Background Subtraction with Background Modeling for Visual Surveillance Applications

Yuan-Been Chen and Chang-Ming Wu

Dept. of Electronic Engineering, Chienkuo Technology University, Changhua City, Taiwan

Dept. of Computer Science and Information Engineering, Providence University, Taichung Hsien, Taiwan

Abstract — This work adopts background subtraction to develop automatic extraction of moving objects in video surveillance. Pixels at a specific position of continuous image frames are manipulated by the modified iterative threshold selection technique to determine the background gray-level value. Pixels at all positions employ such an iterative technique to establish the background that is then analyzed and modified by using the relationship of neighboring pixels to avoid inconsistent classification. The original image is subtracted by this background to obtain a difference image. The robust edge-following method processes this difference image to yield the closed-form objects that are then conducted by the morphological operation to obtain complete objects. As compared to the conventional methods, the proposed method is demonstrated to have the best performance of moving object extraction in video sequences.

I. INTRODUCTION

The technique of moving object extraction plays a critical role in computer vision applications such as surveillance, tracking, recognition and indexing. With an increase of computing capability in a computer, more and more object extraction methods have been proposed in content analyses of videos [1]-[6]. For instance, Liu et al. developed a framework to discover video objects [1]. Wren et al. tracked a human body to understand his behavior [2]. Maddalena et al. proposed a method to watch suspicious movements or objects in an automatic surveillance system [3]. Irani et al. extracted objects of a video to do automatic annotation and retrieval [4]. Stern et al. adopted a method to recognize gestures and track eyeballs for the aid of manipulating a computer [5]. Cavallaro et al. real-time followed the trails of moving vehicles to do the flow control and route arrangement in a traffic surveillance system [6].

The commonly-used methodology of moving object extraction is background subtraction. This approach need establish stationary background regions that are subtracted from original images of a video sequence to yield difference images. Moving objects in these difference images can appear apparently. However, there exist difficulties of generating reliable background due to cluttered scenes, shadows and light change. The more precise is the background, the higher is the accuracy of extracted objects. In the literature, many background-subtraction approaches are based on probabilistic models in which probability distributions of pixels’ gray-level values are built [2],[7-8]. In [2], the Gaussian function is used to model the statistics of pixels’ gray-level values where the mean and variance are recursively updated over time. To characterize multimodal distributions, Stauffer et al. employ a mixture of K Gaussians where each Gaussian is classified as foreground or background distribution depending on the frequency of occurrence [7]. In [8], Elgammal et al. proposed a nonparametric Kernel Density Estimation (KDE) method which was suitable for modeling a wide range of probability density functions. The probability density function of background is estimated using a smoothed histogram of N (typically 50-100) recent pixel intensities. Adaptation is done simply by updating the histogram values with new pixel intensities. In [9], Babacan et al. provided an efficient background subtraction method based on a Bayesian formulation by preserving multi-modularity both spatially and temporally. The spatial interactions are controlled by a Gibbs-Markov random field and the temporal interactions are modeled by a mixture of Gaussians. In [6], Cavallaro et al. developed a hybrid strategy that utilizes object and region information to detect video objects.

This work proposes a method to accomplish moving object extraction based on background subtraction. The modified iterative threshold selection technique is employed to identify background. To avoid the inconsistent classification of neighboring pixels to be foreground or background, the relationship between each pixel and its neighboring eight pixels is used to correct inconsistent pixels, yielding a reliable background. The original images are then subtracted by this background to generate difference images. The robust edge-following method that was developed by us before is used to build the moving objects with closed contours [10]. Finally, the morphological operation is applied to fill empty areas of the closed contours and then yield the extracted objects. The proposed method has been successfully demonstrated by using indoor and outdoor scenes with both rigid and deformable objects without changing parameter setting. Additionally, it shows a superior performance to the conventional methods. Therefore, the method proposed herein can be widely used in various video surveillance applications on simultaneous extraction of multiple moving objects.

II. Proposed Method
II.I. Background Model

In this stage, the background model is built from some frames on a video sequence. All of the pixels in each frame are classified to be background or foreground. The more precise is the background, the higher is the accuracy of extracted objects. If the $i$th frame is processed, all of the frames from $(t-N+1)$th frame to $i$th frame are used to calculate the background model. $N$ (typically 50-100) could be selected for different video sequence. The gray-level value of each frame at location $(x, y)$ is represented as

$$V_{xy}(t) = I(x, y, t)$$

(1)

where $I(x, y, t)$ denote the gray-level values at locations $(x, y)$ of frame $t$. The iterative threshold selection technique [16] that was proposed by Ridler et al. to segment two regions is modified as follows.

1. Let $k = 0$, $T_k = \frac{\text{MAX}[V_{xy}(t)]}{2}$, where $\text{MAX}$ is a function to select the maximum value.

2. $T_k$ is adopted to classify all points in $V_{xy}(t)$ into two levels. A point with $V_{xy}(t) \geq T_k$ is higher_intensity_level, while a point with $V_{xy}(t) < T_k$ is lower_intensity_level. The groups of lower_intensity_level points and higher_intensity_level points are denoted by $L$ and $H$, respectively. In these two groups, the averaged $V_{xy}(t)$ is computed by

$$u_k = \frac{\sum V_{xy}(t)}{\#L}$$

(2)

and

$$v_k = \frac{\sum V_{xy}(t)}{\#H}$$

(3)

where $\#L$ and $\#H$ denote the numbers of lower_intensity_level and higher_intensity_level points, respectively.

3. $T_{k+1} = \text{round} \left( w_L \times u_k + w_H \times v_k \right)$

(4)

where round($\lambda$) rounds off the value of $\lambda$ to the nearest integer number. $w_L$ and $w_H$ ranging from 0 to 1, denote the weighting values of $L$ and $H$ regions, respectively. Additionally, $w_L + w_H = 1$.

4. If $T_{k+1} \neq T_k$ ,

then $k = k+1$ and go to Step 2,

else $T = T_k$.

5. The histogram $h_{xy}(z)$ of $V_{xy}(t)$ is then computed.

$$L(x, y) = \max[h_{xy}(z)|_{z=0-(T-1)}]$$

(5)

$$H(x, y) = \max[h_{xy}(z)|_{z=T-255}]$$

(6)

If $\sum_{z=0}^{T-1} h_{xy}(z) > \sum_{z=T}^{255} h_{xy}(z)$,

then $B(x, y) = L(x, y)$,

else $B(x, y) = H(x, y)$.

Where $L(x, y)$ and $H(x, y)$ represent the maximum of $(x, y)$ in lower_intensity_level and higher_intensity_level, respectively. $B(x, y)$ is the gray level of background model $B$ at location $(x, y)$.

6. Repeat step 1–5 for all locations in the frame.

Then, the relations of each pixel and its 8 neighboring points can be used to remove incorrect region as follows.

1. $(x, y)$ is a point from background model $B$.

2. If $F(x, y) = 0$ and $\sum_{d=0}^{2} F(x^d , y^d) \geq 7 - n$

then $B(x, y) = H(x, y), F(x, y) = 1$.

If $F(x, y) = 1$ and $\sum_{d=0}^{2} F(x^d , y^d) \leq n$

then $B(x, y) = L(x, y), F(x, y) = 0$.

When $n=0$, the isolated incorrect point can be
3. Repeat step 2 for all points in the background model.

In Fig. 1, the background model of “hall_monitor” video sequence is computed by using all of the frames from frame #16 to frame #60 as shown in Fig. 1 (a). Even though the walking man appears in each frame, there are no moving objects in the background model as shown in Fig. 1 (b), so that the stationary background can be obtained.

II.II. BACKGROUND SUBTRACTION

Moving object can be extracted by building the background model and then finding deviations from the model for each incoming frame. Any significant change in an image region from the background model indicates a moving object. Usually, a connected component scheme is applied to obtain connected regions corresponding to the objects. The pixels constituting the regions undergoing change are marked for further processing. The initial-point threshold $T_i$ is obtained from current frame as described in our previous work [10]. The positions and moving directions of initial points are recorded as follows.

1. $G_t(x, y) = I(x, y, t)$, \hspace{1cm} (7)

where $(x, y)$ is a point at frame $t$. $G_t(x, y)$ denote the gray-level values at points $(x, y)$ of frame $t$.

2. If $|G_t(x, y) - B(x, y)| \geq T_i$ then $(x, y)$ is labeled as the initial point and $d^*$ is recorded where

$$D_t^{d^*}(x, y) = \max\{D_t^d(x, y), \text{ for } 0 \leq d \leq 7 \}$$

$$D_t^d(x, y) = |G_t(x^d, y^d) - B(x^d, y^d)|$$

(8)

where $(x^d, y^d)$ neighbors to $(x, y)$ in direction $d$, and $G_t(x^d, y^d)$ denote the gray-level values at points $(x^d, y^d)$. Here, $d$ is a value denoting one of the eight compass directions.

3. Repeat step 2 for all points in frame $t$.

II.III. CONNECTED COMPONENT SCHEME

The position and direction of the searched contour point have been computed in the previous stage. The
modified edge-following method [10] is then adopted to connect the discontinuous contour. When all initial points are conducted, the procedure of the modified edge-following method is ended. Finally, the hollow region is filled by the fill morphological operation [11] to obtain the complete objects.

III. COMPUTATIONAL ANALYSES

In the following experiment, two quantities, precision and recall, are employed to evaluate the segmented results from each segmentation method [13], [14]. Precision, $P$, is the probability that a detected pixel is a true one. Recall, $R$, is the probability that a true pixel is detected. The extracted results by using watershed and proposed methods are evaluated according to the manually extracted ground truths.

A frame of a typical low-quality surveillance video is shown in Fig 2(a), where the resolution is low, blocking and artificial lighting artifacts are present. In Fig. 2(b), objects in the video sequences were manually extracted as the ground truth or actual objects. To do fair comparison, all of the following methods do not use any morphological operation. Fig. 2(c) depicts the result of KDE [8]. The KDE algorithm is run by setting the number of history pixels to 100. The high noise level in this result is due to the short initialization time, which is a major drawback of KDE. The result of MoG [7] is shown in Fig. 2(d), where maximum number of Gaussians per pixel set to 4. The noise-level is reduced but still visible. In Fig. 2(e), watershed [15] is a region-based method, which means that the extracted object has the continuous region. The extracted result has also very low noise level, but influenced by the light, too. In Fig. 2(f), although some of the pants of walking man are lost, the results by using proposed method are barely influenced by light and low contrast. Table 1 list the Precision($P$), Recall($R$) and F-measures($F$) in “hall_monitor” video sequence frames # 126 by using the KDE, MoG, watershed and proposed methods. The proposed method clearly outperforms all of the other methods. The most significant property to note is the very high Recall($R$) value of watershed method which is mainly due to the most ground truth are extracted. However, the extracted regions exceed the ground truth regions, causing the lower Precision ($P$) and F-measures ($F$) values.

<table>
<thead>
<tr>
<th>Frame no.</th>
<th>#126</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDE [8]</td>
<td>0.51 0.74 0.60</td>
</tr>
<tr>
<td>MoG [7]</td>
<td>0.63 0.80 0.70</td>
</tr>
<tr>
<td>Watershed [15]</td>
<td>0.71 0.90 0.79</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.87 0.85 0.86</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

Automatic detection and extraction of moving objects is not only widely used in content-based video applications, it is also an important issue for new standards such as MPEG-4, MPEG-7 and MPEG-21. In this work, an automatic method is proposed for moving objects detection and extraction in video sequences. Even though that all frames of the video sequence include some parts of moving objects, the modified iterative threshold selection technique can separate moving objects from the background and then create the stationary background model. The detection of moving objects can be achieved by comparing each new frame with a representation of the background model. The edge_following method is then adopted to connect the discontinuous contour resulted from low contrast. Finally, the empty areas of the closed contours are filled by the morphological operation to obtain the complete objects.

The experimental results of the proposed algorithm are very promising when compared to those of other techniques. Results show that the algorithm accurately detects single/multiple moving objects from indoor/outdoor video shots with cluttered background. Therefore, the proposed method can be widely and effectively employed in various video applications.

REFERENCES


