Multimedia Tools and Applications (2023) 82:23037–23054
https://doi.org/10.1007/s11042-023-14882-4

Visual significance model based temporal signature for video shot boundary detection

Sasithradevi A[1] · S. Mohamed Mansoor Roomi[2] · P. Nirmala[3]

Received: 22 August 2021 / Revised: 5 February 2023 / Accepted: 6 February 2023 /
Published online: 24 February 2023

Abstract

Video shot boundary detection (VSBD) is the fundamental step for video processing algorithms. The goal of any VSBD algorithm is to detect the transitions (abrupt or subtle) in the given video precisely. In this paper, a visual significance model that is suitable for describing transitions in video is introduced. The proposed visual significance model is composed of parameters like color, texture, edge, motion and focus computed over the frames in the video. Once the visually significant region is identified in each frame of the video, the temporal signature is generated through the dissimilarity measure of the visual significance model. The temporal signature is further examined using standard Random Vector Functional Link (RVFL) networks for categorizing the transitions as Abrupt Transitions (AT), Subtle Transitions (ST) and No Transitions (NT). To validate the performance of the proposed visual significance model based VSBD Framework, it is evaluated on benchmarks like VIDEOSEG2004 and TRECVID2001 to detect and categorize the transitions. Comparison of F1-Score measure with prominent early methods reveals that the proposed framework is a promising model for detecting the transitions in videos even in the presence of varying illumination conditions, fast camera and object motion.

Keywords Video transitions · Visual significance · Random vector functional link network · F1-score

✉ Sasithradevi A
sasithradevi.a@vit.ac.in

S. Mohamed Mansoor Roomi
smmroomi@tce.edu

P. Nirmala
nirmalavp@gmail.com

[1] Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, India

[2] Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Madurai, India

[3] School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

1 Introduction

In the era of large-scale data, video plays a vital role in many applications like distance learning, entertainment, infotainment and so on. This paves the way for research in video processing which provides many practical solutions for human action recognition, video analysis and video browsing. The commonly accepted initial step for these kinds of applications is video shot boundary detection. In general, VSBD algorithms are very much useful in identifying the transitions in video. Video transitions are occurring in videos due to the editing technique that is used to connect the jump between story lines. Usually, transition occurs between different shots. A shot is an undisturbed scene captured continuously on a single camera. A sudden swift from one shot to another shot is called an abrupt transition. A gradual movement from one shot to next shot can be referred to as gradual transition. Given a video, the process of detecting and categorizing the transitions as NT, AT and GT is referred as video shot boundary detection (VSBD) or video temporal segmentation.

The fundamental steps in the VSBD algorithm are: i) Feature representation ii)Temporal signature generation iii)Categorizing the transitions in videos [9]. The main focus of this VSBD work is on uncompressed video domains. Any frames of the uncompressed video will have sufficient content within it. The frame content can be represented by features like color, texture, edge and motion. Characterizing these features is the initial step called feature representation followed by construction of temporal signature. The main idea behind the construction of temporal signature is that large discontinuities of the temporal signal signify the presence of transitions in corresponding frames. These discontinuities are useful in categorizing the transitions as abrupt or subtle. Even though a large number of researches are being done in VSBD, these algorithms are lagging in performance due to issues like illumination effects, camera motion, panning, zooming, object motion, etc. An effective VSBD algorithm is one whose performance is not affected despite the influence of aforesaid issues. Towards this direction, we propose a visual significance model-based video shot boundary detection framework for detecting transitions even under the presence of aforementioned disturbances. We represent each frame in the video using visual significance model which comprises five attributes called color, texture, focus, edge and motion feature. It is the feature representation for the frames in the video. The next step is to construct the temporal signature through any of the suitable dissimilarity measures. As a final step, based on the discontinuity values, standard RVFL is trained and the learned features are used to classify the transitions on test videos. In [16], visual significance is used to improve the contrast of the blurred image. To the best of our knowledge, none of the earlier works has used the novel visual significance model to detect the different transitions in videos. The rest of the paper is organized as follows: Section 2 provides the review on the methodologies available in the literature on shot boundary detection. Section 3 describes the proposed visual significance model and framework for VSBD. Section 4 discusses the validation of the proposed framework through experimental results. Section 5 concludes the paper.

2 Background

In recent years, several researches are ongoing to develop efficient algorithms that can detect the shot boundaries exactly. The methodologies proposed so far in shot boundary detection can be broadly grouped as: Deep learning methodologies and Classical methodologies**.**

2.1 Deep learning methodologies

Recently deep learning methodologies have gained attention from researchers in the field of computer vision. Among the deep learning methodologies prevailing so far, Convolutional Neural Network (CNN) is the common technique used in representing the high level characteristics of images and video frames [13]. Jingwei Xu et al. [28], developed an algorithm for detecting shot boundaries which is based on convolutional neural networks (CNN). The authors utilized CNNs for learning low level features from low-stage layers and the higher-level features are obtained by composing lower-level ones. The high-level features are powerful and suitable for distinguishing transitions. Candidate segment selection is used for locating the positions of shot boundaries roughly and it eliminates the non-boundary frames. This technique detects both abrupt transition (AT) and gradual transition (GT). Another CNN based method which is also capable of detecting both the type of transitions was proposed by Tong et al. [26]. Initial preprocessing stage includes CNN that generates probabilities for 1000 classes. Among these, five classes with highest probabilities are chosen as high level frame features. The shot detection accuracy is low due to similar background scenarios even in gradual transition. Later CNN is extended as spatio-temporal CNN for classifying the frames in videos of fixed length as NT, AT or GT. The high-level interpretable features are learned through 3D-CNN and classified by the SVM classifier. Though the final result is pruned by post processing step to reduce the false alarms in gradual transition detection, 3D-CNN is more computationally complex than 2D-CNN. Moreover, 3D-CNN is not scale invariant. A fully convolutional neural network is also used for video shot boundary detection [6]. The authors augmented the samples in the dataset upto one million frames. The major flaws in this architecture are its difficulty in detecting transitions under scenarios like partial scene change, untrained samples and rapid moving scenes with motion blur. A three stage CNN framework is proposed by Shitao Tang et al. [23] in which the initial stage is the preprocessing step, followed by 2D-CNN learning step to relate the similarity between the adjacent frames to detect cut transitions and final stage includes a 3D-CNN to detect the subtle transitions. In [5], video frames were represented as tensors. The dynamic tensor analysis method was used for building tensor models. The developed algorithm follows efficient operation when compared to the standard tensor decomposition method. The authors stated that auxiliary research will be oriented in the direction of enhancement of accuracy and speed of the technique by including more fitness functions and by developing the parallel tensor decomposition methods. Lifang Wu et al. [27] have reported that the fusion of color histogram and learned deep features works well in the detection of abrupt transition. A 3D-convolutional neural network has been used to detect the gradual shots. In [20], the authors proved that a scalable deep architecture can be used for the detection of shots in the video. A simplest architecture based on TransNet and a dilated convolutional neural network has been used to detect the transitions. Also, its upgraded version called TransNetV2 [19] has been introduced to detect the shot transitions in video. TransNetV2 relies on dilated version of CNN, factorization of convolutional kernels and similarity score between frames.

The methods proposed so far in the Literature to detect the shot boundaries based on CNN, use the CNN architecture for feature learning purpose alone and not for classification. The prominent growth of deep learning methodologies in object detection and image classification has attracted the researchers in video processing. One of the major drawbacks of deep learning-based video processing is the need for large amounts of annotated data. Even though CNN is capable of generating synthetic shots, it cannot compete with the real data which contains shots within the same scenes. On the other hand, numerous approaches have been proposed for detecting shot boundaries using traditional methods.

2.2 Traditional methodologies

Dalton Meitei Thounaojam et al. [24], developed a method which is based on Genetic Algorithm and Fuzzy Logic. The membership functions of the fuzzy system were computed using the Genetic algorithm by taking the observed values for shot boundaries. The classification of shot transition types was done by the fuzzy system. Normalized color histogram difference was utilized for extracting features and to find the variation between the consecutive frames. Chongke et al. [3], proposed a technique for shot boundary detection which is based on dynamic mode decomposition (DMD). DMD is used for detecting shot boundaries from three different shot transitions in hard cuts, dissolves and fades. The authors also discussed the limitation of the technique that the threshold should be set through by computing the mutual analysis. It can be resolved by adaptive threshold computation. The technique proposed in [15], detected the edges using the Sobel operator to obtain gradient frame from gray scale frame. The crisp data were converted into Fuzzy data and the authors extract block based mean cumulative sum histogram from every edge gradient fuzzified frame. Relative standard deviation is utilized for detecting the abrupt and gradual shots in the video. Sawitchaya Tippaya et al. [25], developed a transition detection methodology which is based on multi modal visual features. In this technique, candidate segments selected based on the inter frame distance and threshold. This approach is used to identify the position of shot boundaries and neglecting the non-boundary video frames. Kar and Kanungo [7], detected the shot boundaries in a video sequence in the presence of both motion and illumination variation. The authors utilized the advantages of Hilbert transform as well as DWT for obtaining a new feature space. SSIM metric is used to detect the shot boundaries. This method detects video shot boundaries by extracting the SIFT key points and applying point distribution histogram. It detects shot boundaries under various levels of illumination with different motion effects and camera operations.

Sadiq H. Abdulhussain et al. [1], implemented an algorithm for shot boundary detection (SBD). Features were derived from the orthogonal transform domain for detecting the hard transitions in the video sequences. Smoothing frames were used for eliminating the noise. The limitation is that the computational cost was increased when convolving multiple image kernels with each video frame and the algorithm is not generalized for detecting the different transitions. Also, a dual framework was proposed in [8] to detect shot transitions based on two categories of gradient features namely gradient magnitude and orientation. The authors analysed the performance of dual framework under unsupervised learning process. In [4], the researchers used ant lion optimization algorithm to optimize the weights assigned in neural network for classifying the transitions in video. Convolutional neural networks (CNN) are used to extract the frame features in parallel fashion, the extracted CNN features are fed to LSTM [2]and the transitions are categorized based on Euclidean distance measure.

Referring to the literature, it is evident that representation of the key content of the frames plays a major role in the decision of the categories of transitions as NT, GT or AT available in videos. Video is a combination of three axes namely horizontal, vertical and time. Given a video V, t^{th} frame can be expressed mathematically as F_t. The initial step in VSBD called content representation projects the frames from image space onto feature space. Now the feature space is accountable for the accurate results in classifying the transitions in the scenes.

The feature space must be able to decide the transitions appropriately even in the presence of fast camera and object motion, illumination variations. Though most of the recently proposed algorithms for VSBD provide satisfactory results, these methods fail to ensure the trade-off between sensitivity and invariance. Also, the trade-off between complexity and accuracy is another vital parameter which must be taken into account for any efficient VSBD algorithm. Also, most of the algorithms focus only on detecting the cut transitions rather than detecting both transitions. Hence, the proposed method attempts to tackle the aforementioned issues based on the fact that visual significance model captures the transitions as the human cognition intelligence because it involves a reasonable factor called focus. Hence, this paper proposes a novel visual significance model that takes vital features like color, texture, shape, focus and motion into account for modeling the frame content. The key contributions of this paper are:

- Introducing a novel visual significance model for representing the frame content in videos
- Generating temporal signature by observing the dissimilarity of visual significance model over the frames
- Classification of the scene transitions in the videos as NT, GT or AT
- Evaluating the proposed methodology on benchmarks for detecting the abrupt and subtle transitions in video.

3 Materials and methods

This section articulates the VSBD problem, describes the proposed framework and addresses the visual significance model as a tool for detecting shots in video.

3.1 Problem statement and hypothesis

Early solutions for VSBD based on traditional features are blind in the sense that it uses some set of the low-level attributes without any prior knowledge about the content in the frame. Hence, our proposed framework put forward a visual significance model which includes an additional attribute called ‘Focus’ in addition to color, texture, shape and motion. Falling on the fact that focus changes as the transition occurs, the proposed framework will be efficient in identifying transitions in the scenes.

Consider the training domain *‘Tr_d’* and video ‘*V*’*, where V€* T_d, the corresponding feature space *‘F’* is defined through the visual significance model ‘V_S’ for accomplishing VSBD task *Tr_d’*. In the testing domain *‘Ts_d’*, the feature space $\widehat{F}$ defined through ‘V_S’ should be able to achieve high precision results on *‘Ts_d’* . To achieve the task, both F *and* $\widehat{F}$ must adhere to common space, in-order to attain that, F *and* $\widehat{F}$ are mapped to temporal signature space S *and* $\widehat{S}$. Hence, the prime goal of this work is to complete the task of *‘Ts_d’* using S_L *and* $\widehat{S}_L$, learned temporal signatures using standard RVFL network.

3.2 Methodology

A video shot boundary detection framework, based on visual significance model, has been proposed to classify the visual link between scenes as any of the three categories viz., No transition or abrupt transition or subtle transition. The work flow of the proposed framework has been shown in Fig. 1. The proposed framework consists of training and testing modules. It is evident that, both the training and the testing modules share common steps which includes four vital building blocks of the proposed framework:1) Visual significant modeling 'V$_S$' of the all frames in the given video, 2)Temporal signature is constructed based on the dissimilarity between 'V$_S$' of the consecutive frames in the video, 3) A simple statistical threshold approach is employed on S_L *and* $\widehat{S}_L$ to select the salient candidates (SC) for shot boundary detection 4) SC is further learned by standard RVFL network to determine the type of transitions.

3.2.1 Visual significant modeling of the frames

Given a video *'V'*, which is fragmented into a number of frames *'F_n'* and the tth frame is represented by *'F_t'*. The frame is partitioned into a number of segments/blocks which is further divided into small windows. To maintain the same aspect ratio, the common size for horizontal and vertical partitions is sustained. The model utilizes statistical parameters like mean pixel intensity, mean absolute deviation, edge (horizontal and vertical) count and magnitude to define the vital influencing factors of a visual significant model like color, texture, shape and focus. The visual significant model for a segment in *'F_t'* is represented by:

$$V_s = F_s[\alpha C_s + (1 - \alpha) T_s] \tag{1}$$

Where α is the weighting factor, F_s is the focus parameter, C_s is the color parameter and T_s is the texture parameter. The mathematical expressions to calculate color, texture and focus parameter as well as the algorithm to model a frame through visual significance model is described in algorithm 1.

Algorithm 1 Computation of T_s, F_s, *and* C_s.

```
Data:$F_{W(k,l)}$; Result: $V_s$[]
Initialize:$V_s = 0$
for k=0;k<M;k++ do
    for l=0;l<N;l++ do
        mean of window : $\mu_{F_{W(k,l)}} = \frac{1}{M \times N} \sum_1^M \sum_1^N F_{W(k,l)}$
        deviation of window : $\sigma_{F_{W(k,l)}} = \frac{1}{M \times N} \sum_1^M \sum_1^N \left| F_{W(k,l)} - \mu_{F_{W(k,l)}} \right|$
        segmental mean: $\mu_{F_S} = \frac{1}{k} \sum_1^k \mu_{F_W}$
        segmental deviation of mean: $\sigma_{\mu_{F_s}} = \frac{1}{k} \sum_1^k \left| \mu_{F_W} - \mu_{F_S} \right|$
        segmental color parameter: $C_s = \mu_{F_S} + \sigma_{F_s}$

        Focus parameter: $F_s = \mu_{E_{F_{W(k,l)}}}$
            $\left| E_{F_{W(k,l)}} \right| = \frac{1}{\sigma_{F_W}} \left( \frac{|E_H(F_W)| + |E_V(F_W)|}{\#|E_H(F_W)| + \#|E_V(F_W)|} \right)$
        Texture parameter: $T_s = \frac{1}{\sigma_{\sigma_{F_S}}} + \frac{1}{\min(\sigma_{\#|E_H(F_W)|}, \sigma_{\#|E_V(F_W)|})}$
        $V_s = F_s[\alpha C_s + (1 - \alpha) T_s]$
    end
end
```

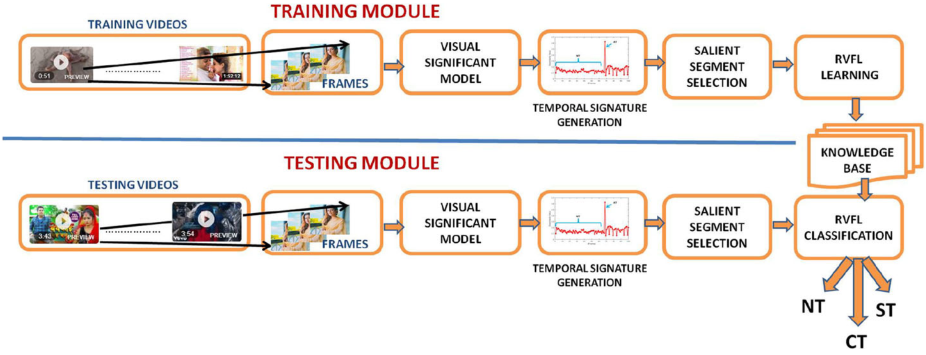


Fig. 1 Proposed Framework for VSBD based on visual significance model

After the computation of all these parameters of the window, mean and mean absolute deviation over the entire segment is computed as listed in algorithm 1. The first term of color parameter C_s represents the average local contrast measure within each window. The second term of C_s denotes the contrast of the segment globally. Referring Eq. 1, focus is given much importance as it figures out the intent (transitions) of the video viewer. The focus measure of the segment in frame is computed through normalized mean of edge magnitude over the entire segment in the frame. To obtain the texture component in the frame by ensuring complexity reduction, texture of a segment is calculated by the sum of reciprocal of deviation of segments of window deviation and reciprocal of minimum of segmental deviation of counts of horizontal and vertical edges. The Edge parameter $\left|E_{F_{W(k,l)}}\right|$ plays a vital factor in computing focus measure. The horizontal edge count is calculated by,

$$\#|E_H(F_W)| = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{l=1}^{N}\rho(|W(k,l)-W(k,l+1)|,\delta(W)) \tag{2}$$

The vertical edge count is calculated using,

$$\#|E_v(F_W)| = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{l=1}^{N}\rho(|W(k,l)-W(k+1,l)|,\delta(W)) \tag{3}$$

(Where) $\rho(i,j) = \begin{cases} 1, & \mathit{if}\ i > j \\ 0, & \mathit{otherwise} \end{cases}$

The magnitude of horizontal edges is calculated as,

$$|E_H(F_W)| = \sum_{k=1}^{M}\sum_{l=1}^{N}\gamma(|W(k,l)-W(k,l+1)|-\delta(W)) \tag{4}$$

$$|E_V(F_W)| = \sum_{k=1}^{M}\sum_{l=1}^{N}\gamma(|W(k,l)-W(k+1,l)|-\delta(W)) \tag{5}$$

(Where) $\delta(W) = \widetilde{\sigma(W)}$

The reason behind computed texture (refer algorithm 1) through this method is that repetitive textures will always result in low deviations only which results in high texture value of windows. After computing the color, texture, focus and edge parameters in a frame, all these parameters are normalized to get the relative nature of these segments within the frame.

3.2.2 Construction of temporal signature

A video with consecutive N frames can be represented as $\{F_1, F_2 \ldots F_N\}$. The visual significant model of these frames is denoted as $V_s = \left\{F_1^S, F_2^S \ldots.. F_N^S\right\}$. The absolute values of difference between the two consecutive frames $\{F_{N\ -\ 1},\ F_N\}$ projected to the visual significance model $\left\{F_{N-1}^S, F_N^S\right\}$ have been utilized for generating the temporal signature. The absolute difference is very much useful in predicting the visual discrepancies and video rhythms. Hence, generation of temporal signature is an important step in the proposed VSBD task. The temporal signature for video is denoted as,

$$\boldsymbol{T}_V\left(F_i, F_{i+j}\right) = \sum \left| F_i^S - F_{i+j}^S \right| \tag{6}$$

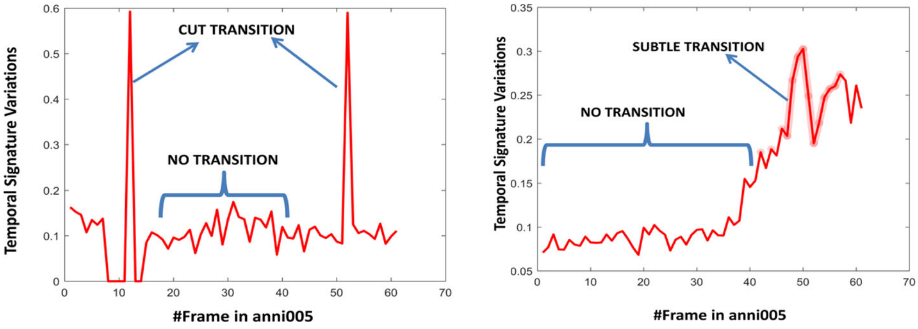
Where i and j denotes the consecutive frame number. Temporal signature computed using (7) finds the sum of absolute difference between ith and (i + j)th frames. In order to avoid scaling effects between frames, the normalized temporal signature is needed and given by,

$$T_V\left(F_i, F_{i+j}\right) = \boldsymbol{T}_V\left(F_i, F_{i+j}\right)/\max(\boldsymbol{T}_v) \tag{7}$$

The samples of temporal signature corresponding to categories like NT, CT and ST is shown in Fig. 2. The temporal signature is generated using (8) for the video named 'Anni005' of TRECVID2001 dataset using MATLAB platform. Figure 2a shows the temporal signature for abrupt and no transitions. The sharp spikes in the signature denotes the abrupt transition, whereas temporal signature of gradual transition resembles bell shape (refer Fig. 2b).

3.2.3 Selection of salient candidates for training purpose

The next step in the proposed framework is selection of salient temporal signature segments which are appropriate for training. These segments provide a better understanding about the transition categories towards the learning algorithm called Random Vector Functional Link Network. The segment selection procedure follows a sliding window approach. The window is designated as $W \in \{t,\ \ldots.\ t + n\}$, n denotes the skipping size of the window and it is set as 25, general frame rate of the video. The algorithm iterates throughout the entire sample N_s of the



(a) AT & NT detected by proposed work

(b) GT & NT detected by proposed work

Fig. 2 Temporal signature generated by proposed method for NT, AT and GT

temporal signature for jth video to collect the salient segment which has transitions contained within it. The algorithm is designed as follows:

Algorithm 2 Selection scheme for obtaining salient candidates from T$_v$ for RVFL learning

Input: $T_v \leftarrow$ Temporal signature of a given video with $T_v\{N_j^i\}$ samples. N_j^i denotes the ith sample of the temporal signature in jth video.

Output: $[T_v\{N_j^l\}, l < i]$←Salient segments of the temporal signature of jth video.

```
Begin
    Initial: j←1, $N_j^i$←1
    While $|N_j^i| \geq 1$ do
        Compute:
            $\nabla^i = |[T_v\{N_j^i\} - T_v\{N_j^{i+1}\}]|$
            If $\nabla^i < 0.2, T_v\{N_j^i\} \in NT$
                Else $T_v\{N_j^i\} \rightarrow T_v\{N_j^k\}$
            end
        end
            i←i+1
    end
    j←j+1
end
```

As elaborated in algorithm 2, the gradient between the temporal signature value of the adjacent frames determines whether the segment is salient or not. Salient segment selection is a preprocessing step for RVFL learning to reduce the irrelevant test samples apart from the abrupt transition and subtle transition categories. Thanks to the salient segments, the computational overhead on the testing module is highly reduced.

3.2.4 RVFL learning for classifying transitions

The proposed framework puts forward an unbiased training set with appropriate positive and negative candidates. To ensure fair distribution of samples in all categories, the training samples are selected manually. Manual selection will maintain a non-overlapping quantity of samples in all categories like no transition, cut transition and subtle transition. Due to the unbiased training samples, the complexities like redundant training, computational and memory overhead is alleviated in the proposed VSBD framework. This unbiased training incorporated in this framework is inspired by [11, 29].

The computational overhead on the VSBD framework is further reduced by the righteous nature of RVFL- like randomization on neural networks that leads to minimal training time. Also, RVFL transforms supervised learning into a linear problem. The unbiased training video samples were uncompressed, modeled by visual significance, transformed to temporal signature and the salient temporal signature segments are used for the purpose of learning in RVFL network. The learning is processed through 100 hidden neurons. The learned knowledge is saved as feature database to categorize the test video samples.

4 Experimental results and discussion

In this section, we present experimental validation on different benchmarks for the purpose of evaluating the proposed visual significance model-based framework for shot boundary detection.

4.1 Evaluation protocols

All the videos in the benchmark dataset are uncompressed and all the frames are transformed into a uniform size of 128 × 128 to maintain the common baselines. All the extensive experimentations are carried out on the MATLAB platform on a PC of i3 core CPU with 16GB RAM. In-order to compare the proposed methodology with competing techniques, it is mandatory to use similar performance evaluation metrics. Hence, we use the commonly used metrics like Precision, Recall and F-Score for evaluation and comparison. Precision can be formulated mathematically as

$$Precision\ (P) = \frac{\#correctly\ detected\ transitions}{\#Total\ \ detected\ \ transitions} \times 100 \tag{8}$$

The metric called precision reveals the ability of the technique to detect the transitions correctly in the given video. Another metric called recall relates the correctly detected transitions to the total transitions available in the ground truth. Recall measures the false alarm provided by the algorithm and can be written as:

$$Recall\ (R) = \frac{\#correctly\ detected\ transitions}{\#Total\ \ transitions\ in\ ground\ truth} \times 100 \tag{9}$$

F_1 measure is calculated from the harmonic average of Precision and Recall. Any VSBD algorithm can be validated by F_1 measure which depicts the detection rate of the algorithm. The F_1 score can be defined as:

$$F_1(F) = \frac{2 \times precision \times Recall}{Precision + Recall} \times 100 \tag{10}$$

4.2 Performance evaluation on dataset 1

The commonly used Benchmark dataset for the evaluation of the VSBD framework is TRECVID2001 dataset (http://trecvid.nist.gov/).

The dataset comprises totally eight video data and few sample frames are shown in Fig. 3. The other details of the videos in TRECVID2001 dataset like Frame resolution, duration, number of frames and additional details about transitions are briefed in Table 1. As seen from Table 1, TRECVID2001 dataset consists of eight videos of varying frame resolution and frame numbers.

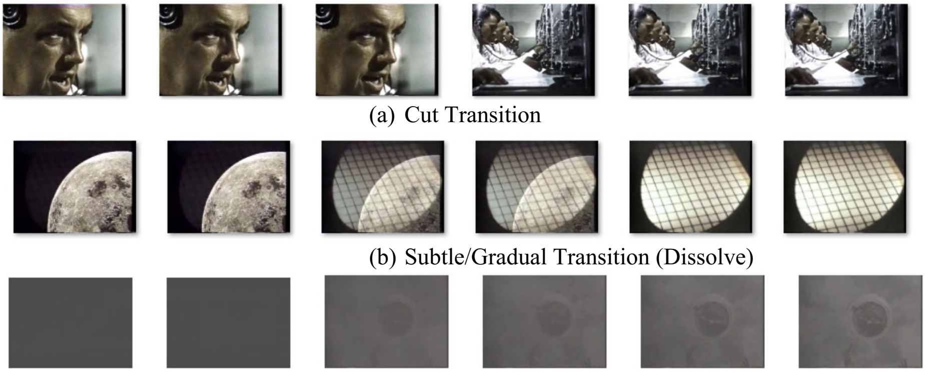
The proposed visual significance based VSBD framework is evaluated on TRECVID2001 to judge the performance in categorizing the transitions present in the video.

The entire responsibility of the proposed VSBD framework lies on salient temporal signature in the sense that the RVFL will learn and classify the information packed in the salient temporal signature. The performance evaluation of the proposed framework and the comparison with other existing solutions is enlisted in Table 2 (bold entries for higher values). It is evident from Table 2 that the proposed framework is an efficient model for characterizing the transitions present in the benchmark videos. Even though the POCS method shows fairly competing performance than the proposed method, it works on pyramid levels which make it computationally complex. The running time for detecting the transitions in video using the proposed framework is 23.5 milliseconds, whereas the POCS method takes 54.9 milliseconds.

Table 3 (bold entries for higher values) shows the performance comparison of the proposed framework in detecting the subtle transitions. Detecting subtle transitions is difficult because all these effects are introduced with the help of editing software. The process becomes even more tedious if it also includes subject and camera motion too. Since the subtle transition comprises dissolve, fade and wipe effects, all these corresponding temporal patterns will also vary from each other. As enumerated in Table 3, the proposed framework shows astounding performance in detecting the subtle transitions. The average F1 score for the proposed framework is 93.04% with a minimal deviation of 5.4. Our method has achieved a better score than existing approaches like POCS method −91.6%, Geometric method- 89.6%, WHT Method - 84.9%.

4.3 Performance evaluation on Dataset2

The proposed VSBD framework is also evaluated on the VIDEOSEG2004 dataset [29] comprising 10 videos, all the videos have different resolutions and are collected from different genres. We have chosen this dataset to prove that our method also works better in detecting transitions in all kinds of genres like horror, action, commercial, etc. TRECVID2001 dataset


(a) Cut Transition

(b) Subtle/Gradual Transition (Dissolve)

(c) Subtle/Gradual Transition (Fade)

Fig. 3 Sample Transition frames from TRECVID2001

Table 1 Details about TRECVID2001 dataset

Video	Anni005	Anni009	Nad 31	Nad 33	Nad 53	Nad 57	Bor 03	Bor 08	Total
Duration	06:19	06:50	29:08	27:39	14:31	06:57	26:56	28:07	144:67
#Frames	11,363	12,306	52,395	49,734	26,115	12,510	48,450	50,568	263,441
Resolution	240×320	240×320	240×352	240×352	240×352	240×352	240×352	240×352	–
#AT	38	38	187	189	83	45	231	380	1191
#GT	27	65	55	26	75	31	11	151	441
#OT	65	103	242	215	158	76	242	531	1632

consists of only documentary videos. Further details like number of frames, cut transitions, duration which are related to the dataset is enumerated in Table 4.

Table 5 shows the performance of the proposed algorithm on VIDEOSEG2004 dataset. As enlisted in Table 5, our method shows a fair performance in detecting the cut transitions in 4 videos. In these 4 videos, our algorithm detects all the cut transitions present without any missed and false alarms. Our method shows a good average F1 score of 97.69%. In order to prove the efficacy of our method it is mandatory to compare our method with the competing

Table 2 Performance evaluation of proposed framework on TRECVID2001 and comparison with state of the art approaches

Video	PROPOSED			POCS METHOD [17]			Geometric Method [11]			WHT Method [9]		
	P	R	F	P	R	F	P	R	F	P	R	F
Anni005	97.44	100	**98.7**	97.4	100	98.7	**97.5**	**100**	**98.7**	94.9	97.4	96.1
Anni009	97.37	100	98.67	97.4	100	**98.7**	97.4	97.4	97.4	86.5	82.1	84.2
Nad 31	99.47	93.5	96.39	99.5	98.9	**99.2**	97.2	93.5	95.6	96.8	94.7	95.7
Nad 33	98.94	99.4	**99.17**	98.4	99.4	98.9	95.4	99.4	97.3	94.7	91.8	93.2
Nad 53	100	100	**100**	98.8	100	99.4	**100**	**100**	**100**	90.4	75.0	82.0
Nad 57	95.45	100	97.67	100	100	**100**	88.3	100	93.8	93.5	95.6	94.5
Bor 03	96.20	97	96.60	98.7	97.0	**97.9**	96.2	**99.1**	97.6	98.7	97.0	97.9
Bor 08	94.33	94.3	**94.33**	97.2	93.1	**95.1**	**97.2**	93.1	**95.1**	96.1	93.1	94.6
Average	97.40	98.0	97.69	**98.4**	**98.6**	**98.5**	96.1	97.8	96.9	93.9	90.8	92.3
Std	2.00	2.74	1.84	1.03	2.43	1.50	3.44	2.92	2.02	3.87	8.04	5.86

Table 3 Performance comparison of proposed VSBD Framework with existing solutions for GT detection

Video	Proposed			POCS METHOD [17]			Geometric Method [11]			WHT Method [9]		
	P	R	F	P	R	F	P	R	F	P	R	F
Anni005	96.43	100	**98.18**	93.1	**100**	96.4	95.1	85.2	89.9	**97.9**	96.0	96.9
Anni009	87.30	84.62	85.94	85.5	81.5	83.5	**89.3**	**96.3**	**92.7**	88.7	85.9	87.3
Nad31	98.15	96.36	97.25	**98.2**	**98.2**	**98.2**	88.4	74.1	80.6	76.8	78.2	77.5
Nad33	**87.50**	80.77	**84.00**	84.0	**80.8**	82.4	82.2	79.4	80.8	80.0	76.9	78.4
Nad53	96.10	**98.67**	**97.37**	96.1	98.7	97.4	**100**	89.2	94.3	88.5	92.0	90.2
Nad57	96.77	**96.77**	**96.77**	90.6	93.5	92.1	95.1	92.5	93.8	93.8	93.8	93.8
Bor03	**90.91**	**90.91**	**90.91**	90.9	90.9	90.9	85.5	**100**	**92.2**	70.0	63.6	66.7
Bor08	**95.86**	92.05	93.92	93.8	90.1	91.9	94.3	**97.3**	**95.8**	88.1	88.1	88.1
Average	**93.63**	**92.52**	**93.04**	91.5	91.7	91.6	92.0	89.2	89.6	85.5	84.3	84.9
Std	**4.38**	**6.87**	**5.53**	4.9	7.5	6.0	5.9	9.1	6.0	9.2	10.8	10.0

Table 4 Details of Videoseg2004 Dataset

Video	Genre	Duration (MM:SS)	Frame Size	#Frames	#AT
D1	Cartoon	00:21	144×192	649	7
D2	Action	00:36	144× 32	957	8
D3	Horror	00:53	288×384	1618	54
D4	Drama	01:45	272× 336	2630	34
D5	Science Fiction	00:17	288×384	535	30
D6	Commercial	00:07	112× 160	235	0
D7	Commercial	00:16	288×384	499	18
D8	Comedy	03:25	240×352	5132	38
D9	News	00:15	288×384	478	4
D10	Trailer	00:36	180×240	871	87
Total	–	08:31	–	13,604	280

approaches and it is shown in Table 6 (bold entries for higher values). Figure 4 shows the visual comparison of performance of our method along with existing solutions in the literature for video shot boundary detection. As illustrated in Fig. 4, our method outperforms others in the state of art for detecting the transitions in VIDEOSEG2004. This performance is achieved by the proposed method because it involves visual significance model based on various features like color, texture, edge and mainly the vital parameter called focus.

Apart from meaningful F1score, another goal of our proposed method is to maintain the trade off between good F1 score and minimal time overhead in detecting the overall transitions in the given video. For the sake of proving the reality, we have compared the run time of our algorithm with other existing solutions.

The run time taken by the proposed visual significance model in processing a frame in the video as well as by other existing solutions for VSBD like NSCT+Threshold, WHT, NSCT+SVM, NSCT+SVM is also reported in Table 7. As shown in Table 7, our method takes minimal time compared to other techniques in state-of-the-art. Hence, the tradeoff between time complexity and good F1 score measure is managed well by the proposed method. The transitions that were correctly detected by the proposed algorithm are shown in Fig. 5. The scenario in the first row which is a segment from TRECVID2001 database is difficult for any algorithm to detect but ours had detected it perfectly owing to the visual significance model. Some of the missed detections are also shown in Fig. 6. These transitions are introduced through the editing effects which involves very slow object motion that disappears gradually.

Table 5 Performance of proposed VSBD framework on VIDEOSEG2004 Dataset

Video	P	R	F
D1	100.00	100.00	100.00
D2	88.89	100.00	94.12
D3	100.00	100.00	100.00
D4	97.14	100.00	98.55
D5	93.55	96.67	95.08
D6	100.00	100.00	100.00
D7	100.00	89.00	94.12
D8	95.00	97.37	96.17
D9	100.00	100.00	100.00
D10	100.00	97.70	98.84
Average	97.46	98.06	97.69

Table 6 Comparison of proposed method with existing approaches

Video	[14]	[12]	[21]	[17]	Proposed
D1	100.00	87.5	100.00	100.00	**100.00**
D2	**100.00**	94.10	**100.00**	94.12	94.12
D3	72.30	81.90	93.58	97.20	**100.00**
D4	97.40	**100.00**	**100.00**	95.65	98.55
D5	87.20	54.20	96.55	95.08	95.08
D6	100.00	100.00	100.00	100.00	**100.00**
D7	90.70	94.70	**100.00**	97.14	94.12
D8	86.80	96.10	96.10	94.87	**96.17**
D9	100.00	80.00	100.00	100.00	**100.00**
D10	75.10	47.30	79.74	97.08	**98.84**
Average	90.95	84.83	96.60	97.11	**97.69**

Hence, the performance can be further improved by incorporating motion parameters in the visual significance model.

4.4 Performance evaluation on Dataset3

The proposed framework is validated on a well known movie song namely, 'Ninukori Varanam' from the Tamil movie-Akni Natchathram. The length of the song is 256 seconds and has 7736 frames in total. The song shows a high variation in terms of lighting conditions. Sample frames are shown in Fig. 7. The song is split into five partitions for the validation purpose. The number of transitions available in the video is detected manually to set the ground truth data. Details about the dataset are enlisted in Table 8. The video frames are modeled by the novel visual significance model and classified by RVFL. The F_1-score of the proposed framework on 'Ninukori Varanam' song compared to the state-of-the arts is enumerated in Table 9.

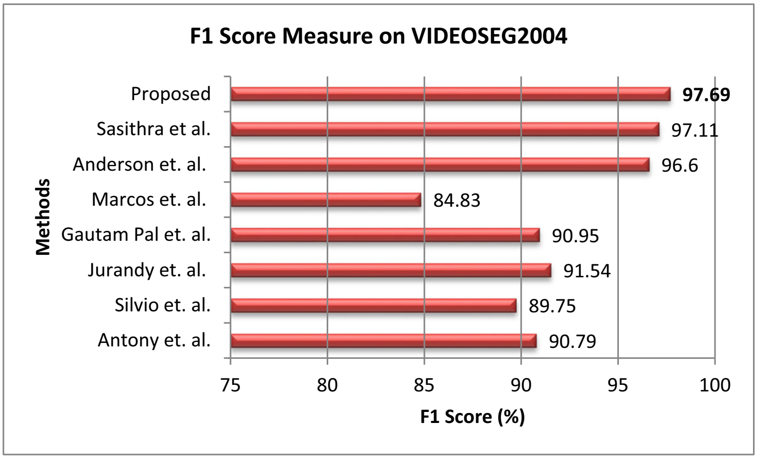


Fig. 4 Performance Comparison of our method with stateof the art techniques in detecting the CTs in VIDEOSEG2004

Table 7 Run time Comparison

Existing Solutions	NSCT+Threshold [22]	WHT [9]	NSCT+SVM [11]	POCS [17]	Proposed
Computational Time (msec) per frame	33,948.73	96.63	142.48	54.933	23.5

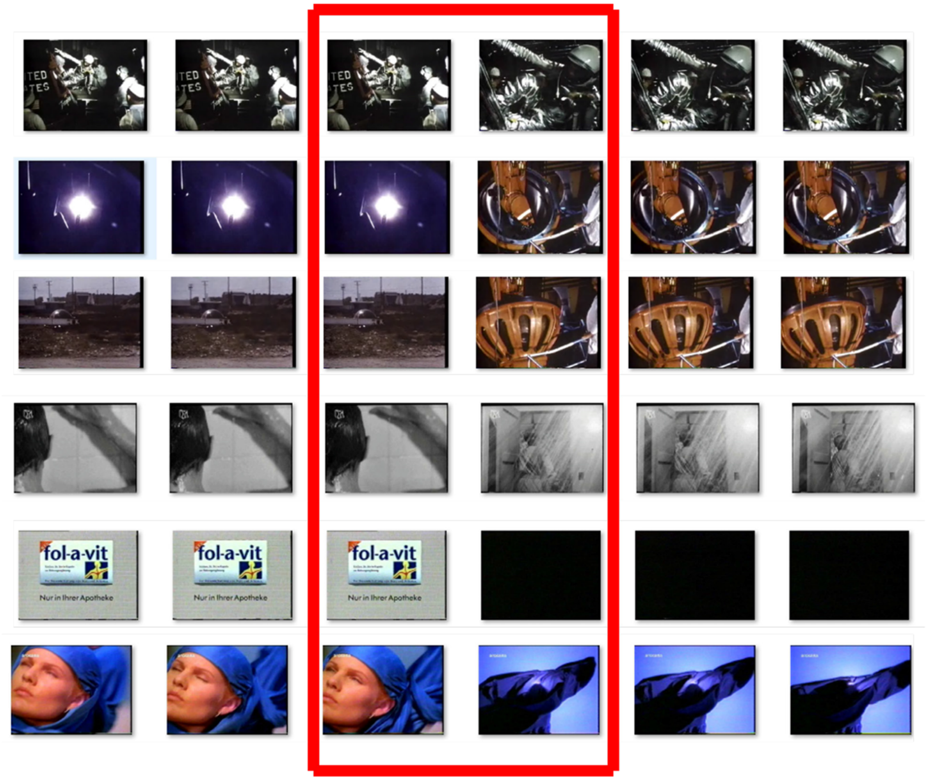


Fig. 5 Samples of successfully detected CTs by proposed framework

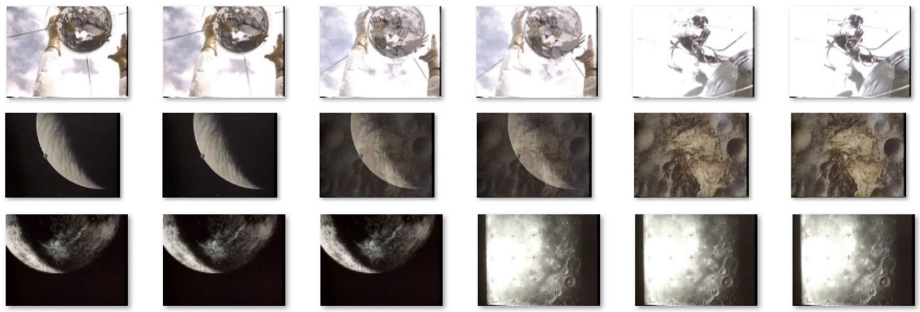


Fig. 6 Samples of missed CTs & STs by proposed VSBD framework

Fig. 7 Sample frames of Ninukori Varanam song

Table 8 Details of the video data from Ninukori Varanam song

Details	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Duration	60	56	60	52	30
#Frames	1799	1679	1799	1599	900
#ATs	12	6	10	20	0

Table 9 F1 Score (%) - comparison with existing solutions

Methods	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Total
WHT [9]	79.50	82.35	68.57	77.37	100.00	81.56
NSCT [22]	70.30	78.75	64.60	80.25	100.00	78.78
Proposed	96.00	90.00	80.00	87.18	100.00	90.64

As shown in Table 9, it is clear that the proposed method shows better F_1score than WHT and NSCT. The performance of all the existing solutions is similar to the proposed model in segment5. A huge difference can be noted in the segments 1, 2 and 3.

As enumerated in Table 10, it is noted that our method leads the recent deep methodologies in the literature. It is obvious that our method attains the state-of-the-art performance when compared with the recent works proposed in transition detection. The focus parameter included in the visual significance model has supported the proposed framework to attain good F1 Score (%).

5 Conclusions and future work

The ultimate goal of this work is to develop a framework to detect the varieties of transitions present in different video genres like News, cartoon, horror, comedy, generic and so on. A visual significance-based model is proposed to detect and categorize the transitions in the video. The proposed model characterizes the transitions well and the same is learned by the

Table 10 Comparison of the performance of proposed framework with recent works

Methods	TSSBD [27]	TRANSNET [20]	TRANSNETV2 [19]	GEBD [18]	DTCW [10]	Proposed
F_1Score	93.2	94.3	96.2	81.7	94	97.69

RVFL network through the temporal signature. This framework also provides a good F_1Score results. For future research, we will explore more contributing parameters like saliency and motion. It is also planned to use the proposed model for video summarization other than video shot boundary detection. We also plan to extend the visual significance model in classifying the sub categories of subtle transitions namely dissolve, fade and wipe.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. Abdulhussain SH, Ramli AR, Mahmmod BM, Saripan MI, al-Haddad SAR, Jassim WA (2019) Shot boundary detection based on orthogonal polynomial. Multimed Tools Appl 78:20361–20382. https://doi.org/10.1007/s11042-019-7364-3
2. Benoughidene A, Titouna F (2022) A novel method for video shot boundary detection using CNN-LSTM approach. Int J Multimed Inf Retr 11:653–667. https://doi.org/10.1007/s13735-022-00251-8
3. Bi C, Yuan Y, Zhang J, Shi Y, Xiang Y, Wang Y, Zhang RH (2018) Dynamic mode decomposition based video shot detection. IEEE Access 6:21397–21407. https://doi.org/10.1109/ACCESS.2018.2825106
4. Chakraborty S, Thounaujam DM, Singh A, Pal G (2022) ALO-SBD: a hybrid shot boundary detection technique for video surveillance system. In: Edge Analytics: Select Proceedings of 26th International Conference—ADCOM 2020. Springer Singapore, Singapore, pp 685–696. https://doi.org/10.1007/978-981-19-0019-8_51
5. Cyganek B, Woźniak M (2017 Tensor-based shot boundary detection in video streams. New Gener Comput 35(4):311–340. https://doi.org/10.1007/s00354-017-0024-0
6. Gygli M (2018) Ridiculously fast shot boundary detection with fully convolutional neural networks. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI). IEEE, pp 1–4. https://doi.org/10.1109/CBMI.2018.8516556
7. Kar T, Kanungo P (2017) Video shot boundary detection based on Hilbert and wavelet transform. In: 2017 2nd International Conference on Man and Machine Interfacing (MAMI). IEEE, pp 1–6. https://doi.org/10.1109/MAMI.2017.8307865
8. Kar T, Kanungo P (2022) A gradient based dual detection model for shot boundary detection. Multimed Tools Appl https://doi.org/10.1007/s11042-022-13547-y
9. LPG G, Dominic S (2014) Walsh–Hadamard transform kernel-based feature vector for shot boundary detection. IEEE Trans Image Process 23:5187–5197. https://doi.org/10.1109/TIP.2014.2362652
10. Mishra R (2021) Video shot boundary detection using hybrid dual tree complex wavelet transform with Walsh Hadamard transform. Multimed Tools Appl 80:28109–28135. https://doi.org/10.1007/s11042-021-11052-2
11. Mondal J, Kundu MK, Das S, Chowdhury M (2018) Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine. Multimed Tools Appl 77:8139–8161. https://doi.org/10.1007/s11042-017-4707-9
12. Mussel Cirne MV, Pedrini H (2018) VISCOM: a robust video summarization approach using color co-occurrence matrices. Multimed Tools Appl 77:857–875. https://doi.org/10.1007/s11042-016-4300-7
13. Nishani E, Cico B (2017) Computer vision approaches based on deep learning and neural networks: deep neural networks for video analysis of human pose estimation. In: 2017 6th Mediterranean conference on embedded computing (MECO). IEEE, pp 1–4
14. Pal G, Acharjee S, Rudrapaul D, Ashour AS, Dey N (2015) Video segmentation using minimum ratio similarity measurement. Int J Image Min 1:87. https://doi.org/10.1504/IJIM.2015.070027
15. Rashmi BS, Nagendraswamy HS (2021) Video shot boundary detection using block based cumulative approach. Multimed Tools Appl 80:641–664. https://doi.org/10.1007/s11042-020-09697-6
16. Roomi SMM, Prakash VJ, Karthikeyan S, Shankar KR (1999) A contrast enhancement technique based on visual significance. J Indian Inst Sci 79(2):89

17. 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. https://doi.org/10.1016/j.jvcir.2020.102754
18. Shou MZ, Lei SW, Wang W, Ghadiyaram D, Feiszli M (2021) Generic event boundary detection: a benchmark for event segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8075-8084
19. Souček T, Lokoč J (2020) Transnet V2: an effective deep network architecture for fast shot transition detection. arXiv preprint arXiv:2008.04838
20. Souček T, Moravec J, Lokoč J (2019) TransNet: a deep network for fast detection of common shot transitions
21. Sousa e Santos AC, Pedrini H (2017) Shot boundary detection for video temporal segmentation based on the weber local descriptor. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp 1310–1315
22. SRRR AS (2016) Non-subsampled Contourlet transform based shot boundary detection. IJCTA 9:3231–3228
23. Tang S, Feng L, Kuang Z et al (2019) Fast video shot transition localization with deep structured models. Pp 577–592
24. Thounaojam DM, Khelchandra T, Singh KM, Roy S (2016) A genetic algorithm and fuzzy logic approach for video shot boundary detection. Comput Intell Neurosci 2016:1–11. https://doi.org/10.1155/2016/8469428
25. Tippaya S, Sitjongsataporn S, Tan T, Khan MM, Chamnongthai K (2017) Multi-modal visual features-based video shot boundary detection. IEEE Access 5:12563–12575. https://doi.org/10.1109/ACCESS.2017.2717998
26. Tong W, Song L, Yang X et al (2015) CNN-based shot boundary detection and video annotation. In: 2015 IEEE international symposium on broadband multimedia systems and broadcasting. IEEE, pp 1–5
27. Wu L, Zhang S, Jian M, Lu Z, Wang D (2019) Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks. IEEE Access 7:77268–77276. https://doi.org/10.1109/ACCESS.2019.2922038
28. Xu J, Song L, Xie R (2016) Shot boundary detection using convolutional neural networks. In: 2016 Visual Communications and Image Processing (VCIP), Chengdu, China, pp 1–4. https://doi.org/10.1109/VCIP.2016.7805554
29. Zhang L, Suganthan PN (2016) A comprehensive evaluation of random vector functional link networks. Inf Sci (N Y) 367–368:1094–1105. https://doi.org/10.1016/j.ins.2015.09.025

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.