An efficient method for lecture video retrieval
Laxmikant S.Kate*1, Prof. M.M. Waghmare*2
*1(STUDENT of Computer Engineering DGOI, BHIGWAN, PUNE)
*2(STUDENT of Computer Engineering DGOI, BHIGWAN, PUNE)
lskate@hotmail.com *1,monawaghmare25@gmail.com*2

Abstract— The expanded accessibility of broadband connection has as of late prompted an increment in the utilization of Internet. Most webcasts are saved and accessed various times. One test to skimming also browsing through such archives is the absence of text transcripts of the webcast’s audio channel. Video is turning into a predominant medium for e-learning. Lecture videos contain text data in both the visual and aural channels: the presentation lecturer's and slides speech. Optical Character Recognition (OCR) innovation on key-frames and Automatic Speech Recognition (ASR) on lecture audio tracks is expert. In this paper, we researched two different frameworks associated just to unmatched question words. In any case method delivers another arrangement of n-gram strings to match the unedited OCR Transcriptions. These n-grams join strings with a changed partition of 1 character and all possible n-gram substrings with no fewer than 3 characters. Second framework for changing OCR incorporated the expression reference of spelling adjustment methodology gave in MS Words. The peculiarities of MS Word 2000, an OCR saw string was stretched out through an application program interface into its correct spellings. An outstandingly dynamic style, simply developing words that MS Word had flag as wrongly spelled which we depict some time recently. This fundamentally diminished the amount of spurious word competitors and kept up a distance from matches which are not true matches.

Keywords—lecture videos; automatic text indexing; OCR; spoken text, textual queries.

I. Introduction
A video retrieval traditionally based on visual feature extraction can't be just connected to lecture recordings on the grounds that of the homogeneous scene organization of lecture videos. Varying variables may bring down the nature of this arrangement. For instance, movement changes of the camera may influence the size, shape what's more the brilliance of the slide; the slide can be halfway discouraged when the speaker moves before the slide; any progressions of camera center might likewise influence the further slide recognition process. These days individuals have a tendency to create lecture videos by utilizing multi-scenes format, by which the speaker what's more his presentation are shown synchronously. Digital video has turned into storage and medium as exchange because of the quick improvement in recording innovation, enhanced video compression strategies what's more fast systems in the most recent few years.

Text is a high state semantic feature which has frequently been utilized for content based data recovery. In lecture videos, texts from lecture slides serve as a background for the lecture and are imperative for understanding. In this manner next, fragmenting a video document into an arrangement of key frames, the text recognition methodology will be executed on every key edge, and the separated text objects will be further utilized as a part of text recognition furthermore slide structure investigation forms. Particularly, the separated structural metadata can empower more adaptable video browsing and video search capacities. Discourse is a standout among the most imperative transporters of data in video lectures. Accordingly, it is of different focal point that this data can be applied programmed lecture video indexing. A lot of textual metadata will be made by utilizing OCR und ASR strategy, which opens up the content of lecture videos. To empower a sensible access for the client, the keywords are further removed from the OCR and ASR results. For content based video search, the search files are made from distinctive data, including manual annotations, OCR and ASR keywords, worldwide metadata, and so on. Here the changing recognition precision of diverse analysis may bring about strength and consistency issues, which have not been considered in most related work [1].

Lecture transcription is a troublesome undertaking, both from an acoustic and a linguistic perspective. Speech, foundation noise, diverse and changing talking rates and numerous spontaneous speech phenomena, are all qualities of lecture discourse that make acoustic displaying troublesome. Language displaying is hampered due to the inadequacy of suitable information and the blended style of lecture language which spoken, joining informal interpretations with formal language.

In the work [2], principle center is on exertion on language demonstrating. A baseline Language Modeling was evaluated utilizing different sorts of information, which were all imperfect, yet utilized as a part of such a route, to the point that their qualities were highlighted and not their insufficiency. Automatically lecture translation is emerging as an essential thing for both applications and research. It is a test for speech recognition as, as opposed to broadcast news, lectures regularly present a higher variability regarding talking style, semantic space, and speech familiarity. From the application purpose of view, spoken record recovery in fact of programmed transcripts
has indicated to the guaranteeing mean for getting to substance in audiovisual digital libraries. Hence forth, imagining digital repositories of recorded discourses and lectures, which would he be able to looked and perused through the net, is truly regular at this point.

There are a few contrasts in the middle of news and lecture discourse such as a literary style and an accessible asset. These days, numerous show organizations give news clips a relating script through online administration. Since we can undoubtedly assemble a preparing corpus utilizing this, show news retrieval has been a real concentrate in talked report recovery range. Then again, they as of now give keyword-search administrations utilizing a text search engine taking into account the news script. Not at all like news, can’t we without much of a stretch get a script of lecture speech. In a business training site, a search is performed utilizing a physically constructed record. In paper [3], a Korean spoken data recovery framework for search lecture. We naturally manufacture a general transformed list table from spoken archives, and we additionally extract data from a textbook or a slide note. We incorporate these two data sources for a search process. The methodology in [3] is like joining of a metadata. The material utilized as a part of framework is a secondary school math lecture videos.

In the previous decade, we have seen a sensational increment in the accessibility of on-line academic lecture material. These resources of educational can possibly change the way individuals learn — understudies with incapacities can upgrade their experience of education, experts can stay aware of late progressions in their field and individuals of all ages can fulfill thirst for information. In complexity to numerous other informative exercises then again, lecture transforming has as of late long ago delighted in little advantage from the improvement of human dialect innovation. Recent endeavors to gather a corpus of spoken lecture material that will empower research coordinated towards quick, precise, what’s more simple access to lecture content. Hence far, we have gathered a corpus of 270 hours of discourse from an assortment of college classes. In [4] it write about a beginning analysis of the spontaneous discourse phenomena present in these information and the vocabulary use designs over some course.

The paper is composed as takes after. Section 2 quickly depicts the related work studied till now. In Section 3, we give a depiction of the Content Based Image Retrieval model with K-means in our proposed work. Test results and summary are introduced in Section 4. At last, Sections 5 talk about our conclusion and future works.

II. RELATED WORK

In [5] they implemented strategies for visualizing, segmenting and indexing presentation videos by independently considering audio and visual information. The audio track is divided by speaker, and enlarged with key expressions which are separated utilizing an Automatic Speech Recognizer (ASR). The video track is divided by visual dissimilarities and enlarged by illustrative key frames. An intelligent client interface consolidates a visual representation of audio, video, text, and key frames, and permits the client to explore a presentation video. It additionally investigates clustering and labeling of speaker information and present preparatory results. It also has exhibited systems for division, text increase, and visualization of presentation videos. Methodology of independently breaking down and picturing audio and video demonstrates that the two media are not one or the other comprehensive nor selective, however corresponding. It improves the division by utilizing record expression sifting to give further signals to visual browsing and looking of presentation video content.

Recording lectures and putting them on the Web for access by understudies has turned into a general pattern at different educational levels. To take full pick up of the learning database that is manufactured by these archives extensive pursuit usefulness must be given that goes past hunt on meta-information level yet performs a point by point investigation of the relating multimedia reports. In paper [6], exhibited a few trials they did towards setting up a Web-based web crawler for audio recordings of presentations. They assess standard, speech recognition programming and additionally achievable recovery execution. Furthermore, it contrasts the discourse recovery results and a customary, text-based methodology for looking to assess the estimation of discourse handling for retrieval of lecture.

The paper [7] depicts a system for prototyping an ASR framework that produces reasonable transcripts of any desired Word Error Rate (WER), hence defeating the disadvantages of both model based and Wizard of Oz simulation. It utilized such a framework as a part of a client study demonstrating that transcripts with WERs fewer than 25% are adequate for utilization in webcast documents. As present ASR frameworks can just convey, in reasonable conditions, Word Error Rates (WERs) of around 45%. It likewise portray an answer for lessening the WER of such transcripts by captivating clients to team up in a "wiki" form on altering the imperfect transcripts acquired through ASR.

The paper [9] presents a brought together approach in breaking down and organizing the content of videotaped lectures for separation learning applications. By organizing lecture videos, it can support indexing and semantic questioning of multimedia records caught in the customary classrooms. The principle objective in this is to naturally develop the cross references of lecture videos and textal archives in order to encourage the synchronized browsing and presentation of multimedia data. The significant issues included
in this methodology are topical occasion recognition, video text investigation and the matching of slide shots and outer records.

In [9] propose a collaborating tagging framework that is joined with a mechanized annotation framework for synchronized multimedia presentations. MPEG-7 metadata are utilized for the annotation of single scenes with client arranged labeling data in mix with metadata given straightforwardly by the creator or by other annotation frameworks. Thus, they proposed a framework having the capacity to question inside multimedia information that can further be reached out to query inside any sort of (incomplete) record to accomplish a harder centered and customized query.

In [10] model predicts signal remarkable quality as a hidden variable in a restrictive system, with noticeable features from both the visual and textual modalities. This methodology essentially beats focused baselines that don't utilize motion data. A model framework was assembled to test the algorithms and the utility of the internet searcher. Clients can skim arrangements of lectures, slides in a specific lecture, or play the lecture video [11]. In [12] presents the CONTENTUS approach towards a robotized media preparing chain for social legacy associations and content holders. Our work process considers unattended handling from media ingest to accessibility careful query and recovery interface. It means to give a set of tools for the handling of digitized print media, audio/visual, speech and musical recordings.

III. Proposed Work

A. System Architecture

In our proposed system following Figure 2 shows proposed system architecture. The detailed description is as follows:

Training Phase:

We can take numbers of videos in for training process as input. At training phase initially we train videos one by one, different frames are extracted from selected video and segmentation of this frames are performed.

OCR:

Optical Character Recognition is applied on the extracted frames. ie we extract the text from each frame, after extracting text we can check this text for spelling. Correction of unmatched word can be done in this phase.

We are also retrieving text from audio by applying Automatic Speech Recognition technique.

Indexing is performed on Text extracted from ASR and OCR and this indexing values are further stored in dictionary. It is used to compare it with user query.

Testing Phase:

In this phase user enters the input query text ie Text for video retrieval. If input query is incorrect we can correct it by applying n gram technique.

Indexing of this text is formed and these indexing values are compared with information stored in directory and at the end related videos are retrieved.

Algorithm/ Technique

All non-title text line objects are further classified into three classes: content text, key-point and footline. The classification is based on the height and the average stroke width of the text line object, which is described as follows:

key-point if \[ st > sh \text{ mean } t > h\text{mean} \]

footline if \[ st < Sh \text{ mean} t < h\text{ mean} = y_{max} \]
content text otherwise; where smean and hmean denote the average stroke width and the average text line height of a slide frame, and ymax denotes the maximum vertical position of a text line object.

**Keyword Extraction and Video Search**

The formula for calculating TFIDF score is:

$$\text{tf-idf}_{\text{seg-int}(kw)} = \frac{1}{N} \left( \frac{\text{tf-idfocr}}{\sum_{\text{ntype}=1}^{\text{wi}}} + \frac{\text{tf-idfasr}}{} \right)$$

where kw is the current keyword, tf-idfocr and tf-idfasr denote its TFIDF score computed from OCR and ASR resource respectively, w is the weighting factor for various resources, ntype denotes the number of various OCR text line types. N is the number of available information resources, in which the current keyword can be found, namely the corresponding TFIDF score does not equal 0.

**IV. Results and Discussion**

### A. Dataset

We will use different video lectures which are done online in future. Videos lectures can also be downloaded from various standard websites available on web.

### B. Results

In following Table I show precision, recall and f1 measure value. Various setups are considered for that. This bar graph shows the accuracy Evaluation of Task1 of the existing system, accuracy is measured by Recall, precision and F1 Measure. This shows that the processing accuracy of the proposed system is greater than that of existing system.

**TABLE I. TABLE FOR EXISTING SYSTEM**

<table>
<thead>
<tr>
<th>Setup</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyframes and video</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>Keyword and video</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>All Features and video</td>
<td>0.96</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>Outline and video</td>
<td>0.87</td>
<td>0.95</td>
<td>0.91</td>
</tr>
</tbody>
</table>

**Figure 2. Graph for Existing System**

Following Table II describes table and graph for proposed system. This bar graph shows the accuracy Evaluation of Task1 of the proposed system, accuracy is measured by Recall, precision and F1 Measure. This shows that the processing accuracy of the proposed system is greater than that of existing system.

**TABLE II. TABLE FOR PROPOSED SYSTEM**

<table>
<thead>
<tr>
<th>Setup</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyframes and video</td>
<td>1.85</td>
<td>1.3</td>
<td>1.85</td>
</tr>
<tr>
<td>Keyword and video</td>
<td>1.23</td>
<td>1.15</td>
<td>1.74</td>
</tr>
<tr>
<td>All Features and video</td>
<td>1.05</td>
<td>1.54</td>
<td>1.57</td>
</tr>
<tr>
<td>Outline and video</td>
<td>1.14</td>
<td>1.25</td>
<td>1.62</td>
</tr>
<tr>
<td>Video only</td>
<td>1.7</td>
<td>1.4</td>
<td>1.03</td>
</tr>
</tbody>
</table>

**Figure 3. Graph for Time required for Training**

**V. Conclusion and Future Work**

In our proposed approach, for a content based video of lecture indexing and recovery in expansive files of
lecture video is presented. In order to check the examination hypothesis we apply visual and also sound resource of location features for concentrating metadata of content based. A couple of novel indexing features have been made in a video lectures gateway by using those metadata and a client study has been driven. In our work, we utilize procedures for correcting mistakes as a part of OCR translations. This procedure first creates n-gram strings for matching OCR unedited translations. The n-gram string contains all possible for sub-strings having least 3 characters. The second system additionally contains word reference of spelling check adjustment. For that MS word 2000 is utilized. It gives checking and correcting while spelling check. Our outcome demonstrates that our method reduced the quantity of false word furthermore it maintains a strategic distance from unauthenticated matching. In future one can use another comparative approach for improving this system.

Acknowledgment
The authors would like to thank the researchers as well as Publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule Pune University and concern members of conference, organized for their constant guidelines and support. We are also thankful to reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

References


