IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 June 2023, accepted 24 July 2023, date of publication 28 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299872

RESEARCH ARTICLE

KDE-Based Simultaneous Background Model Learning and Entropy-Based Fusion of Cascaded Features for Video Object Segmentation With Shadow Removal

SUBHALUXMI SAHOO, (Member, IEEE), AND PRADIPTA KUMAR NANDA, (Senior Member, IEEE)
Image and Video Analysis Laboratory, Department of ECE, Siksha 'O' Anusandhan, Deemed to be University, Bhubaneswar, Odisha 751030, India

Corresponding author: Subhaluxmi Sahoo (subhaluxmisahoo@soa.ac.in)

This work was supported by Siksha 'O' Anusandhan, Deemed to be University.

ABSTRACT Object detection with shadow removal is one of the challenging issues in computer vision. Dynamic shadow resembles a moving object's properties, so separating this shadow from the object is a challenging task. This dynamic shadow if not eliminated, distorts the shape of the object. In this paper, a novel scheme for moving object detection and shadow removal is proposed based on the background modeling in fused feature space, and these models learn to take care of the scene dynamics. Initially, in KDE space, temporal modeling of the spatial KDE (TMS-KDE) is carried out and cascaded features of Gabor and HOG are obtained. Besides, the original video frame is transformed into YCbCr color space and LBP features are extracted. The LBP and cascaded features are fused probabilistically to generate fused feature frames which are used in background modeling. The weights for the feature fusion are determined by the proposed entropy based measure. Background modeling and model learning is a pixel based approach and the pixel is classified as either background or foreground during the learning process. We have tested our proposed method on a wide range of datasets which includes ATON-CVRR, LASIESTA, CD-net, Kaggle, PETS 2006, SGM-RGBD, SBMI 2015, SBMnet 2016 and VIRAT. The proposed scheme is found to take care of different shadow conditions while detecting the moving object. The performance of the proposed scheme is found to be superior to that of many existing schemes.

INDEX TERMS Dynamic shadow, TMS-KDE, LBP, cascaded feature, feature fusion.

I. INTRODUCTION

Video object detection is one of the emerging areas in the field of computer vision and pattern recognition [1], [2]. Detection of objects in complex videos is difficult due to the presence of shadow, dynamic movement of background entities, changes in illumination, noise, motion blur, occlusion, etc. [3], [4], [5], [6], [7], [8]. Many methods have been proposed in the literature to address the above challenges [9], [10], [11], [12], [13]. Out of these factors, shadow detection and removal from moving video objects is challenging [14]. Different schemes utilizing specific shadow features have been proposed for shadow detection and removal [9]. The shadow features include illumination ratios in different color spaces, texture, gradient, and morphological features. It is found that a combination of different features also enhances the detection capability for shadows [10], [11], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

Shadow is formed due to obstruction in the light path by the object, which may be cast shadow or self shadow. Self-shadows are the shadows that are on actual physical objects casting it and cast shadows are the shadows that the objects cast upon other objects. It is detected along with the moving object and hence it distorts the object's shape. Our proposed work is for video object detection with shadow removal, which helps retain the object shape.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

Besides, the shadow created due to the background entities with varying illumination conditions poses a problem. One of the approaches to video object detection in a complex scene is to model the complex background and thereafter adhere to the notion of background separation. These moving and static cast shadows are modeled by Gaussian mixtures [18], [26]. Martel-Brisson and Zaccarin [26] have proposed Gaussian mixture shadow models (GMSM) and learned these shadow distributions for shadow removal. Specifically, the moving shadows are detected by a scheme comprising support vector machine and co-training algorithm proposed by Joshi and Papanikolopoulos [27]. Additionally, the notion of ratio edge is used to take care of the moving cast shadows for object detection [24]. Because of the instant illumination changes and dynamic background entities in a real world scene, conventional object detection methods may be unable to detect the object. Hence, Kim et al. [28] have proposed an accurate and instant background modeling (AIBM) method which utilizes the spatio-temporal information. Moving shadow with the foreground object is also modeled by Dynamic Conditional Random Field (DCRF) model for moving object detection [23]. Because of the uncertainties in the real world scene, the above stochastic framework-based schemes with appropriate learning strategies proved to be quite effective in detecting the foreground object. It is also found in the literature [29], [30] that spatio-temporal modelings with appropriate features could take care of the static and moving cast shadows. The illumination changes and moving shadows are also taken care of by the bit plane method proposed by Lin et al. [31] while detecting the moving object in a real world scene. This method has used the color characteristics in spatio-temporal framework. Though several methods are proposed for handling a wide variety of shadows, the problem of moving object detection in the real world and some typical indoor scenes pose a challenge with different shadow conditions and complex background conditions [32]. Hence, in this paper, we have addressed the problem of moving object detection with different shadow conditions and also used the notion of background modeling in a stochastic framework exploiting the spatio-temporal dependencies. These background models learn the scene dynamics and detect the foreground. Towards this end, a new scheme is proposed to detect the moving object in video and remove the shadow occurring either due to foreground or background. This problem is compounded because of the varying illumination condition over the scene. The problem is addressed using background modeling and model learning in the fused feature space. The proposed scheme is a pixel based process. Initially, the spatial KDE (S-KDE) of the frames are determined and using the S-KDE frames, the temporal modeling is carried out. This modeling is called the TMS-KDE model of a given frame. This modeling differs from our previous work, where the spatio-temporal modeling (ST-KDE) is carried together on the KDE frames. The spatial KDE will reinforce all the entities in the spatial domain, and its temporal modeling will preserve the moving object part along with its static shadows. From these TMS-KDE modeled frames; first Gabor features are extracted, and thereafter HOG features are extracted from the Gabor featured frame. Hence, the combined feature obtained from the TMS-KDE frame is viewed as the cascaded feature of a frame. The Gabor feature is used to remove the dynamic shadow due to the moving object, while the HOG is used to preserve the object's shape. Thus, the cascading of Gabor and HOG will eliminate the dynamic shadow while maintaining the shape of the object.

Besides, the input frames are transformed into YCbCr color space, and the LBP features are extracted from this YCbCr transformed frame. The LBP features thus obtained are fused with the cascaded features probabilistically to result in the fused feature frame. The YCbCr color model is expected to preserve the chromaticity while partially eliminating the shadow as shadow has a low chromaticity value. Besides, the LBP feature preserves the textural features of the entire frame, including that of the moving object. The entropy based fusion of the cascaded feature attempts to eliminate shadow while preserving the object's shape. Thus, the fused feature space is expected to preserve the object's shape in its entirety while removing the static and dynamic shadows to a great extent.

For feature fusion, the weights of the respective features are determined using the proposed entropy based measure which is different from our previous work, [33] which is based on the similarity measure of the distributions. Additionally, the cascaded feature extraction and its fusion differs from our previous work where the individual feature is extracted and fused with another feature. Background modeling and learning is carried out for every pixel of the fused feature frame. The model learning removes the residual shadow components and classifies the object, thus segmenting the moving object. For a given pixel, model histograms are obtained considering a few fused feature frames. These model histograms serve as the background model that learns the information from the new input frame and then classifies the pixel as either background or foreground. Model learning happens with the bin level updation of the histograms together with the updation of the weights of these model histograms. The proposed scheme is tested with eight different data sets with different conditions and is found to possess better shadow detection and discrimination ability as compared to other existing algorithms. The proposed algorithm could also remove the shadow due to the moving object and dynamic entities of the background besides the static shadows. The proposed scheme demonstrates improved performance as compared to other existing algorithms.

II. RELATED WORK

The scene complexity is increasing daily in computer vision, making object detection a difficult task. Shadow of the moving object is one of the complex entities in any video scene. It gets detected along with the object hampering the shape of the object. Work in this direction has been in continual progress. Initially, Prati et al. [9] have given a comparative evaluation of the different algorithms for shadow detection and removal in video scenes. They have classified the methods into deterministic (both model and non model based) and statistical approaches (parametric and non parametric). Deterministic approaches work based on a certain decision making mechanism but statistical approaches work on the basis of probabilistic models for deciding membership of a class.

A. MODEL AND NON-MODEL BASED SHADOW DETECTION

Onoguchi [34] has proposed a deterministic model based method where the height of the shadow is captured using a set of two cameras and then a simple background subtraction technique is utilized for shadow removal. Model based methods are usually complex and time taking for complicated scenes, hence non-model based methods were also considered by many researchers. Deterministic non model based method was used by Jiang and Ward [35] where shadow parameters like shadow intensity and shadow geometry are utilized for shadow removal. Probabilistic models can better analyze complex scenes, hence statistical models became the next area of research. One of the statistical parametric based methods is used in the form of an incremental version of EM algorithm in combination with a mixture of Gaussians for shadow pixel removal in traffic scenes [36]. Model parameter selection is an issue in these models, hence many research works are pursued using nonparametric statistical models. Statistical non parametric approach was used by Horprasert et al. [37] in the form of a computational color model for separating brightness from the chromaticity component in the shadow pixels. Although these methods coud remove the shadow from video scenes, their performance deteriorated with increase in the scene complexity. This happened because none of the shadow properties are utilized during shadow removal. Shadow properties are different from the object properties; hence they provided a separate classification method.

B. FEATURE BASED SHADOW DETECTION

Another broad way of classifying shadow detection and removal in video object scenes is presented by Sanin et al. [38], [39]. They have provided the feature based taxonomy for shadow. The specific shadow features used are intensity, chromaticity, and physical properties [40], [41], [42], [43], [44]. Shadow is always darker than the object and hence this property is used for its detection but this method fails in complex scenarios. Zhang et al. [45] used normalized coefficients of the image block to differentiate shadow from object. But the system complexity is increased due to computation of the coefficients; hence other properties were considered. Chromaticity implies the color component, but the scene shadow is devoid of this. This property is used in conjunction with many color models for shadow detection and removal [10], [11], [15], [16]. Similarly, different physical attributes are utilized for detecting, learning, and removing the shadow pixels [26], [27], [46], [47], [48], [49]. Different geometry based methods like shadow orientation, shape, and size of the shadows are also helpful in detecting and removing shadows from video objects [17], [18], [19], [20], [21]. Textural properties of the shadow also help in its detection because usually shadow has no textural attribute. This can be utilized in video object scenes where object and background have their inherent textures [22], [23], [24], [25].

C. SHADOW DETECTION IN REMOTE SENSING IMAGES

Shadow removal has wide applications in the field of aerial image processing and remote sensing. Extensive research has been carried out in aerial remote sensing images where shadow creates a hindrance. In this regard, Luo et al. [50] have proposed a novel edge-aware spatial pyramid fusion network (ESPFNet) along with a multitask learning framework for salient shadow detection in aerial remote sensing images. Shadow in the case of the multitemporal data is also a concern and is taken care of by the surface reflectance based cloud shadow detection algorithm (SRCSD) proposed by Sun et al. [51]. Shadows cause flaws in object detection by aerial images. Statistical descriptors are extracted from the image for effective shadow removal and are used for shadow detection [52]. Specifically for VideoSAR data, shadow of the moving object is detected using background reconstruction [53]. A scheme based on semantic background subtraction in real-time mode is found to perform well for most of the generalized video scenes, including scenes with shadows [54]. All the above methods are concerned with removing shadows from the static scenes.

D. NEURAL NETWORK BASED SHADOW DETECTION

In the recent past, Convolutional Neural Network (CNN) has also been used for shadow detection [55]. In this regard, a deep-learning method for shadow detection at the pixel level is proposed by Mohajerani and Saeedi [56] that is suitable for single RGB images. This CNN-based method utilizes a novel architecture through which global and local shadow attributes are identified using an efficient mapping scheme. Training of a Kernel Least Square Support Vector Machine (LSSVM) is used for labeling the regions separating the shadow and non shadow portions [57]. Also Support Vector Machine (SVM) based on color saliency space and gradient field is used for shadow detection and removal for on-road visual inspection, where the nonlinear SVM classifier analyzes its color saliency space and gradient information, and in the sequel, reconstructs road shadow descriptor to distinguish shadowed regions [58]. Shadow detection is also achieved by an attentive feedback feature pyramid network (AFFPN) proposed by Kim and Kim [59], and a novel deep neural network named Mask-ShadowNet is proposed for shadow removal [60]. A different deep learning motion architecture with multi cue autoencoder is proposed by Rahmon et al. [61] which detects motion and change cues using multi-modal background subtraction. Spatio-Temporal data augmentations for video-agnostic supervised background subtraction is proposed by Tezcan et al. [62]. Contour optimizer, a different supervised learning algorithm also helps remove shadow in complex scenes [63]. A novel compact end-to-end convolutional neural network architecture with motion saliency foreground network (MSFgNet), is proposed by Patil and Murala [64] to estimate the background and to extract the foreground from video frames. A universal background subtraction framework using Arithmetic Distribution Neural Network (ADNN) for learning the distributions of temporal pixels is proposed by Zhao et al. [65]. But all these deep learning methods require a large dataset for training. Our proposed research work is novel and different in the sense that we are modeling the background, learning these models and classifying the pixel simultaneously in an online mode. Although the problem of shadow detection for single image cases has widely been addressed in literature, our work is based upon detecting and removing shadows from video images during object extraction.

E. FUSION BASED SHADOW DETECTION

Wang et al. [66] have presented an effective framework for moving cast shadows. They have used multiple ratio techniques to justify shadow type along with feature fusion strategy for detecting shadow. Moving object segmentation (MOS) using a Recurrent Edge Aggregation Module (REAM) has also been proposed [67]. Zhao and Basu have proposed a dynamic deep pixel distribution learning for background subtraction [68]. Real time pixel classification along with parameter updation using an adaptive 3 phase background model is proposed by Roy and Ghosh [69]. Zhang et al. [70] have proposed a moving shadow elimination method using the fusion of multiple features. They fused a dual-channel HSV color space feature and a uniform extended scale invariant local ternary pattern (UESILTP) texture feature to eliminate shadow. Other deep learning and convolutional neural network architectures are also helpful in accurately defining the shadow boundary and its removal [71], [72], [73], [74], [75], [76].

III. PROPOSED SCHEME

The block diagrammatic representation of the proposed scheme is shown in Fig.1. This is based on background modeling and model learning in the fused feature space to detect the moving object by removing the shadow. In this framework, the objective is to develop a background model which will have the attribute of taking care of the shadow of the moving object and the background in a given frame and simultaneously discriminating the foreground from the background. As observed from Fig. 1, two different spaces based on different features are created from the raw image space. In the first case, the original image is transformed into YCbCr color space to embed the attribute of shadow removal and the local texture features of the raw space are extracted using LBP. In the second case, a spatio-temporal modeling is proposed in the KDE space. S-KDE of the original frame is computed and thereafter temporal modeling of the S-KDE frames is carried out which is shown in Fig. 2. Therefore this model is named TMS-KDE. In order to embed the shadow removal attribute and extract local features from the TMS-KDE modeled frame, the Gabor features are extracted and in the sequel, HOG features are extracted from the Gabor filtered version of the TMS-KDE frames. This can be viewed as transforming the TMS-KDE space to cascaded feature space.

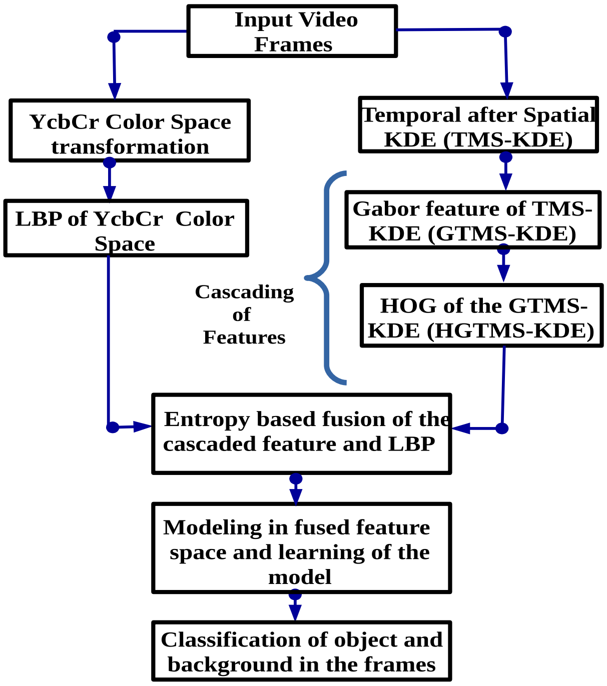

Input Video Frames
YcbCr Color Space transformation
Temporal after Spatial KDE (TMS-KDE)
LBP of YcbCr Color Space
Gabor feature of TMS-KDE (GTMS-KDE)
Cascading of Features
HOG of the GTMS-KDE (HGTMS-KDE)
Entropy based fusion of the cascaded feature and LBP
Modeling in fused feature space and learning of the model
Classification of object and background in the frames

FIGURE 1. Block diagrammatic representation of the proposed approach.

The cascaded features thus extracted are fused with the LBP feature extracted from the raw image space. Feature fusion happens in a probabilistic framework where the weights for fusing the features are determined based on the notion of entropy. Background modeling and model learning happen in fused feature space. The modeling and model learning is pixel based where the model histograms correspond to the pixel in the fused feature frame. Learning of these model histograms takes place with the input histogram of a pixel of the new frame. In learning, both the model histograms and the associated weights of the histograms are adapted. Classification of the pixel as either background or foreground takes place after learning. Learning and classification of all the frames result in the detection of the moving object in the frame while removing the associated shadow of the moving object and the background.

IV. TEMPORAL BACKGROUND MODELING OF SPATIAL KDE FRAMES

In this section, the background modeling in the KDE framework is presented. In the scene, the shadow due to the

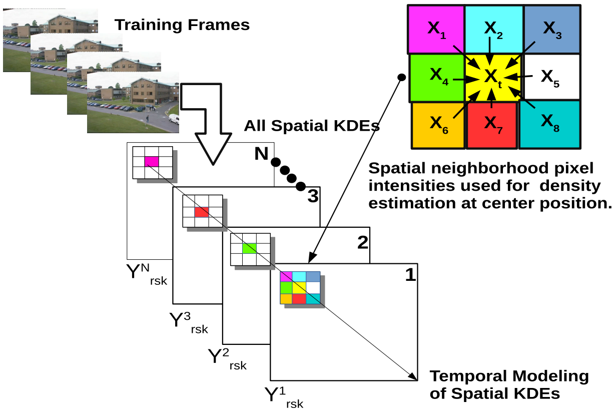

Training Frames
All Spatial KDEs
N
3
2
1
X1
X2
X3
X4
Xt
X5
X6
X7
X8
Spatial neighborhood pixel intensities used for density estimation at center position.
Y^N rsk
Y^3 rsk
Y^2 rsk
Y^1 rsk
Temporal Modeling of Spatial KDEs

FIGURE 2. Spatial and temporal modeling of the pixels in KDE framework.

moving object moves along with the object while the shadow due to background remains stationary. To take care of the moving object, shadow, and, the background, a new spatio-temporal modeling is proposed in the KDE framework. Initially, the S-KDE of the video frames are computed. Here, the S-KDE implies that the probability density of a given pixel is computed using its spatial neighborhood pixels. Let x_t denote the pixel in consideration of t^{th} frame and x_{t_k} denote the k^{th} pixel of the neighborhood structure. The Gaussian kernel is used to compute the density which is given as,

$$P(x_t) = \frac{1}{N_{s_k}} \sum_{k=1}^{N_{s_k}} \frac{1}{\sqrt{2\pi\sigma^2_{spatial}}} e^{-\frac{1}{2}\left(\frac{x_t - x_{t_k}}{\sigma_{spatial}}\right)^2}, \tag{1}$$

where N_{s_k} denotes the number of neighborhood pixels around x_t and $\sigma_{spatial}$ is the bandwidth of the Gaussian kernel. This spatial neighborhood structure is shown in Fig. 2. This spatial KDE modeling is expected to preserve the boundary of the object. In order to model the object in different frames, temporal KDE modeling of the S-KDE frames is carried out. The modeling in the temporal direction will take care of the moving object in different frames as well as the dynamic shadow arising out of the moving object. As shown in Fig.2, we consider the S-KDE frames in the temporal direction numbered as 1,2,3 ……N.

Let y_{s_k} denote the S-KDE value of i^{th} site of the S-KDE frame. Let us consider the corresponding values of r^{th} site of N number of S-KDE frames in the temporal direction which are denoted as $y_{1_{s_k}}, y_{2_{s_k}}, y_{3_{s_k}} \ldots\ldots y_{N_{s_k}}$. The probability density of $y_{i_{s_k}}$ at the r^{th} site of the first frame can be computed as,

$$P(y_{r_{s_k}}) = \frac{1}{N} \sum_{q=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2_{temporal}}} e^{-\frac{1}{2}\left(\frac{y_{i_{s_k}} - y_{(i_{s_k}-q)}}{\sigma_{temporal}}\right)^2}, \tag{2}$$

where N is the number of corresponding sites considered in the temporal direction. Thus, (1) and (2) together correspond to the spatio-temporal modeling of the given pixel of the video frame. Similar process is carried out for all the pixels of a frame to obtain the Temporal Modeling of the Spatial KDE frames (TMS-KDE). The TMS-KDE models of the original frames of Fig. 3(a) and 3(d) are shown in Fig.3(c) and 3(f) respectively. As observed from these two figures, the boundary of the object and the object itself are preserved. Thus, this modeling has the attribute of taking care of the static and dynamic shadows as well. Further, it may be observed from Fig.3(b) and Fig.3(e) that the S-KDE modeling preserves the object boundaries of a frame.

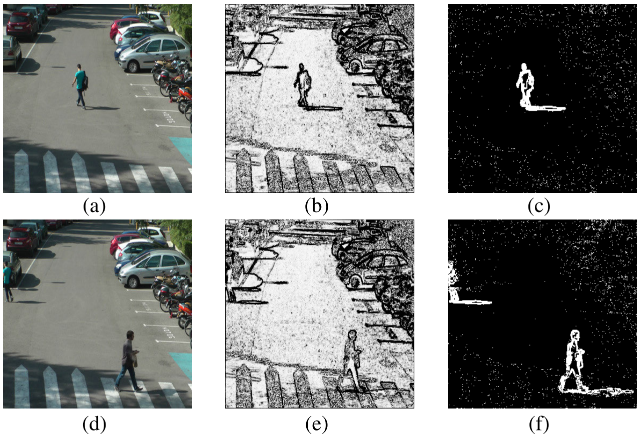


FIGURE 3. TMS-KDE modeling of LASIESTA 2016 dataset (352 × 288). (a) Original image (frame 200), (b) S-KDE, (c) TMS-KDE, (d) Original image (frame 300) (e) S-KDE, (f) TMS-KDE.

After obtaining the TMS-KDE of a frame, cascaded features are extracted and are fused with the LBP features obtained from the original frame and the background modeling is carried out in the fused feature space. Background modeling is a pixel based approach and is performed in the stochastic framework. For a given pixel at the r^{th} site of the frame, a window of a given size (w×w) is considered around the given pixel. The histogram of this window is considered as the model histogram for the given pixel of the frame. To have a set of k model histograms for the given pixel, the corresponding k sites of the previous k TMS-KDE frames are considered. Windows of the same size are considered around these sites. The histogram of each site is a model histogram and hence k such sites result in k model histograms. These are shown in Fig.4 and these histograms of the feature space are considered as the background model of the given pixel. For example, if we consider k to be 3, then three such histograms are considered as the model histograms of a given pixel. The updation of the model happens with each incoming frame. The corresponding pixel is considered and the same size of window is taken around the pixel. The histogram in the 2nd row of Fig. 4 is the input histogram of the pixel and the model histograms learn this input histogram.

V. CASCADING OF LOCAL FEATURES AND ENTROPY BASED FEATURE FUSION

A. CASCADING OF LOCAL FEATURES

It is known from the previous section that the TMS-KDE is able to preserve the object boundary and the object, but along with these the moving shadow is also preserved. To further

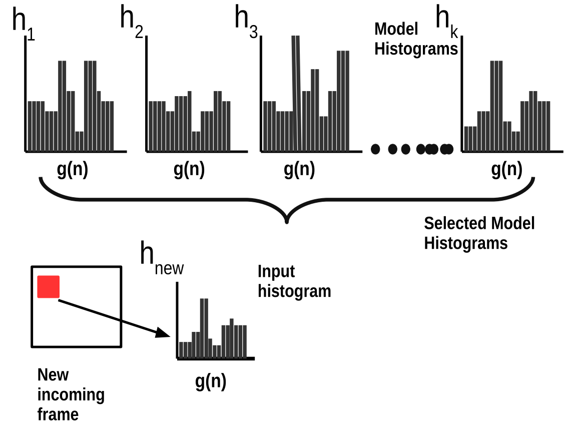


FIGURE 4. Learning of the model histograms for a given pixel of a frame with the input histogram of a given pixel.

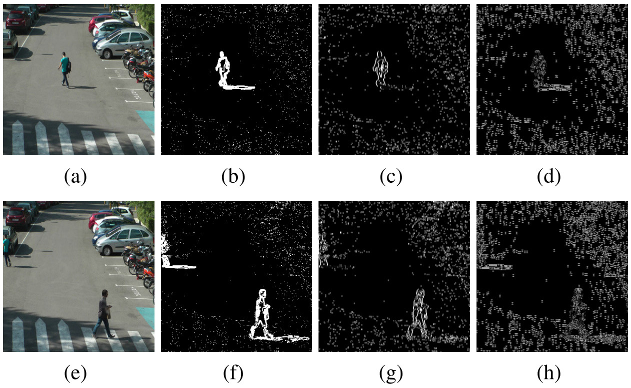


FIGURE 5. Results obtained for cascaded feature in LASIESTA 2016 dataset (352 × 288). (a) Original image (frame 200), (b) TMS-KDE, (c) GTMS-KDE, (d) HGTMS-KDE, (e) Original image (frame 300) (f) TMS-KDE, (g) GTMS-KDE (h) HGTMS-KDE.

improve the attributes of the model with a view to reducing the moving shadow, we have used Gabor's 90° feature thus making the model GTMS-KDE model. These are shown in Fig. 5(c) and 5(g) where it may be observed that the shadow is eliminated to a great extent but some of the local features of the background are also present. To further preserve the shape of the object, we have applied the HOG feature on this GTMS-KDE model thus making it HGTMS-KDE model which is the cascaded feature model. The cascading of these two features is expected to model the object while preserving the local attributes. The HOG feature applied on the GTMS-KDE model reinforces the oriented gradients of the extracted local features by the GTMS-KDE model. Because of HOG, many finer details of the background are also preserved together with the object. Fig. 5(d) and 5(h) show the frames obtained with HGTMS-KDE model and as expected, too many details of the background are also retained. This will be taken care of by the fusion of the features and the learning of the background model in the fused feature space.

B. ENTROPY BASED FUSION

We have attempted to build the background model on feature space and particularly in the fused feature space. The features of the frame extracted due to the cascaded feature are fused with the LBP feature of the YCbCr color space. These features of the HGTMS-KDE model are fused with the LBP features of the color model of the raw data of the original frame. In this process, the local features of the original image space are fused with the cascaded features of the TMS-KDE image space. The features are fused in the probabilistic framework. The fusion process is also a pixel based approach. Let us consider x(i,j) site in the LBP color frame and also the corresponding site in the cascaded featured frame.

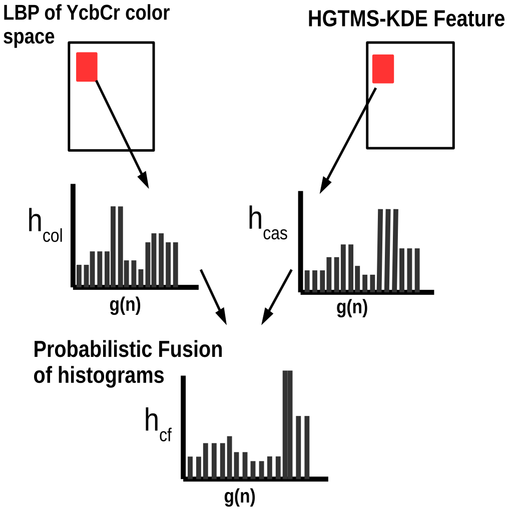


FIGURE 6. Probabilistic fusion of features using corresponding histograms.

Consider a window of size ($w \times w$) around both the pixels of the corresponding frames as shown in Fig.6. The two features of the pixel are fused as follows. Let h_{col} denote the histogram of the LBP of the YCbCr color space and h_{cas} denote the histogram of the cascaded feature of the HGTMS-KDE frame. These two features are fused probabilistically as follows,

$$\overrightarrow{h_{cf}} = w_{col} \times \overrightarrow{h_{col}} + w_{cas} \times \overrightarrow{h_{cas}}, \tag{3}$$

where $\overrightarrow{h_{cf}}$ denotes the cascaded feature histogram, w_{col} and w_{cas} are the weights of $\overrightarrow{h_{col}}$ and $\overrightarrow{h_{cas}}$ respectively. The histogram of the fused feature $\overrightarrow{h_{cf}}$ is used as the background model. The above weights are determined based on the entropy which varies based on the features. Let the entropy over a window of the original frame be denoted as E_o and E_{col} be the entropy over the window of the LBP feature of the YCbCr color frame. Let E_{cas} denotes the entropy over the window of the cascaded feature frame. The entropy over a given window is computed as,

$$S = \sum_i -P_i \times log(P_i), \tag{4}$$

where i indicates the gray level that varies from 0 to 255 and S denotes the entropy over the window.

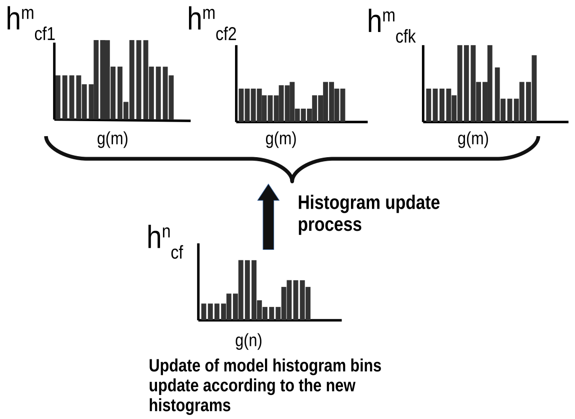

h^m cf1
h^m cf2
h^m cfk
g(m)
g(m)
g(m)
Histogram update process
h^n cf
g(n)
Update of model histogram bins update according to the new histograms

FIGURE 7. Learning of the model histograms for a given pixel with the new input histogram of a given pixel.

The two weights are determined as,

$$w_{col} = \frac{E_{cas}}{E_o}, \tag{5}$$

$$w_{cas} = \frac{E_{col}}{E_o}. \tag{6}$$

Thus, the computed weights are scene dependent. Since the conditions of the background and foreground change over the entire frame, the entropy will also vary over the entire frame. Similar findings are also expected in the two feature frames. Thus the appropriate degree of both the features will be fused to result in a fused feature for a given pixel. This feature fusion happens while learning the background model.

VI. SIMULTANEOUS LEARNING OF BACKGROUND MODEL AND FEATURE FUSION WITH PIXEL CLASSIFICATION

The background feature fusion happens while learning the background model. Learning is a pixel based approach. For a given pixel, initially few frames are chosen and the fused feature histograms are obtained. These histograms are denoted as $\overrightarrow{h^m_{cf_1}}$, $\overrightarrow{h^m_{cf_2}}$, $\overrightarrow{h^m_{cf_3}}$,.......,$\overrightarrow{h^m{}_{cf_k}}$, where k denotes the number of histograms chosen. These histograms are the background model histograms of that given pixel. They are as shown in Fig.7.

Learning of the model histograms takes place as follows. With the arrival of a new input frame, the corresponding pixel is chosen and a window is considered around the pixel. For that window, the cascaded fused feature histogram $\overrightarrow{h^n{}_{cf}}$ is obtained. The model histograms $\overrightarrow{h^m_{cf_1}}$.........$\overrightarrow{h^m_{cf_k}}$ are updated based on the input histogram of the cascaded feature. These model histograms are assigned with random initial weights from 0 to 1. Let these weights be denoted as $q_1, q_2 \ldots q_k$. The new input model histogram $\overrightarrow{h^n{}_{cf}}$ is compared with each of these model histograms $\overrightarrow{h^m_{cf_1}}$.........$\overrightarrow{h^m_{cf_k}}$ and the similarity measures between the new histogram and the model histograms are computed. The similarity measure is obtained as the histogram intersection between two histograms and is given as,

$$\cap(\overrightarrow{h^m_{cf_k}}, \overrightarrow{h^n_{cf}}) = \sum_{k=1}^{K}(\overrightarrow{h^m_{cf_k}}, \overrightarrow{h^n_{cf}}). \tag{7}$$

If the proximity value is found to be less than a preselected threshold T for all the model histograms, the histogram with the lowest weight is replaced with the new histogram and the replaced histogram is assigned with a low value of weight of 0.01. If proximity measures of some of the model histograms with that of the new histogram are found to be above the threshold T, then the new histogram is considered to match with the model histograms. In this case, the best matched model histogram with the highest proximity value is adapted with the new fused histogram $\overrightarrow{h^m_{cf_n}}$ as,

$$\overrightarrow{h^m_{cf_k}} = \alpha_1 \overrightarrow{h^n_{cf}} + (1 - \alpha_1)\overrightarrow{h^m_{cf_k}}, \tag{8}$$

where $\overrightarrow{h^m_{cf_k}}$ denotes the best match fused feature model histogram and α_1 is chosen between 0 to 1. Besides the bin updation of the histogram, the weights of the model histograms are updated as,

$$w_k = \alpha_2 \theta_k + (1 - \alpha_2) w_k, \tag{9}$$

where α_2 is the learning rate. The value of θ_k is 1 for the best matched histogram and 0 for others. α_1 and α_2 are the learning rates which are user-defined parameters and the adaptation of the model histogram is controlled by these learning rate parameters. Each of the model histograms represent either the background or the foreground depending on the assigned weights. After updation, the weights of the adapted model histograms are arranged in decreasing order and the first B model histograms are considered as the background model histograms if the following condition is satisfied.

$$w_0 + w_1 - - - - - - - - - + w_{B-1} \geq T_B \tag{10}$$

where T_B is chosen to be between 0 and 1.

A. CLASSIFICATION OF BACKGROUND AND FOREGROUND

Classification of foreground or background of a pixel is accomplished before updation of the model histograms and the weights. The proximity of the new fused feature histogram $\overrightarrow{h^n_{cf}}$ is determined with each of the background model histograms and if at least one of the model histograms' proximity is higher than the selected T_p, the pixel is classified as background. If proximity of none of the model histograms with that of the new input histograms are above the threshold T_p then the pixel is classified as foreground.

VII. EXPERIMENTAL CONDITIONS, BENCHMARK DATASETS AND PERFORMANCE METRICS

In this section, various benchmark datasets used in our proposed method are presented. Besides, different performance metrics used in our experiment as quantitative measures are also presented.

A. BENCHMARK DATASETS

The objective is to test the efficacy of the proposed method over different scenarios with a wide variety of conditions of shadows and to perform the quantitative analysis to validate the results. In this regard, we have used the following datasets: Scene Background Modeling and Initialization (SBMI 2015) [77], Scene Background Modeling (SBMnet 2016) [78], Scene Background Modeling (SBM-RGBD 2017) [79], PETS 2006, LASIESTA (Labeled dataset for integral evaluation of moving object detection algorithms 2016), Change Detection dataset (CDnet 2014) and ATON-CVRR 2000 [77], [80], [81], Kaggle UCF crime dataset 2019, VIRAT 2020. The overall details of the datasets including brief description, type of shadow, length of the shadow and the class of object are presented in Table 1. Simulation is carried out in Linux platform with coding in C-language using the Machine Specifications which is specified as HP Presario CQ62 model with Intel Core i3- 380M Processor, 2GB DDR3 RAM, 2.5 GHz CPU, and 64 bit OS.

The sources of the above datasets can be obtained at the following URLs https://www.kaggle.com/mission-ai/ucfdatasetforanomaly (kaggle UCF crime), http://rgbd2017.na.icar.cnr.it/SBM-RGBDdataset.html (SGM-RGBD 2017) [79], http://scenebackgroundmodeling.net/ (SBMnet 2016), http://sbmi2015.na.icar.cnr.it/ (SBMI 2015), http://www.cvg.reading.ac.uk/PETS2006/data.html(PETS 2006) https://viratdata.org/(VIRAT 2020 version) https://www.gti.ssr.upm.es/data/lasiesta_database.html (LASIESTA) [80], http://changedetection.net(CDnet) [81], http://cvrr.ucsd.edu/aton/shadow (ATON CVRR) [77].

B. PERFORMANCE METRICS

We have used the following quantitative measures for analyzing our data and evaluating the efficacy of our proposed scheme.

$$Recall(Re) = \frac{T_P}{T_P + F_N}. \tag{11}$$

$$Precision(Pr) = \frac{T_P}{T_P + F_P}. \tag{12}$$

$$Specificity(Sp) = \frac{T_N}{T_N + F_P}. \tag{13}$$

$$Fmeasure(F - score) = \frac{2PrRe}{Pr + Re}. \tag{14}$$

$$FPR = \frac{F_P}{F_P + T_N}. \tag{15}$$

$$FNR = \frac{F_N}{T_N + F_P}. \tag{16}$$

$$PWC = \frac{(F_N + F_P) \times 100}{T_P + F_N + F_P + T_N}. \tag{17}$$

where T_P is the number of true positives, T_N is the number of true negatives, F_N is the number of false negatives, and F_P is the number of false positives, FPR is the false positive rate, FNR is the false negative rate and PWC is the percentage of wrong classifications.

In order to evaluate the shadow detection ability of the proposed algorithm, the shadow detection and removal rate (η) and shadow discrimination rate (ξ) are used. These are expressed as,

$$\eta = \frac{T_{Ps}}{T_{Ps} + F_{Ns}}, \tag{18}$$

$$\xi = \frac{T_{Pf}}{T_{Pf} + F_{Nf}}, \tag{19}$$

where TP and FN indicate true positive and false negative pixels with respect to either shadow or foreground objects. T_{Ps} indicates TP with respect to shadow and T_{Pf} indicates TP with respect to foreground. Similarly, F_{Ns} indicates FN with respect to shadow and F_{Nf} indicates FN with respect to foreground.

VIII. RESULTS AND DISCUSSION

A. EXPERIMENTAL RESULTS AND ANALYSIS ON VARIOUS DATASETS

In this section, we analyze the performance of our proposed method on different datasets both qualitatively and quantitatively. The quantitative measures are evaluated over twenty frames in all examples and the average values are presented in different tables. The qualitative results are post processed using different morphological operations like erosion and dilation to obtain the final results. The first dataset considered is the LASIESTA database and the three video scenes considered are O_SU_01 and O_SU_02 and I_IL_02 which are as shown in Fig. 8. The second row of Fig. 8 corresponds to the ground truth and the segmented results of the proposed scheme are shown in the third row of Fig. 8. As observed, the static and dynamic shadows corresponding to the background and foreground are removed and the object is also detected except a few misclassified pixels of the background. This is because of the presence of shadow in the indoor scene of Fig.8 (c) while the light is fading away from the brighter side of the room. However, the average parameters for quantitative measures over 20 frames are presented in Table 2 where it may be observed that the precision is high as well as the F score besides other parameters. The shadow detection rate η is of a high value indicating that the proposed algorithm could detect the shadow with a high degree and discriminate it from the foreground. The values of other parameters are appreciable in their respective merits. This may be attributed to our feature fusion strategy with the cascaded features and learning the background models in feature space. The performance of the proposed scheme is also compared with other algorithms in the context of the average F score and the different scores are presented in Table 3. It is found that the proposed method yielded the highest F score in two cases and for O_SU_02, our previously proposed method produced the highest one. This demonstrates the superiority of our proposed algorithm over others.

The second example video is considered from ATON-CVRR [77] dataset and the frames are shown in Fig. 9. Though the average quantitative measure parameters are

TABLE 1. Indoor Outdoor sequences in various datasets.

Dataset and Identifier	*Sequence_Type*	*Brief Description and Issues*	*Shadow Length*	*Shadow Type*	*Object Class*
LASIESTA 2016	Outdoor	Dynamic background with hard shadows	Medium	Hard	People
SBM-RGBD 2017	Indoor	Illumination changes with soft and hard shadows	Medium	Hard	People and Object
CDnet 2014	Outdoor	Dynamic background with hard shadows	Large	Hard	People and Object
ATON-CVRR 2000	Indoor, Outdoor	Illumination changes with hard and soft shadows	Medium	Hard	People
UCF Crime Kaggle 2019	Outdoor	Dynamic background with hard shadows	Medium	Hard	People
VIRAT 2020	Outdoor	Dynamic background, illumination changes with hard shadows	Medium	Hard	People
PETS 2006	Outdoor	Hard shadows	Medium	Hard	People
SBMnet 2016	Outdoor	Dynamic background with hard shadows	Medium	Hard	People
SBMI 2015	Indoor	Illumination changes with hard shadows	Medium	Hard	People

TABLE 2. Quantitative analysis on different datasets based on average performance measures.

Datasets	Pr	Re	F-score	Sp	FPR	FNR	PWC	η	ξ
LASIESTA dataset									
O_SU_01	97.68	85.80	91.44	99.96	0.63	31.18	1.07	96.99	85.80
O_SU_02	97.78	86.60	91.85	99.94	0.05	13.39	0.42	98.04	86.60
I_IL_02	87.81	90.03	88.91	98.37	1.62	12.18	1.21	98.69	90.03
SGM-RGB dataset									
Fall01Cam	93.11	94.25	93.67	99.44	0.65	12.61	1.65	96.64	94.25
Genseq2	92.65	94.22	93.42	99.75	0.74	6.43	1.61	93.54	94.22
Shadow_ds	89.72	95.22	92.38	99.81	1.25	7.65	1.67	93.59	95.22
Shadows1	77.41	94.23	84.99	95.12	2.85	13.39	3.31	90.76	94.23
CD-net dataset									
Bunglow	91.53	92.89	92.05	99.68	0.31	8.10	0.53	99.80	92.89
Busstation	81.38	91.89	86.18	94.79	5.20	8.43	5.84	99.81	91.89
Pedestrian	97.75	98.20	97.23	99.28	0.71	3.14	0.75	98.91	98.20
ATON-CVRR dataset									
Campus	91.53	92.89	92.05	99.68	0.31	8.10	0.53	99.80	92.89
Laboratory_ds	92.91	95.80	94.21	96.31	3.68	17.19	4.33	95.13	95.80
Hallway	86.93	90.20	88.53	97.21	3.78	18.39	8.31	93.98	90.20
UCF-Crime Kaggle dataset									
N1	82.53	85.89	84.17	94.40	5.59	34.60	8.20	76.91	85.89
N2_ds	80.81	85.07	82.88	99.74	0.25	31.92	1.01	98.46	85.07
N5	85.72	93.95	89.64	94.92	5.07	6.04	5.14	95.09	93.95
VIRAT dataset									
Virat 01	83.60	86.63	85.09	99.42	0.57	15.43	0.69	88.79	86.63
PETS 2006 dataset									
PETS	85.72	88.51	87.09	97.19	2.80	20.77	4.72	87.62	88.51
SBMnet 2016 dataset									
SBMnet	83.57	91.36	87.29	99.89	0.10	28.63	0.31	89.92	91.36
SBMI 2015 dataset									
Human Body	92.75	89.70	91.20	98.99	1.00	10.29	2.17	99.67	89.70

TABLE 3. Average F-score on LASIESTA database [80].

Method →	Wren [1]	Stauffer [2]	Zivkovik [3]	Maddalena 1 [4]	Maddalena 2 [5]	Cuevas 1 [6]	Haines [7]	Cuevas 2 [8]	REAM [67]	STKDE [33]	Proposed
O_SU_01	0.6808	0.6177	0.5426	0.7467	0.8742	0.6527	0.8115	0.6774	0.9402	0.8928	0.914
O_SU_02	0.8304	0.8304	0.8775	0.8562	0.883	0.8074	0.9021	0.7669	0.9411	0.934	0.918
I_IL_02	0.4568	0.2392	0.3135	0.3750	0.2312	0.7864	0.8122	0.6523	0.9479	0.8345	0.889

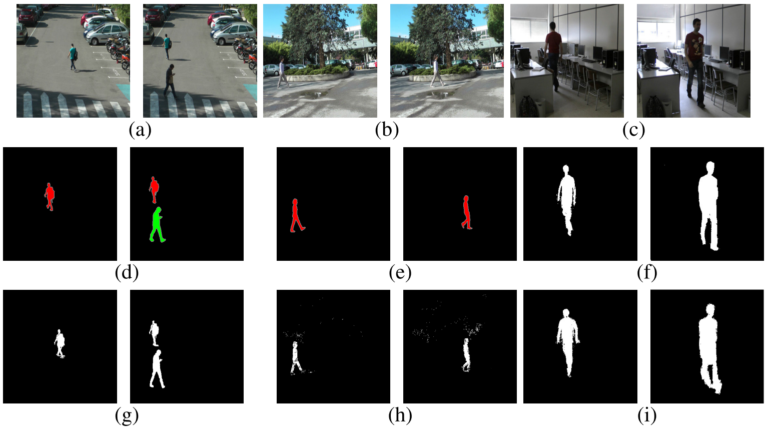


FIGURE 8. LASIESTA data set: 1st row indicates the original image frames (a) O_SU_02 (200, 250), (b) O_SU_01 (130, 180), (c) I_IL_02 (125,350), 2nd row (d)-(f) corresponds to the ground truth and, 3rd row (g)-(i) are the results by the proposed approach.

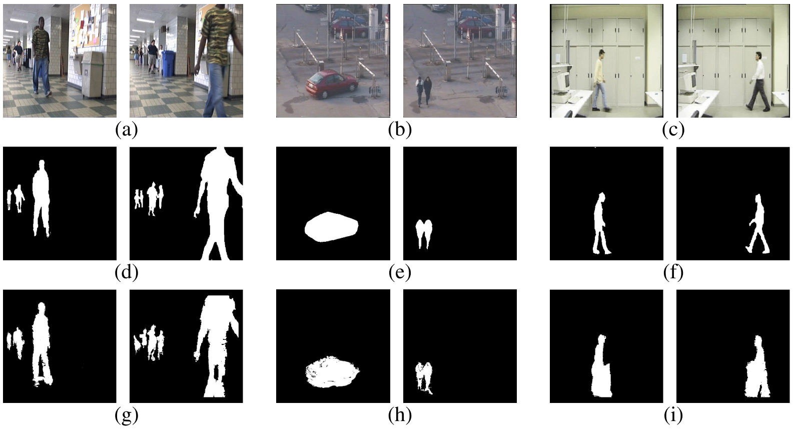


FIGURE 9. ATON-CVRR data set: 1st row indicates the original image frames (a) Hallway (040, 042), (b) Campus (020, 100), (c) Laboratory (025,038), 2nd row (d)-(f) corresponds to the ground truth and, 3rd row (g)-(i) are the results by the proposed approach.

determined over 20 frames, for the sake of visual inspection, two frames from each category of the datasets of Hallway, Campus and Laboratory are shown in Fig. 9. The second row and the third row of Fig.9 correspond to the ground truth and the segmented results respectively. The Laboratory video is an indoor scene having complex background. Besides, Fig. 9(a) and 9(b) have multiple moving objects with shadows. It is found from the visual inspection that the frames of Campus video could be segmented with accurate shapes of the moving objects. However, in other two cases, there are some misclassified pixels despite the object's shape being detected. Different average quantitative measures evaluated over 20 frames are presented in Table 2, where it may be observed that the precision, recall, and F-score values are high indicating the fact that the proposed algorithm could detect the object and remove the shadows. Specifically, the proposed algorithm's shadow detection and discrimination ability is compared with other existing algorithms and are presented in Table 4, where it is observed that for frames of the Campus dataset, the shadow detection rate is the highest for the proposed algorithm. But, for the Laboratory dataset, the shadow discrimination ability of the proposed algorithm is highest among all. This indicates that the proposed algorithm has better shadow detection and discrimination ability than other algorithms and therefore it helps in detecting the moving objects accurately in the complex scene.

The third dataset considered is the change detection dataset (CDnet dataset) [81] where frames are considered from Bunglow, Busstation and Pedestrian videos and are shown in Fig.10. It may be observed from Fig. 10(a) that the frames from the Bunglow video have moving shadows of the object in sharp motion which is a challenge. Similarly, the frames of Busstation video has multiple moving objects with both static and dynamic shadows. The results obtained by our proposed method and others are also shown in Fig.10 where it may be observed that results obtained by our proposed algorithm are comparable to others and in some cases better than other algorithms. Different parameters of quantitative measures are presented in Table 2 while specifically the average η and ξ values are compared with other algorithms and are presented in Table 5. It is found that in all the cases, the shadow detection rate is highest for the proposed algorithm while

TABLE 4. Comparisons based on average quantitative measures on the standard datasets of ATON-CVRR adapted from Lin [40].

Benchmark Datasets	Metric	DNM [35]	GBM [39]	ICF [16]	SNP2 [37]	CCM [41]	ASE [42]	LRT [38]	NTM [43]	SDM [44]	FCN-VGG16 [55]	SMPF [40]	STKDE [33]	Proposed
Campus	η	57.00	66.60	56.22	65.85	62.56	67.50	82.10	53.23	52.09	86.32	86.38	89.66	**99.80**
	ξ	50.30	54.80	82.74	75.36	43.07	70.82	97.70	81.36	97.58	96.97	**98.98**	94.25	92.89
Laboratory	η	**98.90**	49.80	92.22	92.93	91.63	91.37	86.30	68.87	89.17	83.34	95.29	82.19	95.13
	ξ	77.80	67.90	89.69	72.71	84.04	92.36	97.00	67.66	92.66	76.54	95.79	84.10	**95.80**
Hallway	η	**96.70**	53.00	95.32	83.93	96.46	90.09	94.65	85.16	83.45	81.22	94.16	95.07	93.98
	ξ	77.80	73.80	83.02	98.10	68.55	93.19	98.02	83.18	**99.25**	83.27	98.58	94.22	90.20

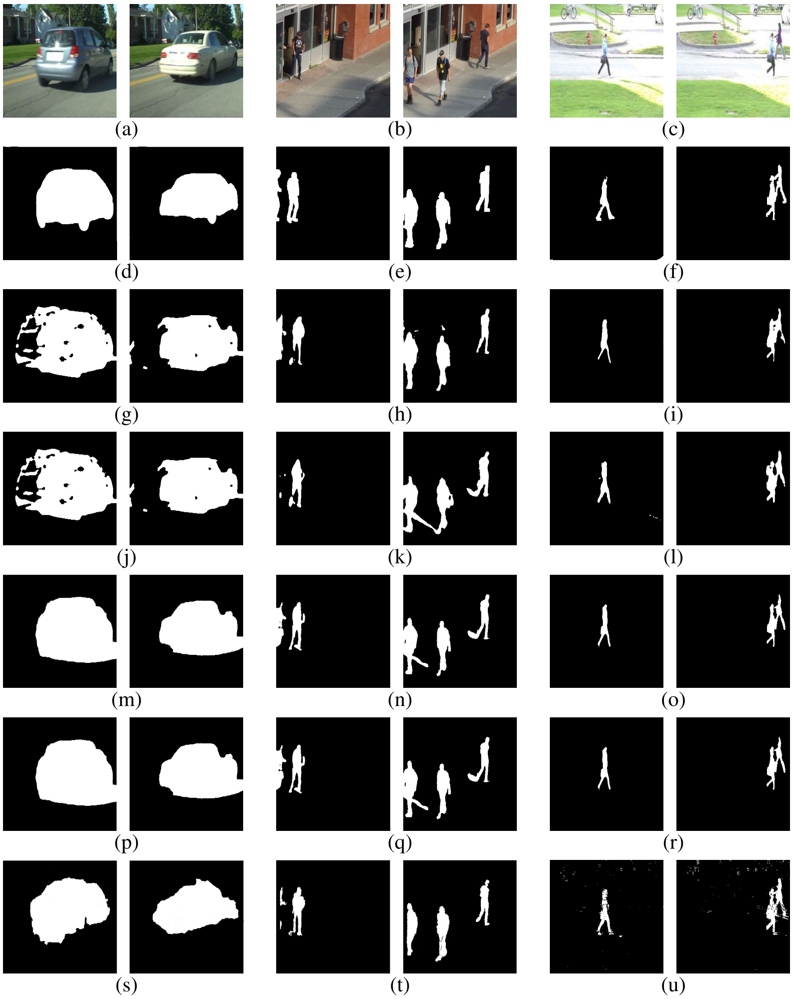


FIGURE 10. CDnet data set: 1st row indicates the original image frames (a) Bunglow (033, 144), (b) Busstation (385, 1045), (c) Pedestrian (375, 425), 2nd row (d)-(f) corresponds to the ground truth, (g)-(i) are the results by MUNet1, (j)-(l) are the results by RTSBSv1, (m)-(o) are the results by Fast BSUVNet 2.0, (p)-(r) are the results by FgSegNet v2 CO and, (s)-(u) are the results by the proposed approach.

the shadow discrimination is the highest for the example of Busstation an Pedestrian videos. This is attributed to the potentialities of the cascaded features and the fused feature. The results of this example also indicate the superiority of the proposed algorithm in the context of η and ξ.

The fourth dataset is the SGM-RGBD dataset [79] where the frames considered have hard shadows cast by the foreground objects. Original frames are shown in the first row of Fig. 11 whereas the second row corresponds to the ground truth and the third row shows the results obtained by our proposed algorithm. The average quantitative results are presented in Table 2 and the comparative results of different algorithm with respect to F-score are presented in Table 6. It is observed that in this case also the F-scores of the proposed algorithm are either close or comparable to other existing algorithms. This example also demonstrates the efficacy of our proposed algorithm. Here also the shadow detection and removal capabilities are appreciable.

In order to test the efficacy of the algorithm, we have considered scenes where dynamism is present in the background entities with moving shadows that otherwise would have been static. This apparent moving shadow camouflages with that of the shadow due to the moving object. Besides, there are multiple moving objects of different sizes and

TABLE 5. Comparisons based on average of quantitatives on standard scenes of CDnet dataset.

Benchmark Dataset	Metric	RTSBSv1 [54]	MUNet1 [61]	Fast BSUVNet 2.0 [62]	FgSegNet_v2_CO [63]	STKDE [33]	Proposed
Bunglow	η	36.1	56.09	76.45	82.76	72.66	**99.80**
	ξ	89.73	89.01	91.76	**92.73**	74.56	91.89
Busstation	η	53.04	98.12	71.48	67.42	84.38	**99.81**
	ξ	84.08	86.48	85.24	85.42	74.63	**89.56**
Pedestrian	η	99.16	99.11	99.02	99.32	**99.48**	98.91
	ξ	97.08	97.32	96.97	96.97	97.88	**98.20**

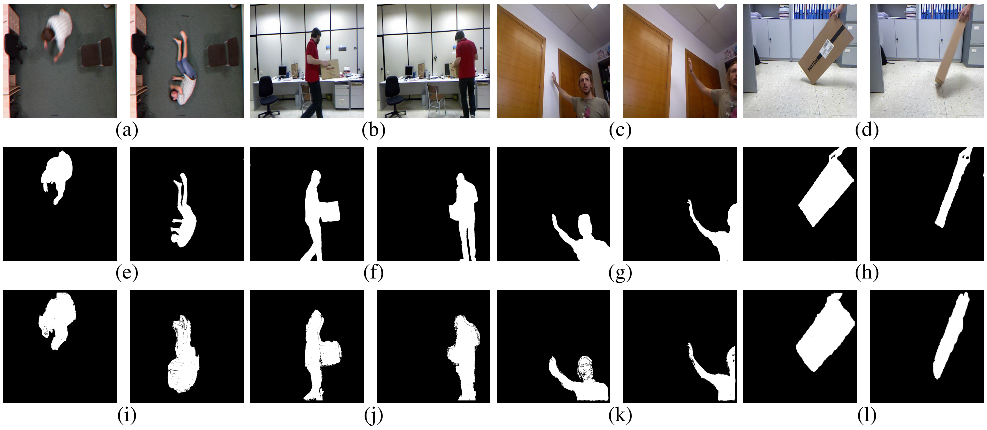


FIGURE 11. SGM-RGBD dataset: 1st row indicates the original image frames (a) Fall01Cam (100, 125), (b) genseq2 (175, 200), (c) shadow_ds (200,250), (d) shadows1 (125,150) 2nd row (e)-(h) corresponds to the ground truth and, 3rd row (i)-(l) are the results by the proposed approach.

TABLE 6. Average F-score on SGM-RGBD database [79].

Method →	RGBD-SOBS [48]	SRPCA [21]	SCAD [49]	cwisardH+ [71]	MFCN [82]	BSABU [83]	STKDE [33]	Proposed
Fall01Cam	0.92	0.76	0.91	0.91	0.97	0.91	0.89	0.87
genseq2	0.93	0.77	0.94	0.92	0.98	0.90	0.91	0.91
shadow_ds	0.95	0.75	0.97	0.90	0.98	0.95	0.90	0.95
shadows1	0.95	0.75	0.94	0.92	0.98	0.95	0.90	0.96

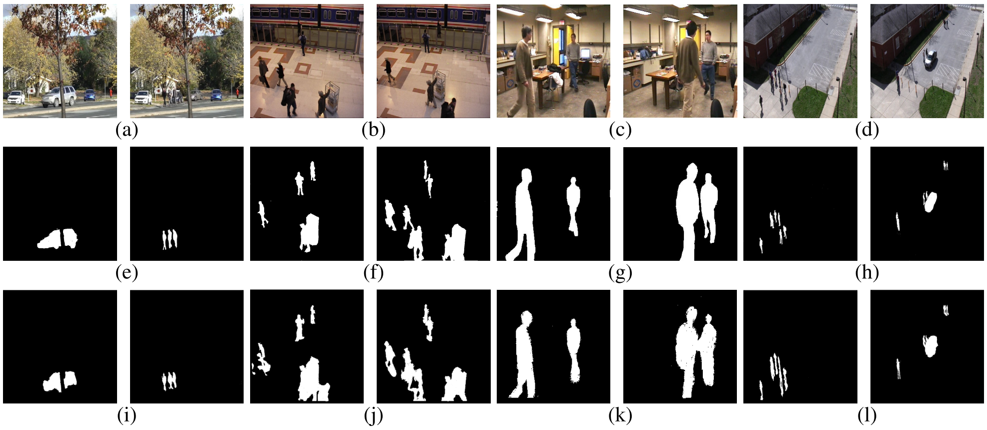


FIGURE 12. Different datasets: 1st row indicates the original image frames (a) SBMnet-2016 Fall (200, 300), (b) Pets 2006 (385, 1015), (c) SBMInet 2015 Humanbody (275,300), (d) VIRAT1 (16, 22), 2nd row (e)-(h) corresponds to the ground truth and, 3rd row (i)-(l) are the results by the proposed approach.

hence the moving shadows are also of different sizes. Some of these frames are shown in the first row of Fig. 12. It may be observed from the third row that objects, single and multiple, could be detected with different scenes and shadow conditions. Different average quantitative measures for these datasets are presented in Table 2 where it may

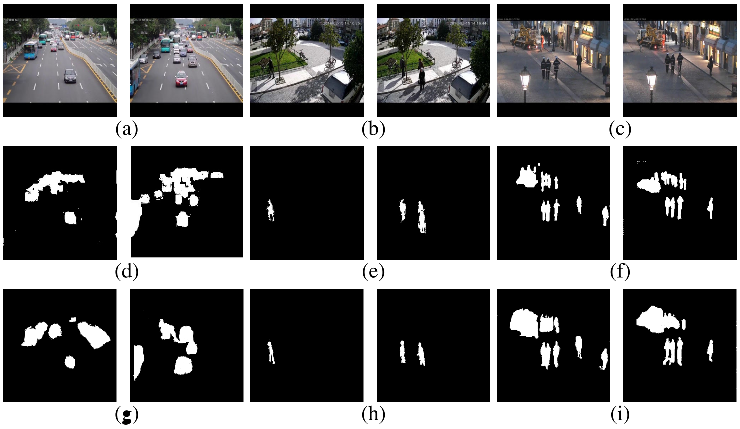


FIGURE 13. Kaggle UCF crime datasets: 1st row indicates the original image frames (a) N1 (50,71), (b) N2 (17, 36), (c) N5 (12, 15), 2nd row (d)-(f) corresponds to the ground truth and, 3rd row (g)-(i) are the results by the proposed approach.

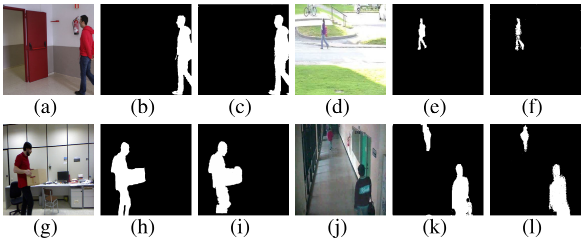


FIGURE 14. Non Shadow cases of the algorithm: (a) Camouflage (LASIESTA), (b) Ground truth, (c) Segmented result, (d) Pedestrian (CDnet), (e) Ground truth, (f) Segmented result, (g) genseq2 (SGM-RGBD), (h) Ground truth, (i) Segmented result, (j) IPPR2 (SBMnet), (k) Ground truth, (l) Segmented result.

be observed that the values of η and ξ are appreciable thus indicating the fact that the proposed algorithm has been embedded with these two attributes as it could handle different shadow conditions. There are appreciable values of precision, recall, and F-score which demonstrate the object detection attribute. The proposed background model learning in the feature space together with the feature fusion takes care of this. Similar observations are also made for the frames of Kaggle UCF crime dataset as shown in Fig. 13. The object detection and shadow removal attribute is reflected in the quantitative measures tabulated in Table 2. Thus, in all the cases, the proposed algorithm demonstrated superior shadow handling and object detection capability as compared to others. Besides, to validate the proposed scheme for both cases of images with and without shadows, the proposed scheme is tested with non shadow images from four databases and the segmented results are provided in the Fig. 14. The average value of different quantitative measures are provided in Table 7. As observed from Table 7, the F-score measures are above 90% and are comparable to the values of Table 2 for shadow cases. Similar observations are also made for precision and recall measures. Hence the proposed scheme works well for detecting moving video objects in both the environments having shadow and non shadow images.

TABLE 7. Quantitative analysis on different non shadow datasets.

Identifier	*LASIESTA*	*CDnet*	*SGM-RGBD*	*SBMnet*
F-score	93.45	92.49	91.05	90.49
Precision	95.69	97.81	96.45	92.23
Recall	89.11	87.73	86.14	83.23

B. TIME COMPLEXITY

The proposed method is a pixel based approach and hence the computational time per pixel with the given machine specifications which is provided in subsection A of section VII is found to be 1.43 milliseconds per pixel. For example, for a frame of size (100 × 100), the execution time will be 14.3 seconds. The execution time can further be reduced with the help of a machine with enhanced specifications. It is found that the execution time per pixel remains same for all the examples as the window size considered around the pixel is constant for all the cases. Hence, the computational burden per frame is proportional to the frame size. For example, in case of the considered datasets of LASIESTA and ATON CVRR datasets with frame size of (352 × 288) and (320 × 240), the execution times are 144.96 seconds and 109.82 seconds respectively.

IX. CONCLUSION

In this paper, attempts have been made to detect moving video object with dynamic and static shadow conditions. The dynamic shadow which occurs due to the moving object is hard to separate from the object itself. Besides, the shadow caused due to dynamic entities of the background makes the problem more challenging. The proposed scheme is found to take care of the different types of shadows while detecting the moving object. Specifically, it is also found to take care of the moving shadow of the background. This could be achieved because of the fusion of cascaded features with the LBP features. The cascaded features are extracted from the TMS-KDE model of the frame which helps preserve the boundary of the object and the object itself. The cascaded feature could eliminate the shadow to some extent. Learning of the model histograms takes place in the fused feature space, and it occurs in online mode as opposed to usual offline mode of training in deep learning network models. Feature fusion, learning, and classification of foreground and background pixels happen simultaneously. The weights for feature fusion are determined considering the scene dynamics and hence appropriate feature is chosen for modeling the scene. It is also found that learning could take care of shadow while detecting the object.

In our experiments, we have considered eight different and diverse datasets: LASIESTA, ATON CVRR, CDnet, SBMnet, SBMI, kaggle, VIRAT, SGM-RGBD datasets. Besides, these different datasets have different types and degrees of shadow in their scenes. The proposed algorithm is found to outperform many existing algorithms as regards the precision and F-score. However, the other quantitative measures are also either comparable or better than the existing algorithms. Additionally, the proposed algorithm has improved shadow detection and discrimination attribute. Thus this scheme is suitable for a wide variety of indoor and outdoor scenes. In future, efficient background modeling and learning for complex scenes is worth pursuing.

ACKNOWLEDGMENT

The authors would like to acknowledge the research facility provided by Siksha 'O' Anusandhan, Deemed to be University for carrying out this work.

REFERENCES

[1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[2] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[3] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, May 2006.

[4] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[5] L. Maddalena and A. Petrosino, "The SOBS algorithm: What are the limits?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 21–26.

[6] C. Cuevas and N. García, "Improved background modeling for real-time spatio-temporal non-parametric moving object detection strategies," *Image Vis. Comput.*, vol. 31, no. 9, pp. 616–630, Sep. 2013.

[7] T. S. F. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, Apr. 2014.

[8] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognit.*, vol. 74, pp. 156–170, Feb. 2018.

[9] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 918–923, Jul. 2003.

[10] A. Cavallaro, E. Salvador, and T. Ebrahimi, "Shadow-aware object-based video processing," *IEE Proc. Vis., Image Signal Process.*, vol. 152, no. 4, pp. 398–406, Aug. 2005.

[11] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[12] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.

[13] T. Bouwmans, C. Silva, C. Marghes, M. S. Zitouni, H. Bhaskar, and C. Frelicot, "On the role and the importance of features for background modeling and foreground detection," *Comput. Sci. Rev.*, vol. 28, pp. 26–91, May 2018.

[14] S. R. R. Sanches, C. Oliveira, A. C. Sementille, and V. Freire, "Challenging situations for background subtraction algorithms," *Int. J. Speech Technol.*, vol. 49, no. 5, pp. 1771–1784, May 2019.

[15] C.-T. Chen, C.-Y. Su, and W.-C. Kao, "An enhanced segmentation on vision-based shadow removal for vehicle detection," in *Proc. Int. Conf. Green Circuits Syst.*, Jun. 2010, pp. 679–682.

[16] E. Salvador, A. Cavallaro, and T. Ebrahimi, "Cast shadow segmentation using invariant color features," *Comput. Vis. Image Understand.*, vol. 95, no. 2, pp. 238–259, Aug. 2004.

[17] C.-C. Chen and J. K. Aggarwal, "Human shadow removal with unknown light source," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2407–2410.

[18] J.-W. Hsieh, W.-F. Hu, C.-J. Chang, and Y.-S. Chen, "Shadow elimination for effective moving object detection by Gaussian shadow modeling," *Image Vis. Comput.*, vol. 21, no. 6, pp. 505–516, Jun. 2003.

[19] H. Nicolas and J.-M. Pinel, "Joint moving cast shadows segmentation and light source detection in video sequences," *Signal Process., Image Commun.*, vol. 21, no. 1, pp. 22–43, Jan. 2006.

[20] A. Yoneyama, C. H. Yeh, and C.-C.-J. Kuo, "Moving cast shadow elimination for robust vehicle extraction based on 2D joint vehicle/shadow models," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Jul. 2003, pp. 229–236.

[21] S. Javed, T. Bouwmans, M. Sultana, and S. K. Jung, "Moving object detection on RGB-D videos using graph regularized spatiotemporal RPCA," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 230–241.

[22] Y.-L. Tian, M. Lu, and A. Hampapur, "Robust and efficient foreground analysis for real-time video surveillance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 1182–1187.

[23] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 279–289, Feb. 2006.

[24] W. Zhang, X. Zhong Fang, X. K. Yang, and Q. M. J. Wu, "Moving cast shadows detection using ratio edge," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1202–1214, Oct. 2007.

[25] D. Xu, X. Li, Z. Liu, and Y. Yuan, "Cast shadow detection in video segmentation," *Pattern Recognit. Lett.*, vol. 26, no. 1, pp. 91–99, Jan. 2005.

[26] N. Martel-Brisson and A. Zaccarin, "Learning and removing cast shadows through a multidistribution approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 7, pp. 1133–1146, Jul. 2007.

[27] A. J. Joshi and N. P. Papanikolopoulos, "Learning to detect moving shadows in dynamic environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 2055–2063, Nov. 2008.

[28] W. J. Kim, S. Hwang, J. Lee, S. Woo, and S. Lee, "AIBM: Accurate and instant background modeling for moving object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9021–9036, Jul. 2022.

[29] J. Hao, C. Li, Z. Kim, and Z. Xiong, "Spatio-temporal traffic scene modeling for object motion detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 295–302, Mar. 2013.

[30] C.-W. Liang and C.-F. Juang, "Moving object classification using a combination of static appearance features and spatial and temporal entropy values of optical flows," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3453–3464, Dec. 2015.

[31] C.-Y. Lin, K. Muchtar, W.-Y. Lin, and Z.-Y. Jian, "Moving object detection through image bit-planes representation without thresholding," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1404–1414, Apr. 2020.

[32] A. Amato, I. Huerta, M. G. Mozerov, F. X. Roca, and J. Gonzalez, "Moving cast shadows detection methods for video surveillance applications," in *Wide Area Surveillance*. Berlin, Germany: Springer, 2014, pp. 23–47.

[33] S. Sahoo and P. K. Nanda, "Adaptive feature fusion and spatio-temporal background modeling in KDE framework for object detection and shadow removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1103–1118, Mar. 2022.

[34] K. Onoguchi, "Shadow elimination method for moving object detection," in *Proc. 14th Int. Conf. Pattern Recognit.*, vol. 1, Aug. 1998, pp. 583–587.

[35] C. Jiang and M. O. Ward, "Shadow identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1992, pp. 606–607.

[36] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," 2013, *arXiv:1302.1539*.

[37] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. Int. Conf. Comput. Vis.*, vol. 99, Jan. 1999, pp. 1–19.

[38] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 141–144.

[39] A. Sanin, C. Sanderson, and B. C. Lovell, "Shadow detection: A survey and comparative evaluation of recent methods," *Pattern Recognit.*, vol. 45, no. 4, pp. 1684–1695, Apr. 2012.

[40] C.-W. Lin, "Moving cast shadow detection using scale-relation multi-layer pooling features," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 504–517, Aug. 2018.

[41] B. Sun and S. Li, "Moving cast shadow detection of vehicle using combined color models," in *Proc. Chin. Conf. Pattern Recognit. (CCPR)*, Oct. 2010, pp. 1–5.

[42] J. Choi, Y. J. Yoo, and J. Y. Choi, "Adaptive shadow estimator for removing shadow of moving object," *Comput. Vis. Image Understand.*, vol. 114, no. 9, pp. 1017–1029, Sep. 2010.

[43] E. Bullkich, I. Ilan, Y. Moshe, Y. Hel-Or, and H. Hel-Or, "Moving shadow detection by nonlinear tone-mapping," in *Proc. 19th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Apr. 2012, pp. 146–149.

[44] C.-W. Lin, "Scale-relation feature for moving cast shadow detection," in *Proc. Int. Conf. Multimedia Model.* Berlin, Germany: Springer, Dec. 2017, pp. 331–342.

[45] W. Zhang, X. Zhong Fang, and Y. Xu, "Detection of moving cast shadows using image orthogonal transform," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 626–629.

[46] Z. Liu, K. Huang, T. Tan, and L. Wang, "Cast shadow removal combining local and global features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[47] F. Porikli and J. Thornton, "Shadow flow: A recursive method to learn moving cast shadows," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 891–898.

[48] L. Maddalena and A. Petrosino, "Exploiting color and depth for background subtraction," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 254–265.

[49] T. Minematsu, A. Shimada, H. Uchiyama, and R.-I. Taniguchi, "Simple combination of appearance and depth for foreground segmentation," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2017, pp. 266–277.

[50] S. Luo, H. Li, R. Zhu, Y. Gong, and H. Shen, "ESPFNet: An edge-aware spatial pyramid fusion network for salient shadow detection in aerial remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4633–4646, 2021.

[51] L. Sun, Q. Wang, X. Zhou, J. Wei, X. Yang, W. Zhang, and N. Ma, "A priori surface reflectance-based cloud shadow detection algorithm for Landsat 8 OLI," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1610–1614, Oct. 2018.

[52] G. Alvarado-Robles, R. A. Osornio-Ríos, F. J. Solís-Muñoz, and L. A. Morales-Hernández, "An approach for shadow detection in aerial images based on multi-channel statistics," *IEEE Access*, vol. 9, pp. 34240–34250, 2021.

[53] Z. Liu, D. An, and X. Huang, "Moving target shadow detection and global background reconstruction for VideoSAR based on single-frame imagery," *IEEE Access*, vol. 7, pp. 42418–42425, 2019.

[54] A. Cioppa, M. V. Droogenbroeck, and M. Braham, "Real-time semantic background subtraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3214–3218.

[55] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and Bayesian GANs," *Neurocomputing*, vol. 394, pp. 178–200, Jun. 2020.

[56] S. Mohajerani and P. Saeedi, "Shadow detection in single RGB images using a context preserver convolutional neural network trained by multiple adversarial examples," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4117–4129, Aug. 2019.

[57] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 682–695, Mar. 2018.

[58] C. Wang, H. Xu, Z. Zhou, L. Deng, and M. Yang, "Shadow detection and removal for illumination consistency on the road," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 4, pp. 534–544, Dec. 2020.

[59] J. Kim and W. Kim, "Attentive feedback feature pyramid network for shadow detection," *IEEE Signal Process. Lett.*, vol. 27, pp. 1964–1968, 2020.

[60] S. He, B. Peng, J. Dong, and Y. Du, "Mask-ShadowNet: Toward shadow removal via masked adaptive instance normalization," *IEEE Signal Process. Lett.*, vol. 28, pp. 957–961, 2021.

[61] G. Rahmon, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Motion U-Net: Multi-cue encoder–decoder network for motion segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8125–8132.

[62] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction," *IEEE Access*, vol. 9, pp. 53849–53860, 2021.

[63] F. Gao, Y. Li, and S. Lu, "Extracting moving objects more accurately: A CDA contour optimizer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4840–4849, Dec. 2021.

[64] P. W. Patil and S. Murala, "MSFgNet: A novel compact end-to-end deep network for moving object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4066–4077, Nov. 2019.

[65] C. Zhao, K. Hu, and A. Basu, "Universal background subtraction based on arithmetic distribution neural network," *IEEE Trans. Image Process.*, vol. 31, pp. 2934–2949, 2022.

[66] B. Wang, Y. Zhao, and C. L. P. Chen, "Moving cast shadows segmentation using illumination invariant feature," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2221–2233, Sep. 2020.

[67] P. W. Patil, A. Dudhane, A. Kulkarni, S. Murala, A. B. Gonde, and S. Gupta, "An unified recurrent video object segmentation framework for various surveillance environments," *IEEE Trans. Image Process.*, vol. 30, pp. 7889–7902, 2021.

[68] C. Zhao and A. Basu, "Dynamic deep pixel distribution learning for background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4192–4206, Nov. 2020.

[69] S. M. Roy and A. Ghosh, "Foreground segmentation using adaptive 3 phase background model," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2287–2296, Jun. 2020.

[70] H. Zhang, S. Qu, H. Li, J. Luo, and W. Xu, "A moving shadow elimination method based on fusion of multi-feature," *IEEE Access*, vol. 8, pp. 63971–63982, 2020.

[71] M. De Gregorio and M. Giordano, "CWISARDH$^+$: Background detection in RGBD videos by learning of weightless neural networks," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 242–253.

[72] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10212–10221.

[73] Q. Zheng, X. Qiao, Y. Cao, and R. W. H. Lau, "Distraction-aware shadow detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5162–5171.

[74] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7454–7462.

[75] H. Le, T. F. Y. Vicente, V. Nguyen, M. Hoai, and D. Samaras, "A+D Net: Training a shadow detector with adversarial shadow attenuation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 662–678.

[76] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8577–8586.

[77] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, pp. 469–476, 2015.

[78] P.-M. Jodoin, L. Maddalena, A. Petrosino, and Y. Wang, "Extensive benchmark and survey of modeling methods for scene background initialization," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5244–5256, Nov. 2017.

[79] M. Camplani, L. Maddalena, G. M. Alcover, A. Petrosino, and L. Salgado, "A benchmarking framework for background subtraction in RGBD videos," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, Dec. 2017, pp. 219–229.

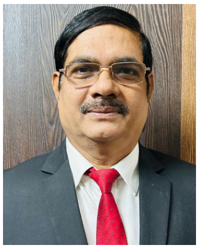
[80] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Comput. Vis. Image Understand.*, vol. 152, pp. 103–117, Nov. 2016.

[81] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 393–400.

[82] D. Zeng and M. Zhu, "Background subtraction using multiscale fully convolutional network," *IEEE Access*, vol. 6, pp. 16010–16021, 2018.

[83] N. Dorudian, S. Lauria, and S. Swift, "Moving object detection using adaptive blind update and RGB-D camera," *IEEE Sensors J.*, vol. 19, no. 18, pp. 8191–8201, Sep. 2019.

SUBHALUXMI SAHOO (Member, IEEE) received the B.Tech. degree in biomedical engineering and the M.Tech. degree in instrumentation engineering, in 2011. She is currently pursuing the Ph.D. degree in video object detection with Siksha 'O' Anusandhan, Deemed to be University, Bhubaneswar, Odisha, India. She has been an Assistant Professor with the Department of Communication Engineering, Siksha 'O' Anusandhan, Deemed to be University, since 2011. Her research interests include video object detection and biomedical image processing.



PRADIPTA KUMAR NANDA (Senior Member, IEEE) received the degree (Hons.) in electrical engineering from the VSS University of Technology (UCE), Burla, Odisha, in 1984, the master's degree in electronics systems and communication engineering from the National Institute of Technology, Rourkela, Odisha, in 1989, and the Ph.D. degree in computer vision from IIT Bombay, in 1996. From 1986 to 2007, he was a Faculty Member with NIT Rourkela, India, with the last position as a Professor and the Head of the Department. He was instrumental in establishing the Center of Excellence on Industrial Electronics and Robotics and served as the Principal Investigator. After serving NIT Rourkela, he joined Siksha 'O' Anusandhan, Deemed to be University, where he is currently discharging his duty as the Vice Chancellor. He was an Academic Visitor with the School of Computing, University of Leeds, U.K., in March 2006, and delivered lectures to their faculties and students. He has delivered several invited lectures at different international/national conferences, workshops/summer, and winter schools. He has published 98 papers in various journals and conference proceedings and has authored five research books and seven book chapters. His research interests include image processing and analysis, bio-medical image analysis, video-tracking, computer vision, soft computing and its applications, and ad-hoc wireless sensor networks. He is a fellow of IETE, India, a member of IET, U.K., and a Life Member of ISTE, India. He was the Chair of the IEEE GRSS Society, Kolkata Chapter.