Color Anomaly Detection and Suggestion for Wilderness Search and Rescue

Bryan S. Morse
Brigham Young University
3361 TMCB
Provo, Utah 84602
morse@byu.edu

Daniel Thornton
Brigham Young University
3361 TMCB
Provo, Utah 84602
danielthornton@byu.edu

Michael A. Goodrich
Brigham Young University
3361 TMCB
Provo, Utah 84602
mike@cs.byu.edu

ABSTRACT

In wilderness search and rescue, objects not native or typical to a scene may provide clues that indicate the recent presence of the missing person. This paper presents the results of augmenting an aerial wilderness search-and-rescue system with an automated spectral anomaly detector for identifying unusually colored objects. The detector dynamically builds a model of the natural coloring in the scene and identifies outlier pixels, which are then filtered both spatially and temporally to find unusually colored objects. These objects are then highlighted in the search video as suggestions for the user, thus shifting a portion of the user's task from scanning the video to verifying the suggestions. This paper empirically evaluates multiple potential detectors then incorporates the best-performing detector into a suggestion system. User study results demonstrate that even with an imperfect detector users' detection increased significantly. Results further indicate that users' false positive rates did not increase, though performance in a secondary task did decrease. Furthermore, users subjectively reported that the use of detector-based suggestions made the overall task easier. These results suggest that such suggestion-based systems for search can increase overall searcher performance but that additional external tasks should be limited.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics—*Operator Interfaces*; I.4.9 [**Image Processing and Computer Vision**]: Applications

Keywords

Wilderness Search and Rescue, Search and Detection, Unmanned Aerial Vehicles, Anomaly Detection, User Study

1. INTRODUCTION

Wilderness search and rescue (WiSAR) is the task of finding missing persons in wilderness areas. This is highly time-sensitive, not only because of increasing danger to the search subject (missing person) but also because the search area increases over time. An efficient way to augment search is to use semi-autonomous unmanned aerial vehicles (UAVs) with cameras that transmit live video and telemetry data to a ground-based team [15]. With sufficient video resolution, searchers can identify traces of the subject from the air.

(a) Grassy Location (b) Desert Location

Figure 1: Examples of video frames from a search scenario. Targets are marked with yellow arrows: (a) blue blanket and (b) white shirt.

Even with the aid of aerial video, it can be difficult to identify signs of a search subject. Figure 1 shows examples of typical search video frames. Image resolution is limited both by the camera (constrained by the payload capacity of inexpensive UAVs) and by the need to cover as much ground as possible. In addition, the UAV and hence the video can move quickly, disorienting searchers and giving them little time to detect targets. This can be compensated for somewhat by enhancing the spatiotemporal presentation of the video [19] or visually enhancing the hue and saturation of color values [24], but detecting objects of interest remains a challenging problem. Consequently, searchers may miss signs of the search subject, even when seen by the camera.

In practice, the task of observing search video consists of two elements: *detection* and *analysis*. An object of potential interest must first attract the user's attention as worthy of inspection, then further analysis must determine whether it is indeed of interest. Objects of interest may include the missing person, abandoned clothing, camping or hiking gear, or other personal items. Such objects may also vary in difficulty of detection; a blanket (Figure 1(a)) may be easier to detect than a shirt (Figure 1(b)).

Detection is the task of quickly identifying possible signs of the search subject. It might be something as simple as seeing something that "looks out of place", especially when searching wilderness areas with primarily naturally-occurring content. When performed by a person, detection may be responded to by a reflexive action, such as a keystroke, accompanying the appearance of an unusual or significant object. The person then tries to determine, through inspection of the imagery, whether the object is likely to be a positive sign. Since detection is less dependent on domain knowledge it is a good candidate for automation, while analysis is best left to the human operator.

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.

Anomaly detection clearly has strong similarities to automated target recognition (ATR) systems, which have also been used to assist human operation of semi-autonomous UAVs (e.g., [8, 11, 12]). However, such ATR systems usually have a notion of a specific target (in our case, the missing person) rather than the more ill-posed problem of finding clues that indicate the recent presence of the search subject.

This work primarily relies on leveraging color information in video of the search area to detect signs of a missing person, a technique 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 empirically. Using the temporal and spatial information in the video stream, unusually colored pixels are aggregated into larger objects.

Because of the sensitive nature of search and rescue, and likely imperfections in automated anomaly detection, the detection task is best treated as a collaborative effort between human and autonomy. Indeed, the very question is ill-posed: find anything that "looks out of place". Therefore, the goal of this work is to use automated detection as an aid for visual search rather than a replacement for it.

The addition of automation into human tasks is not without potential issues. Research has shown that automation can affect users' behavior in various ways, including misuse of the automation [23]. In particular, the results of using automation aids for human-robot interaction can vary greatly depending on the task [16] or even across users [7]. In addition to empirical evaluation of potential detector methods, this paper also contributes the results of a user study designed to evaluate the effects of detector-based suggestions.

2. EFFECTS OF AUTOMATION

Automated detection systems have been studied for many years, including their impact on users' behavior in the presence of imperfect detection. Although such systems can increase user performance in many situations, there are numerous examples of ways in which these systems can be misused or disused [23].

Introducing automation can result in inattention either through a false sense of *complacency* or through *automation bias*, in which the saliency and authoritative nature of the recommendations from the automation affect user behavior and compliance [22]. Because anomaly detectors will inevitably miss potential targets of interest this can result in *errors of omission*, in which users similarly miss potential targets by focusing primarily on the automated suggestions. Because the detector will also inevitably make make incorrect suggestions, false alarms can serve both as a source of distraction and as a source of *commission errors* (more false alarms similarly raised by the user).

The effects of false alarms can be even more pronounced for detection tasks with rare targets or events [23]. If targets appear relatively infrequently, as is the case with search, even a detector with a low false-alarm rate will result in many more false alarms than correct detections (hits). The effects of these false alarms is an important question to address when evaluating the use of detector-based suggestions.

Some studies have shown that imperfect automation has the potential to detract rather than enhance users' performance [11, 12], especially when performing multiple tasks simultaneously. An abundance of either false alarms or missed detections can lead to disuse of the automation. Similarly, frequent misses by the detector may lead to users not making use of it (and thus not gaining the benefit of the suggestions).

Other studies have shown that for UAVs with ATR systems, "human-robot teams can benefit from imperfect [ATR] automation even under high workload conditions" [8, 9]. This suggests that incorporation of imperfect (and even ill-posed) anomaly detectors may behave similarly.

3. COLOR ANOMALY DETECTION

The literature on detection in video is full of various methods, but many of these are not suitable for the particular needs of wilderness search and rescue. Background subtraction [18, 28] can be used to find objects not normally found in a scene, but only if the background has been previously observed. This is not the case when searching a new area, though it does have potential for repeated sweeps. Detectors designed for specific objects (e.g., [21, 27]) also do not apply since the nature of the objects of potential interest is typically not known ahead of time. Detectors based on other domain knowledge [4, 14] might be of use—for example, the method in [4] is useful for finding large man-made objects in natural scenes by using texture distributions for large natural regions versus large man-made regions. However, this reliance on texture means that this method works only for objects significantly larger than those in WiSAR search video, which may only be several pixels in size.

Of course, the one known target is the missing person, so it is tempting to use methods for detecting humans [20, 21, 31]. However, the body of work in this area has focused almost exclusively on detection of humans that are moving and/or upright relative to the camera. Since aerial video sees the person only from above (or perhaps at a slightly oblique angle), these methods cannot be used here unless by chance the person is lying on the ground. Similarly, the speed at which the aircraft passes over the area makes detection and analysis of much-slower human motion difficult.

Finally, while the term *anomaly detection* occurs in the video-processing literature (e.g., [17]), it is commonly used to describe *behavioral* anomalies (such as might be associated with a security threat) rather than simple visual anomalies.

For these reasons, we choose to focus here on anomalous color detection: the finding of unusually colored objects in a scene. This is a subclass of the broader problem of spectral anomaly detection, most of the work for which has focused on hyperspectral images. We adapt here some of these methods for use with RGB video images.

3.1 Spectral Anomaly Detection

A common approach to hyperspectral anomaly detection is to model the statistical distribution of spectral signatures with one or more multivariate normal distributions [3, 25, 26]. This model is then used to identify pixels whose spectral signatures are statistical outliers [5]. The normal distribution is most often 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 [13]. In a multivariate normal distribution, the Mahalanobis distances are distributed according to the chi-square distribution with cumulative distribution function

$$F(d_M; k) = P(k/2, d_M/2) \tag{1}$$

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, $k = 3$.

A multivariate normal distribution rarely characterizes all of the colors in a natural scene, though. This means any effective spectral anomaly detector must perform some trans-

formation on or clustering of the data for it to fit the assumption of normality. Such procedures are referred to here as *normalization*. Once the values have been normalized, the mean vector and covariance matrix are estimated in order to calculate the Mahalanobis distance of each pixel.

3.2 The RX Algorithm

Perhaps the simplest normalization method is the RX algorithm [25]. It assumes 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 [6]. The mean is usually calculated within a window near, but not including, the 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 radius of the outer included neighborhood, R, and the radius of the inner excluded neighborhood, r, of the local neighborhood.

These steps can be thought of as an unsharp masking operation, resulting in a residual error image. Next, a color transformation is applied to the error image with the inverse of its color covariance matrix. Finally, the dot product of each transformed error vector with the corresponding untransformed error vector is computed and stored, resulting in a gray-scale image of Mahalanobis distances, to which is applied a threshold chosen with Equation 1.

Adaptations of the RX algorithm include exchanging the covariance matrix for the correlation matrix [6] or combining local parameter estimation with clustering [1].

3.3 Clustering Methods

A common normalization method is to divide the image pixels into clusters using methods such as vector quantization and k-means [1, 3, 26]. 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, 26] explicitly chooses the distance threshold using the chi-square distribution, as discussed previously.

Gaussian Mixture Modeling (GMM) can be considered a form of fuzzy clustering. It assumes that the true distribution can be modeled by a mixture of normal (Gaussian) distribution components. The GMM for a set of data is usually found using Expectation-Maximization (EM) [10]. If properly estimated, each mixture component may be considered a fuzzy cluster. Unlike clusters produced by k-means or vector quantization, these clusters are designed to be normally-distributed, but this is a much more costly process.

3.4 Robust Methods

Most clustering approaches, such as CBAD [3] 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, 26]. 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 also 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.

4. DETECTOR EVALUATION

To evaluate which detection methods work best for this domain, we implemented the following four methods:

1. The RX algorithm [25]
2. Vector quantization (as used in CBAD [3])
3. K-means clustering
4. The EM algorithm [10], 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.

4.1 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.

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, each typical of wilderness search environments. 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 a digital camera mounted on a small, remote-controlled plane. The photographer then flew the camera over the area, capturing both high-resolution still images and standard-resolution digital video.

The video and images of each scene were then reviewed carefully by visual inspection and with the aid of temporally local mosaics [19]. In addition to the target objects placed in the scene, a number of other objects that could reasonably be considered foreign to the environment were seen. For the grassy location, these included the pilot and two other people, two vehicles, and multiple nearby buildings (video only). For the desert location, these included the pilot and a vehicle (video only), a plastic grocery bag (photographs only), a white box, and a bright orange object.

All anomalies, including the accidental objects listed, were manually labeled in the digital stills on a per-pixel basis. These label maps were then used to tune and compare the different spectral anomaly detection methods.

4.2 Spectral Detector Evaluation

An automated test suite was built for fast and repeatable evaluation. 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 object sizes similar to those seen in the video but not so small as to hinder visual detection, and the images were subdivided to produce stills comparable to the video frames. Each of the 278 still images produced 24 video-

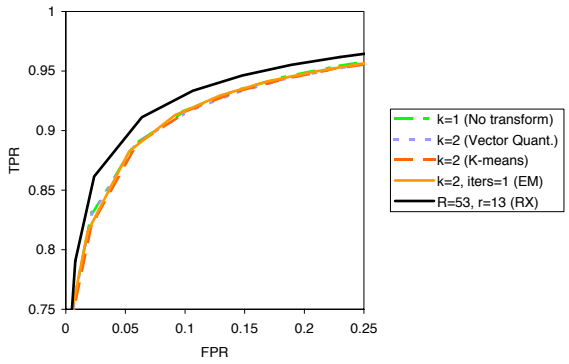


Figure 2: ROC curves for each method

frame-sized subimages, for the equivalent of about 3.7 minutes of manually-labeled high-quality video.

In order to cover as near as possible the full range of false positive values, the target false positive rate was varied from 0% to 100%. This target value was used, in connection with Equation 1, to determine the Mahalanobis distance threshold for each test. Each ROC curve was comprised of 40 tests.

4.3 Detector Evaluation Results

The detector that performed the best overall was the RX algorithm with $R = 53$ and $r = 13$ (Figure 2). While there were significantly worse settings for RX, comparable results were found in a fairly broad range of the parameter space.[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 $k = 2$, but larger values of k showed a decrease in performance overall, even though more clusters sometimes performed better for selected images.

4.3.1 Why Clustering Approaches Struggle

In further analyzing the performance of the various cluster-based approaches, values of k ranging from two to four would often outperform the degenerate $k = 1$ case on many images. In fact, these would often outperform the RX method as well. These results show that clustering can work very well in many cases with appropriate parameter tuning, but the degenerate $k = 1$ case still performs best overall for the clustering approaches, and the RX method in general outperformed that. The problem with using clustering in this domain is that it is sensitive to the content of the scene [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, it should be higher to correctly model the background. The correct number of clusters to use will then change as more or fewer types of ground cover are in view.

In contrast, the best window size for RX is primarily determined by the *projected sizes* of the targets and other objects [3], which is a function of true object size, viewing distance, and camera resolution. In aerial search, altitude is controlled to keep targets large enough for detection while maximizing ground coverage [15]. (For our system's camera resolution, this range is 60–100 meters for targets the size of a person.) Thus, the projected 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.

Although there are numerous methods in the literature for dynamically adjusting the number of clusters (e.g., [29]), these methods were not explored further because dynamic adjustment of the number of clusters on a constant basis was considered too slow for processing of live search video.

4.3.2 BACON

The best performing normalization method, RX, was combined with a robust outlier detection method, BACON [2], to try to improve performance. 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.

Although using BACON produced minimally better results, it was also much slower (for reasons given in Section 3.4). It was also harder to control in terms of tuning the desired false alarm rate than RX alone. Therefore, BACON was not further used for detection in this work.

5. SYSTEM AND USER STUDY

To evaluate the effectiveness of suggestions made by an automated detector for unusually colored objects, we implemented a simple system with detector-based suggestions and conducted a user study to compare user performance both with and without the aid of such suggestions. In particular, we designed the study to investigate potential effects identified previously for automated systems (Section 2):

- How does the introduction of detector-based suggestions affect the user's overall detection sensitivity?
- Do false positives from the detector increase the user's false positive rate?
- Do these false positives distract in such as way as to cause the user to miss potential targets of interest?
- How do these false positives affect a user's performance on a secondary task?

5.1 Presentation and Interface

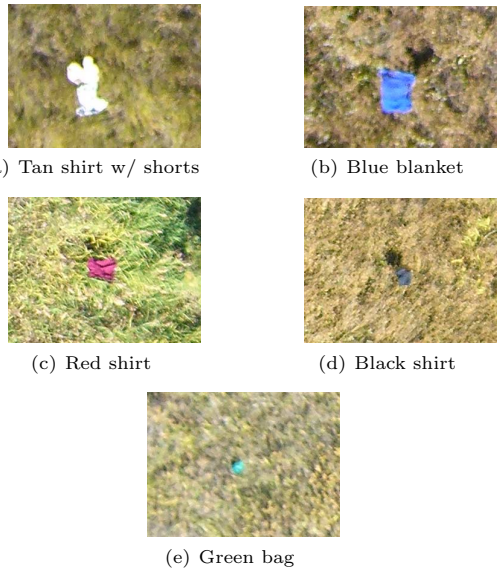
There are numerous ways that the search imagery and detector's suggestions could be presented to a searcher, and the choice of presentation will certainly have an effect on the searcher's performance. In order to keep the implementation and analysis tractable, one simple user interface was implemented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

Each participant was asked to view a series of eight aerial video clips and mark foreign or man-made objects. Participants placed marks (displayed as red circles) on the video with a single mouse-click and could similarly remove marks. To reduce effects of hand-eye coordination, participants were given the option of freezing the video display to examine or mark objects, after which display would resume with the "live" video search. 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 using temporally-local mosaics, a presentation method previously developed in order to expand the spatiotemporal window of observation opportunity for the user [15]. 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.

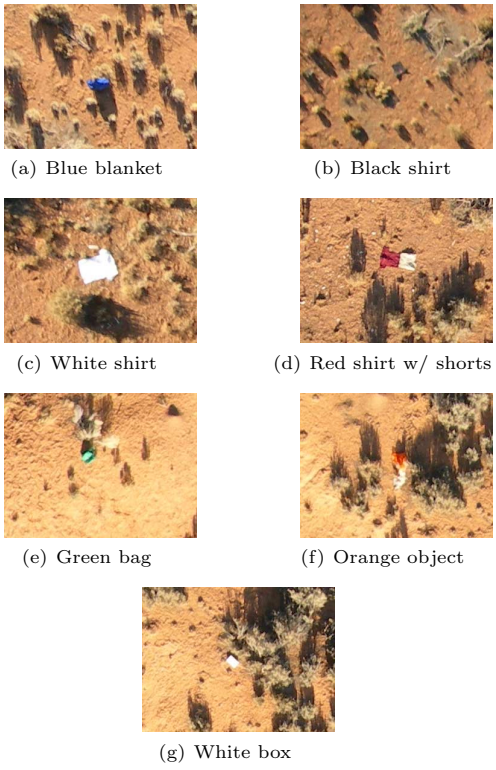
In addition to the primary task of target detection, participants were also given a secondary task in order to assess

[1]A more complete reporting of the performance of each method for various parameter settings can be found in [30].



(a) Tan shirt w/ shorts (b) Blue blanket (c) Red shirt (d) Black shirt (e) Green bag

Figure 3: 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: Target objects at the desert location

the effect on this of primary task automation. For the secondary task, users were asked to count discrete 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. 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. We feel this to be an ecologically valid secondary task, as search-and-rescue personnel are often required to monitor audible communications while performing their primary tasks.

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

After completing a brief demographic survey, each user walked through a set of on-screen instructions. This consisted of a number of explanatory example images 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.

5.2 Video Clips Used

As explained in Section 4.1, aerial video was taken at both a grassy and a desert location, with many of the same target objects being used at both locations. 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. The complete set of eight video clips included two appearances each of 12 target objects for a total of 24 targets. See Figures 3 and 4 for higher-resolution images of each target, and Figure 1 for examples of their placement and size in full video frames.

Each of the 12 target objects consisted 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 at both locations. For example, the same blue blanket was laid out on the ground in the grassy location (Figures 3(b) and 1(a)) and draped over sage brush at the desert location (Figure 4(a)). The same red shirt is laid out by itself in the grass (Figure 3(c)) but paired with the tan shorts in the desert (Figure 3(c)).

In addition to intentionally-placed targets, several other man-made objects were discovered at each location, and 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 (Figures 4(g) and 4(f)). (The detector does not require solid-colored objects, but we used them here for consistency.)

Because some targets appear twice and several physical components are reused between targets, a training effect is possible. In a pilot version of the study, 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 and that the randomized order of presentation was sufficient.

Four of the eight video clips presented to each user included target suggestions, and the other four did not. Suggestions were presented as light blue circles (Figure 5) 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 for the pixel-wise detector was adjusted to 1 in 10 million pixels.

To compensate for noisy imagery (either from acquisition

Figure 5: Suggestions as blue circles

or transmission), we also implemented a spatiotemporal filtering step in order to present to the user only object-level anomalies rather than stray pixels. Potential anomalies were ignored if they appeared for less than 3 frames or in less than 91% of their known temporal extent. Objects of potential 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. The best parameters for the object filtering step were determined empirically using the ground-truth and the object lists from this aggregation step. The 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. Of the 24 target objects, 11 overlapped with suggestions for a 45.8% true-positive rate.

With eight video clips and four presentation methods, there were 32 unique exercises to choose from. The presented sequences were generated so 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 included preventing any presentation method, exercise, or two-clip subsequence from occurring more frequently than the others.

5.3 Data Gathering Methods

The final set of results consisted of data logs from 35 users. Seven clips were shown to each of the participants, with one clip accidentally skipped by one participant. 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.

In reviewing the data for the secondary task, two types of input errors became apparent: so-called "fat-finger" errors (pressing two adjacent number keys when only one was intended), and skipping a tone-count question (which was erroneously recorded as zero). These errors appeared to be random, and we attribute them to the interface for gathering the users' responses rather than the task itself. Data exhibiting such obvious errors (e.g., a response of "87" when the correct answer was "7") were removed manually prior to evaluating the users' performance.

	No Suggestions	Suggestions
Detection Rate	52.57%	61.14%
False Positive Rate	2.88%	2.44%
Sensitivity (d')	1.96	2.25
Response Bias (c)	0.92	0.84

Table 1: User performance on the primary task both with and without automated suggestions.

5.4 Results and Statistical Analysis

An analysis of variance using least-squares means showed that the effect of the suggestions on the primary task performance was significant ($F(1, 33) = 5.69, p = 0.02$). Users' detection rates increased by 16.30% when aided by suggestions from the anomaly detector ($M = 61.14\%, SE = 2.77$) as compared to normal search without suggestions ($M = 52.57\%, SE = 2.57$). The difficulty of the secondary task did not produce a significant effect on the detection rate.

No statistically significant difference was found for either the false positive rate (FPR) or positive predictive value (PPV) for the different presentation methods or level of secondary-task difficulty. Of the four combinations, 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 significant.

Secondary task performance was estimated using 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, suggesting that this aspect did not affect their performance. However, the log MSE increases from 0.5269 to 0.8843 when suggestions are added ($p = 0.0047$), suggesting that the presence of suggestions did have an effect on user attention.

5.5 Discussion and Findings

The likely reason that users detected more targets is that the anomaly detector suggested ones that they otherwise would have missed. But it is also worth exploring the effect of automated suggestions on targets not identified by the anomaly detector, specifically whether the presence of other suggestions distracted the users from these targets or caused them to exhibit complacency. This could potentially decrease the user detection rate for targets that the anomaly detector missed and thus trade increased detection of suggested targets for decreased detection of non-suggested ones.

To see if this effect existed, we broke down user performance further by whether the target was suggested by the detector (11 targets) or was not (13 targets), the results of which can be found in Table 2. For targets the automated detector missed, there was no significant difference in user detection whether the detector was otherwise making suggestions (38.0%) or not (37.3%). This suggests that the presence of other suggestions did not impair the detection of non-suggested targets.

For targets found by the automated detector, users exhibited a significant increase in detection with suggestions

	No Suggestions	Suggestions
Not Marked	37.2%	38.0%
Marked	68.1%	87.9%

Table 2: User detection rates by whether the automated detector identified the target.

(87.9%) compared to without suggestions (68.1%). This suggests that for targets suggested by the detector the contribution of the suggestions is even stronger (29.1%).

It is interesting to note that users still missed some targets even if were correctly identified by the automated detector. This suggests that, even if automatically detected, these targets were still difficult for human users to verify in such a live-flight scenario.

These results plus those from Section 5.4 suggest the following regarding the questions identified previously:

- User detection rates increased by 16.3% from 52.6% to 61.1% when using detector-based suggestions. We would not expect user detection to increase for objects not identified and suggested by the detector, and this was indeed the case (consistently approximately 38%). For objects suggested by the detector, user detection increased 29.1% from 68.1% to 87.9%.
- Although the anomaly detector made frequent false suggestions, we did not detect a significant change in the users' false positive rates when using these suggestions. This suggests that, given modest operational limits, users were able to effectively and quickly filter out incorrect suggestions.
- It did not appear that the presence of false suggestions served as a distractor and caused the users to miss targets of interest. As shown in Table 2, their performance on non-suggested targets appeared to be unaffected by the presence of false suggestions.
- While the use of suggestions increased performance on the primary detection task, it also decreased performance on the secondary tone-counting task. This suggests that filtering false suggestions might require increased user attention. However, it might also suggest that the suggestions shifted more attention to the primary task and away from the secondary task. While the latter is a good thing in terms of the search, it is a factor that should be considered when considering other roles the video searcher might be performing. It should be further noted that 60% of the users subjectively reported that they found the overall combination of tasks easier with suggestions than without.

5.6 Study Limitations

These specific results are for a single simple interface only, and clearly many other factors might be considered. For example, the form in which the suggestions appeared in the interface—their size, coloring, or contrast—would likely affect user performance. Indeed the interface used to present the video itself would also affect the performance.

The parameters of the detector, especially the false positive rate, would also affect the amount to which suggestions aid the searcher. If the false positive rate is too high, it is likely that users would be overwhelmed with processing suggestions, resulting in both decreased detection and increased false positives. In practice, the searcher would also likely be given the ability to adjust this to suit their preferences and comfort level, but this was not included in this study.

The study was limited to fairly consistent (mid-day outdoor) illumination, though this should not generally affect the proposed methods. As long as the variation in lighting is not so extreme as to cause the object's color to no longer appear different from the surrounding environment, the anomaly detector can still function correctly. Changes in daytime lighting also affect shadows, which we found frequently appear as near-black regions of the image and comprise a sufficiently large enough fraction of the image so as to be considered part of the normal environment.

Finally, this study was limited to two search environments typical for Wilderness Search and Rescue and did not encompass a wider range of potential search environments. We have found that some types of search environments generate more potential anomalies than others or are simply more difficult to search. The findings here should generalize to search environments where the targets of interest can be distinguished from the surroundings by their coloring. The environment may consist of a variety of colors, which need not be known ahead of time. Similarly, the targets can consist of multiple colors, which also do not need to be known, as long as one or more of them is atypical of the environment. The methods here would struggle when the target was distinguished by other characteristics (shape, texture, etc.). They would also generate frequent false positives for environments that routinely contain non-target anomalies– i.e., colors that are natural to the scene but do not comprise large parts of it. For these situations, it might be beneficial to allow users to mark these irrelevant anomalies and to explicitly mask these colors in the detector.

6. CONCLUSION

This paper has presented a method for aiding users performing UAV-assisted wilderness search and rescue by incorporating a spectral anomaly detector to suggest unusually colored objects. The contributions include an empirical evaluation of candidate spectral anomaly detectors to evaluate how well each works within this domain, a method for filtering out stray anomalously colored pixels and grouping pixels into potential targets, and a simple method for presenting these potential targets as suggestions to the user.

The empirical evaluation of the various detectors shows that while all of the methods compared here can achieve nearly comparable performance with appropriate parameter tuning, the detector with RX normalization had slightly better overall performance and was much more robust to parameter settings. This robustness relative to clustering-based approaches is likely due to the dependence of these more on the size of potential objects (which stays fairly consistent under search conditions) than on the variation in ground cover (which does not and affects clustering).

The results of the accompanying user study suggest that the use of automated detectors to aid human observers of aerial search video can increase user performance in WiSAR tasks. Even with a partially effective detector (which is perhaps the best that can be hoped for in many situations) that has a much higher false-positive rate than hit rate, user detection can increase without corresponding increase in false positives. User detection of targets suggested by the anomaly detector increased while their detection of non-suggested targets was unaffected. However, the filtering of incorrect suggestions (or a potential shift of attention caused by the suggestions) might require limiting secondary tasks placed on the user.

These findings should transfer to similar search domains where objects of potential interest might be identified by unusual coloring relative to the normal search environment. Actual performance will of course vary depending on the task, the relative clarity of the targets, the detector used,

and the interface used to present the suggestions, but incorporation of suggestions from an automated detector should be considered when designing interfaces for searching video.

7. ACKNOWLEDGMENTS

The work was partially supported by the National Science Foundation under grant numbers 0534736 and 0812653. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank Dr. Dennis Eggett of BYU's Department of Statistics for his help with the design and statistical analysis of the user study. We would also like to thank Jim Walker, Wallace Barrus, and Craig Randall for volunteering their time and expertise with collecting the images and video used for the empirical and user studies.

8. REFERENCES

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

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

[3] M. Carlotto. A cluster-based approach for detecting man-made objects and changes in imagery. *IEEE Trans. Geoscience and Remote Sensing*, 43(2):374–387, 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] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):1–58, 2009.

[6] C.-I. Chang and S.-S. Chiang. Anomaly detection and classification for hyperspectral imagery. *IEEE Trans. Geosci. and Remote Sensing*, 40(6):1314–1325, 2002.

[7] J. Y. C. Chen. Individual differences in human-robot interaction in a military multitasking environment. *Journal of Cognitive Engineering and Decision Making*, 5(1):83–105, 2011.

[8] E. de Visser and R. Parasuraman. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, 5(2):209–231, 2011.

[9] E. J. de Visser and R. Parasuraman. Effects of imperfect automation and task load on human supervision of multiple uninhabited vehicles. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 51(18):1081–1085, 2007.

[10] 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.

[11] S. R. Dixon and C. D. Wickens. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, 48(3):474–486, 2006.

[12] S. R. Dixon, C. D. Wickens, and D. Chang. Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 2004.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

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

[15] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, M. Quigley, J. A. Adams, and C. Humphrey. Supporting wilderness search and rescue using a camera-equipped mini UAV. *Journal of Field Robotics*, 25(1-2):89–110, 2008.

[16] R. C. Johnson, K. N. Saboe, M. S. Prewett, M. D. Coovert, and L. R. Elliott. Autonomy and automation reliability in human-robot interaction: A qualitative review. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 53(18):1398–1402, 2009.

[17] 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.

[18] A. Mittal and D. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 160–167, 2000.

[19] B. Morse, D. Gerhardt, C. Engh, M. 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.

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

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

[22] R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010.

[23] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[24] N. Rasmussen, D. Thornton, and B. 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.

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

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

[27] J. Solka, D. Marchette, B. Wallet, V. Irwin, and G. Rogers. Identification of man-made regions in unmanned aerial vehicle imagery and videos. *IEEE Trans. PAMI*, 20(8):852–857, Aug 1998.

[28] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[29] C. A. Sugar and G. M. James. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, 98:750–763, January 2003.

[30] D. Thornton. Unusual-object detection in color video for wilderness search and rescue. Master's thesis, Brigham Young University, December 2010.

[31] L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. *IEEE Trans. Intelligent Transportation Sys.*, 1(3):148–154, 2000.