A Dynamic Threshold Approach For Video Object Extraction

Nijo John¹, V. Sakthivel²

¹M.E-Final year, Paavai Engineering College, Pachal, Namakkal
²Assistant Professor, Paavai Engineering College, Pachal, Namakkal

¹nijojohnk@gmail.com
²vsakthivelme@yahoo.co.in

Abstract - This paper proposes efficient motion detection and people counting based on background subtraction using dynamic threshold. Here three different methods are used effectively for object detection and compare these performance based on accurate detection. The techniques frame differences, dynamic threshold based detection and mixture of Gaussian model will be used. After the object foreground detection, the parameters like sensitivity, correlation, speed, velocity and angle of motion will be determined. In dynamic threshold based object detection, morphological process and filtering also used effectively for unwanted pixel removal from the background. Along with this dynamic threshold, we introduce a background subtraction algorithm for temporally dynamic texture scenes using a mixture of Gaussian, which has an ability of greatly attenuating color variations generated by background motions while still highlighting moving objects. Finally the proposed method will be proved that effective for background subtraction in dynamic texture scenes compared to several competitive methods and parameter of moving object will be evaluated for all methods.

KeyWords - Frame Separation, Frame Subtraction, Dynamic Threshold Approach, Morphological Filtering, Object Detection

I. INTRODUCTION

A video object extraction system for real-time applications requires the following criteria.

- Segmented object should conform to human perception i.e., semantically meaningful objects should be segmented.
- Segmentation algorithm should be efficient and achieve fast speed.
- Initialization should be simple and easy for users to operate.

In Video Object (VO) segmentation methods, which are using mathematical morphology and perspective motion model, objects of interest should be initially outlined by human observer. From the manually specified object boundary, the correct object boundary is calculated using a morphological segmentation tool [1]. The obtained VOP is then automatically tracked and updated in successive frames.

It has difficulty in dealing with a large non rigid object movement and in the presence of occlusion, especially in the VOP tracking schemes [2]. The algorithm based on edge change detection [3], allows automatic detection of the new appearance of a VO. The edge change detection for inter-frame difference is another stream of popular schemes because it is straightforward to implement and enables automatic detection of new appearance. This ability enables to develop a fully automated object-based system [4], such as an object-based video surveillance system.

It is found that the algorithms based on inter frame change detection render automatic detection of objects and allow larger non rigid motion compared to mathematical morphology and perspective motion model methods [5], [6], [7].
The drawbacks are small false regions detected by decision error due to noise. Thus, small whole removal using morphological operation and removal of false parts like uncovered background by motion information are usually incorporated.

Another drawback in edge change detection is that object boundaries are irregular in some critical image areas, which must be smoothened and adapted by spatial edge information. Since spatial edge information is useful for generating VOP with accurate boundaries, a simple binary edge difference scheme may be assumed to be a good solution. In order to overcome boundary inaccuracy multiple features, multiple frames and spatial-temporal entropy methods are used [8], [9]. In addition, it gives robustness to noise and occluding pixels.

In the existing systems, the Video Object Extraction is done by saliency based methods. Visual and Motion saliency information are used to separate foreground and background regions within and across video frames. For combining these saliency induced features [10], a conditional random field is applied. But instead of assuming that the background is static over short time periods. However, structured motion patterns of the background which are distinctive from variations due to noise, are hardly tolerated in this assumption and thus still lead to high-level false positive rates when using previous models. In dynamic threshold based object detection, morphological process and filtering also used effectively for unwanted pixel removal from the background

A. Frame Separation
A method to shorten the time required for the frame separation in the time-division-multiplexed delta modulation system is described. Employing successive "1"s as the sync pattern, the system detects the sync channel out of the delta modulated information pulses which occur at the rate of one half in average.

Using a memory device with the capacity of one frame, the detection is performed by taking successive frame correlation of each channel in parallel. For example, in the system with 20 channels, the frame separation is established in approximately six frame periods. The system includes such facilities as to stabilize the frame separation against the error of the sync pattern and against the occurrence of the information pattern similar to the sync pattern.

B. Frame Subtraction
In digital photography, dark frame subtraction is a way to minimize image noise for pictures taken with long exposure times. A dark frame, or an average of several dark frames, can then be subtracted from subsequent images to correct for fixed-pattern noise such as that caused by dark. Dark-frame subtraction has been done for some time in scientific imaging; many newer consumer digital cameras offer it as an option, or may do it automatically for exposures beyond a certain time
B. Frame Subtraction

In digital photography, dark frame subtraction is a way to minimize image noise for pictures taken with long exposure times. A dark frame, or an average of several dark frames, can then be subtracted from subsequent images to correct for fixed-pattern noise such as that caused by dark. Dark-frame subtraction has been done for some time in scientific imaging; many newer consumer digital cameras offer it as an option, or may do it automatically for exposures beyond a certain time.

Visible fixed-pattern noise is often caused by hot pixel–pixel sensors with higher than normal dark current. On long exposure, they can appear as bright pixels. Sensors on the CCD that always appear as brighter pixels are called stuck pixels while sensors that only brighten up after long exposure are called hot pixels. Background subtraction is the first step in the process of segmenting and tracking people. Distinguishing between foreground and background in a very dynamic and unconstrained outdoor environment over several hours is a challenging task.

Figure 2: Flowchart of Background Subtraction
In figure 2, the first k video frames are used to train the background model to achieve a model that represents the variation in the background during this period. The following frames (from k + 1 and onwards) are each processed by the background subtraction module to produce a mask that describes the foreground regions identified by comparing the incoming frame with the background model.

Information from frames k + 1 and onwards are used to update the background model either by the continuous update mechanism, the layered Updating, or both. The mask obtained from the background subtraction is processed further in the post processing module, which minimizes the effect of noise in the mask.

1. Background model:
   The principle of the codebook background model is to quantify the variation at each pixel into a set of vectors. Each vector is called a codeword, and a set of code words forms a codebook. The quantization into codeword is not based on any parametric assumptions. If a background pixel has very little variation then only a single codeword is used to model the background at that pixel. If a background pixel has large variations then multiple code words are used to model the background at that pixel. Each codeword basically describes a mean color, a minimum intensity and a maximum intensity. Besides the color and intensity information, the codeword also holds temporal information used during training, updating, and background subtraction.

   When a pixel from a new frame is classified as either foreground or background, it is done based on the distance from the value of that pixel to each of the code words in the codebook, i.e. the pixel with image coordinates (x,y) is compared to the codebook at position (x,y) in the background model. If the distance from the pixel value x1 to the codeword c is sufficiently small, then x is said to be accepted by c, or in other words, c is said to be activated by x. One of the main features of the codebook background subtraction method is the separation of chromaticity and intensity, and the following sections will first describe this color representation, and then describe the actual codebook background model.

2. Subtraction and Update of the Background Model: After initialization, temporally subsequent samples are fed to the network. Each incoming pixel P, of the sequence Frame I, is compared to the current pixel model C to determine if there exists a weight vector that best matches it. If a best matching weight vector Cm is found, it means that P belongs to the background and it is used as the pixel encoding approximation, and the best matching weight vector, together with its neighborhood, is reinforced. Otherwise, if no acceptable matching weight vector exists, it will be discriminated whether P is in the shadow cast by some object or not. In the first case, P should be still considered as background, but it should not be used to update the corresponding weight vectors, in order to avoid the reinforcement of shadow information into the background model; in the latter case P is detected as belonging to a moving object (foreground).

3. Foreground/background classification: When the background model has been initialized through the training period the classification of pixels as either foreground or background is a relatively simple task. If a pixel at image coordinate (x,y) lies within the subspace of RGB space described by any of the codewords in the codebook at position (x,y), then the pixel is classified as background, otherwise it is foreground. Whether or not a pixel lies within the subspace of a codeword is determined in the same way as in the training where the difference in chromaticity must be smaller than the subspace radius, and the intensity Ix must be within the intensity limits Ilo and Ihi. High utilization of the background model gives the foreground/background. In the classification the ability to handle shadows and highlights since the model accounts for the intensity variation. The background model’s way of dealing with chromaticity differences also gives the foreground/background classification a high detection sensitivity to foreground objects that are similar in color to the background.

The preprocessing, mentions that smoothing of the images can be used to reduce camera noise and remove transient environmental noise such as rain. Many algorithms use a Gaussian blur first to average out fluctuating pixel values to alleviate big differences. Alternatively, when temporal data can be exploited in a video, if a pixel’s value is constantly changing over time then it can be assumed it is part of a non-static background object. The background model can deal with events such as objects changing positions by implementing an effective update rule to change the model over time.
C. Dynamic Threshold Method

Fixed decision boundaries classification approaches are successfully applied to segment human skin. These fixed thresholds mostly failed in two situations as they only search for a certain skin color range:

1. Any non-skin object may be classified as skin if non-skin objects’ color values belong to fixed threshold range.
2. Any true skin may be mistakenly classified as non-skin if that skin color values do not belong to fixed threshold range. Instead of predefined fixed thresholds, novel online learned dynamic thresholds are used to overcome the above drawbacks.

In the general case, a threshold is produced for each pixel in the original image; this threshold is then used to test the pixel against, and produce the desired result. According to this, the general definition of a threshold can be written in the following manner:

\[ T = T[x, y, p(x, y), f(x, y)] \] (1)

Where \( f(x, y) \) is the gray level of point \((x, y)\) in the original image, and \( p(x, y) \) is some local property of this point. Actually, this is one of the more important components in the calculation of the threshold for a certain point. In order to take into consideration the influence of noise or illumination, the calculation of this property is usually based on an environment of the point at hand.

An example of a property may be the average gray-level in a predefined environment, the center of which is the point at hand. Local adaptive thresholding selects an individual threshold for each pixel based on the range of intensity values in its local neighborhood. allows for thresholding of an image whose histogram doesn’t contain distinctive peaks.

D. Morphological Filtering

Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. Morphological operations rely only on the relative ordering of pixel values, not on their numerical values, and therefore are especially suited to the processing of binary images. Morphological techniques probe an image with a small shape or template called a structuring element.

The original image is divided into an array of overlapping sub-images. A gray-level distribution histogram is produced for each sub-image, and the optimal threshold for that sub-image is calculated based on this histogram. Since the sub-images overlap, it is then possible to produce a threshold for each individual pixel by interpolating the thresholds of the sub-images.
1. The matrix dimensions specify the size of the structuring element.
2. The pattern of ones and zeros specifies the shape of the structuring element.
3. An origin of the structuring element is usually one of its pixels, although generally the origin can be outside the structuring element.

When a structuring element is placed in a binary image, each of its pixels is associated with the corresponding pixel of the neighbourhood under the structuring element. The structuring element is said to fit the image if, for each of its pixels set to 1, the corresponding image pixel is also 1. Similarly, a structuring element is said to hit, or intersect, an image if, at least for one of its pixels set to 1 the corresponding image pixel is also 1. Zero-valued pixels of the structuring element are ignored, i.e. indicate points where the corresponding image value is irrelevant.

1. Erosion and dilation: The erosion of a binary image \( f \) by a structuring element \( s \) (denoted \( f \ominus s \)) produces a new binary image \( g = f \ominus s \) with ones in all locations \((x,y)\) of a structuring element's origin at which that structuring element \( s \) fits the input image \( f \), i.e. \( g(x,y) = 1 \) is \( s \) fits \( f \) and 0 otherwise, repeating for all pixel coordinates \((x,y)\).

Erosion with small (e.g. \( 2\times2 - 5\times5 \)) square structuring elements shrinks an image by stripping away a layer of pixels from both the inner and outer boundaries of regions. The holes and gaps between different regions become larger, and small details are eliminated.

Larger structuring elements have a more pronounced effect, the result of erosion with a large structuring element being similar to the result obtained by iterated erosion using a smaller structuring element of the same shape. If \( s_1 \) and \( s_2 \) are a pair of structuring elements identical in shape, with \( s_2 \) twice the size of \( s_1 \), then

\[
f \ominus s_2 \simeq (f \ominus s_1) \ominus s_1 \quad (2)
\]

Erosion removes small-scale details from a binary image but simultaneously reduces the size of regions of interest, too. By subtracting the eroded image from the original image, boundaries of each region can be found: \( b = f - (f \ominus s) \) where \( f \) is an image of the regions, \( s \) is a \( 3\times3 \) structuring element, and \( b \) is an image of the region boundaries.

The dilation of an image \( f \) by a structuring element \( s \) (denoted \( f \oplus s \)) produces a new binary image \( g = f \oplus s \) with ones in all locations \((x,y)\) of a structuring element's origin at which that structuring element \( s \) hits the input image \( f \), i.e. \( g(x,y) = 1 \) if \( s \) hits \( f \) and 0 otherwise, repeating for all pixel coordinates \((x,y)\). Dilation has the opposite effect to erosion -- it adds a layer of pixels to both the inner and outer boundaries of regions.

The holes enclosed by a single region and gaps between different regions become smaller, and small intrusions into boundaries of a region are filled in.
Results of dilation or erosion are influenced both by the size and shape of a structuring element. Dilation and erosion are dual operations in that they have opposite effects. Let \( f' \) denote the complement of an image \( f \), i.e., the image produced by replacing 1 with 0 and vice versa. Formally, the duality is written as

\[
f \oplus s = f' \ominus s_{\text{rot}}
\]

Where \( s_{\text{rot}} \) is the structuring element \( s \) rotated by 180°. If a structuring element is symmetrical with respect to rotation, then \( s_{\text{rot}} \) does not differ from \( s \). If a binary image is considered to be a collection of connected regions of pixels set to 1 on a background of pixels set to 0, then erosion is the fitting of a structuring element to these regions and dilation is the fitting of a structuring element (rotated if necessary) into the background, followed by inversion of the result.

\[ E. \ Object \ Detection \]

Objects of interest automatically in digitized pictures drawing the boundaries around objects are essential for pattern recognition, object tracking, image enhancement, data reduction, and various other applications. This constitutes a good survey of research and applications in image processing and picture analysis. The image of an object is more or less uniform or smooth in its local properties and there is detectable discontinuity in local properties between images of two different objects. This will adopt these two assumptions in this paper and assume no textural image.

The work on automatic location of objects in digitized images has split into two approaches: edge detection and edge following versus region growing. Edge detection applies local independent operators over the picture to detect edges and then uses algorithms to trace the boundaries by following the local edge detected. A recent survey of literature in this area.

In this method the two approaches are combined to complement each other; the result is a more powerful mechanism to segment pictures into objects. This will develop a new edge detector and combined it with new region growing techniques to locate objects; in so doing we resolved the confusion in regular edge following that the results where more than one isolated object on a uniform background is in the scene.

\[ F. \ Parameters \ Analysis \]

1. \textit{Correlation:} Correlation and Convolution are basic operations that we will perform to extract information from images. They are in some sense the simplest operations that we can perform on an image, but they are extremely useful. Moreover, because they are simple, they can be analyzed and understood very well, and they are also easy to implement and can be computed very efficiently.

2. \textit{MSE & PSNR:} Image quality assessment is an important but difficult issue in image processing applications such as compression coding and digital watermarking. For a long time, mean square error (MSE) and peak signal-to-noise ratio (PSNR) are widely used to measure the degree of image distortion because they can represent the overall gray-value error contained in the entire image, and are mathematically tractable as well.

<table>
<thead>
<tr>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE = \sum (input - output)^2/N</td>
<td>Mean square error</td>
</tr>
<tr>
<td>PSNR = 10 \log_{10}(\frac{255^2}{MSE})</td>
<td>Peak signal-to-noise ratio</td>
</tr>
<tr>
<td>Or</td>
<td>\textit{PSNR} = 20 \log_{10}(\frac{255}{\sqrt{\text{MSE}}})</td>
</tr>
</tbody>
</table>

MSE works satisfactorily when the distortion is mainly caused by contamination of additive noise. However, the problem inherent in MSE and PSNR is that they do not take into account the viewing conditions and visual sensitivity with respect to image contents.

With MSE or PSNR, only gray-value differences between corresponding pixels of the original and the distorted version are considered. Pixels are treated as being independent of their neighbors.
III. RESULT AND ANALYSIS

The video is divided into frames according to the dimensions and then the frames are processed through above steps. Frames are compared to detect the moving objects and the dynamic threshold values are used to extract the foreground objects.

Noises are removed by Gaussian Modeling and the by considering the parameters like sensitivity, correlation, MSE and PSNR, the quality assessment is done. The Morphological filters are used to smoothening the edges of the extracted objects.

The analysis is done using the Mean Square Error (MSE) and the Peak to Signal Noise Ratio (PSNR). Both these parameters are inversely related. A graph describing the relation of both MSE and PSNR to the background Suppression Amount is shown in Figure 7.

![Figure 7: MSE and PSNR relation](image)

The Values of the parameters are varied according to the each object in the image. A table describing the specific values is given below.

<table>
<thead>
<tr>
<th>Video</th>
<th>MSE</th>
<th>PSNR</th>
<th>SENSITIVITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road</td>
<td>0.484973</td>
<td>51.2736</td>
<td>99.9603</td>
</tr>
<tr>
<td>Man</td>
<td>0.20842</td>
<td>54.9414</td>
<td>97.1142</td>
</tr>
<tr>
<td>Lady</td>
<td>0.19359</td>
<td>550262</td>
<td>95.9027</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

In this project, we proposed an automatic Video Object Extraction approach which utilizes dynamic thresholding and Morphological Smoothening. The Background Subtraction is done by the Dynamic Threshold method. Sensitivity, Correlation, MSE and PSNR are used to extract the information of the image and the objects. Parameters like speed, velocity and angle are determined.
Morphological process and filtering also used effectively for unwanted pixel removal from the background.

Compared with state-of-the-art unsupervised Video Object Extraction methods, our approach was shown to better model the foreground object due to the fusion of multiple types of saliency-induced features. A major advantage of this proposed method is that we do not require the prior knowledge of the object of interest, nor the interaction from the users during the segmentation progress. Extensive testing has been performed on numerous video sequences to demonstrate the effectiveness of the proposed framework, and very satisfactory results have been obtained.

References


