A Transform Domain Approach to Real-Time Foreground Segmentation in Video Sequences

Juhua Zhu, Stuart C. Schwartz, Bede Liu

Dept. of Electrical Engineering, Princeton University
E-Quad, Olden St., Princeton 08544, U.S.A.
E-mail: {juhuazhu, stuart, liu}@ princeton.edu

Abstract

Accurate foreground segmentation is a difficult task due to such factors as illumination variation, occlusion, background movements, and noise. In this paper we present a novel adaptive transform domain approach for foreground segmentation in video sequences. A set of DCT-based features is employed to exploit the spatial and temporal correlation in the video sequences. We maintain an adaptive background model and make a decision based on the distance between the features of the current frame and that of the background model. Additional higher level processing is employed to deal with the variation of the environment and to improve the accuracy of segmentation. The approach is shown to be insensitive to illumination change and to noise. It also overcomes many common difficulties of segmentation such as foreground aperture, and moved background objects. The algorithm can perform in real-time.

1. Introduction

The task of foreground segmentation is to label the regions in an image as moving objects or background. It is the fundamental step in many vision systems including video surveillance, human-machine interface, and very low-bandwidth telecommunications. The challenges facing the segmentation task [11] include: illumination variation, background change, foreground aperture, bootstrapping, camouflage, and shadows. In addition, a complex algorithm may be difficult to implement for real time operation.

Many approaches have been proposed to segment the foreground moving objects in video sequences [1]-[11]. To make the algorithm robust to a change in illumination or in the background, adaptive background modeling approaches have been proposed. Kalman Filtering based methods [8] recover slowly to sudden lighting change and do not handle bimodal backgrounds well [9]. The use of Mixture of Gaussians (MoG) model [9] also adapts slowly to sudden lighting changes. Hidden Markov Model has been used to describe global state changes [10]. The Wallflower system [11] attempts to solve many of the problems with background maintenance.

All the above approaches use the intensity or color information of the pixels, and are susceptible to sudden lighting changes. Recently, efforts have been made to incorporate other illumination-robust features for scene modeling. The intensity and texture information can be integrated for change detection, with the texture-based decision taken over a small neighborhood [7]. The color and gradient information can be fused and robust results have been reported [5]. However the computation of these algorithms is often too intensive for real time implementation.

The temporal correlation in the video suggest modeling the evolution of the pixel values as a parameterized random process, while a block based, rather than pixel based, modeling can take advantage of the spatial correlation. A useful observation is that the background illumination lies mainly in the low frequency range, while the presence of a foreground object in a nearly fixed scene usually produces significant difference in both low-frequency and high-frequency band. Therefore a frequency domain approach is particularly suitable for segmenting the foreground from the background. DCT coefficients were used to combat shadows [1], and to build competing HMM models to handle persistent changes [6].

In this paper, we propose a robust segmentation approach using easily computable intensity and structure features derived from DCT coefficients. By using flexible adaptation of a single Gaussian model to environmental changes, a short or even no training can lead to good performance under difficult conditions, as demonstrated by our experimental results.

2. The proposed method

Each frame of the video sequence is divided into 8x8 blocks, and two features, f_{AC} and f_{DC}, are extracted from the DCT coefficients and each is modeled as an independent Gaussian distribution with different mean μ and variance σ^2. μ and σ^2 are estimated initially from a few frames, possibly with some moving objects, and then are updated depending on the scene evolution. Segmentation of a new frame is obtained initially by thresholding the feature values, and then refined using size filtering, illumination change detection and handling, and filling. Based on the final segmentation of the current frame, the parameters μ and σ^2 are updated for the segmentation of the next frame.

Block-based features incorporate local neighborhood information, so the algorithm using these features is less sensitive to noise and to small scene changes than those using pixel values. This can be seen in figure 1, where the same modeling method works much more robust on block-based features than on pixel values. The use of DCT based features also facilitates compressed domain processing.

0-7803-8874-7/05/$20.00 ©2005 IEEE

2.1 Features

Two features are derived from the DCT coefficients of the luminance component of each block:

$$f_{DC} = DCT(0,0)$$

$$f_{AC} = \sum_{i=0}^{3}\sum_{j=0}^{3}(i^2 + j^2)\cdot|DCT(i,j)|$$

where $DCT(i,j)$ is the $(i,j)^{th}$ coefficient. The DC feature f_{DC} is simply the DC coefficient, and the AC feature f_{AC} is a weighted sum of the low frequency AC coefficients. The high frequency coefficients are not used because they are more sensitive to noise and they relate more to fine details that are more susceptible to small illumination changes. The weight $i^2 + j^2$ de-emphasizes the very low frequency components, which often correspond to smooth transition of intensity, e.g. shadows. So this scheme helps avoiding shadows being taken as part of foreground object, while the de-emphasized low frequency components are still able to help pick out foreground blocks in the case of camouflage.

2.2 Modeling of f_{DC} and f_{AC}

A key step in segmentation is the background modeling. Various approaches have been proposed for modeling video in both spatial and time domain, as well as modeling features derived from video. The choice of a good model must balance various factors, including what quantities are to be modeled, computation requirement, speed of adaptation, and how it affects the final performance of the particular application. For modeling pixel values over a few hours, the use of Mixture of Gaussians (MoG) has been shown to give a better fit than the use of a single Gaussian, even without sudden lighting change [4]. However, we may not need to take into consideration the samples collected over a time period far away from the time instance under discussion, especially when the underlying random process is non-stationary, which is typically the case for surveillance scenario. And we may not need to care much about the small amount of improvement in terms of fitting error brought by adding one modal in the probability distribution function because the presence of foreground objects almost always bring along significant difference from the background distribution. In addition, MoG is slow to adapt to fast changing environment [10], and requires more computation and memory.

In this paper, we use features f_{DC} and f_{AC} of the blocks, instead of the pixel values. We model each feature by a single non-stationary Gaussian random process. That is, with parameters that are varying with time. This is described in subsequent sections. Extensive experiment was carried out using video captures under different conditions. We have concluded that the use of a single non-stationary Gaussian model with a flexible policy allowing for fast and slow adaptation of parameters can accurately model the evolution of these features and ultimately can lead to accurate foreground segmentation under challenging conditions, including repetitive background movement.

2.3 Background model and update

We assume f_{DC} and f_{AC} of each block follow independent normal distribution with means μ_{AC} and μ_{DC}, and variance σ_{AC}^2 and σ_{DC}^2, respectively. That is,

$$f_{AC} \sim N(\mu_{AC}, \sigma_{AC}^2)$$

$$f_{DC} \sim N(\mu_{DC}, \sigma_{DC}^2)$$

Initially, these parameters are estimated from a limited number of training frames, e.g. 20 frames, using robust estimators to reduce the effect of possible outliers:

$$\mu_0 = med(X)$$

$$\sigma_0 = 1.4862 \times (1 + 5/(N-1)) \times med(abs(X - med(X)))$$

where *X* stands for the set of samples of either f_{DC} or f_{AC}, *N* is the number of samples, *med* denotes the median operation and *abs* means the absolute value.

After the segmentation of frame $t-1$, the parameters of the background model for the next frame, frame *t*, are updated using single exponential smoothing.

$$\mu_t = (1-\alpha_\mu)\mu_{t-1} + \alpha_\mu \cdot x$$

$$\sigma_t^2 = (1-\alpha_\sigma)\sigma_{t-1}^2 + \alpha_\sigma \cdot (x - \mu_t)^2, t \ge 1$$

where x denotes either f_{DC} or f_{AC}, and α_μ and α_σ are parameters controlling the learning speed.

At the time of severe change of the environment, it is not fair to update the variance as in the above equations, for the distribution is in fact quite different from the original one (before change occurs), and the mean value used in the variance update equation may be very distinct from the real value due to the delay of its own adaptation. In our current implementation, the new variance value is learned from a few frames as in the training period.

2.4 Initial Segmentation

Once the four parameters μ_{AC}, μ_{DC}, σ_{AC}^2, and σ_{DC}^2, for all blocks of a frame have been determined, initial segmentation is performed by classifying a block as foreground, if, for that block,

$$|f_{AC} - \mu_{AC}| > T_{AC} \qquad (*)$$

or

$$|f_{DC} - \mu_{DC}| > T_{DC} \qquad (**)$$

where T_{AC} and T_{DC} are the threshold values. We call those blocks satisfying (*) "AC foreground blocks" and those satisfying (**) "DC foreground blocks", respectively. A foreground object in general will have both of these blocks. Equation (**) is used because typically there is a large difference in the DC value of the intensity between a moving foreground block and the background block. The use of equation (*) is because the presence of the edge of the foreground objects and the different texture between the foreground and the background will lead to a large difference between the AC values of the current block and those of the background model. The use of these two thresholding operations is equivalent to using both intensity and texture information, and will likely produce more robust and reliable segmentation results.

To handle the effect of small lighting change, we use for the threshold a value derived from both the sample standard deviation and the sample mean:

$$T_{AC} = \delta_{AC} \times \mu_{AC} + 2.5 \times \sigma_{AC}$$

$$T_{DC} = \delta_{DC} \times \mu_{DC} + 2.5 \times \sigma_{DC}$$

where δ_{AC} and δ_{DC} are parameters, whose value depends on how much lighting change is expected to be handled here. It has been found that $\delta = 0.25$ leads to good results.

When training is not allowed or when re-training for variance is taken at the time of severe environment change, i.e. when segmentation should be done without an estimate of the variance, we simple raise δ by a factor of 2.

2.5 Learning parameters

There is a tradeoff between stability and adaptation speed when choosing a learning parameter. When the scene is slowly varying, a small α is preferred as it avoids the impact of outliers. When there is a change of either the camera or the environment, a large α is preferred in order to quickly arrive at a new background model. For the parameter α_μ that controls mean value adaptation, we set a range $[\alpha_{\min}, \alpha_{\max}]$ for it and use the scheme as

$$\alpha_\mu = \begin{cases} 0 & \text{foreground block} \\ \alpha_{\min} & \text{background block} \\ \max(\alpha_{\min}, (0.5)^n \cdot \alpha_{\max}) & \text{fast update needed} \end{cases}$$

where n is frame number starting from when fast update is needed, such as global or local lighting change or moved background objects. We identify a global lighting change if a large portion of the scene is classified as foreground in the initial segmentation result. We identify local lighting change and moved background objects by the persistence of certain groups of 8-connected neighboring blocks classified as foreground.

For the parameter α_σ that controls variance update, as mentioned previously, it does not take effect when fast update is applied to α_μ.

2.6 Size Filtering

After the initial segmentation, all blocks of the frame are labeled 1 (foreground) or 0 (background). To remove false positives, size filtering is applied to each separate blob. Specifically, a blob is removed if: a) its size is smaller than a given value, or b) the ratio of the number of AC blocks to that of DC blocks is less than a given value, or c) the ratio is greater than a certain value. This is not only very effective to remove sporadic false alarms as the usual size filtering scheme does, but also effective in combating local lighting change and repetitive movement, because local lighting change will usually only result in DC blocks, while repetitive movement will only result in AC blocks. Filling can be applied to remove some false negatives, since most interesting objects are compact.

3. Experimental results

We tested our method using sequences representing the typical challenges in foreground segmentation. The four videos are available at:
http://www.princeton.edu/~juhuazhu/Acad/Demo.htm.

The first sequence features a highly reflective wall in the background. The mirror like background shows a considerable number of moving shadows from objects far away from the wall. In addition, the background contains a large area looking very much like the skin-tone in intensity images. There is a walking person wearing trousers with similar color as that of the background. Our method results in very few false negatives and almost no false positives. Two sample frames are shown in Fig. 1. By comparison, the result of the same frames by using a single Gaussian model for each pixel value is shown in the left column. Due to the very short time duration for both training and segmentation, and the very little lighting change, it does not make much sense to use multiple modals. As can be seen from the result, modeling single pixel value fails shortly. If the thresholds are raised higher, there will be more false negatives; if they are lower, a lot of false positives. While the simultaneous modeling of block-based features leads to satisfactory result, with all the parameters kept unchanged, without any adaptation.

The second sequence includes severe and frequent global and local illumination changes, very strong lighting at some locations, some mirror-like background parts, and some cluttered background regions. This is a very difficult environment for foreground segmentation. But our approach produces good segmentation. Sample frames are shown in Fig. 2.

The third sequence contains cluttered background of trees, swaying branches, and a person wearing clothes with the color similar to the background. The swaying branches are detected only when the person shakes the branches violently. But it recovers almost immediately when the shaking stops.

The fourth sequence contains a moving car and a walking person in an outdoor setting. Object sizes vary considerably and the car blocked the person briefly. Our segmentation results show no foreground aperture problem typically present for homogeneously colored cars, and the recovery after occlusion is fast.

In our tests, the number of false positives and false negatives are low. Although some false alarms appear due to a sudden lighting change, they do not persist as our system quickly adapts. Our present C++ code, without optimization, can handle 36 frames per second for a frame size of 352x240 pixels on a PIII 900M HZ machine. It appears that our method can handle well most of the difficulties of real-time foreground segmentation.

4. Summary

This paper presents a novel approach for foreground segmentation for video sequences. It uses two features derived from the DCT of each 8x8 block. A single Gaussian distribution is used to model the evolution of the two features in each block. An updating process is designed to handle effectively the changing background. The judicious choice of features makes our method insensitive to noise and light shadows. The method is also able to successfully handle the problems of gradual and sudden illumination changes, global and local illumination changes, moved background objects, small repetitive background movement, and foreground aperture.

References

[1] N. Amamoto and A. Fujii, "Detecting Obstructions and Tracking Moving Objects byImage Processing Technique," Electron. and Comm. In Japan, Part 3, Vol. 82, No. 11, 1999, pp. 28-37.

[2] T. E. Boult, R. J. Micheals, and X. Gao, “Into the Woods: Visual Surveillance of Noncooperative and Camouflaged Targets in Complex Outdoor Settings,” Proc. of the IEEE, Vol. 89, No. 10, pp. 1382-1402, 2001.

[3] M. Everingham, and B. Thomas, “Supervised Segmentation and Tracking of Nonrigid Objects Using a ‘Mixture of Histograms’ Model,” Proceedings of the 8th IEEE International Conference on Image Processing (ICIP2001), October 2001, pp. 62-65.

[4] X. Gao, T.E. Boult, et al, “Error Analysis of Background Adaptation”, CVPR 2000.

[5] O. Javed, K. Shafique, and M. Shah, “A Hierarchial Approach to Robust Background Subtraction using Color and Gradient Information,” Proc. Workshop on Motion and Video Computing, 2002, pp. 22-27.

[6] M. Lamarre, and J. J. Clark, "Background subtraction using competing models in the block-DCT domain," ICPR 2002.

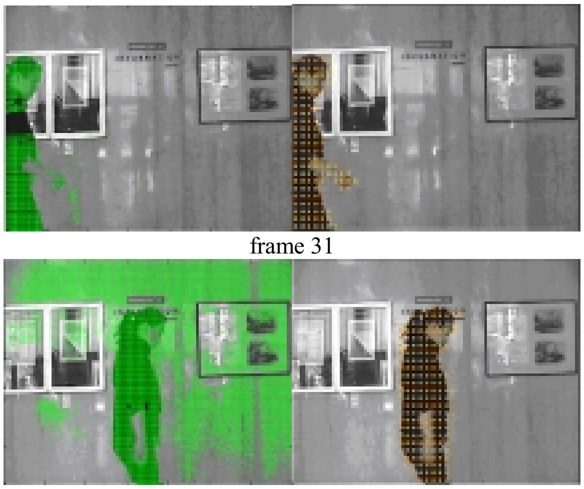
[7] L. Li, and M. Leung, “Integrating Intensity and Texure Differences for Robust Change Detection,” IEEE Trans. on Image Processing, Vol. 11, No. 2, pp. 105-112, 2002.

[8] C. Ridder, O. Munkelt, et al, “Adaptive Background Estimation and Foreground Detection using Kalman-filtering,” Proc. of Intl. Conf. On Recent Advances in Mechatronics (ICRAM), pp. 193-199, 1995.

[9] C. Stauffer, and W.E.L. Grimson, “Adaptive Background Mixture Models for Real-time Tracking,” CVPR 1999, pp. 246-252.

[10] B. Stenger, V. Remesh, et al, “Topology Free Hidden Markov Models: Application to Background Modeling,” ICCV 2001, pp. 294-310.

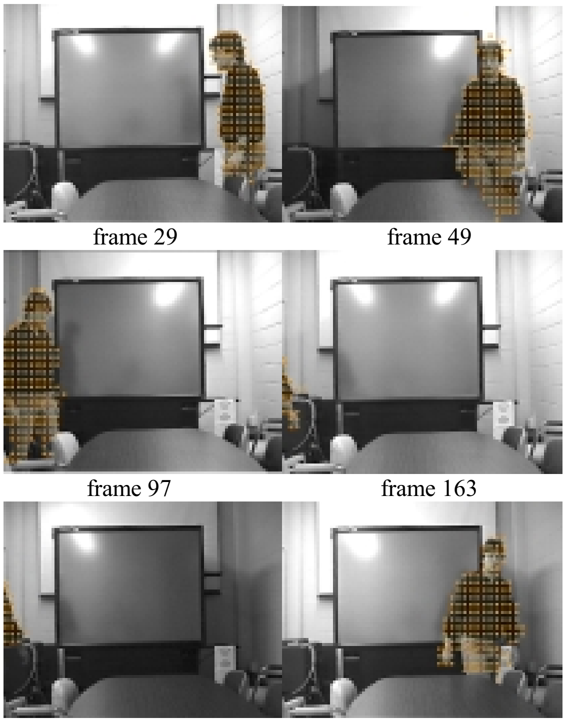
[11] K. Toyama, J. Krunmm, et al, “Wallflower: Principles and Practice of Background Maintenance,” ICCV 1999, pp. 255-261.



frame 31

frame 35

Fig. 1. Left column is the result of using only intensity of single pixels. It fails shortly. The right column is the result of proposed method. It is very robust. First 25 frames are used for background training.



frame 29

frame 49

frame 97

frame 163

frame 164

frame 201

Fig. 2. Segmentation result of proposed method. Severe lighting changes at frames 49, 163, and 164. First 20 frames are used for background training.