Spatial Multimodal Mean Background Model for Real-Time MTI

Jonathan Williford[a] and Chintan Dalal[a] and Minbo Shim[a]

[a]General Dynamics Robotic Systems, 1234 Tech Ct, Westminster, MD, USA;

ABSTRACT

One of the important tasks in video surveillance is to detect and track targets moving independently in a scene. Most real-time research to date has focused on scenarios from stationary cameras where there is limited movement in the background, such as videos taken at traffic lights or from buildings where there is no background proximal to the background. A more robust method is needed when there are moving background objects such as trees or flags close in the camera or when the camera is moving. In this paper we first introduce a variant of the multimodal mean (MM) background model that we call the spatial multimodal mean (SMM) background model that is better suited for these scenarios while improving the speed of the mixture of Gaussians (MoG) background model. It approximates the multimodal MoG background with the generalization that each pixel has a random spatial distribution. The SMM background model is well suited for real-time nonstationary scenes since it models each pixel with a spatial distribution and the simplifications make it computationally feasible to apply image transformations. We then describe how this can be integrated into a real-time MTI system that does not require the estimation of depth.

Keywords: Background Modeling, MTI, Spatial Multimodal Mean, Mixture of Gaussians, Parametric Motion Model

1. INTRODUCTION

Moving target indicator (MTI) is the problem of detecting and tracking objects that are moving in a scene. There have been many publications and research on MTI using radar. In computer vision, where passive sensors such as visible or IR cameras are used, the tracking problem has received a lot of focus, however the initial detection of moving objects has received less focus. Even in the case where the camera is stationary, this problem can be difficult in real-world scenarios due to lighting changes and background clutter for example. Although this paper focuses on the stationary case, the background model that we propose is well suited for use from a moving platform since it models each pixel in the scene with a spatial distribution.

In this paper we outline our approach of detecting and tracking objects from a visible light camera. We show that the spatial multimodal (SMM) background model improves upon the performance of the multimodal (MM) model while conserving some of the computational efficiency properties. It is shown that the SMM model performs better than the mixture of Gaussians (MoG) model both in speed and performance. The overall MTI framework introduced in this paper has been able to detect and track objects that are moving as far away as 700 meters with a 65 degree field of view on days that are not windy. On windy days, targets can be tracked reliably about 300 meters away. Using standard filtering approaches, occlusions with the background and between moving objects, can be handled.

This paper is organized as follows. A survey of related background models is given in section 2. Section 3 and 4 introduce the spatial multimodal mean background model and how we use it to detect moving objects respectively. Several background models are evaluated in section 5.

Further author information: (Send correspondence to M.S.)
M.S.: E-mail: mshim@gdrs.com, Telephone: 1 410 876 9200
J.W.: E-mail: jwilliford@gdrs.com
C.D.: E-mail: cdalal@gdrs.com

2. RELATED WORK

Background subtraction and other backgrounding techniques enable moving objects to be segmented out of scenes from stationary cameras. Because of this ability, it is a useful preprocessing step for many algorithms, especially with MTI. A more thorough review of background modeling than what is found here can be found in [1].

Most background techniques model each pixel independently. For an $M \times N$ image I, a pixel $x_c \in \{1, ..., M \times N\}$ is modeled by a corresponding pixel in the background model x_b, which for stationary scenes it is often modeled as $x_b = x_c$. The models update each background pixel x_b in an online fashion as each new image I_t arrives. What statistics or information are kept and how they are updated is dependent on the model. The weighted mean model, for example, just keeps a running average of the colors $\mu_{t,x_b} = (1 - \alpha)\mu_{t-1,x_b} + \alpha I_t(x_c)$.

In outdoor scenes where a single pixel may be encoding multiple background objects under varying lighting, modeling each pixel as a single distribution is often insufficient. Consequently, there has been a recent movement towards the use of multimodal approaches. The multimodal mixture of Gaussians (MoG) [2] models each background pixel x_b as a mixture of K Gaussians. In this model, each pixel is represented as a mixture of K Gaussians. The probability of observing a pixel is modeled as

$$P(I(x_{c,t})) = \sum_{i=1}^{K} w_{i,t} N(I(x_{c,t}), \mu_{i,t}, \Sigma_{i,t}), \tag{1}$$

where $N(I(x), \mu, \Sigma)$ is the multivariate Gaussian density function of the color distribution of a mode and $w_{i,t}$ is its corresponding weight. Like the weighted mean, MoG updates the background statistics in an online fashion. A pixel is classified as background if it is within 2.5 standard deviations of one of the Gaussians that has sufficient evidence of being background.

Apewokin et al [3] proposed the multimodal mean background model that provides a 6x speedup over the MoG model. We improved their results by adding spatial sampling of the distributions while still achieving a speedup of 2.5x over the MoG model. Further speedups could be achieved by combining the stationary and spatial updates into a single background update. This model is chosen not only because of the speed, but because the calculations are tractable even when warping the background for non-stationary scenarios.

In order to utilize background subtraction from a moving platform, [4] created the spatial distribution of Gaussians (SDG) model. It transforms each pixel x_c to the estimated background location $\hat{x}_b$ by using the transformation matrix Γ between the two image planes by $\tilde{\hat{x}}_b = \Gamma \tilde{x}_c$ where $\tilde{x}$ is the homogeneous coordinates of x. It assumes that the true position x_b is Gaussian distributed about $\hat{x_b}$. They express the probability density function of the true location x_b as

$$p(x_b|\hat{x}_b) = \frac{1}{2\pi|R|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x_b - \hat{x}_b\right)^T R^{-1}\left(x_b - \hat{x}_b\right)\right). \tag{2}$$

The SDG models each pixel as a bimodal mixture of Gaussians. A narrow Gaussian encodes the background and a wide Gaussian (or uniform distribution) encodes the targets. It models the probability density function of observing a $I(x)$ of a pixel x as

$$p(I) = p(I|B)P(B) + p(I|\neg B)P(\neg B), \tag{3}$$

where B stands for background and $\neg B$ is the foreground. They then use the likelihood ratio test for classification. The speed of the SDG approach was not reported.

3. SPATIAL MULTIMODAL MEAN BACKGROUND MODEL

Multimodal background techniques are able to model a point in the background, for example a specific leaf on a tree, that is repeatedly traversing the same path on the background plane. This is a common occurance, largely due to wind that can cause branches and other objects to sway. However, a deviation from the path as

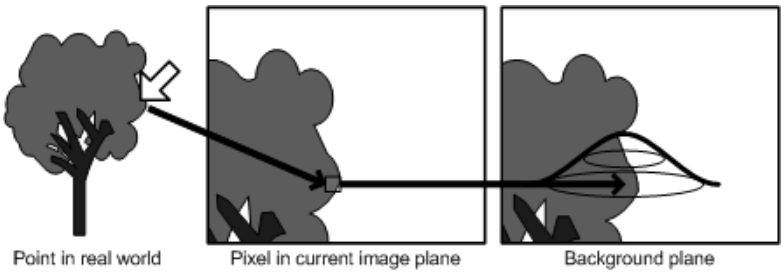


Figure 1. Throughout a sequence of images, a single point x in the background of the scene is constantly being projected onto a pixel x_c of the current image plane. Due to the factors such as wind or movement of the camera, the corresponding pixel location x_c changes throughout the sequence. Given a pixel x_c, the correct corresponding pixel location of the background plane, x_b, is unknown. In addition to keeping a multimodal distribution, SMM models the location of the corresponding background pixel as $x_b = x_c + e$, where e is a bivariate probability distribution and $E[e] = \vec{0}$. Therefore a neighboring point can be projected to a novel neighboring location and still be correctly classified as background.

slight as a single pixel may cause misclassification as foreground. Our spatial multimodal mean (SMM) models the aforementioned scenerio by adding spatial variance, allowing deviation from oscillating paths without being misclassified as foreground. See figure 1.

Let x be a point in the scene, x_c the pixel location in the current image plane, x_b the corresponding pixel location in the background model, and $I(x_c) \in colorspace$ and $B(x) \in \{true, false\}$ be its color in the current image and background classification respectively. We relax the assumption that $x_b = x_c$ with $x_b = x_c + e$ where e is a spatial bivariate distribution with mean $\vec{0}$. We randomly sample pixels from the simulated distribution during the background update.

The multimodal distribution for each pixel x is represented by up to K modes, $m_{x,i}$ where $i \in \{1...K\}$. Each mode $m_{x,i}$ is associated with the integer fields: *count*, *sum*, r_1, and r_2, where *count* is the number of pixels that have been matched with the mode, *sum* is the sum of the corresponding color vectors, and the recency $R = r_1 + r_2$ provides a measure of how often the color distribution was matched within recent history. Table 1 shows the pseudo code of our algorithm for the SMM background model.

A mode is considered background if its count is at least T_{FG} or half of the sum of all of counts of the modes x_b. The latter condition is a small addition to [3] that allows for decent classification after the first several frames of starting the algorithm. T_{FG} is a parameter of the maximum number of times a pixel can be observered in the scene and still be classified as foreground. A pixel is classified as background if it matches a background mode. A match to a mode is defined as each color component in x_c being within a predefined distance E_x of the mean color of that mode.

The function $create\_match\_or\_replace\_mode(I(x))$ first tries to find an existing mode that matches the color of the pixel x and if successful it returns this mode. Otherwise, if there are less than K modes, then it creates a new one. If there are already K modes, then it finds the mode that with the smallest count from the set of modes $\{m|R_m < (w/K)\}$ with low recency values.

For the purpose of adaptation to changes in the scene, we divide the *sum* and *count* fields by two for all of x_bs modes every d frames, where d is a predefined decimation rate. The recency is estimated for each mode by setting r_2 to the value of r_1 and resetting r_1 to 0 every w frames and incrementing r_1 and leaving r_2 untouched for all the other frames.

Pixels with the same distance in the RGB colorspace will be perceptually more similar in the dark regions than in the bright regions. A colorspace that approximates perceptual uniformity, such as CIE 1976 (L*,u*,v*) [5,6], should therefore be used so that the distances in the colorspace are proportional to the perceived color distance. Due to the efficiency that colorspace conversion can be calculated, dealing with the perceptually warped RGB colorspace is not justified.

Table 1. Algorithm for spatial multimodal mean on a pixel at x.

```
update_pixel( x )

for t = 1 to N
  classify(x)                                        ▷ classify pixel

  m_x :=create_match_or_replace_mode(I(x))           ▷ find closest distribution
  update_mode(m_x,I(x))                              ▷ update the distribution

  x_r :=normrnd(x, σ)                                ▷ spatial distribution
  m_{x_r} :=create_match_or_replace_mode(I(x_r))     ▷ find closest distribution
  update_mode(m_{x_r},I(x_r))                        ▷ update the distribution


  if t%d = 0 then                                    ▷ decimation step
    for each mode m_{x,i}                            ▷ d is decimation rate
      count(m_{x,i}) := count(m_{x,i})/2
      for each channel c
        sum(m_{x,i}, c) := sum(m_{x,i}, c)/2
      end
    end
  end


  if t%w = 0 then                                    ▷ recency reset
    for each mode m_{x,i}                            ▷ w is time frame sliding window
      r_2(m_{x,i}) := r_1(m_{x,i})
      r_1(m_{x,i}) := 0
    else
      r_1(m_{x,i}) := r_1(m_{x,i}) + 1
    end
  end
end
```

Table 2. Algorithm for updating the mode m with a pixel x.

```
update_mode( m, I(x) )

count(m) := count(m)+1
for each channel c
  sum(m, c) := sum(m, c)+I(x, c)
end
```

4. MOVING TARGET DETECTION

The SMM background model is used as a part of an MTI framework. Typical techniques such as calculating connected components [7] from the background model and utilizing Kalman filters [8] are used by our MTI system in order to utilize the background classification. Even with a good background model these techniques are not always sufficient to build a robust outdoor MTI system. For example, trees and shrubs which are proximal to the camera can cause significant misclassification in a windy environment. These approaches may incorrectly "see" movement in such areas of misclassification because of the heuristics used to track the foreground segments between frames.

To increase the reliability of the classification of moving targets, we introduce a probabilistic framework in the following subsection. Figure 2 gives the intuition behind the proposed probabilistic approach.

4.1 Probabilistic Classification of Moving Targets

In order to detect moving targets, it is useful to have a precise definition of what it means to be moving. For MTI applications, it is not useful to define objects that sway in the wind or similar oscillating motion as being moving targets. Also, perceived movement from noise should not be considered movement.

Let $U \in \mathbb{R}^2$ be a vector of random variables describing movement of a target, in a single time step, on a stationary image plane. We then define a stationary target to be a target such that $E[U] = \vec{0}$ and correspondingly a moving target is a target such that $E[U] \neq \vec{0}$.

In the rest of this section, we will show how to calculate the probability that a target is stationary, when given the observations of a target. We can estimate the probability distribution of U as

$$U \sim N(\widehat{\mu}, \widehat{\Sigma}), \tag{4}$$

where $\widehat{\mu}$ is the sample mean and $\widehat{\Sigma}$ is the sample variance. Figure 2 (a)-(b) shows two such estimated PDFs where the sampled mean is the same but the sampled variance is different.

We can also approximate the distribution of U with the assumption that the object is stationary by locking the mean to $\vec{0}$:

$$\overset{\circ}{U} \sim N(\vec{0}, \widehat{\Sigma}). \tag{5}$$

See figure 2 (c)-(d).

Let $\star$ be the observation of that the current sample mean is at least as large as $\widehat{\mu}$. We are interested in $Pr(E[U] = 0|\star)$, ie. the probability that the object is stationary given our observation $\star$. We will first show how to calculate $Pr(\star|E[U] = 0)$ and then use Bayes theorem to calculate $Pr(E[U] = 0|\star)$.

Let $\overset{\circ}{V}$ be the random variable for the sample mean of $\overset{\circ}{U}$, in other words,

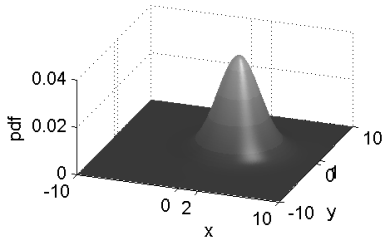
$$\overset{\circ}{V} = \sum_{i=1}^{n} w_i \overset{\circ}{U}^{(i)}, \tag{6}$$

where n are the number of samples, $\overset{\circ}{U}^{(i)} \sim N(0, \widehat{\Sigma})$ is the RV for the ith displacement and w_i is its weight (figure 2 (e)-(f)). The weights correspond to the definition of the sample means being used, which usually means that $w_i = 1/n$. From the properties of multivariate normals,

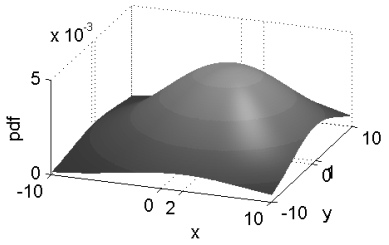
$$\overset{\circ}{V} \sim N\left(\vec{0}, \left(\Sigma_{i=1}^{n} w_i^2\right) \widehat{\Sigma}\right). \tag{7}$$

We can now use $\overset{\circ}{V}$ to calculate the probability of $\star$ under the stationary assumption with

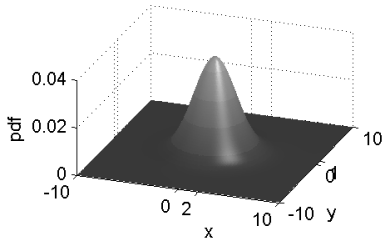
$$Pr\left(\star \middle| E[U] = 0\right) \equiv Pr\left(|X_1| \geq n\,|\widehat{\mu}_1| \cup |X_2| \geq n\,|\widehat{\mu}_2|\right), \tag{8}$$



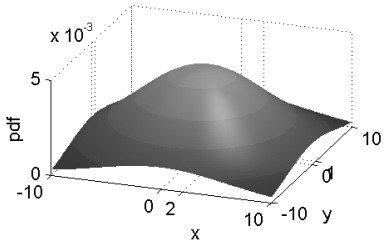
(a) Target 1's estimated PDF of U.



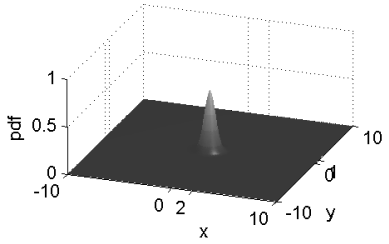
(b) Target 2's estimated PDF of U.



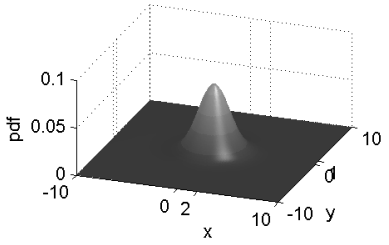
(c) Target 1's estimated PDF of $\mathring{U}$.



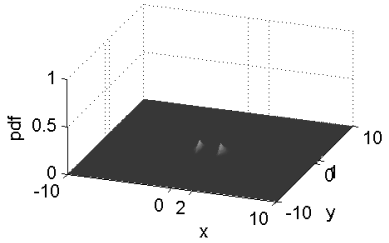
(d) Target 2's estimated PDF of $\mathring{U}$.



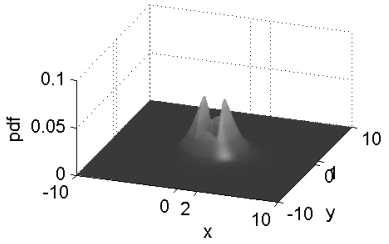
(e) Target 1's estimated PDF of $\mathring{V}$.



(f) Target 2's estimated PDF of $\mathring{V}$.



(g) Target 1's $\mathring{V}$, where $|\ \mathring{V}_1\ | > 2$ or $|\ \mathring{V}_2\ | > 1$.



(h) Target 2's $\mathring{V}$, where $|\ \mathring{V}_1\ | > 2$ or $|\ \mathring{V}_2\ | > 1$.

Figure 2. The probability density functions (PDFs) that are estimated from n sampled movements of two targets with the same sampled mean $(+2, +1)$ are shown in (a) and (b). (a) contains very little noise relative to the movement, while (b) contains a lot more. Intuitively, the probability that the true mean of (a) is the null vector is less than the probability that the true mean of (b) is the null vector. (c)(d) show the PDFs with the assumption that the objects are stationary, (e)(f) show the PDFs of $\mathring{V}$, the random variable for a sample mean with n samples, as defined in (6) with $n = 4$, and (g)(h) have the regions that will not be integrated ($-2 < x < 2$ and $-1 < y < 1$) set to 0. Since the $Pr(\star)$'s are equal for both targets, using (9) we know that the larger area under (h) means that target 2 is more likely to be stationary. The results then match our intuition that target 1 is more likely to be moving than target 2.

where $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \overset{\circ}{V}$, $\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = \hat{\mu}$, and n is the number of samples used to calculate $\hat{\mu}$; hence, $n\hat{\mu}$ is the difference between the last sampled pixel location and the first sampled pixel location.

(8) can be calculated by integrating over $\overset{\circ}{V}$ where $|X_1| \geq |\widehat{\mu}_1| \cup |X_2| \geq |\widehat{\mu}_2|$ or by using an estimate of the CDF of $\overset{\circ}{V}$ [9–13]. See figure 2 (g)-(h).

Using Bayes theorem

$$Pr\left(E[U] = 0 \middle| \star\right) = \frac{Pr\left(\star \middle| E[U] = 0\right) Pr\left(E[U] = 0\right)}{Pr\left(\star\right)}. \tag{9}$$

The probability of an object being stationary, $Pr\left(E[U] = 0\right)$, can be calculated from training datasets, which depends on the platform and background model. $Pr\left(\star\right)$ can be modeled as a univariate or bivariate normal and the parameters can be estimated from real datasets.

5. EXPERIMENTS

We evaluated the weighted mean, MM, and SMM backgrounding methods on four image sequences. Two image sequences, "Waving Tree" and "Bootstrap," are from the Wallflower benchmarks [14]. We include the results of these two benchmarks with the frame differencing and MoG backgrounding techniques as reported by [3]. The third image sequence "Windy" was collected on a windy day. The last image sequence "River" was simulated in the RIVET (Robotic Interactive Visualization and Exploitation Technology) platform.

Each image sequence contains a frame that is used for evaluation. Every pixel in the ground truth image for the evaluation frame is hand-labelled as either background or foreground. We used the ground truths provided with the Wallflower sequence so that our results could be compared with that in [3]. Both MM and SMM produced near zero false positive classifications in the River sequence after sufficient samples, so we evaluated the results of early in the sequence. The benchmark sequences are summarized in Table 3. The Bootstrap sequence is the only one of the included benchmarks that has significant foreground objects in every image.

The parameters used for the different models are shown in Table 4. The same parameters were used as in [3] except for MM, where the number of bins were increased from 4 to 5. We did this so that the increase in performance in SMM would not be attributed to the difference of the number of bins. The E_x distance threshold was made smaller in the SMM since this could be done without a significant increase in the false positives.

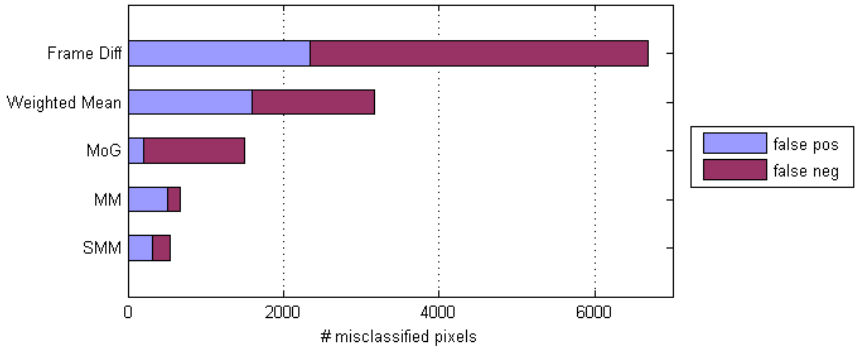
We ran the experiments on a 2.9 GHz Intel Core Duo system running Windows XP. The background models were written in C++. In order to evaluate the computational cost of the MM and SMM algorithms more efficiently, we ran the algorithms on a mosaicked image of size 3380x400 pixels. The MM took 149 ms and the SMM took 361 ms. Assuming a similar speedup of the MM over the MoG as in [3] , the SMM provides a 2.5x improvement in execution speed. Further increases in computational efficiency could be achieved by utilizing the GPGPU.

The accuracy of the background models is shown in figures 3 and 4. The spatial distribution of the SMM allow a tighter threshold to be used without adding additional noise. Because of this, the SMM is able to detect more of the true foreground objects in the Bootstrap sequence. Most of the weighted mean's false positive results in the River sequence are from the area following the Stryker that has been added to be background model, while with the MM and SMM it is from the movement of the river.

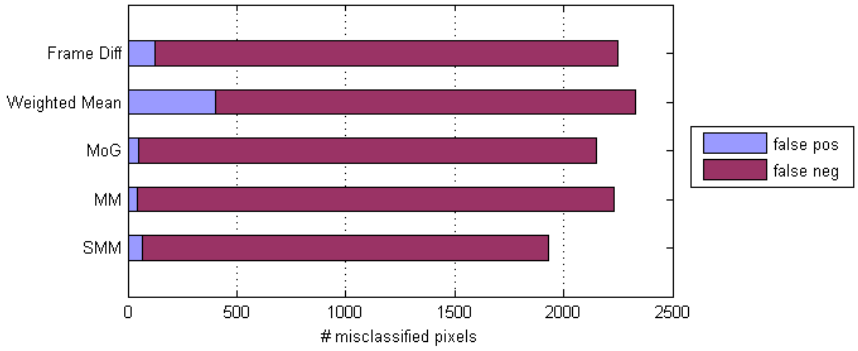
Figure 5 shows the cropped result of the MTI framework being applied to a large mosaicked image.

6. CONCLUSION

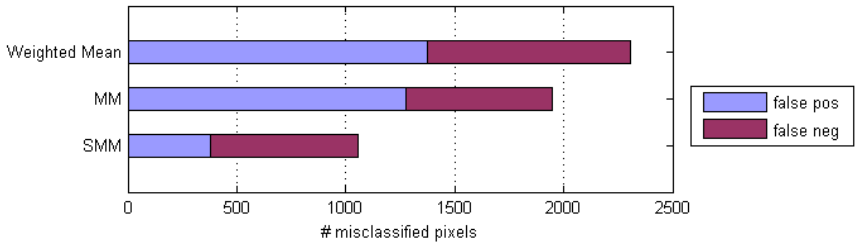
In this paper we proposed the spatial multimodal mean background model and compared it to several other background models, showing that it is more robust towards movement of the background. It is 2.5 times faster than the mixture of Gaussians background model while providing higher accuracy. The new model is used as a major part of a reliable real-time stationary MTI framework that is not dependent on depth estimation.



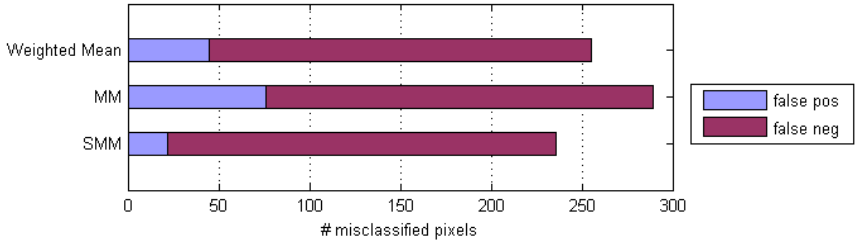
(a) Waving Trees



(b) Bootstrap



(c) Windy



(d) River

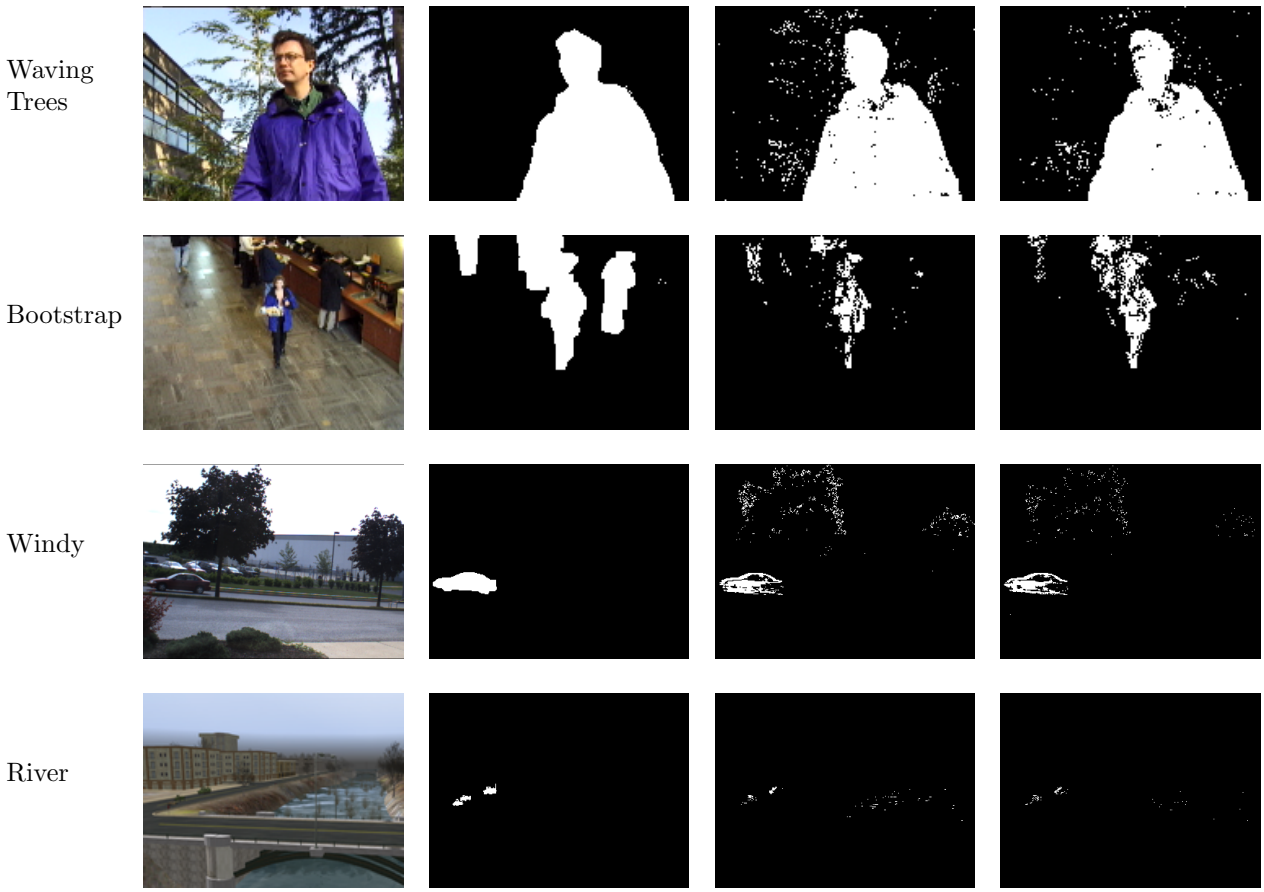
Figure 3. Frame difference and MoG results were reproduced from [3].

Table 3. Benchmarks

Sequence	Sampled Frame	Downsampled Frame Size (width x height)
Waving Tree	247	160 x 120
Bootstrap	299	160 x 120
Windy	571	341 x 256
River	52	341 x 256

Table 4. Background Evaluation Parameters

Algorithm	Parameters
Frame Differencing	$E_x = 30$ for $x \in \{R, G, B\}$
Weighted Mean	$\alpha = 0.1$ in $u_t = (1 - \alpha) * u_{t-1} + \alpha x_t$
Mixture of Gaussians	$K = 4$, initial weight $w = .02$, learning rate $\alpha = 0.01$, weight threshold $T = .85$
Multimodal Mean	$K = 5$, $E_x = 30$ for $x \in \{R, G, B\}$, $T_{FG} = 3$, decimation rate $d = 400$, recency rate $w = 32$
Spatial Multimodal Mean	$K = 5$, $E_x = 20$, $x \in CIELUV$, $T_{FG} = 6$, $d = 400$, $w = 32$, spatial variance $\sigma = 1$ with a 5x5 window



(a) Evaluation Image (b) Ground Truth (c) MM Result (d) SMM Result

Figure 4. Results comparing MM and SMM.

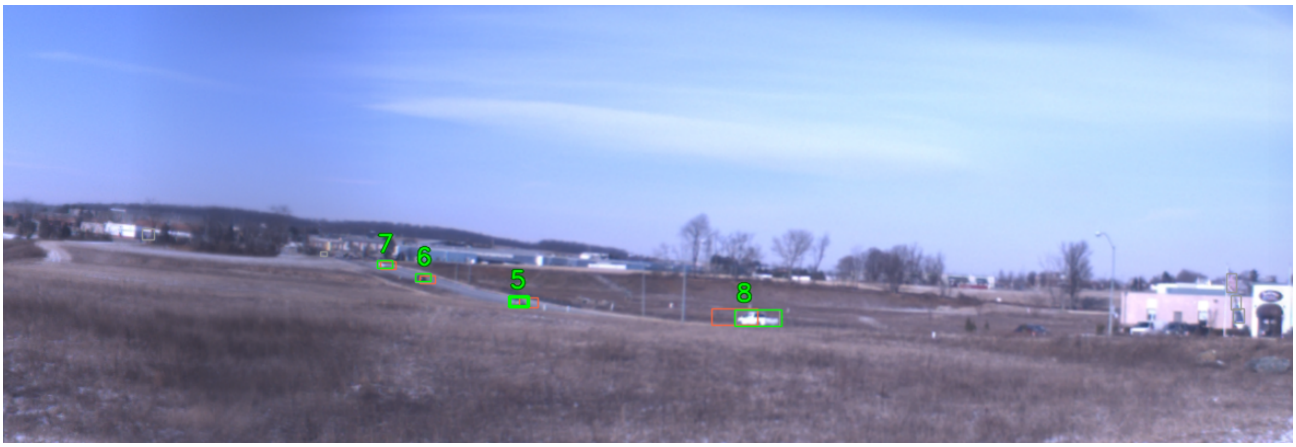


Figure 5. Cropped result from entire MTI system. The dark vertical band in the image is from image mosaicking. Detected moving targets are indicated with two boxes and an ID, where the boxes indicate the current and predicted location in the next frame. The boxes without any ID indicate areas of interest. The boxes with IDs indicate moving objects. The flag on the right moves around too much to be removed from the background model, however it still is not detected as a moving object. A few vehicles on the distant roads are indicated as areas of interest.

ACKNOWLEDGMENTS

This research described in this paper was carried out by the General Dynamics Robotic Systems and was sponsored by the U.S. Army Research Laboratory, under contract Robotic Collaborative Technology Alliance. Reference herein to any specific commercial product, process, or service by trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government.

REFERENCES

[1] Piccardi, M., "Background subtraction techniques: a review," *IEEE Conference on Systems, Man and Cybernetics* **4**, 3099–3104 (2004).

[2] Stauffer, C. and Grimson, W., "Adaptive background mixture models for real-time tracking," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2**, 2246–2252 (1999).

[3] Apewokin, S., Valentine, B., Wills, S., Wills, L. M., and Gentile, A., "Multimodal mean adaptive backgrounding for embedded real-time video surveillance," in [*Proceedings of the Embedded Computer Vision Workshop*], (2007).

[4] Ren, Y., Chua, C.-S., and Ho, Y., "Motion detection with nonstationary background," *Machine Vision and Applications* **13**, 332–343 (2003).

[5] Malacara, D., [*Computer Graphics: Principles and Practice*], SPIE Press, first ed. (2002).

[6] Foley, J. D., Dam, A. V., Feiner, S. K., and Hughes, J. F., [*Computer Graphics: Principles and Practice*], Addison-Wesley, Reading, Massachusetts, second ed. (1991).

[7] Horn, B., [*Robot Vision*], MIT Press, Cambridge, Massachusetts (1986).

[8] Kalman, R. E., "A new approach to linear filtering and prediction problems," *Transactions of the ASME-Journal of Basic Engineering* **82**(Series D), 35–45 (1960).

[9] Cox, D. R. and Wermuth, N., "A simple approximation for bivariate and trivariate normal integrals," *International Statistical Review* **59**, 263–269 (1991).

[10] Divgi, D. R., "Calculation of univariate and bivariate normal probability functions," *Annals of Statistics* **7**, 903–919 (1979).

[11] Drezner, Z. and Wesolowsky, G. O., "The computation of the bivariate normal integral," *Journal of Statistical Computation and Simulation* **35**, 101–107 (1990).

[12] Agca, S. and Chance, D. M., "A comparison of alternative bivariate normal probability estimation procedures for compound and min-max options," *SSRN eLibrary* (1999).

[13] Wang, M. and Kennedy, W. J., "Comparison of algorithms for bivariate normal probability over a rectangle based on self-validated results from interval analysis," *Journal of Statistical Computation and Simulation* **37**, 13–25 (1990).

[14] Toyama, K., Krumm, J., Brummit, B., and Meyers, B., "Wallflower: Principles and practices of background maintenance," in [*Proceedings of International Conference on Computer Vision*], 225–261 (1999). Benchmarks are available online http://research.microsoft.com/∼jckrumm/WallFlower/TestImages.htm.