geodesic active contours." In *IEEE Conference on Computer Vision and Pattern Recognition*, 316–323. Hilton Head, SC. 86

Lou, X., Kang, M., Xenopoulos, P., Muñoz Descalzo, S., and Hadjantonakis, A.-K. (2014). "A rapid and efficient 2D/3D nuclear segmentation method for analysis of early mouse embryo and stem cell image data." *Stem Cell Reports*, 2(3): 382–397. 103, 104

Macqueen, J. (1967). "Some methods for classifications and analysis of multivariate observations." IN *Proc. Symp. Mathematical Statistics and Probability, 5th*, volume 1, 281–297, Berkeley. MR0214227. 103

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman and Co. 85

Moller, J., Pettitt, A. N., Reeves, R. W., and Berthelsen, K. K. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika*, 93(2): 451–458. MR2278096. doi: https://doi.org/10.1093/biomet/93.2.451. 98

Moller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). "Log Gaussian Cox processes." *Scandinavian Journal of Statistics*, 25(3): 451–482. MR1650019. doi: https://doi.org/10.1111/1467-9469.00115. 98

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). "On Spectral clustering: Analysis and an algorithm." In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, 849–856. MIT Press. 103

Orbanz, P. and Buhmann, J. M. (2008). "Nonparametric Bayesian image segmentation." *International Journal of Computer Vision*, 77(1–3): 25–45. 85, 103

Pollard, J. (1971). "On distance estimators of density in randomly distributed forests." *Biometrics*, 27: 991–1002. 94

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association*, 66(336): 846–850. 104

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). "Logistic stick-breaking process." *Journal of Machine Learning Research (JMLR)*, 12: 203–239. MR2773552. 85

Rotondi, R. and Varini, E. (2003). "Bayesian analysis of a marked point process: Application in seismic hazard assessment." *Statistical Methods and Applications*, 12(1): 79–92. MR2081753. doi: https://doi.org/10.1007/BF02511585. 86

Rue, H. and Hurn, M. A. (1999). "Bayesian object identification." *Biometrika*, 86(3): 649–660. MR1723784. doi: https://doi.org/10.1093/biomet/86.3.649. 87

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006). "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation." In *9th European Conference on Computer Vision*, 1–15. Graz, Austria: Springer. 86

Skellam, J. (1952). "Studies in statistical ecology: I. Spatial pattern." *Biometrika*, 39: 346–362. 94

Smith, A. F. M. and Gelfand, A. E. (1992). "Bayesian statistics without tears: A sampling–resampling perspective." *The American Statistician*, 46(2): 84–88. MR1165566. doi: https://doi.org/10.2307/2684170. 96

Sudderth, E. and Jordan, M. I. (2009). "Shared segmentation of natural scenes using dependent Pitman–Yor processes." In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21*, 1585–1592. Curran Associates, Inc. 86

Taddy, M. A. (2010). "Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime." *Journal of the American Statistical Association*, 105(492): 1403–1417. MR2796559. doi: https://doi.org/10.1198/jasa.2010.ap09655. 86

Terenin, A., Simpson, D., and Draper, D. (2015). "Asynchronous distributed Gibbs sampling." arXiv:1509.08999. 97

Vese, L. A. and Chan, T. F. (2002). "A multiphase level set framework for image segmentation using the Mumford and Shah model." *International Journal of Computer Vision*, 50(3): 271–293. 86

Vu, N. and Manjunath, B. S. (2008). "Shape prior segmentation of multiple objects with graph cuts." In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. Anchorage, AK. 86

Wang, X., Wang, M., and Li, W. (2013). "Scene-specific pedestrian detection for static video surveillance." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99: 1. 106, 108

Xiao, S., Kottas, A., and Sansó, B. (2015). "Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences." *Annals of Applied Statistics*, 9(1): 353–382. MR3341119. doi: https://doi.org/10.1214/14-AOAS796. 86

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G., and Micheas, A. C. (2014). "A spatio-temporal point process model for ambulance demand." arXiv:1401.5547. 87

Zhu, S., Wu, Y., and Mumford, D. (1998). "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling." *International Journal of Computer Vision*, 27(2): 107–126. 85

Acknowledgments

This work is partially supported by NIH/NCI RO1CA199673, NSF IIS-0915910 and a MURI grant from the Office of Naval Research. We also thank discussions on Poisson processes with Alican Bozkurt.