Hindawi
Mobile Information Systems
Volume 2022, Article ID 4195905, 14 pages
https://doi.org/10.1155/2022/4195905

Research Article

Efficient Shot Boundary Detection with Multiple Visual Representations

Jasmin T. Jose,[1] S. Rajkumar,[1] Muhammad Rukunuddin Ghalib,[2] Achyut Shankar,[3] Pavika Sharma,[4] and Mohammad R. Khosravi[5]

[1]*SCOPE,Vellore Institute of Technology, Vellore, Tamil Nadu, India*
[2]*School of Engineering & Computing, De Monfort University, Dubai, UAE*
[3]*Dept of CSE, ASET, Amity University, Noida, Uttar Pradesh, India*
[4]*Dept of ECE, Bhagwan Parshuram Institute of Technology, Delhi, India*
[5]*Department of Computer Engineering, Persian Gulf University, Bushehr, Iran*

Correspondence should be addressed to Mohammad R. Khosravi; mohammadr.khosravi@iran.ir

Received 25 April 2022; Revised 4 September 2022; Accepted 15 September 2022; Published 10 October 2022

Academic Editor: Zahid Mehmood

Due to the unlimited growth of video-capturing devices and media, searching and finding a particular video in this huge database becomes a laborious as well as expensive task. Information-rich shots are the inevitable factor of the content-based video processing (CBVP) system. Hence, shot boundary detection (SBD) becomes the basic step of all content-based video retrieval processes. The accuracy of the existing SBD methods highly suffers from false positives and false negatives due to the presence of multiple variants. An efficient SBD method with multiple invariant features is proposed in this paper. A right combination of invariant features such as edge change ratio (ECR), colour layout descriptor (CLD), and scale-invariant feature transform (SIFT) key point descriptors helped to improve the accuracy level of SBD. As the selected features are invariant to most of the variants in video frames, such as illuminance changes, motion, scaling, and rotation, a markable reduction in false detection is possible. Support vector machine (SVM) classifier is used for the classification of frames into transition frames and shot frames. This proposed method is experimented and analysed with the standard SBD dataset TRECVid 2007 videos. The experimental results are compared with some state-of-art methods, and our method shows better performance with a 97% of F1 score.

1. Introduction

The uncontrolled growth of videos due to the mushrooming of multimedia devices has brought the importance of video processing techniques to the concentration of current researchers. So, most of the researchers in the area of video processing are concentrated on CBVP. Content-based image processing (CBIP) techniques can be extended to CBVP[1] by considering the extra useful information in the video such as visual content and audio. CBVP includes the subareas such as video content analysis, video deconstruction into shots or scenes, video summarizing, and video indexing. Both image and video processing systems are working on the basis of different features, which are reflected in the input image or video. The selection of a pertaining feature is the most important factor in both image as well as video processing [2]. Even though CBVP is the prolongation of CBIP, it treats each contiguous frame of video as an image. In CBIP, consistently used features are texture features, colour features, edge features etc. Apart from these conventional basic features, CBVP will consider the visual content features, which will give the continuity details of the video.

As the shot is the basic temporal building block of a video, it will be more powerful, if we could do the processing in this shot [1], rather than working directly with the entire video. Though we are segregating the video into its small patches named shots, it is again a group of image frames with some common features in it. Obviously, processing shots would be much harder and more complicated than processing an image [3]. This makes the CBVP a challenging task. Researchers working on video are having a concept in their mind that a constant feature value exists within a shot, but it will be different in intershots. Shot boundary detection aims to identify the frames with considerable discontinuities in different aspects of their visual contents. These different aspects of visual content are depicted by the features of the frames in the input video. Hence, the selection and extraction of features from video frames should be taken care of.

Illumination, rotation, and scale-variant temporal and spatial features are the greatest challenges in video processing. Miss prediction of shots is possible when the object or camera motion is high in the video [4]. The most commonly used features in video processing are colour, texture, edge, etc. [5]. While considering a video, temporal features, which give more details about the continuity/discontinuity between the frames of the video, are more important. Motion strength, ECR etc. are some of the motion features. Each feature may face one or more factors that lead to some false detection. Eliminating this false detection with the use of invariant features is a big challenge [6].

Video shots or shot transition is of two types [7], abrupt transition (AT) or hard cut and gradual transition (GT) or soft cut. The shot boundaries will be two consecutive frames in the case of AT, i.e., ft and ft + 1, and there will be a large difference in the visual contents of those frames as shown in Figure 1, in which two different shot cuts are visible. But, in the latter one, as the name indicates the transition which is happening gradually throughout several consecutive frames as shown in Figure 2, the natural transition due to the camera close is an abrupt transition, and the gradual transition is due to the visual effects that are applied to the video at the time of video editing. GT is in three different forms, dissolve, fade, and wipe.

As the characteristics of each video are different, the visual contents also will be different. Hence, using a single feature descriptor to represent the whole video will not be sufficient, and it will not give an accurate result. So, in our approach, we are using multiple invariant feature descriptors, and each will extract different representations of visual content. Descriptors, for colour details, texture details, edge, and motion details, are used in our approach. A continuity signal is generated by comparing feature values of adjacent frames of the video. Finally, the classification is done with the help of a well-known supervised classifier, SVM.

The flow of this paper is as follows: Section 2 gives a small review of some of the research works, which are used for getting an idea about the current technologies in the area of the SBD process. The proposed work is explained in Section 3, followed by the experimental results of our work in Section 4. Finally, the paper is concluded by Section 5, with a small description of future work.

2. Literature Review

As the number of videos is increasing rapidly, researchers are more attracted to video processing as well as the SBD process. TRECVid, introduced by the National Institute of Standards and Technology (NIST), is a benchmark for content-based video retrieval (CBVD) processes. SBD was also one of the activities under those workshops. So many researchers participated [1, 4, 8] and contributed a lot to this area of SBD through TRECVid workshops. Multiple datasets were introduced by this TRECVid, which is useful for our SBD experiments. Smeaton et al. [1] have given a detailed overview of the different SBD experiments under this TRECVid workshop. Most of the researchers were using this dataset for their SBD experiments.

Enormous researchers are working on the SBD scheme, with different feature sets and different classification methods. Some of the important research, which helped us to reach this approach, is discussed here. Abdulhussain et al. [2] have done a very good survey on different SBD methods. They have gone through all subsections of SBD processes. Feature selection is very important in the SBD, and a wide variety of local and global features are available. The feature sets start from simple pixel difference and statistical approaches, and now, convolutional neural networks are also used to identify the boundary.

Gargi et al. [3] have given a comparison of different SBD schemes with the colour histogram as a feature. The histogram intersection method is the highlight of their paper. A texture-based SBD approach is proposed by Teng et al. [9], in which they used the local binary pattern as a texture feature. Only with this local feature, it is difficult to get the temporal relationship between the consequent frames. Cosine difference is used as the dissimilarity measure to find the frame difference. In [10], a multilevel difference in the histogram is used for the SBD process. HSV histogram with a single level difference for abrupt transition detection and 5 level differences for gradual transition detection was used. The voting mechanism for the final decision is used. With this multilevel approach, the illumination problems could be minimized in a better way.

Presently, so many researchers were approaching SBD with multiple features, a combination of features. Tippayya et al. [6] and Lin et al. [11] used multiple features. In the first one, a histogram with SURF is used and later, and a histogram as a colour feature and a histogram of gradient as a texture feature were used. Results show better performance in multiple feature models. We have tried a combination of three greyscale features, GLCM (gray level co-occurrence matrix), histogram, and SIFT features in different levels [12] to detect the abrupt transition. The results of the experiment show better performance when compared to the existing single feature-based approaches.

Kumar et al. [13] used the SIFT for SBD and keyframe extraction with some frame elimination to reduce the time complexity. CIEDE2000 colour difference and mean luminance pattern are used in [14] for AT and GT detection, respectively. Lab colour space, in which the approximation of all colours is possible, is used in this work. They have achieved good accuracy by eliminating the effects of illuminance changes. But, no other challenges are considered here.

Image projection and optical flow estimation are used as the features for SBD in [15]. Image projection can consider the shape information, so it is useful for object detection.

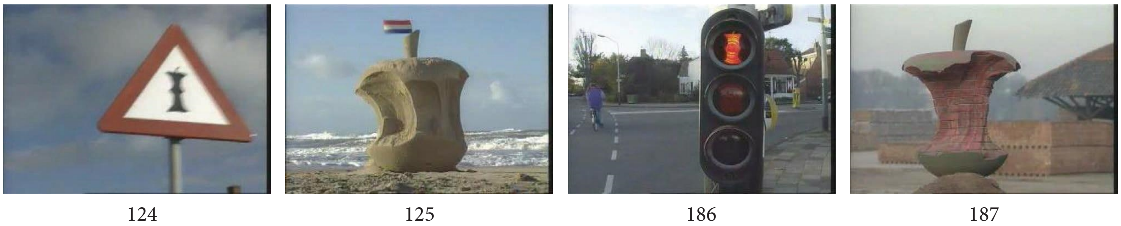


Figure 1: Frames of abrupt transition (CUT) from the video BG 37822 (TRECVid).

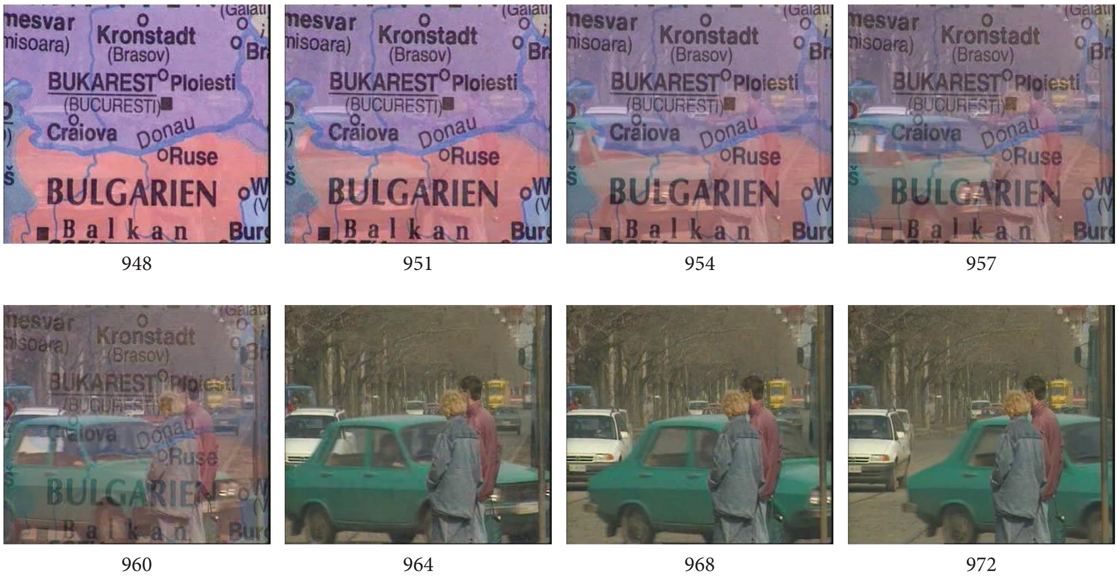


Figure 2: Frames of gradual transition (dissolve type).

Segmenting the video into visual scenes, a group of similar shots, is achieved [16] by the use of the bag of visual word model. Keyframes are selected based on this feature, and then, the feature difference is calculated to identify the similar shots to generate the segment. Better performance in accuracy is achieved by this approach.

Not only the histogram-based or texture-based features used in the SBD process but also edge-based features were used by many of the researchers [5, 17] in this area. Lakshmi and Domnic [5] made use of multiple features such as texture strength, colour strength, and edge strength combined with motion strength in the Walsh Hadamard transformed domain. Motion strength is calculated by the block matching algorithm, which increases the computational cost. The accuracy rate is comparable with top researchers of the TRECVid 2007 workshop. Zhou et al. [18] proposed an SBD with multiple feature sets, such as SURF, colour histogram difference, pixel difference, SIFT along with slice matching. Time complexity is reduced in this approach with the use of candidate frame selection.

Further, finding the similarity/dissimilarity measures for continuity signal generation is important. Selecting a suitable measure is a key point for the SBD scheme. A comparative study on different similarity measures for CBVR is given by [2]. Euclidean distance and city block distance are the commonly used measures in the CBVR system.

Finally, the classification of video frames into different categories, such as cut frames, gradual transition frames, and transition frames, is done with different approaches. Threshold-based approaches and machine learning approaches were used. In threshold-based approaches, adaptive threshold, global threshold, or combined threshold were used by multiple researchers. Most of the aforementioned researchers used different types of threshold mechanisms. Recently, researchers are more interested in machine learning classifications. Zhao and Cai [19] proposed an SBD method with the combination of fuzzy logic and the AdaBoost algorithm for shot classification. The result shows a high precision rate when compared to double-threshold approaches.

Idan et al. [20] proposed a fast SBD method with separable moments and SVM classifiers. Orthogonal polynomial-based moments are used as feature sets on candidate frames, and final classification is done with SVM. Zheng and Zhang [21] have proposed a combined version of the threshold classifier and SVM classifier. The initial classification is done by the threshold approach. Then, SVM is used for confirmation. Researchers on this SBD scheme are more interested in supervised learning classification than unsupervised learning classifiers, as the output classes are well known.

An efficient and accurate SBD method is very important for subsequent CBVP and CBVR. Appropriate selection of features which are invariant to illuminance changes, issues from camera/object motion, scaling, rotation, and new object entry is the major challenge of the SBD process. In fact, existing SBD models are handling a few of the above-mentioned challenges, and they are only able to overcome some of the challenges, but they failed to solve other issues. So, it is essential to identify some set of features which can handle all those issues. Based on this analysis, we have selected some sets of invariant features for an efficient SBD process. We have used three optimal features such as CLD, ECR, and SIFT in our proposed SBD model. The proposed SBD method is explained in Section 3.

3. Proposed SBD Method

Generally, a shot boundary detection process will have three phases: first, feature extraction (representation of visual contents), followed by continuity signal generation from the extracted features, and finally, classification (shot frames and boundary frames) of frames. Video is the amalgamation of contiguous frames, and each frame can be treated the same as an image. Initially, the frames will be extracted from the input video for further processing. Then, these contiguous frames are processed to extract the features. While processing a video, it is important to consider both image content details (spatial) and temporal details of the video. The block diagram of the proposed method is shown in Figure 3. The three phases of our SBD process are as follows.

3.1. Feature Extraction. As mentioned above, our approach used three different features: colour, texture, and edge with motion features. Most of the existing SBD approaches are using any one or two of the features in their SBD process. An efficient SBD method should use features which are invariant to the effects of illumination changes, slow/fast motion in video, scaling, and rotation. Unfortunately, no features are invariant to all these effects. In fact, a combination of features may help to solve this issue. From the research, we have identified a combination of features, which can cover most of the above-mentioned effects. Colour layout descriptor (CLD) is used for colour feature extraction, edge change ratio for motion features, and scale-invariant feature transform (SIFT) for extracting texture features. Each descriptor is described in the following sessions.

3.1.1. Colour Layout Descriptor. The spatial distribution of colour of visual signals can be represented effectively in a compact form by a colour descriptor called colour layout descriptor [7, 22]. It is proven that this CLD is very efficient in the processing of both still images and video. CLD is considered the most precise and fast colour descriptor, and it can describe the colour relation between a group of images. Frames that have a high colour relationship will be coming under the same shot. The discriminability and robustness of CLD make this descriptor optimal for the SBD process. Invariants of CLD under light changes are an added advantage in our algorithm.

Different steps in CLD extraction are shown in Figure 4. The extraction of the CLD feature is performed in four steps. First is the partitioning of frames into 8×8 blocks. Secondly, the dominant colour of each block will be selected, using the average colour method. In the third stage, the colour space of tiny blocks with dominant colour (8×8) is converted to YCbCr colour space, and each of these colour space components will be transformed using DCT $(8 \times 8$ DCT).

The output of this transformation is 3 sets (Y, Cb, and Cr separately) of 64 DCT coefficients. To get the feature vector, a zigzag scanning will be performed on these 3 sets of DCT coefficients, so that all low-frequency coefficients can be grouped. Six from the Y component and 3 from each Cb and Cr, a total of 12 DCT coefficients will be extracted. The main advantage of this descriptor is its less storage cost. CLD extracts only 12 DCT coefficients out of 192 coefficients $(64 * 3)$ from a frame, and the coefficient size is very small. This highlights the CLD among all other colour descriptors. Storage cost is the main challenge of the CBVR system, as it deals with videos, which have more frames, and it will occupy more storage space. Hence, this less storage cost with DLD is very useful in the fast SBD process [22].

3.1.2. Edge Change Ratio. The edge change ratio is a measure that differentiates two consecutive frames in a video. So, this feature can be used to identify the shot transition in the video. Zabih et al. [23] introduced ECF (edge change factor) in their work. Recently, researchers are using this same concept with some slight changes in ECR in their work. ECR is invariant to illumination changes, and it can also minimize the false shot boundary detection due to the camera or object motion in the video. Moreover, it gives the edge change ratio between the consecutive frames, from which the temporal aspects of the video can be extracted. Due to these characteristics, ECR is chosen as an edge feature in our proposed method.

The steps to calculate the ECR from two consecutive frames are as follows:

(1) Extract two consecutive frames F_i and F_{i+1} from the input video.

(2) Convert the RGB frames into YCbCr colour space, F_i and F_{i+1}.

(3) Apply canny edge detection only on the luminance element of YCbCr colour space. Output frames are Fc_i and Fc_{i+1}.

(4) Count the number of edge pixels in Fc_i and Fc_{i+1}. Counts are P_i and P_{i+1} respectively.

(5) Apply dilation then inversion on Fc_i and Fc_{i+1}, and get Fc_i' and Fc_{i+1}'.

(6) Perform subtraction (set subtraction) between the output images to calculate the exiting edge pixels (P_{out}) and entering edge pixels (P_{in}), as follows [23]:

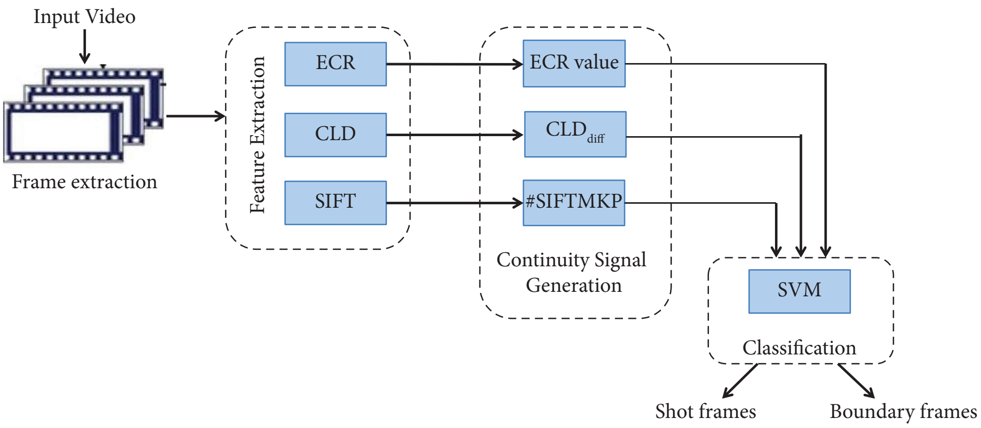


FIGURE 3: Multiple feature-based SBD process.

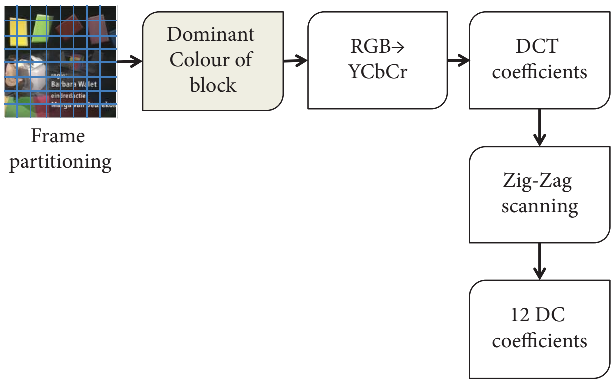


FIGURE 4: CLD feature extraction.

$$\begin{aligned} P_{\text{out}} &= Fc_i' - Fc_{i+1}', \\ P_{\text{in}} &= Fc_{i+1}' - Fc_i'. \end{aligned} \tag{1}$$

(7) Calculate the ECR using the following equation:

$$\text{ECR}(i,k) = \text{Max}\left(\frac{P_{\text{out}}}{P_i}, \frac{P_{\text{in}}}{P_{i+1}}\right), \tag{2}$$

in which P_i and P_{i+1} are the number of edge pixels in the consequent frames F_i and F_{i+1}, respectively and, k is 1 for two consecutive frames. The ECR value is in the range of 0 to 1.

3.1.3. SIFT Key Point Matching. As the name indicates, SIFT descriptor takes the image as an input and transforms it into a large collection of local feature vectors, and these vectors are invariant to the scaling and rotation of images. SIFT can provide correct correspondences between two images. Moreover, SIFT provides information on the content elements. The output of this SIFT algorithm will be a descriptor vector of all SIFT key points. The process of SIFT key point extraction and key point matching is as follows.

```
SIFT key point extraction:
  Input: $N_c$, number of frames ($N_c$)
  Output: Number of SIFT key points
    $S = $ cell (1, 1) Key points extraction
    for $t = 1$ to $N_c$ do
      Find the location and scale of key points
      Strong key point selection
      Consistent orientation of key points
      $\text{kp}(t) = $ key point descriptor
      $S\{t, 1\} = \text{kp}(f_t)$
      $S\{t, 2\} = P_{\text{kp}}(f_t) = \text{count}(S\{t, 1\})$; number of key points
    end for
SIFT key point matching:
    for $t = 1$ to $N_c - 1$ do
      $S\{t, 3\} = \text{match}(S\{t, 1\}, S\{t+1, 1\})$
      $S\{t, 4\} = P_{\text{mkp}}(f_t, f_{t+1}) = \text{count}(\text{S}\{t, 3\})$
    end for
```

where N_c is the number of candidate frames, $\text{kp}(f_t)$ is the key points of frame t, $P_{\text{kp}}(f_t)$ is the number of the keyframe in frame t, and P_{mkp} is the number of matching key point between f_t and f_{t+1}. This number of matching key points of consecutive frames is considered for the continuity signal generation.

3.2. Continuity Signal Generation. The feature difference vector for all these aforementioned descriptors is generated by comparing the feature values of adjacent frames of video. Colour layout descriptor, edge change ratio, and SIFT key point matching are used. In CLD, the matching of feature vectors of consecutive frames can be done with the (3). Let us consider CLDs of two consecutive frames (DY, DCb, DCr) and (DY′, DCb’, DCr’). Then, the frame difference CLD_{Dist} is [24]

$$\mathrm{CLD}_{\mathrm{Dist}} = \sqrt{\sum_i w_{\mathrm{yi}}\left(\mathrm{DYi}-\mathrm{DYi}'\right)^2} + \sqrt{\sum_i w_{\mathrm{cbi}}\left(\mathrm{DCbi}-\mathrm{DCbi}'\right)^2} + \sqrt{\sum_i w_{\mathrm{cri}}\left(\mathrm{DCri}-\mathrm{DCri}'\right)^2}, \qquad (3)$$

where DYi, DCbi, DCri are the *i*th coefficients of Y, Cb, Cr colour components, respectively. wyi, wcbi, wcri denote the weighting factor of the *i*th coefficient.

Equation (2) gives the ECR between two consecutive frames of video. This ratio is considered the continuity signal. In the case of SIFT feature, the number of matching key points (Pmkp from the algorithm), between consecutive frames, is considered for continuity signal. This is calculated by comparing the adjacent frames of the input video. These three continuity signals are considered feature difference vectors for classification. Each feature vector value is normalized in the range of 0 to1. These vectors are given to the SVM as input vectors for classification purposes.

3.3. Classification of Video Frames. Finally, the classification of video frames into two different classes is done, using these aforementioned feature vectors with an SVM classifier. SVM is useful in image classification and segmentation as well [25]. The suitability of SVM in noisy and noise-free environments, less training time due to the use of fewer parameters, and less complexity makes the SVM the best model in our SBD method [20]. This modal can be used not only for linearly separable data but also for nonlinear separable data sets with the help of some known kernel tricks. Most of the researchers are using radial basis function (RBF) (4) as the kernel in SVM [9]. The nonlinear mapping and less complexity of the RBF kernel make it a better choice [20, 25] for SBD. The expression of the RBF kernel with the training vectors v_i, v_j and the kernel parameter γ is as follows:

$$K\left(v_i, v_j\right) = \exp\left(-\gamma\left\|v_i - v_j\right\|^2\right);\ \gamma > 0. \qquad (4)$$

A set of videos from the TRECVid 2007 data set is used for training the SVM. In this data set, the ratio between the boundary frames and shot frames is very high (around 1 : 1200), and it will affect the result. So, a custom data set is constructed from this original data set by removing the redundant frames. SVM of OpenCV library is used for the implementation. 8 videos, which are described in Table 1 (these videos are not considered in training), are used for classification. The cost parameter is tuned with X-fold cross-validation. Binary classification of video frames into shot frames and boundary frames is done with an SVM classifier.

The notations and abbreviations used in this paper are given in Table 1.

Table 1: Acronyms and notations used in paper.

Acronyms and notations used	
CBVP	Content based video processing
SBD	Shot boundary detection
CLD	Colour layout descriptor
ECR	Edge change ratio
SIFT	Scale invariant feature transform
AT	Abrupt transition
GT	Gradual transition
SVM	Support vector machine
GLCM	Gray level co-occurrence matrix
$K(v_i, v_j)$	RBF kernel
γ	Kernel parameter
P_i	Number of edge pixels in f_t
Nc	Number of frames
KP (*ft*)	Key points of f_t
Pkp (*ft*)	Number of key points of f_t
Pmkp (*ft*)	Number of matching key points between f_t and f_{t+1}
$\mathrm{CLD}_{\mathrm{dist}}$	CLD difference between f_t and f_{t+1}

4. Experimental Results and Analysis

In this section, we have done the experimental analysis on the data set TRECVid 2007, which is the standard dataset, particularly for SBD. We have used 8 different videos from 17 video data sets, which are named V1 through V8 in this paper. Table 2 shows a detailed description of those input videos. The abrupt transitions observed by humans are given as the ground truth value. Different types of aberrations such as illumination changes, view changes, scaling, zooming, rotation, etc., are there in the selected videos.

Recall and precision are the important model evaluation metrics usually used in the SBD process. In this paper, we have used these metrics for evaluation along with the *F*1 score, which is the harmonic mean of precision and recall [5]. These metrics can be calculated as follows:

$$\mathrm{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{false negative}},$$

$$\mathrm{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{false positive}}, \qquad (5)$$

$$F1\ \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Table 3 shows the recall, precision, and *F*1 score values for the 8 videos selected for the SBD experiment. Our approach to the SBD process obtained high performance with the use of multiple features and an SVM classifier. As we have used important features such as colour feature, edge feature, and motion feature, we could overcome most of the variance such as illumination changes, rotation variance, scaling, motion effects due to camera motion, object motion etc.

The number of shot frames is very high when compared to the number of cut frames in our data set. *F*1 score is not considering the true negative (TN) cases in the classification. Though we are giving more attention to the positive cases in SBD, as negative cases are very high in the data set, we should

Table 2: Description of input video dataset (TRECVid 2007).

Video representation	Video name	No. of frames	No. of cuts
V1	BG 2408	35892	101
V2	BG 9401	50049	81
V3	BG 14213	83115	106
V4	BG 35050	36999	91
V5	BG 36182	29610	96
V6	BG 37417	23004	76
V7	BG 37822	21960	119
V8	BG 38150	52650	216

Table 3: Average precision and recall values of input videos.

Video input	Total no. of cuts	No. of cuts detected		Proposed method	
	TP + FN	TP + FP	Recall	Precision	$F1$ score
V1	101	99	0.97	0.99	0.979
V2	81	82	0.975	0.98	0.977
V3	106	105	0.98	0.99	0.985
V4	91	93	0.978	0.98	0.979
V5	96	98	0.98	0.98	0.98
V6	76	75	0.94	0.933	0.936
V7	119	118	0.96	0.98	0.97
V8	216	216	0.972	0.972	0.972
Total/Average	**886**	**886**	**0.969**	**0.976**	**0.972**

Table 3 shows the recall, precision, and F1 score values of selected input videos. The total number of actual cuts, number of cuts detected by the proposed method, and average values of recall, precision, and F1 score achieved by the proposed method are shown in bold. These score shows the better performance of our SBD method. This high performance is achieved by the use of multiple invariant features and the SVM classifier. As we have used important features such as color feature, edge feature, and motion feature, we could overcome most of the variance such as illumination changes, rotation variance, scaling, motion effects due to camera motion, and object motion.

consider TN cases also. A metric which gives attention to the TN cases known as balanced accuracy is also calculated for analysis. This metric is calculated as follows:

$$\text{Balance d Accuracy} = \frac{\text{sensitivity} + \text{Specificity}}{2},$$
$$\text{Sensitivity (Recall)} = \frac{\text{True positive}}{\text{True positive} + \text{false negative}}, \quad (6)$$
$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{false positive}}.$$

Table 4 shows the balanced accuracy values for each input video sequence. As the proposed method can reduce the number of false positives at a markable rate, the balanced accuracy is very close to the $F1$ score.

Sample cut frames detected by the proposed method are shown in Figure 5. Figure 6 shows the feature difference plot for those frames. The sudden hikes in feature difference value represent the cut transition. The confusion matrices for each input video are depicted in Figure 7 for more clarification.

To show the comparative performance of our approach, we compared our results with the top three researchers of TRECVid 2007. They are AT & T research [8], THU-ICRC [4], and the University of Marburg [1]. We have also compared our approach with a multifeature-based SBD model [6], and our method performs well. Figure 8 shows the comparison result of the recall, precision, and $F1$ score. The $F1$ score of our approach shows that our approach outperforms in SBD process.

Table 4: Balanced accuracy values.

Video sequence	Recall (sensitivity)	Specificity	Balanced accuracy
V1	0.941	0.987	0.964
V2	0.975	0.991	0.983
V3	0.981	1.000	0.991
V4	0.978	0.988	0.983
V5	0.979	0.980	0.980
V6	0.921	1.000	0.960
V7	0.933	0.977	0.955
V8	0.972	0.991	0.982

The use of ECR and CLD features has great significance in the SBD process. To prove this, we have compared the proposed method with our multiple grey scale features-based SBD method [26]. Gray level cooccurrence matrix (GLCM), histogram, and SIFT key point matching are used as features in [26]. Table 5 shows the comparison results. The average values of recall, precision, and $F1$ score depict the significance of CLD and ECR in the SBD method. The comparison of average values of precision, recall, and $F1$ score in Figure 9 shows the better performance of the proposed method. This difference is only due to the use of CLD and ECR.

Further, to show the superior performance of our method, the result is compared with some recent SBD

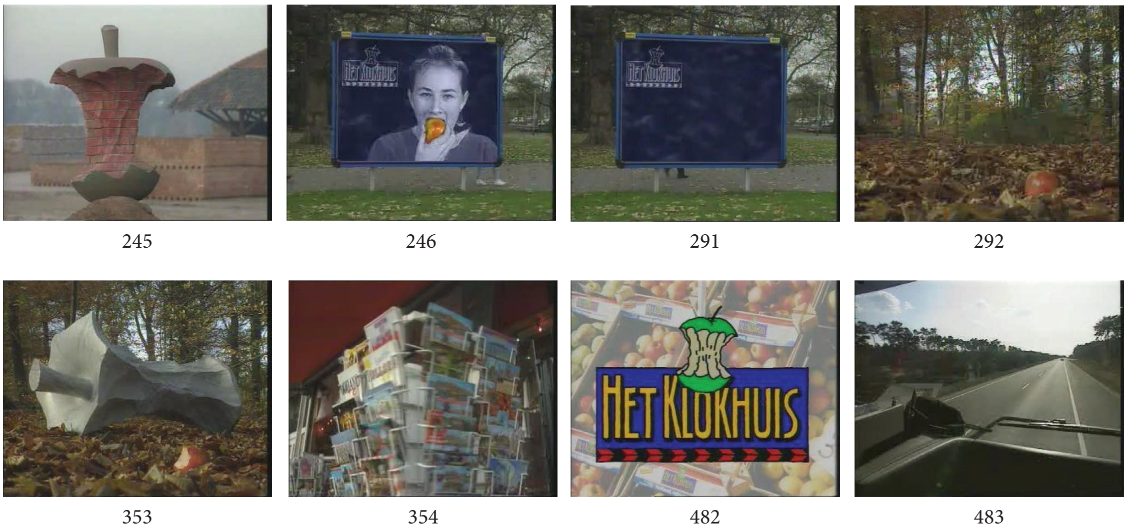


Figure 5: Sample shot cut frames from TRECVid dataset.

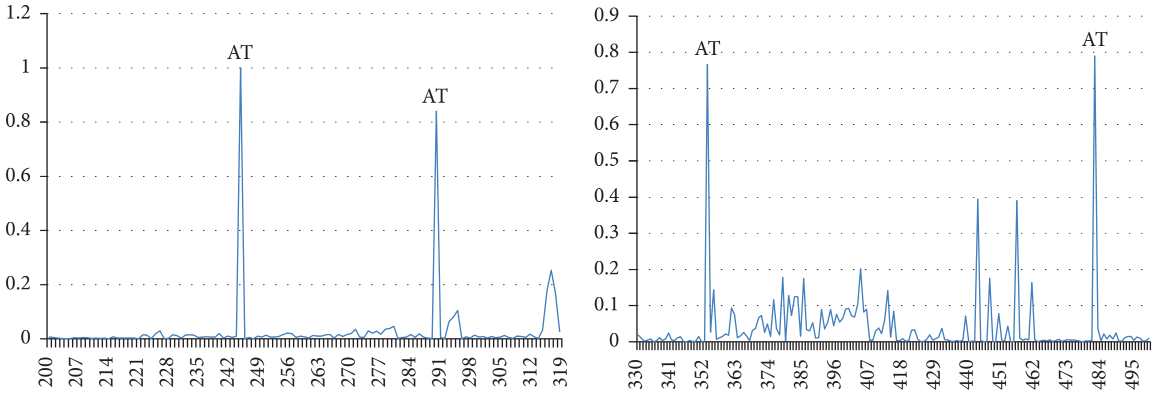


Figure 6: Feature difference value plot for the cut frames in Figure 5.

methods. SBD with nonsubsampled contourlet transform [27], separable moments, and SVM based SBD [20], SBD using block-based cumulative approach [7], local binary pattern histogram Fourier based SBD [24], SBD using pyramidal opponent colour shape [28], and a bifold-stage SBD [29] are considered for comparison. Table 6 shows the comparison result of the recall, precision, and $F1$ score. The average precision rate of the SM-SVM [20] method is higher than the proposed method. A small variation is there in the recall rate between the proposed method and the NSCT [27] method. However, the proposed SBD method has achieved the highest average $F1$ score when compared to other SBD methods which are shown in Figure 10. The selection of invariant features leads the proposed method to overcome the issues such as small level of motion, illuminance changes, rotation, scaling.

New object entry is an unavoidable challenge in the SBD process. Most of the researchers are compromising their accuracy due to the false detection from this new object entry. If the object size is very small compared to the frame size, then it can be managed. But, large-sized object entries will definitely give false positives. Some sample frames are shown in Figure 11(a). As a new object enters the adjacent frame, the possibility of getting high ECR and CLD values is high. It may lead to false detection. From the analysis, it is observed that the use of SIFT matching key points in the proposed method avoids false positives due to the new object entry.

The new object entry is very common in most of the videos. If the new object size is lesser than the frame size, then the SIFT can identify the key points from the remaining area of the frame. A sample set of frames, which avoids the false detection with the new object entry, is shown in Figure 11. Consider the adjacent frames 461 and 462 from TRECVid 2007 videos, a new object is entered into frame 462. Even though the CLD feature difference between those two frames is high as shown in Figure 11(b), the number of SIFT matching key points between those two frames is also very high as shown in Figure 11(c). Because of this, a very high value for SIFT key point matching SVM classifies these frames as shot frames.

Analysis shows that the ECR is invariant to a certain level of camera/object motion. However, the fast motion of the

Predicted Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

(a)

	Predicted Actual	Positive	Negative
V1	Positive	95	6
	Negative	4	35690

	Predicted Actual	Positive	Negative
V2	Positive	79	2
	Negative	3	49887

	Predicted Actual	Positive	Negative
V3	Positive	104	2
	Negative	1	82903

	Predicted Actual	Positive	Negative
V4	Positive	89	2
	Negative	4	36817

	Predicted Actual	Positive	Negative
V5	Positive	94	2
	Negative	4	29418

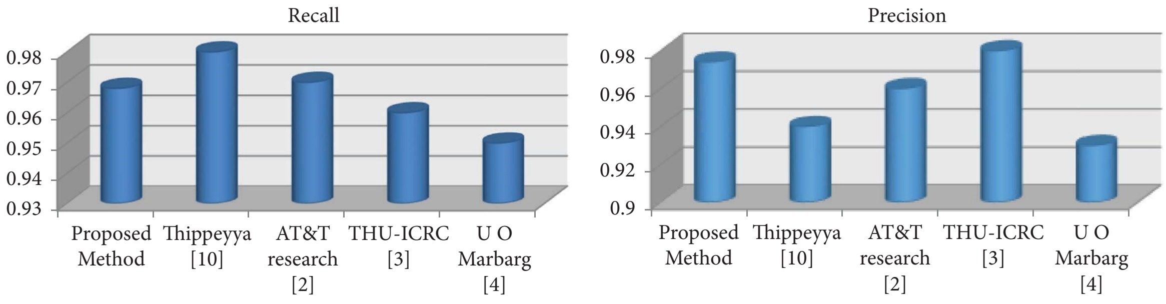
	Predicted Actual	Positive	Negative
V6	Positive	70	6
	Negative	5	22852

	Predicted Actual	Positive	Negative
V7	Positive	111	8
	Negative	7	21722

	Predicted Actual	Positive	Negative
V8	Positive	210	34
	Negative	6	52218

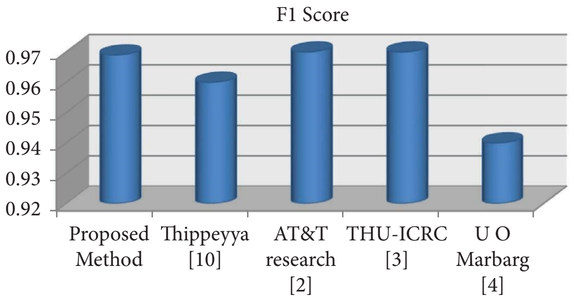
(b)

Figure 7: (a). General confusion matrix. (b). Confusion matrices for the results of each video.



(a) (b)

Figure 8: Continued.



(c)

Figure 8: Result comparison: (a) recall, (b) precision, and (c) *F*1 score.

Table 5: Comparison of gray feature method and proposed method.

Video sequence	Gray feature method			Proposed method		
	Recall	Precision	*F*1 score	Recall	Precision	*F*1 score
V1	0.96	0.951	0.956	0.97	0.99	0.979
V2	0.938	0.95	0.944	0.975	0.98	0.977
V4	0.945	0.956	0.95	0.978	0.98	0.979
V5	0.918	0.944	0.931	0.98	0.98	0.98
Average	**0.94**	**0.95**	**0.945**	**0.975**	**0.982**	**0.978**

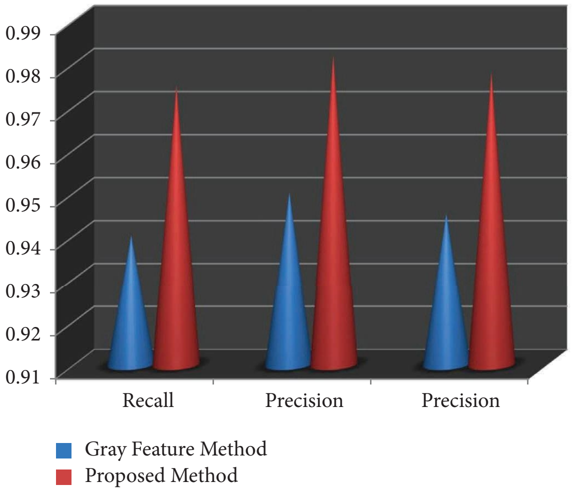


Figure 9: Comparison of gray feature method and proposed method.

Table 6: Comparison of the proposed system with the latest SBD methods.

Algorithm	Precision	Recall	*F*1 score
NSCT [27]	96.36	**97.66**	97.01
SM-SVM [20]	**99.34**	94.71	96.97
BBC [7]	94.7	94.2	94.45
LBP-HF [24]	95	94	94
POCS [28]	94	94	93.67
BIFOLD-STAGE [29]	98.75	95.37	96.97
PROPOSED METHOD	97.6	96.9	**97.2**

Table 6 shows the comparison result of the recall, precision, and F1 score. The dominant values in each metrics among different SBD methods are highlighted. The average precision rate of the SM-SVM [20] method is higher than the proposed method. A small variation is there in the recall rate between the proposed method and the NSCT [27] method.

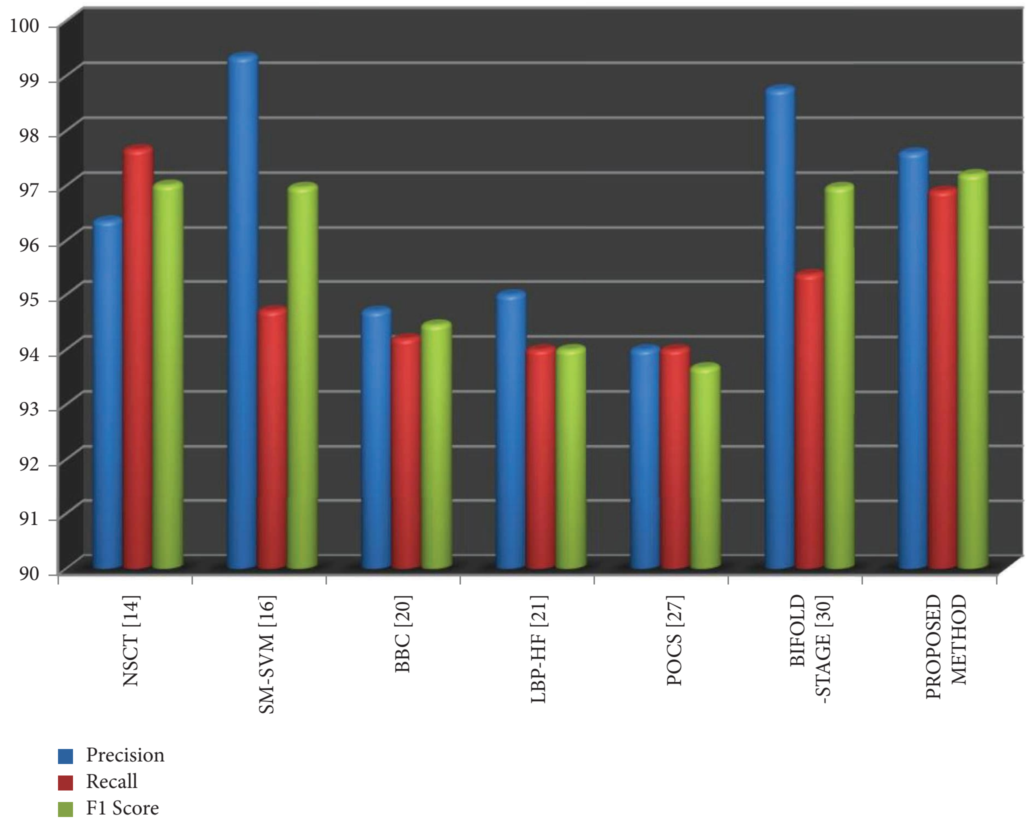
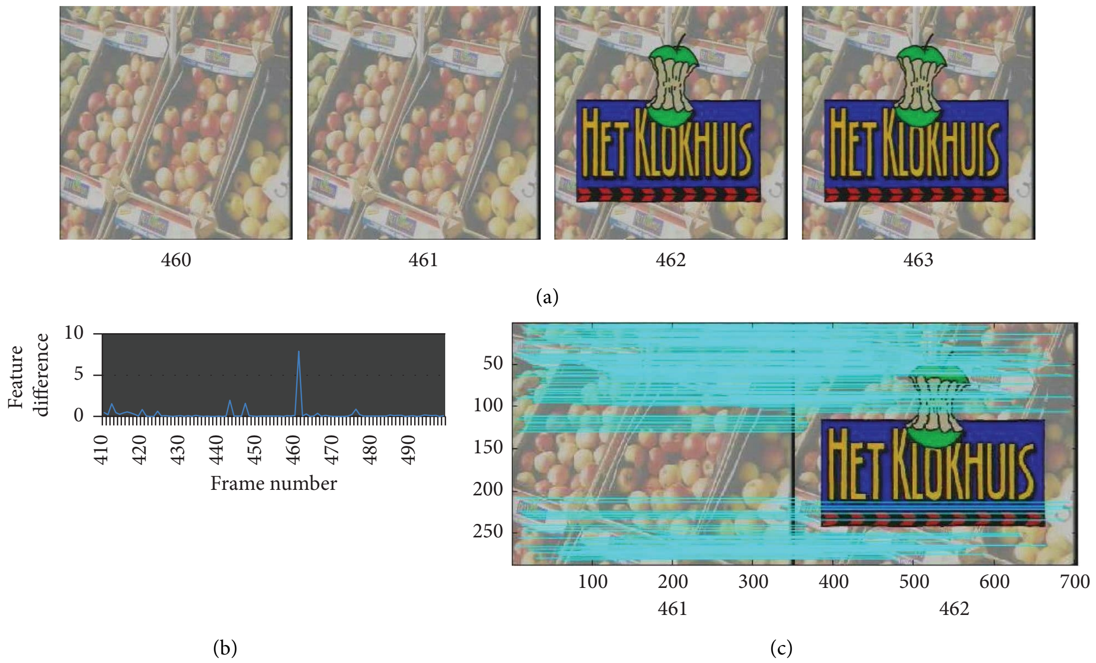


Figure 10: Result comparison: Recall, Precision, and *F*1 score.



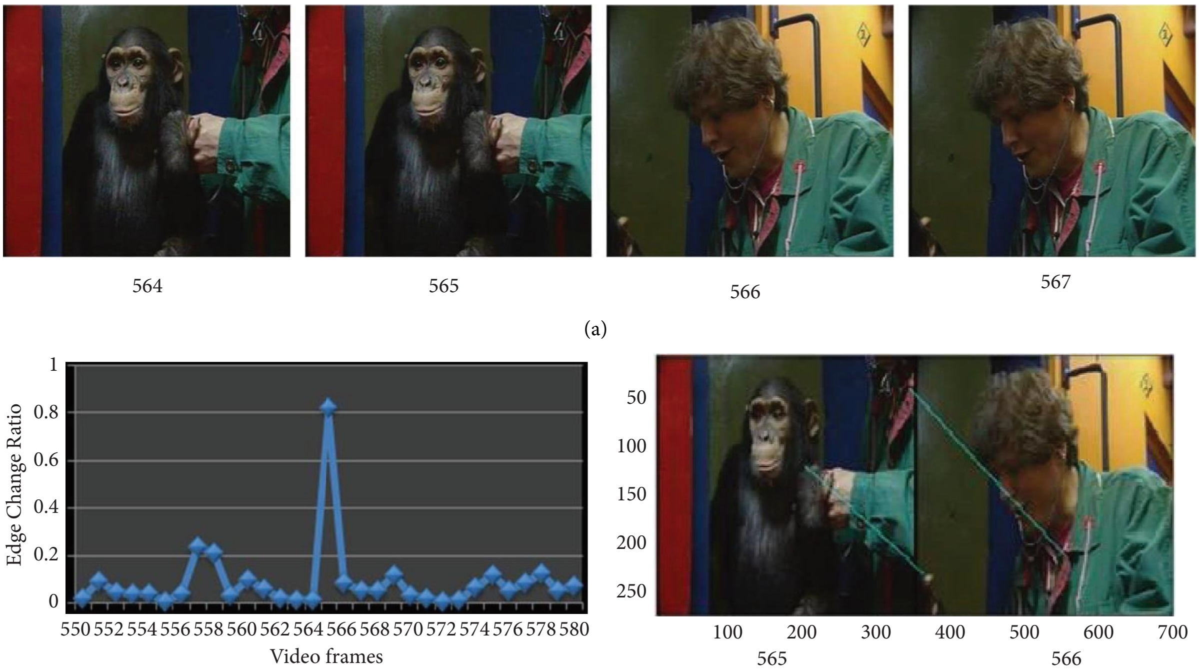
(a)

(b)

(c)

Figure 11: Correct detection in new object entry. (a) Consequent frames of video, (b) feature difference plot for the frames in (a), (c) SIFT matching key points.

camera/object in a video may give a false alarm. Figure 12 shows some sample frames of video, which is a false detection in this SBD process due to the fast movement of the camera. Both ECR value and number of SIFT key point matching in Figures 12(b) and 12(c) show that frames 565 and 566 of Figure 12(a) are boundary frames. It is a false



(a)

(b)

(c)

Figure 12: False positive case in the proposed method. (a) Consequent frames of video, (b) ECR plot for the frames in (a), (c) SIFT matching key points.

detection due to the fast movement of the camera. Hence, it is required to consider some motion features to handle the fast motion of the camera/object in the video. This may be considered in our future work.

Shot boundary detection is a basic step for any CBVR algorithm. A quality SBD process will directly come up with quality CBVR results. So, our proposed method can be used in any CBVR algorithm as a preprocessing stage. Specifically, the algorithms such as video summarisation and video segmentation can use our SBD method for better performance. Gradual transition (GT) detection is not been considered in this approach. Most of the existing methods are utilizing different features for AT and GT detection. The same set of features for both AT and GT detection would be better to reduce the computational complexity. This case also would be considered in future.

A remarkable accuracy improvement in the proposed method is noticed in the results. The main challenge of the SBD process is the multiple variants existing in the video frames. The proposed method achieved this better accuracy by the selection of optimal invariant features as follows.

(1) The discriminability and robustness of CLD help to reduce the false detection due to the illuminance changes. As CLD is selecting only 12 DC coefficients from a frame, the storage cost is very less, and it is useful for a fast SBD model. Further, miss prediction due to the similarity in adjacent frames is reduced by the discriminability nature of CLD.

(2) The invariants of ECR towards illuminance changes and object/camera motion lead to a higher accuracy rate. Moreover, ECR is giving the temporal aspects of the video.

(3) SIFT descriptor is highly invariant to the scaling and rotation effects in the video sequences. SIFT key point matching helped to reduce the false detection due to new object entries in the video frames.

5. Conclusion

An SBD model with an optimal feature set is proposed. The feature vector includes the feature difference from multiple features such as texture, edge/motion, and colour. Colour layout descriptor with DC coefficients, edge change ratio with canny edge detector, and SIFT key point are used as the features. These feature difference vectors are used as the input for the SVM classifier, which classifies the video frames into transition or no-transition frames. Video data from TRECVid, a standard dataset for the SBD process, are used for experiments. Our approach shows a 97.6% $F1$ score, which is a comparatively better result. A comparison of our experimental results with existing methods reveals that the proposed method outperforms in SBD process by reducing the false positives. Some basic materials of this paper have been presented at a 2021 event [26], but the readers should consider the current version as our final work.

Data Availability

The data will be provided at reasonable request to the first two authors (e-mails: jasminlijo@vit.ac.in and rajkumars@vit.ac.in).

Additional Points

The authors have improved the accuracy level with the accurate use of multiple feature descriptors, which can reduce the effects of illumination changes, rotation variance, motion changes, etc. SVM classifier is used in this approach for the final classification of frames into transition frames and normal frames.

Conflicts of Interest

The authors declare that there are no conflicts of interest in preparing this research article.

References

[1] M. Mühling, R. Ewerth, T. Stadelmann, C. Zöfel, B. Shi, and B. Freisleben, "University of Marburg at TRECVID 2007: Shot Boundary Detection and High Level Feature Extraction," in *Proceedings of the TRECVID 2007 workshop participants notebook papers*, Gaithersburg, MD, USA, November 2007.

[2] S. H. Abdulhussain, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim, "Methods and challenges in shot boundary detection: a review," *Entropy*, vol. 20, no. 4, p. 214, 2018.

[3] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *Circuits Syst. Video Technol*, vol. 10, no. 1, pp. 1–13, 2000.

[4] J. Yuan, Z. Guo, L. Lv et al., "THU and ICRC at TRECVID 2007," in *Proceedings of the TRECVID 2007 workshop participants notebook papers*, Gaithersburg, MD, USA, November 2007.

[5] P. G. G. Lakshmi and S. Domnic, "Walsh—hadamard transform kernel-based feature vector for shot boundary detection," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 23, no. 12, pp. 5187–5197, 2014.

[6] S. Tippaya, S. Sitjongsataporn, T. Tan, M. M. Khan, and K. Chamnongthai, "Multi-modal visual features-based video shot boundary detection," *IEEE Access*, vol. 5, pp. 12563–12575, 2017.

[7] B. S. Rashmi and H. S. Nagendraswamy, "Video shot boundary detection using block based cumulative approach," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 641–664, 2021.

[8] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, and P. Haffner, "At&t research at trecvid 2007," in *Proceedings of the TRECVID 2007 workshop participants notebook papers*, pp. 19–26, Gaithersburg, MD, USA, November 2007.

[9] S. Teng and W. Tan, "Video temporal segmentation using support vector machine," in *Asia Information Retrieval Symposium*, H. Li, T. Liu, W. Y. Ma, T. Sakai, K. F. Wong, and G. Zhou, Eds., vol. 4993, pp. 442–447, Springer, Berlin, Heidelberg, 2008.

[10] Z. Li, X. Liu, and S. Zhang, "Shot boundary detection based on multilevel difference of colour histograms," in *Proceedings of the First International Conference on Multimedia and Image Processing (ICMIP)*, pp. 15–22, IEEE Press, Bandar Seri Begawan, Brunei, June 2016.

[11] C.-H. Lin, M.-D. Hsiao, and Li-J. Fu, "Analysis of shot boundary based on color and texture features of frame IOP Conference Series: materials Science and Engineering," *IOP Conference Series: Materials Science and Engineering*, vol. 53, no. 1, Article ID 012004, 2013.

[12] J. T. Jose and S. Rajkumar, "Abrupt transition using Gray scale features," *Journal of Advanced Research in Dynamical & Control Systems*, vol. 11, no. 07, 2019.

[13] G. N. Kumar, V. S. K. Reddy, and S. S. Kumar, "Video shot boundary detection and key frame extraction for video retrieval," in *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, pp. 557–567, Springer, Singapore, Asia, July 2018.

[14] S. Chakraborty, D. M. Thounaojam, and N. Sinha, "A shot boundary detection technique based on visual colour information," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 4007–4022, 2021.

[15] S. Chakraborty and D. M. Thounaojam, "A novel shot boundary detection system using hybrid optimization technique," in *International Journal segment retrieval,* vol. 1, pp. 674–677, IEEE, 2019.

[16] M. Haroon, J. Baber, I. Ullah, S. M. Daudpota, M. Bakhtyar, and V. Devi, "Video scene detection using compact bag of visual word models," *Advances in Multimedia*, pp. 1–9, 2018.

[17] A. M. Jacobs, G. T. Ioannidis, and O. Herzog, "Automatic shot boundary detection combining color, edge, and motion features of adjacent frames," in *Proceedings of the TRECVID 2004 Workshop Notebook Papers*, pp. 197–206, Bremen, Germany, November 2004.

[18] S. Zhou, X. Wu, Y. Qi, S. Luo, and X. Xie, "Video shot boundary detection based on multi-level features collaboration," *Signal, Image and Video Processing*, vol. 15, no. 3, pp. 627–635, 2021.

[19] Z. C. Zhao and A. N. Cai, "Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory," in *Proceedings of the International Conference on Natural Computation*, pp. 617–626, Springer, Berlin, Heidelberg, September 2006.

[20] Z. N. Idan, S. H. Abdulhussain, B. M. Mahmmod, K. A. Al-Utaibi, S. A. R. Al-Hadad, and S. M. Sait, "Fast Shot Boundary Detection Based on Separable Moments and Support Vector Machine," *IEEE Access*, vol. 9, pp. 106412–106427, 2021.

[21] Y. Zheng and Y. Zhang, "Abrupt shot boundary detection with combined features and SVM," in *Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pp. 409–413, IEEE, Chengdu, China, October 2016.

[22] E. Kasutani and A. Yamada, "The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-Speed Image/video," in *Proceedings of the 2001 International Conference on Image Processing*, vol. 1, pp. 674–677, Thessaloniki, Greece, October 2001.

[23] R. Zabih, J. Miller, and K. Mai, "Feature-based algorithms for detecting and classifying scene breaks," in *Proceedings of the third ACM international conference on Multimedia*, vol. 49, no. 9, pp. 3207–3220, Cornell University, San Francisco, California, USA, January 1995.

[24] A. Singh, D. M. Thounaojam, and S. Chakraborty, "A novel automatic shot boundary detection algorithm: robust to illumination and motion effect," *Signal, Image and Video Processing*, vol. 14, no. 4, pp. 645–653, 2020.

[25] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*, National Taiwan University, Taipei, Taiwan, 2003.

[26] J. T. Jose and S. Rajkumar, "Multiple grey-scale feature based shot boundary detection," in *Proceedings of the 2021 Asian*

Conference on Innovation in Technology (ASIANCON), IEEE, PUNE, India, August 2021.

[27] J. Mondal, M. K. Kundu, S. Das, and M. Chowdhury, "'Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8139–8161, 2018.

[28] A. Sasithradevi and S. Mohamed Mansoor Roomi, "A new pyramidal opponent color-shape model based video shot boundary detection," *Journal of Visual Communication and Image Representation*, vol. 67, Article ID 102754, 2020.

[29] S. Chakraborty, A. Singh, and D. M. Thounaojam, "A novel bifold-stage shot boundary detection algorithm: invariant to motion and illumination," *The Visual Computer*, vol. 38, no. 2, pp. 445–456, 2022.