DETECTING MOVING OBJECTS FROM DYNAMIC BACKGROUND WITH SHADOW REMOVAL

Shih-Chieh Wang¹, Te-Feng Su², and Shang-Hong Lai¹,²

¹Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan
²Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

Background subtraction is commonly used to detect foreground objects in video surveillance. Traditional background subtraction methods are usually based on the assumption that the background is stationary. However, they are not applicable to dynamic background, whose background images change over time. In this paper, we propose an adaptive Local-Patch Gaussian Mixture Model (LPGMM) as the dynamic background model for detecting moving objects from video with dynamic background. Then, the SVM classification is employed to discriminate between foreground objects and shadow regions. Finally, we show some experimental results on several video sequences to demonstrate the effectiveness and robustness of the proposed method.

Index Terms—Background subtraction, local-patch Gaussian mixture model, moving object detection, shadow removal, dynamic background

1. INTRODUCTION

Moving object detection is one of the essential tasks in many computer vision applications, including video processing, video surveillance and traffic monitoring. A typical and efficient approach used to achieve such tasks is background subtraction. The idea behind background subtraction is to compare the current image with a reference background model, which is learned and maintained for the background along time.

Much effort has been devoted to developing efficient methods of moving object detection using background subtraction. Some of them estimated the probabilities of individual pixels belonging to background by using Gaussian Mixture Models (GMMs) [1] or labeled each pixel as foreground or background by Markov Random Fields (MRFs) [2]. An improved GMM learning algorithm [3] was proposed to select an appropriate number of components for each pixel on-line, thus fully adapting to the scene. Elgammal et al. [4] proposed to utilize a general non-parametric kernel density estimation technique to build background model for detecting foreground objects. These methods work well under the assumption that the background scene is stationary. However, they are doomed to fail for the case of dynamic scenes, which include repetitive motions like waving trees, rippling water, and camera jitters, etc. Several block-based methods were developed to overcome such problems, which usually divide an image into blocks and calculate block correlation [5] or block-specific features, such as the local binary pattern [6] histogram. However, these block-based approaches allow only coarse detection of the moving objects.

Some recent methods proposed for dynamic background subtraction utilized not only the temporal information of a single pixel but also the spatial information of neighboring pixels. Li et al. [7] extracted foreground objects from a complex video under the Bayes decision framework. A Bayes decision rule was employed for classification of background and foreground from a general feature vector. Sheikh and Shah [8] also utilized a Bayes rule to build the background model based on an MRF framework to enforce the spatial constraint and obtained better results. Zhang et al. [9] proposed a spatial-temporal nonparametric background subtraction method to effectively handle dynamic background subtraction by modeling the spatial and temporal variations simultaneously. Cheng et al. [10] formulated the moving object detection under dynamic background by minimizing a cost function that is motivated from the large-margin principle, and they used the 1-SVM to solve the foreground/background classification problem.

However, shadow is often misclassified as the foreground region in the previous background subtraction methods, and this degrades the accuracy of background subtraction and background color model. Huang et al. [11] formulated the classification problem as a graph labeling over a region adjacency graph based on the MRF framework.
In addition, Martel-Brisson and Zaccarin [12] introduced a nonparametric framework to model background based on physical properties of light sources and surfaces in RGB space. Zeng and Lai [13] used the normalized color space and the brightness gain information to segment the foreground and shadow regions.

In this paper, we propose a two-stage foreground detection algorithm from dynamic background videos. First, we propose a Local-Patch Gaussian Mixture Model (LPGMM) which is extended from the GMM [1] to represent local spatial distribution for each pixel. Next, we use a learning technique that employs SVM to learn the color, texture and intensity characteristics to discriminate foreground from shadow. Fig. 1 gives the flowchart of the proposed method.

The rest of this paper is organized as follows. In Section 2, we introduce the LPGMM background modeling method. Then, the shadow detector by SVM classifier is presented in Section 3. In Section 4, we show some experimental results and quantitative comparisons to demonstrate the superior performance of the proposed method over previous methods. Section 5 concludes this paper.

2. LPGMM BACKGROUND MODEL

The proposed method is a two-stage foreground extraction technique for video surveillance. Unlike the assumption of the traditional approaches that the backgrounds of input surveillance videos are stationary, we consider the problem that the background is dynamic and usually change quickly overtime.

In our system, we first take several frames in the beginning of the input video sequence as the training data for building the background model. Let the recent history of a pixel be \( \{ X_t, ..., X_N \} \), which is modeled by a mixture of \( K \) Gaussian distributions, and \( X \) be an intensity vector for R, G and B color channels. At time \( t \), the probability density function at observing pixel value \( X_t \) is given by

\[
P(X_t) = \sum_{i=1}^{K} \omega_i \cdot \eta(X_t; \mu_i, \Sigma_i)
\]

(1)

where the Gaussian probability density function \( \eta \) is defined as:

\[
\eta(X_t; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (X_t - \mu)^T \Sigma^{-1} (X_t - \mu)}
\]

(2)

The main difference between LPGMM and GMM [1] is that \( X_t \), \( \mu \) and \( \sigma \) for each pixel are vectors formed from observations from its local neighborhood instead of scalar values. The covariance matrix \( \Sigma \) is defined as:

\[
\Sigma = \begin{bmatrix}
\sigma_1^2 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sigma_d^2
\end{bmatrix}
\]

(3)

where \( d \times d \) is the local patch size for each pixel, and \( \sigma_i \) denote the standard deviation for the \( i \)-th pixel.

Let \( f_t(x, y) \) be the intensity value of the observed pixel; \( m_t(x, y) \) and \( s_t(x, y) \) are the corresponding mean and standard deviation of the background model at time \( t \) in the original GMM [1]. For each observed pixel, we set a window around the pixel as the center. Let the window size be \( d \times d \), we can extend the original \( X_t \) that is a single value in the original GMM model to a \( d^2 \)-dimensional vector for each observed pixel as follows:

\[
X_t(x, y) = [f_t(x - d', y - d') \ldots f_t(x, y) \ldots f_t(x + d', y + d')]^T
\]

(4)

where \( d' = (d - 1)/2 \) is assumed to be an integer without loss of generality. The mean \( \mu_t \) and standard deviation \( \sigma_t \) are extended to \( d^2 \)-dimensional vectors for each background pixel, i.e.,

\[
\mu_t(x, y) = [m_t(x - d', y - d') \ldots m_t(x, y) \ldots m_t(x + d', y + d')]^T
\]

(5)

\[
\sigma_t(x, y) = [s_t(x - d', y - d') \ldots s_t(x, y) \ldots s_t(x + d', y + d')]^T
\]

(6)

These vectors represent the local spatial information in a local neighborhood for each pixel, as illustrated in Fig. 2.

When a new frame is processed, we first check if the color values for each pixel are matched to any of the \( K \) Gaussian distributions. The mean difference between \( X_t(x, y) \) and \( \mu_t(x, y) \) is computed as follows:

\[
D(X_t(x, y), \mu_t(x, y)) = \sum_{-d' \leq i, j \leq d'} f_t(x + i, y + j) - m_t(x + i, y + j)
\]

(7)

If the observed pixel is matched to a Gaussian distribution, the mean difference usually falls within 2.5 times the corresponding standard deviation. Thus, a pixel is first classified as background if
\[
D(X_i(x,y), \mu_i(x,y)) \leq 2.5 \sum_{-d' \leq j \leq d'} \sigma_j (x + i, y + j) \tag{8}
\]

In the updating process, we update \(\mu_i(x,y)\) and \(\sigma_i(x,y)\) for those pixels classified to background as follows:

\[
\mu_i(x,y) = (1 - \rho) \mu_{i-1}(x,y) + \rho X_i(x,y) \tag{9}
\]

\[
\sigma^2_i(x,y) = (1 - \rho) \sigma^2_{i-1}(x,y) + \rho [X_i(x,y) - \mu_i(x,y)]^2 \tag{10}
\]

Note that the second term is an element-wise square function, and \(\rho\) is the learning rate set between 0 and 1.

4. SVM Shadow Detector

After the first stage, we can obtain the initial segmentation result that segments the input frame into foreground objects and background regions. Next, foreground objects will be discriminated from shadow regions by using a SVM shadow detector. Feature extraction is very critical to the shadow SVM classification. In this paper, we propose several discriminating features in the SVM classifier to classify between the foreground and shadow. Some of them were proposed in the previous works [7,8,9,13]. These features are detailed in the following.

A. Color projection ratio

This feature [14] is a measure of how closely the current pixel intensity matches that predicted by the background model. This feature is defined as a ratio of the projection of the observed pixel color onto the corresponding background pixel color to itself.

B. Bhattacharyya coefficient

We evaluate the Bhattacharyya coefficient feature value for each pixel as follows:

\[
F_{Bhattacharyya} = \frac{D_{Bhattacharyya}(H_R, H'_R) + D_{Bhattacharyya}(H_G, H'_G)}{2} \tag{11}
\]

where \(H_R^G\) and \(H_G^R\) are the color histograms for R and G channels in the local region of a sliding window in the input image and background model, respectively.

C. Edge magnitude distortion

In addition to the color information, we also utilize the texture information to extract features. If the edge magnitudes of a pixel in the current image and the background model image are \(m_1\) and \(m_2\), then we defined the edge magnitude distortion feature as follows:

\[
F_{Edge} = |m_1 - m_2| \tag{12}
\]

D. Normalized cross correlation (NCC)

We include this feature to describe the local spatial similarity between the local patches in the input and background images. NCC has been widely used to describe the similarity between two patches. The NCC feature is defined as follows:

\[
F_{NCC} = \frac{\sum_{-d' \leq j \leq d'}[f(x+i,y+j) - \bar{m}_i] [m_i(x+i,y+j) - \bar{m}_i]}{\sqrt{\sum_{-d' \leq j \leq d'}[f(x+i,y+j) - \bar{m}_i]^2 \sum_{-d' \leq j \leq d'}[m_i(x+i,y+j) - \bar{m}_i]^2}} \tag{13}
\]

E. Intensity ratio

Generally, there is an attenuation of image intensity to a background pixel when the cast shadow appears at the same location. We simply use the ratio of the observed pixel intensity and the corresponding background pixel intensity as the following intensity ratio feature:

\[
F_{Ratio} = \frac{I_{xy}}{m_i(x,y)} \tag{14}
\]

5. Experimental Results

In our experiments, we utilize the database from [7] and select some videos of size 160x128 or 176x144 pixels for testing the proposed algorithm. The number of Gaussians at each pixel location is empirically set to 3 and the size of the sliding window is set to 7x7. For the SVM training process, we collect some frames from videos and the training accuracy is 84.78%. For the experiments on all the video sequences, we obtain good results with the same parameter setting.

The first video sequence “Fountain” is depicted in Fig. 3(a). There is a fountain in the scene, and it causes intensity variations of many pixels over time. The reason that GMM could have more influences caused by noises is because GMM does not use the spatial information but LPGMM exploits it explicitly to construct the background model. Fig. 3(b) depicts the second video sequence “Waving trees”, which contains intensity variations of many pixels caused by the waving trees. The degree of intensity variations in the dynamic scene is larger than that in “Fountain”, thus the results of GMM have more false-alarm pixels. Fig. 3(c) shows the third video “Rippling water”, which contains a man walking and watching the waving water. Although the motion of waving water is not regular, we can still detect it as a background and obtain better result than the previous methods.

We also evaluated the quantitative result of the first-stage segmentation on three test videos described above. Each test video has its ground truth. “Fountain” has 14 frames with ground-truth segmentations, 20 frames for “Waving trees” and 15 frames for “Rippling water” with ground-truth segmentations. The False Negative (FN) and False Positive (FP) measures used in [7] are calculated to evaluate the accuracies for the comparison with the ground truths. The FN measure denotes the number of the foreground pixels that are classified as background, while the FP measure is the number of the background pixels that are detected as foreground.
Fig.3 Qualitative results for background subtraction with dynamic background videos (i) the input frames, (ii) the ground truths, and the results by using (iii) GMM [1], (iv) Bayes decision rule [7], (v) the proposed LPGMM, and (vi) the proposed LPGMM+SVM.

The results are summarized in Table 1. It is obvious that the proposed LPGMM+SVM has the best results compared with the other methods. Since our method can account for the problem of shadow and illumination variation, this is the reason why it could reduce the FP number in the testing videos. In contrast, the Bayes decision rule proposed in [7] cannot well handle the shadow problem.

6. CONCLUSION

In this paper, we proposed a two-stage system to detect moving object and remove shadow from surveillance videos with dynamic background. The first stage is to detect moving objects by using the proposed LPGMM model, which considers the local spatial information for each pixel. To remove shadow from the extracted foreground region, we utilize the linear SVM classifier to classify each pixel which is detected as moving object at the first stage to foreground or shadow. Experimental comparisons are shown to demonstrate the superior performance of the proposed background subtraction system over the previous competing methods for videos with dynamic background scenes.

7. ACKNOWLEDGEMENTS

This work was partially supported by Ministry of Economic Affairs (project 97-EC-17-A-01-S1-034) and National Science Council (project 99-2220-E-007-016) of Taiwan, R.O.C.

8. REFERENCES


Table 1. Qualitative comparison of different methods.

<table>
<thead>
<tr>
<th>Test Video</th>
<th>Method</th>
<th>FN</th>
<th>FP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fountain</td>
<td>GMM [1]</td>
<td>266</td>
<td>657</td>
<td>923</td>
</tr>
<tr>
<td></td>
<td>Bayes decision rule [7]</td>
<td>60</td>
<td>237</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM</td>
<td>20</td>
<td>481</td>
<td>501</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM+SVM</td>
<td>179</td>
<td>70</td>
<td>249</td>
</tr>
<tr>
<td>Waving</td>
<td>GMM [1]</td>
<td>264</td>
<td>348</td>
<td>614</td>
</tr>
<tr>
<td></td>
<td>Bayes decision rule [7]</td>
<td>126</td>
<td>250</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM</td>
<td>99</td>
<td>278</td>
<td>377</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM+SVM</td>
<td>99</td>
<td>111</td>
<td>210</td>
</tr>
<tr>
<td>Rippling</td>
<td>GMM [1]</td>
<td>1055</td>
<td>590</td>
<td>1645</td>
</tr>
<tr>
<td></td>
<td>Bayes decision rule [7]</td>
<td>135</td>
<td>129</td>
<td>264</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM</td>
<td>101</td>
<td>478</td>
<td>579</td>
</tr>
<tr>
<td></td>
<td>Proposed LPGMM+SVM</td>
<td>107</td>
<td>121</td>
<td>228</td>
</tr>
</tbody>
</table>