Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An attention driven long short term memory based multi-attribute feature learning for shot boundary detection

Swati ChaitandasHadke, PhD Scholar [a,*], Ravi Mishra, Professor [b],
Rushikesh Tukaramji Bankar, Assistant Professor [c],
Sharda Amardas Chhabria, Associate Professor [d],
Shrikant Prakash Chavate, Assistant Professor [e], Latika Shyam Pinjarkar, Associate Professor [f]

[a] Department of Electronics & Telecommunication Engineering, G. H. Raisoni University, Amravati 444701, India
[b] Department of Computer Science and Engineering, Chhatrapati Shivaji Institute of Technology, Durg, 491001, India
[c] Department of Electronics and Telecommunication Engineering, G H Raisoni Institute of Engineering and Technology, Nagpur, 440036, India
[d] Department of Artificial Intelligence, G H Raisoni Institute of Engineering and Technology, Nagpur, 440036, India
[e] Department of Electronics & Telecommunication Engineering, G H Raisoni University, Amravati, 444701, India
[f] Symbiosis Institute of Technology Nagpur Campus, Symbiosis International (Deemed) University, Pune, 412115, India

ARTICLE INFO

Keywords:
Shot boundary detection
Euclidean distance
Attention dilated LSTM
Canonical Appearance Pooling
Motion Squeeze Network
Siamese Style Networks

ABSTRACT

Shot boundary detection is a crucial task in video analysis because it assistsin identifying the transitions between different shots that are fundamental for understanding video structure, editing, and content retrieval. Traditional methods for shot boundary detection often struggle to capture both subtle and complex changes in visual appearance and motion, leading to inaccurate or incomplete detection. This research introduces a novel shot boundary detection method using Siamese Style Networks with Canonical Appearance Pooling (SSN—CAP), Motion Squeeze Network (MSN), and Attention-based Dilated LSTM (AtDLSTM). SSN—CAP improves appearance feature extraction by using canonical appearance pooling, which preserves fine-grained features. MSN is used to extract motion features from keyframes, resulting in a thorough comprehension of dynamic variations. AtDLSTM combines dilated LSTM and attention mechanisms, significantly improving shot boundary detection accuracy. Experimental results demonstrate the superiority of the proposed method, achieving 100 % accuracy across various shot types (fade in, fade out, dissolve, cut) compared to other methods like Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), and Deep CNN (DCNN). The integrated technique takes advantage of the strengths of appearance and motion features, resulting in a reliable solution for precise shot boundary detection in video processing.

1. Introduction

In recent years, social networks and internet platforms have grown in popularity, and a vast volume of video content is created every minute. Researchers are concentrating their efforts on the automated recognition of complex events in web videos for computer vision and multi-media systems [1].The large number of input videos makes it time-consuming to manually search for an intriguing incident, such as a theft or burglary, as well as to observe an event. To detect unusual behaviour, a sophisticated computer vision system has been required to discern between normal and abnormal occurrences without the need for human interaction [2]. Shot boundary detection is an important component of video comprehension. The purpose of shot boundary identification is to identify shots with notable gaps in various visual content areas, as indicated by the attributes of the input video's shots [3].

There are two types of video shots or shot transitions: gradual transitions (GT) (soft cuts) and abrupt transitions (AT) (hard cuts) [4]. The fundamental goal of spatial video segmentation is to detect shot borders, which might include both hard and soft-cut transitions. A hard-cut transition occurs when the camera is turned on and off, but a soft-cut transition is created by modifying techniques. However, the effectiveness of shot boundary identification is substantially limited by the vast range of hard-cut and gradual-change transitions between image frames.

[*] Corresponding author.
E-mail address: swati13hadke@gmail.com (S. ChaitandasHadke).

https://doi.org/10.1016/j.knosys.2025.113379
Received 8 July 2024; Received in revised form 18 December 2024; Accepted 18 March 2025
Available online 1 April 2025

The main challenges in video processing are clarity, rotation, scale-variant temporal and territorial information. Certain feature extraction algorithms use colour or edge properties to identify shot boundaries [5]. Shot boundary detection algorithms are classified into two broad groups: compressed and uncompressed fields, which are determined by the feature extraction domain. If the features are extracted from the compressed domain, rapid shot boundary detection algorithms do not require decoding video frames [6]. However, researchers are giving more attention to the uncompressed domain because video shots contain a large amount of visual information.

Certain feature aggregation or reduction of dimension approaches are used in various models, resulting in low performance when converting features into various feature spaces. When the edges of continuous images fluctuate significantly, edge-based techniques [7] are used to collect the object boundaries for each frame. Pixel-based approaches [8] monitor pixel changes across successive frames, allowing for easier deployment. However, it is unable to detect slight changes quickly and is extremely sensitive to object motion.Using motion vectors to avoid object motion can make a motion-based method [9] more complex. The Fast Averaging Peer Group (FAPG) [10] pre-processing approach was designed to boost frame contrast and remove lighting distortion. Furthermore, multiple invariant feature descriptor techniques are employed with each technique conducting various graphic interpretations.

Traditional unsupervised algorithms [11] for scene boundary recognition do not give competitive rates of accuracy. However, supervised techniques need significant quantities of labelled training data, and hence supervised learning models do not scale effectively [12]. Many self-supervised learning techniques have been used subsequently to develop general graphic representations for images and short videos [13]. Training-based techniques, including genetic algorithms, clustering methods, support vector machines (SVM), and deep learning (DL), generally produce strong identification outcomes. Still, these learning-based techniques often need a lengthy period of training for the identification model [14]. Furthermore, the efficiency of shot boundary identification is highly dependent on enough training data. Deep learning approaches have been utilized to handle recognizing problems in the image and video domains. The combination of CNNs with recurrent neural networks (RNNs) results in a potent design for video categorization [15]. This design can efficiently analyze spatial and temporal data at the same time. However, the varied background and intricate movement changes in videos provide an additional challenge for temporal information encoding. In addition, a cascading architecture based on the Deep Structured Model (DSM) [16] is proposed to improve the effectiveness of shot boundary detection. Deep feature-based approaches have recently made significant advances in a variety of dynamic high-dimensional data domains, including behaviour recognition and video summarization. The application of Generative Adversarial Networks (GANs) to detect discrepancies in videos is significant [17]. This model depicts the standard distribution of videos by combining discriminator and generator approaches with GANs.

Video shot boundary detection is crucial for content analysis, aiding in scene segmentation, video summarization, and indexing. Its applications encompass video editing, surveillance, and multi-media retrieval. Conventional deep learning-based shot boundary detection faces challenges, such as limited labelled data, diverse shot types, variations in lighting and camera angles, and hindering model generalization. Hence, to overcome the challenges, deep learning-based shot boundary detection is introduced. The major contributions are:

- ***SSA-CAP Appearance feature extraction***: Siamese Style Networks-Canonical Appearance Pooling (SSN—CAP) is utilized to extract the appearance features from the chosen keyframes. In this, the conventional pooling is replaced with a canonical appearance pooling layer for obtaining appearance variations within local regions, enabling better preservation of fine-grained details.
- ***MSN-based Motion Feature Extraction***: The motion features from the keyframes are extracted using the motion squeeze network (MSN).
- ***AtDLSTM-based shot boundary Detection***: The shot boundary detection is employed using the AtDLSTM, wherein the Dilated LSTM and attention mechanism are hybridized to enhance the detection accuracy.

The organization of the research is as follows: Section 2 details the related works with the problem statement, and Section 3 elaborates on the proposed shot boundary detection mechanism. Section 4 details the experimental outcome, and Section 5 presents the conclusion with the future scope.

2. Related works

The related works concerning the shot boundary detection methods are reviewed in this section.

In order to overcome the challenges associated with determining shot boundaries, Chen et al. [18] suggested a self-supervised learning method. Initially, the author had been breaking down a full-length input video into its various sets of shots using traditional shot recognition algorithms. After that, contrastive learning was used to produce scene representations that were useful for shot boundary identification since it efficiently encoded the local scene structure. However, the implemented model was less effective in video representation.

Idan et al. [19] suggested a quick video processing strategy for shot boundary recognition based on the frame’s active area and a candidate segment selection mechanism. To reduce computation costs and interruption factors, the active zone of each frame was initially selected, and so the informational content was considered. Furthermore, orthogonal polynomials were used to calculate the moments for every active region. Candidate segments were subsequently retained, and the majority of non-transition shots were eliminated with the help of an adaptive threshold and inequality criterion. In this implemented model, the squared Tchebichef-Krawtchouk polynomials (STKP) were introduced to enhance the effectiveness of the SBD model. However, the developed design was not able to identify various kinds of shot transitions.

Kang et al [20]. had been utilizing the Temporal Self-similarity Matrix (TSM), which was used in the new model of unsupervised/supervised Generic Event Boundary Detection (GEBD) for video formatting. The author suggested a divide-and-conquer strategy of the Recursive TSM Parsing (RTP) algorithm for identifying occurrence boundaries. Additionally, they provided an unsupervised structure for GEBD that combined the novel Boundary Contrasting (BoCo) loss with RTP. However, the implemented model was not able to carry out videos of different lengths.

Nandini et al. [21] used the Local Binary Pattern (LBP) approach to characterize textures and extract binarized edge data from frames to identify abrupt shots. Additionally, an adaptive threshold was employed to identify abrupt shots, and Euclidean distance was employed to generate histogram features. In the following phase, the Sobel gradient function was applied, and its magnitude was evaluated for each frame in a shot to extract keyframes. Furthermore, the suggested approach was unable to analyze additional visual qualities and could not recognize the video’s gradual transitions.

Hadke et al. [22] used a dual-stage fused feature extraction method based on a deep learning model to provide invariant spatial-temporal properties that aid in accurate identification. The author employed the Red Fox Optimization (RFO) approach to use reverse propagation for altering the VGGNet model parameters in order to reduce shot transition identification errors. In order to increase effectiveness, an improved bilateral filter (IBF) was used in pre-processing to remove unnecessary data from video frames. Additionally, the Inter-frame Euclidean Threshold was used to determine the feature similarity among consecutive frames in order to exclude the non-border frames. However, the

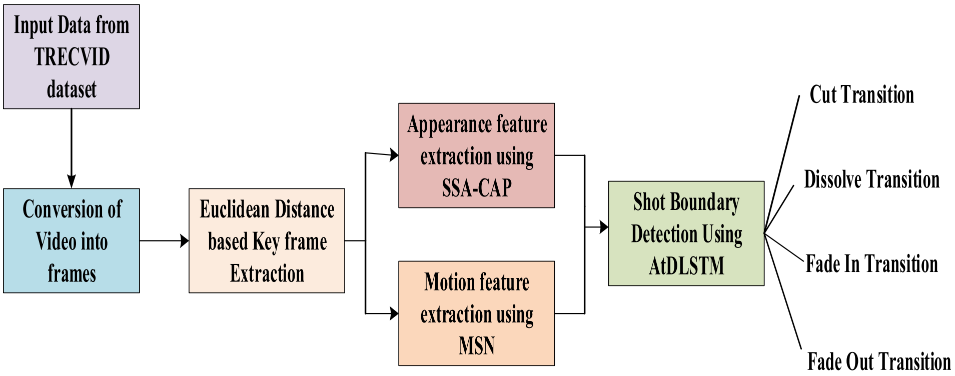

Input Data from TRECVID dataset
Conversion of Video into frames
Euclidean Distance based Key frame Extraction
Appearance feature extraction using SSA-CAP
Motion feature extraction using MSN
Shot Boundary Detection Using AtDLSTM
Cut Transition
Dissolve Transition
Fade In Transition
Fade Out Transition

Fig. 1. Block diagram of the proposed methodology.

suggested meta-heuristic optimization model offered less effectiveness.

Chakraborty et al. [23] designed a pattern-based model based on mean luminance and acquired enhanced outcomes in terms of precision. Still, the abrupt transitions were not effectively detected by the model. Soucek et al. [24] designed a transfer model based on shot boundary detection and obtained better performance with a gradual transition. Still, the computation burden associated with the model was higher. SavranKızıltepe et al. [25] designed a DNN-based model with keyframe extraction. The designed model accomplished enhanced outcomes in keyframe extraction; still, the designed model acquired poor performance compared to the conventional state-of-art methods.

Theiner et al. [26] designed a deep learning based model and obtained enhanced outcomes; still, over-fitting issues limited the performance. Ul Amin et al. [27] designed a shot boundary detection for performing the anomaly detection application from the surveillance video. The designed model utilized a hybrid deep learning model with CNN and LSTM for detecting the boundaries and acquired enhanced outcomes; still, the model was not applicable to real world issues due to the higher computation burden. Wei et al. [28] introduced a Transformer model with shot merger in detecting the boundaries. The vanishing gradient issues led to the instability of learning and degraded performance. Kapre et al. [29] designed a shot boundary detection using the Pearson correlation coefficient (PCC) for performing the image watermarking application. The designed model demonstrated a robust outcome irrespective of the illumination. Still, the designed model was sensitive to noise and was inefficient in detecting gradual transitions.

2.1. Problem statement

Shot boundary detection remains a challenging task due to the complexity of accurately identifying transitions between shots in video content. Existing methods designed several models for addressing these challenges, but limitations still exist. The self-supervised learning method efficiently encoded local scene structures; still, its overall effectiveness in video representation remained limited. The shot boundary detection designed by [19] eliminated non-transition shots to improve the model, but it failed to identify various types of shot transitions. TSM designed by [20] faced challenges in detecting the shot for videos of varying lengths. LBP based approach effectively detects abrupt shot transitions by analysing textures and edges in frames, but the method fails to detect gradual transitions and cannot evaluate additional visual features.RFO+VGGNet model identified the shot transitions; still, it faced issues with overall effectiveness. Thus, the existing models highlight the difficulty in developing a comprehensive solution for shot boundary detection that can handle a wide range of video types and transition styles, while balancing computational efficiency and detection accuracy. To overcome the challenges, this research introduces a novel shot boundary detection using a deep learning model.

3. Proposed methodology

Shot boundary detection is a technique for automatically detecting the margins between shots in a video.This subject has garnered significant attention because nearly all video analysis, indexing, summarization, search, and other content-based activities necessitate it as an essential pre-processing step. CNNs have been widely utilized to learn appearance features from video frames. CNNs have recently been developed to learn temporal characteristics using spatiotemporal convolution over several frames. Motion is the most distinguishing feature between videos and still images. Motion patterns cannot be accurately learned using spatiotemporal convolution. As a result, most contemporary approaches continue to rely on explicit motion aspects, such as dense optical flows. This creates a significant computational barrier in video processing models. The block diagram of the proposed shot boundary detection is portrayed in Fig. 1.

3.1. Data gathering

The input data video for processing the proposed shot boundary detection is obtained from the TRECVID dataset. The captured video is first transformed into frames, and the keyframes are selected in the following phase to reduce the computational weight.

3.2. Key frame extraction

The keyframe extraction employs the Euclidean distance, which calculates the distance between two locations that represent feature vectors of frames and is computed using the formula:

$$KF_{ED} = \sqrt{\sum_{x=1}^{l} (p_x - q_x)^2} \tag{1}$$

here, $p_x$and $q_x$are the corresponding components of the feature vectors for frames p and q, and l is the dimensionality of the feature vectors. KF_{ED} signifies the extracted keyframe using the Euclidean distance.

3.3. SSN—CAP-based appearance feature extraction

The multi-granularity features that are invariant to features are extracted using the Siamese Style Networks-Canonical Appearance Pooling (SSN—CAP) appearance feature extraction technique.SSN—CAP is utilized for appearance feature extraction because it efficiently captures and compares the appearance of keyframes. For learning the model to identify the most important visual features across frames, SSN—CAP effectively detects visual changes between scenes, such as shifts in lighting, object appearance, or background, which are crucial for detecting shot boundaries. Thus, SSN—CAP is utilized for performing

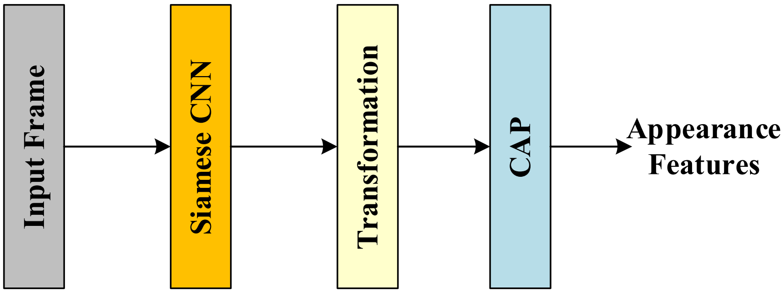


Fig. 2. SSN—CAP-based feature extraction.

the extraction of appearance based features. SSNs are a type of deep neural network architecture used to compare two inputs. When it comes to video frames, SSNs are excellent for feature extraction since they can learn to identify similarities and differences across frames even when faced with variances such as lighting, viewpoint, or object pose.The SSN is normally made up of two identical branches, each processing one of the incoming video frames. These branches have the same weights and architecture, so they can extract similar features from both frames. The branches' outputs are then passed into a similarity metric function, which generates a score representing the similarity of retrieved features from both video frames. Instead of using a traditional pooling layer, the CAP is integrated into the SSN for appearance-based feature extraction. The architecture of the SSN—CAP-based appearance feature extraction is presented in Fig. 2.

Let the input video frame be denoted by$P \in R^{X \times Y \times Z}$, where the channel, width, and height of the image are signified as Z, Y, and$X$, respectively. The appearance features are extracted through the SSN network and are denoted by$A_b \in R^{X' \times Y' \times Z'}$. Here, the granularity index is signified as$b$. The flattening of the features is employed the same as the traditional convolutional networks and is referred to as $A_b \in R^{X' \times Y' \times Z'}$, which is in the form of covariance matrix. Then, the Symmetric Positive Semi-definite form of the matrix is evaluated and is denoted by $W_b^+ \in R^{(Z'+1) \times (Z'+1)}$. For various granularities P_b, the Symmetric Positive Semi-definite is evaluated, and finally, they are concatenated to form a single matrix format. To extract the features based on appearance, the probability distribution is evaluated and formulated as follows:

$$P \rightarrow W^+ \in Sym^+, p(W^+) = a_q(Q \circ W^+) \tag{2}$$

For the input P, the covariance feature $W^+$is accomplished through the factor $P \rightarrow W^+$, wherein the property of the Symmetric Positive Semi-definite matrix is indicated as Sym^+. The softmax layer is signified as a_q, which is utilized for mapping the weighted Symmetric Positive Semi-definite $Q \circ W^+$. Here, Qsignifies the granular feature of W^+, which is formulated as:

$$W^+ = \frac{1}{B} \sum_{b=1}^{B} W_b^+ \tag{3}$$

where, the total count of the granularities is defined as$B$in the Symmetric Positive Semi-definite matrix W_b^+. The extracted features are then fed into the CAP to learn the transformation invariant features.

CAP Layer: The appearance features are extracted using transformation invariant features and are formulated as:

$$A_b^\theta = a_g(\theta(P_b)) \tag{4}$$

where, $\theta \in \Phi$signifies the pre-defined rotation transformation and is defined as$\theta_q = \frac{360^0}{\dim(\Phi)}$, and the transformation length is signified as $\dim(\Phi)$. The feature extraction process is indicated as $a_g(\cdot)$, and the images transformed is defined as $\theta(P_b)$. Here, the optimal extraction of feature is employed using the maximum operator and is formulated as:

$$A_b^\theta \rightarrow \left(W_b^\theta\right)^+, W_b^+ = \underset{\theta \in \Phi}{max} a_d\left(\left(W_b^\theta\right)^+\right) \tag{5}$$

here, the Gaussian covariance matrix is accomplished through the transformation of features through $A_b^\theta \rightarrow \left(W_b^\theta\right)^+$. Besides, the features gathered through the second-order function are signified as$\left(W_b^\theta\right)^+$. Thus, the invariant features are accomplished and are denoted by W_b^+. In this case, the second-order function-based transformation helps to obtain better regional features than the first-order. The high dimensionality of the features may cause computation stress; hence, feature aggregation is used with the channel dimension.

3.4. MSN-based motion feature extraction

Motion changes like sudden movements or transitions that are often used to distinguish one shot from another are extracted through motion feature extraction. However, motion features are finer, and capturing complex information is employed through the motion based feature extraction module. The motion features from the chosen key frames are extracted using the motion squeeze networks (MSN). It is a lightweight and efficient neural module designed for extracting motion features from video frames. The MSN utilizes three various steps: Correlation Computation, Displacement Estimation and Feature Transformation for the extraction of the motion features. The structure of MSN is portrayed

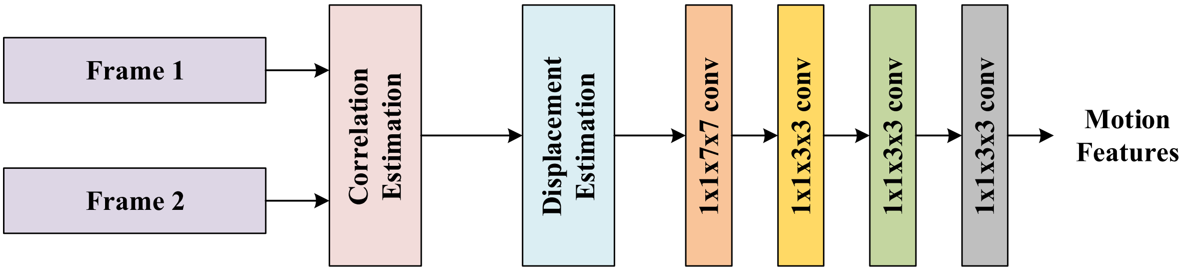


Fig. 3. Structure of MSN for motion feature extraction.

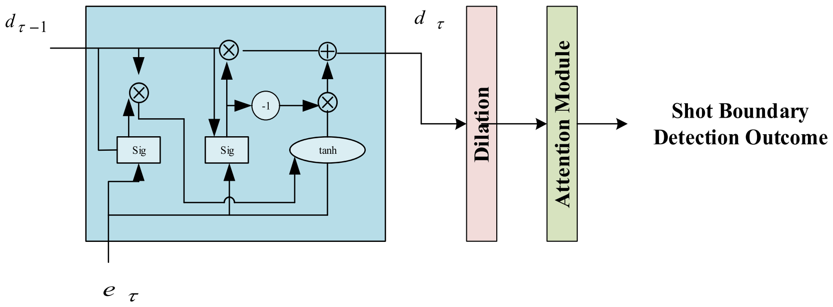


Fig. 4. Structure of AtDLSTM for shot boundary detection.

in Fig. 3.

Estimation of Correlation: It calculates the correlation between feature maps of adjacent frames, identifying areas of similarity and difference. The correlation estimation establishes an initial understanding of motion patterns. The two input frames are provided as input to the MSN and are indicated as $G(k)$and $G(k+1)$respectively with the size of X, Y, and Z. Here, $X \times Y$represents the spatial resolution and Zindicates the dimension. The correlation between the frames is formulated as follows:

$$f(i,j,k) = G_i^k \cdot G_{i+j}^{k+1} \tag{6}$$

here, the position is denoted by i, and its correlation score is indicated as $J = 2m + 1$with the limits of $j \in [-m,m]^2$. The outcome of correlation accomplished at the k^{th}frame is indicated as $X \times Y \times Z^2$. The correlation estimation utilizes the cost equivalent to 1×1 convolution with kernels of J^2 .

Calculation of Displacement: MSN uses the correlations to estimate the displacement of corresponding sections of the objects between frames. The displacement identifies the real motion occurring. Here, the outcome derived from the correlation tensor $F(k)$is fed into the displacement estimation module and is defined as:

$$s(i,k) = \sum_j \frac{\exp(f(i,j,k))}{\sum_{j'} \exp(f(i,j',k))} j \tag{7}$$

The output derived by the displacement function comprises noisy factors that can be eliminated by incorporating the kernel-soft-argmax. The outcome by this stage is estimated as:

$$s(i,k) = \sum_j \frac{\exp(n(i,j,k)f(i,j,k)/T)}{\sum_{j'} \exp(n(i,j',k)f(i,j',k)/T)} j \tag{8}$$

where,

$$n(i,j,k) = \frac{1}{2\pi\sigma} \exp\left(\frac{j - \text{argmax}_j f(i,j,k)}{\sigma^2}\right) \tag{9}$$

A mathematical function often used for smoothing or filtering is employed using the Gaussian filter and is signified as$n(i,j,k)$. The value assigned for the standard deviation is 5 and Tsignifies the temperature factor.

Then, the motion-based features are acquired through the confidence map estimation that is devised based on the maximum pooling and is formulated as:

$$f * (i,k) = \max_j f(i,j,k) \tag{10}$$

After estimating the displacement, the feature transformation is devised and calculated in the following step.

Transformations of features: Based on the estimated displacements, MSN transforms the initial feature maps to capture the motion information. The transformation of features results in specific motion features that depict the video's dynamic aspects. The displacement tensor $S(k)$is initially obtained, representing the spatial shifts or displacements between consecutive frames at time k. The tensor$S(k)$undergoes processing through four depth-wise separable convolution layers. These layers include one $1 \times 7 \times 7$ convolutional layer followed by three $1 \times 3 \times 3$ convolutional layers. Depth-wise separable convolution is chosen for its computational efficiency compared to traditional 2D convolution.

The result of the convolution layers is a motion feature tensor$U(k)$-with the number of channels$Z$as the original input feature tensor$G(k)$. After the point-wise and depth-wise convolution layers, batch normalization and ReLU activation are applied. Batch normalization improves training stability, but ReLU gives the network nonlinearity. The feature transformation process described here is intended to allow users to learn motion characteristics specific to a given activity. It is accomplished by reading the semantics of confidence and displacement, and convolution layers help in capturing these attributes. Two neighbouring appearance features, $G(k)$ and $G(k+1)$, are used to form the Motion Feature $U(k)$, which is subsequently added to the input of the network's next layer. Padding $U(k-1)$results in the final motion feature$U(k)$when working with a series of $k$frames. To ensure that motion information extracted is continuous across frames, the assumption $U(k) = U(k-1)$ is considered. Thus, the motion-based features are accomplished using MSN to perform the video shot boundary detection.

3.5. AtDLSTM based shot boundary detection

AtDLSTM is employed to capture the temporal dependencies across frames and detect shot boundaries by learning from long-term patterns in the input sequence, and the attention mechanism is employed to focus on the most relevant transitions. It is effectively devised using the AtDLSTM, wherein the cascaded LSTM is used along with the attention mechanism. Using the LSTM, long-term dependent characteristics are retrieved to improve detection accuracy. LSTM is well-suited for tasks requiring long-term relationships in sequential data, making it useful for detecting shot boundaries. The ability of LSTMs to selectively keep or forget information from the past makes them resistant to variations in shot transition patterns. The proposed shot boundary detection technique makes use of dilated LSTMs, which help to capture information from a larger range of time steps by employing gaps or skips between connections. The dilation factor determines the space between connections. For example, if the dilation factor is 2, the connections skip one unit; if it is 3, the connections skip two units. The dilation increases the LSTM's receptive field, allowing it to capture both short-term and long-term dependencies in the input sequence. Furthermore, the attention mechanism enables the model to assign varying degrees of importance to distinct parts in the sequence, emphasizing critical information and increasing overall performance. The structure of AtDLSTM is portrayed in Fig. 4.

The three various gates utilized by AtDLSTM are outputs_τ^o, inputs_τ^i and forget s_τ^fgate, which is utilized for controlling the outputd_τ and cell

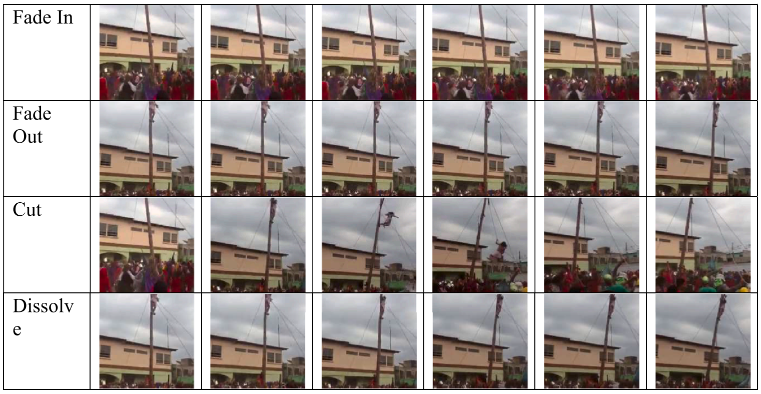


Fig. 5. Experimental outcome.

stateA_τ. The required information is gathered through the cell state for the time step 1 to τ.The outcomes of gate functions ranging from 0 to 1 are shown here to indicate the proportion of reserved information. The forget gate is a sigmoid activation function that accepts a concatenation of the prior concealed state and the current input. The forget gate determines whether the information from the preceding memory cell is discarded or preserved. The outcome of the forget gate is stated as follows:

$$s^f_\tau = sig\left(D_f \cdot [d_{\tau-1}, e_\tau] + k_f\right) \tag{11}$$

Simultaneously, the input gate decides how much of the new information should be added to the memory cell. The combination of the forget gate and input gate operations ensures that the LSTM selectively updates its memory based on the relevance of past and current information. Its outcome is stated as follows:

$$s^i_\tau = sig(D_i \cdot [d_{\tau-1}, e_\tau] + k_i) \tag{12}$$

The cell state stores information from the previous time steps, allowing the model to learn and recognize shot transitions. The cell state is formulated as follows:

$$A_\tau = s^f_\tau \times A_{\tau-1} + s^i_\tau \times \tanh(D_c \cdot [d_{\tau-1}, e_\tau] + k_c) \tag{13}$$

The output gate determines which information from the memory cell should be sent to the next layer of the network. It also helps to control the flow of information. It is outlined as:

$$s^o_\tau = sig(D_o \cdot [d_{\tau-1}, e_\tau] + k_o) \tag{14}$$

where,

$$d_\tau = s^o_\tau \times \tanh(A_\tau) \tag{15}$$

where, the biases and weights are indicated as $k_f, D_f, k_o, D_o, k_i, D_i, k_c$, and D_c, concerning the forget, output, input, and cell state, respectively.

The dilated LSTM utilizes skip connections, and hence, the cell state is defined as:

$$A_\tau(r) = LSTM\left[s^o_\tau(r), A_{\tau-q^r-}(r)\right] \tag{16}$$

where, the dilation of skip length is signified as q^r, and $s^o_\tau(r)$is fed as input to the layer $r$at the time τ. The dilation layer for the skip length is outlined as:

$$q^r = B^{r-1}, r = 1, 2, ...R \tag{17}$$

here, the dilation at various layers is indicated as$B$.

The outcome of the dilated LSTM is passed into the attention layer, which assigns higher weights to the most appropriate features. The attention mechanism enables the model to make predictions based on specific parts of the input sequence, allocating various degrees of importance to different time steps. The attention weights $\delta_\tau$are computed using the softmax function on the output of the dilated LSTM and are formulated as follows:

$$\delta_\tau = soft\max(D_a \cdot d_\tau + k_a) \tag{18}$$

$$d^*_\tau = \delta_\tau d_\tau \tag{19}$$

here, the biases and weights concerning the attention layer are indicated as k_a and D_a respectively, and the hidden feature is denoted by d^*_τ. Thus, using the AtDLSTM, shot boundary detection is employed.

4. Results and discussion

The proposed video shot boundary detection is implemented in the PYTHON programming language and evaluated using the TRECVID dataset. The proposed method is evaluated based on various assessment measures like recall, precision, and F1-score, which are examined for fade-in transition, fade-out transition, cut transition, and dissolve transition.

TRECVID dataset: TRECVID, which stands for Text REtrieval Conference Video Retrieval Evaluation, is an annual international benchmarking activity that evaluates various information retrieval and video analysis tasks. TRECVID is organized by the National Institute of Standards and Technology (NIST) in the United States. The dataset includes videos from different sources, such as broadcast news, documentaries, surveillance footage, and online videos. SBD is the process of identifying the transitions between shots within a video. These transitions can be either abrupt, like cuts, or gradual, like fades and dissolves.

Autoshot Dataset: The Autoshot dataset comprises 853 videos, wherein 103 videos are related to Fade Out, 200 videos are related to Fade In, 250 videos are related to Dissolve, and 300 videos are related to Cut.

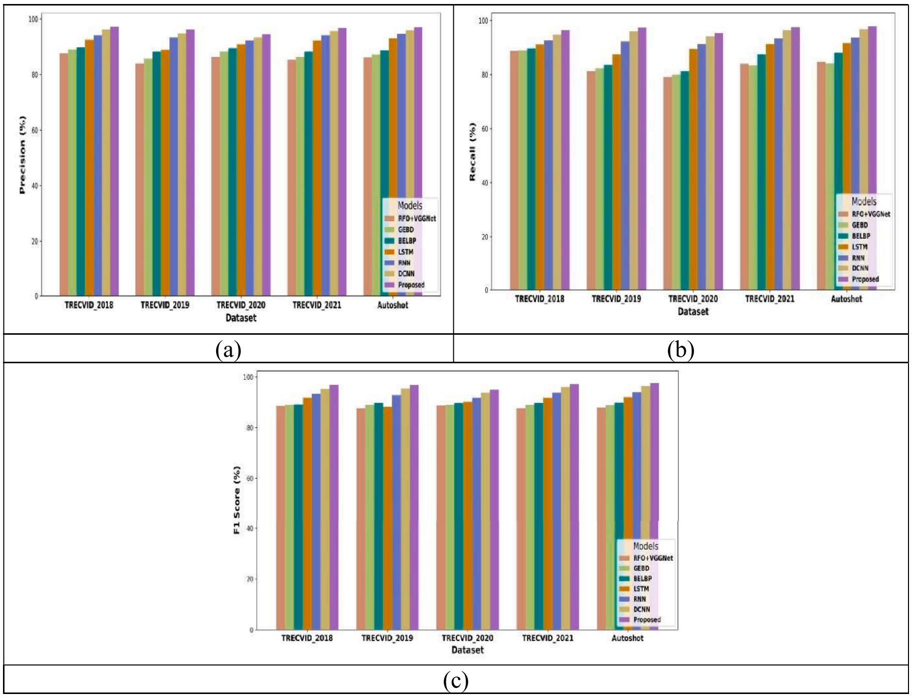
The description of the evaluation measures utilized to measure the performance of the proposed method is:

Precision: Precision, also known as positive predictive value, is the ratio of true positive detections to the total number of positive detections, like true positives and false positives. It represents the algorithm's accuracy in properly recognizing shot boundaries, including incorrect detections. It is defined as:

$$SB_{\text{Pre}} = \frac{SB_{tp}}{SB_{tp} + SB_{fp}} \tag{20}$$

Recall: Recall, also known as sensitivity or true positive rate, is the ratio of true positive detections to the total number of actual shot boundaries like true positives and false negatives. It examines the algorithm's ability to capture all relevant shot boundaries. It is defined as:

$$SB_{\text{Recall}} = \frac{SB_{tp}}{SB_{tp} + SB_{fn}} \tag{21}$$



(a)

(b)

(c)

Fig. 6. Fade in transition analysis based on (a) precision, (b) recall (c) F1-Score.

Table 1
Fade in transition analysis.

Dataset/ Methods	RFO+VGGNet	GEBD	BELBP	LSTM	RNN	DCNN	Proposed
Precision							
TRECVID_2018	87.59	88.98	89.79	92.54	94.12	96.15	97.22
TRECVID_2019	84	85.71	88.24	88.89	93.33	94.74	96.15
TRECVID_2020	86.23	88.24	89.47	90.91	92.31	93.33	94.44
TRECVID_2021	85.24	86.23	88.24	92.31	94.12	95.65	96.77
Autoshot	86.12	87.23	88.67	93	94.56	95.89	97.01
Recall							
TRECVID_2018	88.74	88.98	89.75	91.11	92.59	94.74	96.43
TRECVID_2019	81.25	82.35	83.54	87.5	92.31	96	97.44
TRECVID_2020	79.12	80	81.21	89.47	91.3	94.12	95.35
TRECVID_2021	84	83.33	87.5	91.3	93.33	96.43	97.56
Autoshot	84.67	84.12	88.12	91.67	93.67	96.78	97.89
F1-Score							
TRECVID_2018	88.59	88.98	89.12	91.82	93.33	95.24	96.83
TRECVID_2019	87.59	88.98	89.79	88.19	92.82	95.37	96.79
TRECVID_2020	88.74	88.98	89.75	90.19	91.8	93.72	97.24
TRECVID_2021	89.34	89.23	90.23	91.78	92.91	95.89	97.34
Autoshot	89.45	89.14	90.12	91.12	93.12	96.12	97.89

F1-Score: The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced measure that considers both false positives and false negatives. A higher F1-Score indicates a better balance between precision and recall. It is defined as:

$$SB_{F1-S} = 2 \times \frac{SB_{\mathrm{Pre}} * SB_{\mathrm{Recall}}}{SB_{\mathrm{Pre}} + SB_{\mathrm{Recall}}} \tag{22}$$

here, true positivesSB_{tp} refer to correctly identified shot boundaries, false positives $SB_{fp}$are instances where the algorithm incorrectly identifies a shot boundary, and false negativesSB_{fn} are instances where the algorithm fails to detect an actual shot boundary.

4.1. Experimental outcome

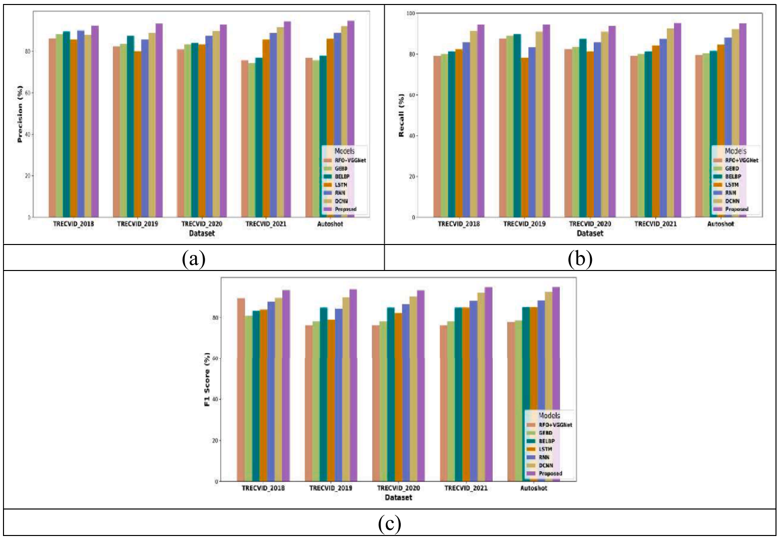
The experimental outcome of the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection model is portrayed in Fig. 5.

4.2. Comparative analysis

The performance of the proposed shot boundary detection based on the cut, fade-in, fade-out, and dissolve transitions for the TRECVID dataset for 2018, 2019, 2020, and 2021, along with Autoshot, is presented in this section.

4.2.1. Analysis based on fade in transition

The fade-in transition analysis using the TRECVID 2018, 2019, 2020, 2021, and Autoshot based on various assessment measures is portrayed in Fig. 6. The highest F1-score evaluated by the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection is 97.34 using the 2021 dataset, which is higher than all existing models. Here, the highest F1score in shot boundary detection signifies a well-performing model that effectively identifies shot transitions while maintaining a good balance between precision and recall. F1-score is valuable for evaluating the overall accuracy and reliability of a shot boundary detection



(a)

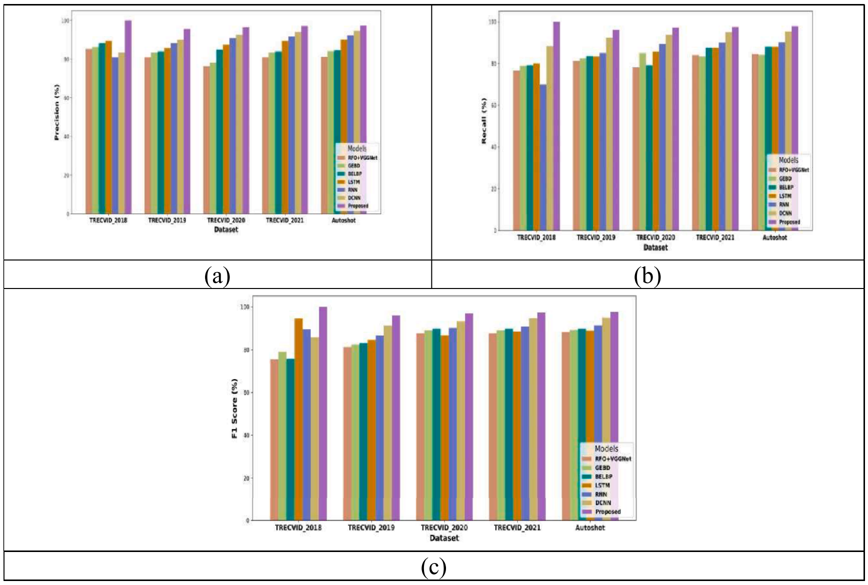
(b)

(c)

Fig. 7. Fade out transition analysis based on (a) Precision (b) Recall (c) F1-Score.

Table 2
Fade out transition analysis.

Dataset/ Methods	RFO+VGGNet	GEBD	BELBP	LSTM	RNN	DCNN	Proposed
Precision							
TRECVID_2018	86.23	88.24	89.47	85.71	90	87.88	92.31
TRECVID_2019	82.35	83.54	87.5	80	85.71	88.89	93.33
TRECVID_2020	80.95	83.33	84	83.33	87.5	89.66	92.86
TRECVID_2021	75.65	74.32	76.98	85.71	88.89	91.67	94.44
Autoshot	76.89	75.67	77.89	86.12	89.01	92.12	94.67
Recall							
TRECVID_2018	79.12	80	81.21	82.35	85.71	91.3	94.44
TRECVID_2019	87.59	88.98	89.79	78.26	83.33	90.91	94.44
TRECVID_2020	82.35	83.54	87.5	81.25	85.71	90.91	93.75
TRECVID_2021	79.12	80	81.21	84.21	87.5	92.59	95.24
Autoshot	79.56	80.34	81.56	84.67	88.12	92.23	95.01
F1-Score							
TRECVID_2018	79.12	80	81.21	82.12	85.65	90.76	94.67
TRECVID_2019	80.21	81.12	83.21	84.17	86.25	90.34	94.01
TRECVID_2020	81.56	82.12	83.12	84.67	86.21	91.67	95.12
TRECVID_2021	84.12	85.01	85.78	86.89	87.12	92.34	96.45
Autoshot	84.78	85.23	86.12	87.23	88.12	92.45	96.89



(a)

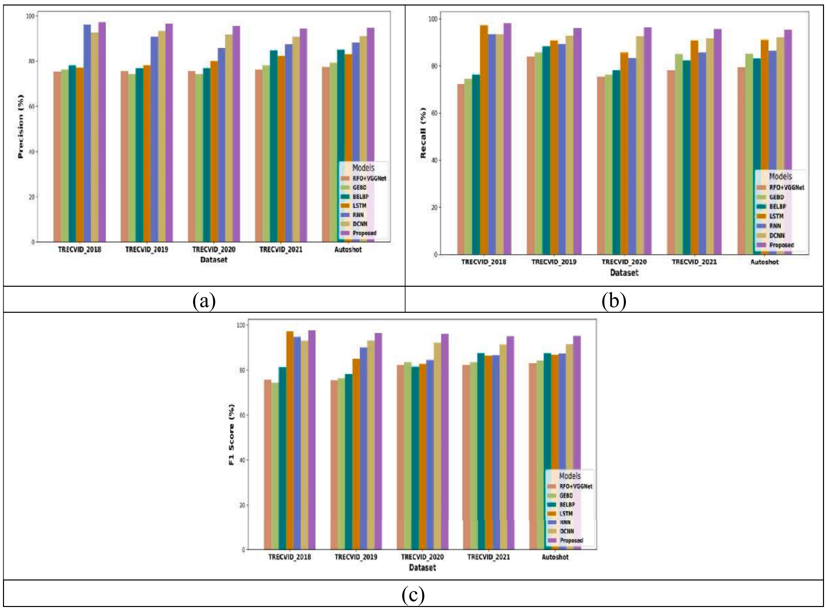
(b)

(c)

Fig. 8. Cut transition analysis based on (a) precision, (b) recall and (c) F1-Score.

Table 3
Cut transition analysis based.

Dataset/ Methods	RFO+VGGNet	GEBD	BELBP	LSTM	RNN	DCNN	Proposed
Precision							
TRECVID_2018	85.24	86.23	88.24	89.47	80.95	83.33	100
TRECVID_2019	80.95	83.33	84	85.71	88.24	90	95.65
TRECVID_2020	76.32	78.21	84.91	87.5	90.91	92.59	96.55
TRECVID_2021	80.95	83.33	84	89.47	91.67	94.12	97.06
Autoshot	81.23	84.12	84.56	90.12	92.34	94.67	97.34
Recall							
TRECVID_2018	76.58	78.84	79.12	80	70	88.24	100
TRECVID_2019	81.25	82.35	83.54	83.33	85	92.31	96.15
TRECVID_2020	78.21	84.91	79.12	85.71	89.29	93.75	97.22
TRECVID_2021	84	83.33	87.5	87.5	90	95	97.5
Autoshot	84.56	84.12	88	88	90.12	95.34	97.8
F1-Score							
TRECVID_2018	75.41	79	75.68	94.44	89.47	85.71	100
TRECVID_2019	81.21	82.35	82.98	84.51	86.57	91.13	95.9
TRECVID_2020	87.59	88.98	89.79	86.6	90.09	93.17	96.88
TRECVID_2021	87.59	88.98	89.79	88.46	90.83	94.56	97.28
Autoshot	88.21	89.12	89.78	88.89	91.22	94.89	97.57



(a)

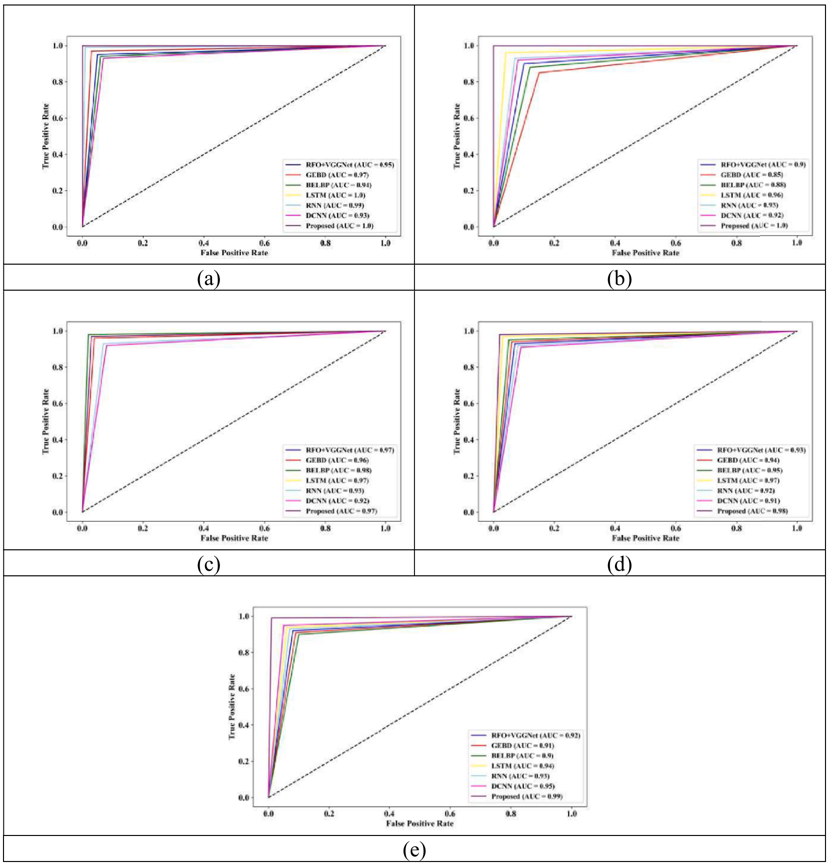
(b)

(c)

Fig. 9. Dissolve transition analysis based on (a) Precision, (b) Recall and (c) F1-Score.

Table 4
Dissolve transition analysis.

Dataset/ Methods	RFO+VGGNet	GEBD	BELBP	LSTM	RNN	DCNN	Proposed
Precision							
TRECVID_2018	75.41	76.32	78.21	77.2	96.15	92.59	97.22
TRECVID_2019	75.65	74.32	76.98	78.26	90.91	93.33	96.67
TRECVID_2020	75.65	74.32	76.98	80	85.71	91.67	95.65
TRECVID_2021	76.32	78.21	84.91	82.35	87.5	90.91	94.44
Autoshot	77.45	79.34	85.12	83.12	88.23	91.12	94.78
Recall							
TRECVID_2018	72.35	74.54	76.38	97.2	93.46	93.46	98.13
TRECVID_2019	84	85.71	88.24	90.91	89.29	92.86	96.15
TRECVID_2020	75.41	76.32	78.21	85.71	83.33	92.59	96.43
TRECVID_2021	78.21	84.91	82.35	90.91	85.71	91.67	95.65
Autoshot	79.34	85.12	83.12	91.12	86.45	92.12	95.34
F1-Score							
TRECVID_2018	75.65	74.32	81.35	97.2	94.79	93.02	97.67
TRECVID_2019	75.41	76.32	78.21	84.91	90.09	93.09	96.41
TRECVID_2020	82.35	83.54	81.52	82.76	84.5	92.13	96.04
TRECVID_2021	82.35	83.54	87.5	86.41	86.6	91.28	95.04
Autoshot	83.12	84.23	87.45	86.78	87.34	91.45	95.12



(a)

(b)

(c)

(d)

(e)

Fig. 10. ROC analysis for TRECVID (a) 2018, (b) 2019, (c) 2020 (d) 2021(e) Autoshot.

algorithm. The detailed analysis based on precision, recall, and F1-Score is presented in Table 1.

4.2.2. Analysis based on fade out transition

The fade-out transition analysis using the TRECVID 2018, 2019, 2020, 2021, and Autoshot based on various assessment measures is portrayed in Fig. 7. The Precision evaluated by the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection is 92.31 using the 2018 dataset, which is higher compared to all conventional methods. The precision metric quantifies the accuracy of the model's positive predictions. Precision measures the fraction of accurately detected shot boundaries among all predicted boundaries. Achieving the maximum level of precision in shot boundary detection is critical to ensure that the recognized shot transitions are accurate and reliable. The detailed analysis of the fade-out transition based on precision, recall, and F1-Score is presented in Table 2.

4.2.3. Analysis based on cut transition

The cut transition analysis using the TRECVID 2018, 2019, 2020, 2021, and Autoshot based on various assessment measures is portrayed in Fig. 8. The recall evaluated by the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection is 96.15 using the 2019 dataset, which is superior compared to all conventional methods. Recallis a crucial metric that measures the ability of a model to capture all actual shot boundaries. Achieving the highest recall in shot boundary detection is important to ensurethat a significant portion of true shot transitions is correctly identified. The detailed analysis of cut transition based on precision, recall, and F1-Score is presented in Table 3.

4.2.4. Analysis based on dissolve transition

The dissolve transition analysis using the TRECVID 2018, 2019, 2020, 2021, and Autoshot based on various assessment measures is portrayed in Fig. 9. The F1-Score evaluated by the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection is 96.41 using the 2019 dataset, which is superior compared to all conventional methods. The detailed analysis of dissolve transition based on precision, recall, and F1-Score is presented in Table 4.

4.3. Analysis based on ROC

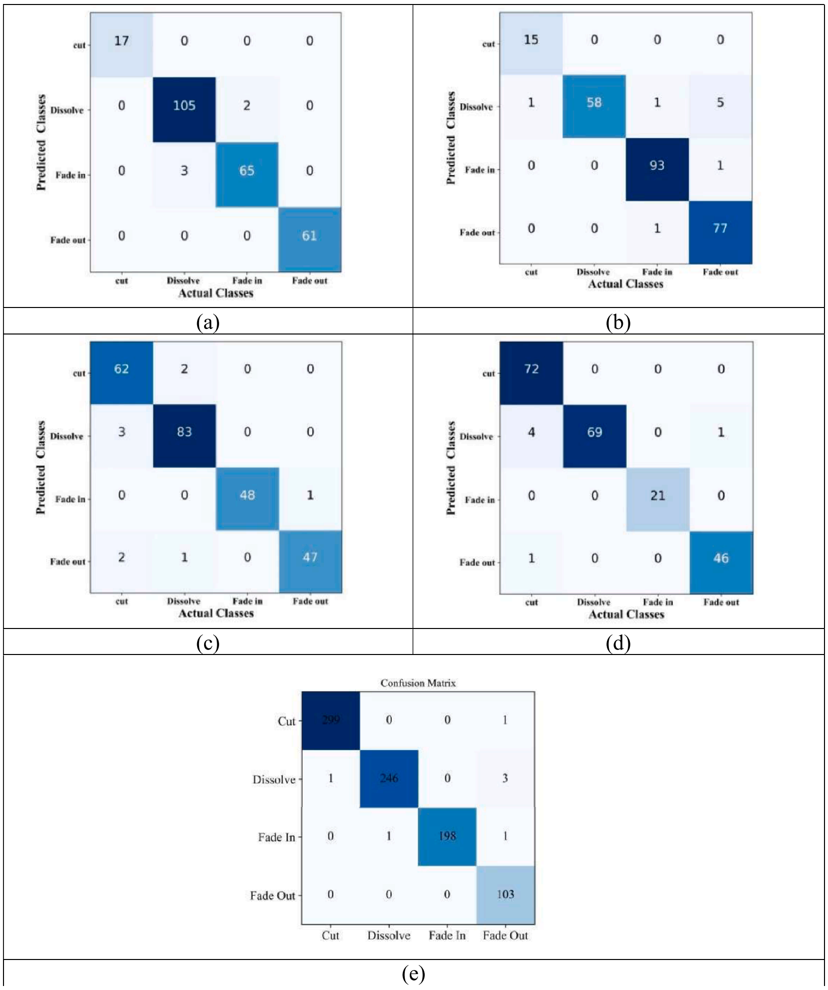
Receiver Operating Characteristic (ROC) analysis is a common method used to evaluate the performance of binary classification algorithms, including shot boundary detection in the context of video analysis. The ROC curve is a graphical representation that illustrates the trade-off between a true positive rate and a false positive rate across different threshold settings. The ROC analysis is portrayed in Fig. 10 for the TRECVID 2018, 2019, 2020, 2021, and Autoshot, respectively, wherein the proposed SSA-CAP+MSN+AtDLSTM-based shot boundary detection acquires better performance.

4.4. Confusion matrix

By examining the confusion matrix and associated metrics, the strengths and weaknesses of the proposed shot boundary detection algorithm can be analyzed, which is portrayed in Fig. 11. The analysis for the TRECVID 2018, 2019, 2020, 2021, and Autoshot with four various transitions, such as fade in, fade out, cut, and dissolve, are illustrated.

4.5. Ablation study

The ablation study of the proposed method with the absence of particular layers is made and the outcome is portrayed in Fig. 12. The analysis of the TRECVID dataset and the Autoshot dataset demonstrates the superiority of the proposed SSA-CAP+MSN+AtDLSTM. The detailed



(a)

(b)

(c)

(d)

(e)

Fig. 11. Confusion matrix for TRECVID (a) 2018 (b) 2019 (c) 2020(d) 2021 (e) Autoshot.

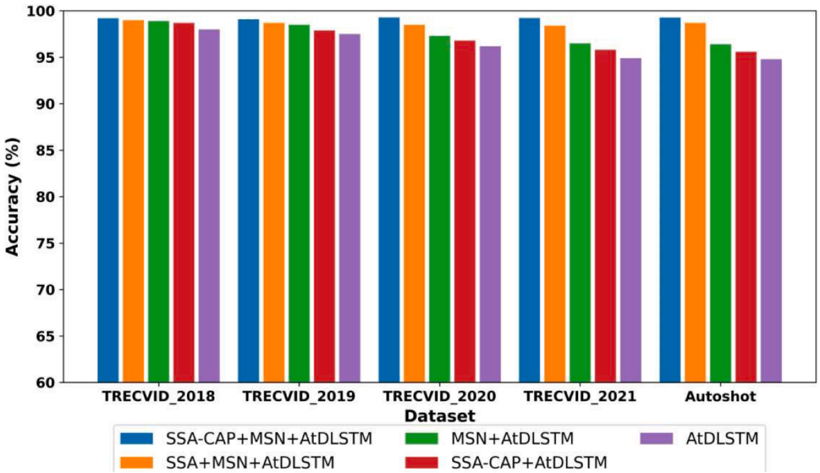


Fig. 12. Ablation study in terms of accuracy.

outcome for various datasets in terms of accuracy measure is portrayed in Table 5.

4.6. Comparative discussion

The comparative discussion of the shot boundary detection based on the best case is presented in Table 6.

Here, the analysis depicts the superiority of the proposed method with 100 % accuracy. The superiority of the proposed method derives from the design of the proposed shot boundary detection using SSA-CAP+MSN+AtDLSTM. The SSA-CAP+MSN+AtDLSTM technique has various advantages for shot boundary identification. First, Siamese Style Networks with Canonical Appearance Pooling (SSN—CAP) improve appearance feature extraction by replacing conventional pooling with

Table 5
Ablation study in terms of accuracy.

Dataset/ Methods	SSA-CAP+MSN+AtDLSTM	SSA+MSN+AtDLSTM	MSN+AtDLSTM	SSA-CAP+AtDLSTM	AtDLSTM
TRECVID 2018	99.20	99.00	98.90	98.70	98.00
TRECVID 2019	99.10	98.70	98.50	97.90	97.50
TRECVID 2020	99.30	98.50	97.30	96.80	96.20
TRECVID 2021	99.23	98.40	96.50	95.80	94.90
Autoshot	99.28	98.70	96.40	95.60	94.80

Table 6
Comparative discussion.

Methods/ Dataset	Fade In	Fade Out	Dissolve	Cut
Precision				
RFO+VGGNet	87.59	76.89	75.41	85.24
GEBD	88.98	75.67	76.32	86.23
BELBP	89.79	77.89	78.21	88.24
LSTM	92.54	86.12	77.2	89.47
RNN	94.12	89.01	96.15	80.95
DCNN	96.15	92.12	92.59	83.33
Proposed	97.22	94.67	97.22	100
Recall				
RFO+VGGNet	84.67	79.12	72.35	76.58
GEBD	84.12	80	74.54	78.84
BELBP	88.12	81.21	76.38	79.12
LSTM	91.67	84.21	97.2	80
RNN	93.67	87.5	93.46	70
DCNN	96.78	92.59	93.46	88.24
Proposed	97.89	95.24	98.13	100
F1-Score				
RFO+VGGNet	89.45	84.78	75.65	75.41
GEBD	89.14	85.23	74.32	79
BELBP	90.12	86.12	81.35	75.68
LSTM	91.12	87.23	97.2	94.44
RNN	93.12	88.12	94.79	89.47
DCNN	96.12	92.45	93.02	85.71
Proposed	97.89	96.89	97.67	100

canonical appearance pooling. SSN—CAP allows better preservation of fine-grained details within local regions, contributing to more robust feature representations. Secondly, Motion Squeeze Network (MSN) is employed to extract motion features from keyframes, providing a comprehensive understanding of dynamic variations in the video content. Thirdly, the hybridization of Dilated LSTM and attention mechanism in AtDLSTM significantly enhances shot boundary detection accuracy. The combination of temporal modelling through Dilated LSTM and attention mechanisms enables the network to capture long-range dependencies and focus on relevant regions, improving overall performance. The integrated approach leverages the strengths of appearance and motion features, making it well-suited for precise and comprehensive shot boundary detection tasks. Overall, SSA-CAP+MSN+AtDLSTM presents a sophisticated and effective solution for video analysis, offering improved capabilities in preserving details, capturing motion variations, and accurately detecting shot boundaries.

5. Conclusion

In conclusion, the SSA-CAP+MSN+AtDLSTM approach provides a significant advancement in shot boundary detection, depicting remarkable achievements in accuracy and robust feature extraction. The combination of SSN—CAP, MSN, and AtDLSTMresults in a system that excels at preserving fine-grained information while accurately recognizing diverse shot types. Achieving a perfect 100 % F1measure in shot boundary identification is a significant achievement that demonstrates the method's usefulness in video analysis. However, it is critical to recognize its limitations, which include potential computational complexity and sensitivity to environmental factors. Future efforts will focus on optimizing the method for real-time applications, addressing challenges related to video quality and scene complexity, and exploring avenues for generalization across diverse datasets. Additionally, advancements such as incorporating contextual information and exploring ensemble techniques could further enhance the method's performance.

Funding information

No funding is provided for the preparation of the manuscript.

CRediT authorship contribution statement

Swati ChaitandasHadke: Writing – original draft, Software, Resources, Project administration, Data curation, Conceptualization. **Ravi Mishra:** Writing – review & editing, Investigation, Funding acquisition. **Rushikesh Tukaramji Bankar:** Methodology, Investigation. **Sharda Amardas Chhabria:** Visualization. **Shrikant Prakash Chavate:** Validation, Supervision. **Latika Shyam Pinjarkar:** Writing – original draft, Project administration.

Declaration of competing interest

Authors do not have any conflict of interest to declare.

Data availability

The authors do not have permission to share data.

References

[1] S. Zhou, X. Wu, Y. Qi, S. Luo, X. Xie, Video shot boundary detection based on multi-level features collaboration, Signal. Image Video Process. 15 (2021) 627–635.

[2] B.S. Rashmi, H.S. Nagendraswamy, Video shot boundary detection using block based cumulative approach, Multi-Media Tools Applic. 80 (2021) 641–664.

[3] X. Li, J. Deng, Y. Fang, Few-shot object detection on remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–14.

[4] S. Chakraborty, A. Singh, D.M. Thounaojam, A novel bifold-stage shot boundary detection algorithm: invariant to motion and illumination, Vis. Comput. 38 (2) (2022) 445–456.

[5] M.Z. Shou, S.W. Lei, W. Wang, D. Ghadiyaram, M. Feiszli, Generic event boundary detection: a benchmark for event segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8075–8084.

[6] V.V. Menon, H. Amirpour, M. Ghanbari, C. Timmerer, Efficient content-adaptive feature-based shot detection for http adaptive streaming, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2174–2178.

[7] J.T. Jose, S. Rajkumar, M.R. Ghalib, A.t Shankar, P. Sharma, M.R. Khosravi, Efficient shot boundary detection with multiple visual representations, Mobile Inf. Syst. 2022 (2022).

[8] R. Mishra, Video shot boundary detection using hybrid dual tree complex wavelet transform with Walsh Hadamard transform, Multi-Media Tools Applic. 80 (18) (2021) 28109–28135.

[9] T. Kar, P. Kanungo, A gradient based dual detection model for shot boundary detection, Multi-Media Tools Applic. 82 (6) (2023) 8489–8506.

[10] K. Kanagaraj, G.G.L. Priya, Curvelet transform based feature extraction and selection for multi-media event classification, J. King Saud Univ.-Comput. Info. Sci. 34 (2) (2022) 375–383.

[11] S. Chavate, R. Mishra, A comparison of different procedures for hardware-based video shot boundary detection. Advances in Image and Data Processing Using VLSI Design, Volume 1: Smart vision Systems, IOP Publishing, Bristol, UK, 2021, p. 14. -1.

[12] S.U. Amin, M. Ullah, M. Sajjad, F.A. Cheikh, M. Hijji, A. Hijji, K. Muhammad, EADN: an efficient deep learning model for anomaly detection in videos, Mathematics 10 (9) (2022) 1555.

[13] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, D. Rueckert, Self-supervised learning for few-shot medical image segmentation, IEEe Trans. Med. ImAging 41 (7) (2022) 1837–1848.

[14] A.S. Parihar, J. Pal, I. Sharma, Multiview video summarization using video partitioning and clustering, J. Vis. Commun. Image Represent. 74 (2021) 102991.

[15] S. Kim, S. An, P. Chikontwe, S.H. Park, Bidirectional rnn-based few shot learning for 3d medical image segmentation, in: Proceedings of the AAAI conference on artificial intelligence 35, 2021, pp. 1808–1816.

[16] J. Jing, S. Liu, G. Wang, W. Zhang, C. Sun, Recent advances on image edge detection: a comprehensive review, Neurocomputing. (2022).

[17] D. Pham, M. Ha, Xiao C, Color structured light Stripe edge detection method based on generative adversarial networks, Appl. Sci. 13 (1) (2022) 198.

[18] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, R. Hamid, Shot contrastive self-supervised learning for scene boundary detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9796–9805.

[19] Z.N. Idan, S.H. Abdulhussain, B.M. Mahmmod, K.A.A. Utaibi, S.A.R. Al-Hadad, S. M. Sait, Fast shot boundary detection based on separable moments and support vector machine, IEEe Access. 9 (2021) 106412–106427.

[20] H. Kang, J. Kim, T. Kim, S.J. Kim, Uboco: unsupervised boundary contrastive learning for generic event boundary detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20073–20082.

[21] H.M. Nandini, H.K. Chethan, B.S. Rashmi, Shot based keyframe extraction using edge-LBP approach, J. King Saud Univ.-Comput. Info. Sci. 34 (7) (2022) 4537–4545.

[22] S.C. Hadke, R. Mishra, Shot boundary detection in video using dual-stage optimized VGGNet based feature fusion and classification, Multi-media Tools Applicat. (2023) 1–28.

[23] S. Chakraborty, D.M. Thounaojam, N. Sinha, A shot boundary detection technique based on visual colour information, Multimed. Tools. Appl. 80 (3) (2021) 4007–4022.

[24] T. Soucek, J. Lokoc, Transnet v2: an effective deep network architecture for fast shot transition detection, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 11218–11221.

[25] R. SavranKızıltepe, J.Q. Gan, J.J. Escobar, A novel keyframe extraction method for video classification using deep neural networks, Neural Comput. Applic. 35 (34) (2023) 24513–24524.

[26] J. Theiner, W. Gritz, E. Müller-Budack, R. Rein, D. Memmert, R. Ewerth, Extraction of positional player data from broadcast soccer videos, in: Proceedings of the IEEE/CVF Winter Conference On Applications Of Computer Vision, 2022, pp. 823–833.

[27] S. Ul Amin, Y. Kim, I. Sami, S. Park, S. Seo, An efficient attention-based strategy for anomaly detection in surveillance video, Comput. Syst. Sci. Eng. 46 (3) (2023).

[28] X. Wei, Z. Shi, T. Zhang, X. Yu, L. Xiao, Multimodal high-order relation transformer for scene boundary detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22081–22090.

[29] B.S. Kapre, A.M. Rajurkarb, Key-frame extraction based video watermarking using speeded up robust features and discrete cosine transform, Comput. Sci. Info. Techn. 4 (1) (2023) 85–94.