Iraqi Journal of Science, 2023, Vol. 64, No. 8, pp: 4260- 4277
DOI: 10.24996/ijs.2023.64.8.43

ISSN: 0067-2904

Extracting Descriptive Frames from Informational Videos

Eman Hato
Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq

Received: 15/8/2022 Accepted: 16/12/2022 Published: 30/8/2023

Abstract

Informational videos are becoming increasingly important among all video types. The users spend so much time browsing the informative videos, even if they are not interested in all their topics. Thence, a new method for extracting descriptive frames is presented in this paper that allows users to navigate directly to the topics of their interest in the video. The proposed method consists of three main phases: video preprocessing, video segmentation, and the video separation phase. Firstly, frames are extracted from the videos, resized, and converted to grayscale. Then, the frames are divided into blocks, and the kurtosis moment is calculated for each block. The videos are segmented based on an examination of the differences between the features of the kurtosis moment. Lastly, the informative frames are grouped into a separate video after they are distinguished from the uninformative ones using the clustering technique. The results demonstrated the functional effectiveness of the proposed method. According to the accuracy and F1-Score measures, it has a performance of up to 100%. Moreover, the video is significantly summarized by reducing the duration to less than 1% of its original time.

Keywords: Informational videos, Moment order, Kurtosis, Skewness, Video summarization.

إستخلاص المقاطع الوصفية من ملفات الفيديو الإعلامية

إيمان هاتو

قسم علوم الحاسوب, كلية العلوم, الجامعة المستنصرية, بغداد, العراق

الخلاصة

تكتسب مقاطع الفديو الإعلامية أهمية خاصة من بين جميع أنواع مقاطع الفديو الأخرى. يقضي المستخدمون الكثير من الوقت في تصفح مقاطع الفديو الإعلامية ، حتى لو لم يكونوا مهتمين بجميع موضوعاتها. ولهذا ، تم تقديم طريقة جديدة لاستخراج المقاطع الوصفية في هذا البحث تسمح للمستخدمين بالانتقال مباشرة إلى الموضوعات التي تهمهم في الفديو. تتكون الطريقة المقترحة من ثلاث مراحل رئيسية: المعالجة المسبقة للفديو ، وتجزئة الفديو ، ومرحلة فصل الفديو. أولاً ، يتم استخراج الصور من مقاطع الفديو وتغيير حجمها وتحويلها إلى الصور الرمادية. بعد ذلك ، يتم تجزئة الصور إلى أقسام ويتم حساب kurtosis moment لكل قسم. يتم تقسيم مقاطع الفديو بناءً على فحص الاختلافات بين ميزات kurtosis moment أخيرًا ، يتم تجميع المقاطع الوصفية في مقطع فديو منفصل بعد تمييزها عن تلك غير الوصفية باستعمال تقنية التجميع. أثبتت النتائج الفعالية الوظيفية للطريقة المقترحة. حيث انها اظهرت أداء يصل إلى 100٪ وفقًا لمقاييس الدقة و F1-

*Email: emanhato@uomustansiriyah.edu.iq

Score. علاوة على ذلك ، يتم تلخيص الفديو بشكل كبير من خلال تقليل المدة الزمنية إلى أقل من 1٪ من وقته الأصلي.

1. Introduction

Education has changed drastically over the years, especially with the advent of online learning. On-line videos such as lectures and informational or educational videos are the more notable resources for keeping learners engaged with the course. The Internet enabled educational institutions and universities to keep video lectures online and accessible at any time [1].

The availability of a large number of educational videos has greatly increased the need for indexing, which is the key to content-based retrieval and distance learning. The most common way to accomplish this is to segment video into meaningful units of desired information. The essence of video segmentation techniques is to compare successive frames by focusing on changing the visual features of the video frames [2, 3].

For video indexing, metadata is associated with the video, which includes linking the description of the topics or contents to the corresponding segments of the video. The metadata is oftentimes prepared manually, and this is impractical. Therefore, metadata generation techniques are required that can look at the video content, extract information, and prepare the video metadata to provide accurate results to the users [4,5].

The importance of video indexing and video segmentation systems has led to many studies. Details can be found in the research [6, 7].
Usually, educational videos come in many forms. For example, the entire frame displays only the lecture slides, the frame is divided into two parts, one for the slides and one for the speaker, or the slides and the speaker are shown in the same frame [8].

Sometimes the required information may be covered only for a few minutes in the whole video, so the user would like to view only that particular part of the video without watching the whole video. For this reason, a new method has been proposed in this paper to extract descriptive frames from informational videos.

2. Related works

Education Despite all the advantages of educational and lecture videos, a considerable amount of time is spent watching and rewatching these videos. Detecting the informational content of videos can help users focus on the important parts, such as annotated slides or subtitles. Therefore, many methods were proposed to automatically extract the topic structure of informational video through performing segmentation processes and analyzing frame features.

H. J. Jeong et al. [9] designed a method to detect forward and backward change in lecture videos. First, the frame regions are detected to extract the SIFT features. Then, the features are compared to determine the similarity between video frames. The mean and standard deviation are used to estimate the threshold. The method achieved 87% and 86% accuracy in forward and backward transition detection, respectively.

V. Balasubramanian et al. [10] proposed a video summarization system that extracts a set of key phrases and topic-based segments from a lecture video. Key phrases are described in a lecture that extracts information based on visual and audio features with two levels of

classification. Naive Bayes classifier is employed at the first level. Then, a rule-based refiner is applied to arrive at a definitive list of key phrases. Topic-based segments represent a set of temporal segments within lecture videos that cover a certain subtopic of lectures. The TexTiling algorithm is used to achieve topic-based segmentation. According to the F1 Score evaluation, the system obtained a score of 0.464 for keyphrase extraction and 0.619 for topic-based segmentation.

In [11], X. Che et al. introduced two methods to highlight the online lectures. In the sentence-level method, the acoustic emphasis of the audio signal is analyzed to determine the importance of the sentences. Sentences with greater importance are presented in subtitle files as the final result. The importance value in the segment-level method is calculated based on the relationship between the lecture speech and the slide content. Subjective evaluation achieved an accuracy of 63.6% for the sentence-level method and 70.2% for the segment-level method.

A framework to retrieve parts from lecture videos was performed by B. J. Sandesh et al. [12]. First, the audio is extracted and converted to a text file. At the same time, the topic keywords are detected from the corresponding web documents using the non-negative matrix factorization (NMF) topic modeling technique. Second, the frequency and position of topic keywords are specified in the text file. Then, the text file is segmented and indexed based on the topic's keywords. Lastly, the video is segmented according to an index text file. The result showed that the average error rate was 25 time units.

B. S. Daga et al. [13] proposed a lecture video retrieval system. The feature mixture database is created from the extracted key frames. The videos are searched using the feature mixture database and a hybrid classifier based on K-Nearest Neighbors and Naive Bayes classifiers. The accuracy was 0.8519, according to F-Score.

Z. Liu et al. introduced a method for detecting slide transitions in [14]. The lecture video is first segmented based on feature detection and matching. Each segment is considered a node in a sparse time-varying graph, which has the ability to model the transition from one slide (segment) to another. Then the adjacency matrix of the graph is generated. The changes between adjacency matrices reflect the slide transition. As the method's average performance, the F-Score was 0.904.

In [15], E. R. Soares and E. Barrere presented a method to segment video lectures. Three feature vectors are generated. The first one represents audio frequency chunks and power spectral density. The second vector represents the audio chunks of the audio transcription features. And the third vector represents the features from knowledge base searches and semantic annotation. Then, the affinity matrix that combines all features is created to cluster the audio chunks into classes that represent the video topics. The spectral clustering algorithm is used to do the clustering process. The F-Score was 0.85 for the performance of the method on YouTube video lectures.

Z. Liu et al. [16] presented a method to detect the slide transitions in lecture videos using the 3D convolutional neural network (3D ConvNet) and residual network (ResNet). The 3D ConvNet was employed to extract the spatiotemporal features of video frames. The ResNet was utilized to facilitate the network training. The method performance according to the F-Score was 0.914 for the lecture video.

M. Guan et al. [17] produced a method to detect slide transitions in lecture videos based on the integration of a 3D convolutional network (3D ConvNet) and a dual path network (DPN). The main idea of the integration is to override the limitations of the 3D ConvNet, like time and memory costs. And take advantage of the DPN, which not only saves memory and time but also extracts more effective features and improves training results. The accuracy result was 0.96 with an F-Score on the lecture in the video library.

A method of slide change detection was performed by A. Sindel et al. in [18]. The proposal consists of two stages. In the first, each video frame is processed by a 2D convolutional neural network to predict slide transition candidates. In the second stage, the transition candidates are refined by 3-D convolutional neural networks. The accuracy of the evaluation results was 89.85 with an F-score. Table 1 summarizes the related works.

Table 1: Summary of related work

Ref.	Techniques	Limitation	Aim	Dataset	Measures
[9]	SIFT	Computation cost is too high.	Detect change in lecture videos	Lecture videos from YouTube	F1-Score Accuracy
[10]	Naive Bayes classifier and rule-based refiner	Rather complicated	Extracting metadata, that best summarizes the lecture content.	Pre-recorded video lectures	F1-Score Precision Recall
[11]	Short-Term Energy	Watching whole video	Highlighting sentences with bold font, or underline	Lecture videos from Massive Open Online Courses	Subjective evaluation
[12]	Non-Negative Matrix Factorization	The indexing is done using the topic keywords extracted from web	Creates index table along with the topic keyword	Lecture videos from internet.	Compared start time of the topic segments and hand annotated segment.
[13]	Naive Bayes, K-Nearest Neighbor classifier and Optical Character Recognition	Very complicated	Extracting the keywords from the video frames.	Lecture videos from internet.	F1-Score
[14]	SIFT	Just segmented the video	Detect slides transitions in lecture videos.	Lecture videos from Yale University Courses and YouTube	F1-Score
[15]	Fundamental frequencies and power spectral density	Matching related content with audio	Topic segmentation in video lectures.	Lectures videos from Videoaula repository	F1-Score, Mean Recall
[16]	3D ConvNet , ResNet and SIFT	Very complicated and time consumer	Detect slide transition in lecture videos.	lecture videos from Stanford University Courses, and YouTube	F1-Score precision, mean recall
[17]	3D Convolutional Networks and Dual Path Network	Costs much training time	Detect slide transition in lecture videos.	Pre-recorded video lectures by author	F1-Score
[18]	2-D convolutional neural network	The extracted slides in the video are not separated	Slide extraction from the lecture videos	Lecture videos from courses in the field of Pattern Recognition.	*F*1 – Score recall, and precision.

Through previous review, one can conclude that the main objectives of the researchers are to enable users to find topics of interest in the video content that may be irrelevant to them. The contribution of this paper is extracting the frames containing the educational information that abstracts the video content and compiling them into a separate video. This makes it easier for users to find the video content they're looking for.

3. Image Moments

Moments are widely used in the fields of image analysis, pattern recognition, and related fields. The moments are able to describe the image completely and encode its contents in a compact way due to their useful mathematical properties, especially geometric invariance. [19].

The moments of an image are certain particular weighted averages (moments) of the pixel intensities. They are used as image features and properties such as color, texture, shape, orientation, or position. The type of described image information is determined by the moment's order. The image's order moment is defined as [20, 21]:

$$m_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} x^p \, y^q f(x,y) \quad p,q = 0,1,2, ... \tag{1}$$

Where (p+q) is moment order, f (x, y) is an image, N is image width, and M is image height. The central moment is calculated as follows [20, 21]:

$$\mu_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} (x-\bar{x})^p \, (y-\bar{y})^q f(x,y) \quad p,q = 0,1,2, ... \tag{2}$$

Were $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$ and they are the coordinates of the gravity center of the image. The normalized central moment is calculated as follows [20, 21]:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}{}^{\gamma}} \qquad \gamma = \frac{p+q}{2} \qquad p+q = 2,3, ... \tag{3}$$

The seven invariant moments are derived using the second-order and third-order normalized central moments, and they are constructed as [20, 21]:

$$M_1 = \eta_{20} + \eta_{02} \tag{4}$$

$$M_2 = (\eta_{20}-\eta_{02})^2 + 4\eta^2{}_{11} \tag{5}$$

$$M_3 = (\eta_{30}-3\eta_{12})^2 + (3\eta_{21}-\eta_{03})^2 \tag{6}$$

$$M_4 = (\eta_{30}+\eta_{12})^2 + (\eta_{21}+\eta_{03})^2 \tag{7}$$

$$M_5 = (\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2 - 3(\eta_{21}+\eta_{03})^2] + (3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2 - (\eta_{21}+\eta_{03})^2] \tag{8}$$

$$M_6 = (\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2 - (\eta_{21}+\eta_{03})^2] + 4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03}) \tag{9}$$

$$M_7 = (3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2 - 3(\eta_{21}+\eta_{03})^2] + (3\eta_{12}-\eta_{30})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2 - (\eta_{21}+\eta_{03})^2] \tag{10}$$

The invariant moments are invariable to image translation, rotation, and scaling. The low-order moments represent the coarse features of the image. High-order moments, on the other hand, represent more detailed features. As the order of the moment increases, the detail of the image representation also increases [22]. The expected value is the first moment, the variance is the second central moment, the skewness is the third standardized moment, and the kurtosis is the fourth standardized moment [23].

The skewness moment is a measure of the asymmetry of the data distribution around its mean. A distribution is asymmetrical when its left and right sides are not mirror images.

Kurtosis moment is a measure of the sharpness of a peak relative to a normal distribution in a frequency distribution curve. Its value describes the heaviness of the distribution [23, 24].

4. The proposed Method Structure

A lot of time can be spent finding topics of interest in informational videos. Extracting informative slides enables the user to navigate in a nonlinear manner through all the topics of the video's educational content. However, extracting educational material is not an easy task and needs automation. The main goal of this paper is to separate the informative frames into an independent video from the rest of the video content. This in turn helps shorten the video time and takes the user to the important topics directly. The general structure of the proposed method with three phases is schematically illustrated in Figure 1. The following subsections provide a detailed explanation of these phases.

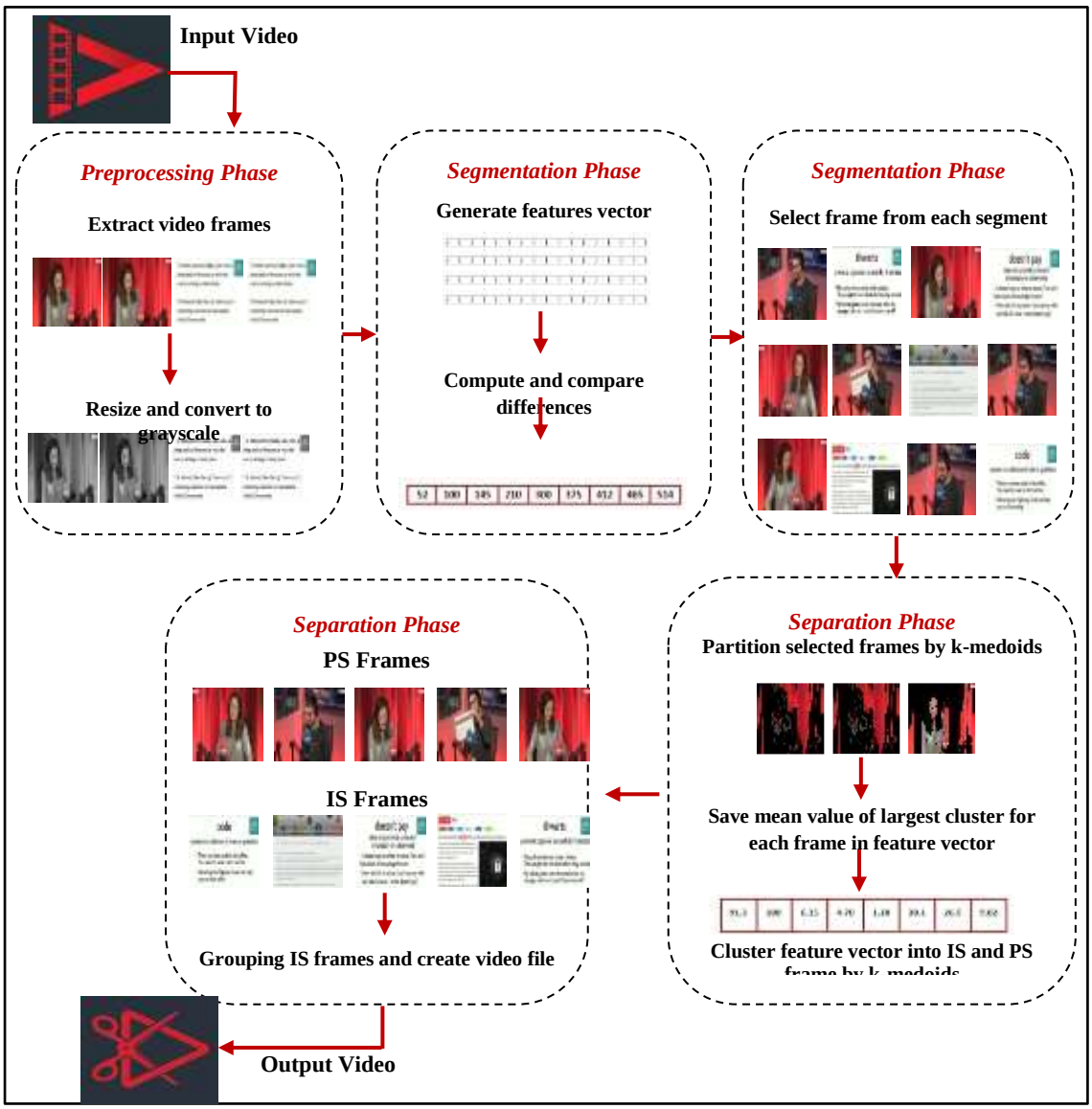

Figure 1: Structure of proposed method

4.1 Video Preprocessing Phase

As mentioned earlier, informational videos come in different formats. In this paper, the informational videos from the BBC Learning English YouTube channel were adopted as test materials. The videos have a relatively fixed structure. They display a series of frames that

include presenters or interviewers and a series of frames that include informative content, often in an alternate fashion.

First, the input video file is read, and the frames are extracted from it. Then all frames are processed by resizing them to 256×256 and converting them to a grayscale image in order to be the input to the next phase. Reducing the frame size helps speed up performance and reduce processing time without affecting the result because of the use of variable scaling features.

4.2 Video segmentation Phase

The purpose of video segmentation is to enable the discovery of the temporal structure of video by splitting the video into meaningful parts. Using the features of moments, the video frame sequence is divided into segments. Measuring the moment similarity between two adjacent frames enables the specification of segment transitions. This is based on the assumption that the moment distributions of frames in the same segment are more similar than those in different segments.

Initially, each frame is divided into 4x4 blocks. Then, the kurtosis moment is calculated for each block. The kurtosis values for each block are aggregated into a vector of length 16. This, in turn, represents the feature vector of the frame, as shown in Figure 2.

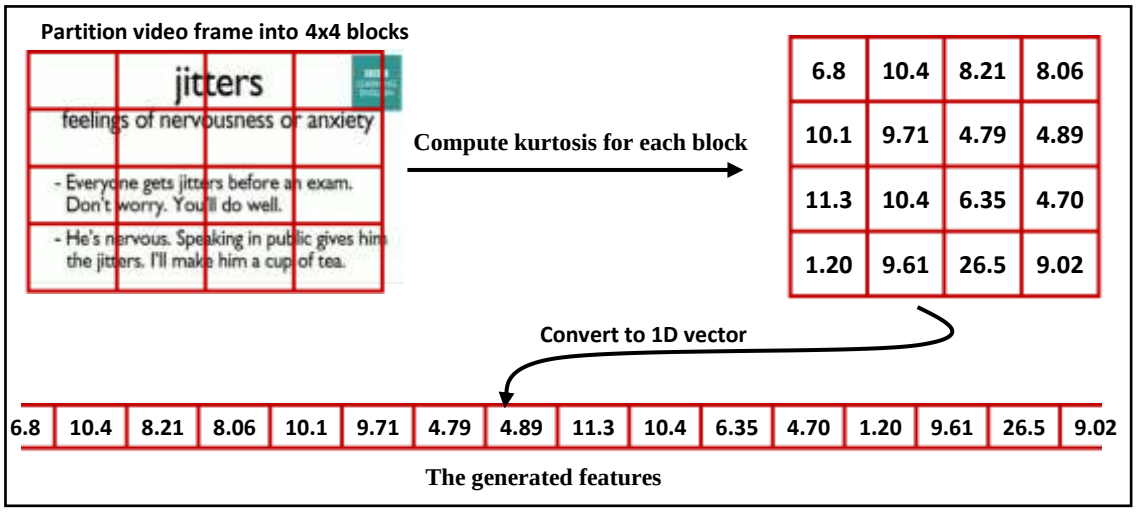

Figure 2: The features vector extraction

Using the mean absolute differences (MAD) measure, the feature vectors of sequential frames are compared in order to capture segment transitions, especially those between frames with interviewers and frames with informative content. Whenever the differences between two successive feature vectors indicate a difference of more than half of the feature vector values, shot transitions are recorded. This means that more than half of the number of frame blocks have difference values greater than zero as a result of the MAD measure, and so the successive frames are dissimilar. It was critical to select frames in order to represent the video segment with its main contents while avoiding excessive repetition between frames in a single segment. Thus, one frame is selected randomly from each video segment.

For simplicity, in this method, the frame representing the video segment with the interviewer or presenter will be referred to as the "Presenter Segment," and the frame representing the video segment with informative content will be referred to as the "Informative Segment." Algorithm 1 illustrates the operation sequences of the video segmentation phase.

Algorithm 1: Video Segmentation

Input: Video File.
Output: Selected Frames Indicators (SFI).

Start
Step1: IV← Extract frames of Video File.
FN ← Extract frames number of IV.
FVector← Initialize features vector of size FNx16.
DVector← Initialize differences vector of size FNx16.
K← Assign value 1 for the counter.
Step2: For I = 1 to I < FN
IV [I] ← Resize IV [I] to 256x256.
Frm ← Convert to grayscale IV [I].
Frm← Partition IV [I] into 4x4 blocks.
// For each block extract kurtosis moment features
For J = 1 to 16
FVector [I , J] ← Extract kurtosis feature from Frm block [J].
End For
End For
Step3: For 1 to I < FN-1
//Compute differences vector
DVector [I, 1:16] = Absolute (FVector [I+1,1:16] - FVector [I,1:16])
// Check the number of DVector row values that are greater than zero
Nonzero ← compute the numbers of non- zero elements values in DVector [I, 1:16]
IF (Nonzero > 7)
// Save the indicators of transitions segments.
SIndicator [k] ← Save I value
K ← K+1
End IF
End For
Step4: For I = 1 to I <= K
//Merge the value 1 and the value FN to begin and end of SIndicator
SIndicator ← [1, SIndicator , FN]
// Select one frame randomly from each video segment
SFI [I] ←Select value randomly from range SIndicator [I] and SIndicator [I+1]
End For
End

4.3 Video Separation Phase

In this phase, the goal is to distinguish the frame PS from the frame IS and then assemble all segments with frames of type IS into a separate video. The video separation phase is done with the help of the K-Medoids clustering technique. K-Medoids is an unsupervised learning algorithm that is better in terms of reducing noise and being non-sensitive to outliers as compared to other clustering algorithms.

By analyzing and comparing the layout patterns of the IS and PS frames, we conclude that the largest part of the IS frame is the white background. This is in contrast to the PS frame, which is often a mixture of different colors, as shown in Figure 3.

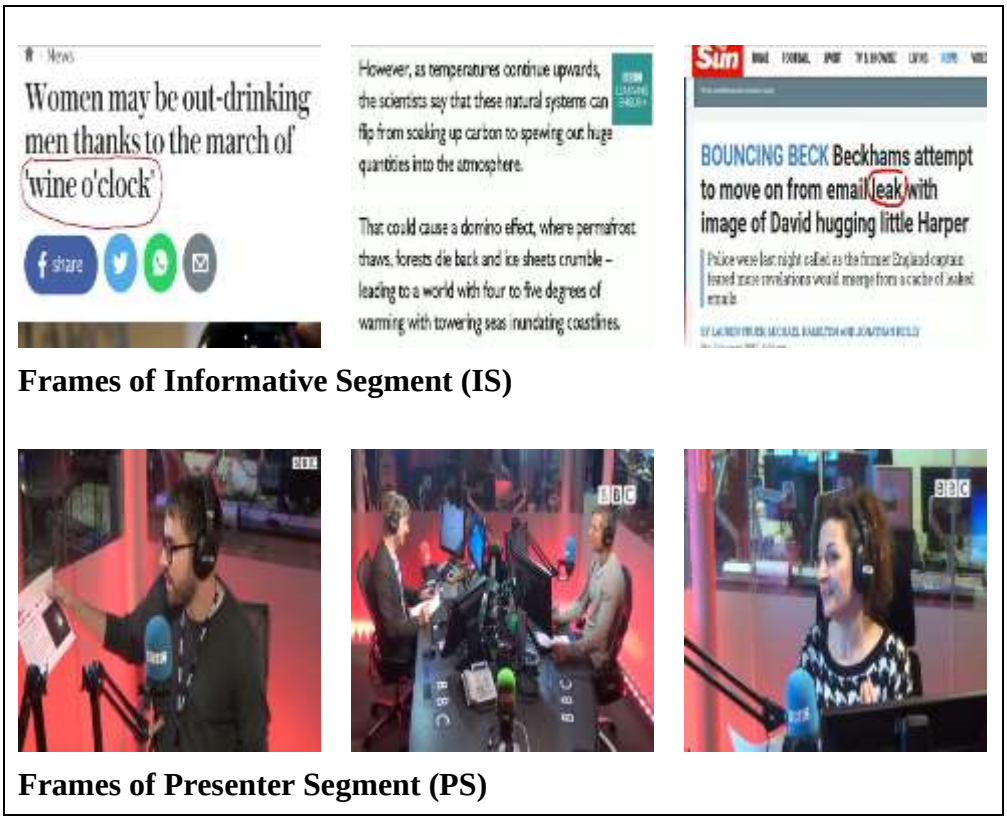

Figure 3: Types of frame in BBC learning English videos

Based on this observation, each selected frame of the video is divided by the K-Medoids algorithm into three clusters according to its color value. The mean for each cluster is calculated. So, each selected frame will have a vector of three values. Then the largest value of this vector is chosen to be associated with the frame as its feature. In other words, each selected frame, whether it is an IS or PS, will be represented by the color mean value of the largest cluster within it.

The largest mean value for the IS frame is high because these pixels belong to the white background, which occupies the frame's largest cluster. And on the contrary, for the PS frame, the mean values are almost identical and distributed according to the presence of colors, which makes their largest value somewhat low compared to the values of the IS frame. Thus, it will be easier to recognize IS from the PS frame. Figure 4 illustrates the graphical visualization of the frame clustering process according to the color distribution.

Once the selected frames for each video are represented by a feature of the mean value of the largest cluster, the K-Medoids algorithm is used again to classify the IS from the PS. In the last step, the video frames of type IS are grouped and stored in a separate video. The separated video can be utilized by users who are interested in the informational content of the video without the rest of the video details. The implemented steps of the video separation phase are given in Algorithm 2.

Figure 4: Clustering process of IS and PS frames

Algorithm 2: Video Separation

Input: Selected Frames Indicators (SFI).
Output: Video File of Informative Segments (VFIS).

Start
Step1: CF← Assign value 3 for the clusters number of the frame.
CVF ← Assign value 2 for the clusters number of the video frames.
Length ← Compute the length of Selected Frames Indicators (SFI).
FVector← Initialize features vector of size Length.
Result ← Initialize output vector of size Length.
Step2: For I = 1 to I < Length
Frm← Get the frame from IV [SF[I]].
// Partition Frm into 3 clusters using k-medoids.
Indexes ← Apply k-medoids (Frm, CF).
Cluster 1 ← Extract cluster 1 (Indexes = 1) from Frm.
Cluster 2 ← Extract cluster 2 (Indexes = 2) from Frm.
Cluster 3 ← Extract cluster 3 (Indexes = 3) from Frm.
// Compute the mean value for each cluster
Clustermean [1] = Apply Mean (Cluster1)
Clustermean [2] = Apply Mean (Cluster2)
Clustermean [2] = Apply Mean (Cluster3)
// Save the mean value of the largest cluster
FVector [I] = Apply Maximum (Clustermean)
End For
Step3:
// Cluster the FVector into 2 clusters (IS and PS) using k-medoids.
Result ← Apply k-medoids (FVector, CVF).
Step4: For I = 1 to I < Length
IF (Result [I]= 'IS')
Save the frame IV [SFI [I]] to the IS Folder.

End IF
End For
VFIS ← Create video file from the frames in the IS Folder.
End

5. Experimental Results and Discussion

To evaluate the proposed method, twelve video files were randomly selected from the BBC Learning English Channel News Review Podcast on YouTube, and their ground truth was manually annotated. The details of the video properties are listed in Table 2.

Table 2: Video characteristics

Videos	Frames Number	Transitions Segment Number	Informative Segment Number	Uninformative Segment Number
IV01	15196	81	8	73
IV02	12453	45	8	37
IV03	14954	86	7	79
IV04	12612	65	7	58
IV05	18302	102	7	95
IV06	11581	53	5	48
IV07	11557	67	5	62
IV08	15310	70	8	62
IV09	14285	85	7	78
IV10	13730	82	7	75
IV11	16591	83	8	75
IV12	15505	94	8	86

To measure the accuracy of video segmentation, the F1-Score and accuracy were used, which are well-known measures of accuracy in statistics. Evaluation tests were performed on the proposed method to detect the transition segment, and the results are listed in Table 3. The values of F1-score and accuracy indicate the high performance of the proposed method.

Table 3: Video segmentation accuracy of transitions detection

Videos	Accuracy	F1 - Score
IV01	1	1
IV02	1	1
IV03	1	1
IV04	1	1
IV05	1	1
IV06	1	1
IV07	1	1
IV08	1	1
IV09	1	1
IV10	1	1
IV11	1	1
IV12	1	1

This is because the proposed method of video segmentation employs kurtosis moments that are not sensitive to the geometric transformations of the image. In addition, kurtosis moments provide sufficient information about data frequency distribution. A comparison of video segmentation results was performed, as illustrated in Table 4, to assess the accuracy of transitional segment detection based on the use of different moments. Since the high-order moments are more complex, it will be sufficient to compare the results of the first four moments. The comparison involved the execution time to extract the moment's features for a single frame divided into 4x4 blocks. The results are listed in Table 5. Also, the comparison results have been depicted as revealed in Figure 5, Figure 6, and Figure 7.

Table 4: Video segmentation results for various moments

Videos	M1		M2		M3		M4	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
IV01	0.92	0.95	0.95	0.97	0.53	0.69	**1**	**1**
IV02	0.93	0.96	1	1	0.95	0.97	**1**	**1**
IV03	0.90	0.95	1	1	0.30	0.46	**1**	**1**
IV04	0.91	0.95	0.94	0.97	0.87	0.93	**1**	**1**
IV05	0.86	0.92	0.92	0.96	0.80	0.89	**1**	**1**
IV06	0.96	0.98	1	1	0.94	0.97	**1**	**1**
IV07	0.98	0.99	0.98	0.99	0.51	0.68	**1**	**1**
IV08	0.78	0.88	0.95	0.97	0.97	0.98	**1**	**1**
IV09	0.96	0.98	0.97	0.98	0.16	0.28	**1**	**1**
IV10	0.82	0.90	0.97	0.98	0.97	0.98	**1**	**1**
IV11	0.93	0.96	1	1	0.98	0.99	**1**	**1**
IV12	0.96	0.98	0.98	0.99	0.80	0.89	**1**	**1**

Table 5: Execution time for various moments

Moments	Time in Sec.
M1	0.00023
M2	0.00041
M3	0.00756
M4	**0.00759**

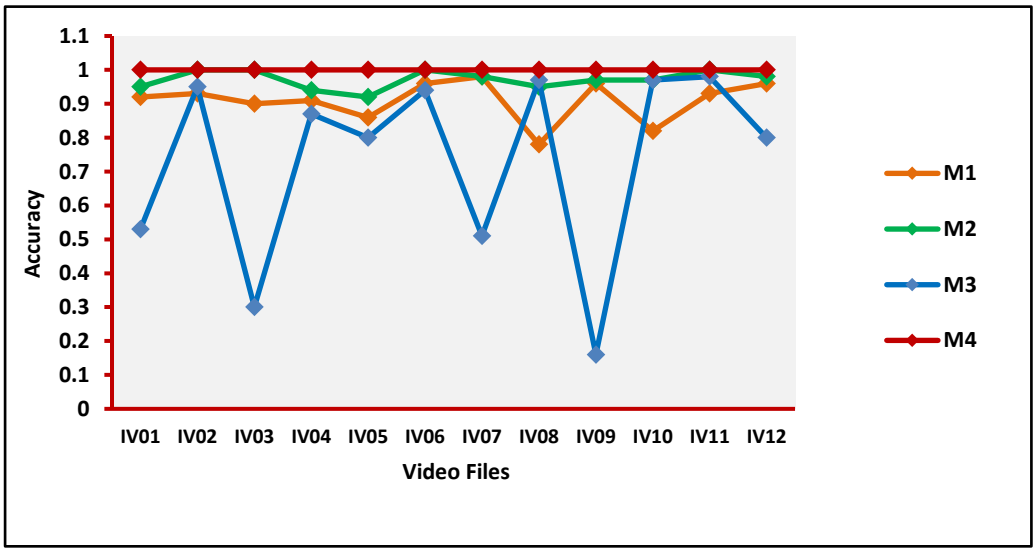

Figure 5: Video segmentation results for various moments according to

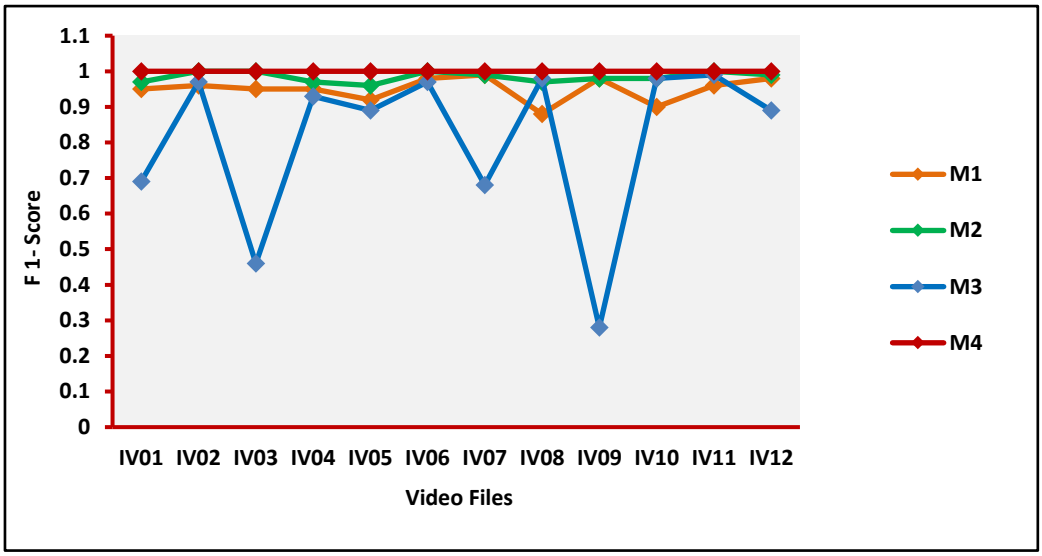

Figure 6: Video segmentation results for various moments according to F1- Score

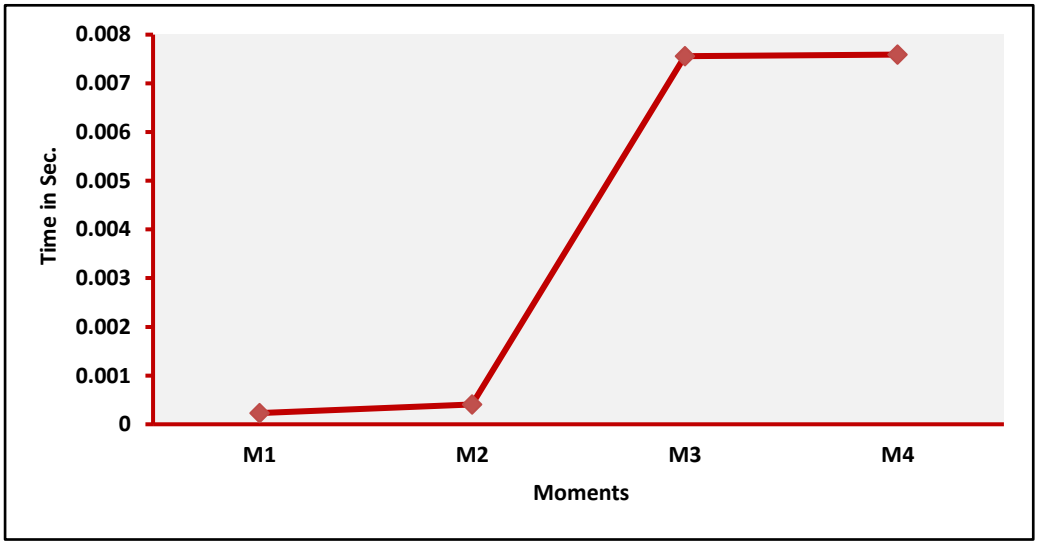

Figure 7: Execution time for various moments

The above two tests indicate that using the kurtosis moment as a feature is a good choice for distinguishing differences between successive frames. It can maintain good stability in geometric distortion. Even if the time consumed is less for some other types, this is at the expense of the high detection accuracy achieved by the kurtosis moments.

The frame partitioning process is another reason for the high accuracy of the proposed method of video segmentation. After applying frame partition, even very small changes are captured. As a result, segment transitions in the video file can be successfully detected. The experimental results show that the best partitioning is 4x4 blocks (i.e., 16 blocks per frame), which provides the highest accuracy. For demonstration, different divisions of the frame were tested, and their impact on the results of the detection accuracy was reported as illustrated in Table 6 and the corresponding Figures 8 and 9.

Table 6: Video segmentation results for various frame partitions

Videos	Without Partition		2x2 Partition		4x4 Partition		8x8 Partition		16X16 Partition	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
IV01	0.01	0.03	0.79	0.88	**1**	**1**	1	1	1	1
IV02	0.46	0.63	0.97	0.98	**1**	**1**	1	1	1	1
IV03	0.02	0.04	0.15	0.26	**1**	**1**	1	1	1	1
IV04	0.53	0.69	0.41	0.58	**1**	**1**	0.98	0.99	1	1
IV05	0.09	0.17	0.96	0.98	**1**	**1**	0.99	0.99	0.99	0.99
IV06	0.07	0.13	0.84	0.91	**1**	**1**	0.56	0.72	1	1
IV07	0.12	0.21	0.19	0.32	**1**	**1**	0.98	0.99	0.98	0.99
IV08	0.03	0.07	0.21	0.35	**1**	**1**	1	1	1	1
IV09	0.12	0.22	0.18	0.31	**1**	**1**	0.98	0.99	0.94	0.96
IV10	0.04	0.08	0.94	0.97	**1**	**1**	0.98	0.99	0.98	0.99
IV11	0.24	0.39	0.41	0.58	**1**	**1**	1	1	1	1
IV12	0.06	0.11	0.29	0.45	**1**	**1**	1	1	1	1

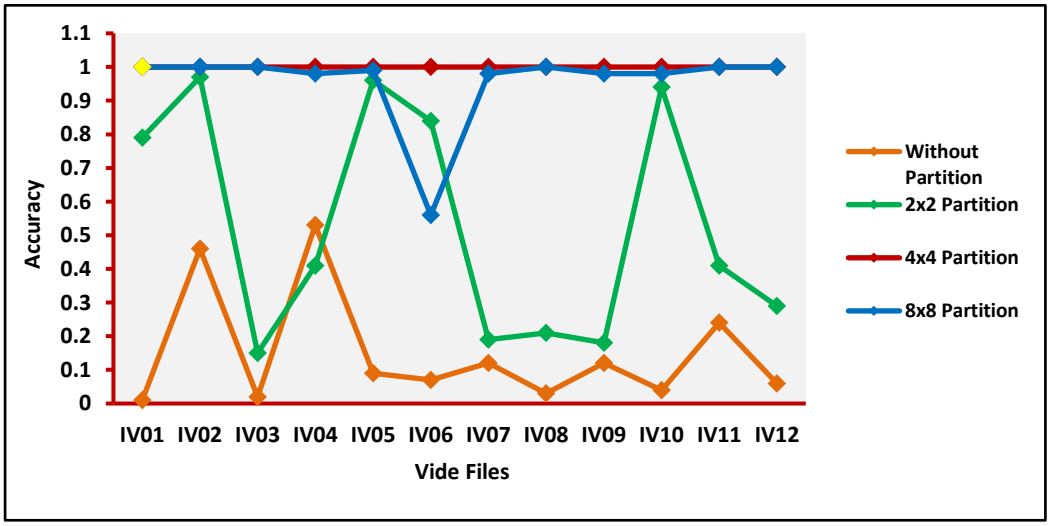

Figure 8: Video segmentation results for various frame partitions

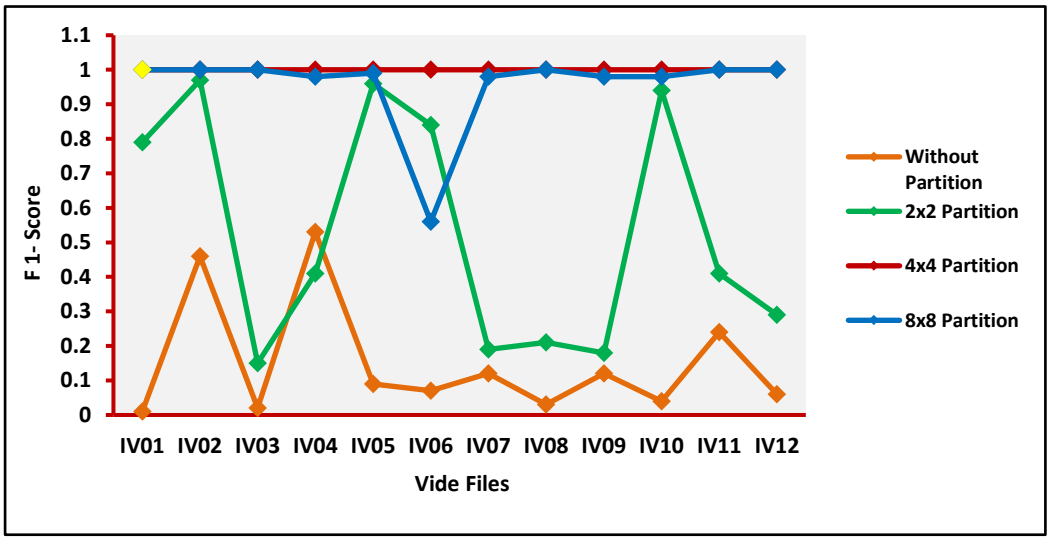

Figure 9: Video segmentation results for various moments according to F1-

By comparing the third column in Table 5, which is bold, with other columns, it is noticeable that 4x4 frame partitions have the highest accuracy detection in comparison with other frame partitions. It is obvious that the execution time also increases slightly when the partition is increased. Table 7 reports the required time for the extraction of the moment's features corresponding to different frame partitions, and a clear picture is given in Figure 10.

Table 7: Execution time for various frame partitions

Frame Partitions	Time in Sec.
Without Partition	0.00886
2x2 Partition	0.04163
4x4 Partition	**0.04282**
8x8 Partition	0.04963
16X16 Partition	0.05586

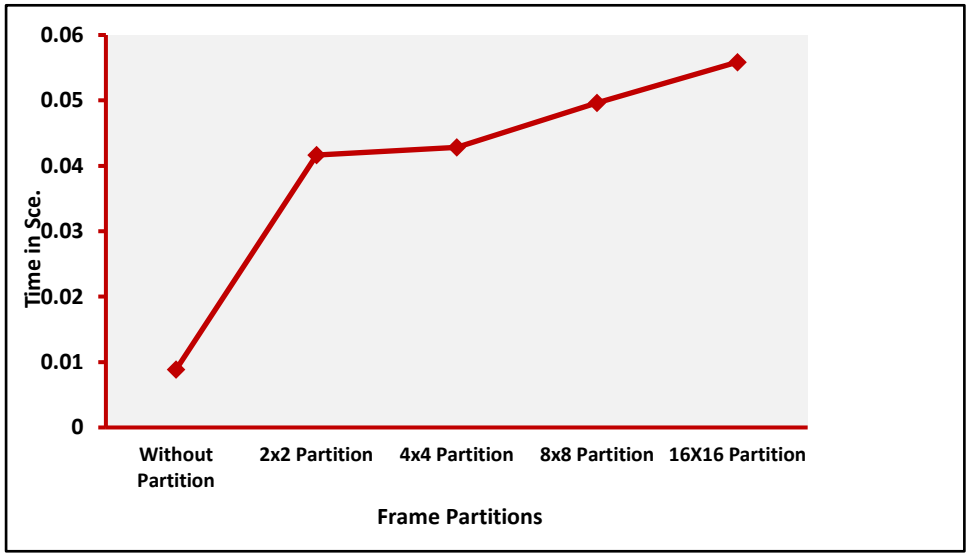

Figure 10: Execution time for various frame partitions

As the aim is frame partitioning, it gave high detection results with reasonable uptime. Therefore, the design of the proposed method uses 4x4 blocks per frame. The other partitions were excluded because they are less accurate. Furthermore, the consuming time on frame partition is roughly equal, so the difference isn't significant.

Distinguishing and grouping IS frames is an essential process for effective video separation and can be experimental in two aspects: evaluating the performance of the proposed algorithm to distinguish between IS and PS selected frames and evaluating the final summarized video. To judge the validity of the detected IS and PS frames, the number of IS and PS was manually calculated for all selected frames as previously mentioned in Table 1. Then F1-Score and accuracy measures were computed to evaluate the performance of the proposed algorithm. Table 8 contains an illustration of the overall performance test results.

Referring to Table 8, it is evident that the proposed method has high detection accuracy. The process of dividing the frame into three regions helped recognize the largest region (i.e., the

background), which is one of the characteristics of the IS frames. This, in turn, was reflected in the accuracy of the distinction between IS and PS frames.

Tables 8: The results of distinguish descriptive frames

Videos	IS Frames Detection		PS Frames Detection	
	Accuracy	**F1- Score**	**Accuracy**	**F1- Score**
IV01	1	1	1	1
IV02	1	1	1	1
IV03	1	1	1	1
IV04	1	1	1	1
IV05	1	1	1	1
IV06	1	1	1	1
IV07	1	1	1	1
IV08	1	1	1	1
IV09	1	1	1	1
IV10	1	1	1	1
IV11	1	1	1	1
IV12	1	1	1	1

Time is an important aspect of the proposed algorithm that seeks to extract descriptive frames in order to reduce the time spent by the user. A comparison is made between the time duration of the original video file and the time duration of the video file that was prepared from informative frames only (descriptive frames) to show the effectiveness of the current proposal algorithm. The results are summarized below in Table 8.

The results from Table 9 obviously demonstrate that the time duration of the video files has been significantly reduced, which is in line with the expected results.

Table 9: The videos time duration comparison

Videos	Time Duration of Original Videos in Sec.	Time Duration of Descriptive Videos in Sec.
IV01	601	**0.266**
IV02	496	**0.266**
IV03	596	**0.233**
IV04	502	**0.233**
IV05	730	**0.233**
IV06	462	**0.166**
IV07	460	**0.166**
IV08	612	**0.266**
IV09	570	**0.233**
IV10	548	**0.233**
IV11	662	**0.266**
IV12	620	**0.266**

6. Conclusions

This paper presents a new method for extracting descriptive frames from informative videos and collecting them into an independent video. in order to avoid content that is irrelevant to their topics. The proposed method employed simple but effective techniques. The videos are segmented based on an examination of the differences between the features extracted by the kurtosis moments. The descriptive frames are then differentiated from other frames based on the K-Medoids clustering technique. Lastly, these frames are grouped into a separate video.

The experimental results demonstrated the efficacy of the proposed method, which achieved high accuracy according to the accuracy and F1-score measures. In addition, the separate video contains the essential content in a very short time. This allows users to concentrate on the most interesting part of the video without wasting time. The author seeks, as a future work, to enable the user to identify the important topics for him from the video and then cut them out in a separate video.

Acknowledgements

The author wishes to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) in Baghdad, Iraq, for its assistance with this work.

References

[1] E. R. Soares and E. Barrere, "An optimization model for temporal video lecture segmentation using word2vec and acoustic features," in *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, pp. 513-520, 2019, doi:10.1145/3323503.3349548.

[2] E. Hato, "Temporal Video Segmentation Using Optical Flow Estimation," *Iraqi Journal of Science*, vol. 62, no.11, pp. 4181-4194, 2021, doi: 10.24996/ijs.2021.62.11.36.

[3] E. Hato, and M. E. Abdulmunem, "Fast algorithm for video shot boundary detection using SURF features," in *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, IEEE, pp. 81-86, 2019.

[4] P. Eruvaram, K.Ramani and C. Shoba Bindu, "An experimental comparative study on slide change detection in lecture videos," *International Journal of Information Technology*, vol.12, no. 2, pp. 429-436, 2020, doi: 10.1007/s41870-018-0210-4.

[5] S. Pal, P. K. D. Pramanik, T. Majumdar and P. Choudhury, "A semi-automatic metadata extraction model and method for video-based e-learning contents," *Education and Information Technologies*, vol. 24, no. 6, pp. 3243-3268, 2019 doi:10.1007/s10639-019-09926-y.

[6] M. E. Abdulmunem and E. Hato, "Semantic based video retrieval system: survey," *Iraqi Journal of Science*, vol. 59 no. 2A, pp. 739-753, 2018, doi:10.24996/ijs.2018.59.2A.12.

[7] N. Gayathri and K. Mahesh, "A generic approach for video indexing," in *Proceedings of the International conference on Computer Networks, Big data and IoT*, Springer, Cham, pp. 701-708, 2018.

[8] C.Thomas, K. P. Sarma, S. S. Gajula, and D. B. Jayagopi, "Automatic prediction of presentation style and student engagement from videos," *Computers and Education: Artificial Intelligence*, vol. 3, no. 100079, 2022, doi:10.1016/j.caeai.2022.100079.

[9] H. J. Jeong, T. Kim, H. G. Kim, and M. H. Kim, "Automatic detection of slide transitions in lecture videos," *Multimedia Tools and Applications*, vol. 74, no. 18, pp. 7537-7554, 2015, doi: 10.1007/s11042-014-1990-6.

[10] V. Balasubramanian, S. G. Doraisamy, and N. K. Kanakarajan, "A multimodal approach for extracting content descriptive metadata from lecture videos," *Journal of Intelligent Information Systems*, vol. 46, no. 1, pp. 121-145, 2016.

[11] X. Che, H. Yang and C. Meinel, "Automatic online lecture highlighting based on multimedia analysis," *IEEE Transactions on Learning Technologies*, vol. 11, no. 1, pp. 27-40, 2017.

[12] B. J. Sandesh, S. Jirgi, S. Vidya, P. Eljer, and G. Srinivasa, "Lecture video indexing and retrieval using topic keywords," *International Journal of Computer and Information Engineering*, vol. 11, no. 9, pp. 1057-1061, 2017.

[13] B. S. Daga, A. A. Ghatol and V. M. Thakare, "Semantic enriched lecture video retrieval system using feature mixture and hybrid classification," *Advances in Image Video Process*ing, vol. 5, no. 3, pp. 1-19, 2017.

[14] Z. Liu, K. Li, L. Shen, and P. An, "Sparse Time-Varying Graphs for Slide Transition Detection in Lecture Videos," in *Proceedings of the International Conference on Image and Graphics*, pp. 567-576. Springer, Cham, 2017.

[15] E. R. Soares, and E. Barrere, "Automatic Topic Segmentation for Video Lectures Using Low and High-Level Audio Features," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pp. 189-196. 2018.

[16] Z. Liu, K. Li, L. Shen, R. Ma, and P. An, "Spatio-Temporal Residual Networks for Slide Transition Detection in Lecture Videos," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 13, no. 8, pp. 4026-4040, 2019.

[17] M. Guan, K. Li, R. Ma and P. An, "Convolutional-Block-Attention Dual Path Networks for Slide Transition Detection in Lecture Videos," in *Proceedings of International Forum on Digital TV and Wireless Multimedia Communications*, Springer, Singapore, pp. 103-114, 2019.

[18] A. Sindel, A. Hernandez, S. H. Yang, V. Christlein, and A. Maier, "SliTraNet: Automatic Detection of Slide Transitions in Lecture Videos using Convolutional Neural Networks," *arXiv preprint arXiv:2202.03540*, 2022.

[19] X. Hua, H. Hong, J. Liu and Y. Shi, "A novel unified method for the fast computation of discrete image moments on grayscale images," *Journal of Real-Time Image Processing*, vol. 17, no. 5, pp.1239-1253, 2020, doi: 10.1007/s11554-019-00878-7.

[20] S. S. Gaikwad, S. L. Nalbalwar, and A. B. Nandgaonkar, "Devanagari handwritten characters recognition using DCT, geometric and hue moments feature extraction techniques," Sādhanā, vol. 47, no. 3, pp. 1-14, 2022, doi: 10.1007/s12046-022-01872-9.

[21] Y. Ren, J. Yang, Q. Zhang and Z. Guo, "Ship recognition based on Hu invariant moments and convolutional neural network for video surveillance," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1343-1373, 2020.

[22] N. D. Karampasis, I. M. Spiliotis and Y. S. Boutalis, "Real-time Computation of Krawtchouk Moments on Gray Images Using Block Representation," *SN Computer Science*, vol. 2, no. 2, pp. 1-15, 2021, doi: 10.1007/s42979-021-00536-5.

[23] N. Varish, "A modified similarity measurement for image retrieval scheme using fusion of color, texture and shape moments," *Multimedia Tools and Applications*, pp. 1-33, 2022, doi: 10.1007/s11042-022-12289-1.

[24] B. Attallah, A. Serir, and Y. Chahir, "Feature extraction in palmprint recognition using spiral of moment skewness and kurtosis algorithm," *Pattern Analysis and Applications*, vol. 22, no. 3, pp. 1197-1205, 2019.