Real-time Vehicle Tracking in Aerial Video using Hyperspectral Features

Burak Uzkent
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology
bxu2522@rit.edu

Matthew J. Hoffman
School of Mathematical Sciences
Rochester Institute of Technology
mjhsma@rit.edu

Anthony Vodacek
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology
vodacek@cis.rit.edu

Abstract

Vehicle tracking from a moving aerial platform poses a number of unique challenges including the small number of pixels representing a vehicle, large camera motion, and parallax error. This paper considers a multi-modal sensor to design a real-time persistent aerial tracking system. Wide field of view (FOV) panchromatic imagery is used to remove global camera motion whereas narrow FOV hyperspectral image is used to detect the target of interest (TOI). Hyperspectral features provide distinctive information to reject objects with different reflectance characteristics from the TOI. This way the density of detected vehicles is reduced, which increases tracking consistency. Finally, we use a spatial data based classifier to remove spurious detections. With such framework, parallax effect in non-planar scenes is avoided. The proposed tracking system is evaluated in a dense, synthetic scene and outperforms other state-of-the-art traditional and aerial object trackers.

1. Introduction

Aerial vehicle detection and tracking has attracted considerable interest in the computer vision community due to its growing importance in various applications. Numerous studies considering different sensor modalities such as thermal, and electro-optical have been proposed in the past to consistently track ground objects from aerial platforms [3, 23, 29, 28, 19, 2]. However, most of them fail to achieve persistent tracking in real-time due to unique challenges posed by aerial detection and tracking. Aerial tracking (wide-area tracking) is a more challenging task than traditional tracking since the aerial images are typically lower in resolution and the amount of overlap in between subsequent frames is smaller. Low resolution imagery yields a small number of pixels (100-200 pixels) representing a vehicle that degrades the performance of appearance based detection methods. In addition, the sampling rate in an aerial surveillance platform (1-4 Hz) is small compared to common traditional tracking sampling rates (25-50 Hz). Therefore, the displacement of a moving object in subsequent frames is larger in number of pixels, resulting in larger registration errors due to less spatial overlap. Motion based detection methods rely on compensating for global camera motion such as panning, tilting, and rotation to achieve camera stabilization. However, the low sampling rate together with the parallax factor, occlusions and lighting changes are barriers to their application in aerial tracking.

Full and adaptive hyperspectral sensors hold the potential to outperform state-of-the-art aerial trackers in the future due to their ability to record extended target data. [25, 4, 28, 24]. In this work, we consider an adaptive hyperspectral sensor with two modalities: a panchromatic full frame image and hyperspectral data at desired pixel locations determined by the tracker. The full panchromatic frame of the scene can be used to align the images and generate a background subtraction mask in the detection process if the tracking is done from a fixed platform or at high altitude. The hyperspectral data in the visible to near infrared light range provide unique fingerprints for different materials which can be incorporated into the detection process to remove redundant pixels and identify target. Another major advantage of hyperspectral data is it can help in removing hyperspectrally different moving objects in the same category with the target. Such distinct hyperspectral profiles can be visualized in fig. 1. This way, we can gain huge benefits in dense traffic urban environments. Since this type of sensor is not yet fully developed, we will rely on realistic synthetic data to test the proposed approach.

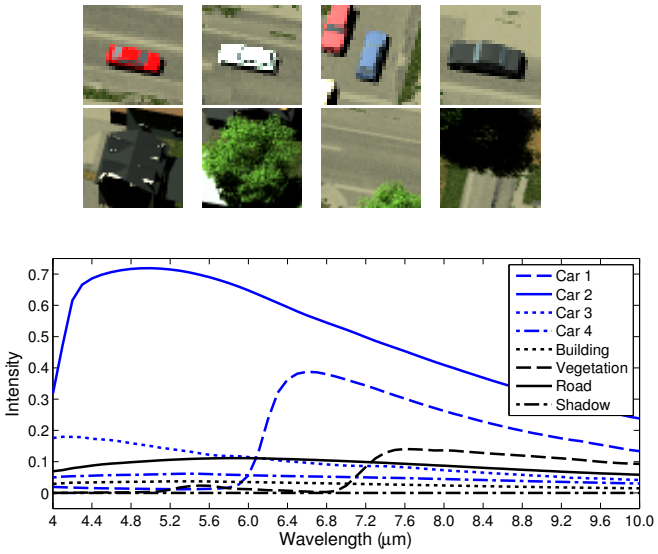


Figure 1: Spectral profiles of different objects. Car 1, 2, and 3 can be separated from the background objects. Central pixel of each object are sampled and shown in the figure.

2. Related Work

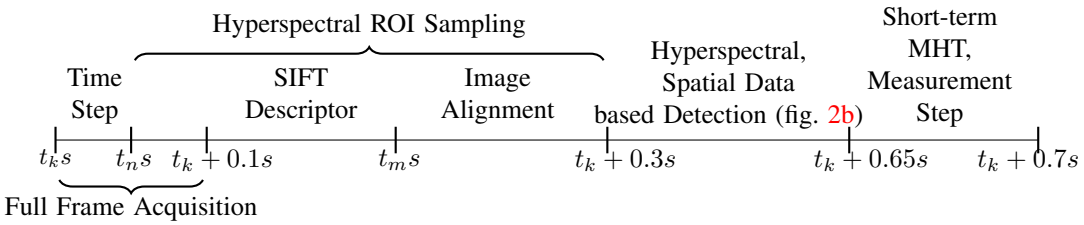
Adaptive hyperspectral sensors are still being developed, but there is a large volume of studies tackling aerial detection and tracking with other sensor modalities such as grayscale, thermal, panchromatic, and LIDAR. Reilly *et al*. [23] proposes detection and tracking of a large number of targets with the Wide-Area Motion Imagery (WAMI) sensor. With the help of six cameras mounted on it, WAMI can provide high coverage and sufficient resolution single-band imagery to accomplish consistent vehicle tracking in long sequences. They detect motion with the well-known median image background model and perform gradient suppression to remove noise due to the parallax effect. Tracking is performed by dividing the scene into equal size grid cells and applying bipartite graph matching within each cell. A set of local scene constraints, such as road orientation is integrated into the graph to improve tracking. However, this method is designed to work in highly planar scenes and assumes minimum parallax error on the road pixels. Palaniappan *et al*. [18] considers the same platform to achieve persistent tracking. Their method is based on an efficient extraction of a set of rich feature descriptors to localize the targets in a region of interest (ROI). They extract region, edge, local shape, and texture based features for the pixels in the ROI. Each pixel in the ROI is classified with the linear binary Support Vector Machines (SVM) and feature likelihood maps are fused in a Bayesian framework. Although they report promising tracking rates, they test the proposed approach with only four vehicles and its feasibilty in real-time tracking is questionable. In the work proposed by Xiao *et al*. [29], a joint probabilistic relation graph with vertex and pairwise edge matching is presented to detect and track vehicles in aerial video sequences. Geographical road structure information is incorporated into model vehicle driving behavior including potential travel direction and velocity. The vehicle driving behavior model is included in the graph structure (vertices and edges) to solve the assignment problem optimally. Motion detection is achieved by the three-frame subtraction approach. As in [23], this motion detection approach might fail in an urban environment where the registration and parallax error can be larger. In addition, such detection methods can not be implemented for an adaptive hyperspectral sensor platform since the detection results lag the actual frame by one time step.

The contribution of this paper is two-fold. To the best of our knowledge, it is the first paper performing an adaptive multi-modal (panchromatic and hyperspectral) sensor based single target tracking without using any of the background subtraction techniques. Second, we will publish the aerial panchromatic video set generated for this study in addition to the full-frame hyperspectral video set ($150\ GB$).

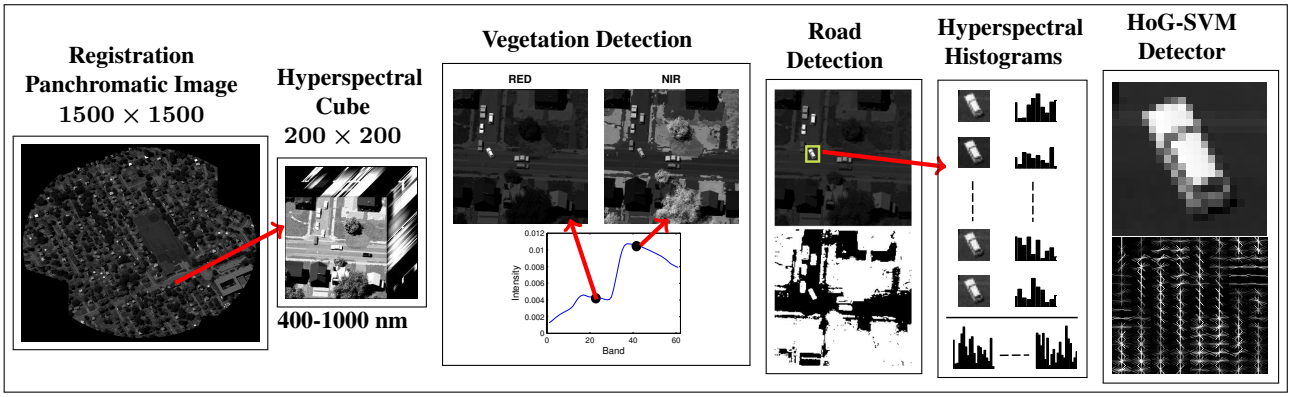
3. Tracking System and Sensor Resource Management

A sensor capable of collecting spatial and hyperspectral data is required to achieve high rate tracking of objects of interest. For this reason, the *Rochester Institute of Technology Multi-object Spectrometer* (RITMOS) proposed by Meyer *et al*. [16] is considered as an adaptive, multi-modal sensor. RITMOS utilizes a Digital Micromirror Array Device (DMD) to reflect the light to one of the two sensors; a spectrograph or a panchromatic channel. The switch from panchromatic to hyperspectral data mode for a pixel can happen very fast due to the compactness and speed of micromirror arrays. To capture a panchromatic image of the scene, an array of micromirrors reflects the light to a panchromatic imaging array. Individual micromirrors imaging the object are then tilted to reflect the light towards the spectrograph and collect the full spectrum of a specified pixel.

A realistic tracking method needs to align well with the RITMOS specifications. First, RITMOS requires about 0.1 s to capture a panchromatic image of a scene. On the other hand, the full hyperspectral profile of a single pixel in the visible to near infrared wavelength takes 1 ms. Spatial and hyperspectral data can be collected simultaneously as long as the micromirror array transfers the light to only one of the two paths. One disadvantage of the hyperspectral imagery is the possibility of spatial misregisteration when there is relative motion between TOIs and surroundings. Some methods have been proposed to precisely calibrate hyperspectral sensors onboard. We ignore this distortion since a



(a)



(b)

Figure 2: The workflow of the proposed tracking system (a) and modules of the detection process in order (b). In (b), a ROI is selected and sampled hyperspectrally for target detection. HoG features are computed in the final step to minimize false alarms. Grayscale imagery for the ROI is computed by summing the bands in visible wavelength.

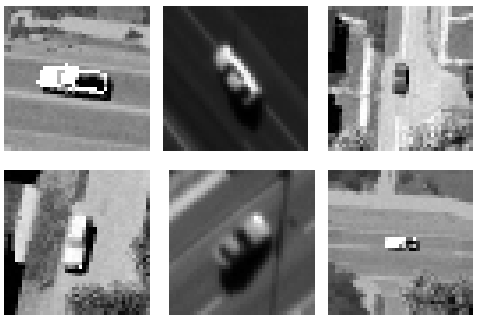
vehicle does not occupy a large number of pixels in width ($\approx$ 15 pixels). With hyperspectral sensors, it is inevitable to make a trade-off between either high spatial resolution imagery with lower hyperspectral resolution or low spatial resolution imagery with higher hyperspectral resolution. We opt for higher spatial resolution imagery, as the vehicle confirmation module relies on an appearance based method. This way, intermixing issues on the subpixel level are mitigated. On the other hand, panchromatic aerial sensors are capable of providing higher resolution imagery. However, in this study we keep the spatial resolution of the hyperspectral and panchromatic modalities the same, as panchromatic images are only used to compute the homography matrices. They are not exploited in moving object detection due to parallax effect and the time lag between the panchromatic image and hyperspectral data acquisition, as seen in fig. 2a.

4. Scenario Generation

The *Digital Imaging and Remote Sensing Image Generation* (DIRSIG) model is used to generate a synthetic aerial video [11]. The scenario used in this paper comes from the DIRSIG Megascene I, which is built to simulate part of Rochester, NY, USA. The simulation uses hyperspectral imaging from an aerial platform orbiting around a specified center in Megascene 1 area. The platform moves with 90 m/s constant velocity at an altitude of 3000 m. The hyperspectral range is 400 to 1000 nm with a hyperspectral resolution of 10 nm, so that generated synthetic images have 61 rectangular wavelength bands. The sensor parameters are also defined to meet the real-world phenomenology. The focal length is set to 225 mm whereas the detector area is $17\times17\mu^2$, matching the Texas Instruments DMD specifications. Pixel pitch is tuned to agree with detector dimensions so that there is minimum gap between the adjacent pixels. The average ground sampling distance is 0.30 m which yields low resolution imagery with 1500×1500 pixels. With these settings, vehicles in both panchromatic and hyperspectral images cover around 150 pixels and this enables us to use appearance based methods. On the other hand, in [27] pixel-to-pixel based matching is used as vehicles occupy fewer number of pixels as seen in fig. 3. More detail in scenario generation is documented in [27].

5. Image Alignment

Since the aerial platform is non-stationary, we need to remove global camera motion. As in most aerial track-



(a) (b) (c)

Figure 3: Two cars are shown in a DIRSIG generated grayscale image in this study (a) and in [27] (b), whereas (c) displays two cars from a WAMI (CLIF 2007) dataset [1]. Vehicles in (a,b) cover 100 - 150 pixels whereas in (c) they are represented by about 50 pixels.

ing studies we employ a keypoint based feature extraction method on two subsequent frames and find the correspondences between extracted points. The Scale Invariant Robust Features (SIFT) is used to compute the descriptor for each keypoint as it is robust to illumination changes, viewpoint difference, and rotation [15]. In the final step, the RANSAC algorithm is employed to robustly fit the homography matrix. The homograpy between the first and following frames (k) is computed by accumulating the homographies (H) for subsequent frames as

$$H_{k,1} = H_{k,k-1} * H_{k-1,k-2} * H_{k-2,k-3}, ..., H_{2,1}. \quad (1)$$

To propagate the $H_{k,1}$ (t_k+0.1s) to the hyperspectral data domain (t_k+0.3s), we factor out the translation and rotation matrices and scale rotation angle and translation components by $0.9/0.7$.

6. Target Detection

Target detection is a major step in persistent tracking. To achieve it, a narrow FOV ROI (200×200 pixels) is determined based on the prior mixture probability density function mean estimated by the prediction stage of the filter following Uzkent *et al.*'s work [27]. Once the ROI is sampled hyperspectrally, the tracking algorithm searches the ROI to detect the target of interest to update the track statistics. Background modeling in a moving platform is a difficult task as the scene constantly changes in addition to frequent stop-then-go motion. However, background subtraction have been extensively used in the aerial tracking studies with additional modules to handle its drawbacks[26, 12, 22, 13, 3, 6]. 3-D stabilization [5] is another way to improve background subtraction prone to parallax effect. The proposed approach stands out from most of the aerial tracking literature by excluding the computationally simple but severely limited methods in the detection process. The modules of the proposed detection approach and workflow can be visualized in fig. 4.

6.1. Vegetation Detection

In the first two steps, the pure spectrum of the individual pixels are considered to filter out as many background pixels as possible to optimize the search space. First, the *Normalized Difference Vegetation Index* (NDVI) is applied [7]. Its uniqueness stems from the fact that vegetation absorbs light extensively in the red wavelengths and reflects most light in the near infrared spectrum, causing a relatively large intensity difference in these bands. Therefore, the NDVI can be formulated as

$$\frac{I_{NIR} - I_{RED}}{I_{NIR} + I_{RED}} \geq T_{NDVI} \quad (2)$$

where I and T_{NDVI} are the pure spectrum of a pixel and empirical threshold, respectively. At each step, T_{NDVI} is selected randomly from an interval bounded by $T_{NDVI} + T_{NDVI}*0.2$ and $T_{NDVI} - T_{NDVI}*0.2$. That is because the performance of NDVI can be sensitive to external factors.

6.2. Road Detection

In the second module, non-vegetation labeled pixels are classified as road or non-road. This is achieved with the pure hyperspectral information of individual pixels as the road pixels show a consistent hyperspectral signature under different conditions [10]. However, asphalt dominated pixels mostly have flat spectrum and they can not be as easily distinguished as the vegetation pixels. For this reason, we train a non-linear SVM to avoid misclassification of target pixels. The goal of this step is to detect as many road pixels as possible without losing any target pixel. In order to

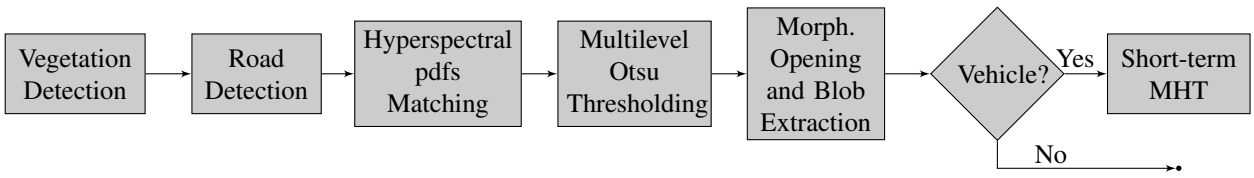


Figure 4: The proposed target detection approach workflow.

train the SVM, the same scene is generated at four different time settings. The road and non-road pixels are sampled from the synthetic scene generated about 30 minutes (11:30 am) before the scene generated for tracking (12:00 am). In total, 300 road and 1200 non-road training samples are collected to train the SVM. The radial basis function (RBF) kernel is applied to perform a transformation to a higher dimensional space as the dimensionality of the input space is the number of hyperspectral bands (61), to prevent underfitting. The RBF kernel parameter is tuned on the validation data collected from the tracking scene where the vehicles of interest do not travel. The final classifier achieves an 87.4% accuracy rate on the 2000 validation samples.

6.3. Local Hyperspectral Histograms Matching

In the third step, local hyperspectral histograms are used on the remaining pixels with a sliding window technique to generate a hyperspectral distance map. The target hyperspectral histogram model is built when the user selects a vehicle initially. For each hyperspectral band, n-bin histograms are built, resulting in a feature vector, p, with $n\times61$ dimensions. Assume $x_i \in R^2$ represents the location of one of the pixels in the detection window with intensity I_i^λ ($\lambda = 400, 410, 420, ...990, 1000\ nm$). A mapping function, b, is designed for x_i to output the bins with intensity values neighboring I_i^λ. Then, a hyperspectral pdf located at y in the ROI for λ is formulated as

$$p(u_{j-1},\lambda) = \sum_{i=1}^{N} \frac{|I_i^\lambda - u_{j-1}|}{|u_j - u_{j-1}|}\delta[b(x_i,1) - u_{j-1}], \quad (3)$$

$$p(u_j,\lambda) = \sum_{i=1}^{N} \frac{|I_i^\lambda - u_j|}{|u_j - u_{j-1}|}\delta[b(x_i,2) - u_j], \quad (4)$$

where N and δ denote the total number of pixels in the detection window and dirac delta function. Once the feature vector is computed, each channel histogram is normalized to improve robustness against lighting changes. Comparison between each pixel's hyperspectral pdf and the target model hyperspectral pdf is done with the Chi-Square distance metric and a hyperspectral distance map (D_{map}) for the ROI is computed. To avoid outlier contributions in histogram computation, vegetation and asphalt dominated pixels are eliminated in the manner described in sections 6.1 and 6.2. The computation of hyperspectral pdfs is expensive due to the large number of hyperspectral bands. To optimize the computation process, we subsample by using the only odd numbered bands, resulting in a $n\times31$ dimension feature vector as shown in fig. 5. This improves the algorithm speed without degrading detection rates dramatically as the neighboring bands correlate largely. Additionally, the integral image theorem is utilized [21]. Integral hyperspectral pdfs are computed for the ROI, resulting in an $n\times31$ hyperspectral histogram integral image. Integrating the mapping kernel into histogram computation with the integral image concept is problematic, however, the benefit of computing a hyperspectral band's pdf in $O(3\times n)$ outweighs the importance of a kernel function.

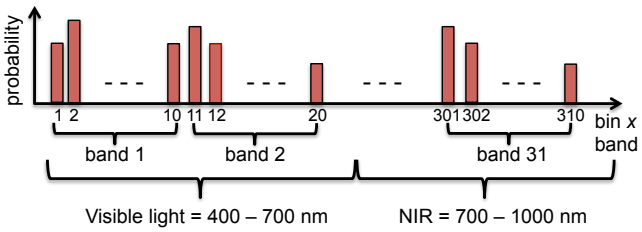


Figure 5: Stacked hyperspectral feature vector with 10 bins in each band.

The size of the sliding detection window is kept fixed since we already know a priori the size of a typical vehicle assuming the platform altitude is fixed. Additionally, pixels classified as road or vegetation are not considered in the ROI integral hyperspectral pdfs. The resultant D_{map} is then applied a threshold determined by the multilevel Otsu's method [17, 14]. With a fixed hyperspectral threshold, ST, in different scenarios, outliers are allowed in the final mask as the level of hyperspectral distinctness of different vehicles shows strong deviation. This adaptive thresholding concept becomes key in removing most of the outliers before the final vehicle verification step. In other words, the idea with multilevel Otsu's threshold method is to incorporate the fact that a TOI matches to itself best hyperspectrally under different conditions. Multilevel Otsu's thresholds are computed by minimizing the intra-class variances as

$$T_f^1 = T_1^1, T_2^1, ..., T_n^1, \quad (5)$$

$$T_f^{k-1} = T_1^{k-1}, T_2^{k-1}, ..., T_m^{k-1} \quad (6)$$

where n and m stand for the number of ROI segmentation levels used in the first and previous time step k-1 and T denotes the thresolds at corresponding levels. The hyperspectral threshold update framework is then designed as

$$ST_1 = T_1^1, \quad (7)$$

$$ST_k = \alpha * ST_{k-1} + (1-\alpha) * T_1^{k-1}. \quad (8)$$

If the target is lost l number of previous frames, the hyperspectral threshold becomes

$$ST_k = ST_{k-l-1}. \quad (9)$$

This way, we avoid relaxing it in the occlusions, as Otsu's method attempts to allow some outliers. Finally, we get the ROI binary mask for the TOI as

$$TOI_{map}(i,j) = \begin{cases} 1 & Dmap(i,j) < ST_k \\ 0 & otherwise \end{cases}. \quad (10)$$

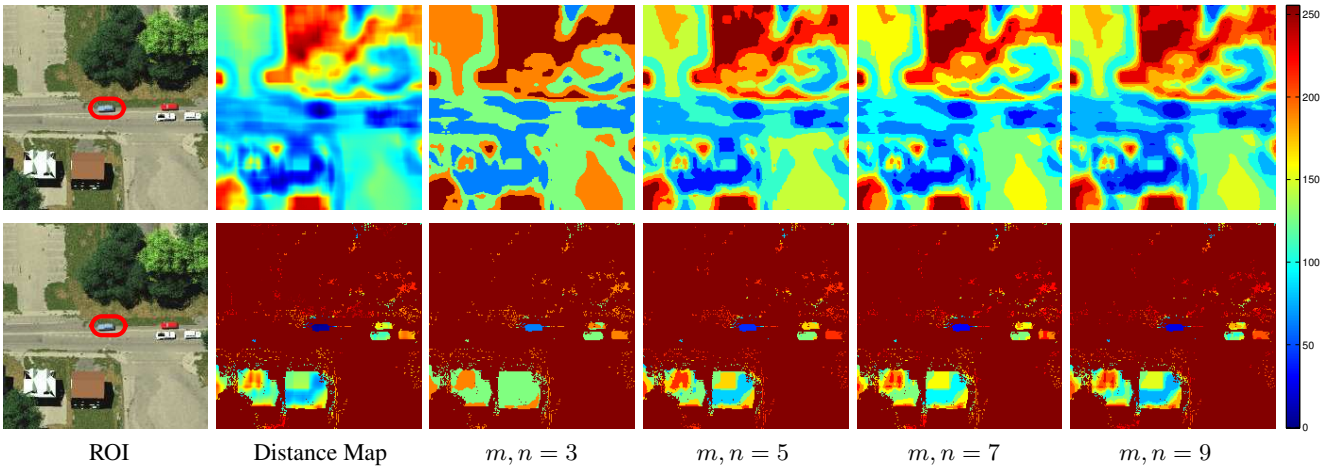


Figure 6: The influence of number of levels (m and n) used in multilevel Otsu's threshold method in ROI segmentation. NDVI and road detection modules are excluded in the first row figures and included in the second row figures.

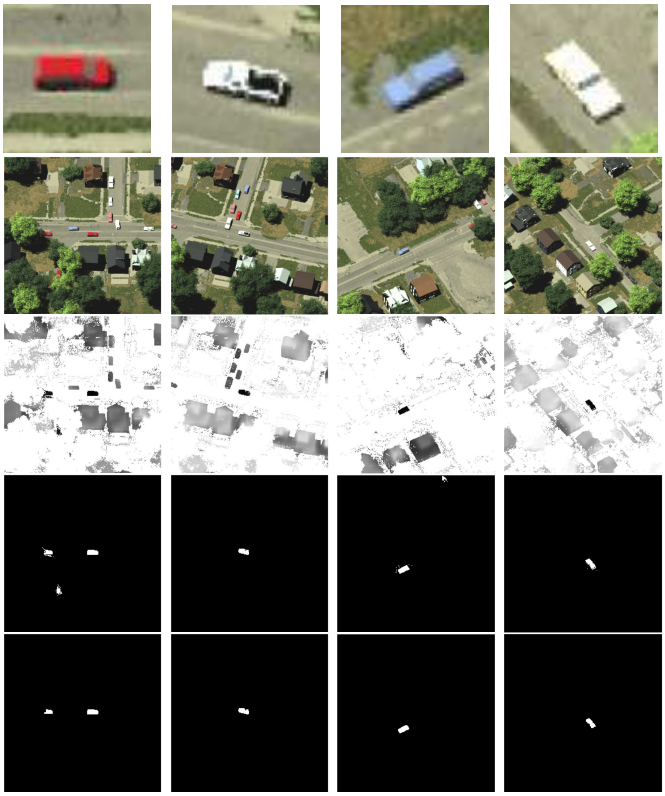
Morphological opening is then performed to get the blobs and IDs are assigned with the connected component labeling algorithm.

The effect of m and n in detection performance is quantified in fig. 6. As seen in fig. 6, without the NDVI and road detection modules, low m and n values tend to oversmooth the target segmentation which results in more background inclusion. However, with the inclusion of NDVI and road detection steps, we locate the target and have no false alarm in all the cases. Therefore, the NDVI and road detection modules not only minimize the false alarms but also reduce the sensitivity to m and n.

Local hyperspectral pdfs can contribute greatly to TOI detection. For instance, in fig. 7a there is a vehicle near the target with much lower similarity values due to differences in the hyperspectral domain. In fig. 7b we are interested in a white truck, and again, hyperspectral pdf matching eliminates the background pixels in addition to all the other moving vehicles. However, vehicles with multiple color materials show relatively weak matching. This non-uniform vehicle structure might result in the inclusion of more outliers into the final mask. In the fourth case, the yellow vehicle shows the strongest discrimination as all the other pixels are assigned low similarity values. However, in the third case, the TOI has blue paint model with a similar hyperspectral profile to some of the building pixels. We can conclude that the local hyperspectral pdf matching shows promising results in adverse scenes.

6.4. Non-Vehicle Blob Removal

In the final step, grayscale and panchromatic images are utilized to extract local, spatial, gradient-based features and detect non-vehicle blobs. The Histogram of Oriented Gradients (HoG) is widely used in the vehicle and human detection literature [9]. It relies on the gradient information of a detection window to compute features highlighting the object contour. It has been very successful in aerial vehicle detection due to their rigid shape. Traditional HoG splits the detection window into a number of blocks and each block is divided into a number of cells. Then, each cell produces a gradient histogram and, as a result, each block outputs a number of gradient histograms. These histograms are stacked and normalization is performed to increase robustness against illumination changes. Finally, the feature vector of each block is stacked to get the final HoG features. Linear SVM is cascaded with the HoG features, as they provide large number of features. Employing a HoG-SVM vehicle detector with a sliding window technique is computationally expensive and not feasible in high frame rate real-time tracking systems. In this study, however, we do not compute HoG features of every pixel in the ROI with a sliding window technique. Instead, it is applied on the candidate blobs to verify if they belong to a vehicle. For this reason, we opt to perform *l-2* normalization to normalize block features unlike the traditional HoG implementation where *l-1* normalization is performed due to its computational efficiency. HoG is computed on panchromatic image chips by summing all the bands in the visible wavelength since it provides the highest contrast in this region. We apply a linear kernel as it is commonly prefered with the HoG features since a non-linear kernel implementation can be costly. In order to train the SVM, we collect 2500 vehicle and non-vehicle chips at five different time settings from the areas of the Megascene 1 where the TOIs do not travel. To find the optimal values for HoG parameters, we collect 1000 positive and negative test samples from the areas of



(a) (b) (c) (d)

Figure 7: First row displays the zoomed images of the TOIs in the RGB images of the ROIs (second row). Hyperspectral similarity distance maps after vegetation and road segmentation are shown in the third row. Fourth and fifth rows show extracted masks after thresholding as in eq. 10 and generated final masks after vehicle blob confirmation step. (Red circles represent the targets.)

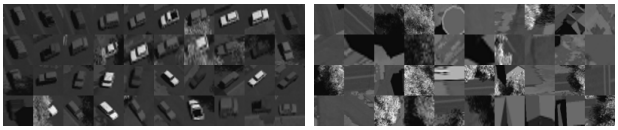


Figure 8: Some of the positive and negative samples in the HoG-SVM based vehicle detection training dataset.

interest. This way, we can tune the HoG parameters such as the size of the cells, number of the cells in a block, and overlap percentage in the neighboring blocks. We resize the image chips to 64×64 pixels as this setting outputs the highest accuracy together with 8×8 cells in a 16×16 block and 50% overlap in adjacent blocks. In total, 1754 HoG features are extracted from the panchromatic image chips. The designed HoG-SVM detector classifies the test samples with 92.75% accuracy.

7. Data Association

The multi-dimensional assignment (MDA) algorithm, first proposed by [20], considers S number of past scans to formulate the data association problem. It is also known as the practical MHT algorithm. A number of target features including kinematic, feature-based similarities, and shape can be integrated into the practical MHT algorithm in different manners. In this study, we insert the kinematic and hyperspectral features-based likelihoods in a weighted sum Bayesian fashion. Kinematic likelihoods can be estimated via a parametric probabilistic modeling approach using the filter. Kinematic likelihoods are assigned lower weight due to low frame rate and hyperspectral likelihoods are given higher weight due to distinctive nature of hyperspectral data. In the detection step, the hyperspectral score f for each blob i is computed. The hyperspectral likelihood F is then formulated as

$$F_i = \frac{ST - f_i}{\Sigma_{m=1}^{M}(ST - f_i)} \tag{11}$$

where M and ST are the number of validated blobs and threshold, determined in sect. 6.3. The hyperspectral likelihood aided MDA algorithm is covered comprehensively in Uzkent *et al.*'s work [27]. In the final step, the track's state space matrix are updated with the assigned measurement. In the filtering stage, we implement the Gaussian Mixture Filter (GMF) to estimate state space matrix parameters as detailed in [27].

8. Results

The experiments were executed on a personal computer with a 2.9 GHz, i7 processor. Table 1 displays the run times of the detection modules. SIFT implementation on a 1500×1500 image can be costly. However, it can be quickly computed on a GPU or we can downsample the image to make the algorithm real-time. Another way to reduce Image Alignment execution time can be to use Harris-Corner detector to find keypoints. We compute the Homographies offline since the primary contribution of this paper is on robust target detection.

We consider the *Track Purity* (TrP) and *Target Purity* (TgP) metrics to measure tracking performance as shown below.

$$TrP[t_j] = \frac{\underset{1\leq i\leq b}{max}\ A_{ji}}{\sum_{i=0}^{b} A_{ji}}, \quad TgP[g_i] = \frac{\underset{1\leq j\leq c}{max}\ A_{ji}}{\#\ frames\ in\ g_i} \tag{12}$$

where b and c denote the number of ground truth platforms g and tracks t and A_{ji} stores the number of times g_i is assigned to t_j. Since we track a single target at separate runs c at most can be one and i=0 represents dummy target assignment. TrP evaluates how many frames t_j is assigned

Module	Veg. Detection	Road Classifier	Spectral Histograms	HoG SVM
Run time	0.002 s.	0.008 s.	0.15 s.	0.05 s.

Table 1: Run time performances of the detection modules.

dominant g_i during the track life whereas TgP measures the ratio of the number of times g_i is associated to dominant t_j to the duration of g_i. The TrP metric favors short tracks and it might be misleading in cases where track terminations occur frequently. On the other hand, the TgP metric considers the life of the ground truth so that potential misleading information due to the TrP is avoided.

We compare the proposed Hyperspectral Feature based Tracker (HFT) to several popular tracking algorithms. Three trackers are considered in the category of the kinematic data based trackers. They are Nearest Neighbor Tracker (NN), Probabilistic Data Association Filter (PDAF), and Multiple Hyphothesis Tracker (MHT). Additionally, two traditionally known object trackers, Mean-shift [8] and a real-time object tracker via an online discriminative feature selection learning (OFDS) [30], are considered. Finally, a recent state-of-the-art aerial vehicle tracker, wide-area aerial tracker via likelihood of features tracking (LoFT) [19], is considered. It should be highlighted that LoFT source code is not available online and we use the results published in the paper since the resolution of the CLIF dataset used in [19] is similar to the generated dataset in this study.

Tracks are initiated interactively with the user selection. As we are interested in a single target tracking, the user is asked to click roughly on the center of the vehicle. Then, a 20×20 pixels size window is used to extract 3-D hyperspectral pdfs. Overall, 43 vehicles are extracted. In fig. 9, initial ROIs for some of the tracks and binary masks produced after vegetation and road detection modules are shown. The proposed track initiation approach is less prone to user error since we do not necessarily need a strict target bounding box.

Figure 9: Initiation of the some of the tracked vehicles and vegetation and road pixels removed binary masks.

As seen in table 2, the proposed HFT outpeforms the other trackers by a large margin in terms of both TrP and TgP. We note that when the tracker fails it is due to the combination of a large density of occlusions and multiple object with similar hyperspectral profiles. This is supported by the fact that on average 20% of the time a vehicle is partially or fully occluded.

Trackers	Track Purity	Target Purity
NN	39.25	34.65
PDAF	26.07	14.19
MHT	39.20	35.07
Mean-shift[8]	8.88	8.88
OFDS[30]	12.66	12.66
*LoFT[19]	60.30	40.50
HFT	69.78	60.30

Table 2: Comparison of the proposed hyperspectral tracker with other trackers. LoFT has not been tested on the generated scenario since its source code is not available. However, it has been tested on a similar CLIF aerial video set (see fig. 3) in Pelapur *et al.*'s work [19] and the results are copied from that work. NN, PDAF, and MHT are provided true measurements from the vehicles in the ROI.

9. Conclusion

We investigated the unique challenges posed by wide-area surveillance from a moving platform and proposed a real-time detection and tracking framework based on an adaptive sensor capable of producing a wide FOV panchromatic image and narrow FOV hyperspectral image. The proposed framework focuses on tracking a single target with higher persistency in complex environments, as the hyperspectral data acquisition and processing are costly. The use of hyperspectral information introduced high false alarm rates for the vehicles with less distinctive hyperspectral profiles. By exploiting the spatial domain in addition to the hyperspectral domain, we reduced the false alarm rates without degrading the recall rates dramatically. In the future, we plan on fusing the likelihoods maps from each band with adapting weights. This way, dependency on the road classifier can be minimized. Also, we will work on testing the proposed approach on a scenario generated from a real platform by using a state-of-the-art hyperspectral camera.

References

[1] WAMI Columbus Large Image Format (CLIF) Dataset. https://www.sdms.afrl.af.mil/index.php?collection=clif2007,. 2007. 4

[2] A. Basharat, M. Turek, Y. Xu, C. Atkins, D. Stoup, K. Fieldhouse, P. Tunison, and A. Hoogs. Real-time multi-target tracking at 210 megapixels/second in wide area motion imagery. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 839–846. IEEE, 2014. 1

[3] S. Bhattacharya, H. Idrees, I. Saleemi, S. Ali, and M. Shah. Moving object detection and tracking in forward looking infra-red aerial imagery. In *Machine Vision Beyond Visible Spectrum*, pages 221–252. Springer, 2011. 1, 4

[4] J. Blackburn, M. Mendenhall, A. Rice, P. Shelnutt, N. Soliman, and J. Vasquez. Feature aided tracking with hyperspectral imagery. *Proc. SPIE*, 6699:66990S–66990S–12, 2007. 1

[5] B.-J. Chen and G. Medioni. 3-d mediated detection and tracking in wide area aerial surveillance. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 396–403. IEEE, 2015. 4

[6] B.-J. Chen and G. Medioni. Motion propagation detection association for multi-target tracking in wide area aerial surveillance. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015. 4

[7] M. A. Cho, A. Skidmore, F. Corsi, S. E. Van Wieren, and I. Sobhan. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation*, 9(4):414–424, 2007. 4

[8] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000. 8

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 6

[10] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, 2013. 4

[11] E. J. Ientilucci and S. D. Brown. Advances in wide-area hyperspectral image simulation. In *AeroSense 2003*, pages 110–121. International Society for Optics and Photonics, 2003. 3

[12] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 441–448. IEEE, 2013. 4

[13] P. Liang, G. Teodoro, H. Ling, E. Blasch, G. Chen, and L. Bai. Multiple kernel learning for vehicle detection in wide area motion imagery. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 1629–1636. IEEE, 2012. 4

[14] P.-S. Liao, T.-S. Chen, and P.-C. Chung. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, 17(5):713–727, 2001. 5

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 4

[16] R. D. Meyer, K. J. Kearney, Z. Ninkov, C. T. Cotton, P. Hammond, and B. D. Statt. RITMOS: a micromirror-based multi-object spectrometer. In *Astronomical Telescopes and Instrumentation*, pages 200–219. International Society for Optics and Photonics, 2004. 2

[17] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 5

[18] K. Palaniappan, F. Bunyak, P. Kumar, I. Ersoy, S. Jaeger, K. Ganguli, A. Haridas, J. Fraser, R. M. Rao, and G. Seetharaman. Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video. In *Information fusion (FUSION), 2010 13th Conference on*, pages 1–8. IEEE, 2010. 2

[19] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan. Persistent target tracking using likelihood fusion in wide-area and full motion video sequences. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2420–2427. IEEE, 2012. 1, 8

[20] A. B. Poore. Multidimensional assignment formulation of data association problems arising from multitarget and multisensor tracking. *Computational Optimization and Applications*, 3(1):27–57, 1994. 7

[21] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 829–836. IEEE, 2005. 5

[22] J. Prokaj, X. Zhao, and G. Medioni. Tracking many vehicles in wide area aerial surveillance. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 37–43. IEEE, 2012. 4

[23] V. Reilly, H. Idrees, and M. Shah. Detection and tracking of large number of targets in wide area surveillance. In *Computer Vision–ECCV 2010*, pages 186–199. Springer, 2010. 1, 2

[24] A. Rice and J. Vasquez. Context-aided tracking with an adaptive hyperspectral sensor. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011. 1

[25] D. Rosario. Hyperspectral target tracking. In *Aerospace Conference, 2011 IEEE*, pages 1–10, 2011. 1

[26] X. Shi, H. Ling, E. Blasch, and W. Hu. Context-driven moving vehicle detection in wide area motion imagery. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2512–2515. IEEE, 2012. 4

[27] B. Uzkent, M. J. Hoffman, and A. Vodacek. Spectral Validation of Measurements in a Vehicle Tracking DDDAS. *Procedia Computer Science*, 51:2493–2502, 2015. 3, 4, 7

[28] T. Wang, Z. Zhu, and E. Blasch. Bio-inspired adaptive hyperspectral imaging for real-time target tracking. *Sensors Journal, IEEE*, 10(3):647–654, 2010. 1

[29] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 679–684. IEEE, 2010. 1, 2

[30] K. Zhang, L. Zhang, and M.-H. Yang. Real-time object tracking via online discriminative feature selection. *Image Processing, IEEE Transactions on*, 22(12):4664–4677, 2013. 8