ABSTRACT In the field of surveillance, effective and rapid video object segmentation is a key technology for video analysis and processing. For the complex scene and noise that affecting segmentation issue in the fixed occasion. On the base of classic and adaptive Gaussian Mixture Background Model presented by Stauffer. In this paper, A new algorithm named the fusion of Spatio-Temporal based on Gaussian Mixture Model is proposed for video object segmentation. The algorithm classifies for each pixel in Time and Space scales. Firstly, the algorithm constructs dynamically Gaussian Mixture Background Model for each pixel and segment foreground objects through background subtraction. Secondly, the algorithm detects synchronously the neighborhood statistic feature of each pixel through two lemmas. Finally, producing a result using the spatial segmentation coupling with the temporal segmentation by “and” operator. Experiments show that the algorithm can segment the moving object effectively and quickly from video sequences and has stronger robustness application prospect contrasted with other algorithms.

KEYWORDS Fusion of Spatio-Temporal Information; Gaussian Mixture Model; Background Subtraction

1 Introduction

Precise subtraction of moving object in video stream is a hot field of interdisciplinary study, which involves many fields such as image processing, artificial intelligence, pattern recognition, machine learning. It’s also a key step to recognize and track objects in video surveillance, present multi-dimensional moving objects, code moving images, search and match content and so on. Background subtraction in video stream is a prerequisite and guarantee of accurate subtraction of moving objects. Quick, precise, and integrated background segmentation is becoming difficulty in research on moving object precise subtraction. Moving object detection and segmentation methods which have been widely applied now mainly include: background difference method, inter-frame difference method, method based on statistical model, optical flow method, etc. Inter-frame method is not sensitive to changes of scene lighting. When grayscale in large area of the object surface is uniform, “holes” will appear and segment the object into many regions. Background difference method detects moving area using difference of current video frame and background image. It can provide the most integrated information of moving object, and has a fast speed, a simple algorithm and a current detection to meet the real-time requirements of the system. But it is sensitive to changes of the moving scenes such as lighting and noise. So it needs to update background constantly. The method based on statistical model can subtract background from frame series well and segment a moving object, but it needs a former assumption of the distribution of background feature density field. Optical flow method uses the physical properties of optical flow that the moving objects change over time to subtract the moving object effectively. Its advantage is that it can also segment the independent moving object under the condition of camera motion. But it has a particular complex and large calculation and can’t meet real-time requirements.

Classical adaptive Gaussian mixture background subtraction model is a background modeling method based on pixel, proposed by Stauffer et al. It constructs a grayscale distribution model of each pixel based on the distribution information of each pixel in time domain and builds a background model of the pixels. Gaussian Mixture Background Model is a weighted sum of finite functions, which can describe the multi-peak state of pixels. It suits to model the complex background such as light gradients and trees swaying. Afterwards researchers do a variety of improvements on the Gaussian Mixture Model (GMM) background subtraction method. This method becomes a usual background subtraction method now. But this orthodoxy GMM models every pixel in every frame, which defaults that each pixel is strictly independent and irrelevant. From this analysis, it ignores the correlation rule of spatial adjacent domain among pixels in the same frame, and reduces the noise immunity and anti-interference. As a result, later researchers tend to fuse some spatial characteristic information of the video object, such as information of intensity, border,
grayscale and texture, to segment the video object. For the above, this article proposes an algorithm for video object segmentation by fusion of spatio-temporal information based on GMM. This algorithm differs from orthodox algorithms based on spatio-temporal information \[13-15\]. Firstly it builds background GMM for each pixel in frame images in the time domain, and segments foreground object from the background by the background difference method in real time. Then by introducing two lemmas in the spatial domain, it synchronously detects the neighborhood statistic characteristics of each pixel on the video stream image sequence from the image grayscale, texture and other characteristics. After this processing, we can couple the results of the spatio-temporal segmentation of the same video frame through “and” operation, and then segment the accurate objects after morphological processing. The experimental result shows that this algorithm has high robustness, separates the foreground moving objects from the background well, has high adaptability to complex background environment, and obtains a good testing effect. At the same time, this algorithm effectively remedies the misjudged issues that the object that stagnates a long time changes from foreground to background in the Gaussian Mixture Background Model proposed by Stauffer \textit{et al.}

2 Video Object Segmentation in the Time Domain

A. Subtraction of Background in the Time Domain Based on GMM

We model each pixel in the frame images in the video stream using GMM. This GMM model uses \( K (K=3,4,5) \) Gaussian distributions to count characteristics of the same pixel in each frame in the video stream. That is to set the pixel value in the latest \( t \) frames to \((X_1, X_2, ..., X_t)\), among them \( X_i \) is the pixel value at the time \( t \). The probability function of \( K \) mixture Gaussian distributions is:

\[
p(X_t) = \sum_{k=1}^{K} \omega_{k,t} \cdot g(X_t; \mu_{k,t}, \Sigma_{k,t})
\]

\[
g(X_t; \mu_{k,t}, \Sigma_{k,t}) = \frac{1}{\sqrt{2\pi|\Sigma_{k,t}|}} \exp\left[-\frac{1}{2}(X_t - \mu_{k,t})^T \Sigma_{k,t}^{-1} (X_t - \mu_{k,t})\right]
\]

The parameters of the probability function are:

\[
(\omega_{1,t}, \omega_{2,t}, ..., \omega_{K,t}; \mu_{1,t}, \mu_{2,t}, ..., \mu_{K,t}; \Sigma_{1,t}, ..., \Sigma_{K,t})
\]

Among them, \( \omega_{k,t}, \mu_{k,t}, \Sigma_{k,t} \) respectively are the weight, mean vector, and covariance matrix of the \( k \) th Gaussian distribution at time \( t \). \( K \) is the number of Gaussian distributions. And \( \omega_{1,t}, \omega_{2,t}, ..., \omega_{K,t} \) meet the condition:

\[
\sum_{k=1}^{K} \omega_{k,t} = 1
\]

Assuming pixel values in the RGB color space are independent of one another and have the same variance, we have the following equation:

\[
\text{Cov}_{k,t} = \sigma_{k,t}^2 I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}
\]

For the new sample point \( X_{t+1} \) in video stream frame image sequence, we get the current value \( \text{ImageData}(X_{t+1}) \) of pixel \( X_{t+1} \) and compare one by one with the mean value \( \mu_{k,t} \) of \( K \) Gaussian model allocated by memory, and see which Gaussian model matches it. The matching principle is:

\[
\left| \text{ImageData}(X_{t+1}) - \mu_{k,t} \right| = \xi \cdot \sigma_{k,t}
\]

Among it \( \xi \) more appropriately values \( 3 \pm 0.5 \) in general. \( \sigma_{k,t} \) is the standard deviation of the \( k \) th Gaussian distribution. According to the judging formula (5) we can find the matching Gaussian distribution. If the \( k \) th pixel exactly matches the \( K (K=1,2,3,4,5) \) th Gaussian distribution, the \( K \) th Gaussian distribution will be updated by \( X_{t+1} \), while other Gaussian compositions remain the same. In actual software design, the structure body directed by the \( K \) th Gaussian distribution pointer will include four important member variables: number of matches, weight, variance, and mean value. Next we will update the weight, variance and mean value of the \( K \) th Gaussian distribution according to \( X_{t+1} \). Update algorithm is as follows:

\[
\omega_{k,t+1} = (1 - \alpha) \cdot \omega_{k,t} + \alpha \cdot (M_{k,t})
\]

\[
M_{k,t} = \begin{cases} 
1 & \text{match successes} \\
0 & \text{match fails}
\end{cases}
\]

Among them is the studying rate of Gaussian model.

\[
\mu_{k,t+1} = (1 - \rho) \cdot \mu_{k,t} + \rho \cdot X_{t+1}
\]

\[
\sigma_{k,t+1}^2 = (1 - \rho) \cdot (\sigma_{k,t}^2) + \rho \cdot (X_{t+1} - \mu_{k,t})^T (X_{t+1} - \mu_{k,t})
\]

Among them:

\[
\rho = \alpha \cdot g(X_{t+1}, \mu_{k,t}, \sigma_{k,t}^2)
\]

After update parameters of each Gaussian distribution, we need to calculate the values of \( \omega_{k,t} \sigma_{k,t} \) in \( K \) Gaussian distributions and sort the Gaussian distributions by \( \omega_{k,t} \sigma_{k,t} \).
from large to small. Determine the Gaussian distribution that can best characterize background characteristics according to sorted Gaussian distributions. The determine algorithm is: we only obtain B Gaussian distributions which have a greater former weight. That is:

$$B = \arg\min_b \left( \sum_{k=1}^b \omega_{k,b} > T \right)$$  \hspace{1cm} (11)

Among them threshold \( T \) represents the prior probability value that the current pixel is judged as background. In the experiment, the \( T \) value is important because if \( T \) is too small, multi-Gauss model will deteriorate into single Gauss model, and if \( T \) is too large, the number of Gaussian distributions that describe the background increases and the system will regard moving objects as the background. Most literatures set \( T = 0.8 \) or so. In this way, we accomplish the subtraction of background in the time domain based on GMM.

**B. Foreground Segmentation Using Background Difference Method**

Known from formula (11) above, if we use B Gaussian distributions that have bigger former weight to represent background distribution, the remaining \( K-B \) Gaussian distributions will be positioned as foreground objects. That is:

$$F = K - B$$  \hspace{1cm} (12)

Through a series of simulation experiments, we know that it appears a result of vague edge contour of segmentation objects to segment foreground objects using formula (12). It will greatly reduce the effect of segmentation. We know, background difference method received wide application for its simple algorithm, accurate detection, immediacy and effectiveness. The premise and key to obtain a fine effect of segmentation is to build a steady and reliable background model. So we can use the background difference method to segment the foreground based on background modeling by GMM. It will obtain a real-time robust preliminary segmentation result in the time domain.

**3 Video Object Segmentation in the Spatial Domain**

In the algorithm of video object segmentation in the spatial domain, since GMM is assumed independent among each pixel, it results in incomplete segmented foreground objects and existence of some small holes, at the same time a small part of the foreground objects will be misjudged as background. In response to these deficiencies, we do the second segmentation in the spatial domain based on the video object segmentation algorithm in the time domain (as shown in Fig.1). Its main purpose is to correct and compensate the background subtracted from the time domain and segmented foreground objects, that is, to secondarily cluster on the basis of the initial classification in the time domain. Correct those background points that are misjudged as foreground points, and those foreground points that are misjudged as background points. Its main segmentation idea is mainly based on the following two lemmas:

**Lemma 1:** if most pixels in a small neighborhood of a pixel belong to foreground in a foreground object, this pixel very likely belongs to foreground.

Lemma 1 illustrates: from the perspective of spatial domain of every frame in the video stream, the grayscale of all pixels in every small neighborhood affects the current pixel magnificently. So from this point of view, we construct a \( 3 \times 3 \) filter (shown as Fig.1) to traverse all of the pixels in each frame. Consider the template of 8-neighborhood around the pixel \((x,y)\), using the Bayesian decision theory in statistics, we model the priori probability \( P_{x,y}(f_g) \) of pixel \((x,y)\) as follows [16]:

$$P_{x,y}(f_g) = \frac{1}{Z} \exp\{-n_{in}B + n_{id}C\}$$  \hspace{1cm} (13)

Among them, \( Z \) is an orthogonal constant, \( n_{in} \) is the number of pixels on horizontal or vertical direction in the foreground template, \( n_{id} \) is the number of pixels on adjacent diagonal direction in the foreground template, \( B \) and \( C \) is corresponding penalty coefficient. Next we give out the likelihood probability \( P_{x,y}(V_i | f_g) \) (among them \( V_i \) is the foreground characteristic vector) of the foreground pixel \((x,y)\). Then the pixel \((x,y)\) clustered as foreground will meet the decision analysis formula as follows:

$$P_{x,y}(V_i | f_g) \cdot P_{x,y}(f_g) > T$$  \hspace{1cm} (14)

Among them, \( T \) is a fixed threshold, valuing 0.7 in general. If formula (15) is satisfied, this pixel will be
clustered as foreground. Conversely, this pixel will be
divided into background.

**Lemma 2:** the pixels in a small neighborhood of a
pixel in foreground segmentation objects have a local
gayscale similarity.

Lemma 2 illustrates: in the foreground object
neighborhood containing small holes obtained from
video object segmentation algorithm in the time domain,
by calculating local brightness similarity of foreground
pixels, we can re-cluster those pixels misjudged as
background to foreground objects. Assume the pixel
gayscale value of foreground object $P$ in the $t$ frame of
the video images is:

$$f(x, y) = f(x, y)|(t, p)$$

Then the gradient vector of this point is:

$$\nabla f(x, y)|(t, p) = f_x(x, y)|(t, p)i + f_y(x, y)|(t, p)j$$

Among them:

$$f_x(x, y)|(t, p) = \nabla f_x(x, y)|(t, p)$$

$$f_y(x, y)|(t, p) = \nabla f_y(x, y)|(t, p)$$

Assume two adjacent pixels in a small neighborhood of
point $P$ are $P_1, P_2$, respectively, and assume the
correspondent gradient vector similarity of the two pixels
$P_1, P_2$ is Similar($P_1, P_2$), then:

$$\text{Similar}(P_1, P_2) = \frac{\nabla f(x_1, y_1)|(t, p_1) \cdot \nabla f(x_2, y_2)|(t, p_2)}{|| \nabla f(x_1, y_1)|(t, p_1)|| \cdot || \nabla f(x_2, y_2)|(t, p_2)||}$$

By the definition of similarity, the following conclusions
are obviously set up:

$$0 \leq \text{Similar}(P_1, P_2) \leq 1$$

$$\text{Similar}(P_1, P_1) = 1$$

$$\text{Similar}(P_2, P_2) = 1$$

$$\text{Similar}(P_1, P_2) + \text{Similar}(P_2, P_1) \geq 2 \text{Similar}(P_1, P_2)$$

From formula (19) we know: if two pixels $P_1, P_2$ both
belong to background or both belong to foreground
objects, the following formula is established:

$$\text{Similar}(P_1, P_2) \approx 1$$

If in the two pixels $P_1, P_2$, one belongs to background
and the other one belongs to foreground object, the
following formula is established:

$$\text{Similar}(P_1, P_2) \ll 1$$

According to this, we introduce a decision-making function $\delta(p_i, p_j)$, its expression is:

$$\delta(p_i, p_j) = 1 - \text{Similar}(p_i, p_j)$$

So we get the pseudo code of lemma 2 algorithm idea

*If $\delta(p_i, p_j) \geq \lambda$ is a threshold, $\lambda$ values
0.5 in general*

**begin**

*If $p_i \in \{(x, y)|(x, y) \in \text{foreground}\}$

then $p_j \in \{(x, y)|(x, y) \in \text{background}\}$

*If $p_i \in \{(x, y)|(x, y) \in \text{background}\}$

then $p_j \in \{(x, y)|(x, y) \in \text{foreground}\}$

**end**

**4 Spatio-Temporal Coupling Based on Gmm**

Through the preliminary segmentation in the time
domain and the following processing in the spatial
domain, we obtain a general background distribution
and foreground object of the video stream. That
is through obtaining general regions of the video
objects by segmentation in the time domain, and then
through gain accurate information of video objects
by secondary segmentation in the spatial domain, we
clear and complete the foreground objects. After above
processing, we can couple the results of spatio-temporal
segmentations of the same video frame through pixel
“and” operation. It successfully restores pixels misjudged
as background to corresponding foreground pixels, and
also makes those pixels belongs to foreground originally
remain the same. Thus, we accomplish a perfect coupling
of a preliminary segmentation based on the time domain
and a secondary segmentation based on the spatial
domain, and accurately segment the foreground objects
in the video at last.

**5 Morphological Processing**

Due to the impact of noise, after the preliminary
segmentation, secondary processing, and spatio-
temporal coupling of video images based on GMM, there inevitably are some local small regions caused by several isolated points and noise. Here we use a 3×3 sliding window (20) to process these processed video images by morphology. The specific processing steps are as follows:

$$W_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$  \hspace{1cm} (20)

1) Firstly do corrosion process twice, removing local small area and isolated points in the video images.

2) Secondly do expansion process for those corroded video images twice, making the edge of the corroded foreground objects smooth and recovered.

6 Experimental Results and Analysis

To verify the actual segmentation effect of the algorithm proposed by this article, we use the outdoor highway traffic video for detection. We did many groups of simulation experiment on the VC++6.0 platform on a Pentium 4, 1.8GHz dual core PC. Here in GMM, parameter $K$ values 5, initial variance values 100, $a$ values 0.06, $\xi$ values 3. Fig.2 and Fig. 3 are comparisons of the experimental results of the segmentation algorithm proposed by this article and Stauffer et al. The first and the second image series correspond to the processing results on frame 536 and frame 585 in the video sequence. Each row from left to right is followed by video image, background subtracted by the adaptive Gaussian Mixture Background Model, segmentation result using Stauffer algorithm, segmentation result using algorithm of this article. From the experimental results we can see: the foreground objects using the segmentation algorithm proposed by this article are more accurate and integrated than the segmentation result using Stauffer algorithm. This algorithm not only overcomes the deficiency of having vague edge profiles and the misjudging problem that the long stagnating objects change from foreground into background in the segmentation result of Stauffer, but also can segment the foreground objects accurately and completely under the condition that the road shakes up and down caused by the slight jitter of the camera, showing strong robustness.

7 Experimental Results And Analysis

This article proposes a video object segmentation algorithm by coupling of spatio-temporal information based on GMM. This algorithm builds Gaussian Mixture Background Model dynamically in the time domain and preliminarily segments the foreground objects through background difference, then introduces two lemmas and secondarily clusters the preliminary segmentations in the spatial domain, at last couples the results of segmentations using the spatio-temporal information and thereby obtains an accurate segmentation result. The experiment verifies the effectiveness of this algorithm: not only improves the completeness and precision of the segmentation result, but also has a strong adaptability to the complex background environment. At the same time, this algorithm effectively makes up the misjudging problem like long stagnating objects change from
foreground into background in the adaptive Gaussian Mixture Background Model proposed by Stauffer et al.

Acknowledgment

This paper is supported by the following grants: Key Laboratory of Robotics and Intelligent System of Guangdong Province (2009A060800016), Guangdong Science Foundation (0860011001) and Shenzhen Fundamental Research (0994151001 and 0889041001).

In addition, all the translation work of this paper was finished by Xuan Wang. I would like to express my gratitude to him.

References


Author

Qingsong Zhu: Research Assistant. Graduated from University of Science and Technology of China in 2010. His research interests include Image Processing, Computer Vision, Machine Learning, Pattern Recognition and Support Vector Machine.