CARTOGRAPHIE ET INTERPRÉTATION DE L'ENVIRONNEMENT PAR DRONE

Martial Sanfourche, Bertrand Le Saux, Aurélien Plyer, Guy Le Besnerais

ONERA The French Aerospace Lab, F-91761 Palaiseau, France

Résumé

Nous présentons dans cet article le système de cartographie géométrique et d'interprétation sémantique de l'environnement pour des applications drone développé à l'ONERA/DTIM. Une cartographie précise en 3D de l'environnement survolé par le drone est réalisée au moyen des données vidéo et Lidar acquises en vol. Notre système comprend ensuite un module pour la cartographie sémantique interactive et la détection d'objets génériques dans le modèle global. Enfin nous proposons des fonctionnalités de détection et pistage des objets mobiles et des évènements sur vidéo, qui permettent de localiser les évènements spécifiques sur la carte, et ainsi avoir une vue globale de la situation.

Mots clés : Modélisation géométrique, Cartographie sémantique, Interprétation, Drone

Abstract

This article presents the processing chain for geometric and semantic mapping of a UAV environment that we developed at ONERA/DTIM. A precise 3D modelling of the environment is computed using video and (if available) Lidar data captured during the UAV flight. Then semantic mapping is performed by interactive learning on the model, thus allowing generic object detection. Finally, event detection and moving-object tracking are performed in the video stream and localized in the 3D model, thus giving a global view of the situation.

Keywords : *Geometric modelling, Semantic mapping, Scene understanding, UAV*

1. Introduction

De nombreux travaux ont été menés pour fournir automatiquement des cartes de l'environnement exploré par des drones : ces cartes contiennent typiquement des représentations géométriques 2D ou 3D, combinées avec de l'information liée au capteur telle que des images texturées. L'hypothèse sous-jacente est que ces modèles constituent un pas vers une autonomie accrue des drones, en permettant l'égo-localisation et la planification de trajectoires (Fraundorfer et al., 2012). L'étape suivante consiste en une cartographie sémantique qui fournit en outre des informations telles que la localisation de certains objets ou d'aires d'intérêt - cf. (Nuechter et Hertzberg, 2008).

Cependant dans de nombreuses situations pratiques comme la surveillance de site industriel ou les opérations de recherche et sauvetage, les drones ne sont pas des robots complètement autonomes, mais des engins au moins partiellement téléopérés. Dans la station-sol, des experts du domaine d'application travaillent en collaboration étroite avec les opérateurs de drone pour concevoir le schéma de l'intervention. Dans ce papier, nous proposons une approche complète pour cartographier et interpréter l'environnement vu par le drone. Notre approche met l'homme au coeur du système en utilisant au mieux l'expertise que peuvent apporter les opérateurs sur le terrain et en leur fournissant en retour une vue globale de l'environnement qui les aide dans leur prise de décisions.

L'article est organisé comme suit. Dans la partie 2, nous détaillons notre méthode pour estimer précisément les trajectoires et obtenir des modèles 3D fins. Nous détaillons la détection des objets mobiles et des évènements d'intérêt dans la partie 3. Enfin, nous proposons une approche pour une cartographie sémantique des objets et des zones d'intérêt génériques réalisée de manière interactive par l'opérateur dans la partie 4, avant de conclure (partie 5).

2. Cartographie hors-ligne de l'environnement pour la détection d'événement et la localisation

Ce premier module produit des orthomosaïques et des modèles numériques d'élévation (MNE) haute résolution à partir d'une séquence vidéo et éventuellement de relevés Lidar acquis par un drone. La chaîne de traitements décrite ici nécessite que chaque donnée capteur soit associée à une information de position et d'attitude du porteur. Cela est réalisé soit par synchronisation des capteurs, soit par l'intermédiaire d'un horodatage de toutes les données enregistrées à bord de l'engin durant son vol ce qui est le cas sur les drones du laboratoire RESSAC de l'ONERA/DCSD qui ont acquis les données illustrant cet article. Par ailleurs, on suppose que le calibrage des capteurs vidéo et Lidar est connu.

Le traitement mis en œuvre enchaîne trois étapes.
(1) En dépit de l'emploi d'un récepteur GPS-RTK qui fournit une précision de localisation décimétrique et d'une centrale inertielle précise, un affinage des paramètres de prise de vue est obligatoire pour assurer la cohérence

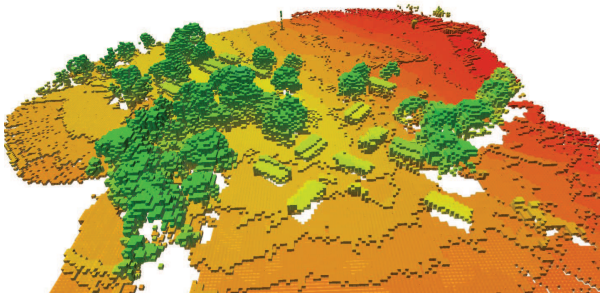

FIGURE 1: Carte 3D de l'environnement construite à partir de la trajectoire précise estimée par ajustement de faisceaux dans les images vidéo.

FIGURE 2: Comparaison d'une image aérienne (source IGN) et du MNE obtenu à partir de données Lidar mises en géométrie commune à l'aide des mesures brutes délivrées par le système de navigation. L'inconsistance de ces mesures produits des bâtiments fantômes autour du carrefour.

géométrique des données vidéo et Lidar le long de la séquence. Sur la figure 2 montrant d'un côté une photographie aérienne d'un carrefour situé dans le village de Caylus, Tarn-et-Garonne (source : IGN, site géoportail) et de l'autre le modèle numérique de la zone obtenu à partir des données brutes (Lidar, position et attitude), les bâtiments fantômes présents sur le MNE illustrent parfaitement l'incohérence géométrique des mesures délivrées par le système de navigation et leur effet sur la production des cartes. L'affinage des paramètres de la séquence vidéo se fera en deux temps. D'abord par un ajustement de faisceaux sur une sélection d'images-clés puis par calcul de pose pour les autres images.

(2) L'affinage des conditions de prise de vue permet de mettre dans une géométrie commune cohérente les points 3D relevés par le Lidar ou les cartes de profondeur obtenues par stéréo-association dense de paires d'images. Notre système produit indifféremment des MNE ou des modèles 3D voxellique (voir Fig. 1) à l'aide de la librairie Octomap décrite dans (Wurm et al., 2010).

(3) Enfin des orthomosaïques superposables aux MNE sont générées en projetant les images de la vidéo sur le MNE tout en tenant compte des occultations.

2.1. Affinage des paramètres de prise de vue

2.1.1. Critère à minimiser

Cette étape obligée consiste à appliquer un traitement d'ajustement de faisceaux à la séquence vidéo - voir à ce sujet (Triggs et al., 2000; Lourakis et Argyros, 2009). En pratique, après avoir détecté des points caractéristiques sur les K images de la séquence vidéo et les avoir appariés afin de constituer N pistes, l'ajustement de faisceaux consiste à rechercher les paramètres de prises de vue des K vues prises en compte et la position dans le repère géographique local des N éléments de la scène auxquels correspondent les pistes de points détectés dans les images minimisant un critère de reprojection dans les images.

Notons :

- T_k^i et Θ_k^i respectivement la position et l'attitude de la caméra correspondant à la prise de vue k à l'itération i du processus de minimisation ;
- p les paramètres de prise de vue connus et fixés tout au long de la séquence (paramètres intrinsèques de la caméra, position et attitude du capteur dans le repère engin) ;
- X_n^i la position dans le repère géographique de l'élément de la scène n ;
- $u_{k,n}^{obs}$ la position de l'amer n dans la vue k fournie par un algorithme de détection et de suivi de points caractéristiques dans les images. Nous décrirons plus bas comment sont extraites les observations image ;
- Π le modèle capteur qui permet de calculer la position dans l'image de X_n^i sachant p, T_k^i et Θ_k^i.

Le critère d'ajustement générique s'écrit alors :

$$J = \sum_{k=1}^{K} \sum_{n=1}^{N} v(k,n) \left\| u_{k,n}^{obs} - \Pi\left(T_k, \Theta_k, p, X_n\right) \right\|_{W_{k,n}} \quad (1)$$

où $v(k,n)$ indique si l'observation $u_{k,n}^{obs}$ existe bel et bien et $\|\bullet\|_W$ désigne la distance de Mahalanobis au sens d'une matrice de covariance W.

Dans notre cas, l'absence de points d'appui de référence nous a conduit à fixer la position associée à la première image et à ajouter au critère précédent un terme d'attache aux données traduisant notre confiance dans les mesures GPS-RTK délivrées par le système de navigation $\{T_k^{obs}\}_{k\in[1\ldots K]}$. Le critère que nous cherchons à minimiser est alors le suivant :

$$J = \sum_{k=1}^{K} \sum_{n=1}^{N} v(k,n) \left\| u_{k,n}^{obs} - \Pi\left(T_k, \Theta_k, p, X_n\right) \right\|_{W_{k,n}} + \sum_{k=1}^{K} \left\| T_k - T_k^{obs} \right\|_{W_T} \quad (2)$$

où W_T est la matrice de covariance décrivant une erreur de localisation de $15cm$ selon les 3 axes. Dans notre cas, l'attitude de la caméra est décrite par les 3 angles d'Euler, ce qui fait 6 paramètres par point de vue et donc $6(K-1)+3N$ paramètres à affiner.

2.1.2. Processus d'optimisation hiérarchique

La minimisation de ce critère non linéaire met en œuvre des algorithmes du type Levenberg-Marquardt ou Dog-Leg. Ces techniques de résolution numérique repose sur l'approximation au premier ordre de la matrice Hessienne comme le produit de la matrice jacobienne associée au critère. La structure éparse de cette matrice [1] permet, lorsque cette propriété est prise en charge par l'optimiseur, de résoudre efficacement des problèmes de grande dimension par un découplage des variables correspondant aux caméras et celles associées à la structure 3D grâce au complément de Schur.

Afin de traiter rapidement des séquences de plusieurs milliers d'images comportant plusieurs milliers d'amers 3D, nous proposons une résolution en deux temps du critère décrit dans l'équation 2.

Traitement d'une sous-séquence d'images-clés. La forte redondance de contenu entre les images successives d'une vidéo nous a conduit à résumer la séquence en un nombre très réduit d'images-clés disposant d'un recouvrement spatial cependant suffisant pour pouvoir associer des primitives entre ces images.

Hormis la première image et la dernière image de la séquence qui sont automatiquement ajoutées à la liste des images clés, les autres images-clés sont sélectionnées sur la base des statistiques du suivi temporel des primitives image, à la manière de nombreux algorithmes de *Simultaneous Localisation and Mapping* (SLAM) basés sur des images-clés (Sanfourche et al., 2013). Dans la première image-clé, C points caractéristiques sont détectés par un algorithme de type Harris. Ils sont suivis dans les images successives par le pisteur KLT de Shi et Tomasi (1994). Afin de détecter les erreurs de suivi, les résultats du suivi KLT $t \rightarrow t+1$ sont confirmés en lançant KLT dans le sens opposé puis en contrôlant l'erreur de dispersion dans la géométrie de l'image acquise à l'instant t. Lorsque le nombre d'associations passe sous un seuil défini comme un pourcentage du nombre de points caractéristiques visibles dans la dernières image-clé visitée (généralement 80 à 90$\%$), l'image courante devient une image-clé et de nouveaux points caractéristiques sont détectés dans l'image afin de suivre à nouveau C primitives image.

Un second traitement permet de faire des liaisons temporelles à long terme entre images-clés correspondant aux multiples passages du drone au dessus d'une même zone. Ce processus s'appelle fermeture de boucle dans la communauté SLAM. Nous exploitons les informations de localisation fournies par le système de navigation pour déterminer les taux de recouvrement entre images-clés sous l'hypothèse d'une scène plate d'altitude nulle. Pour les couples d'images-clés correspondant à des maxima locaux de ce critère (stocké sous la forme d'une matrice carrée de largeur le nombre d'images-clés),

1. qui tient au fait qu'une observation image ne dépend que des paramètres de pose d'une caméra et de la position d'un seul amer 3D

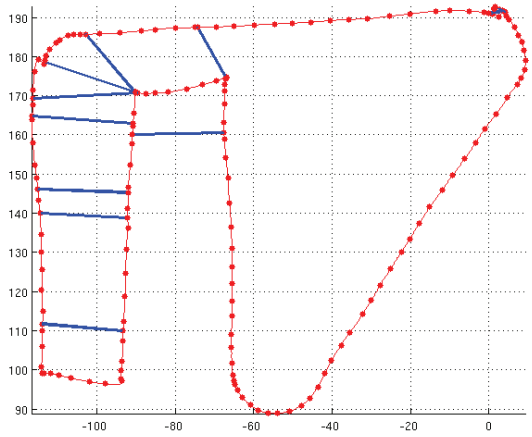

FIGURE 3: Résumé de la séquence vidéo acquise par le drone sous forme d'un graphe d'images-clés. Les points rouges, les arrêtes rouges et bleues correspondent respectivement aux images clés, liens temporels entre images-clés successives et aux liens à long terme de fermeture de boucle (revisite).

les descripteurs SIFT de Lowe (2004) sont associés aux primitives suivies par KLT et un appariement sur la base de ces descripteurs est effectué en reprenant le principe de la sélection mutuelle. On peut ainsi mettre en correspondance des pistes correspondant à un même élément de la scène.

A l'issue de ces traitements vidéo de détection et de suivi de primitives image, on obtient des pistes et un résumé de la séquence d'images comme celui de la figure 3. Le critère décrit par l'équation 2 peut alors être minimisé sur ce résumé par un optimiseur adapté de celui développé par Lourakis et Argyros (2009). En pratique, pour gérer les erreurs d'appariement non détectées par le pare-feu mis en place plus haut, l'optimiseur est relancé successivement 3 fois. Entre chaque exécution, les observations image pour lesquelles le résidu de projection est hors norme (seuil à 3σ) sont éliminées.

Traitement des images restantes. De nombreuses images ont été éliminées lors du processus de sélection des images-clés. Il convient cependant d'en affiner les paramètres de prises de vue pour disposer d'un échantillonnage temporel fin de la trajectoire du drone.

Chaque image est traitée indépendamment et sa pose est affinée à partir des appariements 2D/3D déduits des pistes KLT, la position 3D de l'amer étant celle calculée par l'étape d'ajustement de faisceaux précédente.

2.2. Calcul des modèles numériques d'élévation

Un Modèle Numérique d'Élévation (MNE) représente le relief d'une scène sous forme d'une image dont chaque pixel, correspondant à une cellule d'une résolution donnée dans le plan horizontal, stocke une information de hauteur (par rapport au niveau 0 de référence).

La première étape consiste à collecter les mesures 3D exprimées dans le repère capteur pour les exprimer dans le repère global grâce aux conditions de prises de

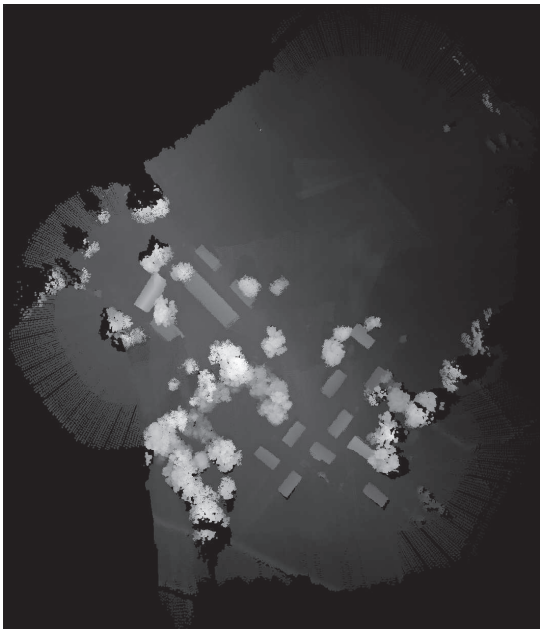

FIGURE 4: MNE du site d'entraînement de la base de Caylus obtenu aux moyens de relevés Lidar aéroportés. La résolution au sol est de 20 centimètres.

vues affinées par ajustement de faisceaux. Deux cas de figure sont possibles selon le capteur utilisé :

- Les mesures de profondeur fournies par le capteur Lidar sont traduites sous forme de nuage de points 3D grâce aux paramètres intrinsèques du capteur (résolution angulaire) ;
- Chaque paire d'images-clés successives (à condition que le mouvement du porteur entre ces deux instants soient principalement de la translation) est traitée par l'algorithme de flot optique *eFolki* de Plyer et al. (2014) pour fournir des associations de chaque pixel et ainsi trianguler des points 3D.

Maintenant, ces données 3D sont combinées dans le MNE en attribuant à chacun des pixels l'altitude maximale mesurée sur l'ensemble des points 3D planimétriquement localisés dans la cellule au sol qui correspond au pixel considéré. La figure 4 présente un MNE obtenu à partir de mesures Lidar alors que la figure 5 présente une portion de MNE obtenue par combinaison de carte de profondeurs produites par eFolki.

2.3. Production des orthomosaïques

Une orthomosaïque est une image aérienne de synthèse corrigée des effets du relief. Produire une orthomosaïque revient donc à appliquer une texture sur le MNE.

Dans le cas d'un MNE produit par relevés Lidar, l'opération est plus complexe puisqu'il faut récupérer l'information radiométrique dans le capteur vidéo. Le centre de chaque cellule du MNE est ainsi reprojeté dans chaque image-clé en tenant compte de leur observabilité géométrique déduite par un procédé de type *Z-buffer*. Le niveau de gris associé à chaque pixel de l'orthomosaïque est obtenu

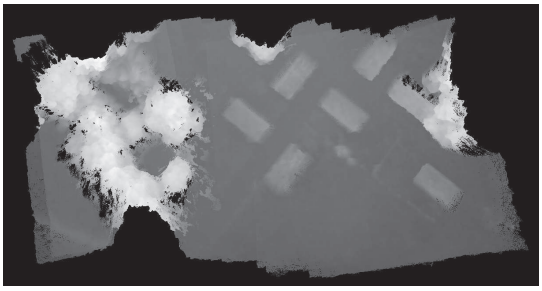

FIGURE 5: MNE du site d'entraînement de la base de Caylus obtenu aux moyens de relevés Lidar aéroportés. La résolution au sol est de 10 centimètres. On dénombre un plus grand nombre d'artefacts que sur les données Lidar.

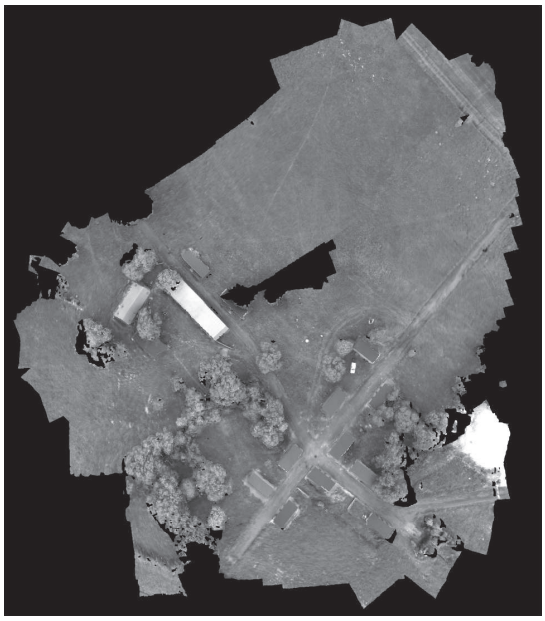

FIGURE 6: Orthomosaïque obtenue en appliquant à chaque pixel du MNE de la figure 4 la moyenne des niveaux de gris relevés dans les images clés dans lesquelles la cellule est visible.

en prenant la moyenne des niveaux de gris relevés dans les images-clés. Un exemple est présenté en figure 6.

Dans le cas d'un MNE produit par stéréovision multivue, l'opération est triviale puisque l'on peut appliquer comme niveau de gris (ou couleur) la moyenne des pixels associés aux points 3D les plus élevés de chaque cellule au sol. Un exemple est présenté en figure 7.

3. Détection, pistage et localisation d'objets mobiles

Nous présentons ici un module offrant une première capacité de détection et de suivi d'objets mobiles. La détection est basée sur l'analyse du flot optique fourni par le même algorithme que celui utilisé pour la génération de MNE : eFolki. Quand un objet mobile est détecté, la position de cet objet d'intérêt et sa taille permettent d'ini-

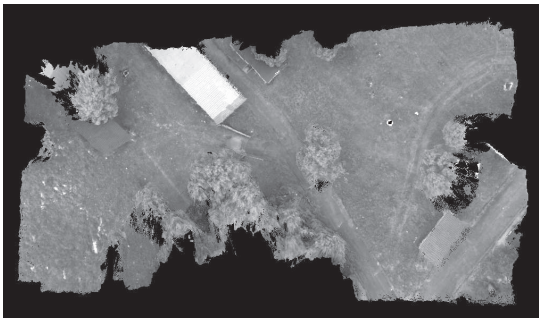

FIGURE 7: Orthomosaïque superposable au MNE de la figure 5.

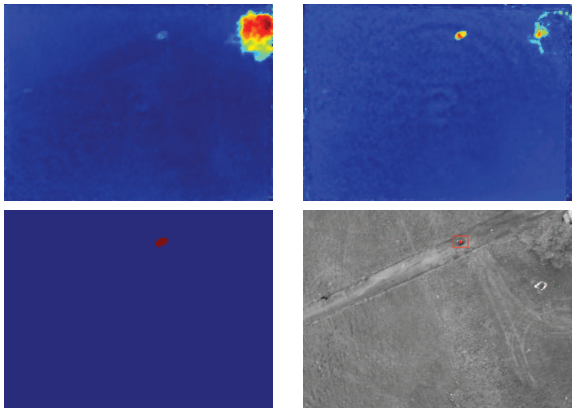

FIGURE 8: Détection d'objets mobiles. De haut en bas, de gauche à droite : moyenne de flot optique résiduel, écart type du flot optique résiduel, blob détecté par seuillage par hystérésis et boîte englobante sur l'image vidéo.

tialiser le suivi par l'algorithme *Tracking-Learning-Detection* (TLD) de Kalal et al. (2011).

3.1. Détection d'objets mobiles

Le flot optique entre deux images combine une composante liée au mouvement du capteur (rotation et translation de la caméra) par rapport à la scène 3D observée et une composante liée à la présence d'objets mobiles. La détection des objets mobiles nécessitent donc que l'on puisse supprimer la première composante.

Comme les objets à détecter évoluer au sol, nous faisons l'hypothèse que la scène observée peut être modélisée par un plan, représentant le sol et majoritaire dans le champ de vue de la caméra, associé à des éléments de structure 3D. Sous cette hypothèse, le mouvement du sol est décrit comme une homographie qui est estimée de façon robuste à l'aide d'un estimateur de type RANSAC exploitant des appariements de points caractéristiques extraits des images. En soustrayant la composante homographique au flot optique mesuré, on obtient la part résiduelle correspondant aux objets 3D fixes et mobiles. En accumulant ces indices sur plusieurs couples d'images $\{I_{k-n}, I_k\}...\{I_{k-1}, I_k\}$, on améliore le ratio signal à bruit et on peut distinguer des objets mobiles par analyse des moments des résidus du flot optique. La figure 8 montre la détection d'un petit robot terrestre évoluant sous le drone par seuillage par hystérésis de l'écart type du flot résiduel.

3.2. Pistage d'objets mobiles et localisation

L'algorithme TLD est initialisé par la boite englobante centrée sur la détection fournie par la méthode décrite dans la section précédente. Les dimensions sont fixées afin de couvrir l'intégralité du blob de détection (voir 8). Le suivi de l'objet est basé sur le suivi de primitives image extraites dans la portion d'images correspondante. Pendant le suivi d'image à image, l'aspect de l'objet peut peut changer, cependant avec de faibles variations. TLD est capable d'apprendre ces différents aspects et ainsi de redétecter l'objet - après une erreur de suivi ou après une sortie de l'objet du champ de la caméra - s'il se représente sous un aspect déjà appris.

A chaque instant, l'objet détecté et suivi est localisé en calculant l'intersection du (1) rayon issu du centre optique et passant par le centre de la boîte englobante et (2) du MNE. La figure 9 montre la trajectoire du robot terrestre après traitement d'une séquence vidéo acquise par un drone de l'ONERA au dessus d'une zone préalablement cartographiée.

4. Détection d'objets génériques et cartographie sémantique

Outre la détection basée sur des primitives de mouvement ou géométriques, nous proposons des outils pour la compréhension de scène afin de permettre la planification opérationnelle. D'abord, un apprentissage interactif d'objets d'intérêts présents dans la carte visuelle est réalisée dans la station-sol afin de générer des cartes sémantiques superposables à l'emprise de la zone observée. Ensuite les détecteurs d'objets ainsi entraînés sont transformés pour être appliqués à de nouvelles vidéos acquises par le drone, permettant ainsi la reconnaissance d'objet en vol (cf. Fig. 10).

4.1. Apprentissage interactif d'objets d'intérêt

Les orthomosaïques informent un opérateur sur la vision globale de la scène et lui permettent d'améliorer les capacités des drones en désignant explicitement des objets d'intérêt : par exemple les arbres qui sont des obstacles au vol ou bien les véhicules ou les victimes qui présentent un intérêt du point de vue de la mission. Sur la station-sol, l'opérateur sélectionne des zones qui contiennent des exemples des objets d'intérêt et des zones qui n'en contiennent pas (fournissant ainsi au système des exemples négatifs). Le système utilise ces exemples d'apprentissage pour estimer un classifieur de primitives de l'image et fournit en retour une carte de classification. Le processus est itératif de manière à ce que l'opérateur puisse fournir de nouveaux exemples pour affiner la classification.

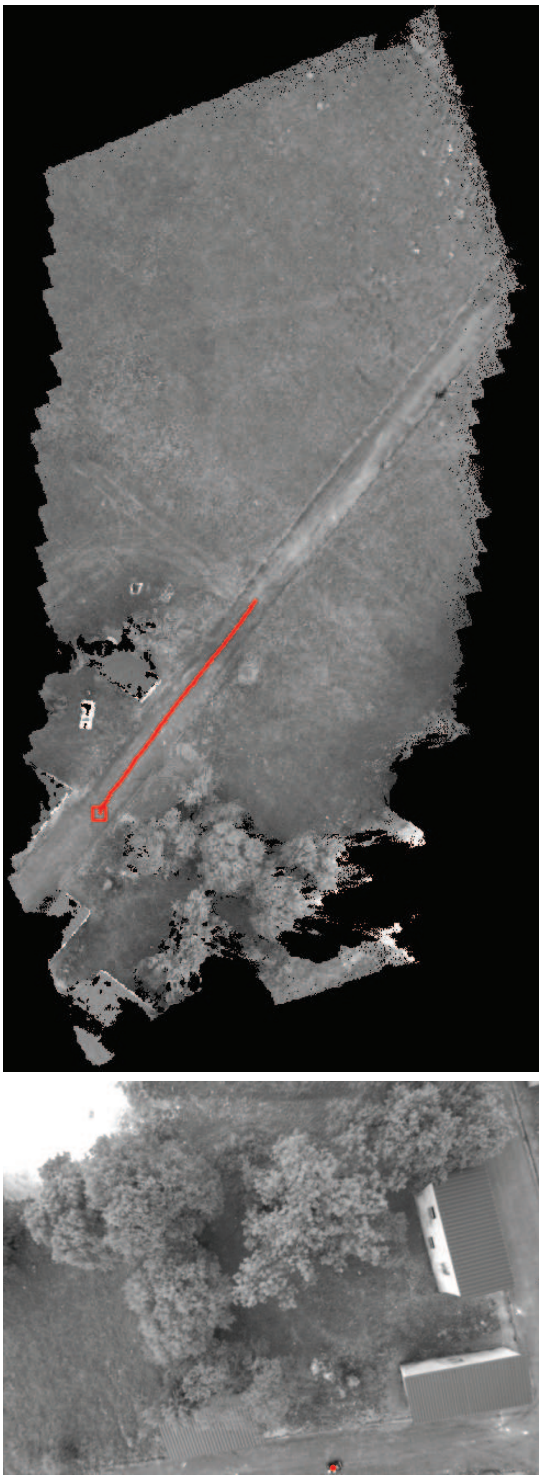

FIGURE 9: Trajectoire d'un objet mobile projetée sur une orthomosaïque (à gauche) et la dernière détection du véhicule avant qu'il ne sorte du champ de vue de la caméra (à droite). Cette détection est marquée par un carré rouge sur l'orthomosaïque.

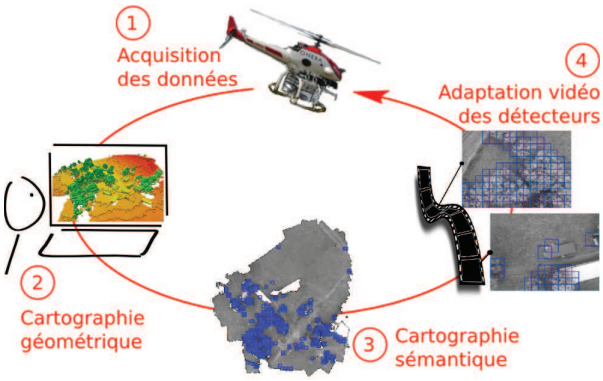

FIGURE 10: Synoptique pour la cartographie sémantique : les détecteurs d'objet sont appris interactivement sur la station-sol, puis appliqués à l'orthomosaïque entière pour créer des cartes sémantiques de la zone survolée, et enfin adaptés au flux vidéo pour une utilisation embarquée sur le drone.

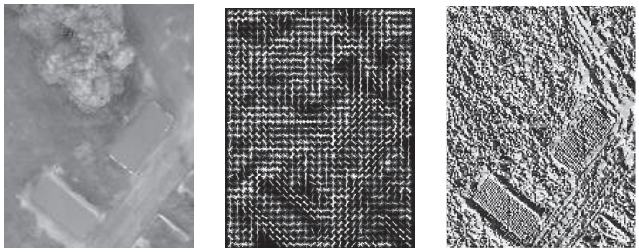

FIGURE 11: (a) Détail de l'orthomosaïque et représentations des descripteurs qui en sont extraits : (b) Histogrammes de gradients orientés comme descripteurs d'orientations des contours et (c) Motifs Linéaires Binaires comme descripteurs de texture.

Nous détaillons maintenant l'implémentation de la méthode de Le Saux (2014). Le système extrait des patchs (imagettes) dans les zones sélectionnées et calcule des descripteurs d'apparence pour chacun. Ces descripteurs constituent l'ensemble d'entraînement (avec des exemples positifs et négatifs) qui sert à apprendre la cible d'intérêt. Nous utilisons une combinaison d'histogrammes de gradients orientés (qui furent initialement introduits pour la détection de personnes mais son maintenant utilisés pour la détection d'objets génériques) et de motifs binaires locaux (qui encodent la texture) (cf. Fig 11). Ces descripteurs ont fait leurs preuves tant pour les images aériennes (Chauffert et al., 2012) que pour les vidéos standards.

L'apprentissage rapide et en-ligne d'un classifieur est ensuite effectué par un algorithme de type *online gradient boost*. Le *boosting* est une approche d'apprentissage machine qui vise à construire un méta-classifieur performant à partir d'un ensemble de classifieurs faibles : dans le cas présent, les composantes du descripteur d'apparence. L'approche *gradient boost* permet de faire face à deux problèmes majeurs soulevés par l'interactivité dans l'apprentissage : les erreurs d'étiquetage (le processus de sélection de zones peut pâtir d'un dessin imprécis ou de l'attribution d'une étiquette erronée, ce qui implique

plus de données mal-étiquetées que dans un ensemble rigoureusement constitué) et le déséquilibre entre les ensembles d'exemples positifs et négatifs (dans les images traitées, les exemples d'intérêt sont rares tandis qu'il est facile de designer des zones sans intérêt particulier).

4.2. Détection pour la cartographie sémantique

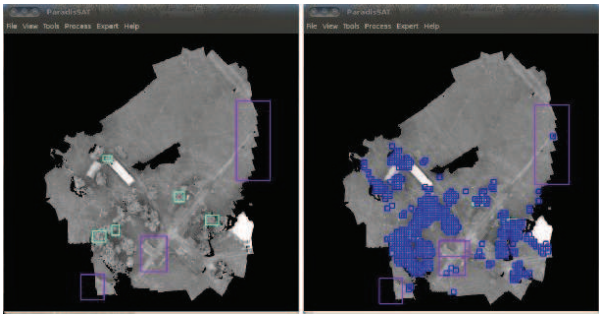

FIGURE 12: Interface d'apprentissage interactif pour un détecteur d'obstacle (à gauche) et la carte de détections résultant de la classification par *online gradient-boost*.

Une fois que l'apprentissage est effectué sur quelques zones, la classification de l'image entière a lieu (cf. Fig. 12). Tout d'abord, les zones visibles de l'orthomosaïque sont segmentées en patchs et indexées par les descripteurs d'apparence. Ces descripteurs sont les données d'entrée du classifieur, qui retourne une valeur indiquant la confiance sur la décision que les patchs testés représentent ou non l'objet cible. En seuillant la sortie du classifieur, nous pouvons produire des cartes de détection superposables à l'orthomosaïque. Dans le cas des drones, deux applications sont visées. Les cartes de détections de cibles (comme les personnes ou les véhicules) sont utiles pour définir l'objectif du vol des drones et planifier le chemin qui y mène. Les cartes de détection d'obstacle (comme les arbres ou les bâtiments) sont utiles pour planifier les chemins qui évitent les principaux dangers, spécialement en phase d'approche d'une cible.

4.3. Adaptation des détecteurs au domaine vidéo

Dans les vols successifs du drone, les paramètres du classifieur sont utilisés dans un détecteur d'objet qui fonctionne sur le flux vidéo et y détecte les objets d'intérêt (cf. Fig. 13). La compacité du modèle permet de le télécharger sur le drone même avec une bande passante limitée. Cependant, l'orthomosaïque est obtenue par synthèse d'image et a des caractéristiques (point de vue, résolution, etc.) différentes des images acquises par la caméra embarquée du drone, si bien qu'une adaptation de domaine est nécessaire pour transférer les classifieurs dans une géométrie différente. En outre, au cours des vols successifs, le drone peut explorer des zones nouvelles avec des motifs jamais vus auparavant : cela doit être pris en compte pour conserver des performances de détections satisfaisantes.

L'adaptation géométrique consiste à recaler les patchs extraits des images vidéo dans le plan de l'orthomosaïque. L'homographie exacte qui relie les deux plans-image peut être calculée grâce aux paramètres intrinsèques (qui sont connus a priori) de la caméra et de la position 3D du drone (donnée par le GPS) - cf. Le Saux et Sanfourche (2013). En pratique, étant donné que le point de vue de la caméra est vertical par rapport au drone, les rotations sont dues seulement aux mouvements du drone (dont le vol est régulier et sans inclinaison trop marquée), donc la plupart des zones de l'image vidéo peuvent être classées.

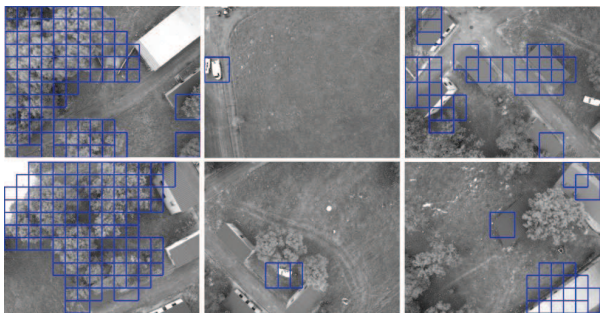

FIGURE 13: Détections après adaptation de domaine dans le flux vidéo pour différents classifieurs : végétation (à gauche), véhicules (au milieu) et bâtiments (à droite).

L'adaptation de domaine statistique vise à transférer des classifieurs appris sur un domaine source étiqueté vers un domaine cible sans étiquettes, où les données de test peuvent être plus ou moins différentes. Nous proposons une méthode pour adapter les détecteurs d'objet aux caméras du drone en utilisant de nouveaux exemples extraits des nouvelles images de test (acquises en vol). Notre approche utilise des contraintes externes sur les instances de nouveaux exemples potentiels (par exemple une piste reliant le même objet dans plusieurs images vidéo successives) pour choisir des échantillons non-étiquetés qui vont vraisemblablement améliorer les classifieurs. De telles contraintes peuvent être extraites d'une manière non-supervisée dans le domaine cible. Par exemple un objet détecté dans une image doit être pistable dans plusieurs image consécutives, et les différents objets de la piste doivent également être des détections valides. Ces nouveaux échantillons sont ajoutés aux originaux pour former un nouvel ensemble d'apprentissage, sur lequel est entraîné un détecteur mis à jour qui est adapté à l'environnement dans lequel évolue le drone à ce moment.

5. Conclusion

Nous avons présenté les composantes nécessaires à un schéma complet de perception de l'environnement des drones. Ces différentes fonctionnalités permettent de résoudre le problème de la détection d'objets et leur localisation dans un modèle 3D reconstruit de l'environnement. L'équipement embarqué sur le drone pour induire cette capacité de perception consiste a minima en une caméra calibrée et un GPS, un Lidar pouvant être utilisé en sus pour une meilleure précision. Avec cette

configuration nous sommes à même de fournir divers produits :

- Orthomosaïques ;
- Modèle numérique d'élévation (MNE) ;
- Cartographie sémantique de la végétation et des bâtiments ;
- Détection et pistage d'objets mobiles résolus.

Ces produits sont géo-localisés grâce au MNE avec la précision fournie par le GPS embarqué.

L'objectif d'un tel système est d'une part de fournir des outils aux opérateurs de drones pour appréhender avec justesse l'environnement dans lequel évolue leur drone et planifier leurs missions, et d'autre part de mettre à profit la présence et la compétence des experts humains pour ajouter du sens aux modèles géométriques générés.

Références

Chauffert, N., Israël, J., Le Saux, B., july 2012. Boosting for interactive man-made structure classification. Dans : IEEE International Geoscience and Remote Sensing Symposium. Munich, Germany.

Fraundorfer, F., Heng, L., Honegger, D., Lee, G.-H., Meier, L., Tanskanen, P., Pollefeys, M., 2012. Vision-based autonomous mapping and exploration using a quadrotor mav. Dans : IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vilamoura, Portugal.

Kalal, Z., Mikolajczyk, K., Matas, J., 2011. Tracking-learning-detection. Pattern Analysis and Machine Intelligence.

Le Saux, B., 2014. Interactive design of object classifiers in remote sensing. Dans : Proc. of International Conference on Pattern Recognition. Stockholm.

Le Saux, B., Sanfourche, M., 2013. Rapid semantic mapping : Learn environment classifiers on the fly. Dans : Proc. of International Conference on Intelligent Robots and Systems. Tokyo.

Lourakis, M., Argyros, A., 2009. SBA : A Software Package for Generic Sparse Bundle Adjustment. ACM Trans. Math. Software 36 (1), 1–30.

Lowe, D. G., novembre 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60 (2), 91–110.

Nuechter, A., Hertzberg, J., 2008. Towards semantic maps for mobile robots. Robotics and Autonomous Systems 56, 915–926.

Plyer, A., Le Besnerais, G., Champagnat, F., 2014. Massively parallel lucas kanade optical flow for real-time video processing applications. Journal of Real-Time Image Processing.

Sanfourche, M., Vittori, V., Le Besnerais, G., 2013. evo : A real-time embedded stereo odometry for mav applications. Dans : Proc. of International Conference on Intelligent Robots and Systems. Tokyo.

Shi, J., Tomasi, C., 1994. Good features to track. Dans : 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94). pp. 593 – 600.

Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W., 2000. Bundle adjustment - a modern synthesis. Dans : Proceedings of the International Workshop on Vision Algorithms : Theory and Practice. ICCV '99. London, UK, pp. 298–372.

Wurm, K. M., Hornung, A., Bennewitz, M., Stachniss, C., Burgard, W., mai 2010. OctoMap : A probabilistic, flexible, and compact 3D map representation for robotic systems. Dans : Proc. of the ICRA 2010 Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation. Anchorage, AK, USA.
URL `http://octomap.sf.net/`