Multimedia Tools and Applications (2024) 83:34417–34444
https://doi.org/10.1007/s11042-023-16982-7

Shot boundary detection in video using dual-stage optimized VGGNet based feature fusion and classification

Swati Chaitandas Hadke[1] · **Ravi Mishra**[1]

Received: 1 June 2022 / Revised: 6 June 2023 / Accepted: 11 September 2023 /
Published online: 26 September 2023

Abstract
Shot boundary detection (SBD) in video sequences is a key process in the analysis, retrieval, and summarization tasks of video content. The major goal of SBD is to detect transition and their boundaries among the subsequent shots by analyzing the spatial appearance and temporal motion information of the video. This paper proposed a deep learning-based intelligent SBD model, which can detect abrupt transition (AT) and gradual transition (GT) concurrently from video sequences. The proposed model follows the Dual Stage Fused Feature Extraction(DSFFE) process using an optimized VGGNet architecture. Initially, the input video data is converted into several frames, and then a pre-processing step is performed using the Improved Bilateral Filter (IBF). Then, Dual Stage Fused Feature Extraction is performed using VGGNet for extracting deep and spatial appearance-temporal motion features from video sequences. Further, a continuity matrix is created using the Inter-frame Euclidean Threshold (IET) to find the dissimilarity measure. Finally, shot transitions are classified using the Softmax Classifier, which categorizes AT and GT transitions as Fade in/out, cut, and dissolve. Especially the VGG network model weights are updated using an algorithm called Red Fox Optimization (RFO) to minimize the loss function. The proposed model is implemented using TRECVID and VideoSeg datasets on the MATLAB platform. The performance outcomes show that the proposed SBD model achieves an average recall, precision, and f1-score of 98.89%, 98.15%, and 98.86%, respectively, which is comparatively better than other models.

Keywords Short Boundary Detection · Abrupt transition · Gradual transition · Dual Stage Fused Feature Extraction · Inter-frame Euclidean Threshold

✉ Swati Chaitandas Hadke
swati.hadke@raisoni.net

Ravi Mishra
ravi.mishra@raisoni.net

[1] Electronics & Telecommunication Engineering Department, G. H. Raisoni University, Amravati, India

1 Introduction

Recently, social media and Internet platforms have been global, generating surplus video data every minute [1]. Video data is becoming a vital part of today's world due to the emergence of 5G technology. On the Internet, video is considered a heavily used data type and is used to access entertainment applications, security systems, web conferencing, etc. [2, 3]. The advancements in video processing have created enormous video repositories on storage mediums. Video data combines audio, images, and text, which constitute a huge volume of information and require more space for storage [4, 5]. Due to the rise in video availability, searching for a specific video from a large database seems very complex. The manual searching process requires considerable effort and time for the desired event retrieval. Using Content-Based Video Retrieval (CBVR), users can search and retrieve the relevant video that matches the given query [6, 7].

The CBVR procedure intends to automate the task of video indexing, retrieval, and management. The significant applications of CBVR are browsing news events, video folders, keyframe extraction and partitioning video events [8, 9]. The design of an effective CBVR model requires the process of Shot Boundary Detection (SBD) [10]. This SBD process segments the video data into several structural elements (SE), such as frames, shots, and scenes [11]. The other name of SBD is video temporal (VT) segmentation mainly focuses on performing video indexing and summarization process [12]. SBD can detect the shot transitions and their boundaries among successive shots, and the shots containing richer information are further used in CBVR [13].

In the video, a single image has termed a frame. The shot represents the series of interconnected successive frames grabbed using a single camera. The scene indicates multiple or single shots that define a story in a video. Shot boundaries are specified depending on the type of interruptions done among the camera operations [14]. The shot boundaries are categorized into different types as abrupt (hard) and gradual (soft) transitions based on the video length, editing effects, and properties [15]. Abrupt transition refers to the quick change in image content among neighbouring frames. It is one of the simple transitions and is otherwise called a hard or cut transition. The gradual transition type is the continuous and slow change in image content among various frames. The various editing effects in the gradual transition are fade (fade-in/fade-out), Wipes, Dissolves (fade-in + fade-out), and Mattes.

Due to advancements in machine learning [16, 17] and deep learning techniques, SBD-based video summarization has received greater attention. SBD based on deep learning has improved the accuracy and efficiency of video summarization compared to conventional methods. Deep learning can potentially improve detection accuracy, which would be extremely beneficial [18–20]. However, these learning strategies have shown promise in these domains for autonomously expressing the feature space using new data qualities to assist learning in high-dimensional feature space. Deep learning uses a vast amount of data to learn how features behave during training and predicts the class of data that hasn't been seen before. SBD is crucial for gaining a better understanding of different transitions, which in turn aids in the design of decisive measures and improves the detection procedure. Digital video processing (DVP) has become a remarkable area of investigation due to its enormous applications in data mining, video summarization, surveillance, etc. In video summarization, SBD is considered one of the most significant tasks. Multimedia video data are accessible now in huge volumes. Indexing and manual annotation of video data are quite ineffective. Henceforth, a powerful and automated video structure analysis is

substantial for executing SBD. The SBD process aims to identify the frames with significant discontinuities in different representations of their visual sequences. These different appearances of the visual sequence are represented by the characteristics of the frames in the input video. Therefore, extracting spatial appearance and time motion features in video processing is a major challenge. First, the existing SBD algorithms cannot efficiently grip the shot boundaries due to rapid changes in lighting. Secondly, when the video has slight lighting frames, these algorithms cannot achieve clear-cut boundary detection.

Moreover, false detection of shots is also possible when the motion of an object is high in the video. Every feature may realize one or more aspects that lead to certain false detection. Removing this false detection with invariant features is a complex task. Hence, employing a single feature descriptor to describe the complete video will not be adequate or provide a precise result. Thus, the proposed model employing deep learning-based dual-stage fused feature extraction for obtaining invariant spatial–temporal features assists in attaining precise detection. The major contributions of the paper are as follows:

- To present an optimized VGGNet-based dual-stage fused feature extraction and classification for shot boundary detection in videos.
- To eliminate the non-boundary frames, Inter-frame Euclidean Threshold is employed to find out the feature similarity between the adjacent frames.
- To minimize the false detection of shot transitions, the VGGNet model parameters are optimized by the RFO algorithm through the backpropagation process.
- Fused Feature extraction and classification within a single network minimizes the network complexity and offers effectual outcomes.

The remaining structure of the paper is discussed as follows: Section 2 lists the related work of some research. Section 3 describes the proposed methodology in detail. Subsequently, the implementation results of the proposed model are explained in Section 4. Finally, Section 5 highlights the conclusion and future work.

2 Related work

In the field of video processing, the need for image boundary detection has increased significantly in recent times. Many of the authors have examined different models for shot boundary detection. This section reviews some of the more recent models examined in the literature.

Soucek, and Jakub [21] developed an effective TransNet V2 architecture based on the deep network for performing fast SBD. The TransNet V2 model includes 6 DDCNN (Dilated Deep CNN) cells integrated with the BN (Batch Normalization) function to stabilize the gradients. The similarity evaluation was processed using learnable features and RGB color histograms. The performance of the TransNet V2 model was executed on the three datasets such as BBC plant earth, Clip-shots and RAI datasets. The major drawback observed with the largest Clip-shot data collection as it was very challenging to perform manual testing of ground truth, unannotated video portions, and incorrectly labelled frames.

Rashmi and Nagendraswamy [22] presented a video SBD using a cumulative block-based technique called MCSH (Mean Cumulative Sum Histogram). Local and global features were extracted from the fuzzified Sobel gradient edge frame. The statistical metric

utilized to identify the gradual and abrupt transitions was RSD (Relative Standard Deviation). The datasets used for experimentation were VideoSeg, and TRECVID (2001, 2007). The major challenge was higher false detection when identifying gradual and abrupt transitions.

Sasithradevi and Mohamed [23] introduced a new approach POCS (Pyramidal Opponent Colour-Shape), for video SBD. The transitions, such as gradual and abrupt, were detected with reduced complexity and improved accuracy. The variation among the features was constructed using TCS (Temporal Continuity Signal). The BTC (Bagged Trees Classifier) classifier was utilized to select SS (Suitable Segment) through parallel processing. The datasets used for analysis were VideoSeg 2004, TRECVID 2001, and 2004. The accurateness of POCS was affected when the dataset included videos comprising flashlights, flickering effects and fire.

Zhou et al. [24] presented video SBD by collaborating the multi-level features. The presented approach uses the top-down zoom rule, local descriptors, image color features and extraction algorithm based on motion area for attaining SBD. The color histogram was used to select the segments of candidate transitions to speed up the robust features. Cut transition detection was performed using the combination of pixel difference, color histogram and uneven matching of slices. Next, a gradual type of transition detection was done by extracting motion area, SIFT (Scale-Invariant Feature Transform), and matching even slices.

Chakraborty and Dalton [25] developed a detection model SBD-Duo considering the effects of motion and illumination for effective detection of video SBD. The luminance and gradient feature's similarity were evaluated using a quality SM (Similarity Measure) to detect the possible number of transition frames. The lab features were integrated with the adaptive threshold to identify the transitions. The performance was measured with an F1 score, and the experimentation was done on the TERCVID 2001 and 2007 datasets. Some limitations affecting the detection performance were camera obstruction and effects caused by non-uniform illumination.

The deep neural network-based SBD model was proposed by Benoughidene and Titouna [26]. The proposed model was a pre-trained network merged with long short-term memory (LSTM) network and Euclidean distance measure. The spatial features were extracted by employing a dual pre-trained network in a parallel manner by using similar weights. After that, extracted features were inputted to the LSTM and the Euclidean distance measure to categorize the video frame sequence into relevant classes. Especially the segment selection procedure was adopted to detect the shot boundaries. The proposed model was implemented on the TERCVID 2001 and 2007 datasets and achieved improved performance in terms of F1 score over the existing comparative models.

Li et al. [27] presented a shot boundary detection algorithm using global and target features. Initially, the RGB colour histogram features were extracted from video frame sequences. Then, the Gaussian Mixture Model (GMM) was employed to detect the foreground of the object in the video frames; further, the foreground targets were extracted using scale-invariant features transformation (SIFT). At last, the fusion of global and target features was achieved through weights, and the difference between adjoining frames was also computed to make a pattern distance map for detecting the transitions.

Kar and Kanungo [28] proposed an automatic dual detection based shot boundary detection model using a gradient feature. In the initial stage, abrupt transition (AT) detection was recognized in the attendance of illumination variation and motion in the shots of video frames. For this purpose, a joint feature representation was created by combining the histogram of gradient magnitude and orientation features of the video frame. In the second

stage, gradual transition detection (GT) was addressed only on the frames within two AT frames fulfilling the particular frame distance measures. The proposed algorithm achieved good performance in terms of F1 score for both AT and GT detection. The Systematic review shot boundary detection techniques are presented in Table 1.

Based on the above literature, it is determined that the existing models have several problems associated with SBD, including dim lighting frames, illumination effect, outliers, object motion, and camera operation. These effects limit the detection of transitions and are critical to the performance of SBD. Moreover, the issues of SBD involve threshold-free algorithms and effective feature descriptors to attain a higher detection rate in identifying any shot transition. The gradual transition is more complex than the abrupt transition because the gradual type includes certain variations among successive frames and continuous multi-frames. Most SBD techniques rely on handmade features greatly, and the major limitation is causing high false positives and miss detection during the shot transition detection. So, an effective SBD in video data accurately detects the different types of transitions.

3 Proposed methodology

This section provides a detailed description of the proposed SBD model using two-stage fusion feature extraction (DSFFE) with optimized VGGNet. The proposed model aims to develop an accurate SBD leading to the production of a compact and short video that eliminates the erroneous shot detections that occur mainly due to camera operation, lighting, object noise and movement. The proposed SBD model determines shot boundaries in a hierarchical manner and accordingly identifies abrupt and gradual shot boundaries. The schematic diagram of the proposed model is shown in Fig. 1.

The input data is initially converted into various frames in the proposed model. A method that permits noise (artifacts) removal, edge highlighting, and contrast enhancement is applied secondly using an Improved Bilateral Filter (IBF). Further, a Dual Stage Fused Feature Extraction using VGGNet (DSFFE-VGGNet) is carried out to extract spatial–temporal and deep features. DSFFE involves fused feature extraction of Abrupt (hard) shot and Gradual (soft) shot transitions. At first, deep feature extraction is performed on the individual frame with the first stage VGGNet.

In contrast, spatio-temporal (motion and appearance) features are extracted from the second stage VGGNet. Both extracted features are fused in the FC8 layer of VGGNet. Next, a continuity (mapping) matrix is constructed using Interframe Euclidean Threshold (IET) to find the dissimilarity measure. Finally, the classification of shot transitions is performed using the Softmax layer. The Softmax classifier categorizes shot transitions as Hard (Cut) and Soft (Fade in/out, dissolve). The weights of the VGG network model are updated using an algorithm called Red Fox Optimization (RFO) to minimize the false detection of shot transitions.

3.1 Pre-processing

The pre-processing in video SBD refers to improving significant information and filtering outliners in digital video to enhance the overall system's performance. Generally, the pre-processing video stage encompasses converting input video $V(u)$, within the same dimensional space α. The proposed model has employed IBF as a pre-processing video signal that consents

Table 1 Comparison of shot boundary detection techniques

Author name	Technique	Colour Space	Transition Type	Database	Average F1-Score
Soucek, and Jakub [21]	DDCNN	RGB	Cut and dissolve	TRECVID and ClipShots	96.2%
Rashmi and Nagendraswamy [22]	MCSH and RSD	Grayscale	Fade-in, Fade-out and Dissolve	TRECVID and VideoSeg	94.45%
Sasithradevi and Mohamed [23]	POCS	RGB	Abrupt and gradual	TRECVID and VideoSeg	93.67%
Zhou et al. [24]	Multi-level features collaboration	RGB and Grayscale	Cut and gradual	TRECVID	85.1%
Chakraborty and Dalton [25]	Adaptive Thresholding	Grayscale	Abrupt and gradual	TRECVID	88.35%
Benoughidene and Titouna [26]	CNN-LSTM	Grayscale	Cut transition	TRECVID and RAI	95.62%
Li et al. [27]	GMM and SIFT	RGB	Cut and gradual	RAI	91.58%
Kar and Kanungo [28]	Gradient based dual detection	Grayscale	Abrupt and gradual	TRECVID and VideoSeg	90.01%

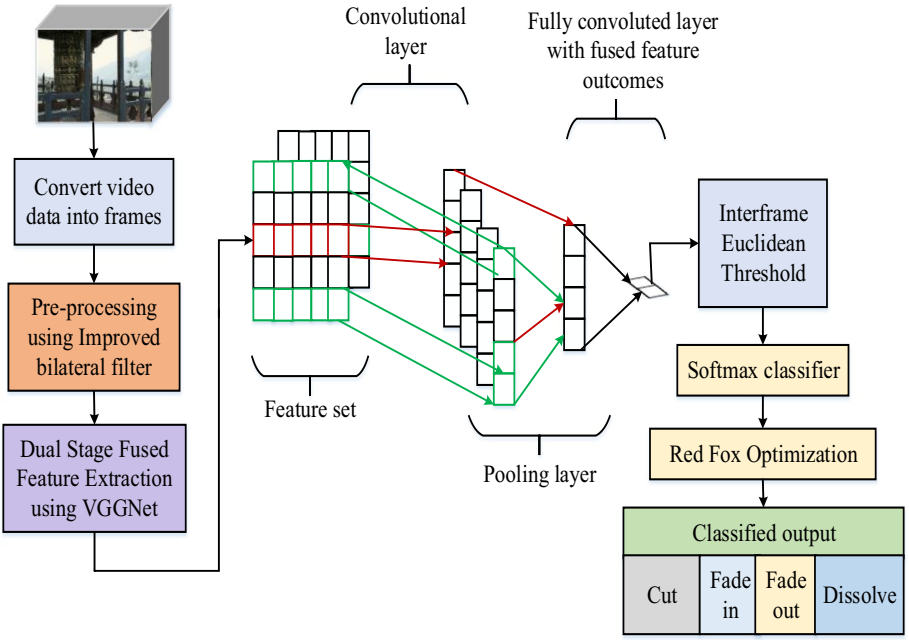

Fig. 1 Block diagram of the proposed model

to signal-to-noise (SNR) enhancement, edge highlighting, and contrast enhancement in each input frame. IBF is that of unsharp masking and bilateral filtering. It improves the significant image features and excludes non-relevant information to lift the overall performance of a system. The IBF is stated as follows:

$$\widehat{g}_{um} = h + \gamma a \tag{1}$$

$$a = m^{*}h \tag{2}$$

$$m = \begin{bmatrix} -1 & -1 & -1 \\ - & 18 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{3}$$

$$\widehat{g}_{dbl}(q) = \left(\sum_{n}^{X} i(q,n) \right)^{-1} \left(\sum_{n}^{X} i(q,n)\widehat{g}_{um}(n) \right) \tag{4}$$

$$i(q_o,n) = exp\left(-\frac{(n-q_0)^2}{2v_t^2} \right) exp\left(-\frac{(h(n)-h(q_0))^2}{2v_s^2} \right) \tag{5}$$

where, $\widehat{g}_{um}$ signifies the unsharpened masking for de-blurring an image, a indicates the high-pass filtered image, normally depth with a Laplacian filter m, h represents the blurred image and γ resembles the gain coefficient. Moreover, Eq. (4) $i(q,n)$ relates to the bilateral

filter kernel provided by Eq. (5), where the pixel n iterates over a window X. The range deviation is represented as υ_s, and the spatial deviation is indicated by υ_t, respectively.

3.2 Dual stage fused feature extraction

Considering frame size imbalance and the distinctive altering character of abrupt shots in the contradiction of gradual shots, a feature fusion module has been constructed initially to detect the abrupt shot change and further differentiate the others. In this proposed model, a Dual Stage Fused Feature Extraction (DSFFE) using VGGNet is carried out to extract spatial–temporal and deep features. DSFFE covers fused feature extraction of gradient (soft) shot and abrupt (hard) shot transitions. Deep feature extraction is initially carried out on a distinct frame with the initial stage VGGNet. Still, the extraction of Spatio-temporal information (appearance and motion) features from the second stage VGGNet. Subsequently, both deep and spatial–temporal features are fused.

Recently, the methods based on convolutional networks (ConvNets) [29] have gained feasible outcomes over classical hand-crafted features. Motion and appearance are the two major sources of information in the video. The proposed model contains three major streams: a temporal stream to capture the motion, a spatial stream to capture appearance, and a Spatio-temporal stream to simultaneously capture both motion and appearance information. The proposed VGG Model holds 16 convolution layers, three fully connected layers, 5 MaxPool layers, and 1 SoftMax layer. The convolution layers perform the basic processes of extracting spatial–temporal features from the video frames by employing successive filters. Especially, the number of convolutional layers is limited by the spatial size of the input samples, and the window is empirically set to be 7×7 in this study. The max-pooling layer is used to decrease the feature map dimensionality and works on each feature map individually. The fully connected layers are used to ensure the ability to fit the features of video frames, and finally SoftMax layer act as the classifier.

Figure 2 presents the block diagram of the feature extraction using DSFFE-VGGNet. In Fig. 2, the proposed model independently extracts the features containing spatial information for all three networks for a given input video. The image with 224×224 resolution and three channels are provided as an input of VGGNet for colour information.

After extracting the particular frames from a video, the extracted frames are resized consequently to the required input size of the network. The output of the last convolutional layer, along with spatial information, pool5, is taken for each frame. The local deep features can be extracted by taking the layer with spatial information for each video frame, along with the information holding almost the spatial membership of the features. The feature map with 512 channels and spatial size of 7×7 is signified as the output of pool5. Besides, to extract the local deep features from the feature map, the proposed model uses each spatial location separately and concatenates the values amongst all 512 channels, attaining local features with 512 dimensions. Hence, $7 \times 7 = 49$ local deep features are obtained from each frame, and then each feature is about a 512-dimensional vector. Thus, for each video, totally $\#\mathrm{frames} \times 49$ local deep features are obtained. The features extracted with a spatial convolutional network are signified as SCN.

The input to the temporal ConvNet contains 10-stacked optical flow fields with a single image for horizontal and vertical motion. So, 20-stacked optical flow images are a single input toward the network. The OpenCV implementation of the TVL1 algorithm [30] is utilized for extracting the optical flow fields. The outcomes of the last convolutional layer and pool5 (structural information) have also been taken for temporal ConvNet. The pool5 layer

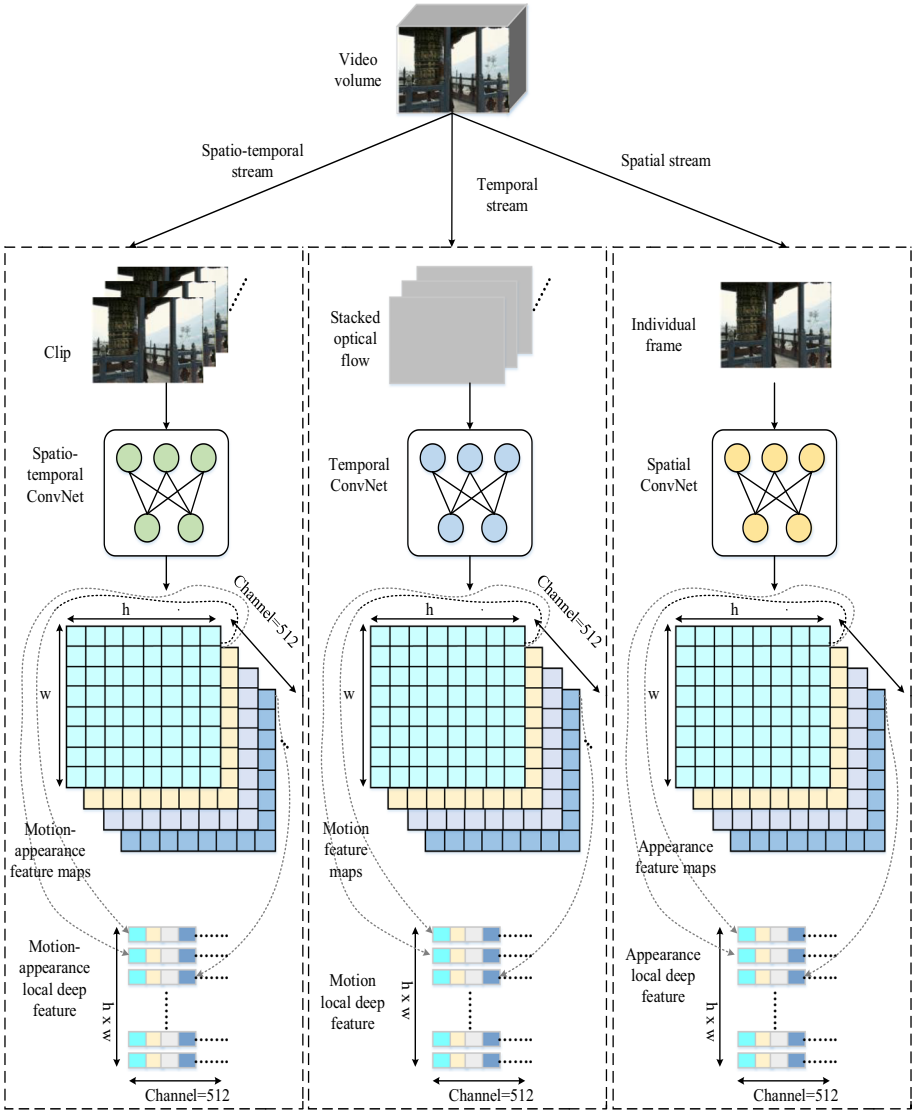

Fig. 2 Dual Stage Fused Feature Extraction using VGGNet

contains 512 channels and the spatial size of 7×7 feature maps. For input, the final local deep features have been gained by concatenating the values from all spatial locations with each channel, ensuring 49 local features. This outcome in $49 \times (\#\text{frames} - 9)$ local deep features by utilizing temporal ConvNet for a video. Besides, the features extracted with the temporal convolutional network are signified as TCN.

For a spatial–temporal stream, the proposed model used 3D ConvNet and consisted of 16 layers. A 3D convolutional kernel is employed in the network design to capture motion and appearance. Then, a 16 frame-long clip extracted from a video is considered an input of the network. Relating to the preceding two networks utilized in the proposed model, the

Spatiotemporal stream has utilized a step size of a single frame for iterating over video frames to make input clips. Besides, the final layer of the network with spatial information only contains the feature maps' size 4×4. The conv5b layer partakes an identical spatial feature maps size as the preceding two networks, which means 7×7 an identical number of channels 512. Although the conv5 layer has two feature maps, each of them encompasses $7 \times 7 \times 512$. In DSFFE-VGGNet, only a single feature map of $7 \times 7 \times 512$ was constructed for input in this network, utilizing the least value from conv5b for each location of both feature maps.

Further, the local deep features that are the same as the preceding networks are extracted. For an input video, the total number of local deep features in a 3D network is nearly $49 \times (\#\text{frames} - 9)$. All the resulting local deep features are signified as a vector along with dimensions of 512.

3.3 Inter-frame Euclidean Threshold

After extracting the features, a mapping (continuity) matrix is generated using the Inter-frame Euclidean Threshold (IET) model for finding the dissimilarity measure. Initially, the proposed model has adopted a suitable distance measurement. The investigation in related fields has stated that Euclidean distance or Manhattan metric is greatly effective. Hence, the proposed model has used the Euclidean distance and initially evaluates the frame difference. The histogram difference of resultant sub-blocks of neighbouring matrices for colour features is given as follows.

$$Sub - block\ difference_{lj} = Euclidian(block(Y_{lj}), block(I_{(l+1)j}) \tag{6}$$

Where, $Sub - \text{block difference}_{lj}$ indicates the histogram Interframe difference of j^{th} the sub-block of the frame l, and $\text{block}(I_{lj})$ signifies the colour histogram information of j^{th} the sub-block of the frame l. So, the histogram difference $E_I(l, l+1)$ between the frame l and frame $l+1$ is provided as follows.

$$E_I(l, l+1) = \frac{\sum_{j=1}^{9} sub - \text{block difference}_{lj} \times X_j}{\sum_{j=1}^{9} X_j} \tag{7}$$

where, X_j indicates the weight of j^{th} the sub-block. For deep features, when D_l utilized to resemble the feature of the frame l, then $E_D(l, l+1)$ is stated as:

$$E_D(l, l+1) = \text{Euclidean}(E_l, E_{l+1}) \tag{8}$$

Further, the weighted similarity between frame l as well as frame $l+1$ is expressed as:

$$E(l, l+1) = \varepsilon E_I(l, l+1) + (1 - \varepsilon) E_D(l, l+1) \tag{9}$$

Where, ε is set to 0.4.

The Interframe difference in abrupt shot variations is frequently higher than in the non-shot position. Nevertheless, the Interframe difference between several videos can be changed to the video content. So, the artificial thresholds cannot be entirely appropriate for all videos. Thus, the thresholds have been attained to the mean of Interframe difference in a single shot. Subsequently, the total Interframe difference in the recent shot is expressed as:

$$\mathrm{TIFD}_l = \sum_{j=1}^{l-1} E(j, j+1) \tag{10}$$

Where, TIFD_l signifies the total frame difference.

Further, the adaptive threshold U in a shot is expressed as:

$$U = \varphi \frac{\mathrm{TIFD}_l}{l-1} \tag{11}$$

Where, φ is in the range of 5 to 10.

3.4 Classification of shot transitions

Several existing methods for gradual SBD prominently depend on hand-crafted features. The low-level visual features are not capable of defining high-level semantics entirely. Many authors have preferred to utilize the deep neural network (DNN) to extract the features since the deep features may resemble the data-conferring task. Even though attaining better outcomes, the researchers have not been able to provide a deeper theoretical understanding of the field. The shot transition features must be extracted appropriately at times, which makes the performance reliant on more extracted temporal features in the SBD process. So the proposed model has used the VGG 19 network to extract deep and spatial–temporal features for categorizing the shots automatically.

Compared to 3D CNN, the VGG 19 network splits the video into frames and applies the convolutional kernel to both spatial and temporal domains. The combination of both spatial as well as temporal features assists in encouraging feature description capability in SBD. Subsequently, the classification of shot transitions is performed using the Softmax layer. The Softmax classifier categorizes shot transitions as Hard (Cut) and Soft (Fade in/out, dissolve).

Consider $(y_1, z_1), \ldots\ldots, (y_p, z_p)$ being the set of p labeled data connected to one of m the classes $D_1, D_2, \ldots D_m$, where $y_j \in S^Q$ and $z_j \in S^l$ are one-hot encoded vectors such that $z_{jl} = 1$ if $y_j \in D_l$. The y_j's can be the input of l_2—regularized classifier together with regularization parameters $(\omega_l)_{l\in[m]} \in S^+$ as well as class-wise weight vectors $x_1, x_2, \ldots\ldots, x_m \in S^q$ set to reduce the loss as follows:

$$L(x_1, x_2, \ldots\ldots, x_m) = -\frac{1}{p}\sum_{j=1}^{p}\sum_{l=1}^{m} z_{jl} \log q_{jl} + \frac{1}{2}\sum_{l=1}^{l} \omega_l \| x_l \|^2 \tag{12}$$

With $q_{jl} = \frac{\beta(x_l^U y_j)}{\sum_{k=1}^{m}\beta(x_k^U y_j)}$ for real-valued function $\beta : S \to S$. Especially, $\beta(u) = e^u$ for the Softmax classifier. Revoking the gradient of loss relating to each weight vector x_l yields, for each $l \in [m]$,

$$\omega_m x_m = -\frac{1}{p}\sum_{j=1}^{p}\left(z_{jl}\tau(x_l^U y_j) - \frac{\beta(x_l^U y_j)}{\sum_{k=1}^{m}\beta(x_k^U y_j)}\sum_{k=1}^{l} z_{jk}\tau(x_k^U y_j)\right) y_j \tag{13}$$

Where, $\tau \equiv \frac{\beta\prime}{\beta}$. In proper statistical postulation on data matrix $Y \equiv [y_1, y_2, y_3, \ldots, y_p] \in N_{q,p}$, and supposing q, p are large, it consequently displays that the stacked vector $X \equiv [x_1^U, x_2^U, \ldots, x_m^T]^U \in S^{qm}$. This vector contains a discriminative representations, which are assists to predict the outcomes of the Softmax classifier correctly.

3.5 Red Fox Optimization

The classification is considered the most significant part of SBD to classify the types of shot transitions as hard (cut) and soft (fade in/out, dissolve). After performing the feature extraction, the extracted features are fed into the classifier layer to classify the type of shot transitions. As mentioned above, VGGNet utilizes a backpropagation model for learning. The proposed model employs a Red Forest optimization (RFO) algorithm to properly select weight in the VGG 19 network by reducing the mean squared error. RFO is a novel metaheuristic optimization algorithm that mimics the hunting lifestyle among red foxes. During hunting, the red foxes hide behind the bushes and slowly come near the prey, and then the prey is unpredictably attacked. In contrast with other metaheuristic algorithms, RFO also embraces exploitation and exploration phases.

The exploration phase is defined by selecting fox prey at positions far from the prey. In contrast, the exploitation phase is signified depending on the closeness of the fox to the prey at any time probable to attack the prey. In the proposed model, the red foxes are considered a set of weights for the neural network. The initialization of RFO is expressed through random individual generation as:

$$Y = \left(y_0, y_1, y_2, \ldots\ldots, y_{p-1}\right) \tag{14}$$

Where, j resembles the number of population, $\left(Y_k^j\right)^u$ indicates the y_j in iteration u and k represents the dimension in searching space. The fitness function is considered as,

$$F = Min(\text{loss function}) \tag{15}$$

Considering g as a condition function in S^p, where p signifies for parameters of range $[b, c]^p$ then,

$$Y^j = \left[\left(y_0\right)^j, \left(y_1\right)^j \left(y_2\right)^j, \ldots\ldots, \left(y_{p-1}\right)^j\right] \tag{16}$$

Where, $b, c \in S$.

Thereby, the optimal solution has been attained while $g\left((Y)^j\right)$ recommending the global optimum. Each individual is presumed as a certain assignment to support the crew in exploration. Besides, if the region has not contained sufficient prey, the individuals can move to another area to obtain a better chance for prey. If the region with sufficient prey is determined, the location of the prey has been shared with others. So, the individuals are adjusted depending on the cost value. In this concern, the Squared Euclidean distance is practised as follows:

$$E\left((Y)^j\right)^u, \left(Y_{\text{Best}}\right)^u = \sqrt{\left((Y)^j\right)^u - \left(Y_{\text{Best}}\right)^u} \tag{17}$$

So, each candidate travels through the optimum solution is expressed as follows:

$$\left((Y)^j\right)^u = \left((Y)^j\right)^u + \varepsilon \times sgn\left(\left(Y_{\text{Best}}\right)^u - \left((Y)^j\right)^u\right) \tag{18}$$

Where, ε resembles the random value.

The candidate's new location can suggest a suitable solution; if not, the previous location has been retained. When the red foxes find their prey, it comes close to the prey.

It was specified as exploiting RFO, which is by considering the random value s of the range [0, 1].

$$\begin{cases} Stay\ and\ hide\ \ if\ s\ \leq \frac{3}{4}, \\ Move\ closer\ \ \ if\ s\ > \frac{3}{4}. \end{cases} \tag{19}$$

At that moment, the member's motion can be discovered using an enhanced cochleoid formula. The succeeding term has been adapted through a variable, specified as radius, which depends on two primary variables, namely, β_o and$b$. β_o indicates the value of range $[0, 2\pi]$ that specifies the observation angle of foxes and b represents the random number of ranges [0, 0.2], respectively. Statistically, this term is expressed as:

$$s = \begin{cases} b\frac{\mathrm{Sin}(\beta_0)}{\beta_0}, & \text{if } \beta_0 \neq 0, \\ \delta, & \text{if } \beta_0 = 0 \end{cases} \tag{20}$$

Where, δ indicates the random value between 0 and 1.

Then the fox population nearing the prey is mathematically expressed as follows:

$$\begin{cases} y_0^{New} = b \times s \times Cos(\beta_1) + y_0^{Actual} \\ y_1^{New} = b \times s \times Sin(\beta_1) + b \times s \times Cos(\beta_2) + y_1^{Actual} \\ y_2^{New} = b \times s \times Sin(\beta_1) + b \times s \times Sin(\beta_2) + b \times s \times Cos(\beta_3) + y_2^{Actual} \\ \vdots \\ y_{p-1}^{New} = b \times s \times \sum_{l=1}^{p-1} Sin(\beta_1) + b \times r \times Cos(\beta_{p-1}) + y_{P-2}^{Actual} \\ y_{p-1}^{New} = b \times s \times Sin(\beta_1) + b \times s \times Sin(\beta_2) + \cdots + b \times s \times Sin(\beta_{p-1}) + b \times s \times Sin(\beta_{p-1}) y_{p-b}^{Actual} \end{cases} \tag{21}$$

Nearly 5% of the worst members in the created population have been disregarded, and additional members were added to the individuals to provide a fixed-size population. Similarly, two optimum members have attained as $(Y(1))^u$ and $(Y(2))^u$ as an alpha couple in the iteration u. Further, the center of the territory has attained as follows:

$$I_d^u = \frac{1}{2}(Y(1))^u - (Y(2))^u \tag{22}$$

Whereas, the diameter of the territory by Euclidean distance is expressed as:

$$I_e^u = \sqrt{(Y(1))^u - (Y(2))^t} \tag{23}$$

In this process, a random number υ is selected between the value 0 and 1.

$$\begin{matrix} Alpha\ couple\ reproduction, & if\ \upsilon \leq 0.45 \\ New\ nomadic\ candidate & if\ \upsilon > 0.45 \end{matrix} \tag{24}$$

In the search area, the random location is captured after this new member is set up by the alpha pair as follows:

$$\left(Y^{Rep}\right)^u = \frac{\upsilon}{2}(Y(1))^u - (Y(2))^u \tag{25}$$

For RFO, the utilized parameters are as follows: $\beta_0 = 1$ and $b = 0.2$. Then the Pseudocode for the proposed algorithm is provided below.

Input: *Input video stream*
Output: *Cut, fade in, fade-out and dissolve transitions*
procedure shot boundary detection
  for every dataset video streams **do**
    Convert videos into frames
    Employ pre-processing for contrast enhancement
    Apply Dual Stage Fused Feature Extraction using VGG ConvNet
      Inter-frame Euclidean Threshold calculation
    Apply softmax classifier for shot transition detection
  End for
  Backpropagation using Red Fox Optimization
    Initialize the red fox population within the search space
  Specify coefficients for iteration b and β_0
  Calculate the fitness function
    Sort individuals based on the fitness
  Choose y_{best}^{j}
    Determine the rearrangement of individuals
    If rearrangement is better than the preceding position
    Move the fox,
    Else
      Fox remains at his location
    End if
    Sort the population based on fitness
      Worst fox killed by hunters or leave the heard
    New fox has switched in the population to a nomadic fox
  End
  Return the fittest fox y_{best}
  End
End procedure

Algorithm 1: Proposed shot boundary detection algorithm

4 Result and discussions

This section will discuss the implemented results of the proposed model for SBD. The proposed method is simulated using the Matlab platform, and the outcomes are compared with recent approaches to analyze the efficiency of detection. Performance metrics such as precision, recall, and f-measure are employed to evaluate the performance. The video shot boundary includes various transitions such as cut, fade-in, fade-out, dissolve, etc. The proposed model detects dissolve, cut and gradual (Fade in and fade out) transitions from the shot boundary of the video. The following sub-sections discuss the proposed model's dataset description, performance metrics, and performance evaluation.

4.1 Dataset descriptions

The proposed model's datasets employed for SBD include TREC Video Retrieval Evaluation (TRECVID) 2018, 2019, 2020, 2021, and Videoseg datasets. All experiments are performed on an Intel(R) Core i7 CPU @ 2.70 GHz with Windows 10.1 environment.

TRECVID datasets are employed to determine the proposed network and are obtained from http://trecvid.nist.gov/. Every available MPEG video in the dataset has been segmented manually by detecting the shot boundaries. The TRECVID 2018 dataset encompasses 4593 videos of Internet archives with a duration between 6.5 min and 9.5 min and a mean duration of 7.8 min. In contrast, the TRECVID 2019 dataset contains 7475 Vimeo videos with an average duration of 8 h. TRECVID 2019 has been segmented into 1 million shots of a short video. In addition, the TRECVID 2020 dataset also comprises 7475 videos with a mean duration of 8 min. The TRECVID 2021 dataset includes 28,450 videos with a total duration of 3,801 h. Further, a VideoSeg dataset [31] encompassing 10 different videos and different resolutions and quality is employed for the experimental analysis.

4.2 Performance metrics

Some performance metrics, such as precision, recall, and f-measure, are considered to evaluate the proposed model's performance. These performance metrics display the efficiency of the proposed model while comparing it with existing approaches.

Recall The recall is determined by correctly predicted positive observations to the total number of observations. It is expressed as follows:

$$R = \frac{D}{D+N} \tag{26}$$

Where, R indicates recall, D represents the correctly detected shot boundary, N signifies the non-detected boundary.

Precision Precision determines the number of correctly reported transitions to the total number of transitions.

$$P = \frac{D}{D+G} \tag{27}$$

Where, P indicates precision, G represents the incorrectly detected shot boundary.

F-Measure F-measure is signified as a weighted average between precision and recall. It is given as follows:

$$FM = 2^{*}\frac{(P \times R)}{P+R} \tag{28}$$

4.3 Performance analysis

In this section, the simulation outcomes of the proposed model are evaluated for TRECVID and VideoSeg Dataset using a few performance metrics and related the proposed model with recent existing methods to display the efficacy of the proposed model. At first, the proposed model is compared to deep learning architectures like DBN, Deep Convolutional Neural Network (DCNN), Recurrent Neural Network (RNN), Deep Neural Network (DNN), and Convolution Neural Network (CNN) using TRECVID Dataset. Further, the existing methods, such as Mean Cumulative Sum Histogram with Relative Standard

Deviation (MCSH-RSD) and POCS model, are compared with the Videoseg dataset. The proposed model detects transitions such as cut, fade in, fade out, and dissolve. Thus, the brief outcomes of transitions are shown in the resulting sub-sections.

4.3.1 Evaluation of performance metrics using the TRECVID dataset:

Gradual transition The simulation outcomes of the proposed model have considered fade in and fade out from the gradual transition. The proposed model effectively detects the gradual transition, and the investigation of the resultant performance is stated as follows.

Fade in transition:

The metrics such as recall, precision, and f-measure are estimated from the video sequence in the TRECVID 2018, 2019, 2020 and 2021 datasets for detecting fade-in transition. The proposed model is compared with a recent existing model, such as CNN, DNN, RNN, DBN, and DCNN, regarding the recall, precision, f-measure and ROC. The graphical representation is displayed to show the superiority of the proposed model.

Figure 3 signifies the fade in the recall, precision, and accuracy results from TRECVID 2018, 2019, 2020 and 2021 datasets. The fade-in transition is the sub-category of gradual transition. The video sequence contains numerous faceplates under various images. The proposed model can efficiently identify the fade in frames from several frames.

Figure 3(i) illustrates the analysis of recall determined by existing models and the proposed model. The proposed model has obtained a recall of 94.5%, 95%, 94.5% and 92% compared to existing models such as CNN, DNN, RNN, DBN, and DCNN for TRECVID 2018, 2019, 2020 and 2021 datasets. Since the proposed method detected the fade-in transition with less false rate, the existing models obtained less recall than the proposed model due to more false rates. Thus, it is observed that the proposed model can differentiate the false rates from the input video and achieve a higher recall value for the fade-in transition.

Figure 3(ii) describes the precision obtained by the proposed and existing models for fade-in transition. The graph shows that the proposed model has achieved a higher precision than existing CNN, DNN, RNN, DBN, and DCNN models. The proposed model has obtained 96.04%, 92.5%, 92.74%, and 89.12% precision using TRECVID 2018, 2019, 2020 and 2021 datasets for fade-in transition detection. The model's performance is normally considered more efficient if the precision value is higher. In this evaluation, the proposed model achieves a higher precision value. The figure proves that the proposed model outperformed the true positive more than the existing models.

Figure 3(iii) resembles the f-measure value obtained by the proposed and existing models for the fade-in transition. The proposed model can accurately detect the classes for the given input video and find the true positive. The graph shows that the proposed model has accomplished an extreme f-measure value over existing models. The f-measure value obtained by the proposed model is 92.5%, 92.92%, 91.61% and 91.01% for fade-in transition detection using TRECVID 2018, 2019, 2020 and 2021 datasets, whereas the existing models have secured less f-measure value. Besides, the efficacy of the proposed model can be simply determined through the f-measure. If the value of the f-measure is more, then the model can accomplish better transition detection. Overall, the proposed model has achieved a higher f-measure value than existing models from the above analysis.

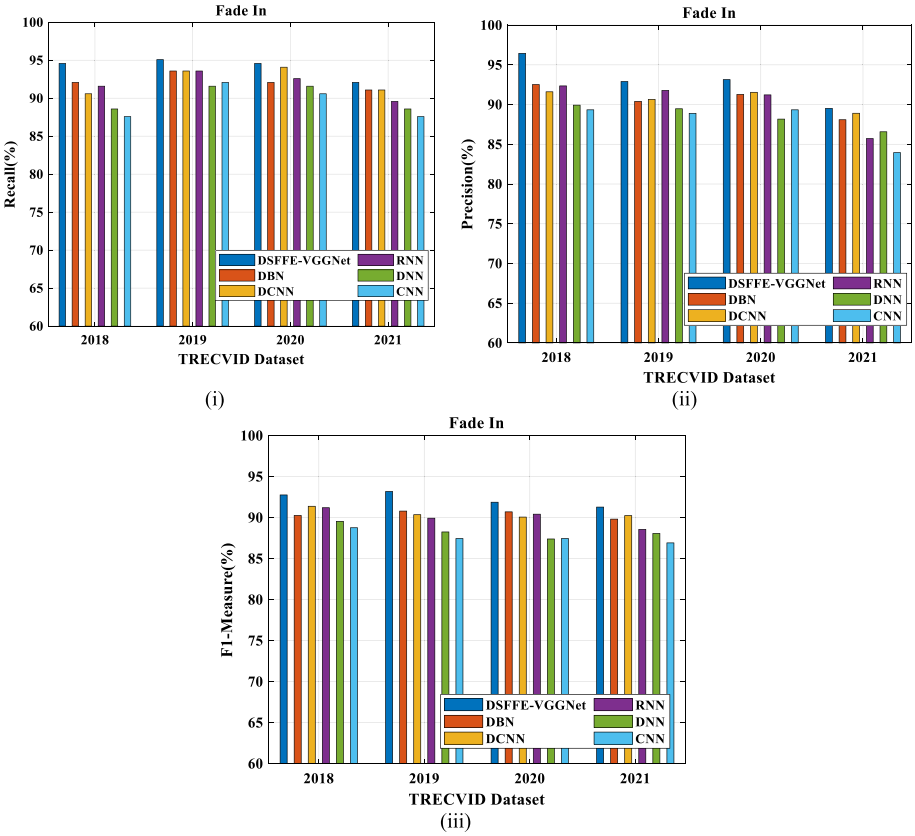

Fig. 3 Experimental outcomes of proposed and existing model for fade-in transition using TRECVID datasets (**i**) Recall (**ii**) Precision (**iii**) F-measure

Table 2 illustrates the experimental results of the proposed and existing model for the fade-in transition. The existing models, such as CNN, DNN, RNN, DBN, and DCNN, are employed for comparison. The proposed model has gained better results than the existing model for detecting a fade-in transition in the above table. The proposed method has enriched the ability of the system to detect class labels.

Fade out transition:

This part evaluates the performance of recall, precision and f-measure gained by proposed and existing models for fade-out transition using TRECVID 2018, 2019, 2020 and 2021 datasets. To measure the effectiveness of the proposed model, a comparative analysis of proposed and existing models is carried out.

The performance measure of the proposed and existing model for fade out images using TRECVID 2018, 2019, 2020 and 2021 datasets are depicted in Fig. 4. Like fade-in transition, fade-out transition is also a sub-category of gradual transition. There are

Table 2 Comparative analysis for fade-in transition

Metrics	TRECVID dataset	Methods					
		DSFFE-VGGNet (Proposed)	DBN	DCNN	RNN	DNN	CNN
Recall	2018	94.5	92	90.5	91.5	88.5	87.5
	2019	95	93.5	93.5	93.5	91.5	92
	2020	94.5	92	94	92.5	91.5	90.5
	2021	92	91	91	89.5	88.5	87.5
Precision	2018	96.04	92.11	91.2	91.94	89.52	88.93
	2019	92.5	90	90.25	91.38	89.06	88.5
	2020	92.74	90.87	91.14	90.81	87.77	88.93
	2021	89.12	87.68	88.5	85.32	86.17	83.55
F-measure	2018	92.5	90	91.11	90.93	89.28	88.5
	2019	92.92	90.53	90.08	89.66	87.98	87.18
	2020	91.61	90.43	89.8	90.15	87.13	87.18
	2021	91.01	89.55	89.97	88.3	87.8	86.65

multiple fade-out frames in a video sequence among various frames. However, the proposed model can efficiently discover several frames' fade-out frames.

Figure 4(i) presents the performance analysis of the recall with the proposed and existing models. The recall value gained by the proposed model is higher than existing models such as CNN, DNN, RNN, DBN, and DCNN. The recall value acquired by the proposed model for TRECVID 2018, 2019, 2020 and 2021 in fade-out detection is 93%, 95%, 94%, and 92%, respectively. The graph shows the maximum recall value gained by the proposed model for detecting fade-out. So the proposed model has the potential to detect the fade-out transitions effectively.

Figure 4(ii) demonstrates the proposed model's precision values for fade-out transition detection. The proposed model has achieved better precision for detecting the fade-out transition observed in the graph. The precision value achieved by the proposed model is 94.04%, 95.2%, 92.15%, and 90.73% for fade out detection using TRECVID 2018, 2019, 2020 and 2021 datasets. As a result, it is seen that the precision obtained by the proposed model is better than existing models for the detection of fade-out transitions.

Figure 4(iii) evaluates the f-measure value acquired by the proposed and existing fade-out transition models. The proposed model acquired f-measure values for TRECVID 2018, 2019, 2020 and 2021 datasets is 93.32%, 92.1%, 94.77% and 93.81%. The existing models, such as CNN, DNN, RNN, DBN, and DCNN, have achieved a lower F-measure value than the proposed model. From the graphical representation, it is proved that the proposed model has improved the performance of the fade-out transition detection.

Table 3 demonstrates the comparative analysis of proposed and existing models such as CNN, DNN, RNN, DBN, and DCNN in terms of recall, precision and f-measure for fade-out transition. The proposed model attained better recall, precision and f-measure than the existing model for detecting fade in transition using TRECVID 2018, 2019, 2020 and 2021 datasets.

Cut transition For cut transition detection, the metrics such as recall, precision, and f-measure are computed from the video sequence in the TRECVID 2018, 2019, 2020 and 2021 datasets. The graphical representation of the cut transition is demonstrated to show the superiority of the proposed model.

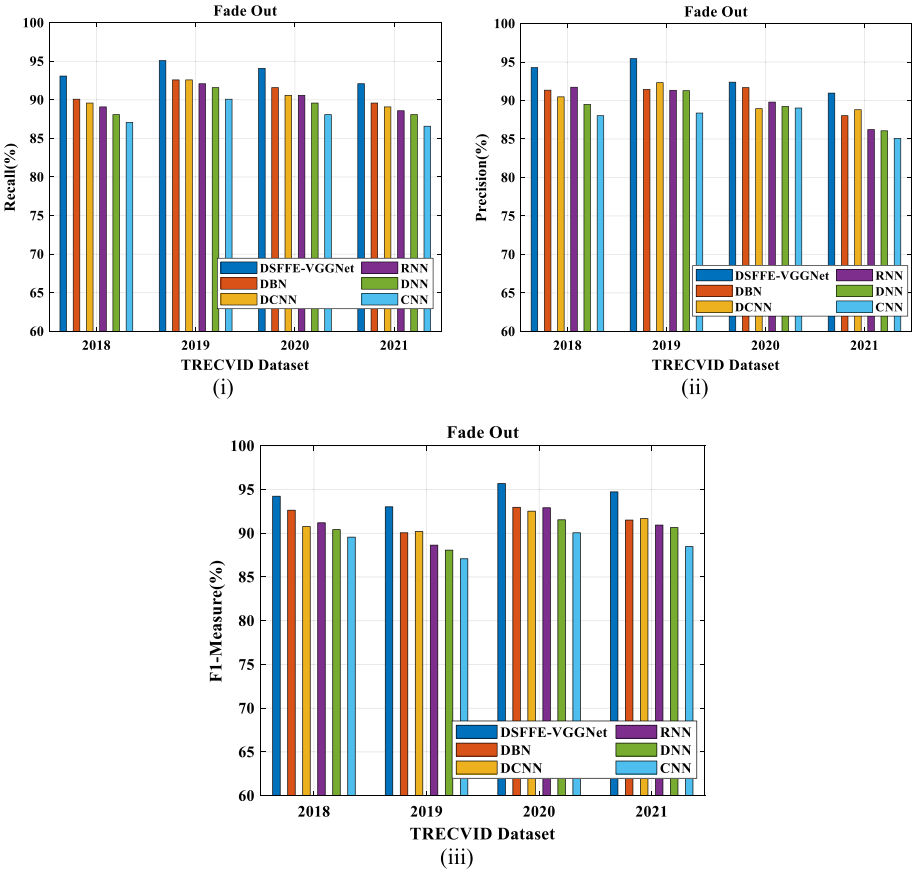

Fig. 4 Experimental outcomes of proposed and existing model for fade-out transition using TRECVID datasets (**i**) Recall (**ii**) Precision (**iii**) F-measure

Figure 5 demonstrates the results of cut transition in terms of recall, precision and accuracy from the TRECVID 2018, 2019, 2020 and 2021 datasets. Figure 5(i) describes the performance comparison of the recall with the proposed and existing models. The proposed model has accomplished a better recall value than existing models to detect cut transitions due to selecting the right classifier. The recall obtained by the proposed model is about 95%, 92.5%, 92.5% and 91.5% for the TRECVID 2018, 2019, 2020 and 2021 datasets. But, the existing models like CNN, DNN, RNN, DBN, and DCNN obtained less recall value than the proposed model.

Figure 5(ii) presents the performance comparison of precision with existing models such as CNN, DNN, RNN, DBN, and DCNN, respectively. The graph shows that the proposed model has gained better outcomes in terms of precision than existing models. The proposed model achieved high precision of 95.76%, 96.14%, 92.5% and 91.34% for TRECVID 2018, 2019, 2020 and 2021 datasets, so the false detection rate is very low than existing models.

Table 3 Comparative analysis for fade-out transition

Metrics	TRECVID dataset	Methods					
		DSFFE-VGGNet (Proposed)	DBN	DCNN	RNN	DNN	CNN
Recall	2018	93	90	89.5	89	88	87
	2019	95	92.5	92.5	92	91.5	90
	2020	94	91.5	90.5	90.5	89.5	88
	2021	92	89.5	89	88.5	88	86.5
Precision	2018	94.04	91.11	90.24	91.5	89.26	87.79
	2019	95.2	91.21	92.08	91.1	91.04	88.14
	2020	92.15	91.44	88.7	89.56	89	88.79
	2021	90.73	87.79	88.56	85.99	85.83	84.85
F-measure	2018	93.32	91.72	89.84	90.28	89.5	88.64
	2019	92.1	89.13	89.28	87.72	87.15	86.16
	2020	94.77	92.05	91.61	92	90.62	89.13
	2021	93.81	90.59	90.75	90.02	89.73	87.55

Figure 5(iii) presents the performance comparison of the f-measure proposed and existing models. The proposed model may be able to correctly recognize the class labels for the input data. The graph shows that the proposed model has determined a maximum f-measure value than existing models. In work, the f-measure value obtained by the proposed model is about 93.21%, 91.4%, 93.62%, and 93.28% for TRECVID 2018, 2019, 2020 and 2021 datasets. The existing models have achieved a considerably lower f-measure.

Table 4 illustrates the experimental results of the proposed and existing models for cut transition detection. The existing models like CNN, DNN, RNN, DBN, and DCNN are considered for a fair comparison. Table 4 shows that the proposed model has obtained superior outcomes than the existing model for providing cut transition detection.

Dissolve transition The recall, precision, and f-measure of the proposed model and recent existing CNN, DNN, RNN, DBN, and DCNN are computed for dissolve transition from TRECVID 2018, 2019, 2020 and 2021 datasets to compute the efficiency of the proposed model. The graphical representation of the dissolve transition is revealed to express the potentiality of the proposed model. Figure 6 demonstrates the results of cut transition in terms of recall, precision and accuracy from TRECVID 2018, 2019, 2020 and 2021 datasets.

Figure 6(i) resembles the performance analysis of recall acquired by proposed and existing models from TRECVID 2018, 2019, 2020 and 2021 datasets for dissolve transition detection. For TRECVID 2018, 2019, 2020 and 2021 datasets, the proposed model considerably achieved a recall of 95%, 94%, 94% and 91.5%. However, the existing models gained less recall value due to no proper feature extraction and classification.

Figure 6 (ii) compares precision with proposed and existing models. The proposed model has reached a precision of 96.51%, 92.55%, 91.8% and 92.49%, in contrast with existing models like CNN, DNN, RNN, DBN, and DCNN. If the precision value is more, the model's performance is efficient for dissolve transition detection. The graph shows that the proposed model can distinguish the false rates from the input video.

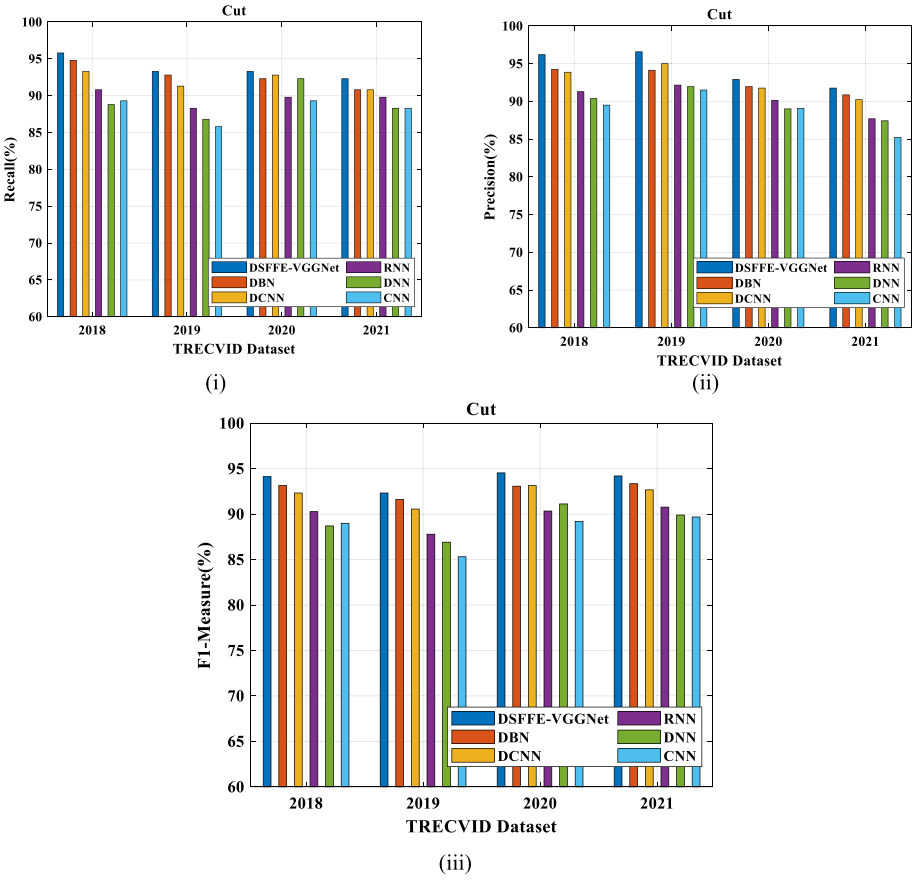

Fig. 5 Experimental outcomes of proposed and existing model for cut transition using TRECVID datasets (**i**) Recall (**ii**) Precision (**iii**) F-measure

Figure 6 (iii) demonstrates the comparison of the f-measure with proposed and existing models for detecting dissolve transition. The f-measure value achieved by the proposed model is greater than the existing models. In dissolve transition detection, the f-measure value gained by the proposed model is 92.88%, 91.46%, 92.77% and 93.4% for TRECVID 2018, 2019, 2020 and 2021 datasets. The comparative analysis proved that the f-measure value obtained by the proposed model is higher than existing models for dissolve cut transition.

Table 5 illuminates the experimental results of the proposed and existing models for dissolve transition detection. To analyze the performance of proposed model, some of the existing models like CNN, DNN, RNN, DBN, and DCNN are deliberated for a fair comparison. From Table 5, it is clear that the proposed model has obtained superior outcomes than an existing model for detecting dissolve transition.

Table 4 Comparative analysis for cut transition

Metrics	TRECVID dataset	Methods					
		DSFFE-VGGNet (Proposed)	DBN	DCNN	RNN	DNN	CNN
Recall	2018	95	94	92.5	90	88	88.5
	2019	92.5	92	90.5	87.5	86	85
	2020	92.5	91.5	92	89	91.5	88.5
	2021	91.5	90	90	89	87.5	87.5
Precision	2018	95.76	93.83	93.44	90.85	89.95	89.07
	2019	96.14	93.7	94.58	91.73	91.52	91.08
	2020	92.5	91.53	91.34	89.72	88.58	88.66
	2021	91.34	90.43	89.8	87.26	87	84.8
F-measure	2018	93.21	92.23	91.4	89.35	87.79	88.07
	2019	91.4	90.69	89.64	86.87	86	84.39
	2020	93.62	92.16	92.22	89.42	90.2	88.28
	2021	93.28	92.43	91.75	89.85	88.98	88.77

Performance analysis based on learning rate Figure 7 shows the performance analysis of average precision in the learning rate ranges from 0.1 to 0.001. The average precision is computed based on the values obtained for fade in/out, cut, and dissolve transitions on the TRECVID 2021 dataset. The graph depicts that the average precision of the proposed model on TRECVID 2021 attained a higher range when the learning rate is fixed to the minimum range. The higher range of precision represents the fair performance of the proposed model. The analysis clearly shows that the proposed model performs fairly on the learning rate of 0.001.

4.3.2 Evaluation of performance metrics using VideoSeg dataset

The task of SBD is performed on a video sequence of the VideoSeg dataset is stated. The dataset is challenging because it consists of a large distinction of shot breaks. However, the proposed model efficiently detects the shot boundaries. The comparative analysis of the proposed model has been carried out on the VideoSeg dataset with some metrics and compared. Figure 8 compares recall, precision and F-measure with proposed and existing models using the VideoSeg dataset. It demonstrates the performance of SBD. The proposed model can discriminate different classes such as fade-in, fade-out, cut and dissolve. Figure 8 shows that the proposed model has considerably achieved better recall, precision and f-measure than other existing methods like MCSH-RSD and POCS model. The overall outcomes of the proposed methods in terms of recall, precision, and f1-score are 98.89%, 98.15%, and 98.86%, respectively. The proposed model has enriched the potentiality of the model for detecting the class. The existing methods, such as MCSH-RSD and POCS, have obtained a recall of 98.21% and 97.91%, a precision of 97.05% and 96.43% and an f-measure of 97.6.5% and 97.11%. Henceforth, the proposed model is superior in detecting shot boundaries.

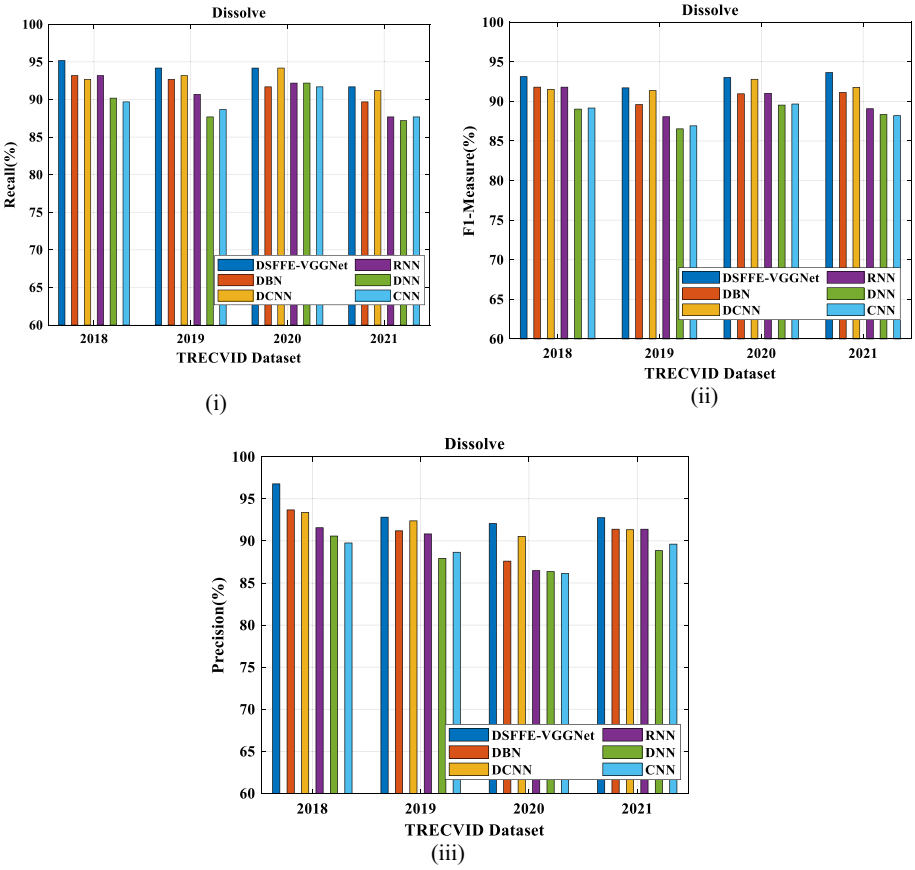

Fig. 6 Experimental outcomes of proposed and existing model for dissolve transition using TRECVID datasets (**i**) Recall (**ii**) Precision (**iii**) F-measure

4.4 ROC analysis

The receiver operating characteristic (ROC) is a probability curve created by plotting the false positive rate against the true positive rate at various threshold settings. Figure 9 illustrates the ROC curve analysis with proposed and existing models such as CNN, DNN, RNN, DCNN and DBN. It illustrates the performance of the classification model. The proposed model has the propensity to differentiate between the given classes of SBD. In Fig. 9, the graph has been strategized between a true positive rate and a false-positive rate for perceiving the ROC. The proposed model has reached better ROC than existing models like CNN, DNN, RNN, DCNN and DBN since it has enriched the potentiality of SBD class detection based on input. Henceforward, the proposed model is superior for SBD to existing models.

Table 5 Comparative analysis for dissolve transition

Metrics	TRECVID dataset	Methods					
		DSFFE-VGGNet (Proposed)	DBN	DCNN	RNN	DNN	CNN
Recall	2018	95	93	92.5	93	90	89.5
	2019	94	92.5	93	90.5	87.5	88.5
	2020	94	91.5	94	92	92	91.5
	2021	91.5	89.5	91	87.5	87	87.5
Precision	2018	96.51	93.41	93.11	91.3	90.31	89.49
	2019	92.55	90.93	92.11	90.56	87.65	88.39
	2020	91.8	87.33	90.27	86.21	86.08	85.86
	2021	92.49	91.12	91.06	91.12	88.57	89.35
F-measure	2018	92.88	91.56	91.28	91.56	88.78	88.92
	2019	91.46	89.36	91.12	87.82	86.29	86.67
	2020	92.77	90.71	92.55	90.78	89.29	89.43
	2021	93.4	90.88	91.53	88.83	88.1	87.96

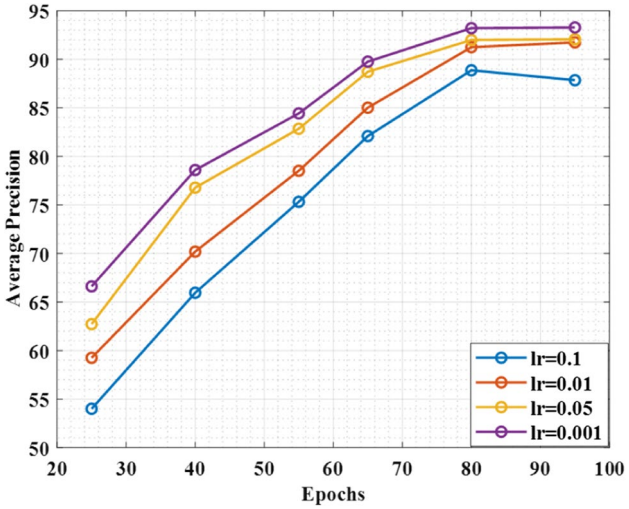

Fig. 7 Evaluation of Average Precision based on learning rate

4.5 Comparison of the proposed model with the recent SBD models

Table 6 shows the performance comparison of the precision, recall, and F1 scores. The average precision, recall and F1 score of the proposed model are greater than the existing models. A small variation exists in the precision between the proposed model and the BIFOLD-STAGE [5] model. However, the proposed SBD model has achieved the highest average F1 score compared to other SBD methods.

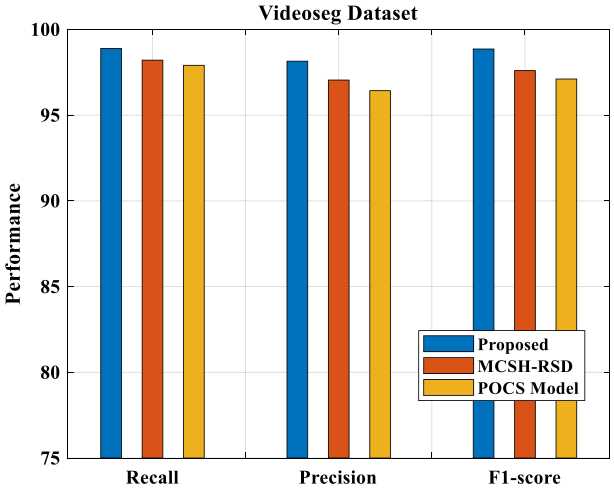

Fig. 8 Comparative analysis of recall, precision and f-measure using the VideoSeg dataset

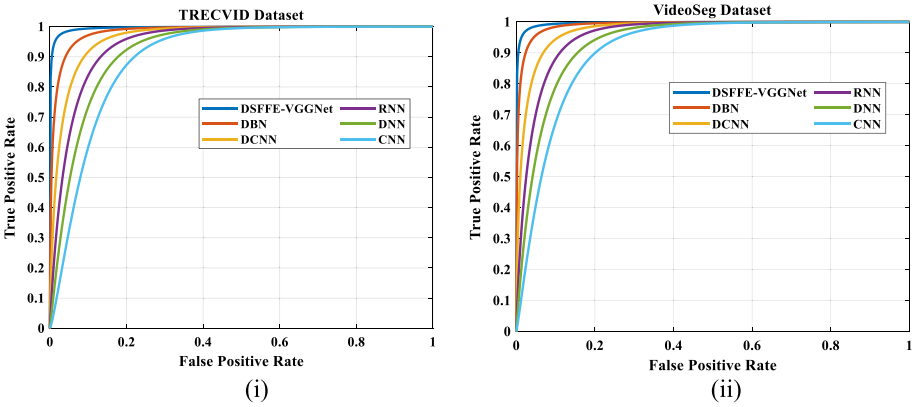

Fig. 9 Analysis of ROC with proposed and existing models (**i**) TRECVID dataset (**ii**) VideoSeg dataset

Table 6 Performance comparison of the proposed model with other SBD models

Model	Precision	Recall	F1-Score
LBB-HF[13]	95%	95%	94%
MCSH and RSD[22]	94.7%	94.2%	94.45%
BIFOLD-STAGE [5]	98.75%	95.37%	96.67%
POCS[23]	94%	94%	93.67%
Proposed Model	98.15%	98.89%	98.86%

4.6 Discussions

As mentioned above, multimedia data sharing over the Internet has increased exponentially due to digitization. This excessive progression tends to accomplish an efficient video retrieval and indexing tool. However, to establish an effective tool, the video content must

be organized appropriately, hence, video segmentation is necessary. SBD is the video segmentation process through spotting the transitions between the consecutive frames as well as the transition highlights the boundary between two successive shots considerably. To provide effective SBD, several methods are introduced; however, the histogram-based model [32] and colour histogram are prominently employed due to the motion-invariant property and computational cost. A logarithmic transform is used for pre-processing, and DCT handles the lighting effect. Structural similarity and complex dual-tree wavelet transform are applied in some models to minimize false positives, mainly due to the OCM and lighting effects.

Besides, a fuzzy logic-based model and a genetic algorithm are employed based on SBD pixels. Fort abrupt transition detection, a fast SBD method, has been presented using pixel-based schemes. Then, a simple model using standard deviation and SSIM has been accomplished for SBD to minimize the motion effect, illumination, and features such as quantized HSV color space. Additionally, an object tracking model is suggested for SBD to locate several frames where the specific object disappears. However, these methods suddenly disappear the objects from the frame, and the movement of large objects is misguided as uneven illumination and wipe transition. CNN based SBD [19] schemes are presented using adaptive thresholds, but these have problems such as misunderstanding abrupt changes. The proposed model achieved better performance by adapting Dual Stage Fused Feature Extraction as follows.

(1) The discriminability and robustness of the dual stage VGGNet model assist to minimize the false detection due to the illumination changes. As dual stage VGGNet extracting spatial–temporal features from a frame, it is beneficial for a fast SBD process. Moreover, the non-boundary frames are eliminated using Inter-frame Euclidean Threshold calculation, leading to detecting the shot transitions precisely.
(2) The RFO algorithm is highly suitable for optimizing the VGGNet model through the backpropagation process, which helped reduce the false detection of shot transitions.

5 Conclusion

This article proposes an effective SBD model using the DSFFE process executed by optimized VGGNet. First, pre-processing is performed by an enhanced bilateral filter (IBF) to eliminate the irrelevant information in video images for quality improvement. After pre-processing, a DSFFE process with optimized VGGNet is performed to extract deep and spatiotemporal features. DSFFE is a fusion feature extraction of abrupt (hard) shot transitions and gradual (soft) shot transitions. First, deep feature extraction is performed for each frame using the first stage of VGGNet. In contrast, spatio-temporal (motion and appearance) information feature extraction is done from the second stage of VGGNet. Both extracted features are merged in the fully connected layer of VGGNet. Next, a continuity matrix (association matrix) is constructed using IET to find the dissimilarity measure. Finally, the shot transitions are classified using the softmax level. The Softmax classifier categorizes the types of shot transitions as Hard (Cut) and Soft (Fade in/out, dissolve). The weights of the VGGNet model are updated using an optimization algorithm called RFO to minimize the false detection of shot transitions. The proposed model can only detect abrupt changes, fade in/out and dissolve, and does not adapt well to other niche shot transitions. In the future, other metaheuristic optimizations will be employed to enhance the performance

of SBD. Moreover, the SBD model will integrate object recognition and trajectory tracking techniques.

Authors Contributions All authors read and approved the final manuscript.

Data availability Data sharing is not applicable to this article.

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate All the authors involved have agreed to participate in this submitted article.

Consent to publish All the authors involved in this manuscript give full consent for publication of this submitted article.

Conflict of interest Authors have no conflict of interest to declare.

References

1. Idrees SM, Alam MA, Agarwal P (2019) A study of big data and its challenges. Int J Inf Technol 11(4):841–846
2. Tiwari V, Bhatnagar C (2021) A survey of recent work on video summarization: approaches and techniques. Multimed Tools Appl 80(18):27187–27221
3. Sreeja MU, Kovoor BC (2019) Towards genre-specific frameworks for video summarisation: A survey. J Vis Commun Image Represent 62:340–358
4. Bhaumik H, Bhattacharyya S, Chakraborty S (2019) A vague set approach for identifying shot transition in videos using multiple feature amalgamation. Appl Soft Comput 75:633–651
5. Chakraborty S, Singh A, Thounaojam DM (2022) A novel bifold-stage shot boundary detection algorithm: invariant to motion and illumination. Vis Comput 38(2):445–456
6. Yoon H, Han JH (2022) Content-Based Video Retrieval with Prototypes of Deep Features. IEEE Access 10:30730–30742
7. Pinge A and Gaonkar MN (2021) A Novel Video Retrieval Method Based on Object Detection Using Deep Learning. In Computational Vision and Bio-Inspired Computing, Springer, Singapore 483–495
8. Parihar AS, Pal J, Sharma I (2021) Multiview video summarization using video partitioning and clustering. J Vis Commun Image Represent 74:102991
9. Yan C, Li X and Li G (2021) A new action recognition framework for video highlights summarization in sporting events. In 2021 16th International Conference on Computer Science & Education (ICCSE). IEEE 653–666
10. Abdulhussain SH, Ramli AR, Saripan MI, Mahmmod BM, Al-Haddad SAR, Jassim WA (2018) Methods and challenges in shot boundary detection: A review. Entropy 20(4):214
11. Chakraborty S, Thounaojam DM (2019) A novel shot boundary detection system using hybrid optimization technique. Appl Intell 49(9):3207–3220
12. Ji Z, Xiong K, Pang Y, Li X (2019) Video summarization with attention-based encoder–decoder networks. IEEE Trans Circuits Syst Video Technol 30(6):1709–1717
13. Singh A, Thounaojam DM, Chakraborty S (2020) A novel automatic shot boundary detection algorithm: robust to illumination and motion effect. SIViP 14(4):645–653
14. Chakraborty S, Thounaojam DM, Sinha N (2021) A shot boundary detection technique based on visual colour information. Multimed Tools Appl 80(3):4007–4022
15. Nandini HM, Chethan HK, Rashmi BS (2021) An efficient method for video shot transition detection using probability binary weight Approach. Int J Comput Vis Image Process (IJCVIP) 11(3):1–20
16. Kumar K, Shrimankar DD (2017) F-DES: Fast and deep event summarization. IEEE Trans Multimedia 20(2):323–334

17. Kumar K, Shrimankar DD, Singh N (2017) Event bagging: A novel event summarization approach in multiview surveillance videos. In IEEE International Conference on Innovations in Electronics, Signal Processing and Communication (IESC) 106–111
18. Kumar K, Shrimankar DD, Singh N (2018) Eratosthenes sieve based keyframe extraction technique for event summarization in videos. Multimed Tools Appl 77:7383–7404
19. Kumar K, Shrimankar DD (2018) ESUMM: event summarization on scale-free networks. IETE Tech Rev 36:265–274
20. Mishra R (2021) Video shot boundary detection using hybrid dual tree complex wavelet transform with Walsh Hadamard transform. Multimed Tools and Appl 80(18):28109–28135
21. Souček T and Lokoč J (2020) TransNet V2: An effective deep network architecture for fast shot transition detection. arXiv preprint arXiv:2008.04838
22. Rashmi BS, Nagendraswamy HS (2021) Video shot boundary detection using block based cumulative approach. Multimed Tools Appl 80(1):641–664
23. Sasithradevi A, Mohamed Mansoor Roomi S (2020) A new pyramidal opponent color-shape model based video shot boundary detection. J Vis Commun Image Represent 67:102754
24. Zhou S, Wu X, Qi Y, Luo S, Xie X (2021) Video shot boundary detection based on multi-level features collaboration. SIViP 15(3):627–635
25. Chakraborty S, Thounaojam DM (2021) nSBD-Duo: a dual stage shot boundary detection technique robust to motion and illumination effect. Multimed Tools Appl 80(2):3071–3087
26. Benoughidene A, Titouna F (2022) A novel method for video shot boundary detection using CNN-LSTM approach. Int J Multimed Inf Retrieval 11(4):653–667
27. Li Q, Chen X, Wang B, Liu J, Zhang G, Feng B (2023) Shot Boundary Detection Based on Global Features and the Target Features. Symmetry 15(3):565
28. Kar T, Kanungo P (2023) A gradient based dual detection model for shot boundary detection. Multimed Tools Appl 82(6):8489–8506
29. Han Y, Zhang P, Zhuo T, Huang W, Zhang Y (2018) Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recogn Lett 107:83–90
30. Yu S, Park B, Park J and Jezz J (2020) Joint learning of blind video denoising and optical flow estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 500–501
31. VideoSeg (n.d.) http://www.site.uottawa.ca/~laganier/videoseg/. Accessed 7 Feb 2022
32. Liu T, Chan S (2014) Automatic shot boundary detection algorithm using structure–aware histogram metric. In: International Conference on Digital Signal Processing 541–546