PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Compact multimodal multispectral sensor system for tactical reconnaissance

Trym Haavardsholm, Thomas Opsahl, Torbjørn Skauli, Annette Stahl

Trym V. Haavardsholm, Thomas O. Opsahl, Torbjørn Skauli, Annette Stahl, "Compact multimodal multispectral sensor system for tactical reconnaissance," Proc. SPIE 12235, Imaging Spectrometry XXV: Applications, Sensors, and Processing, 122350B (30 September 2022); doi: 10.1117/12.2632701

SPIE.

Event: SPIE Optical Engineering + Applications, 2022, San Diego, California, United States

Compact multimodal multispectral sensor system for tactical reconnaissance

Trym V. Haavardsholm[a,b], Thomas O. Opsahl[a], Torbjørn Skauli[a], and Annette Stahl[b]

[a]Norwegian Defence Research Establishment (FFI), NO-2007 Kjeller, Norway
[b]Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway

ABSTRACT

Multispectral imaging is an attractive sensing modality for small unmanned aerial vehicles (UAVs) in numerous military and civilian applications such as reconnaissance, target detection, and precision agriculture. Cameras based on patterned filters in the focal plane, such as conventional colour cameras, represent the most compact architecture for spectral imaging, but image reconstruction becomes challenging at higher band counts. We consider a camera configuration where six bandpass filters are arranged in a periodically repeating pattern in the focal plane. In addition, a large unfiltered region permits conventional monochromatic video imaging that can be used for situational awareness (SA), including estimating the camera motion and the 3D structure of the ground surface. By platform movement, the filters are scanned over the scene, capturing an irregular pattern of spectral samples of the ground surface. Through estimation of the camera trajectory and 3D scene structure, it is still possible to assemble a spectral image by fusing all measurements in software. The repeated sampling of bands enables spectral consistency testing, which can improve spectral integrity significantly. The result is a truly multimodal camera sensor system able to produce a range of image products. Here, we investigate its application in tactical reconnaissance by pushing towards on-board real-time spectral reconstruction based on visual odometry (VO) and full 3D reconstruction of the scene. The results are compared with offline processing based on estimates from visual simultaneous localisation and mapping (VSLAM) and indicate that the multimodal sensing concept has a clear potential for use in tactical reconnaissance scenarios.

Keywords: Spectral imaging, image sensors, image reconstruction, optical filters, robot vision, reconnaissance

1. INTRODUCTION

Small unmanned aerial vehicles (UAVs) have had a large impact on the accuracy and timeliness of tactical surveillance, target acquisition and reconnaissance (STAR) brought on by developments in robotics, sensor technology, machine learning and military tactics. A similar development is taking place in civilian applications such as situational awareness (SA) for search and rescue (SAR) and disaster response. The primary sensor in these applications is the visual camera, which produces imagery commonly used for target detection, localisation and guidance, as well as terrain characterisation and mapping. By exploiting platform movement, visual imagery may also sense the motion of the sensor itself as well as the 3D structure of the ground terrain.

It is generally beneficial to incorporate multiple sensor modalities in order to exploit the widest possible range or target signatures. For a small UAV, weight, size and power limitations tend to dictate use of sensor payloads whose imaging capability consists of only a daylight-based camera, at most in combination with a low-resolution thermal camera. Another sensing modality of increasing interest is spectral imaging, which is particularly useful for discriminating objects and surfaces with low visual contrast, and has a wide range of applications that include tactical target detection, land use mapping and precision agriculture. Although hyper- and multispectral imaging sensors for small UAVs exist, tactical systems often leave these out in favour of high-performance conventional cameras.

We have previously presented a novel imaging concept following this multimodal approach based on six bandpass filters arranged in a periodically repeating pattern.[1,2] Cameras based on patterned filters in the focal

Further author information: (Send correspondence to Trym V. Haavardsholm)
Trym V. Haavardsholm: E-mail: trym.haavardsholm@ffi.no

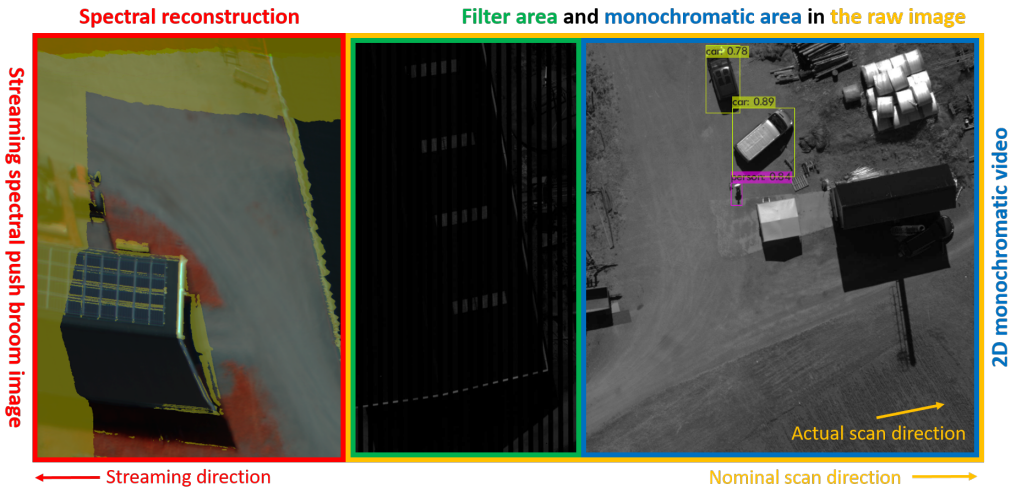

Figure 1. Raw image frame (right part) and a multispectral "push broom" still image (left part) reconstructed from this and preceding frames. The spectral filter strips are seen as a darker part in the image frame on the right. The unfiltered part produces conventional video, which can be used for target detection as well as estimation of camera pose and 3D scene structure. The multispectral image on the left is built up in "push broom" fashion by the platform movement. In this example, a NIR band is mapped to the red colour channel to highlight natural vegetation. Spectra detected as inconsistent or incomplete in the repeated sampling are marked in transparent yellow. The first (left-most) lines in the push broom image are incomplete because the scanning has just started, and that area has not been covered by all sets of filters. The top is incomplete since the camera is moving slightly upwards. There is an inherent offset in space and time between the current video frame and the current line in the push broom image given by the width of the filter area and the time it takes to scan all filters over the scene.

plane, such as conventional colour cameras, represent the most compact architecture for spectral imaging. In our case, most of the focal plane is left unfiltered for high-resolution monochromatic video capture. In effect, a part of the field of view (FOV) in the conventional video camera is converted to recording of imagery with moderate spectral resolution. Platform movement enables us to scan the filters over the scene to capture irregular spectral measurements of the ground surface. This enables sensing modalities associated with hyperspectral cameras, but trades some performance for an extreme degree of compactness, essentially as a retrofit into an existing camera. However, the filter-based approach presents significant challenges in spectral image reconstruction since the different filter regions see different parts of the scene at a given instant, so that multiple recorded images must be coregistered to obtain consistent spectra.

We have previously demonstrated consistent spectral reconstruction with accurate filter alignment based on offline visual simultaneous localisation and mapping (VSLAM) for pose and structure estimation and a locally planar world assumption, where the repeated sampling lets us test consistency in each spectrum.[3] Here, we present results from a more advanced reconstruction chain to investigate tactical applicability of the imaging concept with the following contributions:

1) Improved spectral reconstruction accuracy by taking the local 3D terrain structure into account.
2) A locally consistent online procedure adapted to the tactical scenario by reconstructing in sensor view.
3) An efficient GPU implementation based on OpenGL.

Fig. 1 shows an example result from the tactical pipeline. To the right, we see the current image used for pose and structure estimation, as well as a demonstration of target detection using YOLOv4.[4] For illustration, we also see the six filter strips repeated four times. To the left, we see the result of spectral reconstruction over several thousand frames represented as a push broom image in sensor perspective.

In the rest of this paper, we give a brief overview of the multimodal sensing system in Sec. 2, describe the new tactical reconstruction method in Sec. 3, present experimental results in Sec. 4, and end with the conclusion in Sec. 5.

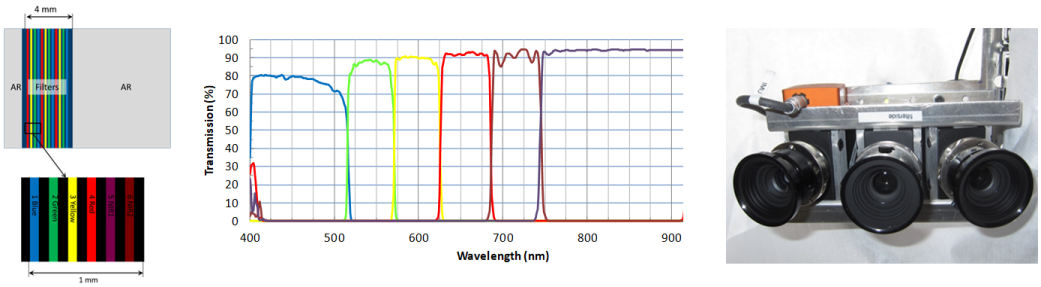

Figure 2. Left: Layout of the filter array. Centre: Transmission spectrum for the different bandpass filters. Right: The UAV payload prototype. We are here only using the centre camera.

2. MULTIMODAL SENSING SYSTEM

The imaging concept considered here combines conventional 2D video imaging with repeated spectral sampling on a single image sensor. The sensor is a regular camera with a specialised filter layout in the focal plane. Different spectral bands are recorded in succession thanks to the platform movement, similar to *push broom* imaging spectrometers commonly used for hyperspectral imaging (HSI). The resulting system 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. This concept and the associated prototype system have been thoroughly discussed in the previous papers,[1,2] and a complete spectral reconstruction processing chain was presented in Ref. 3. We will here give a quick review of the key points relevant to the tactical scenario, and refer to the previous papers for details.

As shown in Fig. 2, left, six bandpass filters are arranged in a periodically repeating pattern. The repeated spectral sampling provides multiple viewing angles for each band and enables spectral inconsistency (SIC) testing, robustness to shadowing, and averaging to improve the signal-to-noise ratio (SNR). When scanning the camera over the scene, we reconstruct a spectral image by coregistering filter image strips to form filter image mosaics. For precise image alignment, taking image rotation, translation and parallax into account, we generally require both known camera motion and scene structure, as well as an accurate camera calibration for modelling the camera projection.

Small and lightweight UAVs naturally exhibit a jerky orientational motion which is difficult to compensate for with a gimbal. Therefore, we want accurate camera pose estimates at every frame to align the filter strips as precisely as possible. Fortunately, the conventional 2D video imagery is well suited to support motion and structure estimation using image-based navigation methods. In Ref. 3, VSLAM with ORB-SLAM[5,6] was used for pose and structure estimation, and the scene structure was represented locally as a planar surface fitted to the 3D point cloud from the VSLAM map. The spectral image was formed in a global world plane, resulting in a planar alignment procedure represented by a homography transformation computed from the current camera pose, the current terrain plane, the world plane and the camera calibration.

The multimodal concept is implemented in the multi-camera UAV payload prototype shown to the right in Fig. 2. Each camera is based on a Sony IMX174 monochrome CMOS image sensor with 1920×1200 pixels. We will only consider the centre camera when developing our tactical approach, and keep the adaptation to multi-camera capture for future work. The system also contains a global navigation satellite system (GNSS) receiver and a MEMS inertial measurement unit (IMU). 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, and are accompanied with GNSS data at 10 Hz and IMU data at 100 Hz

3. SPECTRAL RECONSTRUCTION IN TACTICAL APPLICATIONS

The system presented in the previous section has several shortcomings with respect to tactical applications:

1) The pose and structure estimation based on VSLAM is significantly slower than the frame rate and performs global updates to the estimated variables when correcting for loop closures.
2) The spectral reconstruction method is slower than the frame rate, even with the simplistic planar world approximation, and expects a globally consistent map and navigation. Overlapping areas are overwritten by the newest measurements.
3) The resulting spectral image is represented in global map coordinates with a chosen metric resolution, which is wasteful and cumbersome for spectral processing.

We seek to overcome these challenges and even increase the accuracy of spectral reconstruction by presenting a locally consistent real-time method capable of taking the 3D structure of the scene into account.

3.1 Real-time pose and structure estimation

The relatively high frame rate makes it challenging to estimate camera poses in real-time. We alleviate this problem by transitioning to more efficient pose estimation methods and incorporating IMU measurements.

The arguably simplest, most efficient and most direct way to incorporate IMU measurements for real-time pose estimation is to combine them with GNSS measurements in an inertial navigation system (INS).[7] The resulting globally consistent navigation in absolute coordinates also allows the direct use of georeferenced digital elevation models (DEMs) for structure estimation. Although this is a very robust approach, it is expected to be less precise and consequently lead to less consistent image alignment than image-based approaches, which directly observe camera motion with respect to the ground. Absolute inaccuracies will also cause the navigation to be somewhat misaligned with the structure, potentially further aggravating consistent image alignment. In addition, the DEM is likely to be outdated. Finally, the INS approach will severely depend on the availability of both GNSS signals and georeferenced DEMs in the area of operation.

A very efficient image-based alternative is visual odometry (VO).[8] Like VSLAM, it simultaneously estimates pose and structure in a locally consistent map, but only within a limited horizon without support for loop closure detection and global map correction. The approach is consequently highly efficient and locally precise but also exposed to track loss and global drift in scale, position and orientation. The estimated local structure is inherently well aligned with the navigation but typically sparse and with poor coverage in the fringes of the FOV. VO can be combined with priors from IMU measurements to exhibit more robust, efficient and accurate performance in certain situations.

Visual-inertial odometry (VIO)[9] is VO tightly coupled with IMU measurements. Since the IMU can observe both absolute scale and absolute horizontal orientation, VIO can give estimates with a globally consistent scale and orientation with respect to the local tangent plane. This global consistency and added robustness in otherwise challenging situations come at the expense of higher computational complexity. Furthermore, VIO is still susceptible to drift in position and heading.

Although the structure recovered by VO and VIO is typically in the form of sparse points, a dense 3D surface model may be computed from the point cloud, e.g. by first performing a 2D Delaunay triangulation over tracked feature points in keyframes, and then back-project the triangulation to generate a 3D mesh.[10,11] In applications where the camera is always approximately nadir, the triangulation may be performed directly in the horizontal plane.

Compared to VSLAM, the efficient solutions presented here come at the price of decreased local precision or global accuracy. Since VO is similar to VIO but more efficient and more susceptible to global drift, we choose to concentrate on INS and IMU-aided VO when developing our tactical approach, which will focus on local consistency while depending less on global accuracy.

3.2 Push broom spectral image representation

A straightforward approach to avoid the problems with the global map-based image representation is to instead form the spectral image in the sensor perspective as something equivalent to a traditional push broom image, which is captured line-by-line by accumulating data from a single-lined camera. Although push broom images are distorted by the camera motion, smooth motion ensures that local consistency is preserved, and the resulting stream of spectral data at the original sensor resolution may be processed directly for tactical applications. This

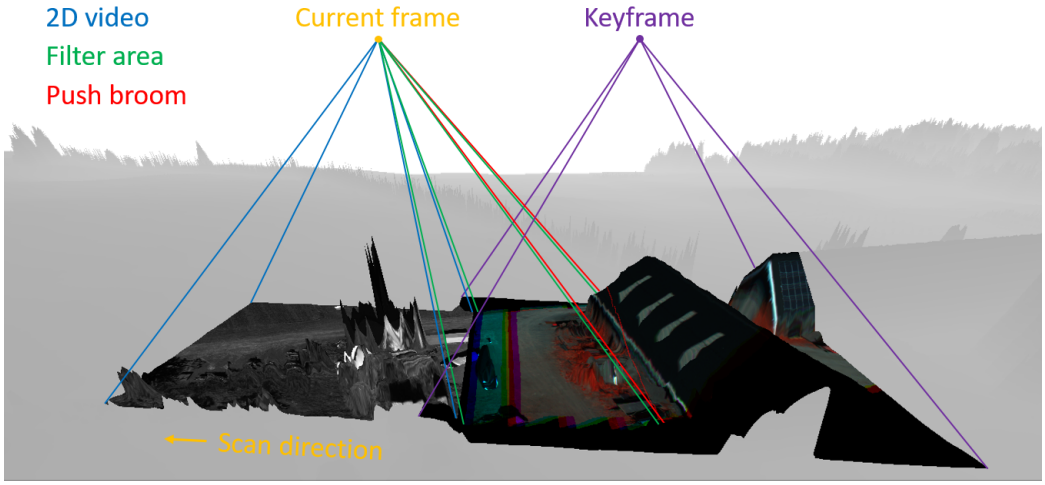

Figure 3. The emulated push broom imaging geometry. The filter area (green) in the current frame (orange) is projected onto the ground surface and back into the current keyframe (purple), where the filter measurements are accumulated. The resulting spectral bands are then projected back into the chosen push broom line in the current frame (red). The current 2D video area (blue) is also shown projected onto the surface for clarity, and the geometry has been slightly exaggerated.

is also the usual way spectral images are captured, and therefore has the additional benefit of fitting optimally with existing processing pipelines in the spectral imaging community.

We can emulate push broom imaging in the spectral reconstruction by accumulating all filter measurements in the current sensor view and extract a chosen line. But reprojecting the filter mosaics to every new frame is inefficient and will quickly accumulate resampling errors. Taking advantage of the smooth scanning motion, we instead accumulate the measurements in periodic keyframes, which are made slightly larger than the original image to account for motion across the scanning direction. For every new frame, we then 1) project the current frame into the keyframe and add the filter measurements, 2) project the mosaics back from the keyframe to a virtual single-lined push broom camera corresponding to the current frame and 3) append the push broom line to the push broom image. Keyframes are periodically updated to the newest sensor view by projecting the mosaics to the current frame. The push broom line is chosen as the first line after the last (left-most) filter strip (see Fig. 3). The procedure is summed up in Fig. 4.

The key technique in this reconstruction procedure is the relative reprojection of images from *projector frames* $\mathcal{F}_p$ onto the scene and back into *observer frames* $\mathcal{F}_o$. In step 1) above, the current frame will act as the projector, while the keyframe is the observer and in step 2) the roles are switched. In general, we can project pixel coordinates $\mathbf{u}_i^p$ given in $\mathcal{F}_p$ to pixel coordinates $\mathbf{u}_i^o$ given in $\mathcal{F}_o$ with

$$\mathbf{u}_i^o = \pi_o\left(\mathbf{T}_{op} \cdot \pi_p^{-1}(\mathbf{u}_i^p, z_i^p)\right) \qquad \mathbf{T}_{op} = \begin{bmatrix} \mathbf{R}_{op} & \mathbf{t}_{op}^o \\ \mathbf{0}^T & 1 \end{bmatrix} \in SE(3), \tag{1}$$

where $\pi_o : \mathbb{R}^3 \to \Omega_o$ is the geometric camera model for the observer camera, projecting 3D points $\mathbf{x}^o \in \mathbb{R}^3$ in $\mathcal{F}_o$ onto pixels $\mathbf{u}_i^o \in \Omega_o$ in the observer image, and $\pi_p^{-1} : \Omega_p \times \mathbb{R}^+ \to \mathbb{R}^3$ is the inverse geometric camera model for the projector camera, backprojecting pixels $\mathbf{u}_i^p$ with the corresponding depths z_i^p back to 3D points in the projector frame. $\mathbf{T}_{op}$ is the relative pose of the projector frame $\mathcal{F}_p$ given in the observer frame $\mathcal{F}_o$ and the $\cdot$ operator represents the action of the pose on 3D points so that $\mathbf{x}^o = \mathbf{T}_{op} \cdot \mathbf{x}^p$.

Using the perspective camera model with calibration matrices $\mathbf{K}_o$ and $\mathbf{K}_p$, the corresponding homogeneous reprojection is given by

$$\tilde{\mathbf{u}}_i^o = \mathbf{K}_o \left[\mathbf{R}_{op} + \frac{\mathbf{t}_{op}^o \mathbf{e}_z^\top}{z_i^p}\right] \mathbf{K}_p^{-1} \tilde{\mathbf{u}}_i^p = \mathbf{H}_{op}^{z_i^p} \tilde{\mathbf{u}}_i^p, \tag{2}$$

where $\mathbf{e}_z$ is the unit vector in the z-direction and the resulting homography $\mathbf{H}_{op}^{z_i^p}$ has to be computed for each unique depth. If the scene is planar with plane equation $\Pi_p : a^p x + b^p y + c^p z + d^p = 0$ given in the projector frame, we can find a common reprojection homography for all pixels on the plane:

$$\tilde{\mathbf{u}}_i^o = \mathbf{K}_o \left[\mathbf{R}_{op} - \frac{\mathbf{t}_{op}^o \mathbf{n}^{p\top}}{d^p} \right] \mathbf{K}_p^{-1} \tilde{\mathbf{u}}_i^p = \mathbf{H}_{op}^{\Pi_p} \tilde{\mathbf{u}}_i^p, \tag{3}$$

where $\mathbf{n}^p = [a^p, b^p, c^p]^\top$ is the plane unit normal and $-d^p$ is the signed distance from the plane to the origin of $\mathcal{F}_p$.

Since the poses involved here are relative, we are less dependent on global accuracy, but still able to exploit local precision for optimal alignment accuracy. In fact, as long as there is insignificant drift within the time it takes to scan all filter sets over a point in the scene, there should not be any noticeable decrease in accuracy in the presence of significant global drift, and the spectral images should look qualitatively the same. Even track loss and re-initialisation should only result in a local and temporary reconstruction failure. Furthermore, the reconstructed spectral image may be processed as a locally consistent push broom image, but later georectified using INS data with good global accuracy, but poor local precision, a procedure commonly followed with ordinary HSI data.[12]

Another interesting feature is that, since we accumulate filter mosaics in stable keyframes, there is no need to reproject back into the original, tumultuous camera frames. We instead apply a kind of "digital stabilisation" by reprojecting into virtual push broom cameras with a smoother motion, where only the keyframes are fixed. The remaining frames are given a smooth trajectory on the pose manifold by interpolating along the Lie tangent space vector between the fixed poses $\mathbf{T}_0$ and $\mathbf{T}_1$

$$\mathbf{T}_\alpha = \mathbf{T}_0 \oplus \alpha(\mathbf{T}_1 \ominus \mathbf{T}_0) = \mathbf{T}_0 \operatorname{Exp}(\alpha \operatorname{Log}(\mathbf{T}_0^{-1}\mathbf{T}_1)) \qquad \alpha \in [0, 1], \tag{4}$$

where we have borrowed the notation from Ref. 13. We also adjust the number of poses to interpolate in order to avoid elongation or shortening distortions due to oversampling or undersampling lines with respect to the motion. The number of interpolated poses are chosen so that the translation between them correspond to the ground sample distance (GSD) across the scan direction. This results in push broom images with approximately the same GSD in both directions. The GSD across the scan direction can be estimated as

$$d_{GSD} \approx \frac{\bar{z}}{f_y}, \tag{5}$$

where $\bar{z}$ is the current average depth to the scene in meters and f_y is the scale parameter in pixels for the y-dimension in the camera calibration matrix.

3.3 Spectral reconstruction with OpenGL

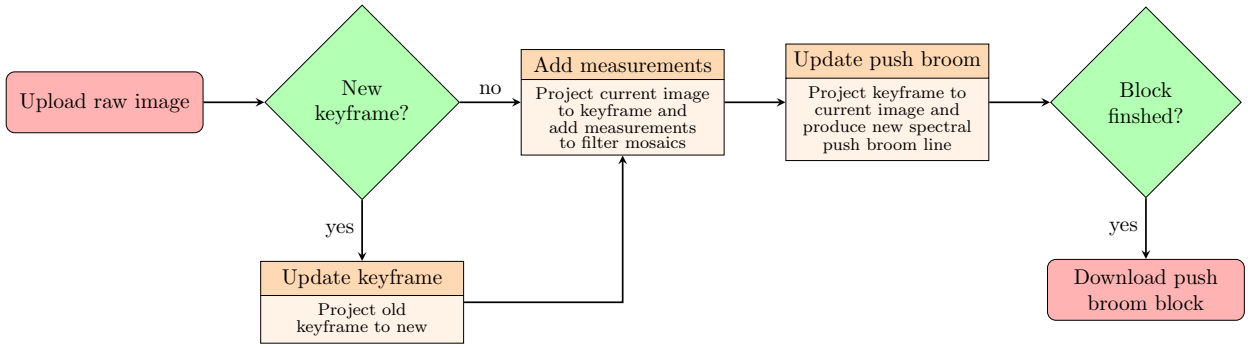

Figure 4. Flowchart for the emulated push broom spectral reconstruction with OpenGL.

A very efficient way to execute the push broom reconstruction procedure in Sec. 3.2 is to exploit computer graphics hardware to perform the reprojections in an extremely parallelised and optimised manner. Furthermore, the computer graphics approach makes it straightforward to represent the scene as full 3D meshes.

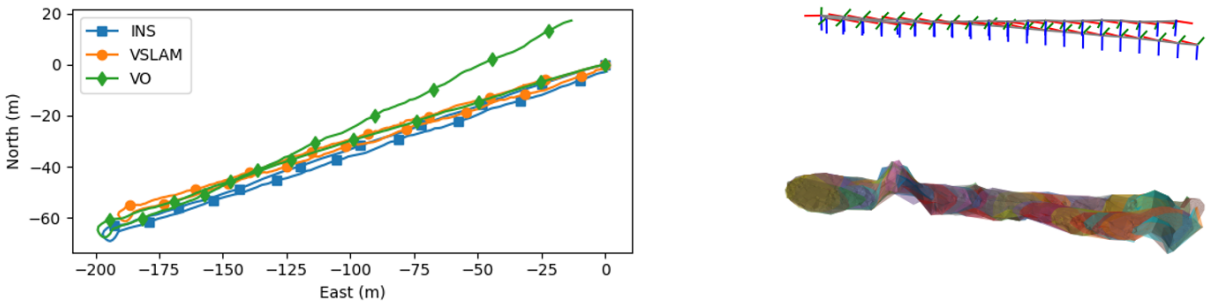

Figure 5. Left: The different pose estimates aligned to INS-data. Right: VO pose estimates and the local surface meshes shown in separate colours.

Our implementation is based on the widely used OpenGL* computer graphics library, which should make it compatible with most relevant computing platforms available for small UAVs. The reprojections in Eq. (2) are performed over 3D meshes by following the *projective texture mapping*[14] technique. The square, orange processing steps in Fig. 4 run on the graphics processing unit (GPU) as consecutive shader programmes executing the related image reprojection with additional processing pixel by pixel. The resulting push broom image contains the six spectral bands averaged over all four sets, plus supplementary channels such as the SIC metric, the world position of each pixel and the depth to the scene. To avoid unnecessarily expensive data transfer between the GPU and the central processing unit (CPU), push broom images are collected as blocks of a chosen size, and downloaded from the GPU when filled.

The projective texturing approach also allows taking shadowing into account by detecting surfaces not visible to the projector, but this involves twice the number of shader executions to estimate depth maps for the projector and will presumably not contribute significantly to the reconstruction when the camera is pointing almost nadir. We have therefore ignored shadowing in this implementation. It is also worth noting that the images formed on the GPU may be accessed directly by other processing algorithms running on the GPU.

4. EXPERIMENTS

The following experiments were carried out on about 33 k images on the East-West flight lines along the road in the dataset presented in Ref. 3, captured by the UAV payload prototype introduced in Sec. 2. All experiments were performed on a HP Z-book laptop running Ubuntu 18.06 with an Intel Xeon CPU @ 2.90 GHz and an NVIDIA Quadro M3000M GPU.

4.1 Real-time pose and structure estimation

Pose estimation using the INS approach was carried out on GNSS and IMU data with NAVLAB[15] in real-time mode. The corresponding structure was represented using a LIDAR-based DEM covering an area of about 1.5 km^2 at a resolution of 25 cm. This resulted in a surface mesh with about 23 M vertices at full size (this surface mesh is used in Fig. 3).

VO was executed with SVO[16,17] using the open source SVO Pro† implementation. The method processed the dataset in real time using the rosbag playback feature in the robot operating system (ROS).[18] 100 of the about 33 k images were not successfully tracked. Their poses were instead estimated based on neighbouring camera positions and angular IMU data. A sparse point cloud of up to a few hundred tracked 3D features were extracted for each new VO keyframe and converted to a local 3D surface mesh using Delaunay triangulation.

For reference, we also use the VSLAM results from Ref. 3 based on ORB-SLAM,[5,6] which processed the data at about 15% of the frame rate. The full sparse point cloud map was converted to a global 3D surface mesh using Delaunay triangulation in the horizontal plane with about 30 k vertices.

*https://www.khronos.org/opengl/wiki/Main_Page
†https://github.com/uzh-rpg/rpg_svo_pro_open

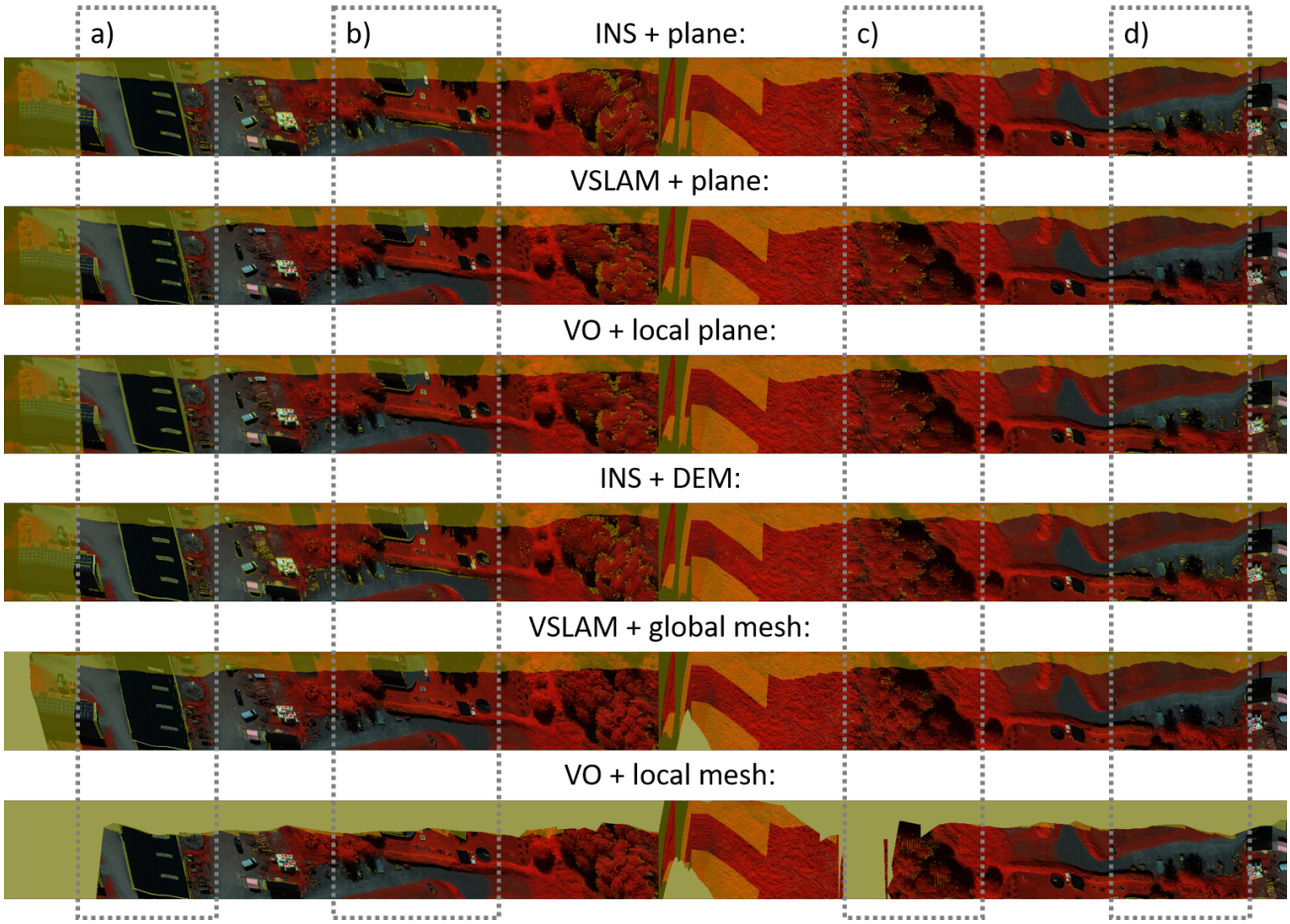

Figure 6. Push broom images in (NIR, G, B) for the different combinations of pose estimation methods and structure representations. Incomplete or inconsistent spectra are flagged in transparent yellow. Sections a) to d) highlight interesting areas referred to in the discussion.

Figure 5 left shows the resulting pose estimates from the different methods aligned with the INS results. As expected, VO exhibits significant global drift compared to the other methods. Figure 5 right shows a 3D visualisation of the VO poses and local surface meshes, where the different meshes are shown in separate colours. Even though the VO pose estimates are drifting with respect to the global frame, local consistency with respect to poses and structure close in time is very good, thanks to the windowed reprojection optimisation over keyframes in SVO.

4.2 Tactical spectral image reconstruction

The proposed tactical push broom spectral image reconstruction presented in Sec. 3.2 was tested using the OpenGL implementation introduced in Sec. 3.3 based on the pose and structure estimates from the previous section. For reference, a planar structure representation for each pose and structure estimation approach was established by fitting planes to the corresponding surface meshes. Figure 6 shows a comparison of the resulting spectral push broom images for the different combinations of pose and structure methods. The spectral images are visualised by mapping the (near-infrared 1 (NIR1), green (G), blue (B)) bands to the (red (R), G, B) channels in the output image. Areas covered by fewer than four filter sets or exceeding a common SIC threshold are flagged in transparent yellow.

In general, we see that all the push broom images have the same global structure, even though the global pose estimates in Fig. 5 are significantly different. It is also clear that the local meshes computed by the VO approach gives the worst structure coverage, aggravated by the fact that the actual scan direction is slightly

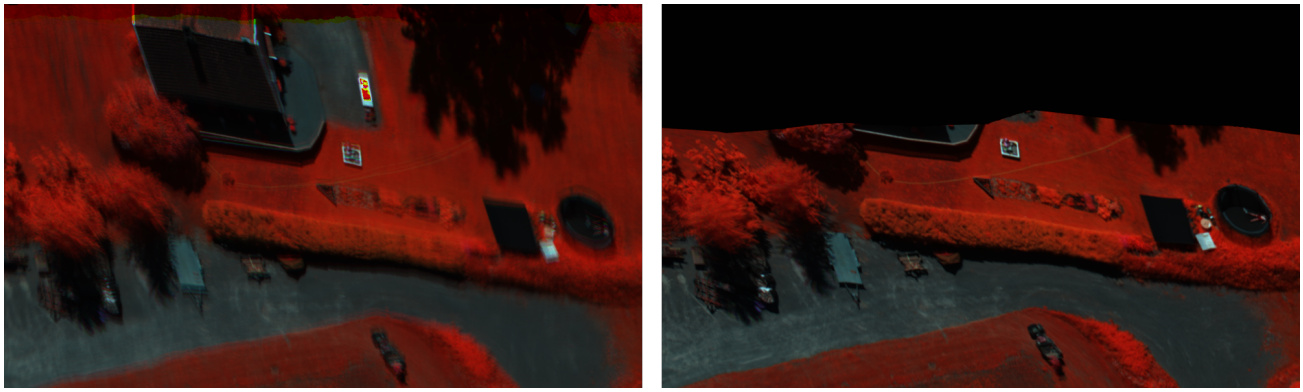

Figure 7. Enlarged view of the push broom images for INS + DEM (left) and VO + local mesh (right) corresponding to section b) in Fig. 6.

upwards (see Fig. 1), which also results in a high number of incomplete spectra in the upper part of all the push broom images.

Sections a) and c) in Fig. 6 highlight areas with significant variations in depth due to buildings and trees. It is clear from the change in consistent spectra and perceived sharpness that the full 3D structure representation generally results in more accurate image reprojections for all methods compared to the planar representation, and thereby a higher quality spectral reconstruction. In section c), the DEM seems to represent the structure of the trees poorly, even seemingly representing trees that are not actually there. This results in worse reprojection accuracy compared to the other mesh-based approaches, and demonstrates the benefits of using the images themselves for estimating observed structure, even though the detail and accuracy may be lower. However, the image-based methods are vulnerable to track loss and estimation failures, aptly demonstrated by the lack of data for VO in parts of this section.

When it comes to the differences in pose estimation, the high number of inconsistent spectra for the INS approach compared to the other pose estimation methods in the relative flat sections b) and d) suggests that the local precision indeed is better for the image-based approaches. The differences in consistency between the INS and VO approaches is clearly visible in the enlarged view of section b) in Fig. 7. Figure 8 demonstrates the benefit of applying our stabilisation approach compared to using every original camera frame for constructing the push broom image. The push broom image without stabilisation on the right is clearly elongated and distorted by variations in camera orientation, while these effects have been largely removed in the stabilised push broom on the left.

The OpenGL spectral reconstruction implementation processed full resolution images at about 3× the frame rate for surface models with up to hundreds of thousands of vertices, and down to about 0.6× the frame rate for the excessively large DEM mesh at about 23 M vertices.

5. DISCUSSION AND CONCLUSIONS

We have presented a complete solution for accurate and efficient spectral reconstruction with a multimodal camera concept adapted to tactical scenarios. The concept offers exploitation of spectral signatures using sensor hardware that amounts, in principle, to a small modification of existing cameras. Reconstruction of spectral imagery is offloaded to software. By introducing a spectral image representation that emulates traditional push broom images, we retain sensor resolution, reduce dependence on global navigation accuracy and focus on exploiting local consistency in efficient image-based pose and structure estimation methods, such as VO. Sample results indicate that high quality spectral reconstruction is feasible even for real-time applications.

More work is still needed to explore to what degree this approach is applicable for more resource-constrained computing platforms on-board small UAVs, and the achievable performance of multispectral target detection in tactical situations. Future work also includes processing data from a multi-camera setup for a broader FOV and more robust image-based pose and structure estimation. The outlook appears promising, since the current

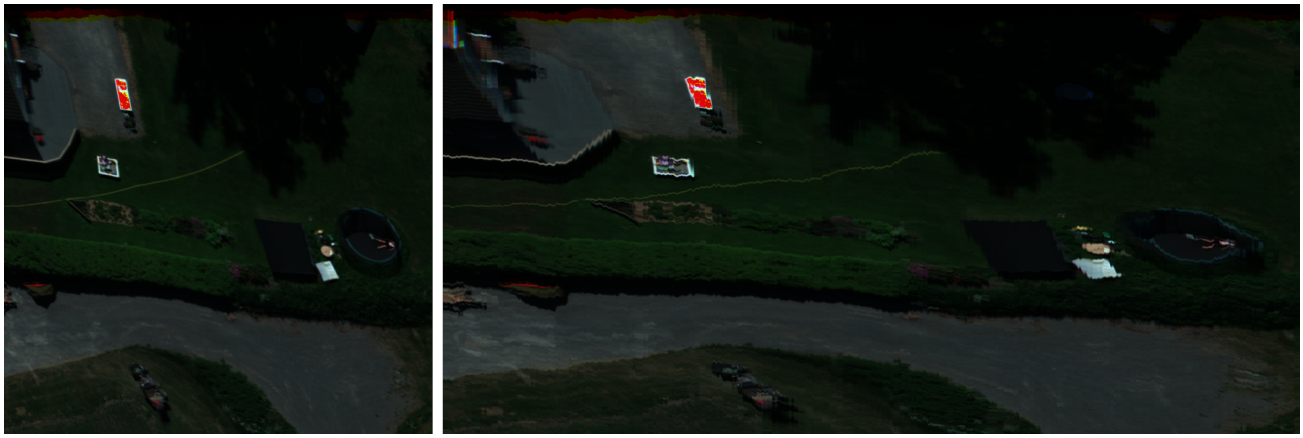

Figure 8. Reconstructed (R, G, B) push broom image based on VSLAM + global mesh with (left) and without (right) our stabilisation approach.

implementation is highly parallel and easily expandable. Tight integration of GNSS data together with VO/VIO for global consistency and georeferencing is another interesting direction of research.

In conclusion, the results demonstrate that a practically relevant performance can be achieved in practice, and indicate that the multimodal sensing concept has a clear potential for use in tactical reconnaissance scenarios.

REFERENCES

[1] Torkildsen, H., Haavardsholm, T., Opsahl, T., Datta, U., Skaugen, A., and Skauli, T., "Compact multispectral multi-camera imaging system for small UAVs," in [*Proceedings of SPIE - The International Society for Optical Engineering*], **9840**, 491–498 (2016).

[2] Skauli, T., Torkildsen, H., Nicolas, S., Opsahl, T., Haavardsholm, T., Kåsen, I., and Rognmo, A., "Compact camera for multispectral and conventional imaging based on patterned filters," *Applied Optics* **53**(13), C64–C71 (2014).

[3] Haavardsholm, T. V., Skauli, T., and Stahl, A., "Multimodal Multispectral Imaging System for Small UAVs," *IEEE Robotics and Automation Letters* **5**(2), 1039–1046 (2020).

[4] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M., "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:2004.10934* (4 2020).

[5] Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D., "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics* **31**, 1147–1163 (10 2015).

[6] Mur-Artal, R. and Tardos, J. D., "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics* **33**, 1255–1262 (10 2017).

[7] Gade, K., "The Seven Ways to Find Heading," *Journal of Navigation* **69**(5), 955–970 (2016).

[8] Scaramuzza, D. and Fraundorfer, F., "Tutorial: Visual odometry," *IEEE Robotics and Automation Magazine* **18**(4), 80–92 (2011).

[9] Forster, C., Carlone, L., Dellaert, F., and Scaramuzza, D., "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry," *IEEE Transactions on Robotics* **33**, 1–21 (2 2017).

[10] Rosinol, A., Abate, M., Chang, Y., and Carlone, L., "Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping," in [*2020 IEEE International Conference on Robotics and Automation (ICRA)*], 1689–1696, IEEE (5 2020).

[11] Rosinol, A., Sattler, T., Pollefeys, M., and Carlone, L., "Incremental Visual-Inertial 3D Mesh Generation with Structural Regularities," in [*2019 International Conference on Robotics and Automation (ICRA)*], 8220–8226 (5 2019).

[12] Opsahl, T., Haavardsholm, T. V., and Winjum, I., "Real-time georeferencing for an airborne hyperspectral imaging system," in [*Proceedings of SPIE - The International Society for Optical Engineering*], Shen, S. S. and Lewis, P. E., eds., **8048**, 80480S (5 2011).

[13] Solà, J., Deray, J., and Atchuthan, D., "A micro Lie theory for state estimation in robotics," Tech. Rep. IRI-TR-18-01, Institut de Robòtica i Informàtica Industrial, Barcelona (2018).

[14] Segal, M., Korobkin, C., van Widenfelt, R., Foran, J., and Haeberli, P., "Fast shadows and lighting effects using texture mapping," *Computer Graphics (ACM)* **26**(2) (1992).

[15] Gade, K., "NAVLAB, a Generic Simulation and Post-processing Tool for Navigation," *European Journal of Navigation* **2**(4), 51–59 (2004).

[16] Forster, C., Pizzoli, M., and Scaramuzza, D., "SVO: Fast semi-direct monocular visual odometry," in [*2014 IEEE International Conference on Robotics and Automation (ICRA)*], 15–22, IEEE (5 2014).

[17] Forster, C., Zhang, Z., Gassner, M., Werlberger, M., and Scaramuzza, D., "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems," *IEEE Transactions on Robotics* **33**, 249–265 (4 2017).

[18] Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., and Ng, A., "ROS: an open-source Robot Operating System," in [*Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*], (2009).