2010-08-20

Unusual-Object Detection in Color Video for Wilderness Search and Rescue

Daniel Richard Thornton
Brigham Young University - Provo

Unusual-Object Detection in Color Video for Wilderness Search and Rescue

Daniel R. Thornton

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Bryan Morse, Chair
Michael Goodrich
Dan Olsen

Department of Computer Science

Brigham Young University

December 2010

ABSTRACT

Unusual-Object Detection in Color Video

for Wilderness Search and Rescue

Daniel R. Thornton

Department of Computer Science

Master of Science

Aircraft-mounted cameras have potential to greatly increase the effectiveness of wilderness search and rescue efforts by collecting photographs or video of the search area. The more data that is collected, the more difficult it becomes to process it by visual inspection alone. This work presents a method for automatically detecting unusual objects in aerial video to assist people in locating signs of missing persons in wilderness areas.

The detector presented here makes use of anomaly detection methods originally designed for hyperspectral imagery. Multiple anomaly detection methods are considered, implemented, and evaluated. These anomalies are then aggregated into spatiotemporal objects by using the video's inherent spatial and temporal redundancy. The results are therefore summarized into a list of unusual objects to enhance the search technician's video review interface.

In the user study reported here, unusual objects found by the detector were overlaid on the video during review. This increased participants' ability to find relevant objects in a simulated search without significantly affecting the rate of false detection. Other effects and possible ways to improve the user interface are also discussed.

Keywords: Wilderness search and rescue, anomaly detection, aerial imagery, user study

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Bryan Morse, for his unwavering confidence and support. I would also like to thank everyone else who has contributed to the BYU WiSAR research group, especially Dr. Michael Goodrich and Ron Zeeman. I'd like to thank Nathan Rasmussen, Carson Fenimore, Doug Kennard and everyone in the Graphics and Vision lab for their insight in answering countless questions and helping me work out solutions to many problems. I am indebted to Jim Walker, Wallace Barrus, and Craig Randall for volunteering their time, equipment and expertise in collecting aerial imagery. I am especially grateful for my wife who encouraged me to get into research and pursue a graduate degree. She has stood by me through this whole process and without her constant reassurance, I may not have realized my own potential. I am eternally grateful to my Savior, Jesus Christ. Without Him, I would never have overcome the many obstacles that stood in my way.

Contents

List of Figures vii

List of Tables ix

1 Introduction 1

1.1 An Unmanned Aerial System for Wilderness Search and Rescue 1

1.2 Challenges of Aerial Search 3

1.3 Solution Overview 5

2 Background 6

2.1 Video Enhancement 6

2.2 Spectral Anomaly Detection 7

2.2.1 Modeling the Spectral Distribution 7

2.2.2 Normalization Methods 8

2.2.3 Summary 10

2.3 Related Detection Methods 10

2.3.1 Background Subtraction 11

2.3.2 Object Detection 11

2.3.3 Human Detection 11

2.3.4 Behavioral Anomaly Detection 12

2.4 Summary 12

3 A Detector for Unusual Objects 14

3.1 Anomaly Detection Implementation 14
3.2 Data Collection 15
3.2.1 Photography 15
3.2.2 Image Labeling 15
3.3 Spectral Detector Evaluation 16
3.3.1 Evaluation Results 18
3.3.2 BACON 24
3.3.3 Summary of Results 26
3.4 Object-based Detection 26
3.4.1 Problems with Pixel-based Detection 27
3.4.2 Aggregation 28
3.4.3 Object Filtering 28
3.5 Summary 28

4 User Study **30**

4.1 User Interface 30
4.2 Data Used 31
4.2.1 Targets 33
4.2.2 Suggestions 34
4.2.3 Sequence Generation 35
4.3 Results and Statistical Analysis 36
4.3.1 True Positives 37
4.3.2 Tone Counting Error 37
4.3.3 False Positives and Positive Predictive Value 38
4.3.4 Distraction Effects on True Positives 38
4.4 Summary 40

5 Conclusion **41**

5.1 The Solution 41

5.2 User Study Results 42

5.3 Limitations and Future Work 43

5.3.1 Detector vs. Interface 43

5.3.2 Other Detectors 44

5.3.3 Other Interfaces 44

5.4 Summary 45

A Pre-study Questionnaire **46**

B User Study Instructions **47**

C User Study Follow-up Questions **48**

D User Study On-screen Instructions **49**

References **52**

List of Figures

1.1 UAV equipped for aerial search 2
1.2 UAV control station 2
1.3 Example video frames 4

2.1 RX convolution kernel 9

3.1 Target FPR 17
3.2 Anomaly detector results 18
3.3 RX results 19
3.4 Vector quantization results 20
3.5 K-means results 20
3.6 EM results 21
3.7 Results for EM with too many iterations 24
3.8 BACON results 25
3.9 Target and measured FPR for BACON 26

4.1 Target objects at the grassy location 32
4.2 Target objects at the desert location 32
4.3 Suggestions as blue circles 34

D.1 Instruction slide 1 49
D.2 Instruction slide 2 49
D.3 Instruction slide 3 50
D.4 Instruction slide 4 50

D.5 Instruction slide 5 50
D.6 Instruction slide 6 51
D.7 Instruction slide 7 51
D.8 Instruction slide 8 51

List of Tables

3.1 RX results 19
3.2 Vector quantization results 20
3.3 K-means Results 21
3.4 EM results 22

4.1 Video clip content 33

Chapter 1

Introduction

Wilderness search and rescue (WiSAR) is the task of finding missing persons in wilderness areas. Many people go missing in wilderness areas every year. Utah's Grand County Search and Rescue performed 29 searches in 2008 alone [20]. This task is highly time-sensitive, not only because of increasing danger to the search subject (missing person), but also because the search radius increases over time. The purpose of this work is to aid searchers in finding traces of the search subject.

1.1 An Unmanned Aerial System for Wilderness Search and Rescue

A fast and efficient way to cover a large search area is with an aircraft-mounted camera. With enough resolution (pixels per square meter), searchers can identify traces of the search subject from the air quickly and effectively. The system presented in this work was specifically designed for a search platform being developed by the BYU WiSAR research group [9] (Figures 1.1 and 1.2).

Small aircraft such as mini Unmanned Aerial Vehicles (UAVs) (Figure 1.1) are useful for their size and portability but are limited to small, lightweight equipment. In addition to portability issues, the need for immediate feedback from the aircraft necessitates the use of video transmitters.

The planes are outfitted with small NTSC video cameras [9]. The video is transmitted live to a ground station along with the plane's telemetry for control of the UAV as well as locating targets on the ground. The video and telemetry are also logged for off-line review.

Figure 1.1: UAV equipped for aerial search

Figure 1.2: UAV control station

By processing the video collected of the search area, searchers identify possible signs of the search subject.

1.2 Challenges of Aerial Search

Even with the aid of aerial video cameras, it can be difficult to identify signs of a search subject. Figure 1.3 shows examples of simulated search images. Image resolution is limited both by the camera itself and by the need to cover as much ground as possible. In addition, the video can move quickly, disorienting searchers and giving them little time to detect targets before they move out of view. Consequently, searchers may miss signs of the search subject, even when captured by the camera.

In practice, the image processing task consists of two steps: target detection and target analysis. In addition to the missing person, targets include abandoned clothing or other personal items. Targets vary in difficulty. For example, a blanket (Figure 1.3a) may be easier to detect than a shirt (Figure 1.3b). Depending on available manpower, BYU WiSAR currently depends on one or more human technicians to perform both detection and analysis of possible targets.

Target detection is the task of quickly identifying possible signs of the search subject. When performed by a human technician, detection is characterized by a reflexive action, such as a keystroke, accompanying the appearance of an unusual or significant object. The analyst then tries to determine, through inspection of the imagery, whether the target is likely to be a positive sign. For example, in the interface described in [13], one mouse click was required at initial detection to freeze the display. After a brief analysis of the target, the user would localize it in the image with a second click. The simplicity of detection and its relative lack of dependence on domain knowledge make it a good candidate for automation.

Because of the sensitive nature of search and rescue, the search team will most likely want to have all imagery reviewed by a human, even with the aid of an automated detector. Because time and resources are precious, any unusual objects found by the detector will need

(a)

(b)

Figure 1.3: Examples of video frames from a simulated search. Targets are marked with yellow arrows: (a) blue blanket and (b) white shirt.

to be confirmed by inspection of the imagery before further action is taken. In addition, it must be assumed that the detector will miss some objects of interest. It is therefore possible that an object could be missed by the automated detector but detected by a human, especially a trained and experienced searcher. Therefore, the goal of this work is to use automated detection as an aid for visual search rather than a replacement for it.

1.3 Solution Overview

This work primarily relies on leveraging color information in video of the search area to detect signs of a missing person. This technique is referred to as spectral anomaly detection. For the best chances of success in applying spectral anomaly detection to this domain, multiple detection methods were implemented and compared.

Using the temporal and spatial information in the video stream, unusually colored pixels are aggregated into larger objects. The resulting list of unusual objects can be presented to a technician in many ways. This work includes a user study in which unusual objects are marked in the video to catch the searcher's attention.

Analyzing the results of spectral anomaly detection methods show that it works well in this domain. The results of the user study show that automated detection helps technicians to find objects of interest without increasing false detection.

The layout of this thesis is as follows: Chapter 2 discusses work related to automated detection, with a focus on spectral anomaly detection. Chapter 3 documents the implementation and evaluation of a system for identifying unusual objects in video of natural scenes. In addition to a rigorous comparison of multiple spectral anomaly detection methods, a method of aggregating spectral anomalies into a concise list of objects is presented. Chapter 4 discusses a user study performed to show the effects of the detector as an aid for the search task. Finally, Chapter 5 discusses the conclusions that can be drawn from the results of this work.

Chapter 2

Background

Tools related to this work can be found in multiple fields of research. BYU WiSAR has developed useful tools for visual enhancement of aerial search video. In the field of hyperspectral imagery, various methods have been proposed for anomaly detection, which may be applicable to traditional color images as well. Another closely-related field is video surveillance, including human detection methods. This chapter explores each of these fields and how these methods may be applied for detecting objects of interest in aerial search.

2.1 Video Enhancement

One enhancement method is temporally local mosaics. As the video is played back, transformations are calculated between video frames, and each image is then repositioned to better align with its immediate neighbors. In addition to extending the length of time that an object is visible, the mosaic provides more context and some amount of stabilization. Such displays have been shown to improve detection rates when searching for specific objects in the video [13].

Other enhancement methods developed for BYU WiSAR visually enhance objects of interest. The objects of interest can be visually enhanced using the hue and saturation of their color values [18] or by the amount of heat they produce [17]. No study was performed on the effectiveness of color enhancement, and neither of these enhancement methods has been studied in conjunction with temporally local mosaics.

2.2 Spectral Anomaly Detection

Most of the work for spectral anomaly detection has been for use on hyperspectral images. A conventional image consists of a grid of pixels, where each pixel is a triple of brightness values for the three primary colors of light: red, green and blue. Thus, each pixel is a three-vector in the RGB color space. A hyperspectral image is of a similar form, except that each pixel contains a much greater number of samples from a wide range of spectral bands.

2.2.1 Modeling the Spectral Distribution

A common approach [19, 3, 21] to hyperspectral anomaly detection is to model the statistical distribution of spectral signatures with one or more multivariate normal distributions. This model is then used to identify pixels whose spectral signatures are statistical outliers. The normal distribution is most likely used for its simplicity. Once the mean vector and covariance matrix have been calculated, outliers can be identified using a threshold on the Mahalanobis distance

$$d_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})} \tag{2.1}$$

where $\vec{\mu}$ is the mean vector and Σ is the covariance matrix. In a multivariate normal distribution, the Mahalanobis distances are distributed according to the chi-square distribution with cumulative distribution function

$$F(d_M; k) = P(k/2, d_M/2) \tag{2.2}$$

where k is the dimensionality of the multivariate normal and P is the regularized Gamma function. The distance threshold can therefore be chosen to encompass a desired probability. In the case of one-dimensional data, this method yields the well-known bell-curve confidence intervals. When data points are RGB triples, $k = 3$.

2.2.2 Normalization Methods

Unfortunately, a multivariate normal distribution rarely characterizes all of the colors in a natural scene. This means any effective spectral anomaly detector must perform some transformation on or clustering of the data for it to fit the assumption of normality. Such procedures will be referred to here as normalization. Once the data has been normalized, the mean vector and covariance matrix are estimated in order to calculate the Mahalanobis distance of each pixel.

The RX Algorithm

Perhaps the simplest normalization method is the RX algorithm [19]. The basic assumptions of the algorithm are that each pixel is drawn from a multivariate normal distribution, but that the mean and variance of the distribution change across the image. The variance is generally assumed to change more slowly than the mean. So much so, that it is common to use the same variance estimate for the entire image but to calculate this using a spatially-varying mean [5]. The mean is usually calculated within a window near, but not including, the very immediate neighborhood of the pixel. Once this local mean has been subtracted, it is straightforward to calculate the covariance matrix and thereby the Mahalanobis distance of each pixel. Apart from the Mahalanobis distance threshold, the only parameters to this algorithm are the outer radius, R, and the inner radius, r, of the local neighborhood.

In image processing terms, these steps can be thought of as an unsharp masking operation (Figure 2.1), resulting in a residual error image. Next, a color transformation is applied to the error image with the inverse of its color covariance matrix. The next step, which does not seem to have an image processing analogue, is to take the dot product of each transformed error vector with the corresponding untransformed error vector. This results in a gray-scale image of Mahalanobis distances, to which is applied a threshold chosen with Equation 2.2.

-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1
-1	-1	0	0	0	-1	-1
-1	-1	0	40	0	-1	-1
-1	-1	0	0	0	-1	-1
-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1

Figure 2.1: Example of an RX convolution kernel: This unsharp mask kernel computes the residual error for each pixel given a square neighborhood with $R = 4$ and $r = 2$.

Adaptations of the RX algorithm include exchanging the covariance matrix for the correlation matrix [5]. It is also possible to combine local parameter estimation with clustering methods [1].

Clustering Methods

A common normalization method is to divide the image pixels into clusters using methods like vector quantization and k-means [1, 3, 21]. In none of these examples do the authors explicitly state that their clusters are normal in shape, but all use a Mahalanobis distance threshold, which implies an assumption of normality. The BACON algorithm [2, 21] explicitly chooses the distance threshold using the chi-square distribution, as discussed above.

The Gaussian Mixture Model (GMM) can be considered a form of fuzzy clustering. With a GMM, it is assumed that the true distribution can be approximated by a mixture of normal (Gaussian) distribution components. The GMM for a set of data is usually found using the Expectation-Maximization (EM) algorithm [6]. The EM algorithm consists of two alternating steps. In the expectation step, the likelihood of each sample is calculated for each mixture component, divided by the sum of likelihoods for that sample. The result of this step is a set of membership values across all samples for each mixture component. Using these membership values as weights, weighted means and covariances are calculated for each mixture component. If the GMM is properly estimated, each mixture component may be

considered a fuzzy cluster. Unlike the clusters produced by k-means or vector quantization, these clusters are designed to be normally-distributed, but this is a much more costly process.

Robust Methods

Most clustering approaches, such as CBAD and GMM, simply use the sample mean and covariance matrix of each cluster, but a more robust approach to outlier detection is the BACON algorithm [2, 21]. BACON aggregates sample points within a cluster into an inlier set by gradually increasing the threshold on Mahalanobis distance, re-estimating the mean vector and covariance matrix at each step. This iterative estimation is more robust to outliers, thus ensuring that the outliers can be correctly identified. It is more costly than simply calculating the sample mean and covariance of the entire image, since it requires multiple iterations with sample sizes approaching the full set.

2.2.3 Summary

While many possible approaches exist, the basic structure of a spectral anomaly detector is generally the same: (1) Estimate the parameters for one or more normal distributions from which the pixels can be drawn, and (2) compare pixels to this model, flagging statistical outliers as anomalies. For a given application of spectral anomaly detection, a balance must be found between the accuracy of the model and the speed at which it needs to be estimated.

2.3 Related Detection Methods

A field closely related to WiSAR is analysis of surveillance video. Within this domain, solutions have been proposed for background subtraction [12, 23], human detection [14, 15, 25], and behavioral anomaly detection [11]. These methods are discussed here for comparison and to discuss the difficulties in applying them to this domain.

2.3.1 Background Subtraction

Work on background subtraction includes modeling of dynamic backgrounds and creating backgrounds for stationary pan/tilt cameras. In a surveillance system, background subtraction generally consists of modeling the background over time on a pixel-by-pixel basis and identifying those pixels that don't fit their respective background models. This is sometimes called motion detection. The sophistication of a pixel's model varies from a simple average to mixture models [23]. For cameras with pan and tilt motion, a panorama background can be constructed for use in place of a single background image [12]. Since a WiSAR search camera is in constant motion and almost never views the same place twice or for more than a few video frames at a time, a localized model of the scene is of little help. Using a single model for all pixels in the image would be in the realm of spectral anomaly detection (Section 2.2).

2.3.2 Object Detection

All object-detection methods require some prior model of the class of objects being detected. In most cases, the object model is formed from a large base of offline training examples [14, 15, 22, 25]. In other cases, the model may be developed from domain knowledge [4, 8]. Because such methods are derived from specific qualities of the object in question, they are often difficult or impossible to generalize to new applications. For example, the method in [4] is only useful for finding large man-made objects in natural scenes because it relies on the way texture should be distributed in large natural regions versus large man-made regions.

2.3.3 Human Detection

A specific class of object detection that has been the subject of a lot of research is human detection. As promising as human detection sounds, it has drawbacks that make it an unlikely candidate for use in wilderness search and rescue. Firstly, the missing person is not the only thing searchers will be looking for. Identification of objects left behind by the

search subject, such as a jacket or bag, can help guide the searchers in the right direction. Detection of humans alone is therefore insufficient for wilderness search and rescue.

Perhaps more problematic of human detection is that previous work has almost exclusively focused on detection of humans that are moving and/or upright relative to the camera [14, 15, 25]. For this reason, a more accurate name for these systems could be pedestrian detection, though they are not always identified as such. Because there is limited variation in overall orientation and shape, detection of pedestrians such can be very effective with offline training and template matching.

In a search scenario, the position and orientation of the search subject can vary widely and is generally unknown. Also unknown is the background class, since background shapes and colors vary from scene to scene. Generalizing these approaches to detect humans in all positions and orientations and against all natural backgrounds would require a wider base of training examples and probably produce more false positives. It would likely be very difficult to gather sufficient training data for such a system. To construct a template, one would need prior knowledge of how the target objects differ from the scene. These problems make adaptation of existing human detection systems for search and rescue impractical.

2.3.4 Behavioral Anomaly Detection

Behavioral anomaly detection [11] involves the characterization of motion in surveillance video and should not be confused with spectral anomaly detection. Such systems presuppose the detection of humans and/or moving objects. Behavior modeling would be of little use in wilderness search and rescue scenario after the person has been detected.

2.4 Summary

Video enhancement methods can be helpful for WiSAR and form a good foundation for this work. Changing the focus from enhancement to detection can have be a greater benefit to the search task.

Many tools exist in the realm of video surveillance, but these tools are difficult to adapt to WiSAR because certain general assumptions fail to apply. These differences include the content of the scene as well as the position and movement of the camera. Because the domain is sufficiently different from WiSAR, other methods must be explored.

Much more promising than adapting surveillance tools is adaptation of hyperspectral anomaly detection. Multiple methods exist for hyperspectral anomaly detection, any of which could be adapted to the relatively low-dimensional data of color images. A good spectral anomaly detector is key to creating an unusual-object detector for WiSAR.

Chapter 3

A Detector for Unusual Objects

The unusual-object detection system implemented in this work uses a multi-step process to find man-made targets in aerial video of natural scenes. The first step is to process each video frame with a spectral anomaly detector, producing a binary map of anomalous pixels. In the next step, these pixels are aggregated into objects with spatial and temporal extent using connected component labeling and alignment between frames. Finally, various information about each unusual object is computed and used to filter out the most likely candidate objects.

This chapter first discusses the implementation of multiple candidate spectral anomaly detectors. Next, it discusses the collection of image data used to evaluate the candidate detectors and the results of the evaluation. Finally, it explains the methods used for pixel aggregation and object filtering.

3.1 Anomaly Detection Implementation

The first step in our object detection algorithm is to detect spectral anomalies. The following four candidate methods were implemented for comparison:

1. The RX algorithm [19]
2. Vector quantization (as used in CBAD [3])
3. K-means clustering
4. The EM algorithm [6], initialized using k-means

The BACON algorithm [2] for robust outlier nomination was also implemented to see if it could improve the results of the best spectral detector. All of these methods are described in Section 2.2.

3.2 Data Collection

In order to evaluate the different detectors, a set of test images was collected. These images are of natural scenes containing a few foreign man-made objects. A ground-truth labeling of the objects within each image was created for fast and repeatable testing.

3.2.1 Photography

To best control the content of the images, the scenes were set up carefully and deliberately. Two natural scenes were used: a grassy location and a desert location. These two locations were carefully chosen to minimize the likelihood of man-made objects in the scene. A small number of man-made objects were then placed at each location for use as visual targets. These targets ranged in size from a t-shirt to a small blanket. Each target consisted of one or two solid-colored objects. Six targets were placed in the first scene and five targets were placed in the second scene. Thus, each scene contained mostly naturally-occurring objects, with only a few foreign man-made objects.

A professional aerial photographer captured aerial imagery of each scene using specialized equipment. The photographer mounted a digital camera on a small, remote-controlled plane. The photographer then flew the camera over the scene, capturing both high-resolution still images and standard-resolution digital video of the targets.

3.2.2 Image Labeling

The video and images of each scene were reviewed carefully by visual inspection, and with the aid of temporally-local mosaics. In addition to the objects placed in the scene as targets,

the camera captured a number of other objects that could reasonably be considered foreign to the scene:

Other foreign objects at the grassy location

- The pilot and two other people
- Two vehicles
- Multiple nearby buildings (video only)

Other foreign objects at the desert location

- The pilot and vehicle (video only)
- A plastic grocery bag (photographs only)
- A white box
- A bright orange object

All known anomalies, including the accidental objects listed above, were manually labeled in the digital stills on a per-pixel basis. This resulted in one label map image for each high-resolution photograph. These label maps were then used to optimize and compare the different spectral anomaly detection methods.

3.3 Spectral Detector Evaluation

An automated test suite was built for fast and repeatable evaluation of the different spectral anomaly detection methods. The test suite calculates a Receiver Operating Characteristic (ROC) curve for each anomaly detection method by varying the detector's threshold and plotting the true positive rate (TPR) against the false positive rate (FPR). The comparison metric for the different methods is the area under the ROC curve.

At least one method (the RX algorithm) is sensitive to the size of objects in the image. Therefore, the full-size stills as well as the corresponding label images were subsampled to get

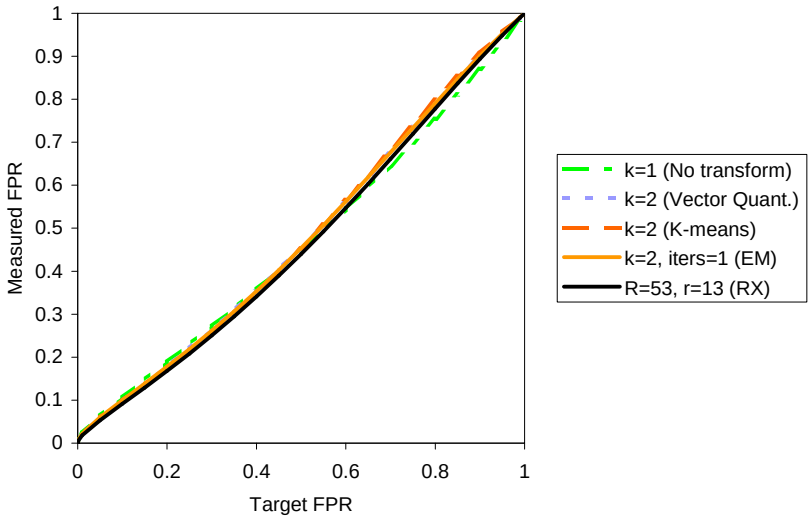

Figure 3.1: Target and Measured FPR for each method

object sizes similar to those seen in the video but not so small as to hinder visual detection. Decimating the still images by a factor of two gives targets of about the right size. Even after decimation, the still images are much bigger than the frames of video. Dividing each decimated image into six regions gives images reasonably close in dimensions (544×612 and 512×576) to that of the video (640×480). Thus, each of the 278 still images produced 24 video-frame-sized subimages, for the total equivalent of about 3.7 minutes of manually-labeled high-quality video.

The anomaly map returned by a given configuration was compared pixel-by-pixel to the label images created during data collection to calculate the TPR and FPR. The TPR is the percentage of labeled pixels correctly identified as anomalous, while the FPR is the percentage of non-labeled pixels incorrectly identified as anomalous.

In order to cover as near as possible the full range of false positive values, a target false positive rate was varied from 0% to 100%. This target value was used, in connection with Equation 2.2, to determine the Mahalanobis distance threshold for each test. Each ROC curve is comprised of 40 such tests. The measured FPR matched closely with the target FPR for all four spectral detection methods (Figure 3.1).

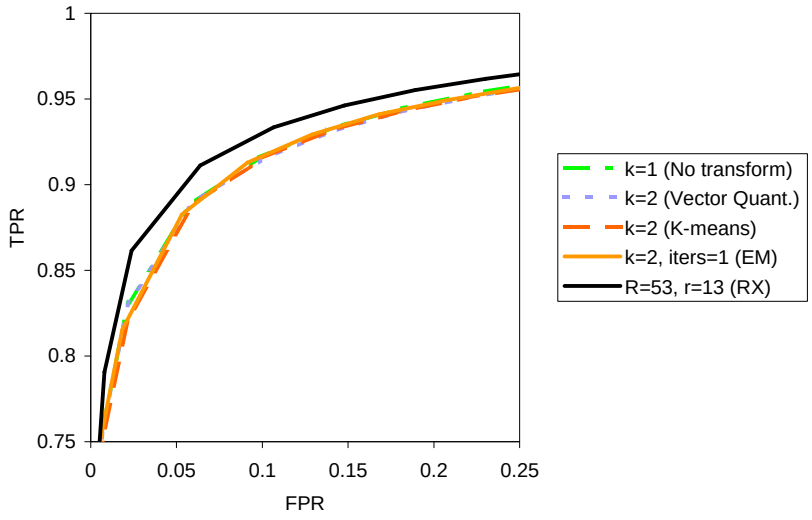

Figure 3.2: ROC curves for each method

3.3.1 Evaluation Results

The spectral detector that performed the best overall was the RX algorithm with $R = 53$ and $r = 13$ (Figure 3.2). While there were significantly worse settings for RX (Figure 3.3), comparable results could be found in a fairly broad range of the parameter space (Table 3.1).

The second best detector was the degenerate clustering case of $k = 1$. This case is the same for all clustering methods as it performs no clustering or normalization of the data. Comparable results were found for each clustering method with $k = 2$ (Figure 3.2), but larger values of k showed a decrease in performance (see Tables 3.2, 3.3, and 3.4; and Figures 3.4, 3.5, and 3.6). The reason for this is not entirely clear. To explore this apparent failure of clustering methods, and to rule out the possibility of implementation error, a few detectors were compared in more detail.

Further Analysis of Clustering

In order to dissect the failing points of clustering, RX and the degenerate case were compared to vector quantization and k-means on each of a smaller set of only 742 subimages. These were the only subimages that contained man-made objects, and could thus be used

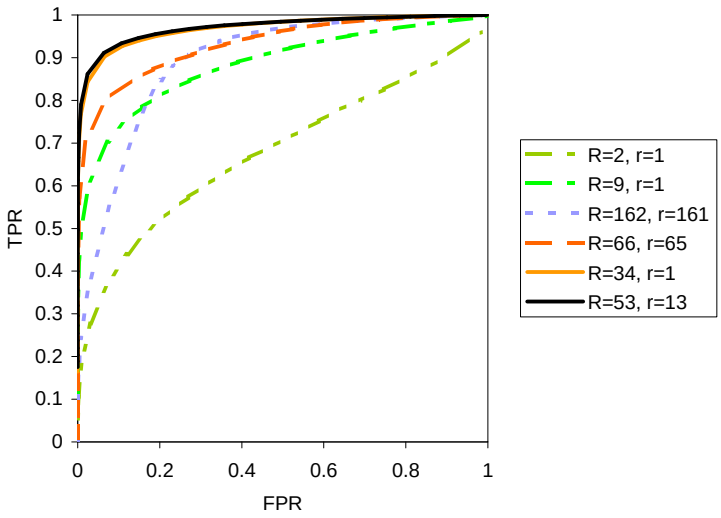

Figure 3.3: Results for RX with various parameters

R	r	ROC curve area
2	1	66.355%
3	1	72.289%
15	1	93.919%
34	33	95.242%
50	11	96.921%
52	27	96.899%
53	13	96.933%
53	15	96.929%
55	15	96.926%
62	21	96.910%
66	17	96.916%
66	33	96.854%
82	1	96.889%
130	65	96.700%
162	161	88.433%
194	129	96.473%

Table 3.1: Results for RX with various parameters

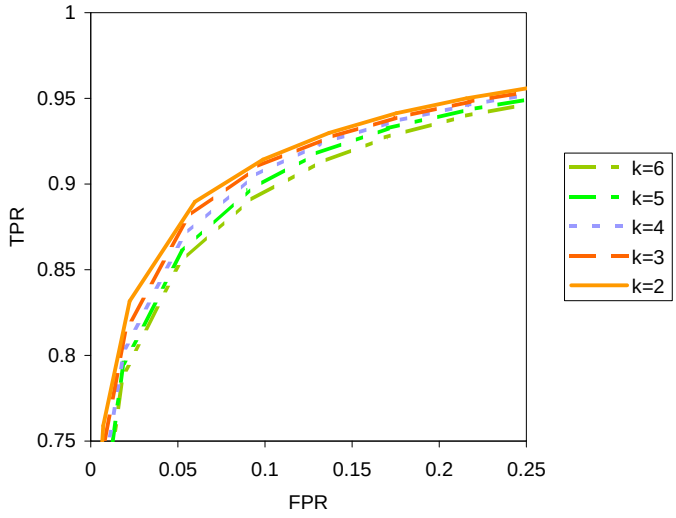

Figure 3.4: Results for vector quantization with different values of k

k	ROC curve area
6	95.564%
5	95.789%
4	96.048%
3	96.21%
2	96.395%
1	96.464%

Table 3.2: Results for vector quantization with different values of k

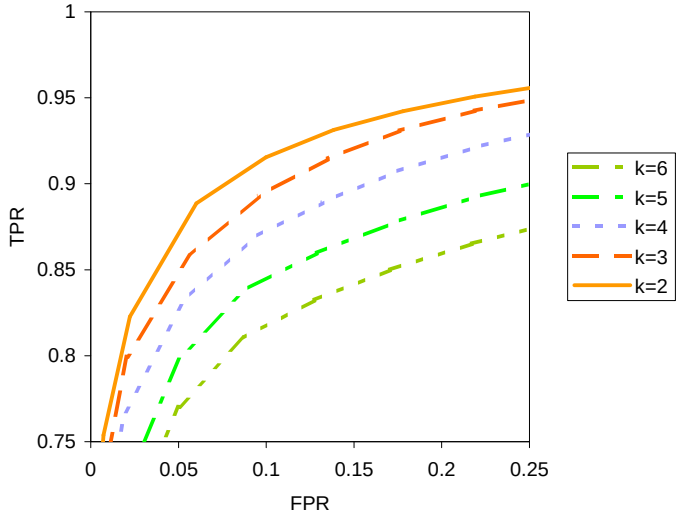

Figure 3.5: Results for K-means with different values of k

k	ROC curve area
6	90.902%
5	92.491%
4	94.544%
3	95.712%
2	96.377%
1	96.464%

Table 3.3: Results for K-means with different values of k

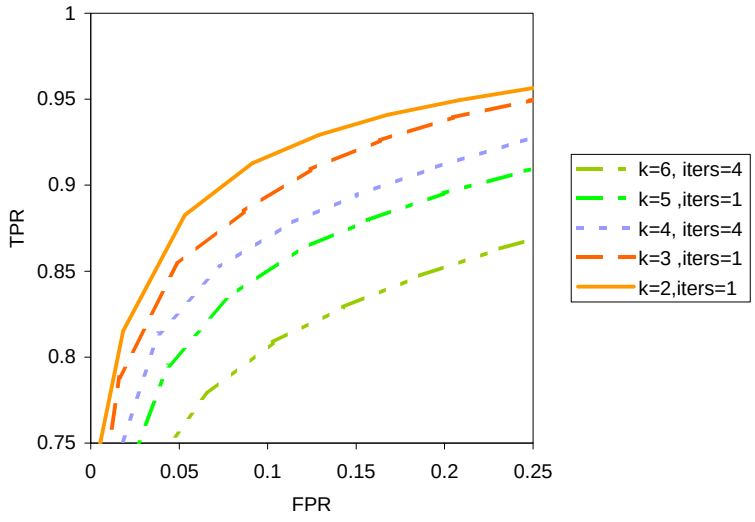

Figure 3.6: Results for EM with various parameters

k	iterations	ROC curve area
6	16	0.002%
5	16	0.037%
4	16	92.971%
3	16	94.128%
2	16	95.309%
6	4	90.612%
5	4	92.743%
4	4	94.602%
3	4	95.370%
2	4	96.406%
6	2	90.913%
5	2	93.000%
4	2	94.929%
3	2	95.643%
2	2	96.438%
6	1	91.219%
5	1	93.161%
4	1	95.103%
3	1	95.770%
2	1	96.447%

Table 3.4: Results for EM with various parameters

to compute ROC curves. For any given value of k between 2 and 6, vector quantization outperformed the degenerate case on over 46% of the images. With $k = 4$, vector quantization outperformed the degenerate case 50.4% of the time. For k-means, the best value of k was 3, outperforming the degenerate case on 52.3% of the images. Clearly, clustering with $k \geq 2$ can be beneficial, but only about half of the time.

In a more specific example, the comparison was performed on a set of 64 subimages of the same target: a white t-shirt at the grassy location. K-means with $k = 3$ outperforms the degenerate case on 42 of these images, or 65.6% of the time. Even more impressively, it outperforms RX 87.5% of the time. Therefore, RX is not always the best choice either.

These results show that clustering can work very well in many cases. They also indicate that the fault is not in the implementation of the clustering algorithms. But if clustering with moderate values of k can outperform $k = 1$ on over half of the images, why does $k = 1$ do better overall? It may be that higher values of k are more likely to overfit. Or it may be that higher values of k are simply less likely to be the right choice. Either way, it is clear that any fixed value of k will perform well on some images, but this benefit will not compensate for failure on the rest of the images.

The real problem with using clustering in this domain is that it is more sensitive to the content of the scene than the size of the target [3]. The content of the scene can change frequently as the plane flies over different areas. If only one type of ground cover is present, k should be very low. For more types of ground cover, the cluster count will need to be higher to correctly model the background. The correct number of clusters to use will then change as the more or less types of ground cover are in view. In contrast to this, parameter selection for RX is mostly dependent on target size [3].

The EM Algorithm

Apart from the problems of choosing the right value of k, the EM algorithm has further complications. In all tests of EM, increasing the number of iterations decreased performance

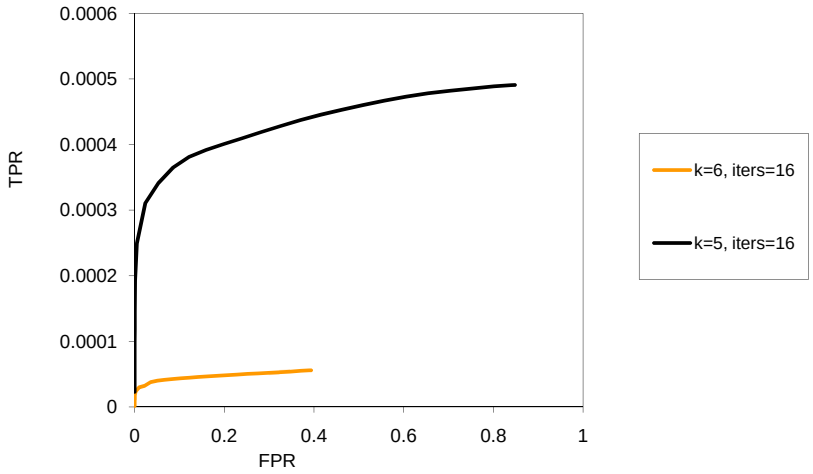

Figure 3.7: Results for EM with too many iterations

(Table 3.4). The reason for this is most likely that with enough clusters and iterations, EM begins to model the outliers as well as the background. For the two highest values of k tested, using 16 iterations resulted in almost no true positives, regardless of the false positive rate (Figure 3.7). With $k = 6$, the false positive rate did not even approach 100%, even with the highest target FPR. From this it is clear that great care should be taken in choosing the parameters for EM to prevent over-fitting the data.

3.3.2 BACON

The best performing normalization method, RX, was combined with a robust outlier detection method, BACON, to try to improve performance. The ROC curves with and without BACON are very similar (Figure 3.8). The ROC curve area with BACON (97.17%) was slightly higher than with RX alone (96.93%). But this increase is less significant than the one between no normalization (96.46%) and RX.

One downside of using BACON is that it is more difficult to control the false positive rate, and therefore the true positive rate. Although the relationship between measured TPR and measured FPR is comparable to RX alone (Figure 3.8), the measured FPR is often much

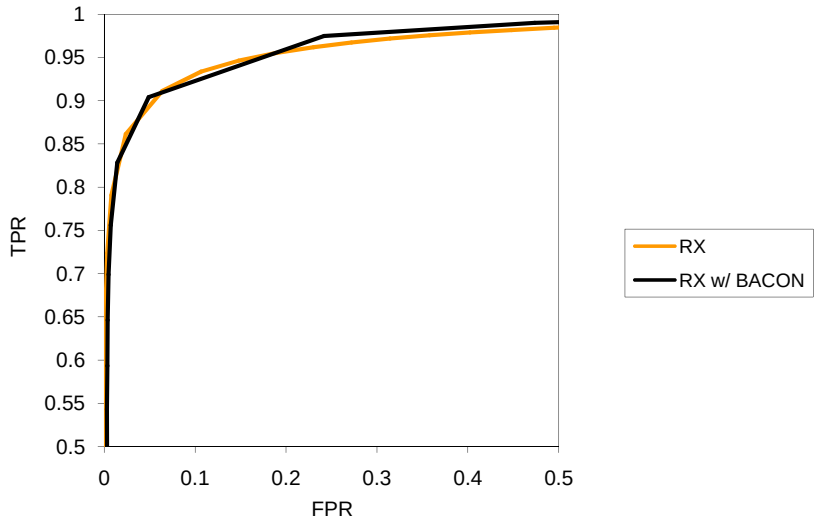

Figure 3.8: Results for RX with and without BACON

higher than the target FPR when BACON is used (Figure 3.9). This is most likely due to BACON's robust nature.

Both methods shown in Figure 3.9 use the target FPR to determine a Mahalanobis distance threshold, but they use them differently. Where the simple method uses this threshold to *exclude* outlying points, BACON uses this threshold to iteratively *include* more points in the set of inliers. Once no additional points fall within the threshold, BACON terminates. This termination can happen at any iteration, regardless of the number of outliers. This means that the number of outliers is not only less predictable, but it has the potential to be much larger than using the simple method.

Another major downside is that BACON's iterative nature is inherently more computationally expensive. Instead of computing the covariance matrix and Mahalanobis distances once, they must be recomputed for each iteration.

Although using BACON produced a slightly better ROC curve, in terms of area, it was also much slower and harder to control than RX alone. Therefore, BACON was not used for detection in this work.

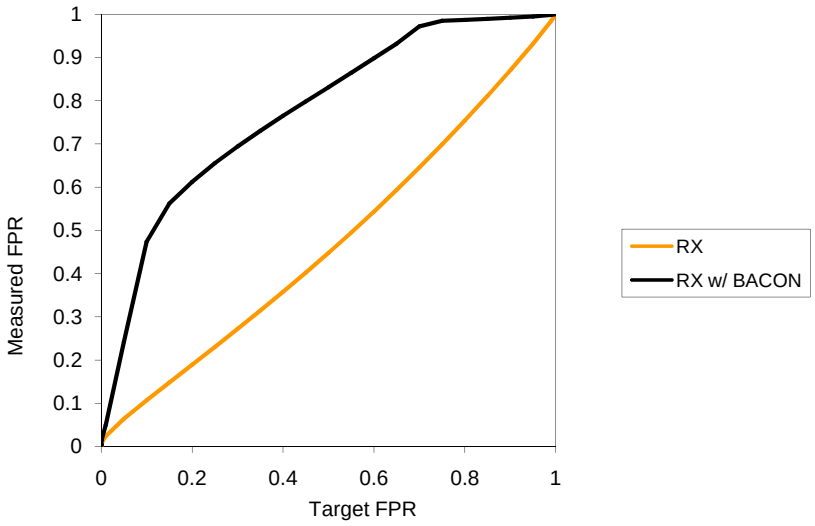

Figure 3.9: Target and measured FPR for BACON

3.3.3 Summary of Results

Clearly, clustering can work well on many images and produce decent results in general, but RX generalizes better in this domain. The reason for this is the difference in parameters. The best value for the clustering parameter k is determined by the content of the scene, specifically, how many color classes are normal to the scene [21]. It does not seem reasonable for one value of k to work best for many different scenes.

In contrast, the best window size for RX is primarily determined by the sizes of the targets and other objects [3]. In an aerial search, the ground resolution would be controlled to keep targets large enough for detection, while maximizing ground coverage [9]. Thus, the target size should easily fall within a predictable range. Since target size is less variable and easier to predict in this domain than scene content, RX should be preferable to a clustering approach.

3.4 Object-based Detection

All of the spectral anomaly detection methods considered in this work perform pixel-wise detection. This means that the best result each of these can give is a binary mask corre-

sponding to the image that shows which pixels are anomalous. A binary anomaly mask is sufficient for video enhancement techniques [17, 18], since an overlay image or filter is all that is required. With a little more work, this information can be summarized for a more clean and consistent user interface.

3.4.1 Problems with Pixel-based Detection

One problem specific to this domain is that one can only expect to detect targets that are significantly larger than a single pixel. The first reason for this is that the objects must be large enough to appear significant to the human reviewer. Even if the detector found a target of single-pixel or sub-pixel extent, it would most likely be dismissed by the human analyst.

Secondly, detection of such small targets using only RGB values is very difficult. Color images contain much less information per pixel than do hyperspectral images. Not only is the number of spectral bands very low, but the color bands significantly overlap each other. This means that separation of the target from the background on a pixel-by-pixel basis is much more difficult for color images. Separation is even more difficult for samples that are a mixture of target and background, such as pixels on the border of a target or targets smaller than a single pixel. Spectral mixing also occurs through digital-to-analog conversion (NTSC) or block compression (JPEG, ATSC) which may be required for transmission or storage of image data prior to processing.

Fortunately, targets are expected to be significantly larger than a single pixel [9] in this domain. As noted above, the target must span a fair-sized region of pixels to be detectable by the reviewer as well as an automated detector. Because of the redundancy inherent in video, objects are also expected to appear in multiple consecutive frames. Leveraging this information, the anomalous pixels are aggregated into spatiotemporal objects.

3.4.2 Aggregation

Anomalous pixels are aggregated spatially by finding connected components in the anomaly map. If the false-positive rate is high, morphological dilation and erosion can first be applied to the anomaly map to filter out speckles. Each connected component in the anomaly map is a unique spatial object.

Spatial objects are aggregated temporally using the frame-to-frame alignment that was computed for the temporally-local mosaic. Each object is warped from frame to frame until it either overlaps with another spatial object or falls entirely outside of view. Since an object can be warped multiple times before overlapping or falling out of view, overlapping objects need not occur in adjacent frames. Spatial objects that overlap are then combined into spatiotemporal objects.

3.4.3 Object Filtering

Information about each spatiotemporal object is gathered to filter out the best candidates. Objects that are too small are likely to be the result of noise, so objects that are not sufficiently large are discarded. The temporal extent of each object is also taken into consideration. Objects that appear too briefly or infrequently are also discarded. Good candidate objects should be detectable soon after they appear and/or shortly before moving out of view. Therefore, objects that do not occur near enough to the edge of the view are likewise discarded. Other information can be calculated as well, such as average color or ground location. Each object that meets the filter criteria is saved in a final list of unusual objects. In this work, the filter criteria were chosen empirically.

3.5 Summary

Unusual-object detection starts with spectral anomaly detection. Of the methods explored here, the RX algorithm adapts best to this domain. While statistically-robust methods like BACON may be useful in some cases, it was not worth the added computational cost in this

application. Using the redundancy of video, spectral anomalies can then be aggregated into spatiotemporal objects and information gathered about them. The result of this process is an information-rich list of objects. One application of the unusual object list is explored in Chapter 4.

Chapter 4

User Study

The best way to evaluate the unusual-object detector is to consider its impact on the search task. While a ground-truth evaluation could be performed, this would not show the benefit of the detector to those performing a search. Therefore, a user study was performed where participants performed visual search in a simulated WiSAR scenario with and without the aid of the detector.

There are numerous ways that the imagery and detector results could be presented to a searcher, and the choice of presentation will certainly have an effect on the searcher's performance. In order to keep the implementation and analysis tractable, one simple user interface was implemented and evaluated for this study. This chapter describes the user interface, the data used, and a statistical analysis of the collected user data.

4.1 User Interface

Each participant was asked to view a series of eight aerial video clips and mark foreign or man-made objects. Participants placed marks on the video with a single mouse-click and could remove marks with a right-click. These marks appeared as red circles (Figure D.3). Participants were given the option of pausing the video to examine or mark objects. Each time a participant marked an object in the video, the location and time were recorded. Unless a mark was removed by the participant, it was also logged in a final list of markings for that participant.

All aerial videos were presented as temporally-local mosaics [9] (Figure D.1). The presentation order was counterbalanced and the order of the videos was randomized. For each participant, four of the eight video clips were randomly selected and marked with suggestions from the detector (see Section 4.2.2 on page 34).

In addition to the primary task of target detection, participants were also given a secondary task in which they counted tones played during each video clip. Some clips contained a series of only low-pitched tones, while other clips contained tones of two different pitches. This secondary task was included in order to evaluate the cognitive load on participants. After each exercise, the participant was asked to report the number of low tones and (if present) the number of high tones played during the exercise.

These options resulted in four presentation methods: suggestions with only low tones, suggestions with both high and low tones, no suggestions with only low tones, and no suggestions with both high and low tones.

Each participant first took a brief demographic survey (Appendix A) and read a set of written instructions (Appendix B). Each participant then viewed a number of explanatory example images (Appendix D) followed by two practice video clips before beginning the exercises. In both practice clips, the participant viewed the same video sequence but with different presentation methods. The presentation method for the first practice clip was generated randomly, with the second being the complement of the first.

4.2 Data Used

Aerial video was taken at both a grassy and a desert location, with many of the same target objects being used at both locations (see Section 3.2). For each location, four one-minute video clips were selected for the user study. The number of targets visible in each clip ranges from zero to seven (Table 4.1). The complete set of eight video clips included two appearances each of 12 target objects (Figures 4.1 and 4.2) for a total of 24 targets.

(a) Tan shirt w/ shorts

(b) Blue blanket

(c) Red shirt

(d) Black shirt

(e) Green bag

Figure 4.1: Target objects at the grassy location

(a) Blue blanket

(b) Black shirt

(c) White shirt

(d) Red shirt w/ shorts

(e) Green bag

(f) Orange object

(g) White box

Figure 4.2: Target objects at the desert location

Clip Number	1	2	3	4	5	6	7	8
Targets	1	4	0	5	2	0	7	5
Suggestions	53	24	37	24	7	4	8	15
Scene	A	A	A	A	B	B	B	B

Table 4.1: Information about the content of each video clip (A is grass, B is desert)

4.2.1 Targets

Each of the 12 target objects consists of one or two man-made objects. Each man-made object is on the order of a person in size and consists of a single color: red, blue, green, orange, tan, black, or white. Some of these physical objects were used for multiple target objects. For example, the same blue blanket was laid out on the ground in the grassy location (Figure 4.1b) and draped over sage brush at the desert location (Figure 4.2a). The same red shirt is laid out by itself in the grass (Figure 4.1c) but paired with the tan shorts in the desert (Figure 4.1c).

In addition to the intentionally-placed targets, several other man-made objects were also discovered at each location. For better control over user-study results, the video clips were carefully chosen to exclude most of these objects. Two unintentional objects were included as targets because they were difficult to exclude, were of the right size, and consisted of solid colors. These were a white box and an unidentified orange object (Figures 4.2g and 4.2f). Because of its proximity to the UAV pilot, a white t-shirt that was intentionally placed at the grassy location had to be excluded from the user study video clips.

Because each target object appears twice and several physical components are reused between targets, a training effect is possible where participants are more likely to detect an object they have already seen. In a pilot version of the study, some participants were asked if they noticed any repeated objects and all of them answered in the negative, suggesting that different appearances of the same object were sufficiently unique. Also, the pseudo-random ordering of the exercises (see Section 4.2.3) should prevent a strong bias from any training effects.

Figure 4.3: Suggestions as blue circles

4.2.2 Suggestions

The purpose of the user study was to evaluate the usefulness of suggestions from the detector described in Chapter 3. Four out of every eight video clips included target suggestions. The other four clips served as a control group with no suggestions. Suggestions were presented as light blue circles (Figure 4.3) with one circle for each anomaly found by the detector. The size and location of each circle was made to encompass a region twice the size of the anomaly's bounding box.

For the pixel-wise detection step, an RX detector was used. The inner radius for the RX convolution kernel should be large enough to exclude most of the target object, while the outer radius needs to be just large enough to accurately sample the surrounding region. While the optimal settings discovered in ground-truth evaluation were 13 and 53, respectively, the inner radius was increased here to 26 to produce better results with some of the larger targets. The false-positive rate was set to 1 in 10 million pixels by trial and error.

The number of unusual objects found was somewhat different between the two locations. The video of the grassy location originally produced many more single-pixel targets than the desert location, significantly slowing down object aggregation. The optional morphological operator was therefore applied to remove these objects earlier. A possible downside of this operator could be a decrease in detection rate, since it could also remove or decrease the size of correctly-detected objects. This does mean that a slightly different process was applied to the two scenes, but this was a direct result of the difference in content.

In practice, the technician could switch this operator on or off, to find which one produces a more reasonable number of false positives.

The best parameters for the object filtering step were each determined empirically using the ground-truth and the object lists from the aggregation step. Objects were ignored if they appeared for less than 3 frames or in less than 91% of their known temporal extent. Objects of interest were also restricted to those that had contained at least 43 anomalous pixels in at least one frame and touched the border of their first or last frame. Of the 24 target objects, 11 overlapped with suggestions for a 45.8% true-positive rate. The filter settings were chosen by varying each parameter, observing the resulting object list in the user study interface, and subjectively choosing a good trade-off point between the number of true positives and the number of false positives. Less restrictive settings produced more true positives but also a good deal more false positives.

4.2.3 Sequence Generation

Each participant was assigned a unique sequence of eight exercises. With eight video clips and four presentation methods, there were 32 unique exercises to choose from. In order to prevent bias in the results, unique semi-random exercise sequences were generated for all participants beforehand. The sequences were generated with a greedy algorithm that guaranteed that each participant would view each of the eight video clips once and each of the four presentation methods twice. If this constraint resulted in multiple choices, exercises were then chosen to ensure equal frequency among the 32 exercises across all participants. Other constraints on the selection algorithm included preventing any presentation method, exercise, or 2-clip subsequence from occurring more frequently than the others. Whenever multiple exercises met all criteria, one was selected randomly.

After reviewing data from the first 30 participants, it was discovered that, through a setup error, the last seven participants received identical sequences. To prevent bias, the last six data logs were ignored. The analysis did not show sufficient statistical significance with

only 24 participants, so 12 more sequences were generated. Another discovery was that for some participants, the targets were not evenly distributed among the presentation methods. In an extreme case, a participant viewed zero true targets in one of the four methods. To prevent these extreme cases, an additional constraint was added to the sequence generation algorithm to try to evenly distribute targets among the four presentation methods for each participant. This additional constraint was only applied to the last 12 sequences.

4.3 Results and Statistical Analysis

The final set of results consisted of data logs from 35 users. One of the scheduled participants did not come, the six participant logs with identical sequences were ignored, and one exercise was accidentally skipped by a participant. Seven clips were shown 35 times, with one clip only shown 34 times. Three presentation methods were shown 70 times, with one shown 69 times. Nine of the exercises were shown 8 times, with 23 shown 9 times.

Ground truth markings were created by hand using the user study interface. One ground truth marking was made for each of the 24 target objects. Once all of the participant data had been gathered, all user markings, suggestions, and ground truth markings were grouped into clusters. Each marking was put in the same cluster as any other markings whose centers lay within its radius, resulting in a total of 535 clusters. Each of the 24 ground truth markings belongs to a unique cluster. Of these ground truth clusters, 11 include one or more suggestions and 21 include one or more participant markings. Out of all 535 clusters, only 132 included one or more suggestion markings. Markings removed by the participant were included in the clustering step but ignored in all other considerations.

Four performance measures were calculated for each participant for each of the four presentation methods. The first three performance measures related to the primary task of detection. These are the true positive rate (TPR), the false positive rate (FPR), and the positive predictive value (PPV). The true positive rate is the percentage of ground truth clusters that a participant clicked on. The false positive rate is the percentage of non-ground

truth clusters that a participant clicked on. The positive predictive value is the percentage of a participant's clusters that included ground truth objects, equivalently known as the *precision*. The fourth performance measure related to the secondary task and was the mean squared error (MSE) of the tone count reported by the participant. The following sections compare the four presentation methods using all four performance measures.

4.3.1 True Positives

Perhaps the most significant measure of performance for participants was the true positive rate. The presence of high tones did not produce a significant effect on the TPR, but the presence of suggestions did. Without the suggestions, the average true positive rate was estimated at 52.57%. With suggestions, the TPR increased to 61.14%. This means that a reviewer aided by the suggestions finds 6 targets for every 5 found by an unaided reviewer. This improvement is statistically significant ($p = 0.0229$).

4.3.2 Tone Counting Error

The participant's cognitive load is indicated by the log of the mean squared error in reported tone counts. No significant difference in log MSE was found between exercises with low tones and those with both high and low tones. However, the log MSE increases from 0.6173 to 1.1788 when suggestions are added. This difference was statistically significant ($p = 0.0072$). The interaction between the presence of suggestions and high tones also appeared to be statistically significant ($p = 0.0202$), but further analysis indicated otherwise, as shown below.

Correction for Input Errors

In reviewing the tone count data, two types of input errors became apparent. The first type of error was what one participant referred to as "fat-finger" errors: pressing two adjacent number keys when only one was intended. The second type of error was when a participant

skipped a tone-count question. This would cause the reported value to be logged as zero. These behaviors were reported by only one participant each, but the data suggests that they both occurred multiple times. Out of all the cases where a zero was reported, the lowest corresponding true value was seven. Since it does not seem reasonable for a user to have not noticed any of seven or more tones, all such entries were considered input errors.

The log MSE was recalculated ignoring all zero entries and two-digit entries where one of the two digits was closer to the correct value. After rerunning the analysis without the input errors, the interaction between suggestions and high tones no longer appeared to have a statistically significant effect on the log MSE ($p = 0.2698$). Without input errors, the estimated log MSE values were 0.5269 and 0.8843 without and with suggestions, respectively. While this is a slightly smaller difference than was estimated before, the result is still statistically significant ($p = 0.0047$). Therefore, the only statistically significant effect on the secondary was that the cognitive load was higher when suggestions were present.

4.3.3 False Positives and Positive Predictive Value

An analysis was run on the log of the FPR and on the PPV, but no difference was found between the presentation methods for either of these measures. Of the four presentation methods, the lowest estimated FPR was 2.44% for suggestions and only low tones. The highest was an FPR of 2.88% for no suggestions with both high and low tones. Estimates for PPV ranged from 49.88% for suggestions and only low tones to 50.84% for no suggestions with both high and low tones. None of the differences in FPR or PPV were statistically or even practically significant. These results indicate that the suggestions did not cause a significant increase in false detection.

4.3.4 Distraction Effects on True Positives

If the detector is functioning as designed, it should detect many of the same sorts of objects that stand out to a human viewer. Even with this overlap, the suggestions can still make

visual detection easier, and the results in Section 4.3.1 support this interpretation. However, non-target objects marked by the detector could distract the searcher, potentially decreasing the detection rate for targets that the detector missed.

As discussed in Section 4.3.2, suggestions reduced accuracy on the secondary task, but the analysis showed a positive correlation between suggestions and the TPR. On this subject, it is interesting to note that 60% of participants reported in the follow-up questionnaire (Appendix C) that the presence of suggestions made the primary task "easier" or "much easier". Since suggestions simultaneously make the primary task easier and the secondary task harder, this may indicate a shift in focus rather than an increase in difficulty. A better analysis can be obtained by recalculating the TPR, distinguishing between which true targets were and were not marked as suggestions.

To examine this, the mean TPR was calculated across participants using only the 13 targets that the detector missed. Without any suggestions, participants averaged a TPR of 37.2% on those 13 targets. With suggestions present (but not on the targets in question) the average TPR was 38.0%, which suggests identical performance.

The average participant TPR for the 11 suggested targets was 68.1% without the suggestion markings. This indicates that the detector is finding objects that stand out on their own. With the suggestions, the TPR rose to 87.9% for the same 11 targets, showing a much more drastic increase than was found when all 24 targets were considered (see Section 4.3.1). This helps support the earlier conclusion that suggestions do, in fact, aid detection.

Unfortunately, no attempt was made to calculate statistical significance for these estimates, but the results are still notable. While the targets found by the detector were already fairly easy to detect visually, marking them in the video still increased the detection rate. Furthermore, there is no evidence that the false markings distracted participants from the unmarked targets.

4.4 Summary

This user study shows the effect on the search task of a fairly simple application of automated detection results. Using the unusual-object detector, markings were added to aerial video as suggestions to aid participants in finding objects relevant to a simulated search. Several randomized trials were performed by participants and the results recorded. Performance on the first task was primarily measured by true- and false-positive rates, while cognitive load was measured by error on a secondary task.

As presented, the suggestions produced by the detector had a significant positive effect on the ability of video analysts to find objects of interest. The presence of suggestions did not increase the participants' false positives or false negatives, indicating that they were able to quickly and effectively distinguish between good and bad suggestions, without hindering detection of unmarked targets. The only notable negative effect is a higher cognitive load on the video analyst, resulting in lower performance on secondary tasks. This effect should be taken into consideration when determining an analyst's workload.

Chapter 5

Conclusion

Wilderness search and rescue is an important topic to many people, not least being those that get lost in wilderness areas each year. Improvement to WiSAR technology has the potential to save lives.

Advances in technology like the UAV-based platform being developed by the BYU WiSAR research team can help WiSAR teams gather data about a search area. The data captured by an aerial camera can be very beneficial for search, but not if items of interest go unnoticed. This work improves on the existing framework by detecting unusual objects to help searchers find relevant items.

This work seeks to improve detection of relevant objects in the video. Before an item can be identified as relevant to the search, it must first be detected. An automated detection system should relieve some of the burden on those performing this task.

Without a perfect detection system, the search team will almost certainly want all image data reviewed by a trained technician. Even for trained technicians, however, object detection is a difficult task. Rather than work in competition with the searchers, the purpose of the detector implemented here is to make video review easier for the searchers.

5.1 The Solution

The most crucial part of the unusual-object detection system is the spectral anomaly detector. A key difference between objects of interest and their surroundings is often their color. By finding which pixels differ from their surroundings, we find those most likely to

belong to objects of interest. This is accomplished using spectral anomaly detection methods developed for hyperspectral imagery.

Using carefully collected and labeled image data, four spectral anomaly detection methods were evaluated: RX, K-means, Vector Quantization, and Expectation-Maximization. All of these methods did well at distinguishing which pixels did and did not belong to naturally-occurring objects. The best performing method, as measured by area under the ROC curve, was then used in the unusual-object detector.

Of the spectral anomaly detection methods explored here, the RX algorithm is clearly the most promising. The RX algorithm performed better than all three clustering methods. While clustering may be more theoretically sound, it requires the selection of the optimal number of clusters. While this can be devised by visual inspection for a single image, choosing one cluster count for a diverse set of images consistently performed worse than no clustering at all. In contrast, the parameters to the RX algorithm are based on the size of the targets, which is fairly easy to predict in this domain. Adding robust outlier detection to the RX algorithm did not improve the results sufficient to consider adding this costly procedure to the detector. Therefore, a simple RX detector was applied in the remainder of this work.

Once the anomalous pixels are identified, they are aggregated into spatial objects using connected-component labeling. These spatial objects are linked together into spatiotemporal objects, using frame-to-frame alignment of the video to compute overlap. The list of unusual objects is then summarized and used to enhance the searcher's video review interface.

5.2 User Study Results

In the user study, the object list was used to overlay the video display with markings to draw the searcher's attention. The results show a significant improvement in the detection rate. The data suggests, but does not positively confirm, that this increase was chiefly among those targets that the detector found, with no decrease in detection among the other true targets.

The presence of the suggestion markings also increased cognitive load on the searchers, as shown by a decrease in accuracy on the secondary task, but most participants reported that these markings made the primary task of search easier. No significant effect was found on the false positive rate.

5.3 Limitations and Future Work

In the user study performed here, the detector results were combined with a specific user interface, but no effort was made to separate the effects of these two factors. It is therefore possible that the improvement in performance is primarily due to the interface, rather than the detector. It is also quite possible that an interface or detector not evaluated here could produce a greater improvement.

5.3.1 Detector vs. Interface

False detection did not increase and it does not appear that the markings distracted the users sufficiently to decrease detection among unmarked targets. It then appears that the incorrect markings in the video had no effect except perhaps increasing cognitive load. It should then follow that the increase in performance is due primarily to the correct markings. If this is truly the case, which seems entirely reasonable, then the improvement in user performance should directly correlate with the detector's true positive rate.

It should be noted, however, that the false positive rate should still be kept relatively low. No predictions can be made from the data about the effect of a change in the detector's false positive rate, but one would expect a decrease in FPR only to have no ill effects, while a sufficiently high FPR would eventually become a burden on the searcher.

It does not then appear that the interface itself is the primary cause of the improvement in user performance. If a detector were found with a higher TPR for the same FPR, it should improve performance, while using a lower TPR setting for this detector would likely be less helpful to searchers.

5.3.2 Other Detectors

In this study, the RX algorithm was found to outperform all clustering methods, but many other possibilities exist for anomaly detection.

It may be worthwhile to compare RX to unsupervised clustering methods that attempt to find the optimal number of clusters. To save on computation, it would probably be best to perform this step infrequently—perhaps when a significant change in scene content is detected or when the number of anomalies gets unreasonably high or low.

Another possibility not explored here would be to use the binary mask used for color enhancement [18]. Unlike the spectral anomaly methods produced here, this method uses a histogram-based approach to find unusual colors. By varying the histogram threshold, it should be possible to compute an ROC curve for this method to compare against other anomaly detectors.

It should be noted that the object aggregation step used here is rather ad hoc, although the results do not appear to be too dependent on the parameter settings. More sophisticated methods of summarizing anomaly detection results could be explored and compared.

5.3.3 Other Interfaces

More sophisticated or even more simple methods could be devised to utilize the results of the unusual-object detector.

One possibility would be to prioritize the video by the number of unusual objects detected or by how much they are expected to stand out. For example, unusual objects that have a short temporal extent may be given higher priority over long ones, since they are less likely to be detected without assistance. These priority levels could be used to change the order or speed of the video. By giving unusual objects more screen time or reviewing them first, the searcher could focus the video review on the best candidate objects.

Other possibilities include switching the display focus to the actual list of unusual objects. The object list could act as an index to the video, with information like color and location listed next to each item. The searcher could then access the video out of order, giving first attention to those objects found by the detector. For this method to work, the detector's false detection rate would need to be sufficiently low or the list would get inordinately long. It would also require some way to review the remainder of the video to find objects that the detector missed.

5.4 Summary

The unusual-object detector presented here would be an excellent tool for wilderness search and rescue using aerial video cameras. The preliminary analysis shows that anomaly detection performs well at detecting portions of man-made objects in color photos of natural scenes. More importantly, the user shows that suggestions produced by the detector improve detection rates by users performing visual search. By assisting searchers in finding unusual objects in the video, they will be less likely to miss signs of missing persons, thereby increasing the likelihood of a successful search.

Appendix A

Pre-study Questionnaire

Please check only one choice per question. Your answers to these questions do not affect your eligibility for the study.

1. Please mark your age group
   - ☐ Under 18
   - ☐ 18-23
   - ☐ 24-30
   - ☐ 31-40
   - ☐ 41-50
   - ☐ Over 50

2. Please mark your gender
   - ☐ Male
   - ☐ Female

3. Do you have any physical limitations that may possibly affect your performance in this user study (i.e. color-blindness, vision impairment, hearing impairment, impaired motor skills, etc.)?
   - ☐ No
   - ☐ Yes (explain) _____

4. How experienced do you feel that you are with using computers?
   - ☐ Expert
   - ☐ Average
   - ☐ Novice

5. How experienced do you feel that you are with wilderness search and rescue tasks?
   - ☐ Expert
   - ☐ Average
   - ☐ Novice

6. How experienced do you feel that you are with tasks involving searching for things on the ground from high up above in the air (aerial searching tasks)?
   - ☐ Expert
   - ☐ Average
   - ☐ Novice

7. How familiar are you with the research related to this study?
   - ☐ I have never heard of this research prior to this user study.
   - ☐ I have heard about the research, but I have never seen the display methods.
   - ☐ I know about the research, and I have seen the display methods before.

8. How familiar are you with others' preferences of the display methods that you will be presented with in this study?
   - ☐ I know many peoples' preferences.
   - ☐ I know a couple other peoples' preferences.
   - ☐ I know somebody else's preferences.
   - ☐ I know nobody else's preferences.

Appendix B

User Study Instructions

The Scenario (fictional)

A person has gone missing in the wilderness. We are flying a plane over the search area to look for traces of the person and the plane is transmitting video to you at a ground station. In addition to the video, the plane is communicating information about its status via a series of high- and low-pitched beeps, or tones.

What You Will See

The interface consists of an enhanced display of the video and sounds played through the speakers or headphones. You will start your session with a few example images and sounds as well as two (2) short practice video clips. The purpose of the practice clips is for you to get comfortable with the tasks before evaluating the system. Please spend as much time practicing as you feel comfortable.

Once you're done practicing, you will be presented with eight (8) enhanced video clips of about one minute each. During each clip we will need you to perform two tasks related to the information we are receiving from the plane.

To assist you in your tasks, our system will enhance the video display by one or more methods. First, the system will stitch the video together into a larger image. Second, the system may suggest objects that it thinks might be traces of the missing person. These suggestions will be marked by light blue circles and will not appear in every video clip. We will tell you before each clip whether or not it contains suggestions. When you click on the video display, a red circle will appear. In the section below, we explain when you should do this.

What You Will Do

Your primary task is to watch the video for man-made objects foreign to the scene. Such unusual objects are indications that a person has been in the area. These objects will be of various colors and could range in size from a shirt to a tent. Whenever you see a man-made or foreign object, you should select it by clicking on it with the mouse. It will thereafter be marked with a red circle. You should disregard all naturally-occurring objects, such as vegetation and rocks, as well as larger man-made elements, such as trails, fences, or permanent structures.

Remember that the light-blue circles provided by the system are merely suggestions for you to consider in your search task. If a natural object has been marked as a suggestion, you should not mark it. You need only select those objects that appear foreign to the scene, whether or not they have been marked as suggestions. Even if some objects are suggested, this does not mean that all of the foreign objects will necessarily be marked with light-blue circles. You must still look for objects in the entire scene.

You may find it useful to pause the video when inspecting and marking objects. You can pause and un-pause the video by pressing the spacebar. Remember that this is a time-sensitive search. Each exercise will be precisely one minute long and pausing the video will not extend the allotted time. When you un-pause the video, the display will very quickly catch-up with the plane's video broadcast. After one minute the exercise will conclude, even if the video is still paused, and you will not be able to view the portion of the video you missed. Because of these time-sensitive effects, pausing for long periods may cause you to miss objects of interest or become disoriented. We therefore recommend that you only pause the video when necessary, and that you un-pause the video as soon as you reasonably can.

While you are viewing the video, we also need you to perform a secondary task. This task will be to count the tones transmitted from the plane. During some exercises, only low tones will be played. During other exercises, both high and low tones will be played, and you will need to count these separately. We will let you know before each exercise whether it will contain both high and low, or only low tones. The tones are not connected to the video or the suggestions, so be sure to maintain focus on your primary task of searching while performing the secondary task.

Appendix C

User Study Follow-up Questions

Please check one choice per question.

How demanding was the primary task?
☐ Very Low ☐ Low ☐ Moderate ☐ High ☐ Very High

For the following elements of the system, how do you feel they contributed to the easiness/difficulty of the primary task?

The secondary task (counting the tones) made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

The pace or speed of the video made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

Image stitching made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

Suggestions (light-blue circles) made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

The viewing angle of the camera made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

The motion of the camera made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

The content of the scene made the primary task:
☐ Much Easier ☐ Easier ☐ No Effect ☐ Harder ☐ Much Harder

Any comments on your answers above?

Any comments on the system in general?

Appendix D

User Study On-screen Instructions

Figure D.1: Instruction slide 1

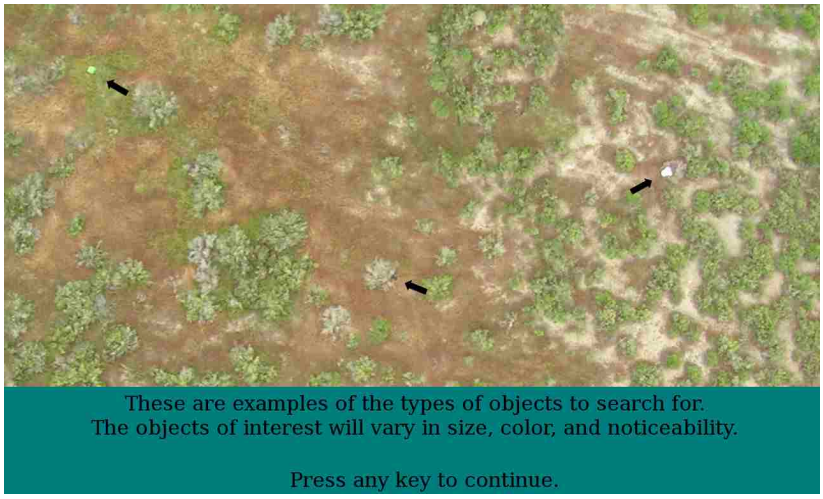

Figure D.2: Instruction slide 2

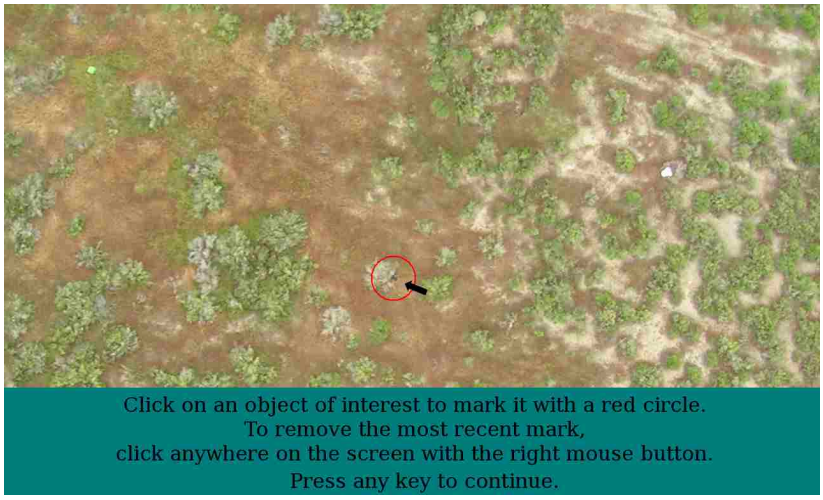

Figure D.3: Instruction slide 3

Figure D.4: Instruction slide 4

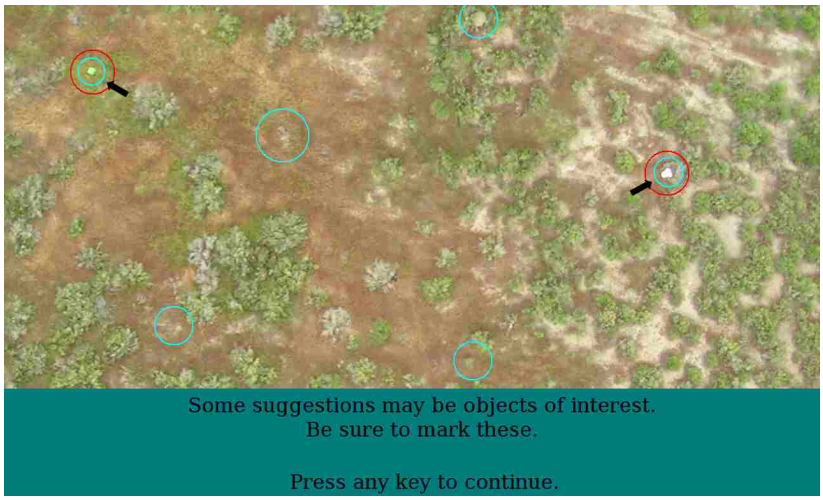

Figure D.5: Instruction slide 5

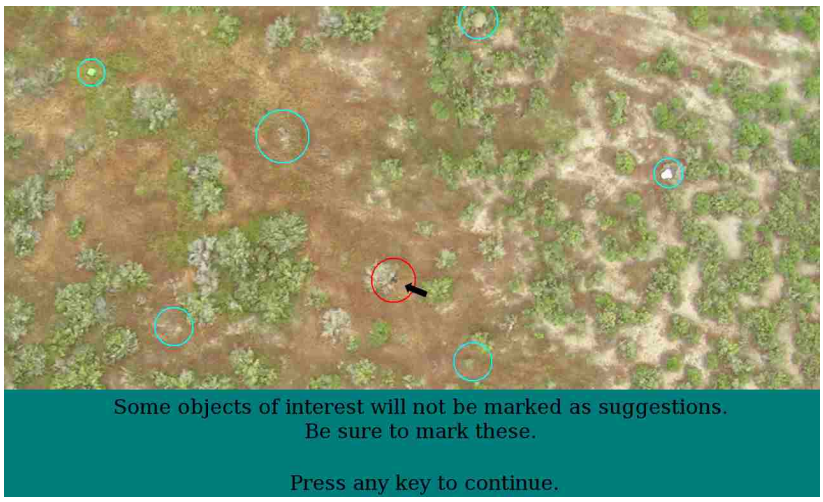

Figure D.6: Instruction slide 6

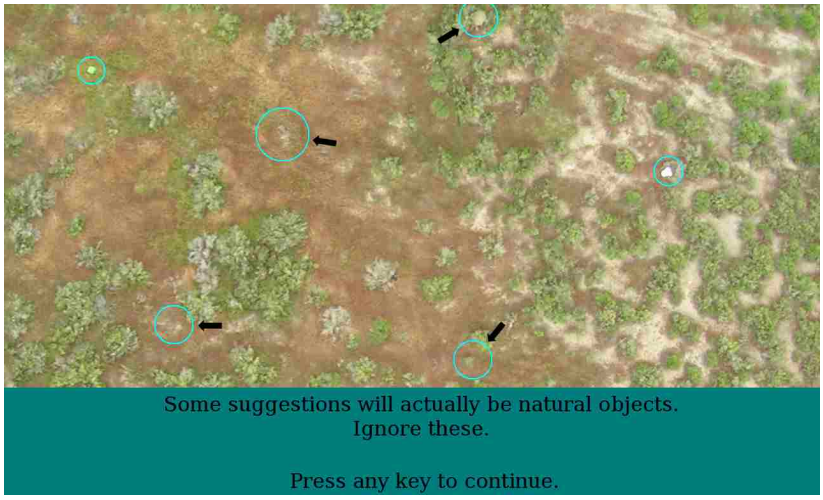

Figure D.7: Instruction slide 7

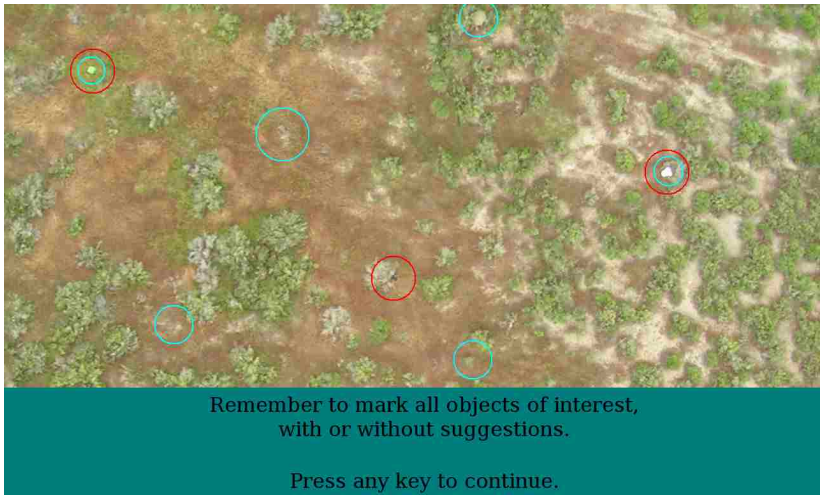

Figure D.8: Instruction slide 8

References

[1] E.A. Ashton. Detection of subpixel anomalies in multispectral infrared imagery using an adaptive Bayesian classifier. *IEEE Transactions on Geoscience and Remote Sensing*, 36(2):506–517, Mar 1998.

[2] Nedret Billor, Ali S. Hadi, and Paul F. Velleman. BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34(3): 279 – 298, 2000.

[3] M.J. Carlotto. A cluster-based approach for detecting man-made objects and changes in imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(2):374–387, Feb 2005.

[4] Y. Caron, P. Makris, and N. Vincent. A method for detecting artificial objects in natural environments. In *16th International Conference on Pattern Recognition*, volume 1, pages 600–603, 2002.

[5] Chein-I Chang and Shao-Shan Chiang. Anomaly detection and classification for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(6): 1314–1325, Jun 2002. ISSN 0196-2892.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] Ryan Dutton, David Caldwell, and Curtis R. Welborn. Anomaly detection for unmanned aerial wilderness search and rescue. In *23rd National Conference on Undergraduate Research*, La Crosse, WI, Apr 2009.

[8] Lynn M. Fletcher-Heath, Lawrence O. Hall, Dmitry B. Goldgof, and F. Reed Murtagh. Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. In *Artificial Intelligence in Medicine*, pages 43–63, 2001.

[9] Michael A. Goodrich, Bryan S. Morse, Damon Gerhardt, Joseph L. Cooper, Morgan Quigley, Julie A. Adams, and Curtis Humphrey. Supporting wilderness search and

rescue using a camera-equipped mini UAV. *Journal of Field Robotics*, 25(1-2):89–110, 2008.

[10] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Towards parameter-free data mining. In *International Conference on Knowledge Discovery and Data Mining*, pages 206–215, New York, NY, USA, 2004.

[11] A. Mecocci, M. Pannozzo, and A. Fumarola. Automatic detection of anomalous behavioural events for advanced real-time video surveillance. In *International Symposium on Computational Intelligence for Measurement Systems and Applications*, pages 187–192, Jul 2003.

[12] Anurag Mittal and Dan Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, page 2160, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[13] B.S. Morse, D. Gerhardt, C. Engh, M.A. Goodrich, N. Rasmussen, D. Thornton, and D. Eggett. Application and evaluation of spatiotemporal enhancement of live aerial video using temporally local mosaics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Jun 2008.

[14] H. Nanda and L. Davis. Probabilistic template based pedestrian detection in infrared videos. In *IEEE Intelligent Vehicle Symposium*, volume 1, pages 15–20 vol.1, Jun 2002.

[15] C.P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Sixth International Conference on Computer Vision*, pages 555–562, Jan 1998.

[16] J. Puzicha, J.M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1165–1172 vol.2, 1999.

[17] Nathan D. Rasmussen. Combined visible and infrared video for use in wilderness search and rescue. Master's thesis, Brigham Young University, Mar 2009. URL `http://contentdm.lib.byu.edu/ETD/image/etd2854.pdf`.

[18] N.D. Rasmussen, D.R. Thornton, and B.S. Morse. Enhancement of unusual color in aerial video sequences for assisting wilderness search and rescue. In *15th IEEE International Conference on Image Processing*, pages 1356–1359, Oct 2008.

[19] I.S. Reed and X. Yu. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(10):1760–1770, Oct 1990.

[20] Grand County Search and Rescue. GCSAR statistics. `http://www.gcsar.org/statistics/statistics.htm`, Jul 2010.

[21] T.E. Smetek and K.W. Bauer. Finding hyperspectral anomalies using multivariate outlier detection. In *IEEE Aerospace Conference*, pages 1–24, Mar 2007.

[22] J.L. Solka, D.J. Marchette, B.C. Wallet, V.L. Irwin, and G.W. Rogers. Identification of man-made regions in unmanned aerial vehicle imagery and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):852–857, Aug 1998.

[23] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages –252 Vol. 2, 1999.

[24] James P. Theiler and D. M. Cai. Resampling approach for anomaly detection in multispectral images. In Sylvia S. Shen and Paul E. Lewis, editors, *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX*, volume 5093, pages 230–240. SPIE, 2003. URL `http://link.aip.org/link/?PSI/5093/230/1`.

[25] L. Zhao and C.E. Thorpe. Stereo- and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 01(3):148–154, Sep 2000.