Motion Segmentation from Surveillance
Videos using Varied Number of Frames

Girisha R & Murali S
P.E.T. Research Center
P.E.S.C.E., Mandya, Karnataka
Email: girisha_raj@yahoo.com, nymurali@yahoo.com

Abstract— Identifying moving objects from a video sequence is a fundamental and critical task in many computer vision applications. We develop an efficient adaptive segmentation algorithm for color video surveillance sequences, which uses varied number of frames in real time with non-stationary background. Traditionally, all segmentation methods proposed so far in the literature is based on fixed frames (i.e., 1, 2 or 3). We propose a new hybrid method, which uses Pearson correlation to segment two frames or multiple correlation to segment three frames depending on input. At runtime, segmentation is performed by checking color intensity values at corresponding pixels P(x,y) in two or three frames using temporal differencing. The segmentation starts from a seed in the form of 3x3 image blocks to avoid the noise. Usually, temporal differencing generates holes in motion objects. After subtraction, holes are filled using image fusion, which uses spatial clustering as criteria to link motion objects. The emphasis of this approach is on the robust detection of moving objects even under noise or environmental changes (indoor as well as outdoor).

Index Terms— Video surveillance, Motion segmentation, Temporal differencing, Pearson correlation, Multiple correlation.

I. INTRODUCTION

Analysis of human movement is currently one of the most active research topics in computer vision. Human Motion Analysis (HMA) includes detection, tracking and recognition of people. HMA can be classified into 3 categories [1, 2], namely low level vision (Detection), intermediate level vision (Tracking) and high level vision (Behavioral Analysis). The application domains where HMA can be applied are video surveillance, content based image retrieval, gait recognition etc.

The automated video surveillance system is expected to detect people and monitor their actions and subsequently need to analyze their behavior in order to prevent any untoward incidents. To analyze the behavior of a person in a given setup, the first step is human detection and tracking. Tracking involves detection of regions of interest in a frame and then finding frame-to-frame correspondence of each region’s location and shape.

Nearly, every system in the HMA starts with segmentation [1]; current motion segmentation methods mainly based on background subtraction or temporal differencing or optical flow or statistical methods [2]. Development of a reliable background models adaptive to dynamic changes in complex environments is still a challenge [1].

Segmented foreground objects generally include their shadows (Self and Cast) as a foreground object since the shadow intensity differs from the background [4]. The inclusion of shadows as foreground points can cause serious problems while extracting moving objects such as object shape distortion, object merging, and even object losses (due to the shadow cast over another object) which also affects surveillance capability while target identification and tracking. To obtain a better segmentation quality, motion tracking algorithms must correctly separate foreground objects from the shadows [4].

In this paper, we propose a novel approach to segment the motion objects from varied number of frames using statistical Pearson and multiple correlation methods. The algorithm obtains stable segmentation results even under varying environmental conditions. The rest of this paper is organized as follows: Section 2 presents a review of the recent and ongoing activity in the domain of motion segmentation; section 3 explains our proposed methodology; section 4 discusses experimental results, finally, section 5 concludes the proposed methodology.

II. RELATED WORK

The first step in HMA is the extraction of motion information through motion segmentation. Motion segmentation in video sequences aims at detecting regions corresponding to moving objects such as humans. A detected moving region provides a focus of attention for later processes such as tracking and behavior analysis. At present, all segmentation methods can be classified into four major groups such as background subtraction, temporal differencing, optical flow and statistical methods.

Background subtraction [3, 6, 7] is a commonly used class of technique to detect moving regions in an image. It is highly dependent on a good background model to reduce the influence of dynamic scenes derived from lighting and extraneous events such as clutter, shadow, occlusion etc. Temporal differencing [8, 9] makes use of pixel-by-pixel difference between two or three consecutive frames in an image sequence to extract moving regions. Temporal differencing is very adaptive to dynamic environment, but generally does a poor job of extracting the entire relevant feature pixels, e.g., generate holes inside the moving entities. Optical flow [10, 11] based motion segmentation uses characteristics of flow to detect independently moving objects even in the presence of camera motion. However most flow methods are computationally complex and very sensitive to noise.
Recently, some statistical methods [12, 13, 14] are proposed to extract change regions from the background and these methods are inspired by the basic background subtraction methods. The statistical approaches use the characteristics of individual pixels or groups of pixels to construct more advanced background models [14]. And the statistics of the background can be updated dynamically during processing. Each pixel in the current image can be classified into foreground or background by comparing the statistics of the current background model. The majority of the statistical methods proposed so far in the literature for background subtraction use either Gaussian or Kernel distribution to model the background [1, 2].

It is very common in real world that the shadow will appear as long as an object is in front of the light source. Shadows occur when objects totally or partially occlude direct light from a light source. According to the classification reported [4] shadows are composed of two parts: self shadows and cast shadows. The self shadow is the part of the object which is not illuminated by the light source. The cast shadow is the area projected on the scene by the object and further classified into umbra and penumbra. The umbra corresponds to the area where the direct light totally blocked by the object, whereas in the penumbra area it is partially blocked.

Fig. 1, depicts an overall overview of proposed system via PETS video of 2006, data set 7 of camera 3 for video frames 109, 112 and 115. The proposed system uses multiple correlation coefficient for three frames while segmenting motion objects from background. After segmentation, we apply inferential statistics to eliminate self shadows as shown in Fig. 1(c) and then cast shadows are eliminated using coefficient of variation [18] as shown in Fig. 1(f) to get final segmented motion objects. However, in this paper, we only present a simple, adaptive algorithm to segment motion objects from the background.

A. Outline of the Approach

Motion segmentation of foreground objects is an active area of research and a number of techniques have been developed over a decade. The proposed technique in this paper uses Pearson or multiple correlation to segment motion objects. Motion segmentation is done, by checking pixel by pixel disparity in RGB color space between two or three video frames simultaneously based on temporal differencing.

Shadows are modeled based on inferential statistics using difference in mean after foreground pixel
extraction. Self shadows are vague in nature, which gradually change intensity and have no clear boundaries. Those parts of the segmented motion objects, which are not illuminated by light source, become self shadow such parts are also eliminated by self shadow removal algorithm as shown in Fig. 1(c).

$$p(x,y)$$ and its eight neighbors $$(N8(p))$$ referred as pixel $$P(x,y)$$ in each frame and 3 frames $$P(x,y)$$ RGB (in total 27 values are used in each calculation.

Figure 5. Pixels selection for calculation.

The temporal differencing is very adaptive to dynamic environment, but generally does a poor job of extracting all relevant feature pixels i.e., segmented regions contains holes in motion objects. To fill the holes of the blob, we use spatial clustering (the criterion is spatial distance between pixels) first applied horizontally and then vertically as shown in Fig. 1(d). By doing this, the holes of the blobs generated by temporal differencing and self shadow elimination algorithm are completely filled. A drawback of this step is that cast shadow area may increase, if regions not linked accurately as shown in Fig. 1(e). The cast shadows removed if any, using co-efficient of variation, that algorithm [18] takes foreground pixels and then attempts to classify them into either motion object or cast shadows as shown in Fig. 1(f). After cast shadow elimination, once again spatial cluster is applied as shown in Fig. 1(g).

Motion segmentation is done in this paper, by checking pixel by pixel disparity (using equation (1) or (5)) in RGB color space between two (K and ) or between three (L, and ) frames simultaneously as shown in Fig. 1(a). Image subtraction is based on temporal differencing (frame gap can be three) using equation (4) or (8). However, this frame gap can also be varied (i.e., frame gap 1 or 2 or 3). But, extensive experiments conducted by us on PET5 data set revealed that, if we do temporal difference with successive frames as shown in Fig. 2 motion of the objects is almost negligible and its waste of processing time. On the other hand, if we increase frame gap beyond three frames than the objects moved very fast in the scene and generated unnecessary cast shadows in the corresponding difference images as shown in Fig. 4. The proposed motion segmentation algorithm in this paper robust to illuminations, complex backgrounds, adapts to dynamic environments and reflections can vary without significantly affecting the result.

III. PROPOSED SEGMENTATION ALGORITHM

A static camera observing a scene is a common case of a surveillance system. Detecting intruding objects is an essential step in analyzing the scene. Even though there exist a myriad of segmentation algorithms in the literature [1, 2]. Most of them follow a simple one or two-frame differencing except [5, 17] and nearly everyone assume that the background does not vary and hence can be captured a priori. This limits their usefulness in most practical applications.

Let, the pixel RGB value on any coordinate $$(x, y)$$ be denoted by $$p(x, y)$$ with $$x$$ in the range from 0 to and $$y$$ in the range 0 to . Where and are the size of the image in the X and Y directions, respectively. Let, a pixel $$p(x, y)$$ along with its eight neighbors $$(N8(p))$$ from now on referred to as pixel $$P(x, y)$$ as shown in Fig. 5. Motion segmentation is done using either two or three frames.

A. Motion Segmentation Using Two Frames

The proposed methodology to segment two frames, based on statistical correlation coefficient (Pearson Product-Moment Correlation Coefficient) method. Motion segmentation algorithm combines background subtraction with temporal differencing (Using equation (4)) to segment motion objects and uses color images in RGB color space. The distance measuring functions (1) and (3) are used to find current pixel say $$P(x,y)$$ either belongs to background or foreground. The two frames pixel $$P(x,y)$$ values are represented using correlation co-efficient $$(r_{ab})$$. Threshold $$(T)$$ is applied to co-efficient of determination $$(r_{ab}^2)$$ and if $$(r_{ab}^2)$$ is greater than the threshold value, then the pixel is classified as background.

$$r_{ab} = \frac{S_{ab}}{S_a \ S_b}$$  \hspace{1cm} (1)

Where,

$$S_{XY} = \frac{\Sigma_{XY}}{N_1}, \quad S_x = \sqrt{\frac{\Sigma_{X^2}}{N_1}}, \quad S_y = \sqrt{\frac{\Sigma_{Y^2}}{N_1}}$$  \hspace{1cm} (2)

Where, $$r_{ab}^2$$ is calculated for each $$(w_X \times b_y)$$ pixel of the frame. $$S_{XY}$$ is a covariance [16], $$S_x$$ and $$S_y$$ are standard deviation of the pixel $$P(x, y)$$, $$N_1$$ total number of RGB values in a pixel.

Let $$I_{K(x,y)}^1$$ and $$I_{1_1(x,y)}$$ are the $$K^{th}$$ and $$i_{1_1}^{th}$$ successive frames respectively. Then the difference images $$D_{K(x,y)}$$ and $$D_{1_1(x,y)}$$ are generated using equations (3), which contains motion objects of frames $$K^{th}$$ and $$i_{1_1}^{th}$$ respectively.

$$D_{k(x,y)} = \{0, 1\}$$  \hspace{1cm} (3)

$$D_{1_1(x,y)} = \{0, 1\}$$  \hspace{1cm} (4)

Where, $$I$$ is a predefined threshold value empirically chosen.

The Pearson correlation coefficient, value varies from -1 to 1 as shown in Fig. 6(b) depending on linear relationship between two pixels RGB values. If deviation of pixel $$P(x,y)$$ RGB values varies in the same direction,
then the product is positive. If the deviation varies in the opposite ways, then the product is negative. If the deviation in pixel \( P(x,y) \) RGB values does not change in a linear pattern, then there is a haphazard association of plus and minus deviations and the is near zero.

The, if interpreted in terms of its squared value (i.e., ) is an estimate of the proportion of the total variation in pixel \( P(x,y) \) RGB values which is explained by the linear relationship between the two values. This proportion is usually converted to a percentage, which is known as the correlation coefficient of determination as shown in Fig. 6(c).

### B. Motion Segmentation Using Three Frames

Multiple correlation, measures the degree of linear relationship between three pixels \( P(x,y) \) RGB values from temporal differencing frames \( K \), \( m_1 \) and \( m_2 \), as shown in Fig. 5. We assume linear relationship between three pixels RGB values at every position \((x,y)\) in all input video frames.

\[
R_{a,b,c} = \sqrt{\frac{\sum_{h_y} \left( w_x \times h_y \right)}{N_2}}
\]

Where, \( R_{a,b,c} \) (where, \( a = L_1 \)) is the coefficient of multiple correlation between the dependent frame \( a \) pixel \( P(x,y) \) RGB value by keeping \( b \) and \( c \) frames pixel \( P(x,y) \) RGB values constant, like this \( R_{a,b,c} \) is calculated for each \( (w_x \times h_y) \) pixel of the frame. Where, \( \tau_{ab}, \tau_{ac} \) and \( \tau_{bc} \) are correlation coefficient \([15]\) computed using equation (6).

\[
S_{GH} = \frac{\sum GH}{N_2}, \quad S_G = \frac{\sum G^2}{N_2} \quad \text{and} \quad S_H = \frac{\sum H^2}{N_2}
\]

Where, \( G = \{a,b\}, \quad H = \{b,c\} \) and \( G \subset H \) \( S_{GH} \) is a covariance \([16]\), \( S_G \) and \( S_H \) are standard deviation of the pixel \( P(x,y) \) depending on \( \tau_{ab}, \tau_{ac} \) and \( \tau_{bc} \). \( N_2 \) total number of RGB values in a pixel.

Let \( J^{m_1} \) and \( J^{m_2} \) are the \( L \)th, \( m_1 \)th and \( m_2 \)th corresponding frames respectively. Then the difference images \( E^L \), \( E^{m_1} \) and \( E^{m_2} \) are generated using equation (7) which contains motion objects of frames \( L, m_1 \) and \( m_2 \) respectively.

\[
E_{m_1} = \{0, \text{RGB of} \ J^{m_1} \text{if} \ R_{a,b,c} \}
\]

Where, \( m = \{L, m_1, m_2\} \),

\[
m_1 = \{(L + 1), (L + 2), (L + 3)\}
\]

\[
m_2 = \{(L + 1), (L + 5), (L + 6)\}
\]

\( L = (9n + 1) \)th frame, \( n \geq 0 \) and \( n \in N \)

Where, \( R_{a,b,c} \) is a predefined threshold value empirically chosen.

A coefficient of multiple correlation, lies between 0 and 1 as shown in Fig. 7(b) depending on linear relationship between three pixels \( P(x,y) \) RGB values as shown in Fig. 7(a). The \( R_{abc} \) if interpreted in terms of its squared value (i.e., \( R_{abc}^2 \)) is an estimate of the proportion of the total variation in pixel \( P(x,y) \) RGB values which is explained by the linear relationship between the three values. This proportion is usually converted to a percentage, (100%)\( R_{abc}^2 \), which is known as the coefficient of multiple determination as shown in Fig. 7(c).
Self shadows are modeled using inferential statistics after foreground pixel extraction. After, spatial clustering, the cast shadows removed if any, using co-efficient of variation, that algorithm [18] takes foreground pixels and then attempts to classify them into either motion object or cast shadows as shown in Fig. 1(f). After cast shadow elimination, once again spatial cluster is applied as shown in Fig. 1(g).

IV. EXPERIMENTAL RESULTS

In this section, we analyze the performance of segmentation algorithm for surveillance video frames of IEEE PETS1 (Performance Evaluation of Tracking and Surveillance) 2000, 2001, 2004 and 2006 data sets. System has been tested using several sequences of PETS data set among which there are different tracking scenarios including indoor and outdoor environments, varied number of people. Results shown here are raw results, without any post treatment. Moreover, represents a snapshot of the algorithm results and are typical of the performance throughout the sequences. For each environments, parameters were set once.

Thirty challenging PETS video sequences are used to test the proposed system. Each of the sequence contains 500 to 4000 frames and resolution varied from one sequence to another sequence. Here the comparison of this approach is made based on the visual interpretation, i.e., by looking at processed frames provided by the algorithm. The false negatives (the number of foreground pixels that were missed) and false positives (the number of background pixels that are marked as foreground).

A false positive pixel either belongs to self or cast shadows as shown in Fig 8. Shadows occur when objects totally or partially occlude direct light from a light source. The self shadow pixels are those which are not illuminated by direct light source. If we decrease the value false negative increases and by reducing false positive pixels as shown in Fig 8. The cast shadow will increase, if the objects move fast in the scene because cast shadow points are usually adjacent to object points and are merged in a single blob on the edge of the moving objects [4]. In addition, cast shadow occurs only at run time (as objects move in the scene). However, self shadow remains almost constant because shadow intensity differs from the foreground as shown in Fig. 8.

Background subtraction [3, 6, 7] and optical flow [10, 11] methods that rely only on color information will most probably fail to detect the moving object correctly because of the similar color of the foreground and the background. However, in this paper we used temporal differencing where background is modeled with Pearson or multiple correlation coefficients; it manages the multi modal backgrounds robustly. Figs. 9 through 15 demonstrate challenges in each sequence and the effectiveness of our algorithm.

In outdoor environments, illumination changes rapidly due to fast changing weather conditions. Figs. 11 and 14 show frames of an outdoor video sequence in which whole image illuminated by direct sun light. This is a particular challenging situation since the object motion is harder to detect in respect to the other cases, especially when the object distance from the camera is large.

There are always variations in the illumination parameters between two frames of the same scene taken even at different times of day. Figs. 9 through 10 show images in indoor environment, corresponding to color video sequences acquired in varying range of fluorescent lighting systems with complex illumination. The drawback of the proposed system is spatial clustering steps. If the objects move close to each other in the scene, they will be linked each other, and it reinserts some of the eliminated cast shadows to the segmented motion objects as shown in Figs. 9 through 14.

V. CONCLUSIONS

In this paper, we have proposed a system capable of segmenting moving objects from surveillance video using varied number of frames. We employed a novel method, which uses Pearson or multiple correlation coefficient to segment two or three video frames simultaneously with each other in temporal differencing method. We combined simple spatial cluster and image fusion to fill the holes in the segmented motion objects.

Extensive experiments conducted on different data sets of PETS reveal that results are stable and satisfactory. The algorithm is robust to large variation in intensity (such as those caused by fluorescent lighting and as well as direct sun light), due to temporal differencing method.

REFERENCES


---

1 PETS data can be found at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/


ACKNOWLEDGMENT
Authors would like to thank DRDO (Aeronautical Research & Development Board). This publication is based on the project sponsored by DRDO file number DARO/08/2021515/M/I.

Figure 9. Results from PETS 2004 data set.

Figure 10. Results from PETS 2006 data set.

Figure 11. Results from PETS 2000 data set.

Figure 12. Results from PETS 2000 data set.

Figure 13. Results from PETS 2001 data set.

Figure 14. Results from PETS 2001 data set.