Processus ponctuels marqués et réseaux de neurones convolutifs pour la détection d'objets dans des images de télédétection

Jules MABON[1], Mathias ORTNER[2], Josiane ZERUBIA[1]

[1]Inria, Université Côte d'Azur
Sophia-Antipolis, France

[2]Airbus Defense and Space
Toulouse, France

jules.mabon@inria.fr, mathias.ortner@airbus.com, josiane.zerubia@inria.fr

Résumé – Cet article présente une méthode combinant processus ponctuels marqués et réseaux de neurones convolutifs pour la détection de petits objets dans des images satellitaires. Dans un tel contexte, la densité des objets est parfois élevée ; la formulation énergétique d'un processus ponctuel nous permet d'intégrer des *a priori* sur les configurations qui modélisent les interactions entre ces objets. D'autre part, les réseaux de neurones à convolution nous permettent d'apprendre les termes de vraisemblance, plus adaptables aux contextes variés, contrairement aux termes classiques dans les approches par processus ponctuels, fondés sur des mesures de contraste.

Abstract – This article presents a method combining marked point processes and convolutional neural networks in order to detect small objects in satellite images. In such a setting, the density of objects can be high: the energetic formulation of a point process allows us to factor in priors on configurations that model object interactions. Moreover convolutional neural networks allow us to learn likelihood terms, making the latter more resilient to varying visual contexts compared to classical point process approach measures based on contrast.

1 Introduction

Dans le cadre de la détection d'objets dans des images satellitaires, la résolution spatiale est généralement telle que les objets d'intérêt ont une taille de quelques pixels seulement, faisant perdre ainsi une grande partie de l'information visuelle. De plus, les objets sont parfois distribués densément, ce qui peut rendre la distinction des instances plus difficile et introduit des interactions entre des objets voisins.

De nombreuses méthodes fondées sur les réseaux de neurones à convolutions (*Convolutional Neural Networks*) (CNN) comme Faster R-CNN [1], YOLO [2] ou RetinaNet [3] proposent de détecter des objets dans des images dites "naturelles", où les objets sont de grande taille et les modèles d'interaction limités. Cependant, la taille des boites de détection ou les limites sur le nombre d'objets rendent ces approches moins fiables dans le cadre de la télédétection.

D'autre part, les approches de type géométrie stochastique proposent de résoudre conjointement le problème de détection et la sélection des objets. Permettant à la fois d'ajouter des modèles d'interaction entre les objets via des *a priori* , et de vectoriser l'information extraite de l'image. Ces approches utilisent classiquement des mesures de vraisemblance construites à partir de mesures de contraste [4-6], mais des situations d'illumination et des contextes visuels changeants rendent la conception d'une telle énergie plus complexe et coûteuse en temps de calcul.

Nous proposons dans cet article de combiner les *a priori* de l'approche par géométrie stochastique avec l'extraction d'information des CNN. Nos contributions principales sont les suivantes :

— nous formulons la tâche de détection comme une minimisation d´énergie, nous permettant d'introduire des *a priori* sur les configurations.
— nous introduisons des mesures de vraisemblance des configurations issues de modèles de CNN simples, pour remplacer les mesures de contraste.
— nous évaluons cette approche sur des images à une résolution spatiale de satellite standard (50cm/px) et comparons les résultats avec [7].

2 Définition du processus ponctuel

Nous définissons le support de l'image comme $S \subset \mathbb{R}^2$. Une configuration de points Y est un ensemble fini non ordonné d'éléments de $S \times M$, avec M l'espace des marques. Dans notre application, un objet $y \in Y$ sera composé de coordonnées i, j dans $\mathbb{R}^2$, et de trois marques paramétrant un rectangle (voir Figure 1) ; taille ($t = \frac{a+b}{2}$), ratio ($r = \frac{a}{b}$) et angle ($\alpha \in [0, \pi]$).

Nous définissons une configuration de points comme la réalisation d'un Processus Ponctuel Marqué (MPP) non-uniforme ; celui-ci est défini par une densité h par rapport au processus de Poisson uniforme [8].

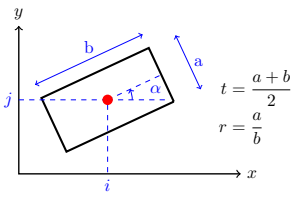


FIGURE 1 – Paramétrisation du rectangle

Le modèle de sélection et d'interaction entre les points découle d'une énergie E, via une densité de Gibbs non normalisée :

$$h(Y) \propto \exp(-E(Y)) \tag{1}$$

Ainsi, pour inférer une configuration d'objets $\widehat{Y}$ à partir d'une image observée X, il faut préalablement construire une énergie $E_{tot}(X, Y)$ dont le minimum est atteint pour la configuration optimale $Y = \widehat{Y}$.

3 Détection sous forme d'énergie

L'énergie totale pour une image X et une configuration de points Y est construite de la sorte :

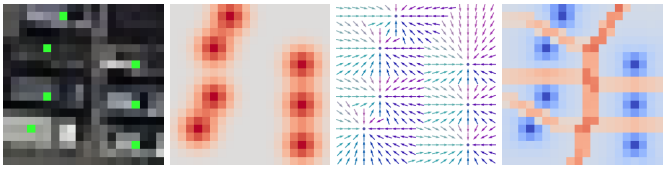
$$\begin{aligned} E_{tot}(X,Y) &= \sum_{y \in Y} e_{\textit{données}}(X,y) + I_{X,y}.e_{\textit{a priori}}(y, \mathcal{N}_y) \\ e_{\textit{données}}(X,y) &= \omega_p.e_p(X,y) + I_{X,y}.\omega_m.e_m(X,y) \\ e_{\textit{a priori}}(y,\mathcal{N}_y) &= \omega_t.e_t(y) + \omega_s.e_s(y,\mathcal{N}_y) + \omega_a.e_a(y,\mathcal{N}_y) \end{aligned} \tag{2}$$

L'énergie d'un point de Y est composée de deux parties, une énergie d'attache aux données $e_{\textit{données}}(X, y)$ qui mesure la vraisemblance du point par rapport à l'image observée X, et une énergie *a priori* $e_{\textit{a priori}}(y)$ qui mesure la cohérence de y avec son voisinage $\mathcal{N}_y = \{\tilde{y} \in Y | \tilde{y} \neq y, \|y - \tilde{y}\| < d_{\max}\}$ dans la configuration Y ; e_p, e_m, e_t, e_s et e_a correspondent respectivement aux énergies associées aux position, marques, *a priori* de taille, *a priori* de superposition et *a priori* d'alignement (paragraphes 3.1 et 3.2). Tandis que ω_p, ω_m, ω_t, ω_s et ω_a pondèrent ces sous-énergies. Enfin $I_{X,y} = \mathbb{1}_{e_p(X,y)<0}$ est l'indicatrice d'une énergie de position négative, ce terme permet de ne prendre en compte les contributions énergétiques de l'*a priori* et du terme de marque seulement si la position est correcte.

3.1 CNN pour l'attache aux données

Classiquement les approches par MPP utilisent des mesures de contraste entre l'intérieur et l'extérieur de l'objet [4-6]. Dans notre cas d'application, l'aspect visuel des objets est très varié, nous utilisons donc des CNN pour construire un terme de vraisemblance plus polyvalent.

Terme de position Un premier CNN infère le terme d'attache aux données propre à la position. Pour ce faire, un Unet [9] infère une carte de probabilité des centres des objets.



(a) image et centres (vert) (b) centres + filtre gaussien (c) vecteurs (d) divergence

FIGURE 2 – Champ de vecteurs (c) et divergence (d) pour estimer les centres (a) (en (b) et (d) : rouge> 0, bleu< 0).

L'identification des centres des objets s'apparente à une tache de détection de points clés [10]. Cependant, les points d'intérêt sont parfois si peu distants qu'une approche directe par probabilité de points clés (*heatmap*) – c'est à dire apprendre une carte des centres dilatés par un filtre gaussien, (voir Figure 2b) – engendre une connectivité entre les objets proches lors de l'inférence, et ne permet pas de bien séparer les instances. De plus, réduire la variance des gaussiennes autour des points, renforce le déséquilibre des annotations, et s'oppose à l'incertitude sur les annotations (qui sont souvent bruitées).

Ainsi, inspiré par [11], pour une image X de taille (h, w), nous apprenons cette carte des centres via une tache intermédiaire d'apprentissage d'un champ de vecteurs unitaires, de taille $(h, w, 2)$, pointant vers le centre le plus proche (Figure 2c). De ce champ de vecteur nous pouvons alors identifier, avec l'opérateur de divergence, les centre des objets comme des zones à divergence négative, et les séparations entre les objets comme des zones à divergence positive (Figure 2d).

Ainsi pour $y \in Y$

$$e_p(X,y) = -2\left(\sigma\left(b + a.\text{div}(\widehat{V})_{y_i,y_j}\right) - s_{det}\right) + 1 \tag{3}$$

où σ est la fonction sigmoïde et $\text{div}(\widehat{V})_{y_i,y_j}$ est la divergence du champ de vecteur inféré $\widehat{V}$ depuis l'image X à la position de y ; a et b sont des paramètres du modèle estimés lors de l'entraînement du réseau de neurones (paragraphe 4.1) et s_{det} est le seuil de détection.

Terme de marques L'énergie $e_m(X, y)$ associée aux marques (ici les paramètres du rectangle t,r,α) est obtenue de la manière suivante : pour chaque marque, on discrétise sa plage de valeurs en N_m intervalles. De la sortie du Unet, pour une image X de taille (h, w), on extrait un tenseur $\widehat{T}^m$ de taille (h, w, N_m) pour chaque marque m (parmi t,r,α). La distribution de probabilité sur les valeurs discrètes de la marque m à la position de y est donnée par $\text{Softmax}(\widehat{T}^m_{y_i,y_j})$. Ainsi pour chaque y, avec y_m la valeur de la marque m, et $\text{indice}(y_m)$ l'indice de l'intervalle contenant y_m, $\text{Softmax}(\widehat{T}^m_{y_i,y_j})_{\text{indice}(y_m)}$ estime la probabilité $P(y_m|X, y_i, y_j)$. Cette approche permet de mesurer des configurations de points où la marque aurait une distribution multimodale, ce qui n'est pas faisable avec une simple régression.

$$e_m(X,y) = \sum_{m \in \{t,r,\alpha\}} -2.\text{Softmax}(\widehat{T}^m_{y_i,y_j})_{\text{indice}(y_m)} - 1 \tag{4}$$

3.2 *A priori* sur les configurations

Nous utilisons divers termes d'*a priori* pour régulariser la configuration des points à partir des propriétés des objets étudiés :

***a priori* de taille** Incite à respecter des limites de taille (minimale et maximale) sur les objets d'intérêt. Ceci permet d'éviter les objets à surface nulle :

$$e_t(y) = \max\{a_0 - \text{aire}(y), \text{aire}(y) - a_1, 0\} \tag{5}$$

***a priori* de superposition** Pénalise la superposition entre objets [5], pour :

$$e_s(y, \mathcal{N}_y) = \max_{\tilde{y} \in \mathcal{N}_y} \left\{ \frac{\text{aire}(\tilde{y} \cap y)}{\min\{\text{aire}(\tilde{y}), \text{aire}(y)\}} \right\} \tag{6}$$

***a priori* d'alignement** Récompense les configurations où les objets sont alignés entre eux :

$$e_a(y, \mathcal{N}_y) = \min_{\tilde{y} \in \mathcal{N}_y} \{-|\cos(|y_\alpha - \tilde{y}_\alpha|)|\} \tag{7}$$

avec y_α l'angle de y dans le repère de l'image (Figure 1).

4 Apprentissage et inférence

En bref, notre modèle s'applique comme suit :

1. apprentissage des termes de vraisemblance (paragraphe 4.1) et ajustement des paramètres du processus ponctuel (paragraphe 5.2) sur le jeu d'entraînement .
2. pour une image X du jeu de test, inférence unique des cartes $\widehat{V}$ et $\widehat{T}^{\cdot}$, définissant alors pour tout $Y \in S \times M$ la fonction $Y \to E_{tot}(X, Y)$ (paragraphe 3.1).
3. minimisation de $E_{tot}(X, \cdot)$ par simulation du processus ponctuel (paragraphe 4.2).

4.1 Apprentissage des termes d'attache aux données

Position Le modèle d'énergie de position est entraîné en minimisant la fonction de coût :

$$L_{pos}(X, Y) = \text{MSE}(\widehat{V}, V) + \text{BCE}(M, \sigma(a.\text{div}(\widehat{V}) + b)) \tag{8}$$

où V est le champ de vecteur dérivé de la configuration vérité terrain Y, M une carte binaire des centres dilatés par un filtre gaussien ($\sigma = 0.6$) dérivée de Y. MSE et BCE sont respectivement la moyenne des erreurs au carré (*Mean Squared Error*) et l'entropie croisée binaire (*Binary Cross Entropy*).

Marques Le modèle d'énergie sur les marques est entraîné en minimisant la fonction de coût, pour une marque m (parmi t, r et α, voir partie 2) :

$$L_m(X, Y) = \frac{1}{|P|} \sum_{p \in P} \text{CE}(\text{Softmax}(\widehat{T}_p^m), \text{Softmax}(T_p^m)) \tag{9}$$

Avec P, l'ensemble des pixels de X et CE, l'entropie croisée (*Crossed Entropy*) entre la distribution estimée $\widehat{T}_p^m$ et la vérité terrain T_p^m pour chaque pixel p.

4.2 Simulation du processus ponctuel

Une fois les termes d'attache aux données appris, on cherche la configuration qui minimise l'énergie totale :

$$\widehat{Y} = \underset{Y \in S \times M}{\text{argmin}}\, E_{tot}(X, Y) \tag{10}$$

Pour ce faire, nous utilisons une chaîne de Markov Monte Carlo à sauts réversibles (RJMCMC) [12]. Cette méthode étend l'algorithme de Metropolis Hastings en permettant d'explorer un espace d'état à dimension variable. La convergence est assurée si on utilise a minima un noyau de naissance-mort uniforme dans $S \times M$. Pour accélérer la convergence, [12] y ajoute des noyaux de translation/rotation/mise à l'échelle.

De plus, nous ajoutons des noyaux de naissance-mort et translation/rotation/mise à l'échelle, proposant des points à partir d'une densité non-uniforme [6] ; dérivée du potentiel $e_p(X, \cdot) + e_m(X, \cdot)$. Ces densités sont facilement échantillonnables car issues des cartes pré-calculées par le CNN à partir de l'image X. La réversibilité des noyaux de transition (condition nécessaire à la convergence) est démontrée dans [6].

$\widehat{Y}$ est obtenu par simulation du MPP de densité proportionnelle à $\exp(-E_{tot}(X, \cdot)/T)$, avec un recuit simulé à température T à décroissance géométrique.

5 Résultats expérimentaux

5.1 Données : DOTA gsd50

Nous nous plaçons dans le cadre de la détection d'objets dans des images de satellites, tels que Pléiades [1] ou CO3D [2], avec une résolution spatiale limitée à 50cm/px. Nous utilisons le jeu de données DOTA [13], dans lequel une partie des images sont issues de prises de vues aériennes (avec une plus grande résolution spatiale). Ainsi nous fixons une résolution à 50 cm/px, en sous-échantillonnant les images plus résolues.

De plus, nous ne nous intéressons qu'à la détection des véhicules terrestres, nous gardons uniquement les classes *small-vehicle* et *large-vehicle*.

5.2 Choix des paramètres

Les paramètres ω (équation 2) sont sélectionnés manuellement $\omega_p = 1$, $\omega_m = 0.25$, $\omega_t = 0.25$, $\omega_s = 0.75$ et $\omega_a = 0.05$,

1. Constellation Pléiades [pleiades.cnes.fr]
2. Constellation Optique 3D [intelligence-airbusds.com]

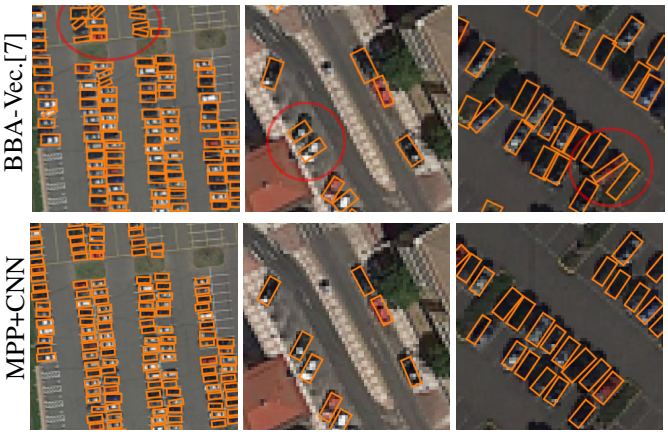


FIGURE 3 – Résultats sur des échantillons d'images.

Méthode	F1	Précision	Rappel
BBA-Vec.[7]	0.65	0.60	0.71
MPP+CNN	0.66	0.58	0.77

TABLE 1 – Comparaison des métriques de détection de boites orientés (seuil IOU : 0.25). Pour BBA-Vec.[7] le seuil de détection est $s = \text{argmax}_{s\in[0,1]}\ \text{F1}(s)$

le premier terme étant arbitrairement à 1 car il importe seulement de connaitre E_{tot} à une constante de normalisation près. Nous utilisons $N_m = 32$, comme compromis entre la résolution et la taille mémoire de la distribution $\widehat{T}^m_{y_i,y_j}$. Le seuil de détection s_{det} de l'équation 3 est choisi de façon à maximiser le score F1 pour la tâche de classification des centres.

5.3 Évaluation des résultats

Nous montrons dans la Figure 3 des résultats dans des imagettes, comparant notre approche avec BBA-Vectors[7]. Notre approche présente des résultats moins bruités grâce à l'utilisation des *a priori*. Cette différence se reflète mal sur les métriques de détection (voir Table 1), car la correspondance vérité-terrain/détection, fondée sur l'intersection sur union (IOU), se focalise principalement sur une correspondance de position et (approximativement) de taille. La faible précision pour les deux détecteurs s'explique par le grand nombre de faux positifs, les objets ayant une apparence très simpliste à cette résolution.

6 Conclusion et perspectives

Dans cet article, nous proposons une combinaison de processus ponctuel et apprentissage profond pour formuler la détection comme une minimisation d'énergie, nous permettant d'introduire des *a priori*, régularisant la détection. Nous obtenons des résultats comparables à l'état de l'art dans des images peu résolues, tout en réduisant le bruit sur la détection. Enfin, si les *a priori* restent faibles dans le contexte présent, nous adapterons notre méthode à des séquences d'images où les dynamiques des objets nécessitent des *a priori* forts.

Remerciements

Les auteurs sont reconnaissants envers l'infrastructure OPAL de l'Université Côte d'Azur (UCA) pour avoir fourni les ressources de calcul nécessaires à ce travail de recherche, ainsi qu'envers BPI France pour le soutien financier dans le cadre du contrat LiChiE.

Références

[1] S. Ren, K. He, R. Girshick et J. Sun, "Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks," *IEEE TPAMI*, 2017.

[2] J. Redmon, S. Divvala, R. Girshick et A. Farhadi, "You Only Look Once : Unified, Real-Time Object Detection," in *Proc. CVPR*, 2016.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He et P. Dollar, "Focal Loss for Dense Object Detection," in *Proc. ICCV*, 2017.

[4] P. Craciun, M. Ortner et J. Zerubia, "Joint Detection and Tracking of Moving Objects Using Spatio-temporal Marked Point Processes," in *Proc. IEEE Winter Conf. on Applications of Computer Vision*, 2015.

[5] X. Descombes, "Multiple objects detection in biological images using a marked point process framework," *Methods*, Image Processing for Biologists, 2017.

[6] C. Lacoste, X. Descombes et J. Zerubia, "Point Processes for Unsupervised Line Network Extraction in Remote Sensing," *IEEE TPAMI*, 2005.

[7] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu et D. Metaxas, "Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors," in *Proc. IEEE Winter Conf. on Applications of Computer Vision*, 2021.

[8] M.-C. V. Lieshout, *Markov Point Processes and Their Applications*. Imperial College Press, 2000.

[9] O. Ronneberger, P. Fischer et T. Brox, "U-Net : Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015.

[10] Z. Cao, T. Simon, S.-E. Wei et Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Proc. CVPR*, 2017.

[11] D. Neven, B. D. Brabandere, M. Proesmans et L. Van Gool, "Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth," in *Proc. CVPR*, 2019.

[12] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 1995.

[13] G.-S. Xia et al., "DOTA : A Large-Scale Dataset for Object Detection in Aerial Images," in *Proc. CVPR*, 2018.