Multimodal Multispectral Imaging System for Small UAVs

Trym Vegard Haavardsholm, Torbjørn Skauli, and Annette Stahl

***Abstract*—Multispectral imaging is an attractive sensing modality for small unmanned aerial vehicles (UAVs) in numerous applications. The most compact spectral camera architecture is based on spectral filters in the focal plane. Vehicle movement can be used to scan the scene using multiple bandpass filters arranged perpendicular to the flight direction. With known camera trajectory and scene structure, it is possible to assemble a spectral image in software. In this letter, we demonstrate the feasibility of a novel concept for low-cost wide area multispectral imaging with integrated spectral consistency testing. Six bandpass filters are arranged in a periodically repeating pattern. Since different bands are recorded at different times and in different viewing directions, there is a risk of obtaining spectral artifacts in the image. We exploit the repeated sampling of bands to enable spectral consistency testing, which leads to significantly improved spectral integrity. In addition, an unfiltered region permits conventional 2D video imaging that can be used for image-based navigation and 3D reconstruction. The proposed multimodal imaging system was tested on a UAV in a realistic experiment. The results demonstrate that spectral reconstruction and consistency testing can be performed by image processing alone, based on visual simultaneous localization and mapping (VSLAM).**

***Index Terms*—Computer Vision for Other Robotic Applications, Mapping, Surveillance Systems.**

I. Introduction

SMALL unmanned aerial vehicles (UAVs) have evolved rapidly over the past decades, offering many new opportunities as platforms for remote sensing. By exploiting the platform movement, UAV imagery can be used to estimate the 3D structure of the ground landscape, as well as the sensor movement itself, using visual simultaneous localization and mapping (VSLAM) techniques. A sensing modality of increasing interest is spectral imaging, which enables a wide variety of applications including environmental monitoring, land use mapping, precision agriculture and forestry, search and rescue operations and military reconnaissance [1].

From a physics point of view, an imaging sensor on a UAV has access to very rich information about the landscape below, through the spatial contrasts and spectral distribution of the incoming light. In principle, the diffraction limit allows a compact camera to resolve the hemisphere underneath a UAV into an image many thousands of pixels across. Furthermore, hyperspectral imaging sensors can in principle resolve the spectral range into hundreds of spectral bands. It is difficult, however, to make sensors that exploit the information carried by incoming light and fit within the size and weight limits of a small UAV. Conventional imaging spectrometers such as HySpex Mjolnir [2] record accurate spectra, but have complex optics and limited pixel count relative to their size. Fabry-Pérot spectral imagers can be compact [3], [4], but are subject to photon noise due to narrow bandwidth, and motion artifacts due to non-simultaneous sampling of bands. Multispectral imagers are available with multiple cameras recording one band each, but such systems grow in size when a large band count and/or good light throughput is needed.

Here we consider a different sensor technology tradeoff that enables the collection of rich spatial information with moderate spectral resolution from a UAV. The sensor is basically a regular camera, but with a specialized filter layout in the focal plane. Different spectral bands are recorded in succession thanks the platform movement, similar to *pushbroom* imaging spectrometers commonly used for hyperspectral imaging. VSLAM techniques are employed for the best possible coregistration of the successively recorded bands. Artifacts in the recorded spectra can still occur due to coregistration inaccuracies, or due to parallax effects or movement in the scene. A novel aspect of the sensor concept is therefore to sample each band multiple times in an interleaved fashion, to enable consistency testing of the recorded spectra. The resulting UAV payload enables efficient collection of spectral and spatial imagery, essentially by making a compromise on spectral resolution and offloading much of the image formation to software.

The imaging hardware concept used here has previously been tested with simplistic imaging geometries and processing methods [5], [6]. In this letter, we build on this work with the following contributions:

1) A new processing chain based on VSLAM.
2) A physics-informed spectral consistency test.
3) Field validation in realistic conditions.

We show that the system accommodates real-world motion and geometry to produce useful image products. This is substantiated by demonstrating signature-specific spectral detection of a challenging target. Thus we demonstrate the feasibility

Manuscript received September 10, 2019; accepted January 6, 2020. Date of publication January 17, 2020; date of current version January 30, 2020. This letter was recommended for publication by Associate Editor T. Peynot and Editor E. Marchand upon evaluation of the reviewers' comments. This work was supported by the Norwegian Research Council through the Centre for Autonomous Marine Operations and Systems (AMOS) at NTNU. *(Corresponding author: Trym Vegard Haavardsholm.)*

T. V. Haavardsholm is with the Norwegian Defence Research Establishment (FFI), Kjeller NO-2007, Norway, and also with the Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim NO-7491, Norway (e-mail: trym.haavardsholm@ffi.no).

T. Skauli is with the Norwegian Defence Research Establishment (FFI), Kjeller NO-2007, Norway (e-mail: torbjorn.skauli@ffi.no).

A. Stahl is with the Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim NO-7491, Norway (e-mail: Annette.Stahl@ntnu.no).

This letter has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/LRA.2020.2967301

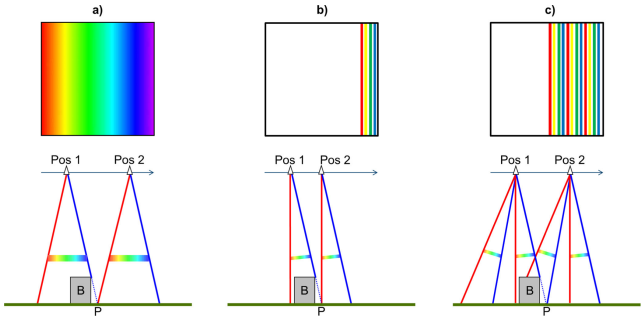

Fig. 1. Different concepts for spectral imaging using a filter in the focal plane (top) in a camera viewing the ground from a UAV (bottom). Coloured lines indicate the first and last band. (a) Linear Variable Filter (LVF). (b) Strips of discrete spectral filters. (c) A set of spectral filters repeated multiple times.

of a concept that can bring spectral imaging within reach for new applications, by physics-informed exploitation of computer vision techniques.

II. Imaging Concept

Cameras with patterned filters on a 2D photodetector array, as found in conventional color cameras, represent the most compact architecture in use for spectral imaging.

A color camera employs a single image sensor with a pattern of spectral filters to record one of the three different primary colors in each pixel. The two missing primary colors are filled in by interpolation, to reconstruct a visually pleasing color image. With a larger number of bands, such simplistic spectral reconstruction requires strong assumptions on the properties of the scene, and is normally not feasible. Instead, it is necessary to collect full spectral information from each image pixel.

This requirement arises from the fact that hyper- and multi-spectral imaging techniques generally need to measure a physically correct spectral distribution, representing the materials present in each pixel. This leads to a relatively stringent requirement on spatial coregistration between bands, in order to avoid a potentially strong crosstalk from spatial contrasts into the recorded spectrum [7].

On a moving platform, different spectral filters can be placed on successive rows of pixel elements in the focal plane of a conventional camera, across the flight direction, so that the spectral bands can be sampled successively along the scan. Fig. 1 illustrates different concepts for such filter-based spectral imaging from an airborne platform.

In Fig. 1(a), a linear variable filter (LVF) is placed in the focal plane, as indicated in the upper part of the figure, enabling recording of hyperspectral imagery [3], [8]. A spectral image is assembled from multiple raw image frames recorded at different positions during the flight. However, as illustrated in the lower part of Fig. 1(a), 3D structures in the scene may cause parallax effects and coregistration errors in the reconstructed spectra: Here the object B obstructs the view to point P only for the blue band, and the reconstructed spectrum becomes an unphysical mixture of the spectra of ground and roof. If the 3D structure and imaging geometry is known, pixel spectra with such artifacts can be identified and labeled, indicating that the spectral information cannot be recovered correctly. However such a technique will depend heavily on the fidelity of the estimated geometry.

Fig. 1(b), illustrates a scheme for multispectral imaging where the focal plane contains strips of discrete spectral filters, one for each band. Contrary to the linear variable filter in a), it is technically feasible to arrange the discrete filters in such a way that they span a narrower range of viewing angles in the flight direction. Parallax effects are then reduced, but not eliminated.

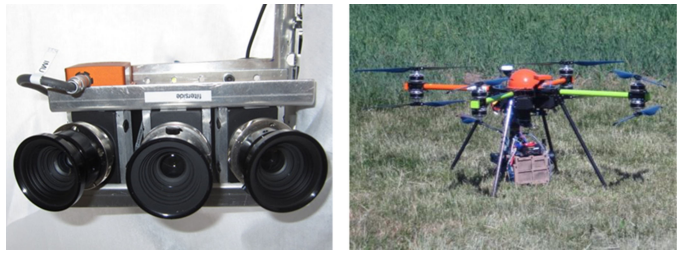

Fig. 2. Left: The three cameras with partially overlapping FOV across track. Right: The sensor payload mounted on a UAV.

A further refinement, central to the imaging concept used in this letter, is shown in Fig. 1(c). Here the set of spectral filters is repeated multiple times. This provides multiple viewing angles for each band. With known 3D geometry, it is then possible to reconstruct spectra correctly for locations in the scene that are visible through only a subset of filters for a given band. Furthermore, it is possible to test spectral data for consistency and integrity by comparing different measurements in the same band in a given location, for example using physical estimates of photon noise to set an acceptance threshold. Interestingly, this consistency test can also be used to aid recovery of 3D information, as mentioned in Section III-A. In addition, multiple consistent samples of the same band can be averaged to improve signal-to-noise ratio (SNR).

An advantage of the concepts presented in Fig. 1(b) and (c) is that with a shorter length of the filter in the flight direction, remaining parts of the image sensor can be used for conventional 2D imaging. The 2D imagery in turn can be used to estimate the imaging geometry, and thereby support the reconstruction of the spectral image. The resulting imaging concept, with repeated spectral sampling combined with 2D imaging on a single image sensor, has potential for efficient collection of information-rich spatial and spectral imagery, provided that it is possible to carry out the required software-intensive reconstruction, which is shown in this letter to be feasible.

A. UAV Sensor Payload

Our UAV sensor payload contains three cameras mounted as shown in Fig. 2, so that the fields of view of the three cameras point nominally along, and to either side, of nadir. The motivation for a multi-camera system is that the size and cost of a single-camera system will not scale well with increasing across-track FOV, when the resolution and per-pixel light collection is maintained (to maintain SNR of the spectral image). We only need data from a strip of pixels across the flight track, but for a single-camera system it would be necessary to use an image sensor which is larger in both length and width (short of an expensive custom design). For a lens with a larger FOV, light level would tend to drop due to the $\cos^4$ dependence on off-axis angle. Furthermore, the angle-dependent spectral shift of filters favors a lens that is approximately image-side telecentric, requiring a lens diameter roughly proportional to image sensor size, and thus to FOV. The cost of lenses and image sensors, and the mass of lenses, grows superlinearly with size. Therefore it is efficient to use a multi-camera system.

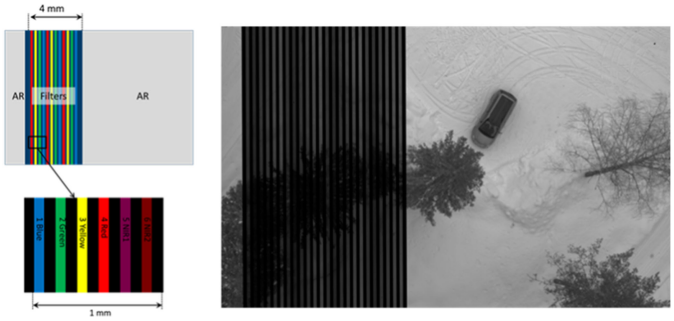

Fig. 3. Left: Layout of the filter array. The gray unfiltered regions permit conventional video imaging, and can for example be used for image-based navigation and 3D scene reconstruction. Right: Example of raw image, where the spectral filter area and video area is clearly visible.

Each camera uses the filter layout shown in Fig. 3, with an array of strip-shaped filters laid out across the FOV. The filter array has six spectral bands in the visible and near infrared, chosen based on tests of material discrimination in a set of hyperspectral images, resulting in bands similar to those used on remote sensing satellites [6]. The set of six bands are repeated four times across the filter array, which extends a total of 4 mm in the scan direction. Each of the 24 individual filters is about 10 pixels wide along the scan direction, an approximate minimum given by camera frame rate and foreseen flying speed. The filter array is placed in close proximity to the image sensor, with the individual filter strips oriented parallel to the short edge of the image sensor. Thus the scan direction is nominally parallel to the long edge of the image sensor. Most of the image sensor area is left unfiltered for recording of panchromatic images. There is also a small unfiltered region between the filters and the FOV edge enabling capture of extra geometrical information. Fig. 3 also shows a single raw image recorded with one of the cameras.

The system employs cameras based on a Sony IMX174 monochrome CMOS image sensor with 1920×1200 pixels. The three cameras are mounted with about 20% overlap, resulting in a total FOV of about 43°. A microcomputer is connected to the cameras. The computer controls the camera exposure adaptively and stores the image data using two solid state drives. The system also contains a GPS receiver and a MEMS inertial measurement unit (IMU). The navigation sensors are read out by a custom FPGA-based navigation synchronization and logging board, which also triggers the cameras.

The camera payload was flown on a Freefly CineStar 8 octocopter, shown in Fig. 2, right. The mass of the key components of the data acquisition and navigation system is 170 g. The mass of a single camera is 170g. Thus, the total mass of the camera system, using off the shelf components, is 680g plus cables, connectors, power supply parts and material for structural assembly leading to a total payload mass well below 1kg.

The output of the sensor system is a stream of raw images from each camera at a frame rate of 80 frames per second (fps), the maximum rate for full camera performance. This allows the FOV to move up to 800 pixels per second without coverage gaps (for 10 pixel wide filters), enabling reasonable ranges of altitude, flight speed and ground resolution. The image data streams also contain metadata such as timestamps, exposure times and gain settings. An additional stream of timestamped sensor data from the IMU and GPS is also available, but will not be exploited in this letter.

III. Spectral Reconstruction

Our aim is to produce an orthonormal spectral map of the scene, that can be used to perform spectral and spatial scene analysis. This task is related to *image mosaicking* [9], [10], where a set of images is coregistered and combined to form a larger image mosaic.

To reconstruct a spectral map image from a set of raw images, we want to coregister each raw image in a common *orthonormal map image* that covers the area of interest. When the scene structure and camera poses are known, we may compute the transformations between each raw image and the map by first projecting the images onto the scene surface, and then perform a chosen geographical map projection onto a planar orthonormal map. This approach requires a detailed georeferenced 3D surface map and highly accurate geographical navigation, but may otherwise be performed efficiently using computer graphics techniques. This opportunity will be left to future studies.

We instead simplify this problem by assuming that the terrain can be approximated locally as a planar structure, and that we can represent the map as the xy-plane in local world coordinates. We will later see that this assumption allows us to represent the transformation between the raw images and the orthonormal map as a simple perspective transformation. The camera poses and sparse terrain structure is estimated from the image sequences. Notably, scene structures that violate the planar assumption are detected later by the spectral consistency test.

Given the transformations between each raw image and the map, we may form an intermediate set of 4×6 *filter image mosaics* for each camera, produced by coregistering sub images recorded through each filter strip in every raw image.

The final step is to combine all filter images into a 6-band multispectral image. Since a pixel in the output map image will typically be covered by several filters corresponding to the same spectral band, we may detect spectral inconsistencies and reduce noise by averaging. The following sections will discuss these reconstruction steps in more detail.

A. Camera Pose and Terrain Estimation

We apply VSLAM to estimate both camera motion and scene structure in 3D, based on correspondences between 2D images [11], [12]. GPS and IMU will in the future be used for increased robustness and for georeferencing the recorded imagery, but lack the precision needed for pixel-accurate coregistrations. For spectral image reconstruction, it is therefore necessary to use image-based navigation, hence we will here explore spectral reconstruction by image processing alone.

VSLAM architectures are typically separated into a front end *tracking* component, which estimates camera motion by tracking the scene structure using a map, and a back end *mapping* component that builds the map based on observations.

Tracking methods can be characterized by their formulation as either *direct* or *indirect*. Direct methods [13]–[15] perform pose estimation by optimizing directly over the *photometric error* [16]. These methods are especially robust in feature-less backgrounds, and are well suited on image sequences with a high degree of overlap.

Indirect, or *feature-based* methods [17], [18] first pre-process the images to extract an intermediate representation such as keypoints, and then optimize over *geometric error* with *bundle adjustment* [19]. This approach is typically less efficient, but

more robust to changing conditions that may affect pixel intensity. There are also hybrid combinations of direct and indirect methods [20], [21] that seek to combine the benefits of both approaches.

The mapping is typically solved through nonlinear optimization, by minimizing for example the global reprojection error, such as in full batch bundle adjustment. A more general formulation allows the incorporation of other types of observations as well, such as IMU and GPS measurements [22]. To enable constant-time mapping, optimization may be performed locally over a subset of images, at the expense of reduced accuracy and drift in pose and scale. Methods based on this mapping scheme are unable to perform large-scale loop closures, and are often referred to as *Visual Odometry (VO)*.

The camera trajectory and scene structure is estimated by employing VO or VSLAM on the unfiltered panchromatic parts of the raw images. Since we need to compute the transformation for every received frame, each camera pose also needs to be estimated. The high frame rate therefore favors very efficient tracking approaches.

In traditional linear scan motions, the camera system is moved in one dominating direction. In this case, local VO adapted to the typical number of frames with overlap will be the most efficient option, and should perform comparably to a global approach. If the sensor system is used to cover overlapping areas, for example by scanning in a lawnmower pattern, VSLAM may be used to exploit loop closures in the overlap between scans. Global updates to the map can change the image-to-map transformations at any time. This has to be taken into account in the spectral reconstruction. We will here only consider offline reconstruction based on VSLAM, and leave online reconstruction to future work.

The resulting 3D model from sparse VSLAM may be used directly in the subsequent steps of the spectral reconstruction, or be refined by estimating a denser 3D map. Even though the frame rate is high, dense 3D modeling of the scene is viable since a high degree of overlap between consecutive frames means that only a small subset of images needs to be processed [23]–[25].

Another possibility is to integrate structure estimation in the spectral reconstruction procedure. One solution is to use the sparse point cloud to guide a plane sweep approach [26] that maps the filter images onto a set of putative depth planes. The depth for each pixel can then be estimated as the depth that maximizes consistency in a neighborhood through bandwise normalized cross correlation. This approach requires a high amount of perspective image warps and correlation computations, all of which are highly parallelizable on GPUs [27]. This is presumably a far less efficient approach, since all frames needs to be part of the structure estimation, but it is slightly less susceptible to errors in dynamic scenes, and it explicitly takes spectral consistency into account.

We will here only consider a simplistic, but very efficient approach, were we assume that the terrain can be described well by a local dominating ground plane. The terrain plane is computed with RANSAC-based plane estimation applied to the part of the sparse 3D model that is visible in each frame. This approximation works well when the area of interest is locally planar, and the height of objects is small compared to the distance to the camera, resulting in little parallax. Since the terrain plane is estimated for each frame, the coarse topography of the terrain will be taken into account. This can be further generalized by estimating several planes for different parts of each frame.

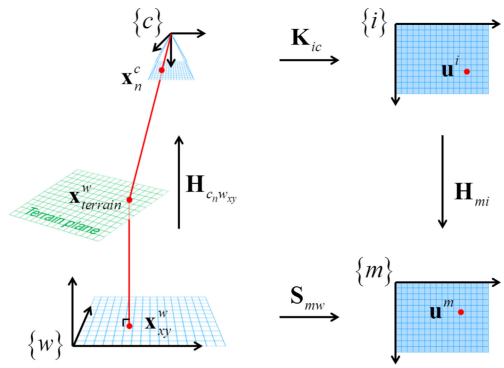

Fig. 4. Illustration of the transformation between image pixels and orthonormal map pixels, via a given terrain plane.

B. Filter Image Mosaics

Given the current camera pose

$$\mathbf{T}_{wc} = \begin{bmatrix} \mathbf{R}_{wc} & \mathbf{t}_{wc}^{w} \\ \mathbf{0}^T & 1 \end{bmatrix} \in SE(3) \tag{1}$$

and terrain plane

$$ax + by + cz + d = 0 \tag{2}$$

in local world coordinates, we can compute a perspective transformation between the raw image and the common map image (see Fig. 4).

If a point $\mathbf{x}_{terrain}^{w}$ on the terrain plane corresponds to the orthographically projected point $\mathbf{x}_{xy}^{w}$ in the xy-plane of the world frame $\{w\}$ as well as the projected point $\mathbf{x}_n^c$ in the normalized image plane of the perspective camera frame $\{c\}$, it can be shown that there is a perspective transformation $\mathbf{H}_{c_n w_{xy}}$, such that $\mathbf{x}_n^c = \mathbf{H}_{c_n w_{xy}} \mathbf{x}_{xy}^{w}$. This homography is given by

$$\mathbf{H}_{c_n w_{xy}} = \left(\begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t} \end{bmatrix} - \frac{1}{c} \mathbf{r}_3 \begin{bmatrix} a & b & d \end{bmatrix} \right), \tag{3}$$

where $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{t}$ are taken from the pose of $\{w\}$ relative to the camera frame $\{c\}$, so that

$$\mathbf{T}_{cw} = \mathbf{T}_{wc}^{-1} = \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3 & \mathbf{t} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

By introducing the intrinsic camera calibration matrix $\mathbf{K}_{ic}$ and a corresponding similarity transformation $\mathbf{S}_{mw}$, which transforms points in the xy-plane of $\{w\}$ to pixels in the map image $\{m\}$, we can compose transformations to produce the homography

$$\mathbf{H}_{mi} = \mathbf{S}_{mw} \mathbf{H}_{c_n w_{xy}}^{-1} \mathbf{K}_{ic}^{-1}, \tag{5}$$

which represents the direct transformation between raw image pixels and map pixels.

With this homography, all 4×6 filter strips in each raw image (cf. Fig. 5, left) may now be warped with a perspective transformation into their corresponding filter mosaic map (cf. Fig. 5, right), resulting in 24 filter maps. To take the dynamic change in exposure into account, all raw image intensity values are simply divided by their corresponding exposure time before being written to the filter mosaics.

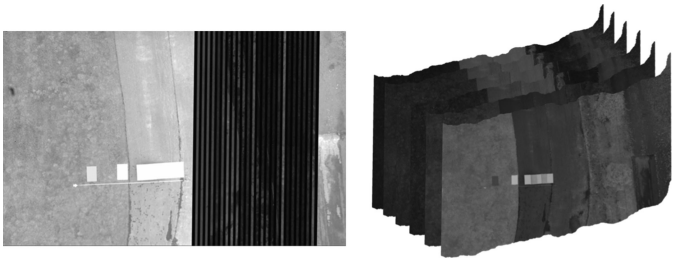

Fig. 5. Left: Raw image. Each filter strip is warped into map coordinates and combined in corresponding filter mosaics. Right: First 6 of 24 filter mosaics, corresponding to each filter strip.

C. Spectral Consistency Testing

Filter mosaics corresponding to the same spectral band can be tested for consistency in order to detect and to compensate for errors, like inconsistent spectral mixing at edges, or uncorrected parallax.

Assume that we have constructed a set of 24 filter mosaics $M = \{M_b^s\}$, where $b = 1, \ldots, 6$ indicate the spectral bands and $s = 1, \ldots, 4$ indicate the filter sets. We can assemble these into a set of 6-band spectral images corresponding to each filter set $S = \{S^1, \ldots, S^4\}$, so that

$$S^i(u,v) = \mathbf{s}_{u,v}^i = \begin{bmatrix} M_1^i(u,v) \\ \vdots \\ M_6^i(u,v) \end{bmatrix} \in \mathbb{R}^6, \tag{6}$$

where $\mathbf{s}_{u,v}^i = S^i(u,v)$ is the spectral vector at a given pixel location (u,v). Given the set S, we want a measure of how spectrally consistent each S^i is for each pixel (u,v).

Let the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_{u,v}, \boldsymbol{\Sigma}_{u,v})$ represent the distribution of the true spectrum at the given pixel location. Assuming that all spectra at the same pixel location measure the same surface material, we estimate $\boldsymbol{\mu}_{u,v}$ as the the mean spectrum

$$\hat{\boldsymbol{\mu}}_{u,v} = \frac{1}{4} \sum_{i=1}^{4} \mathbf{s}_{u,v}^i. \tag{7}$$

Instead of using the sample covariance, which would be unreliable given the few samples, we exploit information about the expected noise given the signal. Assuming that the noise in the pixel measurements are dominated by photon noise, the corresponding Poisson distribution for $\hat{\boldsymbol{\mu}}_{u,v}$ gives us the covariance estimate

$$\hat{\boldsymbol{\Sigma}}_{u,v} = \operatorname{diag}(\hat{\boldsymbol{\mu}}_{u,v}). \tag{8}$$

We can compare each spectrum $\mathbf{s}_{u,v}^i$ with this model using the squared Mahalanobis distance

$$d_{u,v}^i = \left(\mathbf{s}_{u,v}^i - \hat{\boldsymbol{\mu}}_{u,v}\right)^T \hat{\boldsymbol{\Sigma}}_{u,v}^{-1} \left(\mathbf{s}_{u,v}^i - \hat{\boldsymbol{\mu}}_{u,v}\right). \tag{9}$$

The proposed spectral inconsistency (SIC) metric is then the largest squared Mahalanobis distance for each pixel, resulting in the image

$$D_{\text{SIC}}(u,v) = \max_i d_{u,v}^i. \tag{10}$$

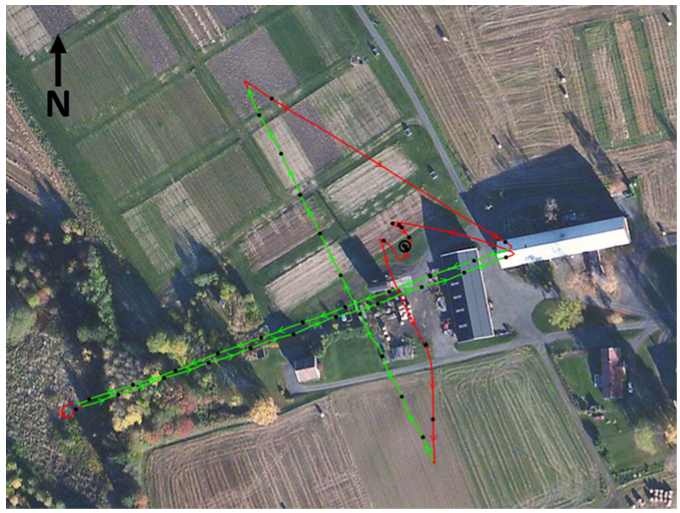

Fig. 6. An overview over the experiment site. The GPS track from the UAV is also shown, where the parts corresponding to the planned flight-lines are marked in green.

We can then detect any pixel that has a SIC value above a certain threshold t to produce the spectral inconsistency veto mask

$$I_{\text{veto}}(u,v) = \begin{cases} 1, & D_{\text{SIC}}(u,v) > t \\ 0, & \text{otherwise} \end{cases}. \tag{11}$$

Even though the estimates (7) and (8) give a rough model for the true spectrum, a χ^2 test on the squared Mahalanobis distances $D_{\text{SIC}}(u,v)$ has empirically shown to result in a reasonable threshold for spectral consistency testing.

D. Spectral Image Composition

Given the set of spectral images $S^i(u,v)$, the resulting 6-band multispectral image may be composed by averaging over each filter set i. The final product is then the spectral image map $S(u,v)$, the corresponding spectral inconsistency veto mask $I_{\text{veto}}(u,v)$, and optionally the SIC image $D_{\text{SIC}}(u,v)$ and the standard deviations over the sets $S_{\text{std}}(u,v)$.

However, taking the maximum distance over all spectral images in (10) is a very conservative approach, which will disregard a pixel even though most filter sets are consistent. A more robust approach is to compute the inconsistency metric over all subsets of 3 filter images, and choose the subset with the lowest SIC score. We can then for each pixel disregard the spectral image that was not part of this subset in the resulting products. This *leave-one-out* approach is employed when a pixel spectrum is detected as inconsistent over all the filter sets.

IV. Experiments

A. Test Data

The camera payload was flown over an agricultural research station. This site consists of flat open fields, vegetation, and tall buildings. A set of known targets were placed in the scene, including a site with colored panels, and a sheet of green canvas in the field north of the buildings. The UAV flew two crossing flight-lines as shown in Fig. 6, at an altitude of about 120 m using

automatic waypoint navigation. We will here mostly consider the flight line going from the southern field, over the buildings and on to the northern field, which we will call the *field-north* flight line.

The three cameras were run at 80 frames per second. As a result, over 200 k raw images were recorded. Of these, only about 180 k images were used, omitting images in the take-off and landing phase.

A field spectrometer was used to record spectral ground truth measurements on the targets in the northern field during the drone overpass.

With sufficient inter-camera overlap and depth observability at shorter distances, Kalibr [28], [29] was used to perform intrinsic and joint extrinsic camera calibration for all three cameras using short-range lab measurements.

B. Spectral Reconstruction With Visual SLAM

Since this dataset has several crossing flight lines, we applied VSLAM for pose and structure estimation. This enables better map consistency through the detection of loop closures, but is expected to be less efficient than a VO pipeline. Although direct and semi-direct tracking methods are assumed to be more efficient on these data, we chose to use the well established indirect ORB-SLAM [18], [30] VSLAM system for its robustness and ease of use. The ORB-SLAM implementation was slightly adapted to be able to read our data and be more likely to initialize in highly planar scenes.

All frames from the centre camera were cropped to cover only the panchromatic part of the raw images, and fed in full resolution to ORB-SLAM. The VSLAM pipeline processed the frames at about 12Hz, which is 15% of of the full frame rate. This resulted in valid poses for all frames, and a sparse 3D point cloud representation of the scene with over 75 k points.

Without additional information, the absolute scale in the VSLAM solution is unknown. Since the distance to the scene is large compared to a baseline of about 6cm between the cameras, a disparity analysis indicated that inter-camera parallax was negligible at the considered flying altitude. We therefore ignored the baseline between the cameras, essentially performing arbitrary scale VSLAM, modelling the camera system as monocular with an extended field of view. This enabled us to use only the rotational part of the extrinsic calibration to transform the trajectory of the centre camera into trajectories for the flanking cameras as well.

Given the estimated camera trajectories and the sparse point cloud, spectral reconstruction was performed on all frames according to the method presented in Section III. The results are shown in Fig. 7. The chosen size of the reconstructed image corresponds to a ground sampling distance of approximately 10 cm.

C. Integrity and Accuracy of the Spectral Signal

From the results presented in Fig. 7 we see that much of the scene is reconstructed with sharp image quality, indicating good coregistration between the filter image mosaics. Those parts of the scene that violate the planar terrain assumption, however, show significant blurring and spectral inconsistencies. The rightmost part of Fig. 7 demonstrates that the leave-one-out approach from Section III-D reduces these effects significantly. Using this approach, we are able to remove what appears to be glare caused by the white-colored panel, as presented in the close up in Fig. 9.

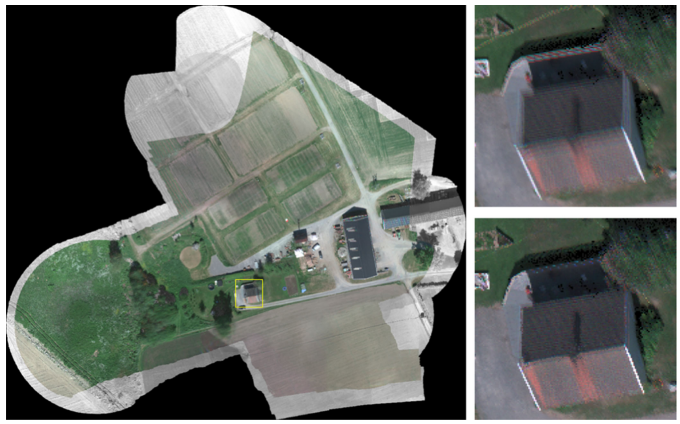

Fig. 7. Left: VSLAM was performed on the entire set of frames, resulting in the spectral image shown as a RGB image. The spectral image is placed on top of a mosaic based on the panchromatic images. Right: Illustration of the spectral image composition strategies in the area shown in yellow. The top image shows the result when using all filter sets, while the bottom image is the result of the leave-one-out approach.

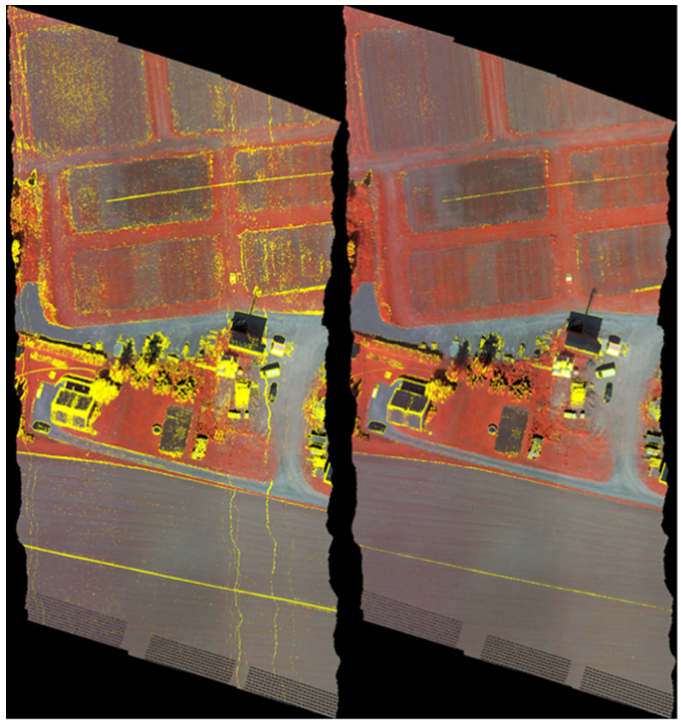

Fig. 8. Spectral inconsistency veto mask in front of a NIR-G-B representation of the reconstructed spectral images over the field-north flight line. Left: Inconsistent pixels when all filter sets are used. Right: Inconsistent pixels with the leave-one-out approach.

Fig. 8 shows the result of spectral inconsistency veto masking, where the threshold has been set so that it corresponds to the 99% quantile of the model in Section III-C. Reconstruction based on all filter sets results in many inconsistent pixel spectra, including inconsistent mixing at edges and inconsistent stripes along the track, caused by defects in the filter arrays. Note also that pixels with valid spectra are mostly available from all the different materials in this scene. The consistency test thus tends to yield

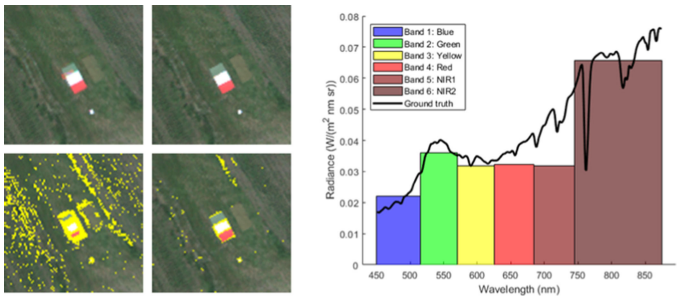

Fig. 9. Left: A close up of the target site. The left column shows the result of spectral reconstruction based on all filter sets, while the right column is the result of the leave-one-out approach. Right: A radiometric comparison between a ground truth measurement of the large green canvas to the right in the target site, and the corresponding reconstructed multispectral spectrum.

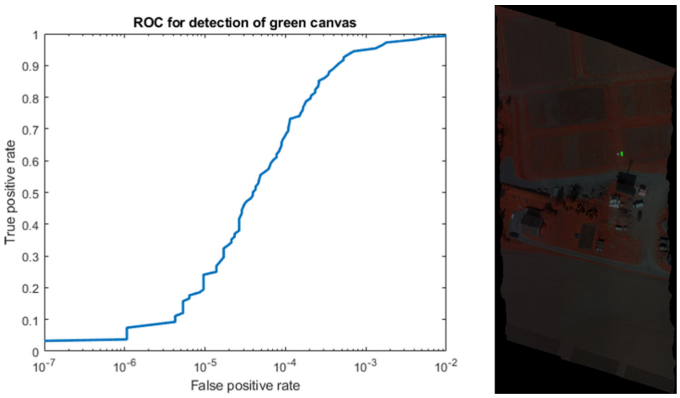

Fig. 10. The result from spectral matched filtering based on the spectral signature of the green canvas. Left: The receiver operating characteristic (ROC) curve with logarithmic false alarm scale. Right: Result when 90% of the target is detected. Detected pixels are shown in green.

good measurements, rather than a large number of spectra. The leave-one-out approach significantly reduces the amount of inconsistent pixels, especially along the edges between different materials. It also removes the artifacts due to the sensor defects.

Fig. 9 right shows an example of a reconstructed spectrum of the large green canvas target to the right at the target site. This spectrum is compared to a ground truth measurement, which was produced by a field spectrometer during the recording. The reconstructed spectrum has been radiometrically corrected and scaled in accordance with an assumed 20% loss in the optics. The comparison indicates that the spectral reconstruction process leads to a plausible spectrum for this target. The responsivity of band 5 (685–745 nm) is less than expected from nominal filter properties, but can be corrected by a full characterization of spectral responsivity.

D. Spectral Analysis Example

We finally demonstrate the utility of the resulting multispectral image with a simple example.

The spectral signature of the green canvas was extracted from a target between the buildings on another flight line. This signature was then used to rediscover the green canvas target in the target site on the field-north flight line, using a simple spectral matched filter approach [1]. The result is shown in Fig. 10. Here, the true positive rate is given as the proportion of target pixels detected, while the false positive rate is given as the proportion of background pixels detected.

The results show that 90% of the target is detected with about 400 pixels of scattered false alarms. By applying a 3×3 morphological closing to ensure spatial consistency, we get 100% detection with no false alarms.

V. Discussion and Conclusions

With this work we have shown that a novel concept for multispectral imaging is feasible for use in the challenging case of low-cost wide area mapping using small UAVs. The sample results indicate that spectral reconstruction for this camera system can be performed by image processing alone, without relying on accurate navigation and preexisting terrain models.

The results show that the spectral data can be reconstructed, even in the presence of residual errors in the estimated scene geometry. In particular, the proposed approach detects and corrects for invalid data using physics-informed spectral consistency tests.

It must be mentioned that the current filter arrangement with 6 bands is an obvious sacrifice of capability compared to hyperspectral imaging. On the other hand, compared to 3-band RGB imaging, a 6-band camera can provide significant capability for automated spectral detection and discrimination.

A major question is whether it is possible to perform spectral reconstruction and target detection on board the resource-limited platforms that are otherwise well suited to carry the system. The current processing system is slower than real-time, but may still be used for on-board processing between recordings. Work towards real-time processing based on more efficient tracking and incorporation of IMU data is ongoing. Future work also includes the integration of 3D scene models into the spectral reconstruction processing, as well as integrating GPS data for georeferencing.

In conclusion, the results presented demonstrate the feasibility of a low-cost and lightweight payload for UAVs providing conventional and multispectral imaging with a high and scalable area coverage rate. It is still a challenge to process the data in real time on a power-limited small UAV, but this may come within reach given the development in algorithms and power-efficient computing hardware. Work is ongoing to further improve the quality of the geometrical reconstruction, taking advantage of the rapid progress in this field of research, but the results here demonstrate a level of performance which is clearly suitable for a number of applications.

References

[1] M. T. Eismann, *Hyperspectral Remote Sensing*. WA, USA: SPIE Press, 2012.

[2] P. Koirala, T. Løke, I. Baarstad, A. Fridman, and J. Hernandez, "Real-time hyperspectral image processing for UAV applications, using HySpex Mjolnir-1024," *Proc. SPIE 10198*, 2017, Art. no. 1019807, doi: 10.1117/12.2267476.

[3] P. Gonzalez *et al.*, "A novel CMOS-compatible, monolithically integrated line-scan hyperspectral imager covering the VIS-NIR range," *Proc. SPIE 9855*, 2016, Art. no. 98550N, doi: 10.1117/12.2230726.

[4] J. Mäkynen, H. Saari, C. Holmlund, R. Mannila, and T. Antila, "Multi- and hyperspectral UAV imaging system for forest and agriculture applications," *Proc. SPIE 8374*, 2012, Art. no. 837409, doi: 10.1117/12.918571.

[5] H. Torkildsen, T. Haavardsholm, T. Opsahl, U. Datta, A. Skaugen, and T. Skauli, "Compact multispectral multi-camera imaging system for small UAVs," *Proc. SPIE – Int. Soc. Opt. Eng.*, vol. 9840, 2016, pp. 491–498, doi: 10.1117/12.2224495.

[6] T. Skauli *et al.*, "Compact camera for multispectral and conventional imaging based on patterned filters," *Appl. Opt.*, vol. 53, no. 13, pp. C64–C71, 2014, doi: 10.1364/AO.53.000C64.

[7] T. Skauli, "An upper-bound metric for characterizing spectral and spatial coregistration errors in spectral imaging," *Opt. Express*, vol. 20, no. 2, pp. 918–933, 2012, doi: 10.1364/oe.20.000918.

[8] A. M. Mika, "Linear-wedge spectrometer," *Proc. SPIE 1298, Imag. Spectrosc. Terr. Environ.*, 1990, pp. 127–131.

[9] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski, "Photographing long scenes with multi-viewpoint panoramas," *Proc. ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 25, no. 3, 2006, pp. 853–861, doi: 10.1145/1179352.1141966.

[10] R. Szeliski and U. Szeliski, Richard (Microsoft Research, "Image Alignment and Stitching: A Tutorial," *Found. Trends Comput. Graph. Vision*, vol. 2, no. 1, pp. 1–104, 2006, doi: 10.1561/0600000009.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2004.

[12] Y. Ma, S. Soatto, J. Košecká, and S. S. Sastry, *An Invitation to 3-D Vision*, (Interdisciplinary Applied Mathematics), vol. 26. New York, NY: Springer New York, 2004.

[13] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 2320–2327, doi: 10.1109/ICCV.2011.6126513.

[14] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale direct monocular SLAM," *Proc. Eur. Conf. Comput. Vision*, (Lecture Notes in Computer Science). Cham: Springer International Publishing, 2014, vol. 8690, pp. 834–849, doi: 10.1007/978-3-319-10605-2.

[15] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018, doi: 10.1109/TPAMI.2017.2658577.

[16] S. Baker and I. Matthews, "Lucas-Kanade 20 Years On: A unifying framework," *Int. J. Comput. Vision*, vol. 56, no. 3, pp. 221–255, Feb. 2004, doi: 10.1023/B:VISI.0000011205.11775.fd.

[17] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality, ISMAR*, 2007, pp. 225–234, doi: 10.1109/ISMAR.2007.4538852.

[18] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[19] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment A Modern Synthesis," *Proc. ICCV '99: . Int. Workshop Vision Algorithms*, 2000, pp. 298–372, doi: 10.1007/3-540-44480-7_21.

[20] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 15–22, doi: 10.1109/ICRA.2014.6906584.

[21] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, Apr. 2017, doi: 10.1109/TRO.2016.2623335.

[22] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Found. Trends Robot.*, vol. 6, no. 1-2, pp. 1–139, 2017, doi: 10.1561/2300000043.

[23] R. Mur-Artal and J. Tardos, "Probabilistic semi-dense mapping from highly accurate feature-based monocular SLAM," *Robot.: Sci. Syst. XI*, Robotics: Science and Systems Foundation, 7 2015, doi: 10.15607/RSS.2015.XI.041.

[24] M. Pizzoli, C. Forster, and D. Scaramuzza, "REMODE: Probabilistic, monocular dense reconstruction in real time," *Proc. IEEE Int. Conf. Robot. Autom.*, 5 2014, pp. 2609–2616, doi: 10.1109/ICRA.2014.6907233.

[25] M. P. T. Hinzmann J. L. Schönberger, and R. Siegwart, "Mapping on the Fly: Real-time 3D dense reconstruction, digital surface map and incremental orthomosaic generation for unmanned aerial vehicles," *Proc. Field Service Robot. - Results 11th Int. Conf.*, 2017, pp. 383–396.

[26] R. Collins, "A space-sweep approach to true multi-image matching," *Proc. CVPR IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 1996, pp. 358–363, doi: 10.1109/CVPR.1996.517097.

[27] D. Gallup, J. M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1–8, doi: 10.1109/CVPR.2007.383245.

[28] "Kalibr repository," [Online]. Available: https://github.com/ethz-asl/kalibr, Accessed: Jan. 11, 2020.

[29] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," *Proc. - IEEE Int. Conf. Robot. Autom.*, 2016, pp. 4304–4311, doi: 10.1109/ICRA.2016.7487628.

[30] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.