Snapshot Spectral and Polarimetric Imaging; Target Identification with Multispectral Video

Brent D. Bartlett[a] and Mikel D. Rodriguez[a]

[a]The MITRE Corporation, 7515 Colshire Dr., McLean, VA, USA

ABSTRACT

As the number of pixels continue to grow in consumer and scientific imaging devices, it has become feasible to collect the incident light field. In this paper, an imaging device developed around light field imaging is used to collect multispectral and polarimetric imagery in a snapshot fashion. The sensor is described and a video data set is shown highlighting the advantage of snapshot spectral imaging. Several novel computer vision approaches are applied to the video cubes to perform scene characterization and target identification. It is shown how the addition of spectral and polarimetric data to the video stream allows for multi-target identification and tracking not possible with traditional RGB video collection.

Keywords: Snapshot, Spectral, Polarimetric, Target Identification, Computer Vision

1. INTRODUCTION

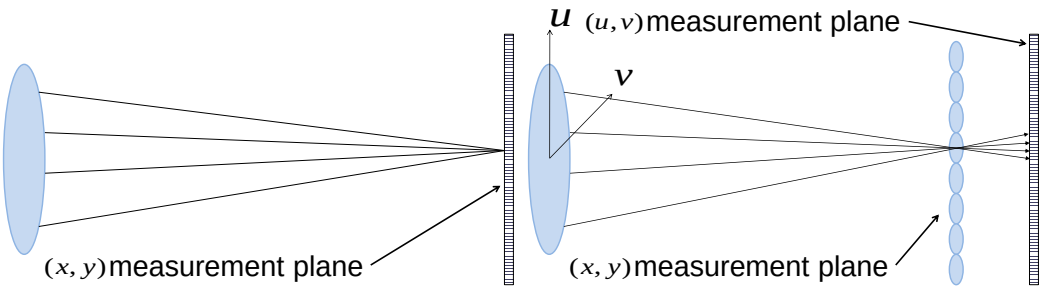
Spectral sensing is a well established modality and has been traditionally implemented through some type of scanning, typically being either temporal or spatial. The types of devices that have been conceived to collect spectral imagery is varied and many approaches have recently been proposed to collect multispectral data over a single integration time of a sensor.[1–3] The majority of these systems trade off spatial resolution to increase the number of spectral bands or even add polarimetric filtering.[4] As these snapshot systems become more prevalent, many applications that were not possible with scanning approaches become feasible in many different fields, such as those demonstrated in various medical applications.[5,6] This work will explore the use of a newly developed snapshot mutlispectral sensor, based on a light field concept,[7] to collect video of dynamic scenes. Traditionally, tracking of moving targets has been an application ideally suited to full motion video (FMV) sensors that collect RGB video. Many computer vision algorithms have been developed around the use of RGB video data and this work will explore the extension of these algorithms to multispectral (MSI) data.

2. SYSTEM DESIGN AND BACKGROUND

The MITRE-developed Spectral and Polarimetric Light Field Camera (SPLiCe) is a system that is capable of collecting data that lies at the intersection of full motion video and hyperspectral systems. It captures imagery with a spatial resolution of $200x134$, a spectral resolution of 20 bands over the spectral range of $0.4um$ to $0.7um$, and 3 band panchromatic linear polarization states. This all occurs within a single integration time, with a maximum rate of four frames per second. As alluded to in the system name, the general approach of this sensor is to utilize a light field imaging concept to facilitate spectral and polarimetric filtering, a brief description of which can be found in the literature.[8] A diagram that contrasts the basic concept of collecting a light field image through the use of a microlens array is shown in Fig. 1. Where a conventional imaging system acquires data in a two-dimensional space, a light field imaging system re-images the lens aperture at each microlens location. Each microlens forms an image of the exit pupil, called a superpixel, on a block of pixels in the FPA. Within each superpixel image, individual pixels measure the magnitude of the ray passing through the corresponding points in the MLA and in the exit pupil. Therefore, the sample rate in (u, v) is driven by the pixel pitch in the FPA. The trade-off that exists between resolution in (x, y) and resolution in (u, v) then becomes apparent.[9]

The current SPLiCe system, shown in Fig. 2, consists of a commercial camera with modified optics. While this system does not collect video at a rate or resolution that would be considered ideal for typical full motion video (FMV) processing, it can be used to explore applying FMV processing techniques to multispectral information. A central element within FMV processing and computer vision in general is the ability to find visually similar matches for a given exemplar, be it a patch of an image or video, an object, or a full image. This fundamental capability represents the basis for most applications of computer vision such as ego-motion compensation and geo-registration, structure from motion, change detection, stereo reconstruction, and tracking. Improvements in the robustness of matching would therefore have beneficial impacts on all of these application domains.

Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIX,
edited by Sylvia S. Shen, Paul E. Lewis, Proc. of SPIE Vol. 8743, 87430R · © 2013 SPIE
CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2016537



(a) (b)

Fig. 1. Conventional imaging (a) vs light field imaging (b). In this architecture, the microlens array forms many sub-images of the object lens exit pupil on the focal plane array.

(a) (b)

Fig. 2. Image of SPLiCe system (a) along with CAD model of aperture filter layout (b).

However, defining a good visual similarity metric can often be surprisingly difficult in typical grayscale/RGB scenarios. Granted, in many situations where the data is reasonably homogeneous (e.g. computing similarity between different patches within the same full motion video sequence), simple grayscale or RGB-intensity based similarity measures work quite well. However, in most real-world datasets a typical query or exemplar of an object of interest is only similar at a more abstract level, but quite dissimilar on the intensity/grayscale pixel level. In this case, pixel-wise grayscale intensity matching fares quite poorly, because small perceptual differences can result in arbitrarily large pixel-wise difference within the intensity/RGB domain.

Currently, the way computer vision researchers address this problem is by using various image feature representations[10–12] that aim to capture the locally salient (i.e. high gradient and high contrast) parts of the exemplars, while downplaying the rest. Despite being successful in certain scenarios, what these features encode are purely local transformations mapping pixel patches from one feature space into another, independent of the global image content. The problem is that the same local feature might be unimportant in one context but crucially important in another. The difficulty of effectively using RGB color as a discriminative feature for recognition in typical FMV scenarios is due in part to strong correlations between the color bands, for example luminance component amounts to approximately 90% of the signal energy.[13] In this work we explore the benefit of using multispectral video processing which should in many cases decorrelate the information between different spectral bands.

3. APPROACH

The hypothesis is that a multispectral / polarimetric video feed will lead to a new class of computer vision processing algorithms that will obtain higher performance than is currently possible through RGB / panchromatic collection. The re-

search question addressed in this study revolves around determining the relative performance increase in tracking obtained through the use of multispectral information vs. traditional RGB imagery. This is tested by collecting a dataset using the SPLiCe imaging system and feeding the full spectral image cubes into a newly developed processing pipeline. The cubes are then spectrally down-sampled to a typical RGB response which is used in a baseline tracking case.

3.1 Test Dataset

A dataset was collected of a parking lot scene in which several cars entered and exited the SPLiCe field of view. A black Jeep and black car were chosen as examples of two different vehicles that have similar signatures and pose a challenge for discrimination. This is a challenging case because black paints in general produce low signal levels relative to more highly reflective car paints and generally have low variability in their spectral signatures. Two video sequences were taken in which several different scenarios were enacted. For example, the car and Jeep were driven past the sensor going in similar and opposite directions, the car was stopped in the field of view while the trunk was open, and both the car and Jeep passed each other at the same time. The first video sequence was captured during the morning and the second was collected in the afternoon producing a total dataset of approximately ten minutes. An exemplar was chosen for each vehicle during the morning video sequence, which was then used by the processing pipeline to find each instance of that vehicle in the rest of the video dataset. Since the illumination conditions vary over the entire dataset, a pre-processing step was performed on the data to mitigate some of these effects, which is discussed in Section 3.2.

3.2 Data Pre-Processing

The raw video data was collected at two times during the day to capture the effects of different illumination conditions. As such, it is important to account for such variations before applying processing algorithms. Since many image cubes were captured, and since in this application the exemplars are found from within the data, the internal average relative reflectance[14] (IAR) method was used. This method has seen success over the years in arid regions that do not contain significant vegetation such as the test scene used in this study while being computationally inexpensive. It is also not crucial to obtain high levels of absolute reflectance accuracy since all algorithms will be searching for exemplars derived from cue frames within the dataset, thus reflectance relative to the scene mean is sufficient. Fig. 3 shows the visual difference in illumination between the morning and afternoon collections and how IAR is effective at minimizing the changes. While this approach does remove the bulk of the illumination changes relative to the mean spectral vector, there are some residual effects left. This residual will be addressed in the processing pipeline which is discussed in Section 3.3.

3.3 Computer Vision Processing Pipeline

Given a video dataset and a set of manually selected targets our goal is to accurately detect the presence of all of the targets across the entire dataset. Fig. 4 shows a diagram that outlines the processing pipeline used, which begins by modeling the background of a scene in order to localize moving blobs in the imagery. We assume a stationary background with the presence of minor dynamic motions, such as moving tree leaves. We then use an adaptive background subtraction method proposed by Stauffer and Grimson.[15] In traditional RGB-based computer vision, background subtracted RGB color values of each pixel are modeled across time, however in this work we extend this traditional modeling approach to include twenty of the multispectral bands that are collected by the SPLiCe system. The spectral bands are modeled by a mixture of K multi-variate Gaussian distributions, termed the Gaussian mixture model (GMM). The probability of k^{th} Gaussian at pixel $p_{i,j}$ is computed as,

$$N\left(x_{i,j}|m_{i,j}^{k},\Sigma_{i,j}\right)=\frac{1}{(2\pi)^{\frac{n}{2}}}e\left(-\frac{1}{2}\left(x_{i,j}-m_{i,j}^{k}\right)^{T}\left(\Sigma_{i,j}^{k}\right)^{-1}\left(x_{i,j}-m_{i,j}^{k}\right)\right) \tag{1}$$

where $x_{i,j}$ is the multispectral vector of pixel $p_{i,j}$. The mean spectral vector $m_{i,j}^{k}$ and the covariance matrix $\Sigma_{i,j}^{k}$ are the k^{th} Gaussian distribution, respectively. For each multispectral pixel, its new value $x_{i,j}^{t}$ at time t is checked against all K Gaussian distributions, and the one that results in the minimum Mahalanobis distance is updated accordingly. If a match between $x_{i,j}$ and the target Gaussian distribution is found, i.e. the distance between them is less than a given threshold, the parameters of the matched Gaussian distribution are updated using an exponential decay scheme. Furthermore, the weight of the matched distribution is incremented by one. If no match is found, the distribution with the lowest weight is replaced with a new distribution having $x_{i,j}^{t}$ as the mean and a pre-defined value as the variance. Given these two conditions

(a) (b)

(c) (d)

Fig. 3. Example RGB images of vehicle driving towards camera in the morning (a) and afternoon (b) along with the corresponding images after pre-processing with IAR (c), (d).

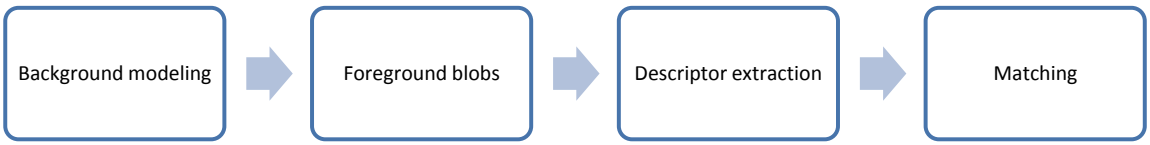


Fig. 4. Processing pipeline used to generate matches of selected targets across entire dataset.

the Gaussian distributions are gradually updated and weights greater than a given threshold, T_w, are incorporated in the set of distributions belonging to the background. In order to obtain segmented foreground blobs we apply a connected-component algorithm over the set of foreground pixels. Spurious foreground regions that arise due to noise are suppressed by eliminating segmented foreground blobs under a minimum area represented by the connected components.

Given a set of foreground bobs a feature vector is constructed by adapting the traditional Histogram of Oriented Gradient (HOG)[11] descriptor to MSI cubes. A grid of uniformly spaced feature point locations, $G = \{(x_i, y_i) : i = 1G, j = 1G\}$ is defined within the foreground blob. Each HOG MSI patch is subdivided into 22 cells. The spectral gradient at a pixel is computed within each cell, and gradients of pixels are accumulated into twelve orientation bins over the range $[0, 2]$. Feature vectors from cells in a patch are concatenated to form a forty-eight dimensional HOG feature vector. This vector is then normalized to unit length for robustness to illumination and contrast changes that were not eliminated by the IAR algorithm described in Section 3.2.

In order to evaluate the detection performance, we follow the PASCAL visual object classification evaluation protocol.[16] A predicted bounding box is considered correct if it overlaps more than 50% with a ground-truth bounding box, where only one detection overlapping with ground-truth bounding box is considered correct, and other overlapping detections are declared as false positives. The performance is measured in terms of precision-recall and average precision (AP) values. This performance is then translated into the true positive and false positive rates, which is used to generate a receiver

operating characteristic (ROC) curve for each method. ROC curves are also generated for two baseline detection methods which will be discussed in Section 3.4.

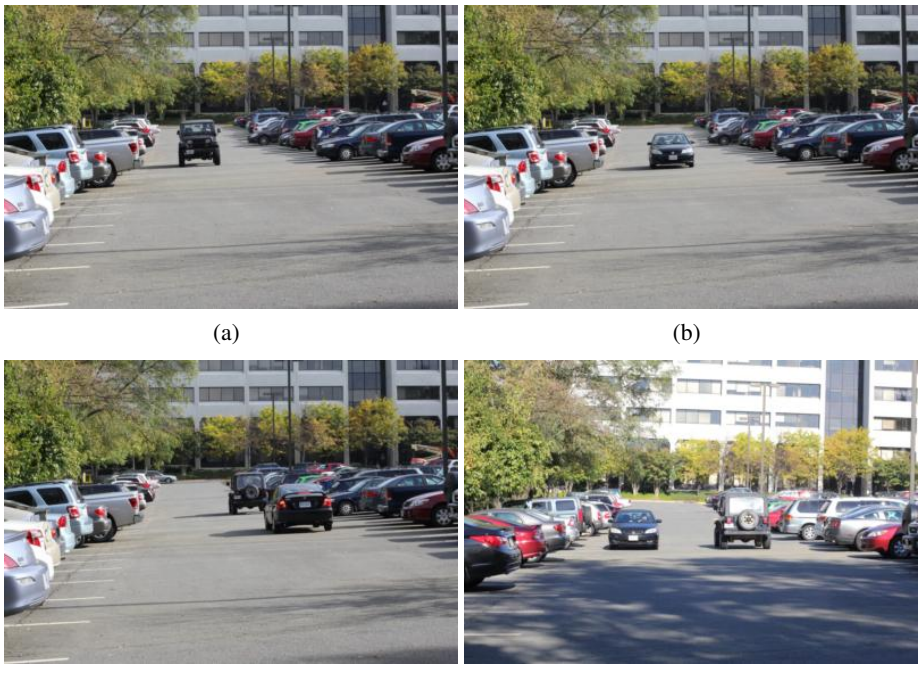
3.4 Baseline Comparisons

In order to demonstrate the advantage of the MSI-based vision processing method we have compared it to two alternative detectors. The first baseline detector is normalized cross correlation in the RGB domain (RGB NCC).[17] Given a set of target templates and an input sequence, normalized cross correlation and non-maxima suppression is performed. The second baseline is the adaptive coherence estimator (ACE) which provides a statistical measure of the likelihood that a given spectral pixel contains a target. In this work, the so called coherent ACE detector is used which operates in de-meaned whitened space which generates the best results for this dataset relative to other ACE implementations.[18]

This provides baseline performance for relatively simple approaches in both the RGB and spectral domains. These algorithms will serve as a benchmark to evaluate the relative performance gains of the newly proposed processing pipeline. The results of applying this pipeline to the multispectral data will show the value of combining computer vision and spectral processing techniques.

4. RESULTS

As indicated in Section 3.1, the SPLiCe system collected a dataset which consists of two video sequences taken of a parking lot in which two vehicles were used as targets to be tracked. A high spatial resolution digital camera was used to collect context imagery of the scene as well. Fig. 5 shows context imagery of the car and Jeep used and two examples of driving maneuvers performed to create a more complex dataset. Two query frames in the morning video sequence were then defined and used throughout the processing. These frames define the target to search the dataset for and the region of interest (ROI) for each are shown in Fig. 6.



(a) (b)

(c) (d)

Fig. 5. Context images captured with high spatial resolution RGB camera of Jeep (a), car (b), following maneuver (c), and passing maneuver (d).

Three new spectral and two baseline detectors were then applied to the dataset as described in sections 3.3 and 3.4. The results of these approaches are presented in the form of a ROC curve, shown in Fig. 7. The baseline methods of RGB-NCC and ACE produce ROC curves indicating that using either RGB spatial information or multispectral information

(a) (b)

Fig. 6. Query frames used along with ROI of Jeep (a) and car (b).

alone yields very similar results. A large gain in performance is achieved through the use of spectral GMM+HOG. Further performance gains are realized by using a combination of traditional computer vision methods, such as GMM+HOG+NCC and spectral methods such as ACE.

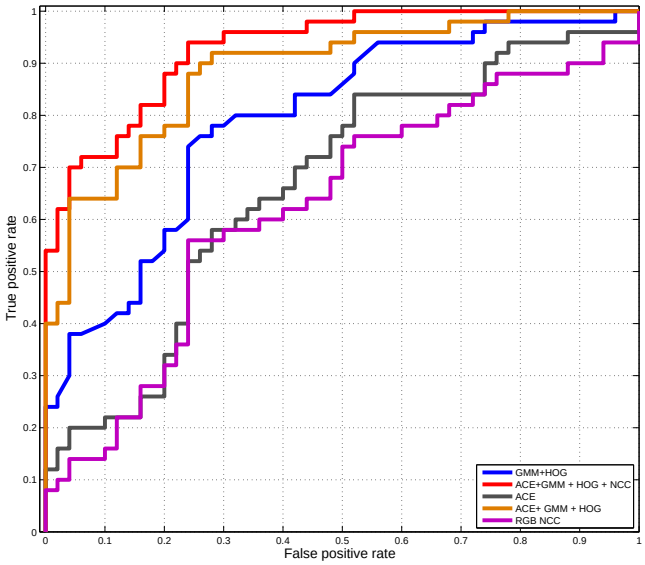


Fig. 7. ROC curve showing detection performance of each detection method over entire dataset.

The SPLiCe sensor also collected broadband polarimetric data. This data can be processed to generate a standard degree of linear polarization (DOLP) image product. This imaging modality can in general be used to discriminate between objects that exhibit high polarization and those that depolarize the collected radiance. One promising avenue for exploiting polarimetric phenomenology is to suppress moving objects that are not cars. For example, a moving dismount next to a moving car is shown in Fig. 8. Both are clearly visible in the grayscale image, but the dismount can be easily removed via simple thresholding of the DOLP image.

5. CONCLUSIONS

A new imaging sensor is described which was used to acquire multispectral and polarimetric video of a parking lot scene containing moving vehicles. A new processing pipeline was developed to exploit this data source that extends computer vision algorithms to utilize multispectral data. The processing technique was applied to tracking two vehicles throughout the dataset which spanned a morning and an afternoon. This new approach shows significant improvement in tracking ROC curve performance relative to two baseline algorithms. Preliminary results also indicate that the use of polarimetric

(a) (b)

Fig. 8. A grayscale video frame is shown in which a dismount is walking near a moving car (a). The same frame is shown in DOLP (b) where the dismount has been removed via a simple threshold.

information could lead to improved discrimination between moving vehicles and dismounts. Future work will explore identifying which spectral bands provide the best improvement in performance as well as the impact of increasing spatial resolution. Methods to incorporate polarimetric information will also be considered.

REFERENCES

[1] Wagadarikar, A., John, R., Willett, R., and Brady, D., "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.* **47**, B44–B51 (Apr 2008). doi:10.1364/AO.47.000B44.

[2] Gorman, A., Fletcher-Holmes, D. W., and Harvey, A. R., "Generalization of the lyot filter and its application to snapshot spectral imaging," *Opt. Express* **18**, 5602–5608 (Mar 2010). doi:10.1364/OE.18.005602.

[3] Gao, L., Kester, R. T., Hagen, N., and Tkaczyk, T. S., "Snapshot image mapping spectrometer (ims) with high sampling density for hyperspectral microscopy," *Opt. Express* **18**, 14330–14344 (Jul 2010). doi:10.1364/OE.18.014330.

[4] Sabatke, D., Locke, A., Dereniak, E. L., Descour, M., Garcia, J., Hamilton, T., and McMillan, R. W., "Snapshot imaging spectropolarimeter," *Optical Engineering* **41**(5), 1048–1054 (2002). doi:10.1117/1.1467934.

[5] Johnson, W. R., Wilson, D. W., Fink, W., Humayun, M., and Bearman, G., "Snapshot hyperspectral imaging in ophthalmology," *Journal of Biomedical Optics* **12**(1), 014036–014036–7 (2007). doi:10.1117/1.2434950.

[6] Kester, R. T., Bedard, N., Gao, L., and Tkaczyk, T. S., "Real-time snapshot hyperspectral imaging endoscope," *Journal of Biomedical Optics* **16**(5), 056005–056005–12 (2011). doi:10.1117/1.3574756.

[7] Horstmeyer, R., Athale, R., and Euliss, G., "Modified light field architecture for reconfigurable multimode imaging," *Adaptive Coded Aperture Imaging, Non-Imaging, and Unconventional Imaging Sensor Systems* **7468**(1), 746804, SPIE (2009). doi:10.1117/12.828653.

[8] Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., and Hanrahan, P., "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR* **2**, 11 (2005).

[9] Georgiev, T., Zheng, K., Salesin, D., and Nayer, S., "Spatio-angular resolution tradeoff in integral photography," in [*Proc. of Eurographics Symposium on Rendering*], (2006).

[10] Lowe, D. G., "Object recognition from local scale-invariant features," in [*Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*], **2**, 1150–1157, Ieee (1999). doi:10.1109/ICCV.1999.790410.

[11] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 886–893 vol. 1 (June). doi:10.1109/CVPR.2005.177.

[12] Tola, E., Lepetit, V., and Fua, P., "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32**(5), 815–830 (2010). doi:10.1109/TPAMI.2009.77.

[13] Geusebroek, J.-M., van den Boomgaard, R., Smeulders, A. W. M., and Geerts, H., "Color invariance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23**(12), 1338–1350 (2001). doi:10.1109/34.977559.

[14] Kruse, F. A., "Use of airborne imaging spectrometer data to map minerals associated with hydrothermally altered rocks in the northern grapevine mountains, nevada, and california," *Remote Sensing of Environment* **24**(1), 31 – 51 (1988). doi:10.1016/0034-4257(88)90004-1.

[15] Stauffer, C. and Grimson, W. E. L., "Adaptive background mixture models for real-time tracking," in [*Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*], **2**, –252 Vol. 2. doi:10.1109/CVPR.1999.784637.

[16] Everingham, M., Gool, L., Williams, C., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision* **88**, 303–338 (2010). doi:10.1007/s11263-009-0275-4.

[17] Lewis, J., "Fast normalized cross-correlation," in [*Vision interface*], **10**(1), 120–123 (1995).

[18] Pieper, M. L., Manolakis, D., Lockwood, R., Cooley, T., Armstrong, P., and Jacobson, J., "Hyperspectral detection and discrimination using the ace algorithm," **8158**, 815807 (2011). doi:10.1117/12.893950.