Index Point Detection for Text Summarization using Cosine Similarity in Educational Videos

Rekha Sharma, Mehul Mahrishi, Sudha Morwal and Girish Sharma
KanoriaPGMahilaMahavidyalaya, Jaipur, Rajasthan, India
SwamiKeshvanandInstituteofTechnology, Management&Gramothan, Jaipur, Rajasthan, India
BanasthaliVidhyapith, Niwai, Rajasthan, India
SwamiKeshvanandInstituteofTechnology, Management&Gramothan, Jaipur, Rajasthan, India
E-mail: rekha04sharma@gmail.com

Abstract. Explosive growth in digital content creates a trend of video-based learning and knowledge sharing. Several educational videos are shared by various trainers/ bloggers/ instructors daily. It is now evident that even during emergencies like the recent ongoing CoVid-19 outbreak, video lectures are the saviors when the whole world comes to rest. This rapid and expansive growth has encouraged the researchers to efficiently and automatically index, browse, and retrieve video data. This research aims to process videos and precisely identify the index points which are discussed within the video. Experimental results show a good performance of the topic detection model in finding keywords and reducing dimensionality.

Keywords: Cosine Similarity, Key-frame Identification, Optical Character Recognition, Text Extraction, Text Recognition, Video Frames

1. Introduction
Segmentation splits a video stream into different camera-takes/scenes or shots. These are nothing but the correlated sets of a contiguous picture frame sequence taken by a camera from the moment it begins to capturing the moment it finishes. [1][2]. Content-wise, shots are homogeneous and possess a degree of visual uniformity. The idea is to extract the video content and analyze it for information and knowledge retrieval. This paper intends to offer a framework that assists in the video summarization technique by extracting keywords as and when they occur in the video. Shot Boundary Detection is done through key-frame identification using Cosine Similarity with a static threshold value.

1.1. Textual content in a Video
Text in a video sequence can exists in one of the two forms viz.
- Scene Text: It is also called graphic text and is displayed in a video sequence during the video making, for example, text written on boards, banners, and hoardings.
- Artificial text: is the text that is introduced explicitly into the video through editing. It is always displayed at a specific position in the video. Examples are names, data about the video, or interpretation of discoursed in video film, captions in a video.
1.2. **Online Video Lectures**

Recording video lectures and presentations for enhancing the teaching-learning mechanism is now longstanding in India.[3] [4]. With the support of the Ministry of Human Resource Development (MHRD), institutes of national importance in India started recording the video lectures, which substantially increased the self-paced learning for students and academicians. This repository is now huge and authentic that many articles have been published, and much research is possible for data scientists by using this data-set [5]. This data-set is very useful in many research domains like natural language processing, speech recognition, image, and video analysis.

2. **Related Work**

Text summarization is the process that automatically gives the precise meaning of the given document and attracting the attention of Natural Language Processing researchers.[1] Proposed an algorithm in which the OpenCV library is used to decode video into frames. [6] proposed that Clustering and dictionary learning together can be employed to determine keyframes, and additional local and global features of these frames can is determined. Color features [7],[8], moments of inertia [9] as well as local binary features[10] are most predominantly used for representing image frames in a video. Table 1 represents the literature regarding text and feature extraction techniques used by different authors in their state-of-the-art research.

Few methods came into existence that uses divide-and-conquer algorithms to segment videos and determines keyframes [11]. [12]Using an intra-modality and inter-modality fusion mechanism to select sub-summaries, the aggregated attention curve was created and all sub-summaries were finally combined into an aggregated summary. [13]proposed a dual-threshold method for video segmentation. Different keyframes are extracted from each segment. Scale-Invariant Feature Transform (SIFT) features are extracted from the keyframes of the segments. It is proposed in SVD-based method (Singular Value Decomposition) ,two video frames with SIFT point set descriptors are suggested to match.

[14] figures out the mutual information method according to information theory and confirms that it improves the extraction efficiency.

3. **Index Point Detection Framework**

This research discovered that the frequency of words, frequency of n-grams, and several first-time words in a video provide valuable information for video segmentation by topic. The research is intended to identify semantic and topic-based index points in educational videos so that the video has discussed topics that can be quickly and efficiently listed. These index points are the topic/subtopic discussed in the video lecture. Data sets of online video lectures, tutorials, open course-ware, and similar resources with variable playback time and specifically the videos that contain lecture slides are used as input. As represented in figure 1, the proposed framework is robust and exhibits good performance for video text detection and recognition.

For similarity measures, cosine similarity is used, and we adopted a threshold value while measuring the similarity among video frames. Videos with most graphs, statistical data, and diagrams are avoided in the data set to maintain the throughput. The research work focuses on selecting Keyframes, and the goal is to make a sequence of keyframes in the original sequence while reducing the redundancy as much as possible.

3.1. **Partitioning of Video into Frames**

Initially, image frames are extracted from a video. Generally, a video contains 24 image frames per second, out of which the majority are redundant frames. Therefore, there is a need to implement some measures to extract critical frames only. Usually, a video of k minutes is partitioned into:
Table 1. Text and Feature Extraction.

| Sr.No. | Title of Research                                                                 | Cite  | Proposed Idea                                                                                                                                                     |
|--------|-----------------------------------------------------------------------------------|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1.     | Movie Tags Prediction and Segmentation Using Deep Learning                         | [15]  | Shot boundaries are computed by intersecting HSV (Hue, Saturation, Value) histograms of the successive frames. Key frames are identified by aggregating spatio temporal features generated by CapsuleNet.G1 continuity error and cubic H Bézier curve fitting for shot boundary detection within local sliding windows. For eight neighbours, the grey difference is calculated between the adjacent pixels. If the threshold is exceeded, 1 will denote the corresponding statistic, and 0 otherwise. The main colours are applied to the adjacent frames using colour feathers to enhance the accuracy of shot detection and to judge the similarity of material between adjacent frames. |
| 2.     | A Novel Key-frames Selection Framework for Comprehensive Video Summarization        | [16]  | SIFT, SURF (Speed up Robust Features), Sobel Edge Detector, MSER (Maximally Stable Extreme Regions) are the key state-of-the-art local feature extraction algorithms used recently, using histograms-based feature extraction techniques. The most popular global attribute of a video are colour histograms. They offer a good balance between precision and time. The colour space of the HSV is more random since it determines a 3-D colour that provides more robust detection results in turn. |
| 3.     | A Novel Shot Detection approach using 8 Neighbors and Key-Colours                  | [17]  | Edge detection for feature extraction. Hybrid features for feature extraction                                                                                     |
| 4.     | An efficient method for video shot boundary detection and key frame extraction using SIFT point distribution histogram: Multi Modal Visual Features Based Video Shot Boundary Detection: Abrupt Shot Detection in Video using Weighted Edge Information: Shot boundary detection using perceptual and semantic Information | [18]  | A Study of Discriminant Visual Descriptors for Sport Video Shot Boundary Detection Improving the Video Shot Boundary Detection Using the HSV Color Space and Image Subsampling Edge Strength Extraction using Orthogonal Vectors for Shot Boundary Detection Shot Boundary Detection in MPEG Videos Using Local and Global Indicators Video shot transition detection using spatio temporal analysis and fuzzy classification |
| 5.     |                                                                                   | [19]  |                                                                                                                                                                 |
| 6.     |                                                                                   | [20]  |                                                                                                                                                                 |
| 7.     |                                                                                   | [21]  |                                                                                                                                                                 |
| 8.     |                                                                                   | [22]  |                                                                                                                                                                 |
| 9.     |                                                                                   | [23]  |                                                                                                                                                                 |
$N = k \times 60 \times 30 \text{frames}$

where $N$ is number of frames generated initially from the video.

It is evident that in educational videos, a topic will be discussed for at least 10 seconds. Therefore, to curtail the number of redundant frames, 10 seconds delay is introduced in frame generation. Figure 2 displays the frame generation of a sample video from the dataset after every 10 seconds.

![Figure 1. Index Point Detection Framework](image)

![Figure 2. Video Name: Data Structure. Total 85 frames are generated after introducing 10 Seconds delay](image)

3.2. Extraction of Text using Optical Character Recognition

[24] Optical Character Recognition (OCR) tools use the binarization algorithm to identify text in image frames. At the same time, frames are containing plain text that often gives a
readable output. Frames with complex images, plots, or screenshots sometimes give incorrect and distorted output. To address this problem, we consider only those videos that have maximum content as plain text. Keyframes are detected after processing the text using the vectorization method. PyTesseract tool is used to extract text from image frames.

### 3.3. Cosine Similarity

Text similarity estimates the similarity index between the text documents using the distance measures or semantic similarity index. In this technique, text in the document acts as the local features used to find similarity between the documents. In our research, we will use the cosine similarity method [25].

Cosine similarity uses the cosine value of two vectors to represent the similarity between two image frames [26]. If frames are identical their cosine value is closer to 1, else varies from [0,1] as per equation 1 where A & B are image frames.

$$\cos(A,B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Values range from -1 and 1 , where -1 is perfectly dissimilar and 1 is perfectly similar. The similarity threshold is typically set between 0.7 and 0.8. It is observed that higher the compression ratio requirement, the smaller the threshold setting.

### 3.4. Index Point Identification

After successful extraction of image text, index points/keywords needs to be identified. The first step in this direction is text pre-processing [27], which involves removal of noisy data like punctuation marks, special character, tags, URL and removing redundant text components. In the second step removal of stop words viz. pronoun, prepositions, conjunctions and connecting verbs is done.

Third and final step are Text Preparation [28] in which the Bag of Words model is used to count word frequencies. The word frequency determines the possibility of a word to be called as ‘Index Point’.

In the process of Keywords extraction, removal of stop words, and conversion of the text into a list of words are also done. Unigram and bigrams are also generated so that accurate index points could be identified. These are converted to a matrix of integers by the process of vectorization. The figure 3 depicts that ”Data structure”, ”Data type”, ”abstract data”, and ”primitive data”, are bi-grams that are more often used by the instructor in the video.

### 4. Result Analysis

Keyframes extraction done using cosine similarity method with experimenting with different threshold values is shown in the table 2. The representation is only dramatic; actual comparisons are made for all videos dataset, as shown in table 3. The table reflects that the actual keyframes and the keyframes generated by our framework are almost equal, which justify that we achieve good precision at the threshold value of 0.8. Table 3 shows the precision and accuracy level for different videos as per the framework that we have developed. A suite of educational videos handpicked (source: NPTEL) is created. The keyframes with the high precision are identified which are indispensable for video summarization based on keywords.

It can be observed in figure 4 that almost all videos have a precision value of more than 70%. In some videos, due to the presence of images and mathematical formulas, the precision value is degraded. Those videos are selected willingly in the dataset to keep a check on the worst-case
Figure 3. Histogram of most frequently occurring Uni-grams & Bi-Grams

Table 2. Threshold Comparison.

| Threshold  | Threshold 1 | Threshold 2 | Threshold 3 | Threshold 4 | Actual Key Frames |
|------------|-------------|-------------|-------------|-------------|-------------------|
| Key frames (Data Structure video) | >.5 | >.6 | >.7 | >.8 | 0.8 |
| Key frames (PERT & CAM video) | 7 | 9 | 9 | 9 | 9 |
| Key frames (---) | – | – | – | – | – |

The average precision value of the framework is 86%.

Figure 4. Actual Keyframes vs Algorithm Keyframes
Table 3. Video wise Precision table.

| Video          | Length     | Total Frames | Videos satisfying Threshold | Original Keyframes (True Positive) | Precision Value |
|----------------|------------|--------------|-----------------------------|------------------------------------|-----------------|
| Cloud Security | 9min38sec  | 58           | 3                           | 3                                  | 1.00            |
| DBMS           | 6min42sec  | 41           | 41                          | 41                                 | 1.00            |
| PERT & CAM     | 7min52sec  | 48           | 24                          | 22                                 | 0.92            |
| IAA            | 5min56sec  | 36           | 22                          | 22                                 | 1.00            |
| Sensor         | 9min39sec  | 61           | 6                           | 5                                  | 0.83            |
| Lifi           | 12min31sec | 80           | 24                          | 21                                 | 0.88            |
| Ecommerce      | 6min32sec  | 40           | 18                          | 8                                  | 0.44            |
| Data Structure | 14min4sec  | 85           | 9                           | 9                                  | 1.00            |
| Array          | 17min14sec | 104          | 28                          | 23                                 | 0.82            |
| TOC            | 9min37sec  | 58           | 13                          | 10                                 | 0.77            |
| Testing        | 14min54sec | 93           | 34                          | 25                                 | 0.74            |
| SE             | 12min23sec | 79           | 30                          | 27                                 | 0.90            |
| Analog         | 9min59sec  | 60           | 8                           | 7                                  | 0.88            |
| NE&IM          | 7min47sec  | 47           | 19                          | 18                                 | 0.95            |
| CRT            | 9min58sec  | 60           | 8                           | 8                                  | 1.00            |
| Concrete       | 24min27sec | 154          | 9                           | 9                                  | 1.00            |
| Appraisal      | 15min54sec | 96           | 17                          | 13                                 | 0.76            |
| 8051           | 17min55sec | 109          | 40                          | 19                                 | 0.48            |
| Appraisal2     | 18min29sec | 111          | 20                          | 17                                 | 0.85            |
| CAM COST       | 11min53sec | 72           | 21                          | 19                                 | 0.90            |

5. Conclusion

Through this research paper, we want to propose a new framework that extracts text from keyframes for keyword-based video summarization. Throughout the article, the keywords are referred to as Index Points. Cosine-Similarity approach is used to verify the index of similarity between the frames. OCR tool PyTesseract is used to extract text from the keyframes. The findings are rather persuasive, and the precision is considerably higher. This system would allow better indexing of videos in the future.

References

[1] Lu W, Sun H, Chu J, Huang X and Yu J 2018 *IEEE Access* 6 40198–40211
[2] Mehmood I, Sajjad M, Rho S and Baik S W 2016 *Neurocomputing* 174 393–403
[3] Ahanger G and Little T D 1996 *Journal of visual Communication and image representation* 7 28–43
[4] Abowd G D 1999 *IBM systems journal* 38 508–530
[5] Mahrishi M, Hiran K K, Meena G and Sharma P 2020 Machine learning and deep learning in real-time applications URL https://www.igi-global.com/book/machine-learning-deep-learning-real/240152

[6] Gharbi H, Bahroun S, Massaoudi M and Zagrouba E 2017 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE) pp 1502–1506

[7] Asim M, Almaadeed N, Al-Máadeed S, Bouridane A and Baghdadi A 2018 2018 Colour and Visual Computing Symposium (CVCS) (IEEE) pp 1–6

[8] Tarare P and Jadhav D 2015 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (IEEE) pp 403–408

[9] Ejaz N, Tariq T B and Baik S W 2012 Journal of Visual Communication and Image Representation 23 1031–1040

[10] Duan Y, Lu J, Feng J and Zhou J 2018 IEEE transactions on pattern analysis and machine intelligence 40 1139–1153

[11] Mahrishi M and Morwal S 2020 Advances in Intelligent Systems and Computing AISC Springer

[12] Mademlis I, Tefas A and Pitas I 2018 Information Sciences 432 319–331

[13] Liu H, Lu H and Xue X 2012 IEEE transactions on knowledge and data engineering 25 1706–1718

[14] Lanying H and Cong-huan Y 2012 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet) (IEEE) pp 2179–2182

[15] Khan U A, Martínez-Del-Amor M A, Altwaijri S M, Ahmed A, Rahman A U, Sama N U, Haseeb K and Islam N 2020 IEEE Access 8 6071–6086

[16] Huang C and Wang H 2019 IEEE Transactions on Circuits and Systems for Video Technology

[17] Liu H B, Tan T H, Huang S C, Wang Z J and Zhou J 2018 2018 7th International Symposium on Next Generation Electronics (ISNE) (IEEE) pp 1–3

[18] Tippaya S, Sitjongsataporn S, Tan T, Khan M M and Channongthai K 2017 IEEE Access 5 12563–12575

[19] Tippaya S, Tan T, Khan M and Channongthai K 2015 2015 10th Asian Control Conference (ASCC) (IEEE) pp 1–4

[20] Liu F and Wan Y 2015 2015 Seventh International Conference on Advanced Computational Intelligence (ICACI) (IEEE) pp 351–354

[21] Priya G L and Domnic S 2012 Procedia Technology 6 247–254

[22] Ren J, Jiang J and Chen J 2009 IEEE Transactions on Circuits and Systems for Video Technology 19 1234–1238

[23] Lin G S, Chang M K and Chiu S T 2009 International Journal of Pattern Recognition and Artificial Intelligence 23 1179–1200

[24] Kaundilya C, Chawla D and Chopra Y 2019 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom) (IEEE) pp 145–150

[25] Hoque M N, Islam R and Karim M S 2019 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (IEEE) pp 1–5

[26] Ma M, Met S, Hou J, Wan S and Wang Z 2017 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) (IEEE) pp 164–169

[27] Yao J 2019 Journal of Physics: Conference Series vol 1187 (IOP Publishing) p 052020

[28] Sarwar A, Mehmood Z, Saba T, Qazi K A, Adnan A and Jamal H 2019 Journal of Information Science 45 117–135