Adaptive background subtraction models to Enhance Moving Target Classification from Real-Time Video

Brian Ly
York University
4700 Keele St, Toronto, ON M3J 1P3
Lybrian1@my.yorku.ca

Abstract

This paper describes an improved end-to-end method for extracting moving targets from surveillance video and classifying them into two predefined categories according to image-based properties. The moving targets in the video are detected using adaptive background subtraction models. A classification metric is applied to each target and classifies them into two distinct categories: humans and cars, while ignoring background clutter. An evaluation of the proposed algorithm will be conducted at the end of this paper.

1. Introduction

1.1. Motivation

Inexpensive surveillance cameras are becoming increasingly available to the public with relatively high-performance video processing hardware. This opens up advancements in motion detection and object classification technologies. This field is very beneficial in our ever-increasing state of surveillance where countries across the world are adopting mass surveillance to monitor their citizens. Moving target detection and classification from real-time surveillance video are becoming more important than ever, especially in developing a technique to process large volumes of video footage on relatively low powered consumer hardware.

This paper will mainly focus on the object classification of pedestrians and vehicles from surveillance video since it is most obvious use of consumer security systems. It is important that the software is able to be robust enough to work under less than ideal circumstances. It is common for surveillance videos to have significant levels of noise especially in low light conditions such as during the night. This is where video preprocessing to reduce noise and using background subtraction techniques may benefit existing object identification and classification methods.

1.2. Related Research

The development of object detection and classification of moving targets can be traced back to the highly cited paper "Moving target classification and tracking from real-time video" by Alan J. Lipton, Hironobu Fujiyoshi, and Raju S. Patil [1]. The paper describes a method for extracting moving targets from real-time video and classifying them into distinct categories such as humans and vehicles with imaged based properties. Once classified, the targets are robustly tracked by temporal differencing and template matching. The final product will continually track objects across the screen and reject background clutter.

Historically, background subtraction involves calculating a reference image, subtracting each new frame from this image and thresholding the result. In non-adaptive methods of background subtraction, a single time-averaged background image is produced and becomes the reference image. This technique requires a training period of constant re-initialization without foreground objects, otherwise errors in the background will accumulate over time. Additionally, this method does not work well with gradual changes in background illumination which makes it impractical to use with surveillance applications.

In the highly influential paper "Adaptive background mixture models for real-time tracking" by Chris Stauffer and W.E.L Grimson, the authors devised an improved adaptive background mixture model that uses an adaptive nonparametric Gaussian mixture model to solve the problems of existing background models. The Gaussian distributions of the adaptive mixture model are continuously evaluated with an algorithm to determine which update in the model is a result of background processes and which are foreground objects of interest.

1.3. Hypothesis

Background subtraction by modeling each pixel as a mixture of Gaussians is proposed to improve the end-to-end method for moving object extraction from a real-time video stream, classifying objects into predefined categories and robustly tracking objects across the scene. It is hypothesized that using the adaptive background mixture model algorithm will greatly improve object detection and

classification in noisy and low illuminated videos. Furthermore, it may improve performance and create a more robust system that does not depend on lighting changes or moving elements in the scene.

2. Approach

2.1. Mixture Model

Using a mixture of adaptive Gaussians per pixel.
First, a time series of pixel values is created.

Denote t as time

Denote $\{x_0,y_0\}$: particular pixel

Denote I as the image sequence

Denote $\{X_1,...,X_t\}$ is modeled by a mixture of K Gaussian distributions

The recent history of a particular pixel at $\{x_0,y_0\}$:
$$\{X_1,...,X_t\} = \{I(x_0,y_0,i) : 1 \leq i \leq t\}$$

Denote K as the number of distributions

Denote $\omega_{i,t}$ as an estimate of the weight of the i[th] Gaussian in the mixture at time t

Denote $\mu_{i,t}$ as the mean value of the i[th] Gaussian in the mixture at time t

Denote $\Sigma_{i,t}$ as the covariance matrix of the i[th] Gaussian in the mixture at time t

Denote η as a Gaussian probability density function

The probability of observing the current pixel value:
$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t})$$

Gaussian probability density function:
$$\eta(X_t,\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_t)^T\Sigma^{-1}(X_t-\mu_t)}$$

The covariance matrix:
$$\Sigma_{k,t} = \sigma_k^2\mathbf{I}$$

Every new pixel value, X_t, is checked against the existing K Gaussian distributions, until a match is found with a pixel value within 2.5 standard deviations of a distribution. If no match is found then the least probable distribution is replaced with a distribution with the current value as its mean value.

Denote α as the learning rate[2]

Denote $M_{k,t}$ as the value 1 for the model which matched and 0 for the remaining models

The adjusted $\omega_{i,t}$ for the prior weights of the K distributions at time t:
$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t})$$

The weights are then normalized.

Denote μ and σ as parameters for unmatched distributions to find the second learning rate[3]

The parameters of the distribution which matches the new observation are as follows:
$$\mu_t = (1-\rho)\mu_{t-1} + \rho X_t$$
$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(X_t-\mu_t)^T(X_t-\mu_t)$$

Then, the second learning rate[3], ρ:
$$\rho = \alpha\eta(X_t|\mu_k,\sigma_k)$$

2.2. Background Model Estimation

Start by determining which of the Gaussians of the mixture are most likely produced by background processes. Note that the variance of the moving objects in the scene is expected to remain larger than a background pixel until the moving object stops moving. Therefore, we look for the Gaussian distributions which have the most, supporting evidence and the least variance to determine the background model.

Denote T as the measure of the minimum portion of the data that should be accounted for by the background

Denote B as the first B distributions that are chosen for the background model
$$B = argmin_b\left(\sum_{k=1}^{b}\omega_k > T\right)$$

The value of T will be chosen by the user. If T is a small value then the background model is usually unimodal, whereas if T is a large value then the background model is usually multi-modal distribution.

2.3. Connected Components

Now that foreground pixels are identifiable in each frame, the pixels can be clustered into motion regions. This can be done using a connected component criterion as mentioned in the paper “Moving target classification of and tracking from real-time video” or by using a two-pass, connected components algorithm [3].

2.4. Target Classification

Temporal consistency can be used to differentiate between objects of interest and background clutter. If a target persists over time, it is likely that it is an object of interest for classification, otherwise it is considered background clutter.

A version of Maximum Likelihood Estimate (MLE) is used to make the classification decision. The classification metrics are collected until a statistical decision can be made about the classification of the target. This technique is ideal since it is very computationally inexpensive and reasonably effective in classifying humans and vehicles. Furthermore, this method can be trivially implemented in a microcontroller or programmed into a low powered security camera.

The idea is based on the knowledge that human bodies have extruding body parts such as limbs and usually have more complex shapes than vehicles. Because the shape of the human body is more complex, pedestrians will have a larger dispersedness than a vehicle.

Figure 1: Typical dispersedness values for a human and a vehicle [1].

Dispersedness is calculated using the following formula:

$$Dispersedness = \frac{Perimeter^2}{Area}$$

The bi-variate classification can be trained with a large dataset to improve the decision making. A linear segmentation was calibrated to produce the best possible accuracy in classification.

2.5. Remaining Steps

Since the purpose of this paper is to evaluate the proposed hypothesis, the description of the method will end here. For more details about the remaining steps and tracking, refer to the paper “Moving target classification of and tracking from real-time video” [1] and “Adaptive background mixture models for real-time tracking” [2].

2.6. Remaining Steps

As mentioned above in the previous sections, a portion of the background subtraction algorithm devised by Stauffer and Grimson was used to detect foreground pixels. The source code for the adaptive background subtraction algorithm was obtained from MathWorks Inc [4]. A binary image is produced for each frame of the video that has only two possible values for each pixel (1 or 0). The binary image is used as a mask, to subtract all the background clutter and only marks pixels making up the moving objects.

The detected pixels were then clustered into motion regions. Each cluster was then cropped out to be individually processed by the object classification techniques devised by Lipton el al to calculate its area and dispersedness of each frame. These values were then compared to a predetermined line segmentation equation. Any points to the left of the segmentation line was classified as a human, whereas the points to the right of the segmentation line was classified as a vehicle.

The classification of each detected target was displayed as a label on a bounding box that highlighted each moving target. The bounding box are outlines of the cropped regions that were processed by the algorithm. The classification is continuously updated as the scene progresses. Finally, the output frames of the mask (consisting of binary images) and the original video with target classification is produced along with a scatter plot of the dispersedness vs area are displayed. The source code for the classification algorithm was not available online and had to be created to extend on the adaptive background subtraction code.

Input: A video sample

Output: A processed motion video with clustered motion regions, a processed video with bounding boxes and a dispersedness vs area scatter plot. Both output videos will have labels with object ID (order in which it appears in the scene), object classification (Human/Car/Unknown), dispersedness and area values (eg. 5: Car D:18.4 A:10444.6).

3. Empirical Evaluations

3.1. Experimental Dataset

A large dataset of surveillance video footage was necessary to conduct the experiment of the purposed algorithm. Ideally the videos should be from stationary cameras with minimal camera shake. The majority of the dataset was obtained from YouTube and other video sharing websites. Lengthy and higher definition videos were reduced to approximately 30-60 seconds at 480p (640×480p) resolution to reduce the computational time of each video.

3.2. Empirical Protocol

Approximately 12 surveillance videos, between 1-2 minutes in length were tested with the implemented algorithm. The segmentation line equation was calibrated for each video to produce the most accurate classification. The segmentation line had to be slightly adjusted for

different camera positions and distances from the targets. For the majority of use cases the most accurate segmentation line equation was determined to be:

dispersedness = (0.0171)*area + x

where, x is typically a number between [-40 +40].
As mentioned previously, all points to the left of the segmentation line is likely to be humans, whereas all points to the right on the segmentation line was likely to be vehicles. A third classification, 'Unknown', was added. It is used to label targets that are likely to be false positive detections caused by the Gaussian mixture model, when either the area or perimeter of the target reaches the value of zero. This is prominent when a target stops moving abruptly (its binary masks starts to fade away) or the target exits the frame. More detail will be discussed in the Experimental Analysis section of this paper.

3.3. Experimental Results

Here we can see a sample screenshot of the output video from the combined algorithm.

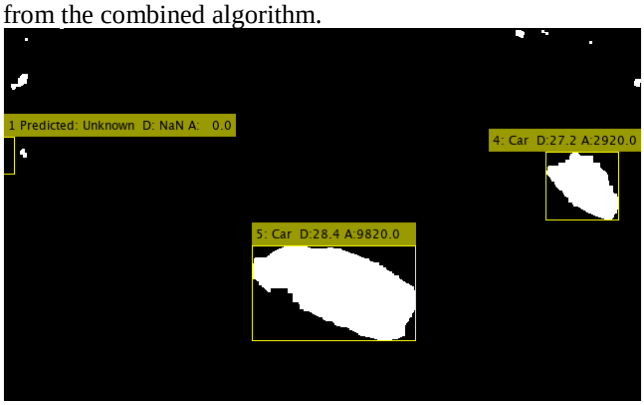

Figure 2: Binary image of foreground targets

Figure 3: Frame with bounding boxes and classification information

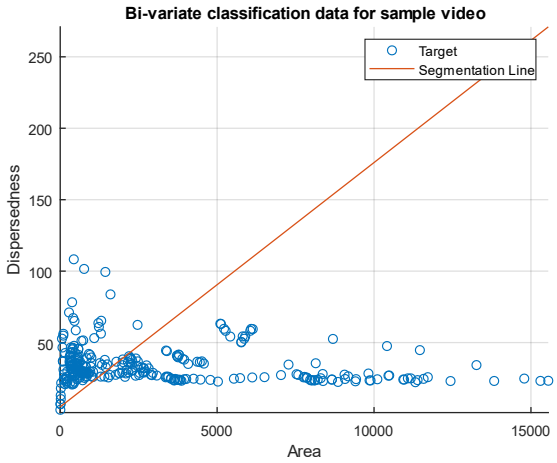

Figure 4: Bi-variate classification data for sample video (parking1.mp4)

In Figure 4, each blue circle represents one target per frame. The segmentation line equation used above was:
dispersedness = (0.0171)*area + 5.0

Approximately 12 similar videos were experimented using the proposed algorithm, results are enumerated about the table below.

Table 1: Result Table

Classification Category	Total Successful classification	Total Failed classification
Humans	21	9
Cars	88	16

As expected, the proposed algorithm was successful at classifying moving targets most of the time. Further analysis of the failures needs to be conducted.

3.4. Experimental Analysis

Table 1 shows that the results were very positive as the proposed algorithm successfully classified 70% of human figures and 85% of vehicles. Additionally, stationary objects and vehicles were correctly ignored. By analyzing the cases where failures were occurring, we can better understand the potential pitfalls of the proposed algorithm. First, targets must constantly be moving to be detected. When targets pause for a brief moment it fades into the background in a matter of a few frames, such as the red SUV in the foreground of Figure 6. As the red SUV fades, the area and perimeter decrease to zero, momentarily labelled as 'Unknown'. Of course, the rate can be adjusted in the algorithm, however there is a tradeoff between persisting targets and increase in background clutter.

Next, when vehicles are relatively far away and at an angle of approach, sometimes it may resemble a human

figure, as shown in Figure 5 and Figure 12. The target labeled as ID 9 is classified as a human briefly because of the angle of its approach relative to the camera and the distance away from the camera which makes it resemble a human figure, much like ID 10. Additionally, as targets move out of frame, its metrics can become momentarily distorted causing the target to be misclassified or labeled as 'Unknown'. An example of this can be seen in Figure 2 and Figure 3, ID 1 where there was a vehicle moving out of frame.

In Figure 6 and Figure 7, we can see a false negative detection of a human figure near the bottom left of the frame. This could be fixed with further calibration of the detection mechanism. To test for noise, a video of a dimly illuminated scene was used and Gaussian noise was added. Furthermore, this particular video footage had some camera shaking caused by the wind. We can see something interesting here as the shadow of the human target is clearly detected in Figure 8, which slightly skews the measurements here. However, in Figure 10, the shadow is no longer detected when Gaussian noise was added. This is one of the benefits of using an adaptive nonparametric Gaussian mixture model. It handles noise extremely well. Overall the proposed algorithm was successful in detecting and classifying the objects on the scene despite the noise and camera movement.

Figure 6: Binary image of foreground targets with classification information

Figure 7: Frame with bounding boxes and classification information

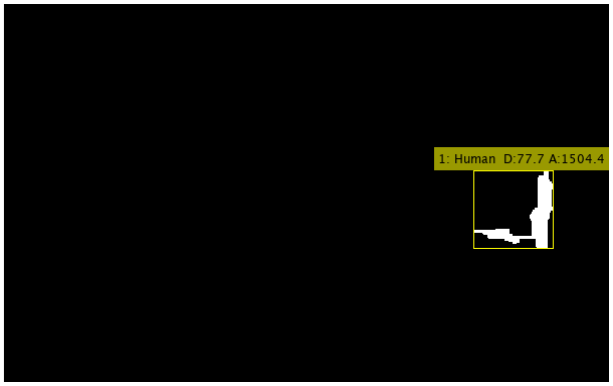

Figure 8: Binary image of foreground targets with bounding boxes and classification information

Figure 9: Frame with bounding boxes and classification information

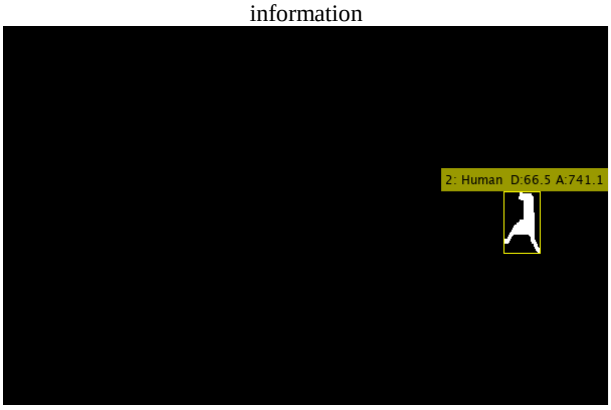

Figure 10: Binary image with bounding boxes and classification information of noisy video

Figure 11: Frame with bounding boxes and classification information of noisy video

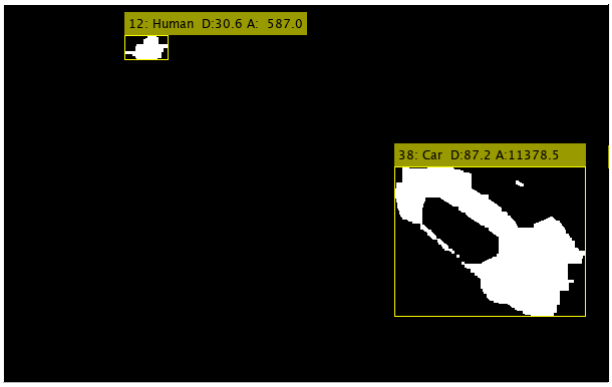

Figure 12: Binary image of noisy video

Figure 13: Frame with bounding boxes and classification information of noisy video

4. Conclusion

Noise and low illumination indeed affect object detection and classification. The proposed algorithm involves combining the adaptive nonparametric Gaussian mixture model with the object classification technique. Using a variety of test videos of different scene conditions, a deep analysis was conducted. Although, there were many pitfalls outlined in this paper, for the most part, the proposed algorithm was very successful in detecting and classifying moving targets in each scene. Moving forward, the effectiveness of the proposed algorithm could be improved with further calibration and additional testing.

References

[1] A. J. Lipton, H. Fujiyoski, and R. S. Patil, “Moving target classification of and tracking from real-time video,” in *Proc. IEEE Workshop Applications and Computer Vision*, Princeton, NJ, 1998, pp. 8–14

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 1999, pp. 246-252 Vol. 2.

[3] B. K. P. Horn. *Robot Vision*, pp. 66-69, 299-333. The MIT Press, 1986.

[4] MathWorks Inc. “Foreground detection using Gaussian mixture models” *MathWorks*, https://mathworks.com/help/vision/ref/vision.foregrounddetector-system-object.html