Passive 3D Scene Reconstruction via Hyperspectral Imagery

Corey A. Miller and Thomas J. Walls

Naval Research Laboratory, 4555 Overlook Avenue, Washington, DC 20375
corey.miller@nrl.navy.mil

Abstract. We present the framework for a novel structure from motion (SFM) pipeline to generate 3D reconstructions of low-resolution hyperspectral imagery (HSI). Generating 3D models from a sequence of raw HSI datacubes, where each image pixel retains its spectral content of the scene, significantly expands the analysis currently possible with HSI. In addition to traditional HSI anomaly detection and spectral matching, a 3D spatial model of the scene allows for additional viewing from previously undefined viewpoints, digital elevation map generation, and enhanced object classification capabilities. State-of-the-art SFM techniques are utilized and enhanced by leveraging the spectral content recorded at each image pixel. We explore the potential of this HSI SFM pipeline using an experimental aerial data set collected using a stabilized, 160-band hyperspectral sensor on an aerial platform.

1 Introduction

Advancements in hyperspectral imaging (HSI) sensor technology has allowed for the integration of HSI sensors into small, lightweight gimballed payloads suitable for off-nadir aerial data collection [1]. With more and more HSI data being utilized for motion-based applications including medical-based imaging [2], geological surveying [3], and surveillance/reconnissance [4], the desire to process this spectral content in an intuitive 3D environment has clear advantages. Several researchers have previously integrated spectral analysis into 3D reconstructions. Nieto et al. fuse hyperspectral classifications with laser-based range data to classify 3D geological maps [5]. Similarly, Kim et al. integrate a hyperspectral data with a 3D scanner to study the spectral reflectance of objects [6]. Liang et al. take this a step further and utilize hyperspectral data to accurately segment plants from their background in order to accurately construct 3D models from individual spectral bands [7].

The method presented here seeks to leverage spectral content directly for building a 3D reconstruction of spectral scene content collected on a gimbal-stabilized aerial platform. Feature points are first extracted from multiple subsets of the HSI spectrum, and then correlated between views using a combination of traditional structure from motion (SFM) techniques and spectral matching. One issue that can lead to the weakening of pose estimation algorithms is incorrect feature matching. The integration of a spectral match verification step helps to

G. Bebis et al. (Eds.): ISVC 2014, Part I, LNCS 8887, pp. 413–422, 2014.

eliminate mismatched feature points. Verified matches from each spectral subset are combined and used to compute the overall camera motion between views. A 3D model is constructed from these camera motions, where the spectral content of each 3D point is stored for any form of post-processing analysis typically performed on HSI data, including anomaly detection and spectral match filter generation. Preliminary results are promising, and show that the spectral content can be utilized to improve feature detection and feature matching steps in a SFM pipeline. The rest of paper is organised as follows. Sect. 2 describes the hyperspectral imaging system. Sect. 3 introduces the proposed 3D modelling method. Sect. 4 presents the experimental results, with conclusions and future work given in Sect. 5.

2 Hyperspectral Imaging

Hyperspectral imaging is a technique which collects densely sampled spectral information for each pixel in the image of a scene. Each data collect generates a three-dimensional (x, y, λ) dataset, called a hyperspectral datacube. Rather than the three-band (red, green, blue) collection of standard visible cameras, the increased spectral sample density of an HSI cube allows for the enhanced identification of in-scene objects using their full-spectral signature.

There are various techniques used to create HSI data cubes, including both spatial and spectral scanning, the choice of which depends on the specific application. We limit our discussion here to data cubes generated by the spatial scanning technique. In this approach, the two-dimensional focal-plane array (FPA) is representative of a full slit spectrum (x, λ), with the third dimension (y) being generated through a line-scanning motion. The hardware used in our experiments is a NRL developed gimbal-stabilized short-wave infrared (SWIR) airborne hyperspectral imaging system, with a focal plane array of 1280×1024 pixels, binning SWIR wavelengths into 190 individual spectral bands. The sensor spectrometer, detector array, and optics are integrated into a Wescam MX-20 gimbal system highly stabilized for cued operations at long standoff distances, and can image objects on line of sight at any relative bearing from the platform.

High-fidelity stabilization is a key factor in the application of our technique without in-scene calibration objects. Without this level of stabilization, the irregular movement of the line-scanning platform introduces instabilities in the intrinsic camera parameters within each collected image. These instabilities make standard SFM techniques inadequate for extracting camera pose information, and therefore a consistent 3D reconstruction of the scene. Various laboratory techniques have been developed in an attempt to calibrate line-scanning cameras, however assumptions are made about the stability of the line-scanning that typically do not extend to aerial HSI imagery, such as a constant scan speed [8]. In the absence of these assumptions, the imaging in-scene of specially fabricated calibration objects is often required.

3 Structure from Motion Pipeline

A typical SFM pipeline consists of five main steps; feature extraction, feature matching, baseline triangulation, adding remaining views, and bundle adjustment [9]. In our proposed pipeline, the first two steps reduce the 3D datacubes into a series of matched feature point correspondences between images. These feature matches can then be used to generate a 3D model using state-of-the-art SFM techniques.

3.1 Spectral Feature Extraction

A variety of 2D images can be generated from each HSI datacube through the combination of spectral bands. The most straightforward way to do this is to average over all available wavelengths (λ), reducing each 3D HSI datacube (x, y, λ) to a panchromatic two-dimensional (x, y) data set, similar to a standard photograph. This spectrally-averaged image provides a basis for feature extraction. Using any combination of feature detectors/descriptors, feature points are identified and then quantified. These feature points and their descriptors can then be fed into feature matching routines.

To explore the spectral variation of in-scene objects, we also divide each HSI datacube into several spectral subsets, essentially binning the spectrum into multi-band mini datacubes. These datacube subsets are each averaged across λ as previously described, resulting in several individual images that are spectrally disjoint. For example, if the user wishes to break a 100-band HSI datacube into four equal subsets, then bands 1-25, 26-50, 51-75, and 76-100 would each be averaged into individual images and used for feature extraction. The user can define how exactly to split the datacube, taking advantage of any expectations they have about the spectral response of the scene. Feature points are extracted from each of these images as well. As feature point descriptors will vary amongst the different spectral representations of the scene, they are kept independent of each other until they are fully matched.

3.2 Spectral Feature Matching

As previously described, we make no assumptions about the camera's intrinsic parameters being known (focal length (f_x, f_y), principal point (p_x, p_y), and skew coefficient γ), and therefore can only relate image points up to a projective transformation via the fundamental matrix F constraint

$$x^v F_{v,w} x^w = 0, \tag{1}$$

for pixel x and a pair of cameras $v, w \in 1, \ldots, N$ out of N total cameras. In order to calculate the F matrix, we need a set of matched image points between views. If a specific image point is correctly tracked from one image to the next, the spectral content of that pixel in the HSI datacube will match in both views up to changes in angular reflection. To this end, we have integrated a spectral

matching step into our HSI SFM pipeline. The goal of this spectral matching is to eliminate outliers based on their differing spectral signature from one camera to the next.

To do this, we first apply a brute-force (BF) feature matcher to the set of feature points for a pair of cameras, considering each spectral subset individually. The BF matcher computes distances between feature points in their descriptor-space, and returns the closest corresponding feature point as a match. The output of this BF matching is a set of correlated image points. We then combine results from the multiple spectral subsets (if any), since the feature descriptors are no longer needed. Together, this provides a baseline set of feature point matches. We then compare the spectral signatures of these matches. The full spectrum is extracted for both pixels in a match, and we compute the modified spectral angle similarity (MSAS) between the two spectral vectors. The MSAS_i between spectrums S_i^v and S_i^w of points x_i^v and x_i^w in match i and cameras v, w is defined by

$$\text{MSAS}_i = \arccos\left(\frac{\sum_\lambda S_i^v \cdot S_i^w}{\sqrt{\sum_\lambda \left(S_i^v\right)^2} \cdot \sqrt{\sum_\lambda \left(S_i^w\right)^2}}\right) \cdot \frac{2}{\pi}, \tag{2}$$

where MSAS values range from 0 to 1, with a value of 0 indicating independence. We set an initial MSAS threshold of 0.990 and remove any outlier matches whose MSAS value falls below it. This removes matches that are spectrally differing from future consideration. We then compute the fundamental matrix via RANSAC for the remaining matches, with an inlier threshold set to 3 pixels (measured from the epipolar line in each image). Next we re-evaluate the MSAS values for the remaining matches, raising the MSAS inlier threshold to 0.997. This time, however, we expand our area of consideration to the neighborhood of pixels immediately surrounding each matching point. If a spectral match is found above this threshold within this search area, then we mark it as an inlier; if the strongest spectral match differs in location from the original pixel (i.e. one of its neighbors has a higher MSAS value), the point is updated to the stronger spectral match location. This allows for corrections to be made to the existing matches based on their spectral content. This serves as a much more strict spectral comparison than the first iteration, returning a set of spectrally-verified feature matches that are used to compute the fundamental matrix.

3.3 3D Modeling

We then transform the computed fundamental matrix to an essential matrix, which is a metric relation between scenes, by providing a baseline estimate for the camera's intrinsic parameters. It is important to note here that our estimates for f_x, f_y, p_x, p_y, and γ are not the true camera parameters, and thus do not result in truly undistorted images. These values will be updated throughout the 3D reconstruction process via the weight function that is optimized in the subsequent bundle adjustment step, so our initial baseline values only need to be sufficiently accurate to extract a baseline camera pose. These intrinsic parameters are represented

by a camera calibration matrix (K) which relates the fundamental matrix to the calibrated essential matrix by

$$E_{v,w} = K^t F_{v,w} K, \tag{3}$$

where ^{t} indicates matrix transpose. We can then extract estimated rotation (R) and translation (T) extrinsic camera parameters from our estimated essential matrix. We take the SVD of E and extract R and T by defining matrix W such that

$$\begin{aligned} E &= USV^t \\ T &= VWSV^t \\ R &= UW^{-1}V^t, \end{aligned} \tag{4}$$

where the subscripts v, w are withheld for simplicity. The extrinsic rotation and translation camera parameters allow us to triangulate the raw image feature points into 3D world coordinates, establishing a baseline for our reconstruction. The image point x_i^j, which is the i^{th} point when viewed from camera $j \in 1, \ldots, N$, can be related to its 3D world coordinate point X_i according to

$$x_i^j \propto K_j \left[R_j | T_j\right] X_i. \tag{5}$$

Each point in the baseline image pair is triangulated [10], and a bundle adjustment (BA) routine is applied to the triangulated points [11]. The BA process optimizes the camera positions and 3D point cloud structure through a sparse Levenberg-Marquardt optimization procedure. Additional views are then added iteratively by determining the matching-potential of each possible view based on the number of inliers, selecting the best match, re-triangulating the potential pairwise matches between the new image and the established 3D data points, and using an iterative 3D-2D perspective-n-point (PnP) RANSAC routine based on Levenberg-Marquardt optimization to extract the new extrinsic camera motion. The BA routine is applied after each added camera. This SFM pipeline generates a sparse 3D point cloud, representative of the scene captured in the camera views of the available data set. Since our SFM pipeline assumes the intrinsic camera parameters are never exactly known, a true metric reconstruction cannot be achieved. By allowing these intrinsic parameters to be optimized at each step in the reconstruction however, the resulting 3D structure is often accurately rectified. Projective ambiguities in the x-y plane are mostly corrected through the correlation between views of a 360° orbit, as is often the case when working with aerial imagery.

4 Experimental Results

We explore the applicability of our SFM pipeline on a data set generated from the stabilized HSI sensor in an attempt to not only generate 3D reconstructions of objects in-scene, but to leverage the additional information inherent to having access to additional spectral content. We look to explore the effect that the in-flight line-scanning has on the ability to extract an accurate camera pose, as well

as introduce spectrally unique features into the SFM pipeline to aid in 3D modeling. Our test data set consists of $N = 52$ images, each 1280×966 pixels, covering a full 360° view centered on an industrial complex. Since we have no intrinsic information about this system, we estimate baseline values for the focal length and principal point of the camera. For the focal length, a common baseline estimation is $1.2 * \max(\text{img height}, \text{img width})$; our images are 1280×966 pixels, giving us a baseline focal length of $F_x = f_y = 1536$ pixels. With modern cameras, the principal point is often assumed to be the center of the focal plane, so $p_x = 640$ and $p_y = 483$. If the stabilization of the line-scanning platform is sufficient, the large standoff distance should minimize the shift in camera center that occurs during a scan and the resulting images should reduce approximately to a similar image captured by a framing camera.

4.1 Spectral Subset Analysis

We try to leverage the HSI data cubes by dividing the full HSI spectrum into equally-sized, multi-band subsets that are averaged into individual images for processing, highlighting the spectral variation of objects in the scene. We average over the multi-band subsets in order to reduce noise and bad pixel irregularities found in single-band images. A visual example of the difference in spectral subsets for a given camera can be seen in Fig. 1. We first compare the number of additional, unique feature points each subset provides. Table 1 shows an example using four different feature point detectors (SURF, BRISK, FAST, and ORB [12,13,14,15]), comparing the full-spectrum images to those generated when the full spectrum is split into three subsets. We can see that there is a minimal advantage to using multiple spectral subsets compared to the full spectrum image when considering the unique *location* of feature points.

(a) Spectral subset 1 (b) Spectral subset 2 (c) Spectral subset 3

Fig. 1. Three spectral subsets from a single HSI data cube. The difference in spectral content is seen here as differing shades of gray; the roof of the building and the grass differ significantly between spectral band subsets (a)-(c).

Simply comparing the number of feature points, however, doesn't take into account the variation within feature point descriptors. For example, while the same scene point may be detected in the full spectrum image as well as several of the spectral subset images, the feature point descriptor at that scene point

Table 1. The average number of additional, unique feature points compared between the full-spectrum image and the images created when the HSI data cube was split into three spectral subsets. Only a handful of uniquely located feature points are added as a result of splitting the spectrum into subsets, regardless of the feature detector used.

Feature Detector	SURF	BRISK	FAST	ORB
Avg Full Spectrum	6243	2829	3852	1000
Avg Subset 1	21	13	62	15
Avg Subset 2	0	5	31	16
Avg Subset 3	0	0	0	11
Avg total	23	24	105	35

can vary greatly, making it more-or-less favorable for feature point matching. To this end, we compute feature point matches between subsequent images using a brute-force feature matching with crosscheck verification. The matches from each spectrum subset are combined into a master list which covers the entire spectrum, and is used to compute the fundamental matrix F for overall scene motion. Since we use a RANSAC algorithm, errant matches that don't fit the overall motion are removed. The total number of inlier feature matches is then used as a metric for comparison. Table 2 shows a comparison between the number of inlier feature matches for the full spectrum image, a division of the HSI data cube into three spectral subsets, and a division of the data cube into six spectral subsets. Results are shown for an individual feature point descriptor (SURF), as well as a combination of several feature point descriptors.

Table 2. The number of inlier feature matches compared between the full spectrum image, a division of the HSI data cube into three spectral subsets, and a division of the data cube into six spectral subsets. The number of matches increases significantly with the additional spectral subsets.

	Full spectrum	3 Spectral subsets	6 Spectral subsets
SURF Only	1375	5279	8996
SURF, BRISK,FAST, ORB	2767	10248	18043

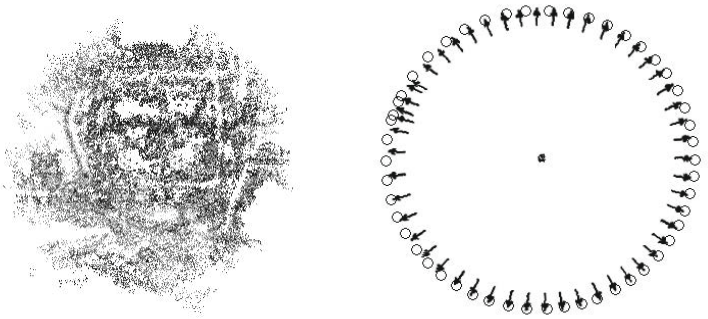
There is a significant increase in the number of accurate feature matches between scenes when considering multiple subsets of the spectrum in combination. This results from the feature descriptors being improved in at least one of the spectral subsets, resulting in more feature matches. A single feature point descriptor (SURF) with all threshold parameters held constant saw a 380% increase in the number of matches between subsequent scenes when comparing the full-spectrum image to three spectral subsets; there was an increase in performance of 650% when combining six spectral subsets. This performance boost was seen across all feature detector/descriptors.

4.2 Spectral Matching

The results of the spectral verification step to the feature matching for our test data set are presented here. On average, the straightforward brute-force feature matching returned 20298 matches between subsequent image pairs. Of these, an average of 1887 (9.3%) were removed in the first round of spectral matching. The resulting 90.7% were used to generate the fundamental matrix relating each image pair, which in turn reduced the 18402 (average) input matches to 13257, a 27.9% reduction. These were then fed into the more strict spectral matcher, which output an average of 8889 matches between views. Of these, there were an average of 320 feature matches (3.6%) which had a point in either image corrected due to a stronger spectral signature match.

4.3 Full-Spectrum SFM

Overall, the 52 input HSI datacubes were fed into our SFM pipeline generated a 3D model consisting of 139,480 points, with 458,627 2D-3D image correspondences and a mean reprojection error of 1.29 pixels. The reconstruction of this data set successfully shows the closed-loop sequence of camera movements as well as the scene structure. A top-view screen shot of the 3D point cloud as well as the extracted camera positions can be seen in Fig. 2. It should be noted again that no a-priori information about the camera's intrinsic parameters ($f_x, f_y, p_x, p_y,$ andγ) or any of the extrinsic camera motions (rotation and translation between views) were used in generating this 3D reconstruction; nothing but the raw, spectrally-flattened 2D images were used as input. When we compare the camera positions to the ground-truth GPS data collected during the data collection orbit, the SFM-generated camera poses are seen to be extremely accurate (Fig 2(b)).



(a) (b)

Fig. 2. (a) A close-up, top-down view of the reconstructed scene; the sparse point cloud is an accurate reconstruction of the original scene structure. (b) The extracted (triangle) and ground-truth GPS (circle) camera positions (enlarged for viewing) can be seen to match well in a circular orbit around the scene.

A second verification of the 3D model's accuracy can be found in the bundle-adjusted intrinsic camera parameters. Since we made no assumptions about the camera's intrinsic values in our reconstruction, we had to estimate values for focal length and principal point. As a result, these are also corrected as part of the optimization step. The 3D reconstruction's final, optimized focal length was within 0.1% from the true focal length that was calculated using the actual lens parameters of the system. The optimized principal point location varied by less than a pixel in both the x- and y-dimensions, confirming the location in the center of the FPA. It's clear that this gimbal-stabilized hyperspectral sensor produces imagery that accurately approximates a framing camera, sufficient to create 3D models using standard SFM techniques in long standoff scenarios in an intelligently constructed pipeline, even though it uses a line-scanning technique to generate one of the physical scene dimensions.

5 Conclusions and Future Work

We have shown the initial development of a 3D reconstruction SFM pipeline for processing HSI data cubes. We have shown that with the appropriate stabilization, hyperspectral line-scanning sensors can be used to generate 3D models of a scene using SFM techniques developed for traditional framing cameras. A full 3D reconstruction was generated using 52 HSI data cubes as input, triangulating over 100,000 3D points into an accurate model of the scene. The resulting camera positions were accurately compared to the ground-truth values, demonstrating the abilities of the SFM pipeline. Additional processing that is unique to HSI data cubes was also explored. We demonstrated the potential that splitting the data cube into smaller, multi-band subsets has on generating useful feature points and feature matches between scenes when compared to the spectrally-flattened data. We also demonstrated the integration of spectral similarity matching for the removal of outlier feature point matches.

Future work includes the development and integration of anomaly detection matching routines. We are currently exploring the usefulness of identifying spectrally unique pixels via anomaly detection that can be used as reliably-tracked feature points for baseline motion estimation. Additionally, we are developing a tool to enable user interaction with the 3D model that allows highlighting spectrally unique areas (via anomaly detection) in the 3D space, along with other post-processing analysis techniques.

Acknowledgements. We would like to acknowledge the contribution from Dr. Eric Allman for his insight into the HSI data structure and processing techniques, as well as Dr. Jonathan Neumann for leading the development of the system sensor used to collect our testing data set.

References

1. Neumann, J., Allman, E.C., Downes, T., Howard, J., Kruer, M., Lee, J., Linne von Berg, D., Leathers, R., Murray-Krezan, J., Nezis, N.: Demonstration of the MX-20SW standoff SWIR hyperspectral imaging ball gimbal system. MSS, Passive Sensors (2008)
2. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. Journal of Biomedical Optics 19, 010901 (2014)
3. Van der Meer, F.D., van der Werff, H., van Ruitenbeek, F.J., Hecker, C.A., Bakker, W.H., Noomen, M.F., van der Meijde, M., Carranza, E.J.M., Smeth, J., Woldai, T.: Multi-and hyperspectral geologic remote sensing: A review. International Journal of Applied Earth Observation and Geoinformation 14, 112–128 (2012)
4. Yuen, P.W., Richardson, M.: An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. The Imaging Science Journal 58, 241–253 (2010)
5. Nieto, J.I., Monteiro, S.T., Viejo, D.: 3D geological modelling using laser and hyperspectral data. In: 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4568–4571. IEEE (2010)
6. Kim, M.H., Harvey, T.A., Kittle, D.S., Rushmeier, H., Dorsey, J., Prum, R.O., Brady, D.J.: 3D imaging spectroscopy for measuring hyperspectral patterns on solid objects. ACM Transactions on Graphics (TOG) 31, 38 (2012)
7. Liang, J., Zia, A., Zhou, J., Sirault, X.: 3d plant modelling via hyperspectral imaging. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 172–177. IEEE (2013)
8. Spinnler, Y.C.K., Wolfsmantel, A.: Calibration of 1d cameras. In: Proceedings of the Vision, Modeling, and Visualization 2004, November 16-18, p. 55. IOS Press, Standford (2004)
9. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2003)
10. Hartley, R.I., Sturm, P.: Triangulation. Computer vision and image understanding 68, 146–157 (1997)
11. Zach, C.: Simple Sparse Bundle Adjustment, SSBA (2011), `http://www.inf.ethz.ch/personal/chzach/opensource.html` (accessed October 2013)
12. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
13. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: Binary robust invariant scalable keypoints. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555. IEEE (2011)
14. Rosten, E., Drummond, T.W.: Machine learning for high-speed corner detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 430–443. Springer, Heidelberg (2006)
15. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571. IEEE (2011)