Learning and scoring Point Process models for object detection in satellite images

Jules Mabon, Mathias Ortner, Josiane Zerubia

To cite this version:

Jules Mabon, Mathias Ortner, Josiane Zerubia. Learning and scoring Point Process models for object detection in satellite images. EUSIPCO 2024 - 32nd IEEE European Signal Processing Conference, Aug 2024, Lyon, France. hal-04601239

HAL Id: hal-04601239

https://hal.science/hal-04601239

Submitted on 4 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Learning and scoring Point Process models for object detection in satellite images

1st Jules Mabon
Inria, Université Côte d’Azur
Sophia-Antipolis, France

2nd Mathias Ortner
Airbus Defense and Space
Toulouse, France

3rd Josiane Zerubia
Inria, Université Côte d’Azur
Sophia-Antipolis, France

Abstract—In this paper we propose a joint Point Process and CNN based method for object detection in satellite imagery. The Point Process allows building a lightweight interaction model, while the CNN allows to efficiently extract meaningful information from the image in a context where interaction priors can complement the limited visual information. More specifically, we present matching parameter estimation and result scoring procedures, that allow to take into account object interaction. The method provides good results on benchmark data, along with a degree of interpretability of the output. The code will be available at github.com/Ayana-Inria/

Index Terms—Object Detection, Point Process, Convolutional Neural Network, Energy Based Model, Remote Sensing

I. INTRODUCTION

For over 20 years of study [1] small object detection in optical satellite images has remained challenging; objects of interest are only few pixels large (thus lack visual information), and the dense scattering of objects increases the difficulty of separating instances (introducing interactions between neighboring objects). In this paper, we aim at extracting the geometrical configuration of vehicles in images with resolutions around 0.5 m, while leveraging the priors on object interaction to compensate for the lack of visual information.

Most Convolutional Neural Network (CNN) based approaches utilize anchor proposals [2], [3] or heatmaps [4] to locate objects; failing to consider interactions other than through a post-processing step. Approaches such as [5] model interactions through a cascade of Transformers, at a great complexity cost.

On the other hand, Point Process (PP) [6] methods allow building lightweight interaction models and jointly solve for the local and interaction contribution of a detection. The previous approaches applied to microscopy [7] or remote sensing data [8], [9] rely on contrast measures to assert the correspondence of points with the image; those fail when objects and backgrounds are varied.

With our proposed approach, we leverage the feature extraction capabilities of CNN within the PP framework as introduced in our previous works [10]. More specifically in this paper, we focus on the scoring for the output detection that takes into account interactions (while CNN approaches score objects independently, and previous PP methods use no score, simply computing precision and recall from the output as is). This score has good properties regarding the parameter optimization method utilized, and allows for interpretable results.

II. POINT PROCESS

Point Processes consider configurations of points (a finite non-ordered set \( y \) of elements of \( S \times M \)) as realizations of a random variable \( \Phi \) in the set of all possible configurations \( \mathcal{Y} = \bigcup_{n=0}^{\infty}(S \times M)^n \) (with an arbitrary amount of points). Space \( S \) corresponds to the image space, and \( M \) to the mark space. A mark can be any random variable from the radius of a circle to a discrete categorization of the object. In our case, a point \( y \in Y \) is composed of coordinates \( y_x, y_y \) in \( S \), and three marks that describe a rectangle : width \( y_w \), length \( y_l \) and angle \( y_\alpha \). We denote \( n(y) \) the number of points in a configuration \( y \). As the number of points is an unknown of the problem, Point Processes require specific sampling procedures we will detail later.

A. Point Process density

The law of the random variable is defined by its density \( h \) relative to the uniform Point Process [6]. The density derives from an energy \( U \), through a Gibbs density :

\[
h(y|X) = \frac{1}{Z} \exp(-U(y, X, \theta)),
\]

with \( Z \) an intractable normalizing constant. The energy measures the compatibility between the image \( X \) and the configuration \( y \) for a given set of parameters \( \theta \); the lower the energy the higher the compatibility (see Energy Based Models [11]).

B. Papangelou conditional intensity

The reference Poisson Point Process has an intensity \( \lambda \) that is either constant or depends on the location [6]. The density in (1) implies the intensity is now a function of the location and neighborhood of a point. Thus, the Papangelou conditional intensity \( \lambda(\cdot|\cdot) \) [6] associated to a Point Process \( \Phi \), is introduced as:

\[
\lambda(y; dy) = \lambda(\cdot|\Phi) = 1_{\Phi}(dy) = \int_\Phi 1(dy) e^y = y \cap (dy)^c, \tag{2}
\]

1 For any compact \( A \subseteq S \), \( \mathbb{E}[n(\Phi)] = \lambda[A] \) if \( \lambda \) is constant.
i.e. the infinitesimal probability to find a point in region \(dy\) around \(y \in S\), given the configuration \(y\) outside \(dy\) (i.e. \(dy^c\)). This conditional density is what allows modeling interaction between points of the PP.

When \(y \in y\), the Papangelou conditional intensity can be computed from the energy function \(U\) as:

\[
\lambda(y; y \setminus \{y\}) = \frac{h(y|X)}{h(y \setminus \{y\}|X)} = \exp(U(y \setminus \{y\}, X, \theta) - U(y, X, \theta)).
\]

III. Energy Model

We define our energy function in \(H\) as a sum of energy terms \(V_e\ (e \in \xi, \xi\) the set of energy terms) for each point \(y\) in the configuration:

\[
U(y, X, \theta) = \sum_{y \not\in X} \sum_{e \in \xi} w_e V_e(y, X, N^y_e, \theta). \quad (4)
\]

The weight parameters \(w_e \in \theta\) encode the relative importance of each energy term. We distinguish between two kinds of energy terms: data terms, written \(V_e(y, X, \theta)\), that measure the compatibility of a point against the image; and prior terms, written \(V_e(y, N^y_e, \theta)\), that measure the coherence of a point itself or against its neighborhood \(N^y_e\). The energy of a point is denoted \(V(y, \ldots)\).

A. Interpreting CNN outputs as data terms

As in our previous work [10], we build data terms by reinterpreting the outputs of a CNN as energies that can be plugged into the energy model \(H\).

\(a\) Position term: CNN based models such as \(H\) use a heatmap to find object key points (in our case: centers). Supposing that the CNN outputs a center probability map \(a, b, \alpha\), denoted \(V_e\) (\(e \in \xi\)), the set of energy terms) for each point \(y\) in the configuration:

\[
V_{pos}(y, X, \theta) = \ln(1 + \exp(-\tilde{Z}_{pos}[y_1, y_j] + t_{pos})), \quad (5)
\]

with \(t_{pos} \in \theta\) a threshold parameter.

\(b\) Mark terms: Similarly, for each mark \(k \in \{a, b, \alpha\}\), supposing we have a CNN trained to perform pixel classification into \(n_k\) discrete classes for mark \(k\). The CNN model provides the probability of the pixel at \(y_1, y_j\) to belong to class \(c_k\) \((k = 1, \ldots, n_k)\) as:

\[
p(c_k|y_1, y_j, X) = \exp\left(\tilde{Z}^{y_1, y_j}_{\kappa}[k]\right) \sum_{k=1}^{n_k} \exp\left(\tilde{Z}^{y_1, y_j}_{\kappa}[k]\right), \quad (6)
\]

i.e. using the Softmax operation. Then, as in [12], we reinterpret the Softmax input as an energy:

\[
V_{\kappa}(y, X) = -\tilde{Z}^{y_1, y_j}_{\kappa} + \ln\left(\sum_{k=1}^{n_k} \exp\left(\tilde{Z}^{y_1, y_j}_{\kappa}[k]\right)\right), \quad (7)
\]

where \(c^\kappa(y_\kappa)\) is the class corresponding to value \(y_\kappa\).

\[}

B. Prior terms

To build the prior and interaction model we combine a set of simple energy functions (that will form more complex behaviors when combined in \(H\)). We illustrate some priors used in Figure 1 with \(V_e\) an energy term computed over a single object \(y\), and \(v_e(y, u)\) an interaction energy between objects \(y\) and \(u\), aggregated for each \(y\) with a max or min over its neighbors. Notations \(\mu_e\) and \(\tau_e\) represent model parameters. Further details can be found in [10].

C. Learning the energy model

To estimate the parameters \(\theta\) of the energy model \(H\), we aim to maximize the likelihood of parameters relative to the annotated data \(D\) \([11]\). To bypass intractable integrals, \([13]\) propose the Contrastive Divergence approach: i.e., within a stochastic gradient descent scheme, minimizing the loss:

\[
\mathcal{L}(\theta_n; y^+, y^-) = U(y^+, X, \theta_n) - U(y^-, X, \theta_n) + \gamma R_V, \quad (8)
\]

where \(R_V\) is a regularization term that avoid exploding energies (average of all \(V(y)\)), \(y^+\) is a configuration close to ground truth, and \(y^- \sim \exp(U(\cdot, X, \theta_n))\). We detail the procedure adapted from [14] to Point Processes in [10].

D. Sampling with Jump Diffusion

The Point Process is sampled by building a Markov chain \((y_t)_{t=1,\ldots}\) which converges towards the stationary density \(h\).

\(a\) Birth and Death: Birth and Death moves \([15]\) allow adding and removing points. For our model we leverage the precomputed tensors \(\tilde{Z}_{pos}\) and \(\tilde{Z}^{y_1, y_j}_{\kappa}\) to propose more relevant points in space \([10]\).

\(b\) Diffusion: To modify the current configuration \(y_t\) at a fixed number of points, we leverage Diffusion dynamics \([16]\):

\[
y_{t+1} \leftarrow y_t + -\beta \nabla y_t U(y_t) + \sqrt{2T_t w_t}, \quad w_t \sim N(0, \beta), \quad (9)
\]

with \(T_t\) the temperature at time \(t\) and \(\beta\) the gradient step.

\(c\) Simulated annealing: When looking for the best fitting configuration for a given image \(X\), i.e. the configuration that minimizes the energy \(U\), we use simulated annealing \([1\) i.e. simulate a chain of stationary density \(h/T_t\) with \(T_{t+1} = 0.998T_t\).

\[\]
IV. SCORING USING POINT INTERACTIONS

Classical CNN based object detection models for object detection (such as \cite{2, 3, 17}) yield a confidence score \( s(y) \in \mathbb{R} \) for each proposed object \( y \) in the image. This confidence score is often interpreted, for each detection, as proportional to the probability of the proposed element \( y \) to be a true positive, \( s(y) \propto p(y|X) \). Applying a score (or confidence) threshold \( t_s \) gives a set of detections, for which metrics such as precision and recall can be computed by matching the detections with the ground truth. This allows adapting the threshold according to the need of the application; i.e. some applications may require few false positive (high precision) while others require less missed detections (high recall).

In order to assess the performance independently of the threshold selection, the Average Precision (AP) metric sums up the performance of a model as the area under the precision-recall curve. Indeed, to trace a precision recall-curve, one only requires the sequence of \( \text{Recall}(t_s), \text{Precision}(t_s) \) pairs, which are obtained by sequentially pruning the lowest scoring points. Equations \( (12) \) provides a score to each point \( y_n \).

C. Contrastive divergence loss and Papangelou intensity

On one hand the energy model is trained by minimizing the loss function in \cite{8} derived from the likelihood maximization of the parameters regarding the annotated data. On the other we evaluate the performance of the inferred configuration with the scoring method in \cite{12} sourced from the Papangelou intensity. Here we show that while the two are derived differently, minimization of the loss function leads to good properties on the score function.

Here we consider a simplified loss with only the two energy terms (as \( \gamma \approx 0 \)). Denoting the energy change induced by the move from configuration \( y \rightarrow x \) as \( \Delta U(y \rightarrow x) = U(x) - U(y) \), we have:

\[
\mathcal{L}(\theta, y^+, y^-) = \Delta U(y^- \rightarrow y^+). \tag{13}
\]

Similarly, the Papangelou intensity can be rewritten as such:

\[
\lambda(u; y) = \exp(\Delta U(y \cup \{u\} \rightarrow y)). \tag{14}
\]

\( a) \) Single point addition: Thus, for a simple negative sample \( y^- = y^+ \cup \{u\} \) in which we add a non-valid point \( u \) to \( y^+ \), we have:

\[
\mathcal{L}(\theta, y^+, y^-) = \log(\lambda(u; y^+)). \tag{15}
\]

This leads into the expected behavior: minimizing the loss \( \mathcal{L} \) leads to minimizing the score of non-valid point \( u \). The same stands for the removal of a valid point \( y \in y^+ \), and maximizing its score.

\( b) \) Arbitrary sequence of moves: This is also valid for the generic case where \( y^- \) is generated from an arbitrary sequence of additions or removal of points from \( y^+ \) (a translation/rotation/scaling can be viewed as removal then addition). This defines a sequence \( (y_k)_{k=0, \ldots, n} \) of \( n \) configurations as:

\[
\forall k = 1, \ldots, n, \ y_k = \begin{cases} y_{k-1} \setminus \{y_k\} & \text{if } y_k \in y^+ \\ y_{k-1} \cup \{y_k\} & \text{otherwise} \end{cases} \tag{16}
\]

with \( y_0 = y^+, y^- = y_n \), and \( y_k \) elements of either \( \mathcal{S} \times \mathcal{M} \) or \( y^+ \). Without loss of generality we can reorder the sequence to match the pruning order defined in \cite{11}. The energy change for one move is given as:

\[
\Delta U(y_{k-1} \rightarrow y_k) = \begin{cases} \log(\lambda(y_k; y_{k-1} \setminus \{y_k\})) & \text{if } y_k \in y^+ \\ -\log(\lambda(y_k; y_{k-1} \setminus \{y_k\})) & \text{otherwise} \end{cases} \tag{17}
\]

As we have (by definition) \( \Delta U(x \rightarrow x') = \Delta U(x \rightarrow x') + \Delta U(x' \rightarrow x') \), the loss is given as:

\[
\mathcal{L} = \sum_{y_k \in y^+} \sum_{y_{k-1} \in y^+} \log(\lambda(y_k; y_{k-1} \setminus \{y_k\})). \tag{18}
\]

By ordering the \( y_k, y_k \) to match the pruning order in \cite{11} each \( \lambda(y_k; y_{k-1} \setminus \{y_k\}) \) can be matched to their respective score.
allows for regularization of the resulting configurations.

Hereby, we showed that minimization of the loss at a configuration level leads to the expected results on object scores.

D. Results interpretability

Due to the decomposition of the total energy into energy terms introduced in (4), the object score can be decomposed similarly:

\[ s(y \setminus \{y\}) = \prod_{e \in \xi} s_e(y \setminus \{y\}) \tag{19} \]

with \( s_e(y \setminus \{y\}) = \exp(w_e \Delta V_e(y \rightarrow y \setminus \{y\})) \), the Papanicolau intensity obtained by considering the single energy term \( e \). This allows viewing the contribution of each component. Moreover, we propose grouping these contributions into the data and prior contributions to respectively obtain \( s_{\text{data}} \) and \( s_{\text{prior}} \) such that the final score is a product of the two:

\[ s(y \setminus \{y\}) = s_{\text{data}}(y)s_{\text{prior}}(y \setminus \{y\}). \]

V. Application

A. Models

In this paper, we show results on three CNN-based models, and our two PP models: \( \text{CNN-PP*} \) and \( \text{CNN-PP}^\star \) correspond to our PP+CNN model, while the first only estimates the weights \( w_e \) through the estimation method (other parameters set manually); \( \text{CNN-LocalMax} \) only uses the CNN part of our model with local maxima applied to extract object; \( \text{BBA-Vec.} \) and \( \text{YOLOV5-OBB} \) correspond to [2] and [17]. We provide further insights in the relative complexity of those models in [20].

B. Results

1) Quantitative and qualitative evaluation on benchmark data: We train and evaluate our models on the DOTA [21] dataset, sub-sampled to a 0.5 m resolution (in order to match satellite sensor specifications from Airbus). To assert the noise resilience, we also evaluate the methods on the same data with additive noise. For every model, we compute the Average Precision (AP) in Table I, the increased performance from CNN-LocalMax. to \( \text{BBA-Vec.} \) and \( \text{CNN-PP}^\star \) shows the PP improves results over the CNN alone. Some results on sample images are shown in Fig. 5. It shows our CNN and PP combination allows for regularization of the resulting configurations:

\[
\begin{array}{|c|c|c|}
\hline
\text{Method} & \text{AP} \text{DOTA} & \text{AP} \text{DOTA+noise} \\
\hline
\text{BBA-Vec.} & 0.82 & 0.19 \\
\text{YOLOV5-OBB} & 0.86 & 0.10 \\
\text{CNN-LocalMax.} & 0.86 & 0.55 \\
\text{CNN-PP}^\star & 0.91 & 0.58 \\
\text{CNN-PP}^\star & 0.92 & 0.62 \\
\hline
\end{array}
\]

2) Qualitative evaluation on ADS data: We evaluate the methods on data provided by ADS, at a 0.5 m resolution. As this data is not labeled, models are trained only on the benchmark data presented above. Results are presented in Fig. 4. Qualitatively, our PP model is able to produce regular configurations of vehicles, while missing fewer objects of interest compared to \( \text{BBA-Vec.} \).

3) Inference interpretability: In Figure 5, we illustrate how the two components of the score \( s \) can help analyze the results: green objects correspond to detection with high prior and data scores, while blue detection have a higher data contribution. The few yellow detection correspond to objects with low data score, often located on ambiguous locations.

VI. Conclusion

Here we propose a novel scoring method for our model that utilizes Convolutional Neural Networks within a Point Process framework. This score allows to measure the detection confidence considering the object interactions. We show that, on top of allowing regularization and robustness on the resulting configuration of points, this enables some explainability of the results through the decomposition of the model into multiple terms.

REFERENCES

Fig. 3. Samples of detection on the test dataset. The score threshold (to not display low score objects) is set to maximize the $F_1$ score for each model.

Fig. 4. Samples of detection on the ADS data. The dataset is not annotated. [© Airbus Defense and Space]

(a) Inferred configuration  (b) Color correspondence

Fig. 5. Inferred configuration on an ADS data sample (a), colored according to their prior/data scores (b): yellow: $s_{\text{prior}} > s_{\text{data}}$; blue: $s_{\text{prior}} < s_{\text{data}}$; purple: low $s$; green: high $s$. Each point in (b) corresponds to a detection in $s_{\text{data}}$-$s_{\text{prior}}$ space (log values scaled to $[0, 1]$).


