Change Detection by Frequency Decomposition: Wave-Back

Fatih Porikli Christopher R. Wren
Mitsubishi Electric Research Laboratories
Cambridge, MA, 02139, USA

Abstract

We introduce a frequency decomposition based background generation and subtraction method that explicitly harnesses the scene dynamics to improve segmentation. This allows us to correctly interpret scenes that would confound appearance-based algorithms by having high-variance background in the presence of low-contrast targets, specifically when the background pixels are well modeled as cyclostationary random processes. In other words, we can distinguish near-periodic temporal patterns induced by real-world physics: the motion of plants driven by wind, the action of waves on a beach, and the appearance of rotating objects. To capture the cyclostationary behavior of each pixel, we compute the frequency coefficients of the temporal variation of pixel intensity in moving windows. We maintain a background model that is composed of frequency coefficients, and we compare the background model with the current set of coefficients to obtain a distance map. To eliminate trail effect, we fuse the distance maps.

1 Introduction

Background subtraction is the most common approach for discriminating a moving object in a *relatively* static scene. Basically, a reference model (background) for the stationary part of the scene is estimated and the current image is compared with the reference to determine the changed regions (foreground) in the image.

Existing methods can be classified as either single-layer or multi-layer methods. Single-layer methods construct a model for the color distribution of each pixel based on the past observations. Wren [6] proposed a single unimodal, zero-mean, Gaussian process to describe the uninteresting variability. The background is updated with the current frame according to a preset weight, which adjusts how fast the background should be blended to the new frame. However, it is shown that such a blending is sensitive to the selection of the learning factor. Depending its value, either the foreground may prematurely blended into the background, or the model becomes unresponsive to the observations. Toyama [5] preferred an autoregressive model, Kalman filter, to capture the properties of dynamic scenes. The various parameters of the filter such as the transition matrix, the process noise covariance and the measurement noise covariance may change at each time step but are generally assumed to be constant. Another drawback of the Kalman filter is its inability to represent multiple modalities.

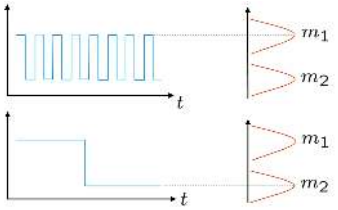

Figure 1: Model based methods, which neglect temporal correlation, cannot differentiate the above sequences.

Stauffer and Grimson [2] suggested to model the background with a mixture of Gaussian models. Rather than explicitly modeling the values of all the pixels as one particular type of distribution, the background is constructed by a pixel-wise mixture of Gaussian distributions to support multiple backgrounds. Instead of mixture approach, Porikli and Tuzel [4] presented a multi-modal background algorithm that models compete each other. Elgammal [1] used a non-parametric approach, where the density at a particular pixel was modeled by Gaussian kernels. Mittal and Pragios [3] integrated optical flow in the modeling of the dynamic characteristics.

A major shortcoming of all the above background methods is that they neglect the temporal correlation among the previous values of a pixel. This prevents them detecting a structured or periodic change, for example shown in Fig. 1. This is often the case, since real-world physics often induces near-periodic phenomenon in the environment: the motion of plants driven by wind, the action of waves on a beach, and the appearance of rotating objects.

The main contribution of this work is an algorithm, called as *Wave-Back* that explicitly harnesses the scene dynamics to improve segmentation. We generate a representation of the background using the frequency decompositions of the pixel's history. For a given frame, we compute the Discrete Cosine Transform (DCT) coefficients and

compare them to the background coefficients to obtain a distance map for the frame. Then, we fuse the distance maps in the same temporal window of the DCT to improve the robustness against the noise and remove the trail artifacts. Finally, we apply a threshold to the distance maps to determine the foreground pixels.

2 Cyclostationarity

The goal of the segmentation algorithm is to decide if the samples are drawn from the background distribution, or from some other, more interesting distribution. By assuming independent increments, the existing algorithms are relying completely on the "current" appearance of the scene. Let's examine the case of a tree blowing in the wind. The multi-modal background models would build up separate modes to explain, say sky, leaf, and branch appearances. As the tree moves, the individual pixel may image any of these. The independent increments assumption says that these different appearances may manifest in any order. However, we know that the characteristic response places constraints on the ways that the library of appearances may be shuffled.

Specifically, given two samples from the observation process: $x[n]$ and $x[m]$, the independent increments assumption states that the autocorrelation function $R_x[n,m]$ is zero when $n \neq m$: $R_x[n,m] \stackrel{\triangle}{=} E\left[x[n]x^*[m]\right]$. This is correct when the process is stationary and white: such as a static scene observed with white noise. For a situation where the observations are driven by some physical, dynamic process, we can expect that the dynamics will leave their spectral imprint on the observation covariance. For instance, if the process is periodic, then we would expect to see very similar observations occur with a period of N samples, in contrast to the previous model:

$$R_x[n, n+N] \neq 0$$

We say that this process is cyclostationary if the above relationship is true for all time. A process is said to be harmonziable if its autocorrelation can be reduced to the form $R_x[n-m]$, that is, so that the autocorrelation is completely defined by the time difference between the samples. It is possible to estimate the spectral signature of harmonziable, cyclostationary processes in a compact, parametric representation.

3 Frequency Decomposition

The Discrete Cosine Transform (DCT) is conceptually similar to the Forurier transform except it decomposes the signal into a weighted sum of cosines instead of sinusoidal frequency and phase content. Given data $x[n]$, where n is an integer in the range $0, .., N-1$, the DCT is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi n(2n+1)}{2N}\right).$$

Note that, the DCT coefficients of similar waveforms are equal. Furthermore, if the waveforms are not similar, the coefficients will be different even if the pixel values at that time instants are equal.

There are several advantages of the DCT over the DFT. The DCT does a better job of concentrating energy into lower order coefficients than does the DFT, thus it enables definition of more stable distance measures for change detection by reducing the effects of the high frequency components in the decomposition. The DCT is purely real, on the other hand, the DFT is complex (magnitude and phase). An N-point DCT has the same frequency resolution as and is closely related to a $2N$-point DFT. Assuming a periodic input, the magnitude of the DFT coefficients is spatially invariant (phase of the input does not matter). On the other hand, we can recover phase differences by using the DCT.

We begin by accumulating sample sequences, $x_t[n]$, for each pixel from a number of frames of video as $x_t[0] = x_t, x_t[1] = x_{t-1}, ...$ where x_t is the value of the pixel in the current frame, x_{t-1} in the previous frame, etc. Each of these sequences serves as an example of the periodic behavior of a particular pixel in the image. These sequences are used to initialize the background model for each pixel. We compute the DCT coefficients $X_t[k]$ using the accumulated $x_t[n]$. We take this as an estimate of the spectral components in the autocorrelation function of the underlying scene process. For each new sample x_{t+1}, we construct a sample sequence $x_{t+1}[n]$ and extract a new set of DCT parameters, $X_{t+1}[k]$. We take this to represent the process underlying the current observations.

To determine if these two samples sequences were generated by the same underlying process, we compute the L_2-norm of the difference between the DCT coefficients:

$$d(t) = \langle X_t[k], X_{t+1}[k] \rangle = \left(\sum_{k=0}^{N-1} (X_t[k] - X_{t+1}[k])^2 \right)^{\frac{1}{2}}$$

Small distances are taken to mean that the samples are drawn from the same process, and therefore represent observations consistent with the scene.

Since the signals we encounter are almost never truly stationary, we add a simple exponential update mechanism to the above algorithm. This consists of combining the current estimate of the DCT coefficients with the estimate of the scene's DCT coefficients:

$$Y_t[k] = \alpha Y_{t-1}[k] + (1-\alpha) X_t[k]$$

where α is the exponential mixing factor that we set to 0.998 in our experiments. Here, $Y_t[k]$ stands for the background,

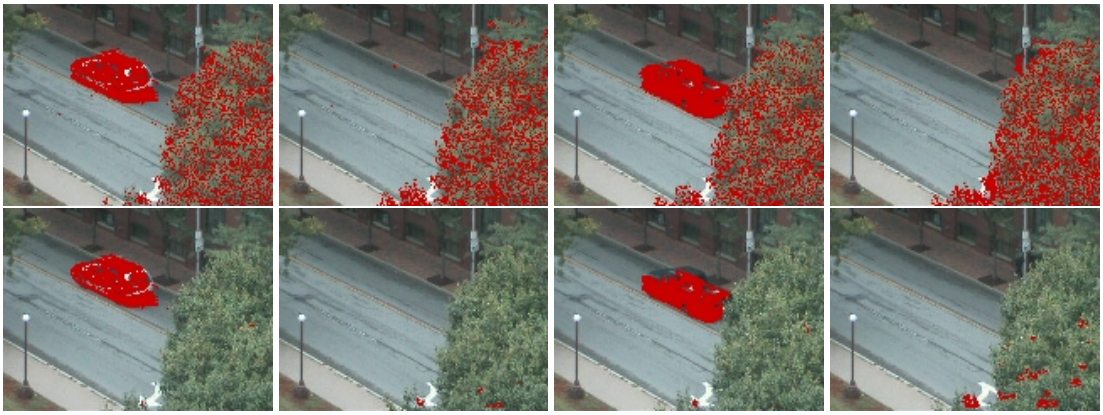

Figure 2: Background subtraction results. **Top:** unimodal approach. **Bottom:** DCT based method.

and in our implementation we compare the current DCT coefficients with the background, i.e. $d(t) = \langle X_t[k], Y_t[k] \rangle$.

The length of the window, N, is a parameter that must be chosen with care. If the window is too small, then low-frequency components will be poorly modeled. However large windows come at the cost of more computation, more lag in the system, and a trail effect. The segment size controls the tradeoff between frequency resolution and time resolution. Choosing a wide window gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution.

The sample sequence $x_t[n]$ includes the values of the pixel in the previous frames. Thus, the pixel values in the past within the temporal window will still contribute to the DCT coefficients of the following frames by causing a drift problem as shown in Fig. 3. Depending on the speed of the object and the window size N, the computed distances will contain a trail of the changed DCT coefficients behind the object. In case the temporal window size is larger, the length of the trail will be longer. As we mentioned above, if we use a wide window then we can model lower frequency cyclostationary behavior, however the amount of the trail will be significant. A narrower window will minimize the trail, however, it will limit the detectable frequency of the temporal changes.

To minimize the trail, we accumulate the distances as

$$d^*(t) = \prod_{m=0}^{M-1} d(t-m)$$

where M is adjusted depending on the amount of the overlap between the regions of moving object in the frames within the temporal window. Thus, M is correlated with the speed and size of the object. A smaller value should be assigned in case the object has smaller overlapping regions. In the shown traffic sequence, we assigned it to the window size as $M = N$ since most of the objects have N frames overlaps. In addition, it is possible to apply a spatial smoothing to the DCT coefficients to improve the robustness against the noise and trail.

4 Results

We tested the Wave-Back algorithm on 3000 frames of a traffic video sequence in which a tree sways drastically due to the high wind in the recording time. The tree occupies almost one-third of the image and causes significant motion (5-20 pixels), which is certainly a severe distraction for any background segmentation algorithm.

We compared several versions of the Wave-Back algorithm using a 16-point DCT algorithm. We also tested unimodal and multi-modal background subtraction algorithms. In an attempt to most directly demonstrate the performance of the background models, we present change detection results in Fig 2. We determined the pixels that have higher distance scores than a given threshold. To make a fair comparison, we tuned the unimodal algorithm (single Gaussian model) using false/miss ROC curve such that the threshold gives the minimum amount of the misclassified pixels.

Figure 4 shows miss detection vs. false alarm graphs for three algorithms: the alpha-blending, a multi-modal model, and the Wave-Back with window size $N = 16$. To obtain these results, we manually segmented ground truth foreground and background regions. The graphs correspond to

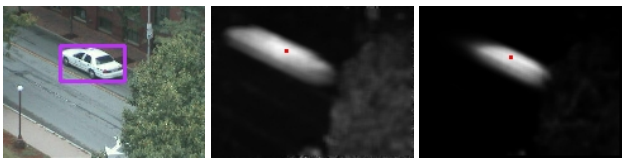

Figure 3: Distance maps before and after trail removal.

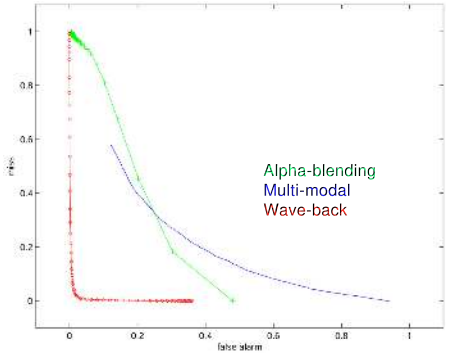

Figure 4: ROC graphs of miss (foreground pixels that are classified as background) versus false alarm (background pixels classified as foreground).

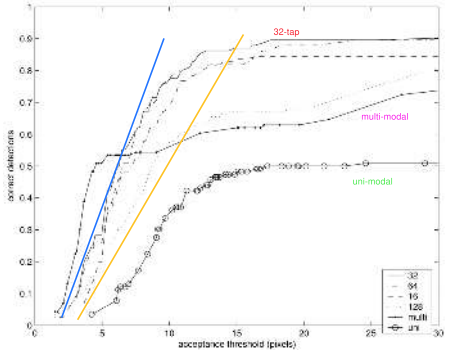

Figure 5: Correct detection graphs (IR sequence). Frequency decomposition method has higher accuracy, however, using temporal window causes 2-12 pixels long trail, which is the lagged region between the blue and yellow lines.

the ROC curves for the percentage of the misclassified foreground pixels vs. misclassified background pixels. Closer to a curve to the lower right hand corner (zero-error percentages) indicates better detection performance. It is visible that the 16-point Wave-Back performs the best among the three algorithms, achieving almost a 2% false alarm rate at a 2% missed point rate, and accurately detecting a 99% of the foreground pixels while detecting 97% of the background as well. We explain the very high accuracy of the proposed algorithm by the fact that it can classify the swaying tree as a part of the background, thus it gives smaller error rates of false detection. For the IR sequence (Fig. 6), we estimated the position of the boat by finding the maximum distance at each frame. We compared the estimation result with the ground truth and accepted the estimation as a correct result if the distance from the ground truth is less than an acceptance threshold. Figure 5 shows the accuracy of the estimation depending on the threshold. We observed that the Wave-Back method has higher much correct detection ratio, and a lag due to the trail effect, which can be eliminated by the proposed fusing method.

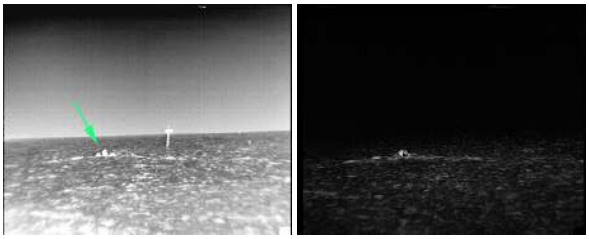

Figure 6: **Left:** A frame from IR sequence (*Courtesy of PETS 2005*) that shows a small boat moving in the waves. **Right:** Corresponding distance map by Wave-Back.

In terms of the computation time, the Wave-Back method using 16-point windows is comparable with the multi-modal approach, which can run in real-time for a 320x240 color video sequence.

5 Conclusion

We have presented a novel algorithm that detects new objects based solely on the dynamics of the pixels in a scene, rather than their appearance. This is accomplished by directly estimating models of cyclostationary processes to explain the observed dynamics of the scene and then comparing new observations against those models. We have presented results that demonstrate the efficacy of this algorithm on challenging video.

References

[1] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction", *In Proc. of European Conf. on Computer Vision*, pp. 751-767, 2000

[2] C. Stauffer and W.Grimson, "Adaptive background mixture models for real-time tracking", *In Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 1999

[3] A. Mittal and N. Paragios, "Motion-Based background subtraction using adaptive kernel density estimation", *In Proc. Int'l Conf. on Computer Vision and Pattern Recognition*, 2004

[4] F. Porikli, O. Tuzel, Human body tracking by adaptive background models and mean-shift analysis, *In Proc. of Int'l Conf. on Computer Vision Systems, Workshop on PETS*, 2003.

[5] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, "Wallflower: Principles and Practice of Background Maintenance", *In Proc. of Int'l Conf. on Computer Vision*, pp. 255–261, 1999

[6] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland "Pfinder: Real-time tracking of the human body", *In PAMI*, 19(7):780-785, July 1997