A Change Detection Approach to Moving Object Detection in Low Fame-Rate Video

Reid Porter*, Neal Harvey, James Theiler

Space and Remote Sensing Sciences Group, Los Alamos National Laboratory,
Los Alamos, NM, USA.

ABSTRACT

Moving object detection is of significant interest in temporal image analysis since it is a first step in many object identification and tracking applications. A key component in almost all moving object detection algorithms is a pixel-level classifier, where each pixel is predicted to be either part of a moving object or part of the background. In this paper we investigate a change detection approach to the pixel-level classification problem and evaluate its impact on moving object detection. The change detection approach that we investigate was previously applied to multi- and hyper-spectral datasets, where images were typically taken several days, or months apart. In this paper, we apply the approach to low-frame rate (1-2 frames per second) video datasets.

Keywords: Moving object detection, background subtraction, change detection, video, wide-area motion imagery.

1. INTRODUCTION

In temporal image analysis, pixel-level classification is used in two main application domains. The first is background subtraction for moving object detection in fixed frame-of-reference video data[1] and the second is change detection for multi- and hyper-spectral image data[2]. These two application domains have similar goals in that they try to identify rare targets within a cluttered and variable background, but they also have differences. Background subtraction assumes a large number of images while change detection is typically only applied to two images. In addition, background subtraction algorithms have been developed for high spatial and temporal resolution data, with low spectral resolution (panchromatic or 3-color), while change detection algorithms have been applied to multi- and hyper-spectral data, with lower spatial and temporal resolutions (images days or months apart). These differences mean that the pixel-level classifiers proposed for each domain focus on different points in the design space: in background subtraction, classifiers are typically temporally invariant, but spatially variable; while in change detection, the classifiers are typically spatially invariant, but temporally variable.

Recently, wide-area airborne imaging sensors have come into practical use. These systems image small city-sized areas at approximately 0.5m / pixel and about 1 or 2 frames per second. This type of data is called Wide Area Motion Imagery (WAMI) and it lies roughly in between traditional narrow-field-of-view video and commercial satellite imagery. Almost all pixel-level classifiers used for moving object detection in WAMI are from the background subtraction literature. These techniques face significant challenges when applied to WAMI: 1) *point-like* moving objects move anywhere from 1 to 200 pixels 2) data is typically acquired at oblique viewing angles, which means buildings and other tall landmarks suffer from parallax introducing a large amount of motion clutter and 3) registration is often required in real-time and is therefore approximate; e.g., stationary objects might appear to move up to 10s of pixels.

In this paper we investigate a variety of pixel-level classifiers for WAMI by drawing from the collective toolbox of background subtraction and change detection domains. We investigate the tradeoffs as a function of spatial, spectral and temporal resolutions with real datasets. In Section 2 we present an overview of the most commonly used pixel-level classifiers. This overview suggests several variants, as well as some new ways to combine detectors which are described in Section 3. In Section 4 we discuss the experimental framework that is based on ground-truthed WAMI datasets at various temporal and spectral resolutions and in Section 5 we present performance results for the various algorithm variants. We conclude in Section 6 by suggesting some future directions.

* {rporter, harve, jt}@lanl.gov

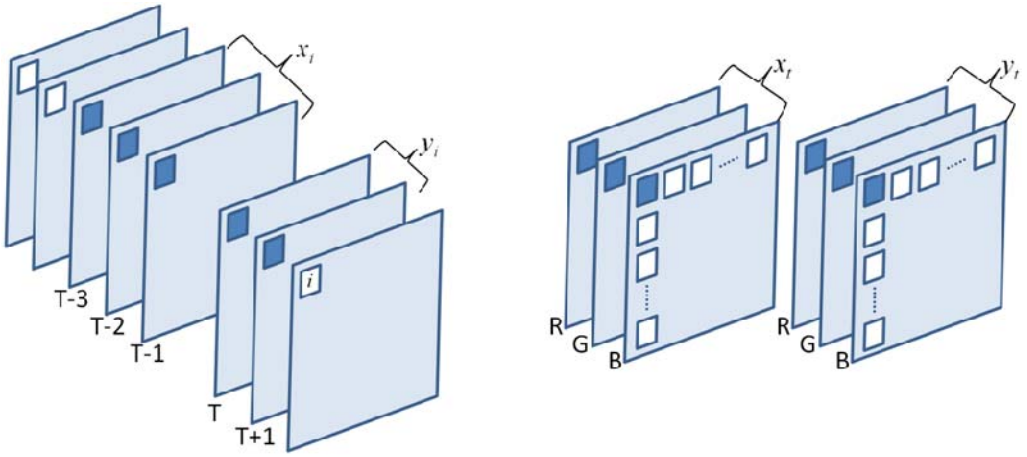

Figure 1. Left) Background subtraction uses spatially variant, temporally invariant models and Right) change detection builds spatially invariant, temporally variant models.

2. PIXEL-LEVEL CLASSIFIERS

The most popular approach to background subtraction involves a set of random variables, where each variable, $x_i \in \mathbb{R}^N$, is associated with a pixel location i. Distributions for the variables, $P_i(x_i)$, are estimated from pixel values sampled from location i, at different points in time. Given these distributions, a new sample at location i is classified as a target if its probability is below a certain constant threshold c. This is a reasonably strategy, since for fixed cameras, and for registered airborne cameras, we typically expect the pixel's value to be relatively constant, unless obscured by a moving object. Another way of interpreting this approach, which is popular in the machine learning community, is as a classifier, where we evaluate the ratio of background and target distributions. Pixels are flagged as moving objects if

$$\frac{P_i(x_i)}{T_i(x_i)} < c, \tag{1}$$

where $T_i(x_i)$ is a target distribution. The most common choice for $T_i(x_i)$ is the uniform distribution, $U(x)$, which is constant for all values of x, and therefore usually omitted in most presentations. We will investigate another possible choice for this distribution in Section 3. A slightly more complicated approach to background subtraction considers a second random variable $y_i \in \mathbb{R}^M$ associated with the same pixel location, but later in time, i.e. x_i and y_i are sliding windows, as shown on the left in Figure 1 for $N = 3$ and $M = 2$. The idea is to exploit the local structure found in the temporal sequence. For example,

$$\frac{P_i(x_i, y_i)}{U(x_i, y_i)} < c, \tag{2}$$

could be used to identify low probability transitions. A popular choice of target distribution in this case is the marginal distribution[3]:

$$\frac{P_i(x_i, y_i)}{P_i(x_i)} < c. \tag{3}$$

This detector is discussed in Section 3 and has also been suggested for fault detection applications[4]. We call the approach used in Equations 2 and 3, temporal change detection, since the models have almost identical form to those suggested for change detection.

In change detection, two random variables, x_t, y_t, are associated with co-registered pixels between two images, which are collectively identified by an index t. The distribution $P_t(x_t, y_t)$ is estimated from co-registered pixel values sampled from different pixel locations spread over the spatial extent of the images. This is illustrated on the right in Figure 1. Similar to Equation 3, a popular strategy for detecting changes in remote sensing imagery is[5]:

$$\frac{P_t(x_t, y_t)}{P_t(x_t)} < c. \tag{4}$$

For change detection applications in remote sensing, we recently proposed and investigated[2]:

$$\frac{P_t(x_t, y_t)}{P_t(x_t)P_t(y_t)} < c. \tag{5}$$

2.1 Parameterization

One of the most popular ways to parameterize the distributions discussed in the previous section is with Gaussians. This leads to quadratic detectors with a simple algebraic form. For example, given sample estimates for the mean $\bar{x}_i$ and the covariance Σ_i associated with variable x_i, we calculate the detector output for background subtraction by taking a negative log likelihood of Equation 1 (assuming the uniform distribution, and ignoring the constant):

$$D(x_i) = (x_i - \overline{x}_i)\, \Sigma_i^{-1} (x_i - \overline{x}_i). \tag{6}$$

Similar expressions are easily obtained for the temporal change detection detectors in Equations 2 and 3. For a comprehensive description of the Gaussian detectors in change detection see Ref. 2.

More complex parameterizations are also used. For example, in background subtraction a Gaussian Mixture Model (GMM) is very popular[6]. The motivation for a GMM is to account for situations where a pixel value changes often due to its location in the scene. For example, a pixel near a tree that sways in the wind, or a pixel close to a high contrast edge. In this paper we are primarily interested in the choice of detector and therefore, for the purposes of comparison, we restrict our attention to Gaussian parameterizations. However, almost all techniques that we investigate can be readily extended to more complex parameterizations.

2.2 Estimation

Background subtraction (Equation 1) aims to capture the long term statistics of the background. Temporal change detection (Equations 2 and 3) aims to capture the long term transition statistics of the background. In both cases the statistics are usually estimated directly from real data and the potential for contamination due to moving objects is typically ignored. This is a reasonable approach since moving objects are typically rare. In principle we can estimate the distribution parameters offline to find a constant distribution for the background which can be used for all time steps. However, in practice it is common to estimate the distribution parameters with an online, or recursive, algorithm. These can be efficiently implemented, for Gaussians and Gaussian mixtures, and also means the distribution can adapt to slowly varying changes in the background. The challenge with this approach is to choose the sliding window sizes and appropriate update-rate for the distribution parameters. When the update-rate is too slow, the approach does not adapt. If the update-rate is too fast, there are limited samples available for estimation and, in the limit, the approach reduces to simple frame differencing. In this paper we use an off-line estimation to determine a constant distribution for the entire sequence. This is because our test data sequences are only 100 frames long (less than 1 minute), and also because it simplifies the experiments.

Loosely speaking, where background subtraction ends, change detection begins. In Equations 4 and 5, the values used to estimate distribution parameters are sampled from different pixel locations, within a single temporal window. This is a good way to suppress pervasive differences that exist between the two images (e.g. due to illumination or mis-registration) and enhance smaller, lower-probability changes. Since moving objects can be expected to be rare spatially (as well as temporally), we expect that this approach will have value for moving object detection as well. A type of change detection was used for moving object detection in Ref. 7. However it differs from this paper in two ways. First, the dimension of x_i was increased by using a small, overlapping patch for each pixel. Second, an online estimation algorithm is used, which updates the distributions in Equation 4 as the detector slides across the image in raster scan order.

3. CLASSIFIER VARIANTS

The main technical challenge is to choose a pixel-level classifier that is most appropriate for the moving object detection problem. This involves choosing the appropriate background and target models which corresponds to choosing the appropriate numerator and denominator for the detectors described in Section 2. In this section we suggest a number of alternatives.

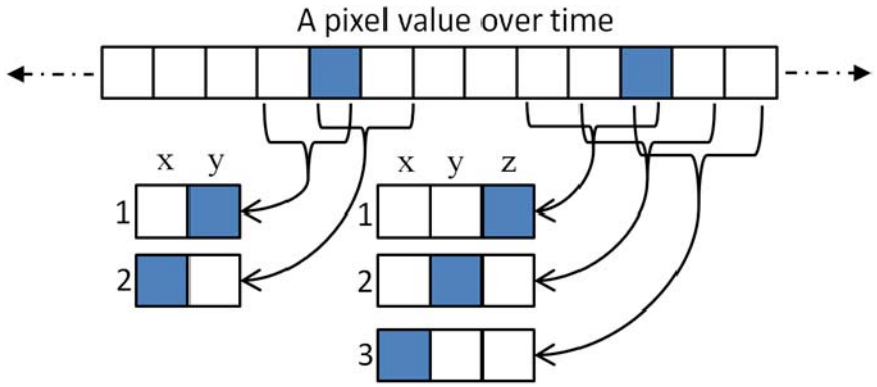

Figure 2. Exploiting local temporal structure can help moving object detection in low frame-rate video.

3.1 Movement of Typical Pixels

Given the background distribution in Equation 1, we are left with the choice of target distribution. The uniform distribution implies we know little about the target which, at a purely pixel (spectral) level, is often true. An alternative target model is the distribution of the image at time t:

$$\frac{P_i(x_i)}{P_t(x_t)} < c \tag{7}$$

This target model biases the detector towards pixels values that are unusual for their location, but not unusual with respect to the rest of the scene. Equation 7 may be particularly appropriate for single channel (gray-valued) datasets where moving objects have similar spectral values to the background. However, this detector also places less emphasis on pixels with values that are fairly unusual for the scene, e.g. sun glint on vehicles, which may be undesirable. We investigate the tradeoff through experiment in Section 5.

3.2 Exploiting Local Temporal Structure

In Figure 2 we provide a simple illustration of why we might expect local temporal structure to help with moving object detection in low frame-rate video. The top row of boxes represents the gray value of single pixel location over time. The pixel's typical value is white and two dark squares represent a low probability moving object passing over the pixel at two different points in time. Underneath this row we show a sliding window of two inputs, on the left (showing 2 specific cases), and three inputs, on the right (showing 3 specific case). The variables x, y, and z are the pixel values in the sliding window at times t-1, t, and t+1 respectively.

For a two-frame detector, $P(x,y)/P(x)$, case 1 has the smallest response: $P(x,y)$ is low for both cases, but $P(x)$ is larger in case 1. This is a useful property since it focuses the detector on the arrival of moving objects and minimizes the effects of ghosting that occurs in simple frame-differencing. By similar reasoning, the three-frame detector:

$$\frac{P(x,y,z)}{P(x,z)} \tag{8}$$

will have the smallest response to case 2, illustrated on the right in Figure 2. The two- and three-frame detectors can be used for both temporal and spectral change detection. We propose that the best choice depends on the characteristics of the dataset and we investigate the performance of both approaches in Section 5.

3.3 Hybrid Approaches

In many ways, the background subtraction and change detection approaches are complementary. Background subtraction can exploit the long-term background statistics of a pixel and can be implemented in a way that adapts to gradual changes in these statistics. Change detection can mitigate sudden pervasive differences by exploiting the fact that these differences affect a large fraction of the image. A potential way to improve performance is to combine aspects from both types of approach. A simple way to do this is to synthesize an estimate for the background image, b_t, and then use this as an additional input to change detection. This leads to the following one-frame, two-frame and three-frame detectors:

$$\frac{P_t(b_t, x_t)}{P_t(b_t)} < c, \tag{9}$$

$$\frac{P_t(b_t, x_t, y_t)}{P_t(b_t, x_t)} < c, \tag{10}$$

$$\frac{P_t(b_t, x_t, y_t, z_t)}{P_t(b_t, x_t, z_t)} < c, \tag{11}$$

where,

$$b_t = \underset{x_i}{\operatorname{argmax}} P_i(x_i) \quad \forall\, i. \tag{12}$$

That is, b_t is an image formed by maximizing the background model probability at each location i. When the distributions are Gaussians, the hybrid detectors can be implemented easily and with a relatively small increase in complexity in comparison to Equation 6. For Equation 12 we simply need to maintain the mean at each pixel location much like we do in Equation 6, and we avoid the covariance estimate at each location. Instead we must estimate the covariance for the distributions in Equations 9-11. The dimension of these covariance matrices is larger than in Equation 6, but we only need to estimate it once for the entire image. The dimension increase has the greatest impact on computation when the detector is applied, however optimizations may be possible since a lot of the multiplications share coefficients.

4. DATASETS

To investigate the classifier variants described in the previous Section we use two WAMI datasets, which we designate A and B. An example of the A dataset is illustrated in Figure 3. It was acquired with a single, relatively narrow field-of-view camera, at approximately 15 frames-per-second with 3-colors, 24-bits / pixel. The B dataset is acquired with a more typical WAMI multi-camera system and produces approximately 2 frames-per-second with 8-bit pixels. Both datasets were taken from airborne platforms and were registered with semi-automated techniques. In a subjective comparison the registration error of dataset A appeared higher than dataset B. To investigate the effect of temporal and spectral resolutions we produced two additional datasets. We use every 5th frame from the first dataset, producing a 3 frames-per-second, 3 color dataset which we designate AR. We convert the images to grayscale to produce a 3-frames-per-second, grayscale dataset (similar resolutions to the B dataset) which we designate AG.

Figure 3. Left) Example image from dataset A and Right) manually identified vehicle tracks within a 150 frame sequence.

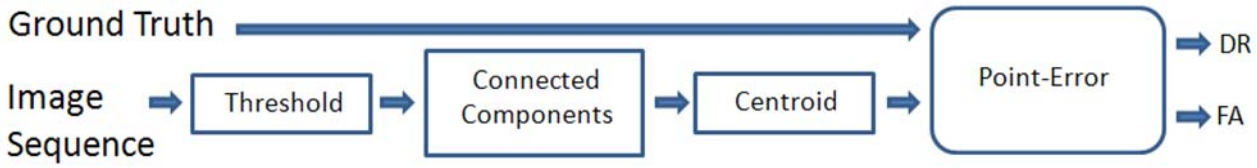

Figure 4. Post processing pipeline used to evaluate performance.

4.1 Performance Metrics

In the A dataset the location of all moving objects were identified in every 5th frame of a 150 frame sequence. This corresponds to all moving objects in 30 consecutive frames of the A-R dataset. Moving objects were identified by manually tracking all vehicles though the sequence (illustrated on the right in Figure 3) and then selecting those track points that moved a minimum distance of 2 pixels. A total of 963 moving objects were identified. A similar method was used for dataset B and 1554 moving objects were identified in a sequence of 100 consecutive frames.

The output produced by the detectors is a real valued image, which is thresholded, at c, to produce a binary image. We post-process the binary image image to produce a candidate set of moving object locations using a standard image processing pipeline, which is illustrated in Figure 4. First, we apply connected component analysis to the binary image, and then calculate the centroid of each component. These centroids are then compared to the list of ground-truth (manually identified) locations to produce an error estimate in terms of the Detection Rate (DR) and the number of False Alarms (FA). These quantities are defined in Equations 13 and 14.

$$DR = \frac{\#\ coincident\ detections}{\#\ ground\ truth\ detections} \tag{13}$$

where the number of coincident detections is the number of ground-truth detections with a connected component centroid within a 5 pixel radius.

$$FA = \#\ centroids - \#ground\ truth\ detections \tag{14}$$

The post-processing steps in Figure 4 are widely used for translating image-based detection images to location-based detections. Most tracking algorithms require location-based detections as input and therefore our performance metric provides a realistic indication of how the detector performance will eventually affect a larger tracking system. In our experiments we will vary the threshold c to obtain a collection of DR and FA numbers which we display in traditional Receiver Operator Characteristic (ROC-curve) form. However, unlike traditional ROC-curves the nonlinear steps that follow thresholding in Figure 4 means that the curves are non-monotonic.

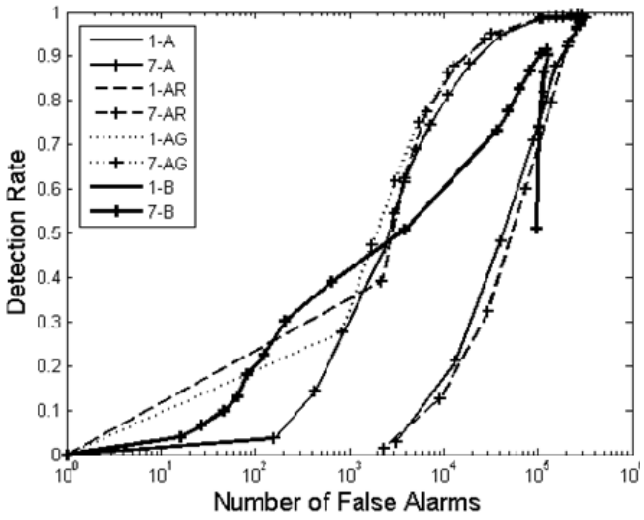

Figure 5. Comparison of Equation 1, with a uniform target distribution, to Equation 7 for all 4 datasets.

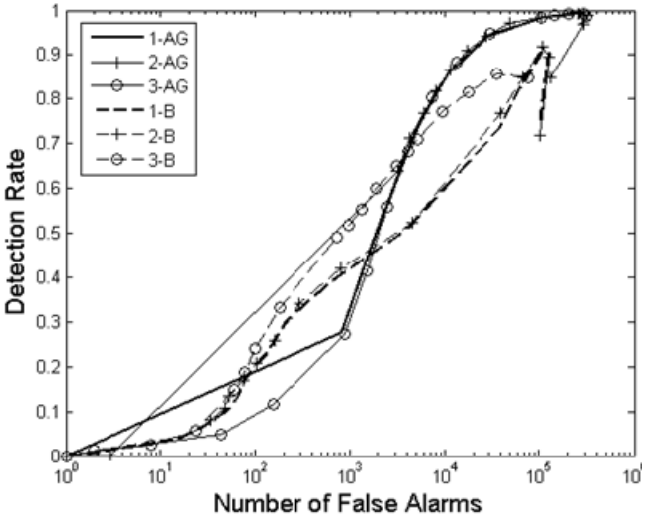

Figure 6. Comparison of Equation 1 to Equations 2 and 3, for the two low frame-rate, single channel datasets.

5. EXPERIMENTS

In the first experiment we investigate the effect of changing the target distribution in Equation 1. We compare the uniform distribution to the movement of typical pixels distribution described in Equation 7 and show the results in Figure 5. We observed that the performance was almost identical in all cases: in Figure 5, plots without markers (associated with the uniform distribution) are completely obscured by plots with markers. We also observe that performance of the detectors is not really affected by the decrease in temporal and spectral resolution. In fact, there appears to be a slight improvement in performance in the AR and AG datasets compared to the original A dataset. We attribute this phenomenon to registration error.

In the second experiment, summarized in Figure 6, we compare the performance of Equation 1 to the temporal change detection algorithms in Equations 2 and 3. The three algorithms obtained very similar performance for dataset AG, but in dataset B there was a performance improvement when using Equation 3 (the conditional probability: $P_i(y_i|x_i)$). We observed the better performance of Equation 3 was due to improved suppression of false alarms at building edges created by registration error.

In the third experiment we compare the performance of background subtraction to the two- and three-frame change detection algorithms in Equations 4 and 8. We also compare the performance of the associated hybrid detectors described by Equations 9-11. We compare the algorithms on all 4 datasets and summarize the results in Figure 7.

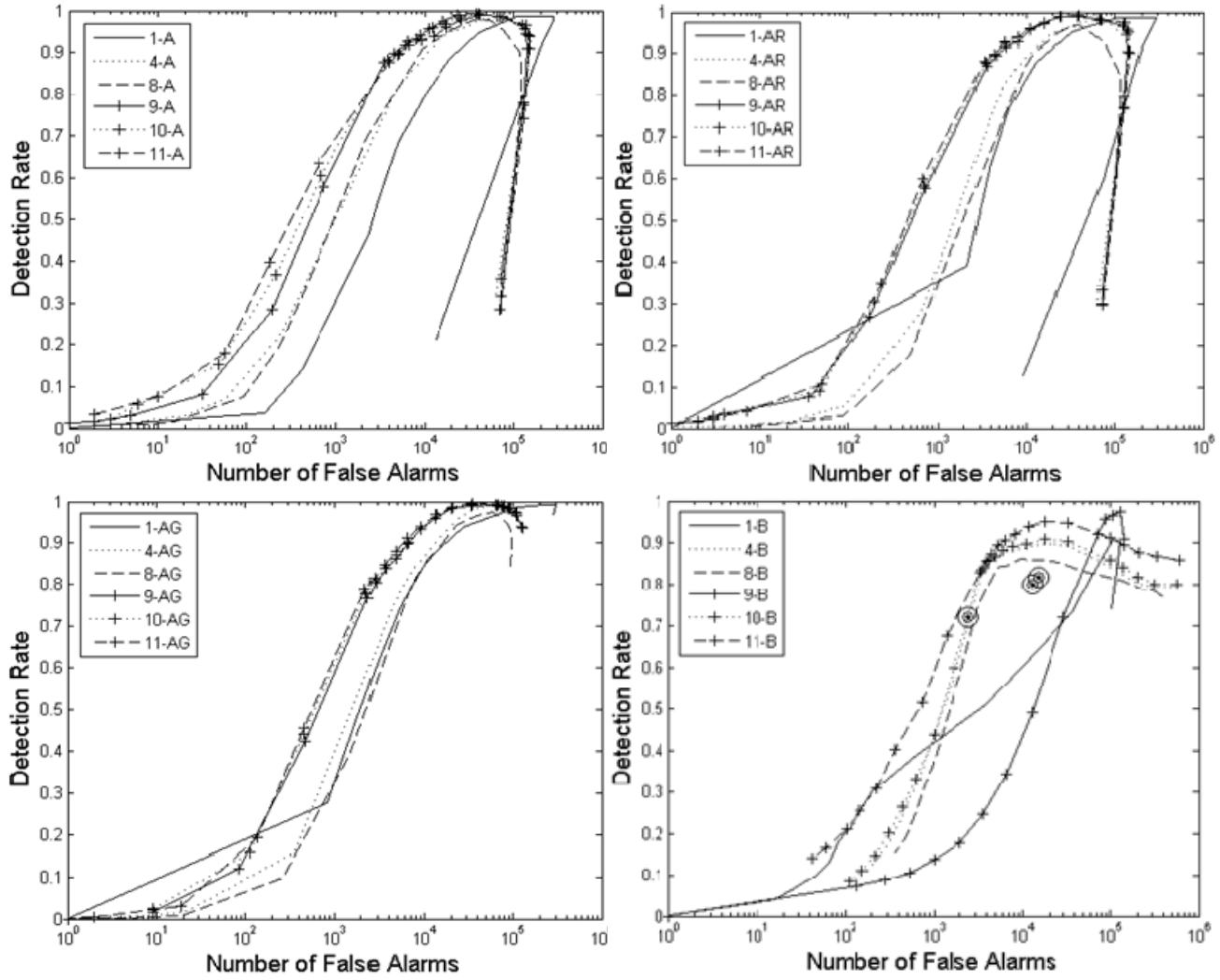

Figure 7. Comparison of change detection and hybrid detectors for datasets
A (top-left), AR (top-right), AG (bottom left) and B (bottom-right).

Our main observations are:

- The two- and three-frame change detection algorithms appear to have equal, or slightly better, performance in comparison to the background subtraction method.
- The two-frame change detection algorithm slightly outperforms the two-frame temporal change detection result obtained in Experiment 2. The two-frame change detection algorithm also appears to slightly outperform the three-frame change detection algorithm in three of the four datasets.
- The hybrid detectors appear to provide significant performance improvement in all 4 datasets. In the A datasets the three-frame version appears to have a slight advantage over the two-frame version, and the two-frame version has a slight advantage over the one-frame version. In dataset B, the three-frame version is clearly the best detector, the two-frame hybrid detector only slightly outperforms the two-frame change detector. And the one-frame hybrid detector performs quite poorly. We note that dataset B contained many more sudden illumination changes compared to dataset A, which may contribute to this result.

In the bottom-right of Figure 7, we also include background subtraction results obtained with a Gaussian Mixture Model implementation of Equation 1, as suggested by Stauffer and Grimson[6]. The implementation produces a binary output and therefore we show three independent results (shown with circles in Figure 7) from three different values of the update parameter. By using a more complex parameterization, we observe that the performance of the background subtraction approach is significantly improved and is approximately equal to the two- and three-frame change detection results.

6. SUMMARY

Background subtraction and change detection application domains have independently developed a number of useful techniques that are applicable to moving object detection in low frame-rate video data. In this paper we have evaluated a number of the design choices on two realistic datasets, and have suggested a hybrid detector that consistently out-performs the independent use of background subtraction and change detection techniques. Future work will investigate more complex parameterizations for these hybrid detectors and also explore other detector architectures that may address a larger range of temporal, spectral and spatial resolutions. For example, incorporating change detection ideas into a hierarchical Gaussian Mixture Model[8] might provide additional performance improvements and a unified approach to a wide-range of applications.

ACKNOWLEDGMENTS

We would like to thank Christy Ruggiero for helping with the ground-truthing of the datasets and Rohan Loveland for supplying and registering one of the datasets. This work was funded by the Laboratory Directed Research and Development program at Los Alamos National Laboratory.

REFERENCES

1. S. Y. Elhabian, K. M. El-Sayed, S. H. Ahmed, "Moving object detection in spatial domain using background removal techniques - State-of-art", *Recent Patents on Computer Science 2008*", No. 1, pp. 32-54, 2008.
2. J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors", *Applied Optics*, **47**(28), pp. F12-26, Oct. 2008.
3. K. Toyama, J. Krumm, B. Brumitt, B. Meyers, "Wallflower: Principles and Practice of Background Maintenance", *Proc. Seventh IEEE International Conference on Computer Vision*, pp. 255-261, 1999.
4. F. Gustafsson, *Adaptive Filtering and Change Detection*, Wiley & Sons, 2000.
5. A. Schaum and A. Stocker, "Long-interval chronochrome target detection", *Proc. 1997 International Symposium on Spectral Sensing Research*, 2008.
6. C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 246-252, 1999.
7. R. Zaibi, A. Enis Cetin, Y. Yardimci, "Small Moving Object Detection in Video Sequences", *Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, pp. 2071-2074, 2000.
8. Y. Sun and B. Yuan, "Hierarchical GMM to handle sharp changes in moving detection", *Electronics Letters*, **40**(13), pp. 801-802, June 2004.