Détection de véhicules en imagerie aérienne par mélange de modèles discriminatifs

Hicham RANDRIANARIVO[1], Bertrand LE SAUX[1], Marin FERECATU[2]

[1]ONERA - The French Aerospace Lab
F-91761 Palaiseau, France

[2]CNAM - Laboratoire Cedric
292 rue Saint-Martin, 75141 Paris, France
{hicham.randrianarivo, bertrand.le_saux}@onera.fr,marin.ferecatu@cnam.fr

Résumé – Nous proposons une nouvelle méthode pour la détection de véhicules dans des images aériennes ou satellites à très haute résolution. Elle est basée sur un mélange de filtres d'orientations de gradients qui décrivent finement l'apparence visuelle des objets. Chaque modèle est entrainé de manière discriminative afin de permettre une spécification implicite de la classe d'objet. Une procédure itérative de fouille des contre-exemples difficiles permet de mieux spécifier le détecteur. Nous validons notre approche sur plusieurs jeux de données de grande taille, et montrons qu'elle répond à de nombreux problèmes en télédétection, changement d'orientation et taille des images.

Abstract – We propose a new method for vehicle detection in very-high-resolution aerial or satellite images. It is based on a mixture of gradient-orientation filters which capture the visual appearance of objects. Each model is discriminately trained so that a class can by implicitely specified. We use an iterative hard-negative mining procedure for focusing the detector on difficult samples. We assess our approach on several large datasets and show it tackles efficiently major problems in remote sensing such as orientation change and data size.

1 Introduction

L'imagerie aérienne aujourd'hui (et satellitaire demain) est transformée par deux innovations fondamentales. D'une part, la très haute résolution (THR, environ 10cm/pixel) permet de distinguer des détails de l'image et donc d'analyser finement son contenu sémantique. Cela conduit à un grand nombre d'objets visuels par image, chacun finement caractérisable. D'autre part, le nombre d'images disponibles, provenant de différents porteurs et de capteurs optiques variés, augmente sans cesse et permet d'accumuler des exemples de chaque objet. En contrepartie, le nombre et la taille des images impliquent d'avoir des traitements automatiques à la fois performants et rapides.

Nous proposons une nouvelle approche pour la détection d'objets dans les images aériennes, et plus précisément de véhicules. De nombreuses applications sont visées, telles que la gestion des flux en urbanisme et la détection de convois de réfugiés dans une contexte humanitaire. Les environnements complexes (notament urbain) en font un problème difficile : de nombreux artefacts sont susceptibles d'être confondus avec des véhicules par la machine. Une seconde difficulté provient des orientations multiples d'un objet dues au point de vue vertical.

Notre approche s'appuie sur plusieurs techniques pour former un détecteur d'objet adapté aux images aériennes qui réalise un bon compromis performance/rapidité. Les objets sont modélisés par un mélange de classifieurs discriminatifs (Machines à Vecteurs de Support - SVM - linéaires). Chacun de ces classifieurs est sur-entraîné par *bootstrapping* pour apprendre finement des modèles d'apparence basés sur des Histogrammes d'Orientation de Gradient (HOG, [1]). À la détection, la combinaison HOG et SVM linéaire permet un filtrage en temps raisonnable des images, tout en s'adaptant à des changements d'échelle et de résolution.

Travaux connexes Dans [7], les auteurs proposent un détecteur de véhicules qui combine un modèle d'apparence de patch basé sur les HOG et un classifieur entraîné par Boosting. [5] utilise le HOG normalisé de son orientation principale et un classifieur de type SVM (noyau à base radiale). Dans les deux cas (testés sur des images de 60 à 70 cm / pixel) les HOG montrent leur fort pouvoir discriminatif. Des alternatives existent cependant, comme les filtres de Gabor utilisés dans [3]. Les HOG sont aussi le descripteur de choix dans les modèles à parties entraînés discriminativement (*Discriminatively-trained Part Models* [2]). Ce modèle d'objet, déjà appliqué avec succès aux images satellitaires [9], est très générique, mais plusieurs de ses caractéristiques permettent de construire un détecteur de véhicules efficace : outre l'apprentissage discriminatif par SVM et les HOG, le mélange de modèles pour représenter chaque catégorie permet de tenir compte des changements d'apparence intra-classe. Enfin, les avancées récentes en classification et détection d'objets par réseaux de neurones profonds nécessitent de grandes bases d'images étiquetées (comme la détection de routes [6] ou de cibles [10]), ce qui freine leur utilisation en imagerie aérienne.

2 Mélange de filtres d'orientations de gradients discriminatifs

2.1 Apprentissage

Notre détecteur est un mélange de classifieurs (HOG et SVM linéaire qui définissent un filtre d'orientations de gradients) entraînés sur un ensemble d'images aériennes. L'apprentissage est basé sur deux mécanismes fondamentaux : la recherche de sous-catégories visuellement similaires pour une même classe d'objet, et une procédure de *hard-mining* : un entraînement itéré sur les exemples difficiles (c'est-à-dire les exemples mal classés par les premières versions du classifieur).

Algorithme 1 Apprentissage du mélange de filtres d'orientations de gradient

1: Sélectionner des exemples positifs dans les annotations $\longrightarrow \{(O_i, y_i = +1)\}$
2: Catégoriser $\{(O_i, y_i = +1)\}$ par mélange de Gaussiennes $\longrightarrow$ sous-catégories $S^{(k)} = \{(O_i^{(k)}, y_i = +1)\}$
3: $\forall$ catégorie $S^{(k)}$: **faire**
4: Estimer la taille du filtre (taille médiane sur $S^{(k)}$)
5: Sélectionner des exemples négatifs (patchs choisis aléatoirement dans les images et ayant un recouvrement inférieur à 50% avec les positifs $\{(O_i, y_i = +1)\}$) $\longrightarrow \{(O_i, y_i = -1)\}$
6: Re-dimensionner $\{(O_i, y_i = +1)\}$ à la taille du filtre
7: Représenter $\{(O_i^{(k)}, y_i)\}$ par les HOG $\longrightarrow \{(x_i^{(k)}, y_i)\}$
8: Entraîner une SVM sur $H^0 = \{(x_i^{(k)}, y_i = +1)\} \cup \{(x_i, y_i = -1)\}$ $\longrightarrow \beta^{(k)}$: le filtre d'orientations de gradients qui modèlise la sous-catégorie $S^{(k)}$
9: **boucle hard-mining** $\forall m \leq M$: **faire**
10: Classifier les images par $f^{(k)}(x) = \beta^{(k)} \cdot x$ pour extraire de nouveaux exemples négatifs "difficiles" $\longrightarrow H^{m+1} = \{(x'_i, y'_i = -1)\} \cup H^m$
11: Ré-entraîner la SVM sur H^{m+1}
12: **fin boucle**
13: **fin boucle**
14: Sortie : $\longrightarrow$ mélange de modèles $\{\beta^{(k)}\}$

Formellement, à chaque image est associée une liste d'objets définis par une boîte englobante : des véhicules et des instances d'autres types d'objets possibles (bâtiments, végétation, etc.), ce qui constitue l'ensemble $\mathcal{T} = \{(O_i, y_i)\}$, $O_i \in \mathbb{R}^{h_i * l_i}$ imagette de taille $h_i \times l_i$ représentant l'objet et $y_i \in \{-1; 1\}$. Chaque imagette est indexée par un HOG $O_i \mapsto x_i$. La procédure d'apprentissage résumée dans l'algorithme 1 permet de construire le mélange de classifieurs :

$$f^{(k)} : HOG \to \mathbb{R}$$
$$x \to f^{(k)}(x) = \beta^{(k)} \cdot x$$

Catégorisation de la classe objet : La première étape vise à trouver les différentes sous-catégories d'aspect de la classe objet visée. Seuls les exemples positifs $\{(O_i, y_i = +1)\}$ de $\mathcal{T}$ sont utilisés. Suivant [2], le critère choisi est le ratio d'aspect des boîtes englobantes $\frac{h_i}{l_i}$, mais pour obtenir des catégories plus homogènes nous optimisons un mélange de distributions Gaussiennes sur ces valeurs. L'ensemble des exemples positifs est alors décomposé en plusieurs ensembles $S^{(k)} = \{(0_i^{(k)}, y_i = +1)\}$ correspondant aux diverses sous-catégories.

Entraînement avec recherche de négatifs difficiles : Nous estimons ensuite un modèle par sous-catégorie. Pour chacune, un ensemble d'exemples négatifs $\{(x_i, y_i = -1)\}$ est construit de manière itérative par une procédure de *bootstrapping*. Il est initialisé en choisissant au hasard des imagettes dans les images, sous la seule contrainte d'un faible recouvrement avec les positifs. Un filtre d'orientations de gradients β_k (HOG appris par la SVM linéaire) est estimé sur $H^{(0)} = \{(x_i^{(k)}, y_i = +1)\} \cup \{(x_i, y_i = -1)\}$, puis le détecteur basé sur β_k est appliqué aux images complètes : les exemples négatifs sont alors enrichis par les faux positifs ayant un score élevé, c'est à dire les objets les plus difficiles pour le modèle courant. Cette procédure est itérée M fois (dans les expériences $M = 4$) sur $H^{m+1} = \{(x'_i, y'_i = -1)\} \cup H^m$ afin de minimiser :

$$L_{H^m}(\beta) = \frac{1}{2}||\beta||^2 + C \sum_{H^m} \max(0, 1 - y_i.\beta_k \cdot x_i) \qquad (1)$$

2.2 Détection

Afin de traiter les images plus rapidement que par une procédure de fenêtre glissante classique, la classification est directement effectuée dans l'espace des HOG. Les modèles β_k y définissent des filtres d'orientations de gradients. L'image entière est indexée par HOG et convoluée par chaque β_k en utilisant la corrélation croisée normée rapide. On obtient une carte de détections dense, équivalente au résultat d'une SVM linéaire appliquée à chaque patch x du HOG (voisinage de la taille de β_k) : $f^{(k)}(x) = \beta^{(k)} \cdot x$. La détection est facilement réalisée à des échelles différentes en redimensionnant l'image avant de calculer le HOG, ce qui permet de traiter des images de résolutions différentes de l'apprentissage, et donc de transférer un détecteur pré-entraîné sur des images acquises dans des conditions différentes.

3 Expériences

3.1 Jeux de données

Deux jeux de données optiques THR couvrant de larges zones d'environnement urbain ont été utilisés pour les expériences. Le premier jeu de données comporte 4 ortho-images aériennes (10cm/pixel, 5000×4000 pixels) acquises sur la ville de Christchurch (NZ) en 2011 [8] : 2357 voitures y ont été annotées. Le deuxième jeu de données (*grss_dfc_2015*) comporte 7 ortho-images aériennes (5cm/pixel, 10000×10000 pixels) acquises sur la ville de Zeebruges (Belgique) en 2011. Il a été rendu pu-

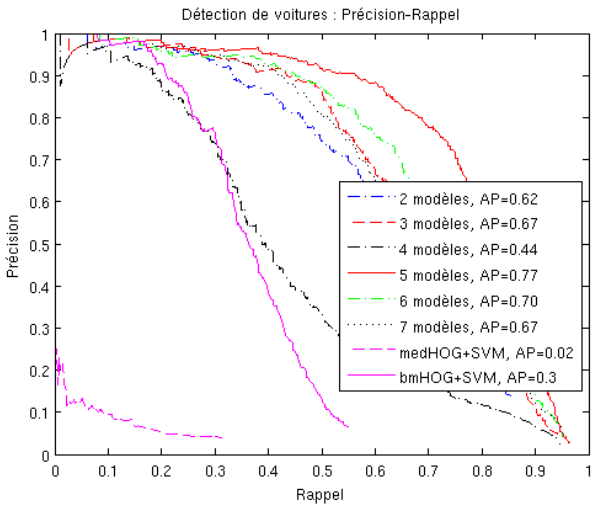


FIGURE 1 – Courbes précision-rappel de détection de voitures en milieu urbain (données Christchurch). Des mélanges plus ou moins grands (de 2 à 7 modèles d'apparence) sont comparés avec des détecteurs simples HOG (sur patch médian et sur le meilleur modèle) et SVM. La meilleure précision moyenne est obtenue pour 5 modèles d'apparence et vaut 0.77.

blic lors du *Data Fusion Contest 2015* [4]. Nous y avons annoté manuellement 955 voitures.

3.2 Performance de détection

FIGURE 2 – Mélange de 3 modèles et détections-type associée.

La figure 1 compare les performances de notre approche pour des mélanges plus ou moins grands : de 2 à 7 modèles d'apparence. Les courbes précision-rappel sont calculées comme suit : les détecteurs sont entraînés sur 3 images, et testés sur la 4ème. Une détection est correcte si elle présente un recouvrement normalisé d'au moins 50% avec la vérité-terrain. Une catégorisation appropriée influe fortement sur les performances : la précision moyenne (ou *Average Precision*, AP) varie de 0.44 (4 modèles) à 0.77 (5 modèles). Plus précisément, les meilleurs mélanges comportent 3 modèles ($AP = 0.67$ correspondant aux 3 orientations principales possibles : horizontale, oblique, verticale, cf. Fig. 2) ou 5 modèles ($AP = 0.77$ correspondant à un découpage plus fin de l'espace des orientations). Notons qu'avec un nombre plus élevé de modèles, la catégorisation des exemples d'apprentissage conduit à 3 catégories effectives (non-vides). Fig. 1 montre également le résultat d'un détecteur

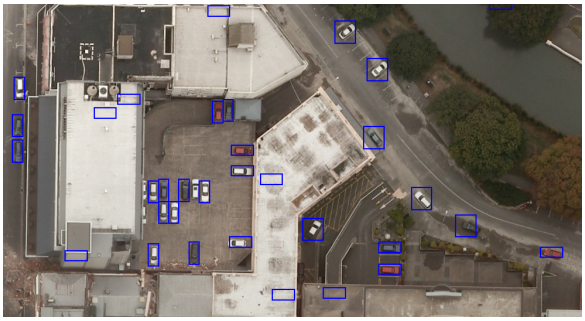


FIGURE 3 – Détections de voitures à diverses orientations en milieu urbain par mélange de modèles (données Christchurch).

standard HOG et SVM (similaire à [5]) entrainé sur patch à la taille médiane des voitures ou suivant un patch de forme allongée, qui obtient des performances limitées.

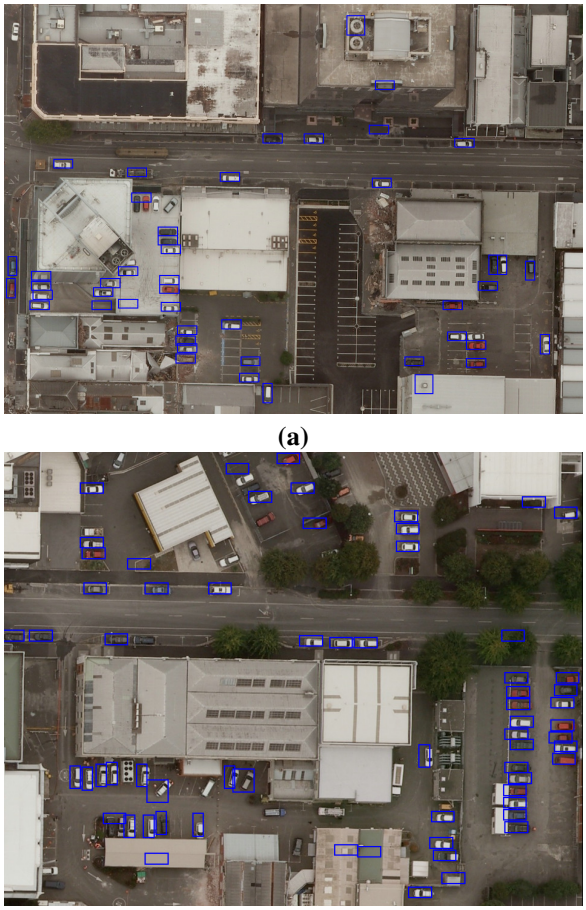
Fig. 4 montre plusieurs résultats de détection. Alors que les bâtiments comportent de nombreux objets de taille et de forme comparables à des véhicules (ventilations, ouvertures, etc.), peu de fausses alarmes apparaissent, et quasiment toutes les voitures sont détectées. Fig. 3 montre que les différents modèles ont répondu correctement, ce qui permet d'estimer l'orientation locale des véhicules. Appliquer le détecteur sur une image $5\text{k} \times 5\text{k}$ pixels prend environ 3 minutes sur un processeur Intel i7 CPU M-640 @ 2.80GHz.

3.3 Transfert de modèle

Alors qu'un biais fort vers la classe cible est recherché, un sur-apprentissage aux données est à éviter. Le modèle appris sur les données Christchurch (ville anglo-saxonne, 10cm/pixel) est appliqué sur une image des données *grss_dfc_2015* (contexte européen, environ 5cm/pixel). Par construction, nos modèles sont facilement applicables à une échelle différente (ou une plage d'échelles si la résolution de l'image de test est inconnue). La figure 5 montre ces détections : Fig 5-a montre que peu de fausses alarmes apparaissent sur les bâtiments ou dans les cours, tandis que Fig 5-b montre que les différents modèles du mélange on été utilisés par le détecteur, permettant de retrouver la plupart des véhicules quelque soit leur orientation.

4 Conclusion

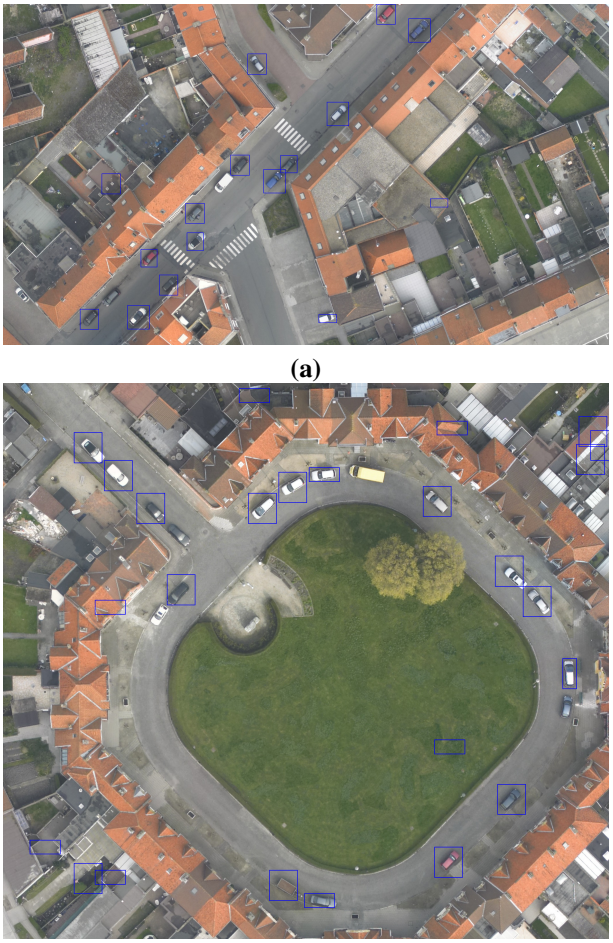
Nous avons présenté un algorithme pour la détection de véhicules dans les images aériennes THR par mélange de modèles discriminatifs, chaque modèle correspondant à une catégorie d'aspect de véhicules. Ces modèles sont des filtres d'orientations de gradients entraînés discriminativement avec une procédure de recherche d'exemples négatifs difficiles. Cette approche est particulièrement efficace, y compris dans les environnements urbains complexes, facile à mettre en oeuvre, et permet la détection multi-échelle.



(a)

(b)

FIGURE 4 – Détections de voitures en milieu urbain par mélange de modèles (données Christchurch).



(a)

(b)

FIGURE 5 – Détections de voitures en milieu urbain par mélange de modèles (données *grss_dfc_2015*).

Remerciements

The authors would like to thank the Belgian Royal Military Academy for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

Références

[1] N. DALAL et B. TRIGGS : Histograms of oriented gradients for human detection. *In Proc. Comp. Vis. and Pattern Rec.*, pages 886–893, Washington DC, USA, 2005.

[2] P. FELZENSZWALB, R. GIRSHICK, D. MCALLESTER et D. RAMANAN : Object detection with discriminatively trained part-based models. *IEEE Trans. Patt. An. Mach. Int.*, 32(9):1627–1645, 2010.

[3] J. GLEASON, A. V. NEFIAN, X. BOUYSSOUNOUSSE, T. FONG et G. BEBIS : Vehicle detection from aerial imagery. *In ICRA*, 2011.

[4] 2015 IEEE GRSS DATA FUSION CONTEST : Online : `http://www.grss-ieee.org/community/technical-committees/data-fusion`.

[5] J. MICHEL, M. GRIZONNET, J. INGLADA, J. MALIK, A. BRICIER et O. LAHLOU : Local feature based supervised object detection : Sampling, learning and detection strategies. *In IEEE IGARSS*, Vancouver, Canada, 2011.

[6] V. MNIH et G. HINTON : Learning to detect roads in high-resolution aerial images. *In Proc. of Eur. Conf. Comp. Vis.*, 2010.

[7] T. NGUYEN, H. GRABNER, B. GRUBER et H. BISCHOF : On-line boosting for car detection from aerial images. *In IEEE Conf. on Research, Inovation and Vision for the Future*, pages 87–95, Hanoï, Vietnam, 2007.

[8] NZAM : New Zealand Aerial Mapping Limited, Aerial Christchurch after earthquake on feb, 22, 2011. `http://nzam.com/`, 2011.

[9] H. RANDRIANARIVO, B. LE SAUX et M. FERECATU : Man-made structure detection with deformable part-based models. *In IGARSS*, Melbourne, Australia, 2013.

[10] S. RAZAKARIVONY et F. JURIE : Discriminative autoencoders for small target detection. *In ICPR*, Stockholm, Sweden, 2014.