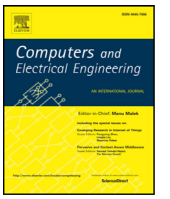
Computers and Electrical Engineering 81 (2020) 106520

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Background modeling and foreground extraction in video data using spatio-temporal region persistence features☆

Satyabrata Maity[a,*], Amlan Chakrabarti[b], Debotosh Bhattacharjee[c]

[a] CSE, ITER, SOADU, Bhubaneswar, Odisha, India
[b] A.K.Choudhury School of I.T, University of Calcutta, JD-Block-II, Sector-III, India
[c] Dept. of CSE, Jadavpur University, 188, Raja SC Mallik Road, Kolkata-700032, India

ARTICLE INFO

Article history:
Received 9 January 2019
Revised 21 November 2019
Accepted 21 November 2019
Available online 10 December 2019

Keywords:
Background modeling
Spatiotemporal region persistence (*STRP*)
Block-wise bin histogram (BBH)
Contrast normalization
Anisotropic smoothing: Foreground extraction

ABSTRACT

The proposed work describes an efficient methodology for adaptive background modeling and foreground extraction from video data using the newly proposed Spatio-temporal region persistence (STRP) descriptor. The STRP background descriptor includes block-wise statistics of intensity bins and their temporal persistency. Blockwise feature extraction helps to consider the local changes, while intensity bins provide consistent output for a group of similar intensities in an intra-regional sense. In this work, we have tried to minimize the effect of different video irregularities like dynamic background, ghosting effect, change in illuminations, video noise, etc. Additionally, adaptive threshold selection and regular adjustment of modeled background descriptors make the procedure robust. Two benchmark datasets, Changed Detection and Scene Background Modeling, and Initialization (SBMI) have been used to verify the efficiency of our work. The results and comparative studies with the related works justify the effectiveness of our proposed technique.

1. Introduction

With the introduction of web-cam and closed-circuit television, video-based applications have been snowballing commendably due to the fast technological advancements in video capturing and storing devices. In the present scenario, the enormous use of video data in plenty of applications leads to the creation of a large number of videos. On the other hand, a video includes Spatio-temporal redundancy to improve the visual impact as well as the quality. Spatial redundancy enhances the clarity of information by increasing the number of pixels per unit area and temporal redundancy improves the quality of smoothness in changing the foregrounds by escalating the frame rate. The redundancy is inversely proportional to the entropy value of the underlying information in a video. Hence, the extraction of essential information to understand the underlying facts of the video is extremely difficult in such cases, which is required to analyze the video contents for further course of action in numerous video-based applications. Most of the time, applications like home, office and shopping mall monitoring, outdoor and indoor surveillance generate a massive amount of no-informative-videos, i.e. the videos contain no effective information, and storing or analysis of the same is wastage of time and resource.

Video information is mostly categorized into two distinct types, specifically background or rigid part, and foreground or moving part. A substantial component of a video is termed as background if it remains unchanged or static throughout

☆ This paper is for regular issues of CAEE. Reviews processed and recommended for publication to the Editor-in-Chief by Area Editor Dr. E. Cabal-Yepez.

* Corresponding author.
E-mail address: satyabrata.maity@gmail.com (S. Maity).

https://doi.org/10.1016/j.compeleceng.2019.106520
0045-7906/

the frame sequence. On the other hand, the moving or dynamic area of a video, which is the active part throughout the scene is termed as the foreground. Therefore, the background is redundant while the foreground is informative. Hence, the elimination of background is required to extract the foreground information for further processing and analysis. Thus, effective background estimation is one of the crucial prerequisites for most of the video processing tasks. There can be several kinds of environmental conditions in outdoor and indoor scenarios in the real-world environment, which creates difficulties to model the background accurately. Some of the common challenges in background modeling, mentioned in [1] and [2], are described as follows:

- **Dynamic background** [1]: When some parts of the background have regular ambient motions like moving tree leaves, fountain, waves in the sea, etc., they are termed as dynamic background.
- **Ghosting effect** [2]: When a part of non-background objects is estimated as a part of the modeled background, it leaves its presence in a time of foreground extraction until the background is updated. This phenomenon is called the ghosting effect, which happens due to slow or late moving foreground(s) while estimating background.
- **Gradual illumination changes** [1]: The change in illumination due to soft deviations of light gradually. For instance, the variation in outdoor illumination over different times of the day.
- **Sudden illumination changes** [2]: An abrupt change in light like switching off/on the light in a room environment, which is very difficult to model and it leads to false detection of foregrounds.
- **Video noise** [2]: Video signal includes some noise in most of the cases. Background subtraction approaches for video surveillance have to cope with the effects of different types of noise.

Considering the challenging situations as described above, the estimation of a proper background in every situation is not an easy task. These situations can catalyze all sorts of possibilities for false detection, which can disturb the entire procedure of video analysis.

The proposed approach includes the Spatio-temporal region persistence (*STRP*) descriptor to estimate the background for handling challenging situations. Region persistence features provide the immobility of information throughout the frames of a video. Each frame is spatially divided into several blocks and bin-wise statistics represent a block. The merging of bin-wise statistical features temporally concerning a fixed block provides the *STRP* features.

The key contributions of this research work can be summarized as follows:

- Proposal of a statistical background modeling technique based on temporal persistence and occurrence probability.
- The newly introduced Spatio-temporal Region Persistence (*STRP*) feature descriptor, which successfully estimates the occurrence distribution of intensity bins in a block-wise fashion.
- Adaptive threshold selection, the decision of foreground extraction depends on the persistence measurement, and updating background at regular intervals.
- Adaptive over time i.e. if gradual changes occur among the frames they are adapted with the current modeled background. Also, scene changes are detected when an abrupt change occurs and sustains over a specified period. The proposed work re-estimates the background after tracking the change in the scene.
- The proposed method can handle several irregularities in a video like ghosting effect, noisy conditions, bootstrapping, camouflaged conditions.

Although deep learning-based methods have been very popular in different scientific areas and mostly in computerized medical image processing described in [3], due to their efficient formulas, there are some drawbacks of these approaches. For instance, training stages are time-consuming, parameters and batch size should be chosen carefully. Therefore, the proposed method in this work is based on adaptive threshold selection and auto-update techniques.

The rest of the paper is organised as follows: Section 2 provides an overview of the relevant works in the literature followed by Section 3 which delves into the details of the propsed work. Next, Section 4 gives the result and analysis for the proposed work. Finally, Section 5 concludes the paper.

2. Related work

Related literature of background modeling approaches is expanded in many directions. Some of the key approaches related to our proposed work have been included in [4], [5], and the curious readers are redirected to the same for the review on the related research domain. The main work is segmentation of background and foreground information. On the other hand, the background is the most static area of the frames throughout the scene. Thus, the modeling of the most static areas among the frames of any scene of a video will produce the background. The static part of a video can be defined by the redundancy in the information of any location along the temporal direction. Three broad kinds of strategies [6] are used to estimate the background of a video in the related domain. They are:

1. Pixel level processing: This kind of approach the approximate intensity of each of the pixels individually using temporal statistics among consecutive frames to estimate the static area among the frames of a scene and thus helps in the modeling of the background frame. Pixel-wise modeling is very noise sensitive. These kinds of strategies apply pixel averaging, median filtering, fuzzy-based selection, temporal histogram, the mixture of Gaussian [7], kernel density estimation [8], etc. to estimate the intensity value of a pixel.

2. Frame level processing: Unlike pixel-level processing, these types of approaches approximate the background frame using global features of consecutive frames in the scene. Foreground extraction is then done by comparing the features of the current frame with that of the modeled background. These kinds of strategies apply feature extraction techniques like global histogram [9], optical flow [10], etc. These strategies are effective in the estimation of global changes, but they show poor performances in the estimation of the local changes. Although, this kind of approaches are hardly affected by noise, yet, the local changes may be overlooked.
3. Region level processing: These types of strategies combine region-based features to model the background of all past frames and compare the same concerning the current frame for extracting foreground information. This is the best strategy amongst the three, but it has high time complexity. Hence, our proposed work uses a block-level method to estimate the background of a scene, which is less noise prone, considers local properties and does not require any sophisticated algorithm for region segmentation.

The proposed background modeling approach includes some key ideas to handle challenging situations.

- Extraction of active pixels between every two consecutive frames while computing the features. Active pixels are those which determine moving areas.
- Block-based processing instead of pixel-based processing to reduce the effect of noise. Moreover, this kind of processing is sensitive to local changes in the temporal direction.
- Generally, a trivial change in intensity values of near by pixels carries the similar information. It would be more effective to take the statistical response of a bin to estimate the features of a block instead of individual frequency.
- A statistical measure for both active and static bins for each of the blocks, so that the removal of active pixels can be done despite irregularities.
- Selection of threshold dynamically according to the contrast of a video, as the change in illumination, has a direct relation to the contrast of an image.
- Occurrence of an object in the same location (block) in consecutive frames defines the persistence. The effectiveness of a region is measured in terms of that persistence rather than occurrence probability.

3. Proposed work

The proposed methodology intends to design a feature-based background modeling and its elimination towards foreground extraction. Moreover, this method is adaptive over time and detects the scene changes in the video information. The work-flow of the whole methodology is described in Fig. 1, which is primarily categorized into four distinguishing parts:

- Preprocessing
- *STRP* Feature extraction and background modeling
- Background elimination and foreground extraction
- Background checking or Scene change detection

3.1. Preprocessing:

The preprocessing step is used to minimize the effect of noise for efficient outcomes in the proposed system. It includes contrast normalization to minimize the effect of illuminations and anisotropy based smoothing to reduce the intra-regional variances. Illumination can have a large impact on the effective intensity values and the normalization, in contrast, can reduce the minor variability of light. On the other hand, minute detailing of intra-regional information can increase the effective cost for processing, so the anisotropy based smoothing is used which can reduce the intra-regional variability to some extent.

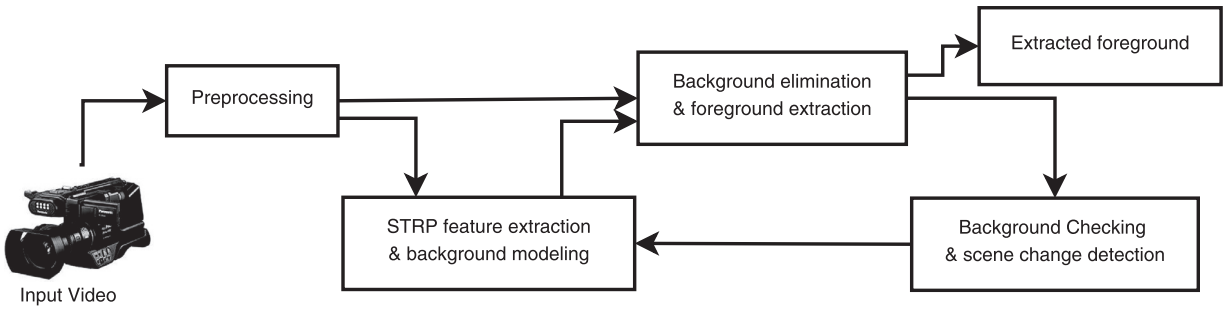


Fig. 1. Framework of the proposed approach.

3.1.1. Anisotropy based smoothing

Any image contains high and low-frequency elements for defining the intra-regional and edge area respectively. General smoothing filters smooth all areas of an image, whereas the anisotropic filter, proposed by Perona and Malik [11] using Eqs. (1) and (2) smoothens only the low-frequency areas while stopping the operations in high-frequency edge areas.

$$I_t = div(c(x, y, t)\nabla I) = c(x, y, t)\Delta I + \Delta c.\nabla I \tag{1}$$

$$c(\|\nabla I\|) = \frac{1}{1 + (\frac{\|\nabla I\|}{k})^2} \tag{2}$$

where, $I(., t) : \omega \to \mathbb{R}$ be a family of 2*D* images, and $\omega \subset \mathbb{R}$ denotes a subset in 2*D* plane. *div*, Δ, and ∇ are the divergence, Laplacian, and gradient operators respectively. $c(x, y, t)$ is the diffusion coefficient, which is generally derived from image gradient, at spatial coordinate (x, y). k is a constant that controls the sensitivity of the edges. $c(x, y, t) =$ image obtained after a diffusion time t, $\|\nabla I\|$ is the gradient magnitude, and $c(\|\nabla \perp\|)$ is an "edge-stopping" function. This function is used to satisfy $c(x) \to 0$ when $x \to \infty$ so that the diffusion is "stopped" across edges.

Many works have been done on the edge stopping function. The interested readers are requested to follow the concepts as described in [12]. A modification of the Perona-Malik anisotropic filter is described in [13], which has used the concept of directional Laplacian to reduce the stair-casing effect for low contrast images to preserve the edges.

Time complexity of anisotropy based smoothing depends on the number of iterations. Since frames of the same scene contain more similar information, there is no requirement for higher-order smoothing. Hence, we restrict the number of iterations to two to save complexity.

3.1.2. Contrast normalization

The luminance value of black is low and it gradually increases while it is transforming towards white. Michelson et al. [14] defined the contrast as $(L_{MAX} - L_{MIN})/(L_{MAX} + L_{MIN})$, where L is the luminance value L_{MAX} is the maximum, and L_{MIN} is the minimum luminance value in a given image respectively. It is always found that the visibility of a given image has a direct relation with the implicit contrast of the same as discussed in [14]. This step focuses to balance the illumination using contrast normalization. Let, the upper and lower limit of a normalized image be *ul* and *ll* respectively; lower and higher values of the current frame are *lv* and *hv* respectively, and P_{in} and P_{out} are the input and output pixel values. The normalization is done using Eq. (3).

$$P_{out} = (P_{in} - lv)\left(\frac{ul - ll}{hv - lv}\right) + ll \tag{3}$$

3.2. STRP feature extraction and background modeling

STRP is a block-wise background descriptor that includes *bin response vector* of each block in a frame. A *bin response vector* contains bin wise occurrence distribution of intensities as bin histogram in the block. The whole intensity range is divided into many small equal-sized groups termed as intensity bin. Since pixels of the same region have similar intensity values, bins are used to suppress trivial variations in intra-regional intensities. Background modeling is done using the following steps:

- Block wise spatial feature or block-wise bin histogram (*BBH*) extraction
- Temporal merging of extracted features to generate the *STRP* descriptor
- Background estimation based on *BBH* features and *STRP* descriptor.

Every frame of a video is divided into many equal-sized squares termed as blocks. The frames, blocks and the bin descriptor of each block are shown in Fig. 2.

Block-based processing prioritizes local features, which are mostly ignored in frame-based techniques. Besides, the presence of noise in frames can affect pixel-based processing but it is reduced effectively in block-based processing.

3.2.1. Block wise spatial feature or BBH extraction

Let us assume that $(l \times m)$ pixels is the size of a frame and it is divided into many disjoint blocks with size $(p \times q)$. The number of blocks for each frame is $m1 \times n1$. The size of a block and the efficiency of a method are proportional to each other; if the block size increases, processing speed increases and vice-versa. A bigger sized block always overlooks the local disparities whereas a tiny block is noise prone and cannot resist small or negligible changes. Here, the resolution of an image frame of the video is considered to set the block size which is not less than 8×8.

Let B_{ij} be one of the blocks located at the i^{th} row and the j^{th} column in a frame. The bin histogram of the block B_{ij} is $\mathbb{Q}_{ij}$, which includes the bin histogram and bin-wise weighted average. All $\mathbb{Q}_{ij}$ is a two-column vector of length l, where each row contains frequency and the weighted average of the corresponding bin. The concatenation of all the $\mathbb{Q}_{ij}$ provides the *BBH* descriptor of a frame (β). Counting sort strategy is used to extract the frequency of each bin as given in Eq. (4). *BIN* is the total number of bins considered for the feature vector and *BS* is the size of each bin. Here, x represents an intensity

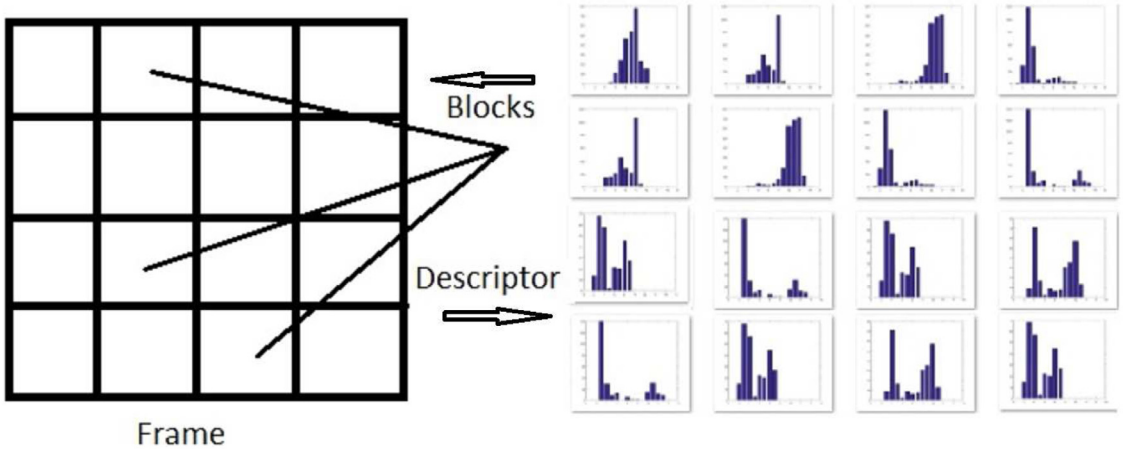


Fig. 2. Frame, blocks and bins.

value of the block under consideration and l ranges from 1 to *BIN*. The weighted mean of any bin for a particular block is computed dynamically to save the computational cost using Eq. (5), where, x_i is the intensity value at index i, $\bar{x}_{i-1}$ is the average mean of previous $i-1$ elements from the same bin observed in the same block. This simple change efficiently reduces the complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ to compute the weighted mean.

$$\mathbb{Q}_{(l,1)} = \text{COUNT(x)}, \ \ \forall\, l \le (x/BIN) \le l+1 \tag{4}$$

$$\bar{x}_i = \bar{x}_{i-1} + (x_i - \bar{x}_{i-1})/i, \ \ where\ i > 1\ \&\ \bar{x}_1 = x_1 \tag{5}$$

Fig. 3 describes the visual comparison of the general histogram and block-wise bin histogram (*BBH*) of a particular block. Both index profiles are very similar which refers to the sustainability of fundamental property despite applying bin histogram. However, the size of a block descriptor is effectively small in the case of bin histogram.

3.2.2. Temporal merging of extracted features

The objective of this step is to estimate the temporal persistence of all β to form the *STRP* descriptor for the initial background of a video. Suppose a t number of consecutive frames is considered to compute the background *STRP* descriptor. $\mathbb{Q}$ features from the first block of all k frames are merged temporally to extract the first *STRP* descriptor of the background frame. Similar temporal merging is followed in the remaining blocks to pursue the *STRP* descriptor of the background frame denoted with β_{BAK}.

The static bin will be more responsive than the dynamic bin of any block to model the background. The spatial distribution of intensity bins is stored in β for the corresponding frame. The temporal persistence response of a bin for a particular block is measured to define the background descriptor. The merging of features is done using the simple bin-wise averaging of the extracted features in temporal direction using Eqs. (6) and (7). The value of t can be tuned according to the situations, but it will not be affected much as the background is updated according to the time. The proposed method takes the value of t, the same as the frame rate of the video.

$$\beta_{BAKij}(\mathbb{Q}(l,1)) = \left(\sum_{k=1}^{t}(\beta_{ijk}(l,1))\right)/t \tag{6}$$

$$\beta_{BAKij}(\mathbb{Q}(l,2)) = \left(\sum_{k=1}^{t}(\beta_{ijk}(l,2))\right)/t \tag{7}$$

3.2.3. Background frame modeling based on persistence checking

The proposed approach computes the intensity value of the modeled background frame (*BAK*) using the temporal pixel statistics for any particular location by combining it with β_{BAK} features of the corresponding block. A temporal *BBH* in t number of frames, which are in the same location and their weighted average are stored in $\mathbb{T}$ using Eqs. (8) and (9). $\mathbb{T}(l,1)$ holds the temporal *BBH* and $\beta_{BAKij}(l,1)$ holds *BBH* of the corresponding block. The maximum value after combining these two *BBH* responses sets the background value of the corresponding location as shown in Eqs. (10) and (11).

$$\mathbb{T}_{(l,1)} = (COUNT(X))/t, \ \ \left\{\forall X \in FRM_{ijt} \mid l = X/BIN\right\} \tag{8}$$

$$\mathbb{T}_{(l,2)} = \sum(X)/\mathbb{T}_{(l,1)} \ \ \left\{\forall X \in FRM_{ijt} \mid l = X/BIN\right\} \tag{9}$$

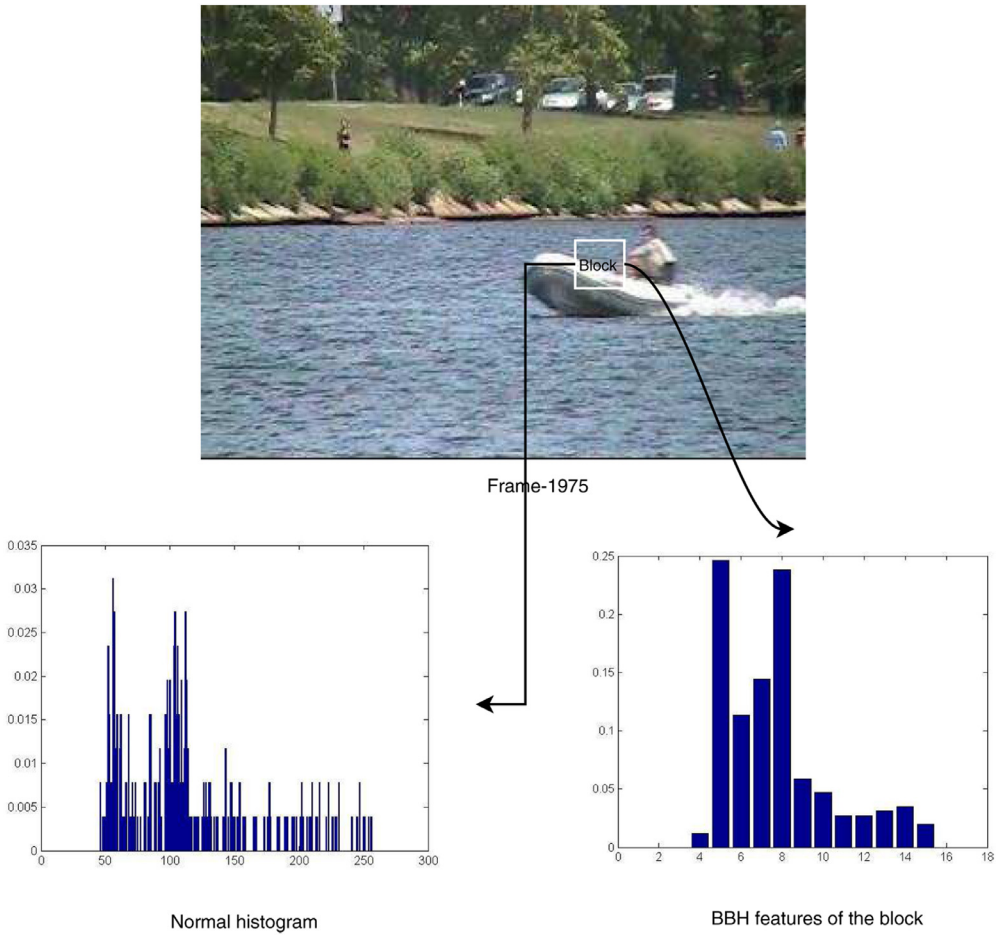


Fig. 3. Normal histogram of the highlighted block and the corresponding bin descriptor of a block.

$$MI = MAX_{l=1}^{BIN}(\mathbb{T}_{(l,1)} \times \beta_{BAKij}(l,1) \tag{10}$$

$$BAK_{xy} = \mathbb{T}(MI, 2) \tag{11}$$

A comparison between the *BBH* descriptor of a block in different temporal frames is described in Fig. 4. The background frame is shown in the middle of the first row in Fig. 4, on the left side same block is highlighted without foreground (Frame-1922), and the right side includes foreground (Frame-1942) in the highlighted area. The row below in Fig. 4 is a block descriptor wise comparison between the background frame and with those two cases respectively. Blue colored bar signifies the block descriptor bins of the modeled background frame while yellow refers to the block descriptor bins of frames in the video. Block descriptors are very close to each other while there is no foreground irrespective of the dynamic background (the wave in the water). The descriptors largely disagree with the presence of the foreground.

3.3. Foreground extraction

The modelled background and current frame are compared with respect to a suitable threshold to extract the foreground of any frame in the scene. The prime objective of this threshold is to estimate the temporal changes in information among the frames.

3.3.1. Threshold ($\mathrm{TH_{chg}}$) computations

A threshold is computed based on the contrast matrices of t consecutive frames since the variability, in contrast, change the visual impact of the image. Contrast matrix (MAT_{con}) of any image contains the maximum divergence (η) value of each pixel in terms of the maximum difference ($MAX(\eta)$) among all of its neighbours using Eqs. (12) and (13). The mean contrast of a frame (M_{ct}) is as given in Eq. (14). The average M_{ct} of first t consecutive frames of any scene defines the threshold TH_{chg},

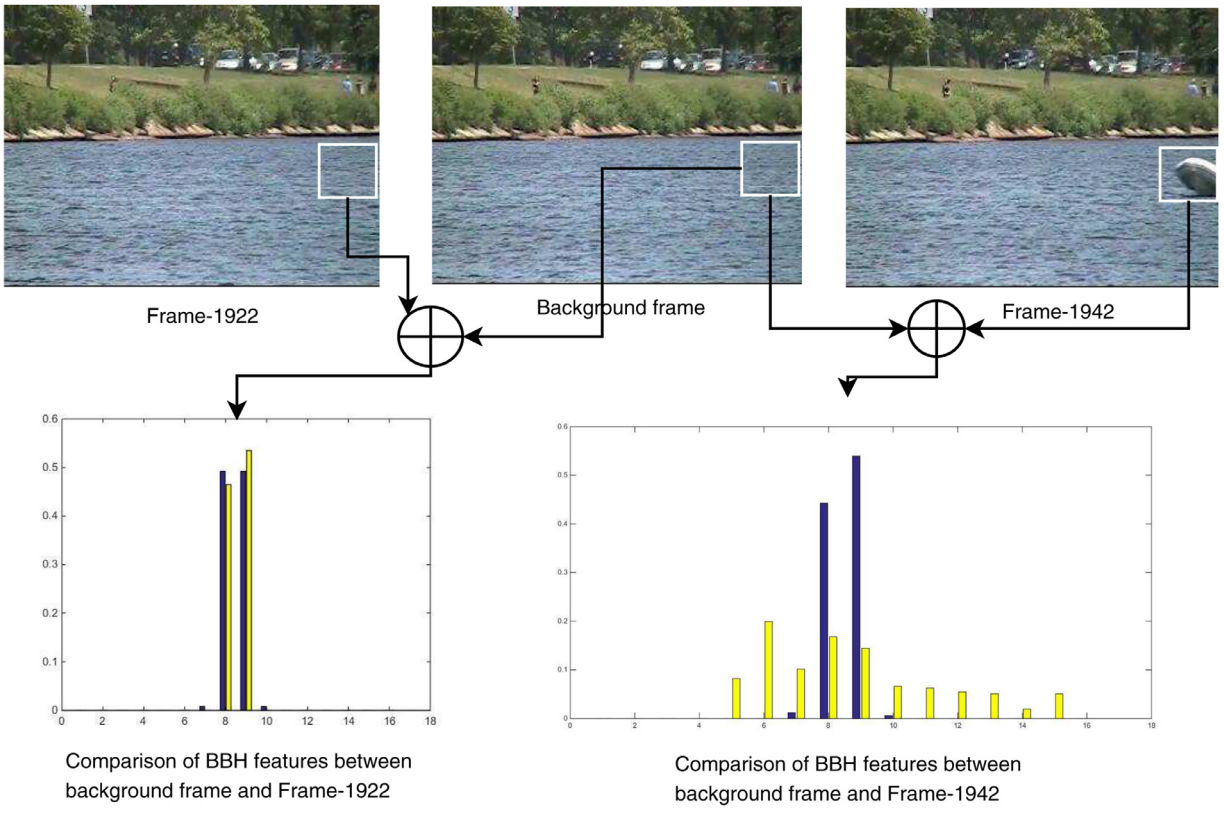


Fig. 4. A comparison among BBH descriptor of a block in temporarily different frames. First row contains the frame without foreground object, modeled background frame, and the frame with foreground object. Highlighted areas are same blocks in temporarily different frames. The third row contains comparative study among the block descriptor of frames with respect to background frame.

which is computed using Eq. (15). The value of parameter t is set to the frame rate of the video. Threshold selection takes place with the change of scene, which makes the procedure adaptive with respect to the scene elements. It is assumed that one second is enough to sense the variable condition on the scene. This parameter can be changed for more varying conditions but as we increase the value of t, the time complexity for the same is also enhanced. This computed threshold assists us to estimate the active areas in consecutive frames and the initial foreground estimation by eliminating the background.

$$\bigtriangledown I(x,y) = |I(x,y) - I(x+i, y+j)| \tag{12}$$

where -1 $\leq$ i $\leq$ 1, and -1 $\leq$ j $\leq$ 1

$$MAT_{con}(x,y) = MAX(\bigtriangledown I(x,y)) \tag{13}$$

$$M_{ct} = \sum_{i=1}^{m}\sum_{j=1}^{n} MAT_{con}(i,j)/(m \times n) \tag{14}$$

$$TH_{chg} = \sum_{i=1}^{t} M_{ct}(i)/t \tag{15}$$

3.3.2. Active region extraction

The active region descriptor embraces the active areas of the current frame by eliminating the background. Difference between current frame (FR_t) and BAK with respect to threshold TH_{chg} results in the background frame difference which is termed as Δ computed using Eq. (16).

$$\Delta(i,j) = \begin{cases} 1 & \text{if}(\mid BAK(i,j) - FR_t(i,j) \mid > TH_{chg}) \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

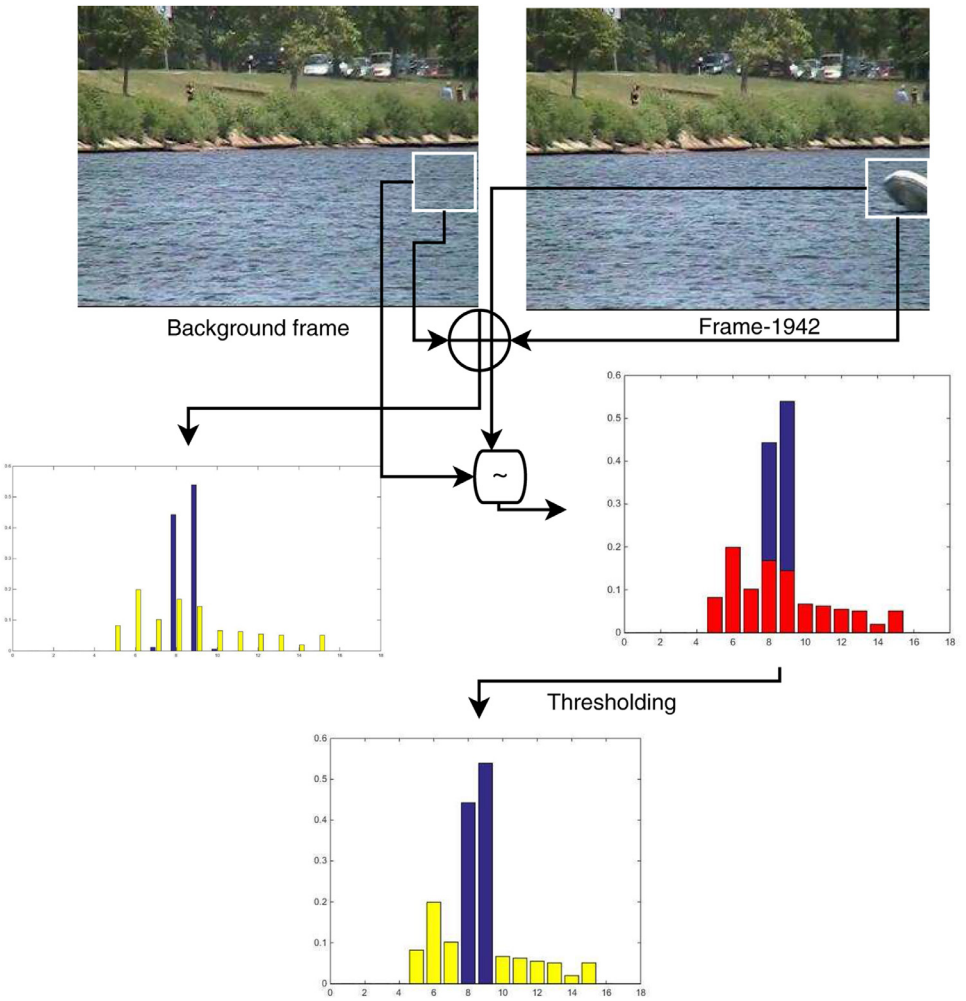


Fig. 5. Estimation of foreground features. Row-1: The same block of background frame and Frame-1942. Row-2: Comparison of both the features. Row-3: Foreground BBH.

3.3.3. Foreground descriptor extraction for current frame

BBH features of any frame FR_t and (β_{BAK}) are used to estimate the foreground feature descriptor ϕ using Eq. (17). If both are equivalent, then the block is static, else foreground elements are present in that block. ϕ determines the foreground bins with '1', '0' otherwise.

$$\phi_{ij}(\mathbb{Q}_l) = \begin{cases} 1 & IF, ((\beta_{FR_{tij}}(\mathbb{Q}_{(l,1)}) - \beta_{BAK_{ij}}(\mathbb{Q}_{(l,1)})) > THC) \\ 1 & IF, \beta_{FR_{tij}}(\mathbb{Q}_{(l,1)}) > 0, \ and \ \beta_{BAK_{ij}}(\mathbb{Q}_{(l,1)}) = 0 \\ 0 & otherwise \end{cases} \tag{17}$$

$$THC = STD_{k=1}^{BIN}(\beta_{BAKij}(\mathbb{Q}_{(l,1)})) \tag{18}$$

The bins of *BBH* feature of any block are included in the foreground *BBH* if those consists of either 0 share in the background *BBH* or the bin share of the current frame is significantly large with respect to the background one (Eq. (18)). Estimation of foreground BBH features by comparing the *BBH* features of a block of background frame and the corresponding block of another frame where the foreground exists is shown in Fig. 5. The bin with a yellow bar of the final figure will contain 1 in ϕ since those are the foreground bins.

3.3.4. Relative activity descriptor

The relative activity descriptor describes the active area between a pair of consecutive frames i.e. the active regions of a frame pair to complete the corresponding action. Pixel wise difference between two consecutive frame pairs with respect to the corresponding changing threshold is stored in δ. δ is the relative activity descriptor and is basically a matrix of same

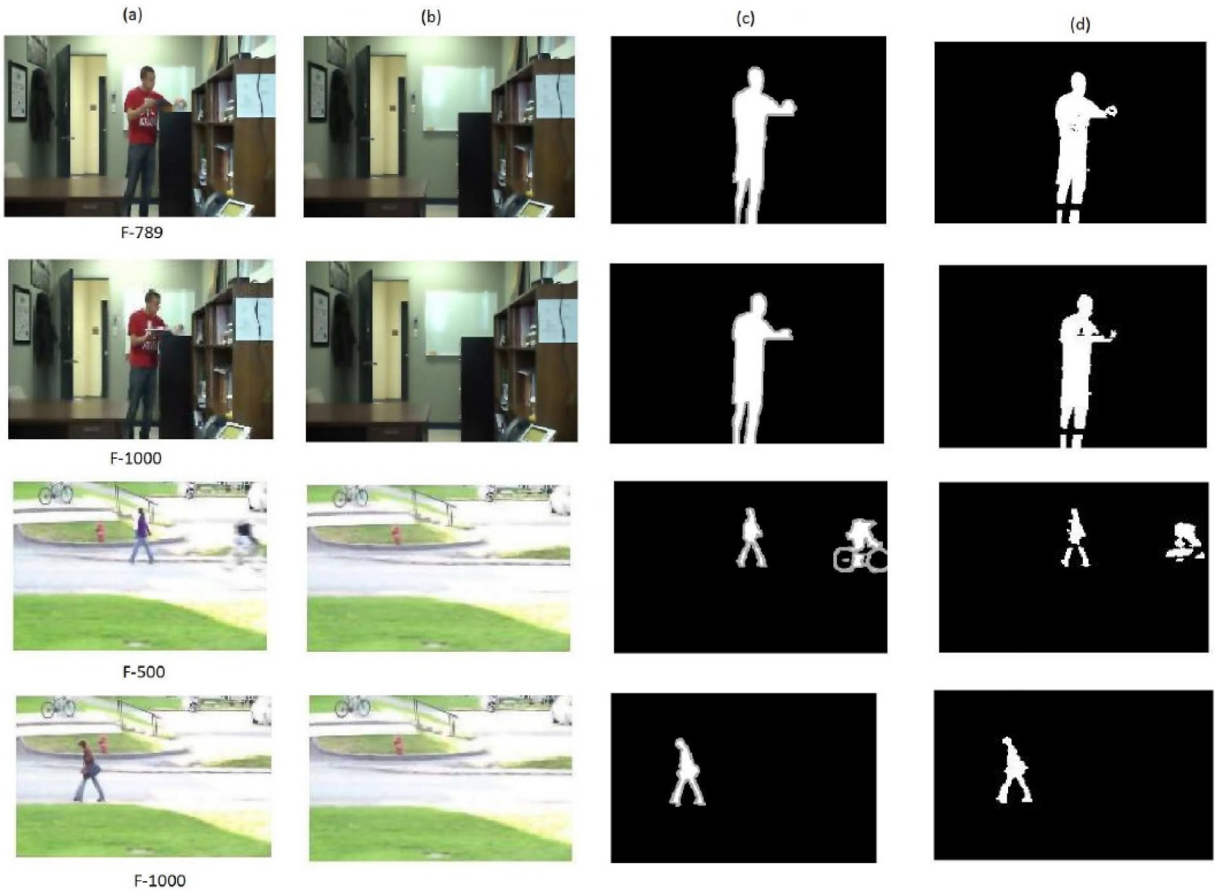


Fig. 6. Foreground detection results: (a)The original frame, (b) the modeled background, (c) Ground Truth and (d) our results after eliminating background .

size like any other frame consisting of binary value 1, when the absolute difference is greater than TH_{chg} and 0 otherwise as given in Eq. (19).

$$\delta(i,j) = \begin{cases} 1 & \text{if } (\mid FR_{t+1}(i,j) - FR_t(i,j) \mid > TH_{chg}) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

3.4. Foreground extraction

The changes among consecutive frames mainly occur due to foreground movements but there are several other conditions as well which may lead to false foreground detection. So, the elimination of background frame from the current frame cannot produce the foreground efficiently. Hence, the proposed approach uses ϕ, Δ and δ to extract the foreground.

3.4.1. Binary foreground frame extraction

The ϕ includes responses from foreground bins with respect to the block. If all the bins of a block are zero, no further processing towards foreground extraction is required for that block as there is no foreground. On the other hand, if all the bins of a block are one, then all the pixels are included in the foreground. For all the other cases, any pixel from the current frame is tested with respect to corresponding values of ϕ, Δ, and δ. If all the comparative results are true, then that pixel will be considered as foreground pixel and denoted with one, otherwise as 0 (background) as given in Eq. (20).

$$FRG(i,j) = \begin{cases} 1 & \text{if } (\Delta(i,j) = 1,\ \delta(i,j) = 1 \text{ and } \phi_{ij}(FR(i,j)/BIN) = 1) \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

3.4.2. Noise elimination & resultant foreground output

There are different types of noise caused by different reasons. Therefore, noise types are usually identified to apply appropriate noise reduction method. For instance, in a computer assisted image analysis [15], different types of images can

be acquired and a different denoising can be applied. However, there are a few possibilities of noise in foreground of a video. This noise removal is done using small sized object elimination and standard morphological dilation and opening operations. Mainly 3×3 structuring element is used to dilate the regions to find the object more effectively. If the size of a object is less than a specific threshold, area opening operation is used to eliminate the object. This threshold depends upon the object size to be considered. Thus, false foreground or the noises after background elimination are reduced.

3.5. Background checking and scene change detection

The proposed approach keeps the track of the background information in a regular interval. The background features are estimated locally eliminating the forground information from each of the frame.

The locally estimated background is then compared to the modeled background and if there are changes, it is updated accordingly. An interval of $t/4$ frames has been taken for considering the local checking point. The background descriptor extraction is done in a similar way used in primary background modeling. After that, the local block-wise descriptor is compared with the modeled one for making the procedure adaptive for trivial local changes. If more than fifty percent of the local block descriptors disagree with the modeled background, it needs complete remodeling as that difference is adequate towards scene change detection. Otherwise, the local feature is merged with the current background to update the same.

4. Result and analysis

Two well-known datasets, *ChangeDetection* [4] and *SBMI* [16] are used to verify the outcomes of the proposed approach. There are several influencing parameters to disturb the background modeling and foreground extraction to handle these issues; the proposed approach uses two adaptive threshold selection procedures:

- TH_{chg} computation for estimating the temporal difference
- *THC* computation for approximating the foreground bin of a block

It is computed using t, which is taken as the frame rate of the video and since the frame rate may change to video, the value of t is dynamic. Local change estimation to update the background $t/4$ numbers of the frame are considered. These two can be tuned up according to the situations.

4.1. Change detection dataset

Receiver operating characteristic is the most used evaluation approach, especially in medical image analysis [17] since it needs precise measurements. Therefore, in this work the traditional performance measuring techniques are used to justify the efficiency of the proposed technique based on TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). These values are computed by comparing the experimental outcome and the ground truth. A brief of the traditional parameters are described below:

- *RE* (Recall) : TP / (TP + FN)
- *PR*(Precision) : TP / (TP + FP)
- *SP* (Specificity) : TN / (TN + FP)
- *FPR* (False Positive Rate) : FP / (FP + TN)
- *FNR* (False Negative Rate) : FN / (TP + FN)
- *PWC* (Percentage of Wrong Classifications) : $100 * (FN + FP)/(TP + FN + FP + TN)$
- F-Measure : $(2 * Precision * Recall)/(Precision + Recall)$

When the true value of experimental outcomes coincides with that of the ground truth, count of *TP* increases, otherwise *FP* rises. On the other hand, if the false value of output and the ground truths agree, *TN* increases, otherwise *FN* does. Hence, *TP* is highly desirable and *FN* is regrettable. *FN* counts the responsible pixels of false foreground. *RE* is the ratio of actual versus total extracted foreground areas in a frame. *FP* is one of the unexpected things in any technique as it counts the misclassified pixels which are supposed to be foreground. *PR* is the statistical ratio between *TP* and the total actual foreground.

The efficiency of any technique can be verified with the above mentioned statistical parameters. It is expected that any method must have higher values of *RE, SP, F-measure, PR* and smaller values of *FPR, FNR* and *PWC*. The high value of those parameters is expected because of the higher accuracy in outcomes. At the same time, a smaller value of the other parameters is expected because it depicts the minimization of the error rate. The proposed technique focuses on effective foreground detection using a *STRP* feature descriptor. Comparative results of the proposed technique with that of the other existing techniques are shown in Table. 1. It shows that the proposed approach produces the best results for *SP, FPR, PWC*, and *F-measure*. The average rank of the proposed method among all the methodologies is shown in Table. 1. Among the 9 different methods, the average rank of the proposed method is 1.33, whereas the second average rank 3 is obtained by *GMM*(*SG*) [4]. The best value of each parameter is shown in bold faces.

Table 1
A comparative study based on performance metric among the proposed approach and the other state of the art approaches on videos of *Changed Detection* Datasets.

Methods	Mean RE	Mean SP	Mean FPR	Mean FNR	Mean PWC	Mean FM	Mean PR	Mean Rank
STRP (Proposed)	0.82	**0.9989**	**0.001**	0.19	**0.58**	**0.87**	0.92	**1.333**
QCH [18]	0.704	0.9923	0.008	0.30	2.21	0.66	0.70	6.555
KDE-ISTF [19]	0.75	0.9954	0.005	0.25	1.81	0.74	0.78	4.44
GMM-RECTGAUSS [20]	0.67	0.9979	0.002	0.33	1.53	0.75	0.92	4.33
KDE-STDC [21]	0.755	0.994	0.006	0.245	1.915	0.755	0.78	4.67
GMM(SG) [22]	0.82	0.995	0.005	0.182	1.53	0.825	0.85	3
pROST[23]	**0.84**	0.994	0.006	**0.159**	1.15	0.83	0.82	3.11

Table 2
Comparative results on six evaluation parameters of the proposed approach and the same of the related research work.

Approach	Average AGE	Mean pCEPs	Mean MS-SSIM	Mean PSNR
LRGEOMC [25]	18.07	0.31	0.93	26.05
RMAMR [26]	18.13	0.31	0.93	26.03
GROUSE [27]	19.17	0.31	0.92	24.37
OR1MP [25]	20.95	0.32	0.89	23.91
IALM [28]	21.12	0.32	0.88	22.93
Outspace [29]	23.92	0.335	0.84	20.09
SC-SOBS [30]	6.32	4.47	0.93	29.89
Proposed (STRP)	**3.76**	**0.81**	**0.98**	**33.41**

4.2. SBMI Dataset

Scene Background Initialization (*SBMI*) [24] dataset is used in this work which includes 14 image sequences and their ground truth backgrounds. The irregularities mentioned before are adopted in this dataset. Hence, this dataset is the appropriate one to verify the proposed methodology. This dataset is adopted from the *SBMI-2015* workshop [24].

4.2.1. Evaluation parameters

The website as proposed in [24] provides the script for evaluating results for six metrics, which are used in the literature for background estimation. *GT* denotes the ground truth image and *CB* is the estimated background of the corresponding background modeling approach. The six metrics are used to estimate the difference between *GT* and *CB* images which are used to evaluate the effectiveness of background modeling.

- Average Gray-level Error (*AGE*): The absolute difference between gray-level *GT* and *CB* images. This is a global estimation of differences and the smaller value is more appreciable.
- Percentage of Clustered Error Pixels (*pCEPs*): This parameter is applied to ensure the number of clustered error pixels (*CEPs*) by checking the error in 4-connected neighbors of an error pixel. The lower error rate is better.
- Multi-Scale Structural Similarity Index (MS-SSIM): This parameter is proposed by is used to estimate the perceived visual distortion by using structural distortion and the greater value determines lesser distortion.
- Peak-Signal-to-Noise-Ratio (*PSNR*): It is defined as $\mathrm{PNSR}=10log-10(L-1)(\times 2/MSE)$, where L is the maximum number of grey levels and the *MSE* is the mean of squared error between *GT* and *CB* images. The value provides the superiority of the information over the noise. Thus, a larger value is desirable.

In this work, the results of the proposed approach have been compared with the corresponding *GT* of *SBMI* dataset. As shown in Table 2, the proposed approach produces the best results in terms of average *AGE*, average *pCEPs*, average *MS-SSIM* and average *PSNR*. The results in Table. 2 include the average values of seven videos for several parameters of the proposed and other methods of the related research. On the other hand, the results are shown in Fig. 7 illustrate the performance of the proposed background modeling approach. *AGE*, cluster, structural similarity and peak signal to noise ratio are better in case of the proposed work. The proposed approach handles the irregularities in a better way to minimize wrong classifications due to late moved, abounded or missing articles, video noise, etc. The proposed method updates the background at a regular interval which includes this type of the wrong area and this remains static in more frames.

4.3. Handling irregularities

Apart from the convincing comparative results described in the presiding part of this section, the proposed technique can handle the following irregularities.

- **Dynamic background:** Since the block-wise features can estimate the occurrence distribution of intensity bins, it can easily prune out the small changes in low-frequency regions as shown in Fig. 4. Besides, the persistence mechanism helps

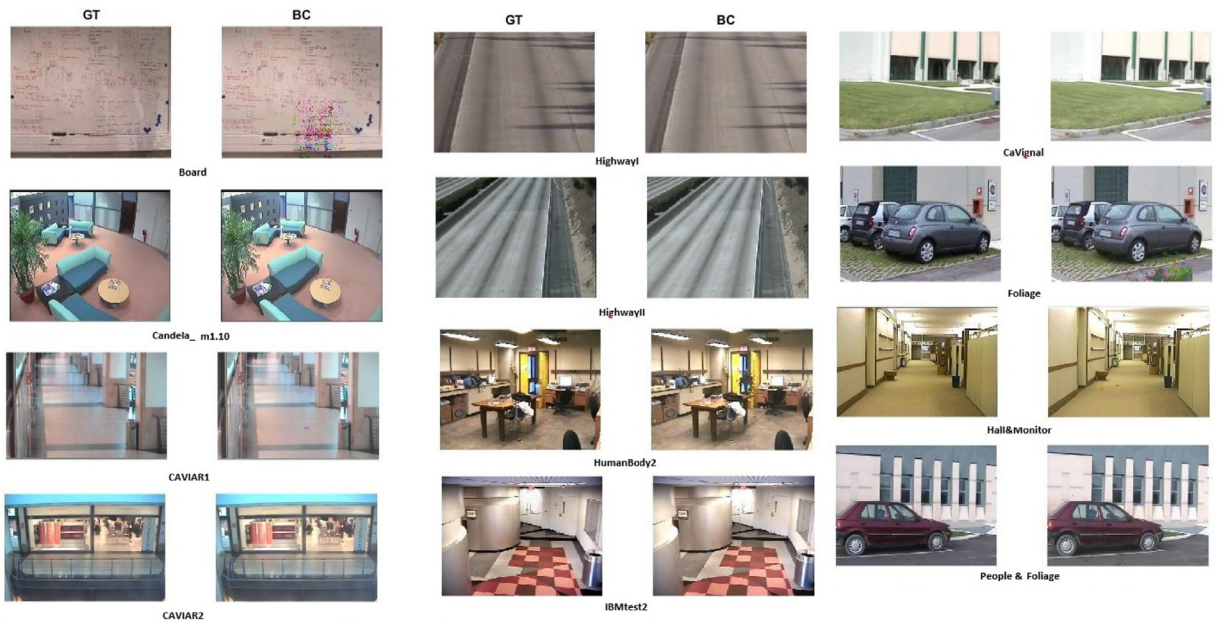


Fig. 7. Background detection results GT is the ground truth and BC is the estimated background.

to accommodate the intensity bins, which can occur consistently or at regular intervals in background *STRP* descriptor. Thus, the proposed approach is robust towards dynamic background which is visually described in Fig. 7.

- **Ghosting effect:** In our approach, this problem can occur in the initial part of foreground extraction due to slow-moving foreground objects in lower frame consideration for background modeling. The tunable parameter for the number of frames can be adjusted to reduce this effect. Otherwise, the adaptive capability of background estimation can easily remove the problem later on.
- **Gradual and sudden changes in illumination:** Contrast normalization are used to standardize the variable trivial lighting changes. Hence, gradual changes can be reduced.
  On the other hand, scene change detection can be the reply towards sudden changes, since the visibility is appreciably changed in such cases.
- **Video noise:** The smoothing operation and block-based features enable the proposed work to deal with the video noise and dynamic background. The noise or dynamism in the background may affect a single pixel but we are taking the gross bin-wise response instead of pixel intensity of every block which can slow the procedure. Besides, the adaptive threshold selection strategy takes care of the camouflage condition to a certain extent.

4.4. Limitations

Though the proposed work tries to handle many key challenges of the domain, some limitations are there, which can be treated as future scope of the current work.

- **Shadow detection:** Shadow detection is one of the key issues towards efficient foreground extraction. The inclusion of shadow detection procedures in the future can make this approach more robust. Shadow always creates confusion to identify an original object. A shadow detection and removal technique is an essential step to be incorporated in the future.
- **Automatic selection of block size:** The selection of block size is a crucial issue in case of any block-based approach. It would be better to use some mathematical models to select the size of the block automatically. The automatic selection of block size will help this work to fit in a resolution-independent approach.
- **Bootstrapping and slow-moving object:** The slow-moving objects always misleading the background estimation, which leads to poor foreground extraction. If we incorporate a module to handle such situations in the future, the background estimation procedure would be more efficient.

5. Conclusions

An adaptive background modeling technique has been presented using the newly introduced *STRP* descriptor, which is used to estimate the background from the video. This modeled background is then used to extract an effective foreground. Besides, the use of adaptive threshold selection and auto-update techniques make the work more robust in challenging

situations like dynamic background, ghosting effect, change in illuminations, video noise, etc. Scene change detection helps the procedure to adapt to the variable environmental conditions. Two different groups of statistical measures have been included in this work to verify the efficiency of the current work. The first group of parameters includes precision, recall, specificity, F-measure, false-positive rate, false-negative rate, percentage of the wrong classification. On the other hand, the second group comprises gray level error, clustered error pixels, multi-scale structural similarity, peak signal to noise ration. The results and comparative study of the work with that of the other related research for *Changeddetection* and *SBMI* dataset based on those two groups of parameters justify the novelty of this work.

The limitations like shadow detection, automatic assessment of block size, and misclassification due to bootstrapping or slow-moving objects can be the future scope to make this work more efficient. Different techniques can be adopted for shadow detection and elimination towards efficient foreground object extraction. like mean-shift filtering, graph cut, watershed segmentation, fisher linear discrimination. Deep learning is an effective tool for classification and prediction nowadays, which will be used to increase the effectiveness of such kind of work.

Declaration of Competing Interest

The authors declare that they do not have any financial or nonfinancial conflict of interests.

Acknowledgement

This research work is funded by DST, Govt. of India, through the INSPIRE project, and is supported by the *TEQIP* phase-II project at the University of Calcutta.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compeleceng.2019.106520.

References

[1] Brutzer S, Hoferlin B, Heidemann G. Evaluation of background subtraction techniques for video surveillance. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '11; 2011. p. 1937–44.
[2] An J, Ha SJ, Cho NI. Probabilistic motion pixel detection for the reduction of ghost artifacts in high dynamic range images from multiple exposures. EURASIP J Image Video Process 2014;2014(1):42.
[3] Goceri E, Goceri N. Deep learning in medical image analysis: Recent advances and future trends. In: International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2017 (CGVCVIP 2017), Lisbon, Portogual; 2017. p. 305–11. doi:10.1109/ICMLA.2015.229.
[4] Goyette N, Jodoin P, Porikli F, Konrad J, Ishwar P. Change detection.net: A new change detection benchmark dataset. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2012. p. 1–8.
[5] Bouwmans T, Maddalena L, Petrosino A. Scene background initialization. Pattern Recogn Lett 2017;96(C):3–11.
[6] Zheng J, Wang Y, Nihan NL, Hallenbeck ME. Extracting roadway background image: mode-based approach. Transp Res Rec 2006;1944(1):82–8.
[7] Wren CR, Azarbayejani A, Darrell T, Pentland AP. Pfinder: real-time tracking of the human body. IEEE Trans Pattern Anal Mach Intell 1997;19(7):780–5.
[8] Jodoin J, Bilodeau G, Saunier N. Background subtraction based on local shape. CoRR 2012;abs/1204.6326.
[9] Zhang S, Yao H, Liu S. Dynamic background subtraction based on local dependency histogram. Int J Pattern Recognit Artif Intell 2009;23:1397–419.
[10] Biswas S, Sil J, Sengupta N. Background modeling and implementation using discrete wavelet transform, a review. Graphics, vision and Image Processing Journal 2011;11:29–42.
[11] Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 1990;12(7):629–39.
[12] Kamalaveni AV, Rajalakshmi R, Narayanankutty KA. Image denoising using variations of perona-malik model with different edge stopping functions. Procedia Comput Sci 2015;58:673–82.
[13] Wang Y, Guo J, Chen W, Zhang W. Image denoising using modified perona-malik model based on directional laplacian. Signal Processing 2013;93:2548–58.
[14] Michelson AA. Studies in optics. University of Chicago science series. The University of Chicago Press; 1927.
[15] Elder JB, Puduvalli VK, Otero JJ, Gurcan MN, Goceri E, Goksel B. Quantitative validation of anti-ptbp1 antibody for diagnostic neuropathology use: image analysis approach. Int J Numer Method Biomed Eng 2017;33(11):e2862. doi:10.1002/cnm.2862.
[16] Wu Y, Lin T, Li S, Wang W, Wang Y, Chen B. Quantification of full left ventricular metrics via deep regression learning with contour-guidance. IEEE Access 2019;7:47918–28.
[17] Gurcan MN, Goceri E, Shah ZK. Vessel segmentation from abdominal mr images: adaptive and reconstructive approach. Int J Numer Method Biomed Eng 2016;33(4),e02811. doi:10.1002/cnm.2811.
[18] Strauss O, Sidib D, Puech W. Quasi-continuous histogram based motion detection. Technical Report, LE2I; 2012.
[19] Nonaka Y, Shimada A, Nagahara H, Taniguchi R. Evaluation report of integrated background modeling based on spatio-temporal features. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2012. p. 9–14.
[20] Riahi D, Bilodeau G, St-Onge P. RECTGAUSS-Tex: Block-based background subtraction. Rapport technique. École polytechnique; 2012.
[21] Yoshinaga S, Shimada A, Nagahara H, Taniguchi R. Background model based on intensity change similarity among pixels. In: The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision; 2013. p. 276–80.
[22] Stauffer C, Grimson WEL. Adaptive background mixture models for real-time tracking. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 2; 1999. p. 246–252Vol. 2.
[23] Seidel F, Hage C, Kleinsteuber M. Prost : a smoothed lp-norm robust online subspace tracking method for realtime background subtraction in video. CoRR 2013;abs/1302.2073.
[24] Bouwmans T. Recent advanced statistical background modeling for foreground detection - a systematic survey. Recent Patents on Comput Sci 2011;4(147).
[25] Wang Z, Lai M-J, Lu Z, Ye J. Orthogonal rank-one matrix pursuit for low rank matrix completion. SIAM J Sci Comput 2014;37.
[26] Ye X, Yang J, Sun X, Li K, Hou C, Wang Y. Foregroundâbackground separation from video clips via motion-assisted matrix restoration. IEEE Trans Circuits Syst Video Technol 2015;25(11):1721–34.

[27] Balzano L, Wright SJ. On grouse and incremental svd. In: 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP); 2013. p. 1–4.
[28] Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. J Struct Biol 2013;181 2:116–27.
[29] Keshavan RH, Montanari A, Oh S. Matrix completion from noisy entries. J Mach Learn Res 2010;11:2057–78.
[30] Maddalena L, Petrosino A. The sobs algorithm: What are the limits?. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 00; 2012. p. 21–6.

Satyabrata Maity is presently working as Assistant Professor in CSE, SOADU, Odisha, India. He received his Ph.D. and M.Tech degree in IT from the University of Calcutta in 2019 and 2009, respectively. His research interests pertain to image and video processing, computer vision, and machine learning. He has published around 20 research papers. He received DST INSPIRE fellowship for Ph.D. and got Gold Medal in M.Tech.

Amlan Chakrabarti is a Full Professor in the A.K.Choudhury School of Information Technology at the University of Calcutta. He was a Post-Doctoral fellow at the School of Engineering, Princeton University, USA during 2011-2012. He has published around 140 research papers in referred journals and conferences. He is a Sr. Member of IEEE and ACM, Distinguished Speaker of ACM.

Debotosh Bhattacharjee is a full professor in Computer Science and Engineering at the Jadavpur University. He received the PhD (Engineering) degree from Jadavpur University, India, in 2004. His research interests pertain to the applications of machine learning techniques for face recognition, gait analysis, hand geometry recognition, histopathological image analysis, and biomedical imaging. He has authored or co-authored more than 220 publications, and articulated with two US patents. He was invited by many renounced Universities in Europe.