Expert Systems With Applications 149 (2020) 113341

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Automatic lecture video skimming using shot categorization and contrast based features

Badri Narayan Subudhi[a,*], Thangaraj Veerakumar[b], Sankaralingam Esakkirajan[c], Santanu Chaudhury[d]

[a] Department of Electrical Engineering, Indian Institute of Technology Jammu, Nagrota, Jammu, 181221, India
[b] Department of Electronics and Communication Engineering, National Institute of Technology Goa, 403401, India
[c] Department of Instrumentation & Control Systems Engineering, PSG College of Technology, 641004, India
[d] Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur, 342037, India

ARTICLE INFO

Article history:
Received 24 January 2019
Revised 3 February 2020
Accepted 25 February 2020
Available online 28 February 2020

Keywords:
Video skimming
Capsule preparation
Histogram features
Contrast

ABSTRACT

Video skimming is one of the recently, getting popular technique for preparing preview for long watching video sequences. Most of the video skimming techniques developed in the literature uses manual intervention of users to prepare the review. Mostly the literature reported video skimming for sports and movie industries. In sports the portion of video where audience claps are used and in movie important contents are manually selected for preparing the preview. However in literature rarely any work reported for skimming of lecture video sequences. Lecture videos are generally, recorded indoor, low illuminated, noisy environment condition and contents of the scene rarely changes much. Hence designing an automatic skimming scheme is quite difficult task. In this article, we put forward an intelligent expert video skimming technique for lecture video sequences, where human intervention is not required. In the proposed scheme, initially the lecture video is segmented into a number of shots. We proposed the use of radiometric correlation technique for lecture video segmentation or finding the shot transitions. After getting the shot transitions in a video, the shots are recognized. The fuzzy K-nearest neighborhood technique is proposed to recognize the shots in a video. The shots are recognized into three categories: title slides, written texts/displayed slides and talking heads/writing hands. Three contrast based features: one existing i.e., average sharpness (AS) and two newly proposed: relative height (RH) and edge potential (EP) are used to find the contents of a frame. The frames with different contrast values are categorized to prepare the video skimming or the capsule. The media recreation is achieved by selecting a set of frames around these selected content frames. The effectiveness of the proposed scheme is demonstrated in this paper using five test sequences, including three NPTEL and two non NPTEL. It is also observed that the capsule prepared by the proposed scheme, provides a better preview of the actual sequence. The performance of the proposed scheme is tested by comparing it against three state-of-the-art techniques. The evaluation of the proposed scheme is carried out by using three evaluation measures. It is also observed that the proposed scheme is found to be better than that of the existing schemes.

1. Introduction

With increase in demand of multimedia in entertainment, advertisement and education process, there is a large demand for creating previews for large duration videos. In a movie mostly high resolution content frames with high frequency audio contents are chosen manually for preparing the trailer or the preview (Dagtas & Abdel-Mottaleb, 2001). It may be noted that for sports video sequence audio is the measure cue which is used for creating the capsule or the highlights. The major sound cue considered for such sequences are when audience usually shout or clap hands during a major events in the sports video or the commentator speaks loudly (Babaguchi, 2000; Rui, Gupta, & Acero, 2000). These parts of the video with sound are used for preparing the highlights of the sports sequence.

In the recent years, education through lecture video is getting its popularity as it helps in improving the knowledge of the stu-

* Corresponding author.
E-mail addresses: subudhi.badri@gmail.com (B.N. Subudhi), tveerakumar@yahoo.co.in (T. Veerakumar), rajanesakki@yahoo.com (S. Esakkirajan), santanuc@ee.iitd.ac.in (S. Chaudhury).

https://doi.org/10.1016/j.eswa.2020.113341
0957-4174/

dents and research scholars (Liu & Kender, 2004). However selection of the desired video to be watched is quite difficult and tedious process as someone needs to watch bunch of the entire duration video to select the desired one. Hence, a quick preview of the actual video may helps in; deciding the usefulness in watching the video before investing time and money (Zelnik-Manor & Irani, 2001).

It is to be noted that, most of the task for creating trailer/preview/highlight/capsule is done manually by the human intervention. The manual creation of the highlight or the capsule from a lecture video is quite time consuming and tedious process. Again watching a long duration sequence for several hours to select the key contents to prepare a capsule is quite complex and sometime may produces erroneous results. For a lecture video sequence most of the contents are similar and audio contents are rarely changing. The technique of manually preparing a trailer/preview for these videos is rather tedious. This motivates the researchers in designing of an expert intelligence system for creating the video capsule. The expert intelligent system is expected to create a knowledge based system that will emulate the human intervention in deciding the key contents to be included in preparation of the video capsule. The automatic way of capsule/highlight creation is popularly called as video skimming and is one of the popular and hot topic of research in the area of expert and intelligent system.

Video skimming has several important applications including: media creation, advertising, story board creation, etc. It is a kind of moving abstract creation and is getting its popularity in current scenario to attract the market. There are two ways of creating video abstract: video summary creation and video skimming. Video summary is a kind of still abstract, where the key frames selected from the video are presented for representing the abstract (Lienhart, Pfeiffer, & Effelsberg, 1997). Similarly, the video skimming is a moving abstract and is a collection of sequence of image frames with or without corresponding audio. Video skimming is popular in different field of video processing like creation of highlight for sports sequence, creation of trailer for entertainment industry like movie, theater & opera, news headline creation, etc (Rav-Acha, Pritch, & Peleg, 2006).

Although capsule creation or video skimming is new; a very few work has been reported in the literature (Subudhi, Veerakumar, Yadav, Suryavanshi, & Disha, 2017). The fundamental work on video skimming is reported by Smith and Kanede in Smith and Kanade (1997), where the authors proposed an integration of the language and image understanding techniques for video skimming. The authors have tried to delete some keywords that appear in close proximity or repeat throughout. The key phrase were selected within a small window. The audio and video synchronization is considered not to be aligned. One of the fundamental works reported in the literature for video key frame selection is reported in Gerek and Altunbasak (1997). In the said scheme, the authors have initially separated the *P, B* and *I* frames of the video. Then the histogram corresponding to the DC coefficients of each macro block of *I* frames are compared to find the scene transition or scene cut. However such an approach fails in large video data. A fast and novel video skimming technique is proposed in Martin and Lozano (2000) to manage large video database, where annotated video segments are used for retrieval task. It is observed that the video skimming technique is also used in many video indexing applications (Gao, Xin, & Ji, 2002). Shipman, Girgensohn, and Wilcox (2003) proposed a video skimming technique where hyperlinks were created for the test video to ease the task of video indexing. A video skimming technique is proposed for creating the highlights for American football sequences by Babaguchi, Kawai, Ogura, and Kitahashi (2004). The correspondence between the textual contents of image frames of the considered video is considered for creating the highlights. A principal component analysis (PCA) based motion synthesis technique is also developed by Egges, Molet, and Magnenat-Thalmann (2004), where irrelevant details in the video are removed from the original video to represent a compact video which represents the moving abstract of the test sequence. Use of latest machine learning scheme deep learning is also explored for video skimming. Yao, Mei, and Rui (2016) proposed a deep learning technique to differentiate between a video and video abstract. The said approach is mainly used to study human activities. Bastan, Gudukbay, and Ulusoy (2008) proposed an object based video indexing techniques for MPEG sequences. The said scheme uses JSEG technique (Deng & Manjunath, 2001) to segment each image frame in to number of meaningful regions. Further color, texture, shape, motion and inter regional features are trained in a support vector machine classifier to detect the important and non-important regions in a frame. The classified important regions are tracked in temporal direction to detect the important objects and hence indexing.

Ram and Chaudhuri (2009) proposed an automatic capsule creation scheme for lecture video sequences. This is one of the early and pioneered work reported in the state-of-the-art literature for lecture video skimming. Here the authors have used the hidden Markov model (HMM) to recognize the activities in earlier detected shots of the test video. Statistical parameters obtained from the probability distributions of the grey values are used to evaluate the objective qualities of the frames in a video and few high quality frames are extracted for creating the capsule. Peker and Divakaran (2004) proposed a video skimming technique where playback speed of the video is varied based on the the visual complexity of the scene. The authors have considered the principles of early vision which is applicable across a wide range of content type and applications. A computational attention methodology is developed by Ma and Zhang (2002) where the authors have proposed a motion attention model, which is able to detect the temporal subsegments of frames with high motion attention and is useful for video skimming. Recently, Subudhi et al. (2017) have proposed a new technique, where the contrast based features extracted from the histogram of the frames are used for shot boundary detection and the contents of the frames are analyzed for skimming of the lecture sequences.

It is to be concluded from the above analysis that; the literature in video skimming and lecture video skimming are very few. Most of the article reported in the area of video skimming are either developed for creating sports highlights or creation of movie trailer. However many of them are developed where human intervention is required. Most of the movie industry choose manual selection of specific scene for creating it capsule or movie trailer. Similarly, for sports sequences the capsule or the highlights are created by considering those parts of the video when audience claps or commentator speaks loudly. The work on lecture video skimming is very few. In lecture video sequences the scene contents change rarely and the only scene changes are due to talking head of the professor or the written texts. Hence creation of the visual abstract/video skimming for lecture video sequence is a challenging task and its automation is also difficult. This motivate us to explore the possibility of applying expert intelligent system in low illuminated, noisy lecture video skimming. It is to be noted that as per author's knowledge no work is reported even in the recent state-of-the-art literature where the contextual contents of the scene are utilized for lecture video skimming. A preliminary version of this work is also reported in Subudhi et al. (2017).

In this article, an automatic intelligent technique for video skimming has been proposed for lecture video sequences. The proposed scheme follows four steps: shot segmentation, shot recognition, video content analysis for frames, and media recreation. In the initial stage of the proposed scheme, the radiometric similar-

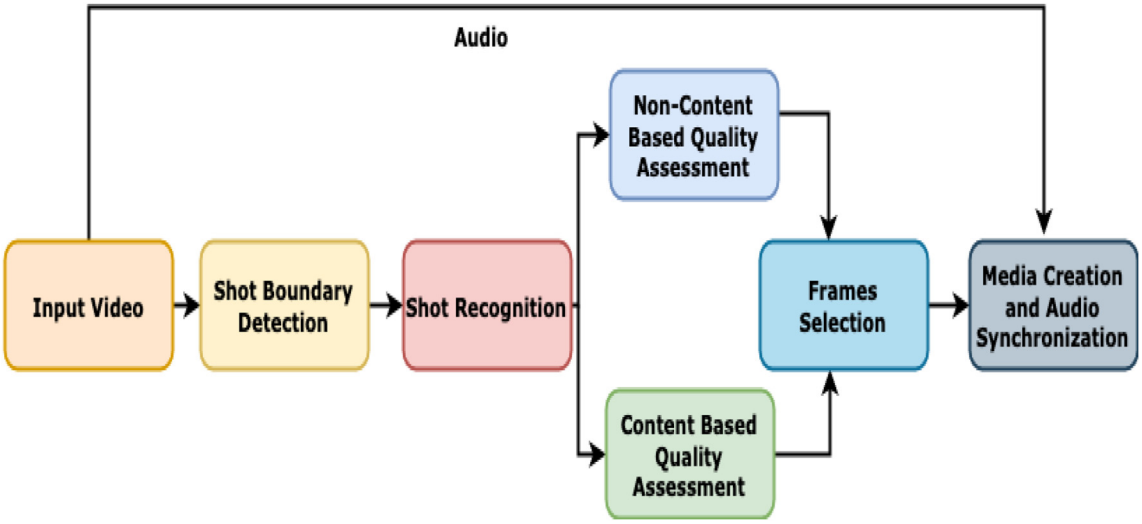

Fig. 1. Block diagram of the proposed scheme.

ity between the frames are explored for video segmentation or to identify the shot transition in the given lecture sequences. In the next stage of processing we have proposed the use of Fuzzy K-Nearest neighborhood algorithm to recognize the shots. Then the content of the shots are analyzed in details with some contrast based features to pick the important frame and prepare the video skimming. The contrast based features: average sharpness (AS), relative height (RH) and edge potential (EP) are introduced to analyze the contents of the frames. The frames with different content values are picked with neighboring frames and media recreation with audio synchronization is used for preparing the video skimming. The proposed scheme is tested on several lecture sequences; however, for page constraint the results on five sequences are reported in this article. The results confirms that the proposed scheme is found to be giving an overall preview of the lecture sequence with proper contents. The performance of the proposed scheme is tested by comparing it against automatic capsule preparation scheme (Ram & Chaudhuri, 2009), motion attention model for video skimming (Ma & Zhang, 2002) and personalized video abstract creation scheme (Babaguchi et al., 2004). The performance of the proposed shot categorization using Fuzzy K-NN is compared with those of the three other classification techniques: minimum distance, Bayes and K-NN. Evaluation of the proposed scheme is carried out by using three evaluation measures: number of frames selected, execution time and accuracy.

Remaining parts of this paper is organized as follows. The step by step description of the proposed scheme with brief description of each stage along with the block diagram is provided in Section 2. Results and discussions are given in Section 3. Finally the conclusion is drawn in Section 4.

2. Proposed automatic lecture video skimming technique

The block diagram of the proposed scheme is shown in Fig. 1. The proposed scheme follows four important blocks: shot boundary detection, shot recognition, content and non-content frames assessment and media recreation. In the initial stage of the proposed scheme, the input video sequence is segmented into number of shots. In this regard, radiometric similarity is explored. After the shot transition identification, the shots are recognized using Fuzzy K-Nearest Neighborhood (K-NN) algorithm to identify as categories: title slides, written texts and talking head. It may be noted that for preparation of video skimming, talking head are non-content frames and are needed to be included in the video capsule preparation. Here it is assumed that the time instant is same as that of the frame instant. For assessing the content and non-content frames an use of three contrast based features are proposed. After selection of the content based frames, some frames from before and after the content frames are chosen with audio synchronization for media recreation. The details on each stage of the proposed scheme is described in the subsequent sections.

2.1. Video segmentation

The proposed scheme follows three steps for detecting the shot boundaries from the given input lecture sequence. In the proposed scheme, we have followed feature extraction, radiometric correlation followed by thresholding operation for finding the shot boundaries. The basic idea is for finding the similarity/correlation between the consecutive frames in the video and point out the discontinuity from it. In this regard, we have considered the radiometric correlation (Spagnolo, Orazio, Leo, & Distante, 2006) based similarity measure to find the correlation between the frames. The radiometric correlation between the frames, can be estimated as,

$$R(I_t I_{t-1}) = \frac{m[\bar{X}_t] \times m[X_t] m[X_{t-1}]}{\sqrt{var[X_t] \times var[X_{t-1}]}}, \tag{1}$$

where X_t represents the input feature vector extracted for tth frame and corresponding mean is represented by $m[X_t]$. Corresponding variance is represented by $var[X_t]$. Corresponding feature value, mean and variance value for $t-1$th frame is represented as, X_{t-1}, $m[X_{t-1}]$ and $Var[X_{t-1}]$. The product vector $m[\bar{X}_t]$, can be defined as,

$$m[\bar{X}_t] = \frac{1}{n} \times [X_t][X_{t-1}]^T, \tag{2}$$

where the feature vector is assumed to be of $n \times 1$ dimension. The radiometric correlation varies in the range [0, 1]. To differentiate the actual shot transition we have used a logarithmic operation on it. From the logarithmic radiometric correlation values, a threshold is required to detect the shot boundary. Hence we have used the Liu's thresholding scheme (Liu, Jiang, & Feng, 2006) on it to differentiate the shot transitions.

2.2. Shots recognition

After detecting the shot boundaries the second step of the proposed scheme is detecting the shots categories. Here, shots are

categorized into three categories: talking head/person, title slides and written texts. In the proposed scheme, we adhered to the use of the Fuzzy K-Nearest Neighborhood (K-NN) classifier (Keller, Gray, & Givens, 1985) for this. We have used the Histograms of Oriented Gradients (HOG) features (Dalal & Triggs, 2005) for training and testing.

2.2.1. Histograms of oriented gradients (HOG)

Histogram oriented gradient (Dalal & Triggs, 2005) is one of the powerful and popularly used feature extraction technique in computer vision. Stating from the object detection (Suard, Rakotomamonjy, & Bensrhair, 2006) and object tracking (Avidan, 2007) to gesture recognition (Freeman & Roth, 1995); HOG is popularly used in many computer vision applications. One of the most important advantages of the HOG feature is that; it can well capture the shape of the object and invariant to both geometric and photometric transformation of the object in the scene. It is also invariant to illumination changes in the scene. To extract the histogram of the gradient direction at each frame of the video is plotted to accumulate the edge orientation of the frame. To reduce the effects of the shadow and unwanted noise in the scene; the histograms are normalized and are referred to as Histograms of Oriented Gradients (HOG) descriptors. To extract these features, at each frame location of the video a 1-D histogram of gradient directions or edge orientations is accumulated.

2.2.2. Fuzzy K-nearest neighborhood

The advantage of use in Fuzzy K-NN classification scheme is, it maps a data point (i.e., the HOG feature extracted from each frame) to different shot types by assigning a suitable membership value rather than assigning directly to a particular shot type. It is assumed that, s_1, s_2 and s_3 are the three shot types: title slides, talking head/person, and display slide or written texts, respectively. Let μ_{ij} be the membership value of i^{th} frame to the j^{th} shot type and $U_j(I_a)$ be the membership value assigned to an unknown data point a.

The assigned memberships of an unknown frame from a shot $'a'$ is given as follows (Mondal, Subudhi, Roy, Ghosh, & Ghosh, 2015):

$$U_j(I_a) = \frac{\sum_{i=1}^{K} \mu_{ij}(1/\|a - s_i\|^{2/(m-1)})}{\sum_{i=1}^{K}(1/\|a - s_i\|^{2/(m-1)})}. \tag{3}$$

where the assigned membership value $U_j(I_a)$ is varied as inverse of the distance from the nearest neighbors. m represents the fuzziness parameter, as m increases, the neighbors are more evenly weighted, and their relative distance from the frames being classified will have less effect. K is a constant and represents the K neighbors.

For shot recognition; after shot boundary detection a randomly chosen frame from each shot is chosen for Fuzzy K-NN classification. For a particular video, initially the different shots are identified and after finding the shot; 0.1% frames from each shot are checked for shot categorization. It may be noted that for preparing a capsule, the frames with talking head/person are not be useful whereas the frames with title slides and display slide or written texts are important and after shot categorization, these shots are properly chosen to prepare the video capsule.

2.3. Contents and non-contents based frame analysis

In the next stage of processing the task is to analyze the content of the video. In this regard, we have used some contrast based features to categorize each image frame. In this work, we have proposed the use of three contrast based features extracted from the histogram of a frame for calculating the contents of an image frame. The extracted features are average sharpness, relative height and edge potential.

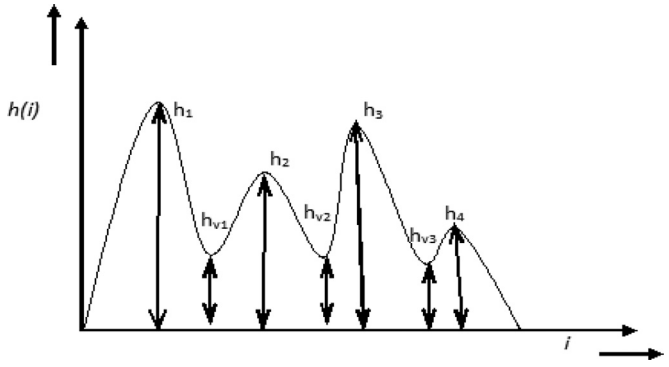

Fig. 2. Peaks and valleys of the histogram.

Average Sharpness (AS): Sharpness is defined as the variation in color or grey values of the image in different region type. One of the best way to measure average sharpness is by using standard deviation of the histogram modes. For calculating this feature, the histogram corresponding to each frame of the video is plotted and the parameters from the multiple modes of the histograms are computed. AS is computed as Ram and Chaudhuri (2009) (Fig. 2):

$$AS = \frac{1}{\sum_{j\epsilon M_j} \sigma_j}, \tag{4}$$

where σ_j represents the standard deviation of grey values corresponding to the jth mode of the histogram. M_j represents the jth mode. The grey value deviation from each mode of the histogram is computed as;

$$\sigma_j = \frac{\sum_{j\epsilon M_j}(j - \mu_j)^2 \times hs(j)}{\sum_{j\epsilon M_j} hs(j)}, \tag{5}$$

where j represents the grey value of the image frame and are in the range $[0, L]$. μ_j represents the mean value of the jth mode. Then,

$$\mu_j = \frac{\sum_{j\epsilon M_j} j \times hs(j)}{\sum_{j\epsilon M_j} hs(j)}, \tag{6}$$

where $hs(j)$ is the jth intensity value frequency of occurrence.

Relative Height (RH): Here we have proposed a new feature is called relative height. The relative height feature is obtained by counting the change in the average height of the distribution from the minimum height of the grey value distribution of a frame. For calculation of this feature, Kurtosis measure is computed on the histogram of the image frame. In statistics, Kurtosis is used to compute the average height of the distribution (Subudhi, Ghosh, Shiu, & Ghosh, 2016). The feature relative height can be computed as:

$$RH = |\, kur(hs(j)) - min(hs(j))\,|, \tag{7}$$

where $kur(hs(j))$ denotes the kurtosis of the grey value distribution $hs(j)$ of a frame and $min(hs(j))$ is the lowest grey value frequency occurrence in that frame. The RH feature describes the difference in height (frequency of occurrence) between the dominant grey value and sharpness of the local grey level distribution.

The Kurtosis measure is obtained as

$$kur(hs(j)) = \frac{1}{M}\sum_{j\epsilon M_j} kur(hs_j(j)); \tag{8}$$

where $kur(hs_j(i))$ denotes the Kurtosis of the jth mode. For any mode it can be obtained as:

$$kur(hs_j(j)) = \frac{E[hs_j(j)(j - \mu_j(hs_j(j))^4]}{\{E[hs_j(j)(j - \mu_j(hs_j(j)))^2]\}^2}, \tag{9}$$

Fig. 3. Detected shots from Lecture video-1 includes: (a) title slide, (b) talking head and (c) written texts.

where $\mu_j(hs_j(j))$ is the mean of the jth mode (M_j) of the histogram and E represents the expectation operation.

Edge Potential (EP): Basically, edge potential for an image frame represents the deviation of the major peaks of the image frame from the minor valley of the histogram. We propose the edge potential feature (EP), which gives a measure for the contrast of the image frame (Fig. 2). The edge potential feature can be computed as;

$$EP = max\{h_1, h_2 ... h_M\} - min\{h_{v_1}, h_{v_2} ... h_{v_{n-1}}\}, \quad (10)$$

where h_1, $h_2 ... h_M$ are the peaks corresponding to M modes of the histogram. $h_{v_1}, h_{v_2} ... h_{v_{M-1}}$, represents the valley corresponding to the $M-1$ valleys in between the M peaks.

In the proposed work, we have considered a combination of all these features and named it as mixed contrast strength (Subudhi et al., 2016). It can be calculated as:

$$f = w_1 \times (AS/AS_{max}) + w_2 \times (RH/RH_{max}) + w_3 \times (EP/EP_{max}), \quad (11)$$

where AS is the average sharpness, RH is the relative height, and EP is the edge potential. w_1, w_2, and w_3 are the weights corresponding to the three considered features and $w_1 + w_2 + w_3 = 1$. We consider $w_1 = w_2 = w_3 = 1/3$. AS_{max}, RH_{max} and EP_{max} are considered to be the AS, RH, and EP features maximum values, respectively.

2.4. Capsule preparation and audio synchronization

The frames with higher potential functions are chosen for creating the the capsule of the video. To prepare the capsule; $\pm$ 200 frames are selected in the vicinity of the frame which posses the highest mixed contrast strength. While preparing the capsule; the original audio of the video is synchronized with that of the selected frames. Finally, we selected such *20* seconds long videos around all of the outstanding high quality frames in a full lecture video to prepare the capsule for the complete sequence.

3. Results and discussions

To validate the proposed skimming technique for lecture video, we have successfully tested it on several lecture sequences. The proposed skimming technique is implemented using Pentium $i7 - X64$ based processor, 3.20 GHz PC with 16G RAM and *Python* programming language in Windows operating system. 'Results and Discussions' section is divided into two subsections: qualitative & quantitative evaluation and discussions and future works. The first part depicts the description on qualitative and quantitative analysis of the proposed scheme. The discussions and future works of the proposed scheme are described in the second part. The proposed scheme is tested on several video sequences, however for page constraints, we have provided the results for five sequences in this article, including three NPTEL (NPTEL, 2018) and two non NPTEL sequences (MLSF, 2018; SAP, 2014). The proposed skimming technique is validated by comparing the results obtained by it with those of the state-of-the-arts-techniques: automatic capsule preparation (Ram & Chaudhuri, 2009), motion attention model for video skimming (Ma & Zhang, 2002) and personalized video abstract creation scheme (Babaguchi et al., 2004). The accuracy of the proposed shot categorization that of Fuzzy K-NN is compared against three other classification techniques i.e., minimum distance, Bayes and K-NN. The performance of the proposed skimming technique is carried out using three evaluation measures: number of frames selected, execution time and accuracy.

3.1. Qualitative and quantitative evaluation

The first sequence considered for the proposed approach is a lecture on Information Theory & Coding and is named as lecture video-1 (NPTEL, 2018). All the shot transitions from this video are extracted in the first part of the experiment. It is observed that in this sequence 120 shot transitions are there. Some important shots of this sequence are shown in Fig. 3. It may be noted that the obtained shots include course title, talking head, and written texts. Use of Fuzzy K-NN scheme is able to distinguish these shots and short listed these shots for further processing. An use of contrast based feature produces results which is able to discriminate the frames content based on the contrast values. It is observed that the contrast based feature values obtained for talking head is found to be highest, written text and written text with hand have medium and title slide has a lowest contrast value. A few frames with high contrast values with $\pm$ 20 seconds are selected for capsule preparation. Some of the selected frames of final capsule are shown in Fig. 4.

The second sequence considered for our experiment is the first lecture of Prof. Ambikairajah's Speech and Audio processing class (SAP, 2014). The experimental analysis finds that the complete sequence has a total of 8 shots. Most of the shot transition happens due to changes in content of the slides. It is most important here to mention that; the informative shots were distinguished easily by Fuzzy K-NN technique. After shots recognition, we applied the contrast based feature extraction and found the eight specific frames as high contrast frames for representing the capsules. Some of the chosen frames for capsule preparation are shown in Fig. 5.

The third example considered in our experiment is carried out on Prof. S. Sengupta's lecture on Video Coding Standard H.264 (NPTEL, 2018). It is found that this sequence consists of 113 shot transitions. Considering the above shots we have used our contrast based features and obtained high contrast frames (some of them are shown in Fig. 6) are used for capsule preparation.

The next example we have considered for our experiment is Prof. Jagannathan's wireless sensor network lecture NPTEL (2018), where the professor have used MSpaint for presenting the figures and the written texts. The selected frames for the capsule preparation are shown in Fig. 7.

The last example considered for our experiment is Prof. Andrew Ng's (MLSF, 2018) first lecture on Machine Learning course, where the professor have used both black board and slides in the lec-

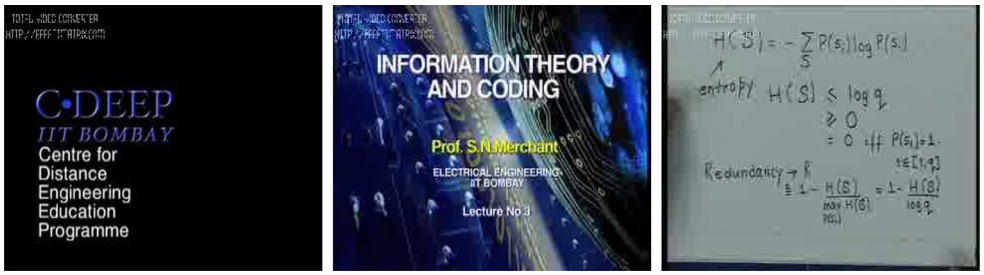

Fig. 4. Selected frames for capsule preparation from Lecture video-1.

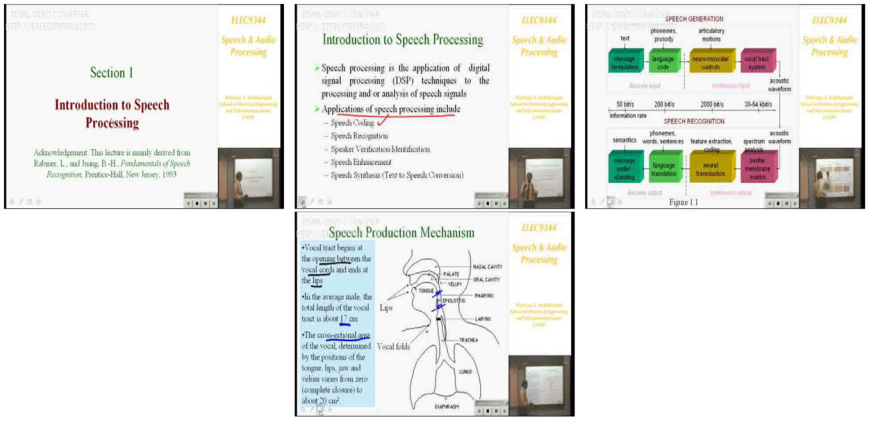

Fig. 5. Selected frames for capsule preparation from Lecture video-2.

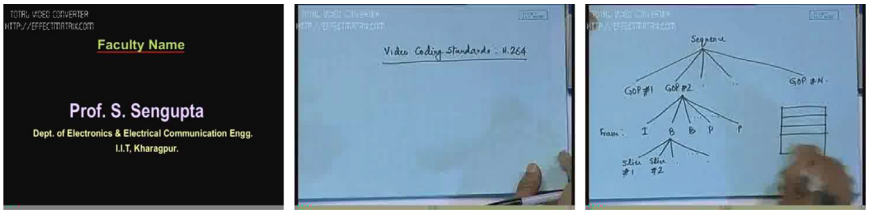

Fig. 6. Selected frames for capsule preparation from Lecture video-3.

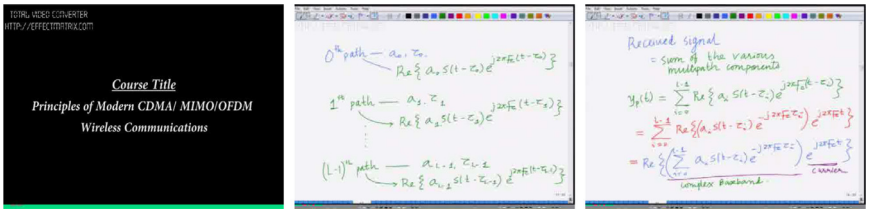

Fig. 7. Selected frames for capsule preparation from Lecture video-4.

Table 1
Contrast based feature values for Lecture video-1 sequence.

Frame instant	*AS*	*RH*	*Ep*	*f*	Category
7	0.211	0.183	0.279	0.224	Low Content
201	0.288	0.268	0.262	0.273	Low Content
6777	0.321	0.357	0.408	0.362	Average Content
11221	0.407	0.404	0.507	0.439	Average Content
14003	0.601	0.61	0.648	0.62	High Content
45001	0.685	0.627	0.768	0.693	High Content
1351	0.86	0.79	0.86	0.837	Non Content

ture. The selected frames for the capsule preparation are shown in Fig. 8.

For all of these above considered video sequences, we have provided the value of the considered contrast based features. Table 1 shows the contrast values of the Lecture video-1 sequence. In this sequence it may be observed that the contrast based feature values for the frames with talking head is obtained with a very high value and found to be a non-content frame. The frames with title slides are having lesser contrast values and are considered as the low content frames. The frames with less written texts will have medium range of contrast values and are corresponding to the average content frames. The frames with full of written texts found to have high contrast values and are considered as the frames with high contents. Similarly the frames contents for other considered sequences are given in Tables 2–5.

Three performance evaluation measures: number of frames selected, execution time and accuracy are considered as the quantitative measures to decide the performance of the proposed scheme. The measure 'number of frames selected' gives; how many frames

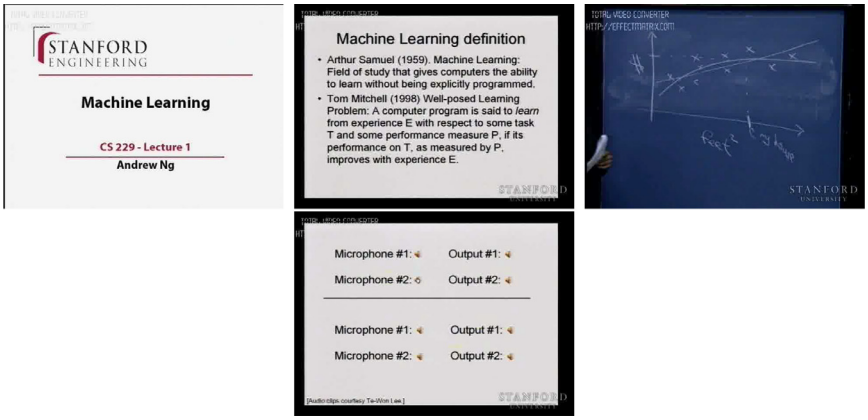

Fig. 8. Selected frames for capsule preparation from Lecture video-5.

Table 2
Contrast based feature values for Lecture video-2 sequence.

Frame instant	*AS*	*RH*	*EP*	*f*	Category
23	0.199	0.174	0.251	0.208	Low Content
303	0.274	0.285	0.277	0.279	Low Content
5075	0.319	0.325	0.394	0.346	Average Content
21030	0.398	0.401	0.497	0.432	Average Content
31003	0.589	0.605	0.651	0.615	High Content
16085	0.662	0.614	0.726	0.667	High Content
901	0.852	0.789	0.851	0.831	Non Content

Table 3
Contrast based feature values for Lecture video-3 sequence.

Frame instant	*AS*	*RH*	*EP*	*f*	Category
100	0.226	0.207	0.296	0.243	Low Content
807	0.305	0.288	0.279	0.291	Low Content
13001	0.334	0.364	0.415	0.371	Average Content
45078	0.428	0.419	0.496	0.448	Average Content
23035	0.594	0.608	0.651	0.618	High Content
31033	0.703	0.659	0.782	0.715	High Content
7005	0.882	0.811	0.875	0.856	Non Content

Table 4
Contrast based feature values for Lecture video-4 sequence.

Frame instant	*AS*	*RH*	*EP*	*f*	Category
305	0.179	0.189	0.305	0.224	Low Content
1222	0.257	0.241	0.245	0.248	Low Content
1768	0.352	0.374	0.428	0.385	Average Content
10909	0.399	0.411	0.524	0.445	Average Content
47098	0.592	0.625	0.661	0.626	High Content
13033	0.708	0.658	0.802	0.723	High Content
23007	0.884	0.804	0.876	0.855	Non Content

Table 5
Contrast based feature values for Lecture video-5 sequence.

Frame instant	*AS*	*RH*	*EP*	*f*	Category
1020	0.210	0.188	0.222	0.207	Low Content
7680	0.388	0.223	0.313	0.308	Low Content
7889	0.440	0.279	0.437	0.385	Average Content
112201	0.456	0.471	0.587	0.505	Average Content
323451	0.671	0.751	0.667	0.696	High Content
753211	0.688	0.823	0.811	0.774	High Content
9807	0.877	0.931	0.889	0.899	Non Content

are selected for preparation of the final capsule for skimming. The 'execution time' defines the amount of time taken by a particular scheme for a single frame of a sequence. The accuracy measure is defined as

$$Accuracy\ (in\ \%age) = \left\{1 - \frac{|u - \hat{u}|}{\hat{u}}\right\} \times 100, \tag{12}$$

where u represents the number of frames selected by a particular algorithm. $\hat{u}$ represents the actual number of frames required or eligible to be taken for capsule preparation, which is obtained by expert knowledge. In the proposed scheme, we have used an average of five separate viewer (experts from the same field of research) to select the actual frames. The results obtained by all these measures for the considered sequences are given in Table 6. It may be observed from this table that except Lecture video-2 the execution time and accuracy of the proposed scheme are more acceptable than the automatic capsule preparation scheme (Ram & Chaudhuri, 2009). Similarly, for all the considered sequences the accuracy of the proposed scheme is better than that of the motion attention model for video skimming (Ma & Zhang, 2002) and the personalized video abstract creation scheme (Babaguchi et al., 2004) with comparable execution time. For Lecture video-2 the accuracy of the automatic capsule preparation scheme and the proposed scheme are same whereas the execution time of the proposed scheme is less.

3.2. Discussions and future works

It is to be noted that for preparing capsule we have considered only title slide and written texts or displayed slides frames. The talking head and frames with only hands are avoided. The frames with talking head and hand are considered as non-content frames and does not carry any information which can be used for preparing the preview of the lecture video. For a particular sequence if a talking head shot will come in between two written texts shots then for capsule preparation we have chosen the frame with high contrast valued from the written texts or displayed slide from two of such shots.

For shot transition, we observed that for some sequences as the amount of written texts on a particular instant of time go on increasing; sometime the proposed scheme found to be detected it as two separate shots. We have avoided such instances of including two frames from two shots for capsule preparation. On those instances two consecutive shots will be recognized as the written text shots. Hence, we have combined such instances into a single shot and a single frame is chosen from such shot for preparation of the capsule.

In the proposed scheme we have used frames from different categories of shots for getting trained using the Fuzzy K-NN classifier. Here it is required to mention that, for all the training we have used 1000 frames from each category of the NPTEL database

Table 6
Comparison of the proposed scheme with existing techniques.

Video	Automatic capsule Ram and Chaudhuri (2009)			Motion attention model Ma and Zhang (2002)			Personalized abstraction Babaguchi et al. (2004)			Proposed		
	frames selected	Avg. exe. time (sec.)	Accuracy in (%)	frames selected	Avg. exe. time (sec.)	Accuracy in (%)	frames selected	Avg. exe. time (sec.)	Accuracy in (%)	frames selected	Avg. exe. time (sec.)	Accuracy in (%)
Lecture video-1	12	1.01	67	09	1.31	59	10	0.62	70	14	0.61	74
Lecture video-2	04	1.17	76	06	1.73	69	05	0.78	72	04	0.69	76
Lecture video-3	13	1.11	56	11	1.44	51	14	0.97	60	16	0.89	66
Lecture video-4	12	1.67	71	08	2.13	67	10	1.37	75	14	0.99	81
Lecture video-5	11	1.22	79	06	1.33	71	08	0.98	66	12	0.91	82

Table 7
Evaluation of classification accuracy in terms of %.

Video	Min. dist.	Bayes	K-NN	Fuzzy K-NN
Lecture video-1	58.2	72.4	74.6	81.3
Lecture video-2	48.1	63.3	66.8	76.9
Lecture video-3	74.8	73.3	84.6	90.1
Lecture video-4	51.7	78.8	71.7	80.4
Lecture video-5	46.7	81.1	70.9	80.7

(NPTEL, 2018). These data are taken based on the manual choice of the frames. The variable m in Eq. (3) is considered as 2. For shot categorization, we have chosen 2% frames from a shot, and if majority of the frames from that shot are found to be classified into a particular category then that shot will be classified into the specific category. In the proposed scheme 5 frames are considered from each shot for categorizing. Majority voting technique is considered for fusing the outputs of five frames. Since 5 frames are considered for categorizing each shot; hence 3 is considered as the threshold to support the voting in favor of a particular category.

The performance of the considered Fuzzy K-NN classifier is further evaluated by comparing the performance of the proposed scheme with that of replacing the Fuzzy K-NN classifier with that of the other three popular classification techniques: minimum distance (min. dist.), Bayes, and K-NN classifiers. Evaluation of all the considered classification techniques are obtained by considering an equal number of training and testing sets. The K values considered for K-NN and Fuzzy K-NN classifiers are considered to be the same. The performance of all these classification techniques are evaluated by considering classification accuracy; i.e., the % of frames correctly classified into desired classes. Table 7 depicts the comparisons of different classification techniques for shot category identification. From this table it may be observed that Fuzzy K-NN is found to be providing higher accuracy.

In most of the instances the literature on Video processing uses contrast based features to deal with the scene with low illuminated and blurred sequences. It is to be noted that for a lecture video sequence the illumination condition of the scene is mostly low. Also, most of the video lectures are recorded mostly with low illuminated environment condition as they are indoor recorded sequences. Hence the selection of proper contrast based features for the lecture video sequences is important task. In this context, it is required to mention that we have tested our algorithm with several contrast based features and obtained that the considered features are found to be providing better accuracy in noisy, low illuminated and indoor environment conditions. In Ram and Chaudhuri (2009); the authors have used bimodal grey-level distribution; where average sharpness is only feature used to define the high content frames. Here it is required to mention that; most of the image frames in a video may not follow bimodal distribution as a scene with class room black board must contains at least the professor, black board, written texts and other equipments. Hence multi-modal distribution may fit such cases. Hence, in the proposed scheme we have extended the the grey-level distribution to multi-modal and extracted three features: average sharpness, relative height and edge potential for our analysis. In Babaguchi et al. (2004); the authors have proposed a video skimming technique where the textual contents of the scene are used to describe the important contents of the game. Then some manual selection of the highlights to select which are required to be included in the skimming. The major disadvantages of video skimming using Babaguchi et al. (2004) is it mostly depends on the textual parts to detect the important contents in the video. It may be noted that in a lecture video most of the scene contains textual parts. Hence detecting important contents of the lecture video using (Babaguchi et al., 2004) not able to produce good results. One of the approach for video skimming (Ma & Zhang, 2002), where the authors have used motion attention model to create the moving abstract. The motion information is obtained by considering: intensity coherence, spatial coherence and temporal coherence. However in a lecture video sequence there will be rarely any motion due to the professor's movement. Hence may not provides good solution. The proposed scheme finds the contrast based information from the video to prepare the moving abstract. Since it uses the modes of the distribution to extract the contrast based information; it can detects the underlying texts in the lecture video frames for further analysis.

It may be noted that the proposed scheme is able to reduce the duration of the video by 17 times of the original sequence. Hence, the preview is reduced by a great factor with providing all the essential details from the video. The efficiency of the algorithm is tested with computational complexity. The time required for execution of the proposed scheme for a single frame of size 480×368 is 0.4 second. Hence, execution time for the proposed scheme is found to be faster.

For more critical analysis of the proposed scheme, we also applied the proposed technique on TRECVID 2007 dataset. The obtained transition over different datasets of the TRECVID 2007 is evaluated using three performance evaluation measures: precession, recall and F-measure (Subudhi et al., 2016). For a better performance all these measures should be high. The results obtained by the proposed scheme on TRECVID 2007 are provided in Table 8. It may be observed from this table that the proposed scheme gives a better overall transition over TRECVID 2007 dataset.

To check further in critical applications, where a professor teaching a course, highlights few contents in the shown teaching materials, the proposed scheme is applied to check its performance. Few results where the key frames selected and contains highlighted text/figures are provided in Fig. 9. This demonstrate the capability of the proposed scheme in finding/extracting meaningful frames from the video for capsule preparation.

In the proposed scheme, we have not checked the quality of the audio for the capsule preparation. However a suitable selection of the audio integration may yield a better and accurate preview of the lecture sequence. The future work involves selection of high content frames with the best audio quality and its integration to

Table 8
Performance evaluation of the proposed method on TRECVID 2007 dataset.

Dataset		BG2408	BG9401	BG11362	BG14213	BG35187	BG36182	BG36506	BG36537	BG36628	BG37359	BG37822	BG38150	BG35050
Overall	Pr	1.00	0.98	0.94	0.96	0.95	0.93	0.94	0.99	0.94	0.91	0.98	0.99	0.96
transition	Re	0.97	0.96	0.96	0.96	0.96	0.97	0.91	0.93	0.98	0.96	0.96	0.92	0.98
	F	0.98	0.97	0.95	0.96	0.95	0.95	0.93	0.96	0.96	0.93	0.97	0.95	0.97

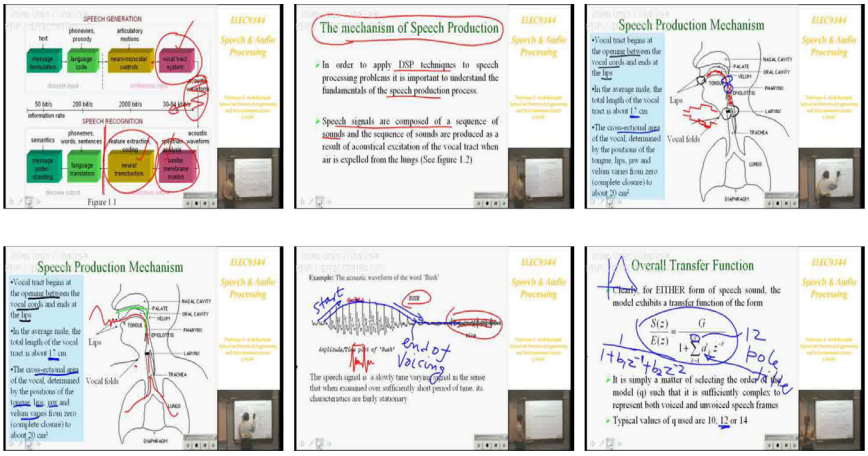

Fig. 9. Example for highlighted details from lecture video-2 sequence.

prepare the video capsule. In future we would also like to go for subjective testing about content of the capsule.

It may also be observed from the results of the proposed scheme, the selection of key frames to represent the important contents of a video; is mainly based on the visual contents that conveys important concepts conveyed by the professor. It is also true that many professors used to express the important concepts of the lecture through their high pitch or voice. The proposed scheme not able to identify those part of the lecture sequence when professor deliver some important concepts through the high pitch voice. Usually it is observed that in a particular lecture, when a professor used to cover any important part of the lecture, he use to modulates the voice pitch so as to make the observer as attentive to his lecture. Hence in future, some algorithms will be developed where the modulated pitch in the lecture will be fused with the corresponding contextual contents of different shots. To find the qualitative pitch, where professor gives an important part of lecture; some speech and audio processing techniques will be developed where formants of the audio contents of the video will be shifted to obtain the contextual contents of the voice part and hence fused with the corresponding visual contents to make the more informative capsule.

It is also important to synchronize the voice contents of the video capsule with the original video. In real life it may fails to do so. In future some re-sampling mechanism for voice signal will be used so that it can exactly synchronize the voice and visual contents in the capsule. The utmost care will be taken to reconstruct the motion blur and low contrast image frames to be included in the capsule preparation. Also the proposed technique, sometime perform low if some parts of the texts are highlighted or marked with less contents in the writing board. It may also produce less efficiency in a case where the professor tries to demonstrate by showing some specific objects/materials. In this context, we expect and salient object detection scheme as proposed in Bastan et al. (2008) may be utilized to detect the highlighted texts and further use of it for video skimming.

4. Conclusions

In this article, we have proposed a skimming technique for lecture video sequence. In the proposed scheme the input lecture sequence is initially segmented into number of shots by using radiometric similarity analysis. After that each shot is recognized using a pre-trained Fuzzy K-NN classifier. The contents of the frame is analyzed using three contrast based features namely, average sharpness, relative height and edge potential function. The high content frames extracted by using analysis of contrast based features are combined with previous and next frame instants with synchronized audio for preparation of the video capsule. The proposed scheme is tested on several lecture sequences, however for page constraint the results of the proposed scheme is provided for five sequences here, including three NPTEL and two non-NPTEL. The obtained results found to be representing a shorter preview of the actual lecture sequence. It is also observed that the prepared capsule by the proposed skimming technique is found to be providing actual details of the original sequence. The performance of the proposed scheme is tested by comparing it against three state-of-the-art techniques. The evaluation of the proposed scheme is carried out by using three evaluation measures. It is observed that the proposed scheme is found to be better than that of the existing scheme.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2020.113341.

Credit authorship contribution statement

Badri Narayan Subudhi: Conceptualization, Formal analysis, Writing - original draft. **Thangaraj Veerakumar:** Conceptualization, Formal analysis, Writing - original draft. **Sankaralingam Esakkirajan:** Conceptualization, Formal analysis. **Santanu Chaudhury:** Conceptualization, Formal analysis.

References

Avidan, S. (2007). Ensemble Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(2), 261–271.

Babaguchi, N. (2000). Towards abstracting sports video by highlights. In *Ieee international conference on multimedia and expo: 3* (pp. 1519–1522).

Babaguchi, N., Kawai, Y., Ogura, T., & Kitahashi, T. (2004). Personalized abstraction of broadcasted american football video by highlight selection. *IEEE Transactions on Multimedia, 6*(4), 575–586.

Bastan, M., Gudukbay, U., & Ulusoy, O. (2008). Automatic extraction of important objects for an mpeg-7 compliant video database system. In *2008 IEEE 16th signal processing, communication and applications conference* (pp. 1–4).

Dagtas, S., & Abdel-Mottaleb, M. (2001). Extraction of TV highlights using multimedia features. In *IEEE fourth workshop on multimedia signal processing* (pp. 91–96).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of ieee computer society conference on computer vision and pattern recognition: 1* (pp. 886–893).

Deng, Y., & Manjunath, B. S. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(8), 800–810.

Egges, A., Molet, T., & Magnenat-Thalmann, N. (2004). Personalised real-time idle motion synthesis. In *12th Pacific conference on computer graphics and applications* (pp. 121–130).

Freeman, W. T., & Roth, M. (1995). Orientation histograms for hand gesture recognition. In *IEEE international workshop on automatic face and gesture recognition* (pp. 296–301). Zurich, Switzerland.

Gao, X., Xin, H., & Ji, H. (2002). A study of intelligent video indexing system. In *Proceedings of the 4th world congress on intelligent control and automation: 3* (pp. 2122–2126vol.3).

Gerek, O. N., & Altunbasak, Y. (1997). Key frame selection from MPEG video data. In J. Biemond, & E. J. D. III (Eds.), *Visual communications and image processing '97: 3024* (pp. 920–925). International Society for Optics and Photonics SPIE.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, 15*(4), 580–585.

Lienhart, R., Pfeiffer, S., & Effelsberg, W. (1997). Video abstracting. *Communications of ACM, 40*(12), 54–62.

Liu, D., Jiang, Z., & Feng, H. (2006). A novel fuzzy classification entropy approach to image thresholding. *Pattern Recognition Letters, 27*, 1968–1975.

Liu, T., & Kender, J. R. (2004). Lecture videos for e-learning: current research and challenges. In *IEEE sixth international symposium on multimedia software engineering* (pp. 574–578).

Ma, Y.-F., & Zhang, H.-J. (2002). A model of motion attention for video skimming. *International conference on image processing*: 1. pp. I–129–I–132.

Martin, H., & Lozano, R. (2000). Dynamic video abstract generation using an object DBMS. In *2000 IEEE international conference on multimedia and expo. icme2000. proceedings. latest advances in the fast changing world of multimedia (cat. no.00th8532): 3* (pp. 1523–1526vol.3).

MLSF (2018). Machine learning course: Center for professional developement video database. https://see.stanford.edu/Course/CS229.

Mondal, A., Subudhi, B. N., Roy, M., Ghosh, S., & Ghosh, A. (2015). A study on nonlinear classifier-based moving object tracking. In *Intelligent computing, communication and devices, volume 1* (pp. 571–578). Springer India.

NPTEL (2018). NPTEL video database. http://nptel.ac.in/.

Peker, K. A., & Divakaran, A. (2004). Adaptive fast playback-based video skimming using a compressed-domain visual complexity measure. In *Ieee international conference on multimedia and expo: 3* (pp. 2055–2058).

Ram, R., & Chaudhuri, S. (2009). Automatic capsule preparation for lecture video. In *2009 international workshop on technology for education* (pp. 10–16).

Rav-Acha, A., Pritch, Y., & Peleg, S. (2006). Making a long video short: Dynamic video synopsis. In *IEEE computer society conference on computer vision and pattern recognition: 1* (pp. 435–441).

Rui, Y., Gupta, A., & Acero, A. (2000). Automatically extracting highlights for TV baseball programs. In *Eighth ACM international conference on multimedia* (pp. 105–115).

SAP (2014). Speech and audio processing course: The University of New South Wales Course video database. http://freevideolectures.com/Course/2504/ELEC9344-Speech-and-Audio-Processing.

Shipman, F., Girgensohn, A., & Wilcox, L. (2003). Creating navigable multi-level video summaries. In *International conference on multimedia and expo.: 2* (pp. II–753).

Smith, M. A., & Kanade, T. (1997). Video skimming and characterization through the combination of image and language understanding techniques. In *Ieee computer society conference on computer vision and pattern recognition* (pp. 775–781).

Spagnolo, P., Orazio, T. D., Leo, M., & Distante, A. (2006). Moving object segmentation by background subtraction and temporal analysis. *Image and Vision Computing, 24*(5), 411–423.

Suard, F., Rakotomamonjy, A., & Bensrhair, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. In *Proceedings of IEEE conference on intelligent vehicles* (pp. 206–212).

Subudhi, B. N., Ghosh, S., Shiu, S. C., & Ghosh, A. (2016). Statistical feature bag based background subtraction for local change detection. *Information Sciences, 366*, 31–47.

Subudhi, B. N., Veerakumar, T., Yadav, D., Suryavanshi, A. P., & Disha, S. N. (2017). Video skimming for lecture video sequences using histogram based low level features. In *IEEE 7th international advance computing conference* (pp. 684–689).

Yao, T., Mei, T., & Rui, Y. (2016). Highlight detection with pairwise deep ranking for first-person video summarization. In *IEEE conference on computer vision and pattern recognition* (pp. 982–990).

Zelnik-Manor, L., & Irani, M. (2001). Event-based analysis of video. *IEEE computer society conference on computer vision and pattern recognition*: 2.