Environment Mapping & Interpretation by Drone

Martial Sanfourche, Bertrand Le Saux, Aurélien Plyer, Guy Le Besnerais
ONERA The French Aerospace Lab
F-91761 Palaiseau, France

***Abstract*—In this paper we present the processing chain for geometric and semantic mapping of a drone environment that we developed for search-and-rescue purposes. A precise 3D modelling of the environment is computed using video and (if available) Lidar data captured during the drone flight. Then semantic mapping is performed by interactive learning on the model, thus allowing generic object detection. Finally, tracking of moving objects are performed in the video stream and localized in the 3D model, thus giving a global view of the situation. We assess our system on real data captured on various test locations.**

I. Introduction

Active research have been carried out for automatically producing maps of the environment explored by a drone. Typically, those maps show 2D or 3D geometric representations, combined with sensor-based information such as textured images. The underlying assumption is that these models constitute a first step toward drone autonomy, by allowing ego-localization and trajectory planning [1]. The next step is semantic mapping that provides informations object localization and regions of interest [2].

Though, in many real-life situations, such as urban monitoring or search-and-rescue (SAR) operations after an industrial accident or a natural disaster, drones are not fully autonomous but at least partially remotely operated. In the ground control station, qualified professionals collaborate closely with the drone operators to conceive the intervention scheme. In this paper, we propose a complete workflow for mapping and interpreting the drone environment. Our approach aims at setting the users at the center of the system by using their expert knowledge and provinding them in return with a global view of the situation which helps them making decisions.

The article is organized as follows. In part II, we present our approach for estimating precisely the drone trajectory and modelling in 3D the environment: buildings, trees, obstacles, ground, etc. We detail our method for remote detection of moving objects and events of interests from a moving platform in part III. Finally, we propose interactive tools for interpretation of objects and areas of interest in part IV, before presenting the results of experiments in part V.

II. Offline 3D mapping

This module produces high resolution geographical data from video (and eventually Lidar) sensors mounted on the drone. The main objective is to obtain a 3D model of the observed scene (cf. Fig. 1), that will allow to produce easily interpretable products like orthomosaics and Digital Elevation Models (DEMs). The proposed processing chain decomposes in precise drone trajectory estimation, calculation of the 3D voxel model and DEM, and finally DEM texturing by video mosaicing and orthorectification.

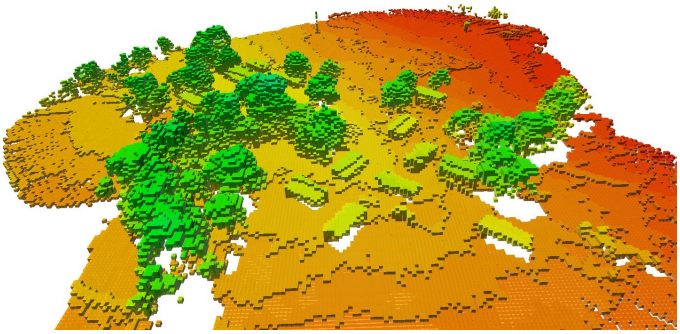

Fig. 1. 3D model of the area explored by the drone, built using the precise flight-trajectory estimate obtained by bundle adjustment of the video images.

A. Refinement of trajectory parameters by bundle adjustment

In spite of drone localization using a precise GPS RTK receiver and good (but let precise) attitude measurements, raw navigation measurements often lead to geometric inconsistencies and reconstruction artefacts, as shown in Fig. 2. We use bundle adjustment [3], [4] to refine simultaneously the trajectory and 3D structure parameters (such as a map of 3D landmarks) using heterogeneous data: video-frames, GPS positions and 3D measurements if available.

Practically, we first detect points of interest in the K video frames and match them to create N tracks. Then, bundle adjustment consists in estimating the K image-capture parameters and the N 3D positions in the local reference system of the landmarks defined by the tracks. In the following, we use the notations defined in Tab. I

1) Objective function: The standard objective function that bundle adjustment aims to minimize is the cumulative matching cost of back-projection of the landmarks in the images (first term of Eq. 1). For better precision, we added a constraint that

TABLE I. Notations for Precise Trajectory Estimation

T_k^i, Θ_k^i	Estimates (at iteration i) of position and attitude (3 Euler angles) of the camera (at capture time k)
X_n^i	Estimate (at iteration i) of the landmark n position
$u_{k,n}^{obs}$	Position of landmark n as seen in image k
p	Known camera parameters (camera intrinsic parameters, camera position on the drone, relative pose between sensors)
Π	Sensor model to compute the projection of X_n^i in image k (given p, T_k^i et Θ_k^i)

978-1-4799-6652-3/15/$31.00

Fig. 2. Comparison of an aerial image (source: IGN) and a altitude map obtained using Lidar data set in the same reference frame from raw navigation measurements. Ghost buildings appear and the scene geometry is distorted, which means a localization drift occurs.

implies position estimates can not drift too much away from the GPS RTK measures $\{T_k^{obs}\}_{k\in[1\dots K]}$. The new objective function is then:

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

where $\delta(k,n)$ indicates if $u_{k,n}^{obs}$ exists (i.e. if X_n^i appears in image k), $\|\bullet\|_W$ is the Mahalanobis distance defined by the covariance matrix W, and W_T is a covariance matrix that allows a localization error of $15cm$ (based on the GPS precision) with respect to the 3 directions.

2) Hierarchical optimization: In order to process video sequences of several thousand images and even more 3D landmarks, we propose a 2-step approach.

a) Bundle adjustment on key-frames: Due to strong redundancy between adjacent video-frames, the sequence can be summarized by a limited numer of key-frames. However, enough spatial covering is needed to allow matching of features between key-frames. Following a procedure previously used for Simultaneous Localisation and Mapping (SLAM) [5], key-frames are chosen according to statistics of feature tracking over the sequence. This iterative algorithm proceeds as follows. In the first image, C points of interest are detected using a Harris corner algorithm. Those points are tracked in subsequent images using a KLT tracker [6]. Tracking errors $k \rightarrow k+1$ are detected by reversing the KLT tracker from $k+1 \rightarrow k$ and checking the dispersion error in image k. When the matching number drops below a given ratio (typically 80% of the tracked points), the current image defines a new key-frame. C new points of interest are detected in this key-frame and the procedure iterates.

Then a loop-closing procedure allows to integrate global structure across the sequence. GPS information is used to link key-frames which are distant temporally but cover the same geographic area, under the assumption of a flat scene. Points of interest in both key-frames are matched using SIFT descriptors [7], thus allowing to extend tracks over the whole sequence. Eq. 1 is finally minimized on a reasonable number of variables using a Levenberg-Marquardt-type algorithm that benefits from the sparsity of the Hessian matrix associated to the objective function [4].

b) Fast processing of remaining images: Each image is now treated separately to insure a fast process. Given the 3D landmark locations estimated during the previous step, the pose of each image is computed using the 2D/3D matching induced by KLT tracks.

B. Digital elevation models and orthomosaïcs

A DEM represents the scene relief as an image in which each pixel corresponds to a cell of given resolution in the horizontal plane and stores the height of the 3D measurements falling into the corresponding cell. For the Lidar sensor, the depth is directly provided and the location of reflecting 3D points can be easily deduced from the sensor parameters (angular resolution). If only the monocular camera is mounted, we use stereovision for triangulating 3D points. We exploit the viewing parameters refined by bundle adjustment and a very efficient GPU-based optical flow algorithm: eFolki [8] to obtain robust matchs between two successive keyframes. Each cell is then given the maximum altitude of all the 3D points that fall inside.

Orthomosaics are synthetic aerial images corrected of the relief effects. Practically, generating an orthomosaic consists in texturing the DEM. In case of Lidar-based DEM, one needs to get the additional colorimetric information from the camera. The centre of each DEM cell is projected in the key-frames (while checking they are observable using a Z-buffer-like procedure) and we take the average of the intensity levels of the retrieved image pixels. For a camera-only DEM, the orthomosaic is easy to build by considering the average intensity level of the pixels with the highest altitudes that fall into the corresponding cell.

III. Event and moving object detection

We present here a module for detection and tracking of moving objects. Detection is based on the analysis of the image motion or optical flow. We exploit here the same optical flow algorithm as used to compute the depth maps in the 3D mapping module (see section II). Motion-based detections are then used to initialize a more robust appearance-based tracking, using the Tracking-Learning-Detection (TLD) algorithm [9].

A. Moving object detection

Detection of a moving object in stereo could be obtained by normalizing 3D residuals of the sene reconstruction with the hypothesis of a static environement such as in [10]. The probleme is more difficult with a monocular camera. We have to consider the residual which respect to the 3D structure of the scene which consists of an epipolar flow, and possible missed-detection cases (for example when target and drone have coplanar trajectory).

As we are interested to ground moving vehicles, we model the scene as a ground plane with some 3D elements. Under the the assumption that the ground plane is mainly in the image field of view, the ground motion is modelled by a homography which is robustly estimated by RANSAC from image features matched in successive views. By subtracting the estimated

homographic optical flow to the raw image flow, we obtain the residual part that corresponds to 3D objects and moving objects. We accumulate these indicators for multiple image couples in order to improve the signal to noise ratio and (in case of erroneous labelling of 3D object) for distinguishing between 3D objects and moving objects by analysis temporal statistical moments of the residual optical flow. The detection is done by hysteresis thresholding of the standard deviation and by taking the local maximum.

B. Moving object tracking and localization

Successive detections at close image locations indicates a probable target, and trigger the tracking algorithm. TLD tracks image features located inside a bounding box that is initialized around the detection blob. During the frame-to-frame tracking, the object changes its aspect smoothly. TLD learns the different aspects of the object to be able to redetect it in case of object occlusion (behind a 3D element of the scene or when the object goes outside of the camera field of view). At each moment, the object detected and tracked is located in the global scene model by computing the intersection of (1) the ray coming from the optical center and passing by the center of the bounding box and (2) the DEM.

IV. Generic object detection

The third component of our system aims at developping scene-understanding tools which help the operational planning: object classifiers applied the geometric model and to images of the drone camera.

A. Interactive learning and semantic mapping

Orthomosaïcs give a global overview of the scene. They are used to learn interactively classifiers of objects of interest. The operator selects examples of interesting areas and useless areas, from which small patches are extracted and then indexed by appearance features (histograms of oriented gradients) to constitute a training set. Fast online learning is then performed by online gradient boost, a variant of boosting which allows us to deal with two major problems raised by interactive learning: mislabelling (due to imprecise selection or wrong labelling) and unbalanced datasets (in the images we are dealing with, positive samples are often scarce) [11].

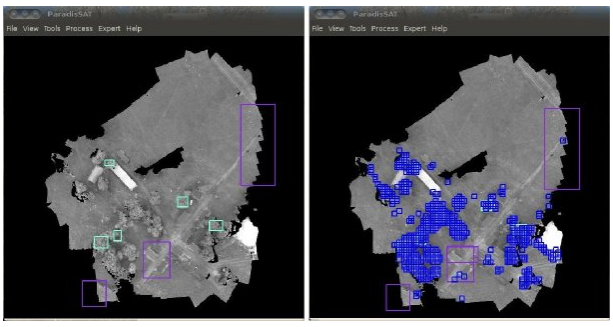

Fig. 3. Interactive learning interface on the orthomosaïc (left) and resulting semantic map by *online gradient-boost* (right).

Once the training is done on a few areas, classification can be processed on the whole image by a standard sliding-window approach (cf. Fig. 3). For drone planning, the interest is two-fold. Target-detection maps (like cars or persons) are useful for defining the target of the drone flight and thus planning the path that leads to it. Obstacle-detection maps (such as trees or buildings) are useful for planning paths that avoid potential dangers, especially when approaching the target.

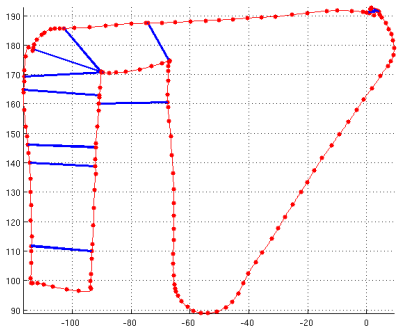

Fig. 4. Graph of the key-frames of the video sequence captured by the drone, displayed according to the drone location. Temporal links between key-frames are shown in red, while geographical (i.e. loop-closing-obtained) links are shown in blue.

B. Adaptation of detectors to the video domain

During successive flights, the classifier parameters are then used in a detector that performs on frames of the video flow and detects the objects of interest in them. These images have viewing angle and resolution that differ from the orhomosaïc ones. Geometric adaptation is performed by rectifying patches extracted in the video-frames, using the exact homography that depends on the camera parameters and the drone position given by the GPS [12].

V. Experiments

The platform used for experiments is a Yamaha RMAX helicopter with a 1,3MP monochrome camera for video and a 4-line-scan laser measurement sensor for range data. The localization of the helicopter is given by a decimeter-class GPS system. Several flights were performed in various locations that present peri-urban scenes with buildings, trees and open areas.

c) 3D mapping: Fig. 4 shows the result of the bundle adjustment step of the 3D mapping (cf. section II-A) performed over selected key-frames captured by the drone. Fig. 5 which shows $10cm$ resolution DEM and orthomosaïc built using only images and monocular stereovision.

d) Event and moving object detection: Fig. 6 and Fig. 7 show respectively the motion-base detection and appearance-based tracking of targets of interest (for example someone trying to catch the drone's attention in an emergency situation). Video of the whole process can be viewed online [1].

e) Generic Object Detection: In Fig. 8, we present some results of the adapted classifiers of section IV used on new images captured by the drone: vegetation, vehicles, and buildings. Video of the whole process can be viewed online [2].

[1] https://www.youtube.com/watch?v=JyHaeBkvKTQ
[2] https://www.youtube.com/watch?v=OTxaLcouOHE

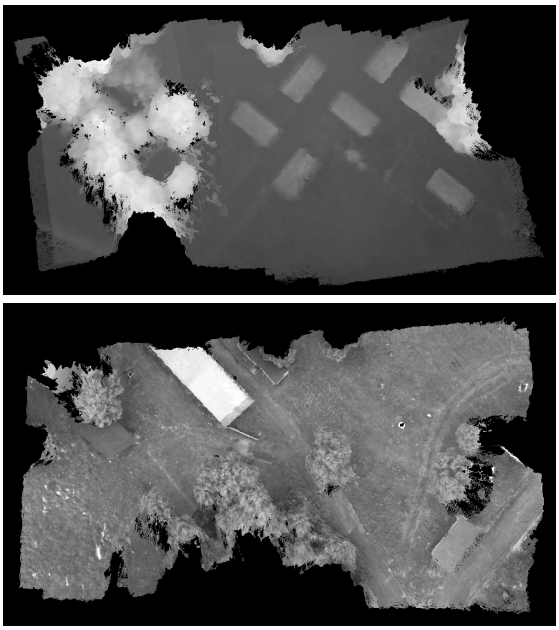

Fig. 5. DEM with image-only data ($10cm$ resolution) and corresponding orthomosaïc obtained using stereovision.

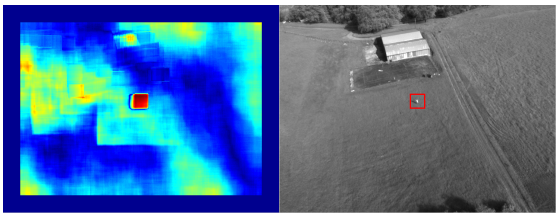

Fig. 6. Residual motion detection by optical flow (left) and corresponding detection of moving object(right).

VI. Conclusion

We have presented a complete workflow for environment mapping using remote-sensing from a drone. This system allows to sense and understand objects and regions of interest, and to localize them in a 3D model. On-board equipment is minimal: a calibrated camera and a standard GPS is enough, even if a Lidar can be added to obtain a better precision. With this setup, the system is able to deliver geo-localized orthomosaïcs, DEMs, semantic maps, alarms for moving objects.

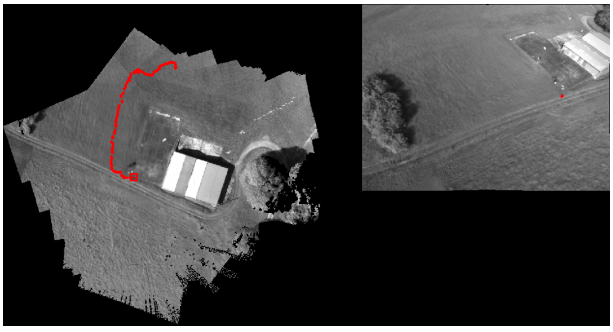

Fig. 7. Tracking of the objects in the video (right) and simultaneous localization on the orthomosaïc (left).

Such a system aims first at providing drone operators with tools for a better understanding of the environment in which the drone evoluates, and second, to benefit from qualified experts for adding knowledge to the geometric models.

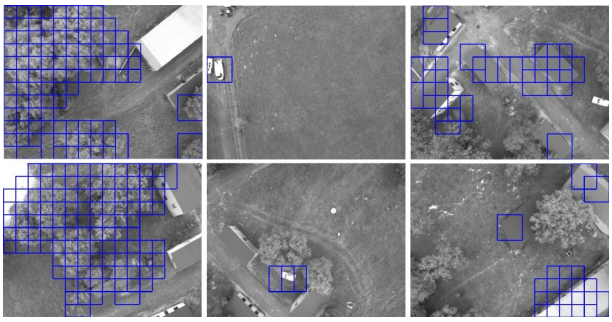

Fig. 8. Detections after adaptation to the video-domain for various objects classifiers: vegetation (left), vehicles (middle), and buildings (right).

Acknowledgment

This research was partially funded by the European project DARIUS. The authors gratefully acknowledge the Department of Systems Control and Flight Dynamics of ONERA for data captured in real-world conditions.

References

[1] F. Fraundorfer, L. Heng, D. Honegger, G.-H. Lee, L. Meier, P. Tanskanen, and M. Pollefeys, "Vision-based autonomous mapping and exploration using a quadrotor mav," in *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, 2012.

[2] A. Nuechter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, vol. 56, pp. 915–926, 2008.

[3] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - a modern synthesis," in *Proc. of ICCV/W. on Vision Algorithms: Theory and Practice*, London, UK, 2000.

[4] M. Lourakis and A. Argyros, "SBA: A Software Package for Generic Sparse Bundle Adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 1–30, 2009.

[5] M. Sanfourche, V. Vittori, and G. Le Besnerais, "evo: A realtime embedded stereo odometry for mav applications," in *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013.

[6] J. Shi and C. Tomasi, "Good features to track," in *1994 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94)*, 1994.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

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

[9] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *Pattern Analysis and Machine Intelligence*, 2011.

[10] M. Derome, A. Plyer, M. Sanfourche, and G. Le Besnerais, "Real-Time Mobile Object Detection Using Stereo," in *Proc. of Int. Conf. on Aut. Rob. and Comp. Vis. (ICARCV)*, Marina bay sands, Singapore, Dec. 2014.

[11] B. Le Saux, "Interactive design of object classifiers in remote sensing," in *Proc. of Int. Conf. on Patt. Rec. (ICPR)*, Stockholm, Sweden, 2014.

[12] B. Le Saux and M. Sanfourche, "Rapid semantic mapping: Learn environment classifiers on the fly," in *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013.