Video Retrieval using Textual Queries and Spoken Text

Laxmikant Kate
Department of Computer Engineering
Dattakala Faculty of Engineering
Swami Chincholi Daund, Pune India
lskate@hotmail.com

Prof. M. M. Waghmare
Department of Computer Engineering
Dattakala Faculty of Engineering
Swami Chincholi Daund, Pune India
monawaghmare25@gmail.com

***Abstract*— The expanded accessibility of broadband associations has as of late prompted an increment in the utilization of Internet webcasting. Most webcasts are filed and got to various times reflectively. One test to skimming also searching through such documents is the absence of text transcripts of the webcast's channel. Transcription of lectures is challenging assignment, both in acoustic and in language demonstrating. Recording lectures and putting them on the Web for access by understudies has turned into a general pattern at different colleges. To take full pick up of the information database that is manufactured by these records involved inquiry usefulness must be given that goes past pursuit on meta-information level however performs an itemized examination of the relating multimedia reports. Videos and Texts demonstrated in lecture are nearly identified with substance of the lecture, gives important source for recovering lecture videos and indexing. Text substance may be separated, then analyze and deducted consequently by OCR (Optical Character Recognition) strategies. In this paper, for remedying lapses in the OCR Transcriptions, we investigated two separate systems connected just to unmatched question words. In the first place strategy produces a new set of n-gram strings to match the unedited OCR Transcriptions. These n-gram incorporate strings with an altered separation of 1 character and all conceivable n-gram substrings with no less than 3 characters. Second system for redressing OCR included the word reference of spelling adjustment strategy gave in MS Words. The peculiarities of MS Word 2000, an OCR perceived string was extended through an application program interface into its corrected spellings. An exceptionally progressive style, just growing words that MS Word had flag as erroneously spelled which we depict before. This significantly decreased the quantity of spurious word competitors and maintained a strategic distance from false matches**

Keywords-lecture videos; automatic text indexing; n-gram string; OCR; lecture video.

I. INTRODUCTION

In the previous decade, we have seen a drastic increment in the accessibility of on-line academic lecture material. These educational assets can conceivably change the way individuals learn students with any type of disability can upgrade their educational experience, experts can stay aware of late progressions in their field and individuals of all ages can fulfill their hunger for knowledge. In complexity to numerous other informative exercises be that as it may, lecture transforming has as of not long ago delighted in little advantage from the advancement of human language innovation. In spite of the fact that there has been noteworthy exploration controlled to audio indexing and retrieval. Embedded content in a feature grouping gives important data of foremost essentialness. Messages more often than not show up as logos, subtitles, inscriptions or pennants in the feature grouping. Illustrations of such educational embedded writings can be to a great extent found in the news and other famous TV broadcastings. In spite of the fact that messages give extra data, not every one of them are essential as they may block critical allotments of a video. There are a few contrasts in the middle of news and lecture speech such as an abstract style and an accessible asset. These days, numerous telecast organizations give news cuts a comparing script through online administration. Since we can without much of a stretch form a preparing corpus utilizing this, telecast news retrieval has been a significant concentrate in talked report retrieval territory. In any case, they as of now give watchword hunt administrations utilizing a content search engine focused around the news script. Dissimilar to news, we can't undoubtedly get a script of lecture speech. In a business education site, an inquiry is performed utilizing a physically fabricated record.

Text is a high state having semantic features which has regularly been utilized for content-based information retrieval. In lecture videos, writings from lecture slides serve as a layout for the lecture and are critical for comprehension. In this manner in the wake of segmenting a video document into a set of key frames, the content detection method will be executed on each one key edge, and the extricated content articles will be further utilized as a part of content recognition also slide structure investigation forms. Particularly, the removed structural metadata can empower more adaptable video browsing and video pursuit capacities. Speech is a standout amongst the most essential transporters of information in video lectures. Consequently, it is of unique focal point that this information can be sought programmed lecture video indexing. Furthermore, the majority of the current lecture speech recognition frameworks in the explored work can't attain to a sufficient recognition result. A lot of literary metadata will be made by utilizing OCR and ASR system, which opens up the content of lecture videos. To empower a sensible access for the client, the delegate keywords are further concentrated from the OCR and ASR results. For content-based video look, the look lists are made from

distinctive information assets, including manual annotations, OCR and ASR keywords, worldwide metadata, etc. nowadays individuals have a tendency to create lecture videos by utilizing multi-scenes position, by which the speaker and his presentation are shown synchronously. This can be accomplished either by showing a video of the speaker and a synchronized slide record, or by applying a condition of the lecture recording framework, for example, tele-Teaching [10].

Anyplace Solution Kit such a framework which conveys two principle parts of the lecture: the principle scene of lecturers which is recorded by utilizing a video cam and the second which catches the desktop of the speaker as machine during the lecture through an edge grabber apparatus. The key advantage of the recent one for a lecturer is the adaptability. For the indexing, no additional synchronization in the middle of video and slide records is needed, and we don't have to deal with the slide position [11][12].

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

II. RELATED WORK

A texture-based system for detecting messages in images was exhibited by K. I. Kim, K. Jung, and J. H. Kim. The framework examines the textural properties of writings in images utilizing a SVM and spots the content regions by working CAMSHIFT on the texture classification results. The proposed system can encourage quick content detection, despite the fact that it doesn't accept the sort of media or the shade and textural properties of writings and is moderately uncaring to image determination. It additionally functions admirably in concentrating writings from mind boggling and textured foundations and was found to create a superior execution than some different methods. Nevertheless, the texture classifier did experience issues arranging little content or content with a low differentiation [1].

Content identification in image and feature with complex foundations and layering impacts is a troublesome and testing issue. D. Chen, H. Bourlard, and J. Thiran had introduced a quick content identification algorithm based on support vector machine. The algorithm first incorporates the edge, and heuristic proofs to concentrate the hopeful content lines and after that recognizes these applicants by utilizing SVM. The algorithm portrayed in this paper does not utilize shade, albeit numerous frameworks additionally make utilization of color data in detecting content in color images. The principle reason is that the begin purpose of our framework, the edge proof, is basically originating from force in layered image. Changing the RGB color image to YUV shade space and performing edge detection in U or V image can without much of a stretch discover this. No worldly data is utilized as a part of our algorithm. Since content may have diverse developments in feature, content identification is typically performed before following the content among the feature frames. The algorithm exhibited here attains to high identification rate and additionally low false caution rates. In quick content line extraction express, this algorithm is quicker than (or equal to) other quick content identification systems, despite the fact that the entire identification methodology is more CPU serious than region-based and edge-based routines. The assessment paradigm of the identification result introduced in this paper is on the premise of right benchmark restriction. This foundation is stricter than complete spread paradigm utilized as a part of past studies. With this basis, we can quantify the identification execution correctly without needing to demonstrate the last character distinguishes result [2].

In this study, another Farsi/Arabic text detection and localization methodology is proposed. Initially, with the assistance of edge extraction, fake corners are acquired and text dimension estimation is performed. Second, by consolidating discrete cosine transform coefficients, texture intensity based picture is made. Subsequently, another Local Binary Pattern (LBP) picture is acquainted with depicts the acquired texture design. The info picture is then separated into macro squares and a few features are extricated from them and sustained into Support Vector Machine (SVM) classifier to sort them into text and non-text groups. Trial results exhibit that the proposed half and half approach can be utilized as a programmed text detection framework, which is powerful to text dimension, textual style shade and foundation unpredictability [3].

Creator exhibit a two stage framework for programmed feature text removal to identify and remove installed feature texts what's more fill-in their remaining districts by fitting information. In the feature text detection stage, text areas in each one frame are discovered through an unsupervised clustering performed on the associated parts created by the stroke width transform (SWT). Since SWT needs a precise edge map, we create a novel edge identifier which profits from the geometric peculiarities uncovered by the bandlet transform. Next, the movement examples of the text objects of each one frame are broke down to confine feature texts. The recognized feature text areas are deleted, then, the feature is restored by an inpainting plan. The proposed feature inpainting approach applies spatio-temporal geometric streams removed by bandlets to remake the missing information. A 3d volume regularization algorithm, which exploits bandlet bases in exploiting the anisotropic regularities, is acquainted with complete the inpainting scheme. The strategy does not require additional courses of action to fulfill visual consistency [4].

E. Leeuwis, M. Federico, and M. Cettolo had focused their exertion on dialect demonstrating. A LM baseline was evaluated utilizing different sorts of information, which were all defective, however utilized as a part of such a path, to the point that their qualities were highlighted and not their lacks. Utilizing the ITC-irst WSJ AM adjusted on 8h of TED preparing information, it brought about a WER [5]. D. Lee and G. G. Lee had introduced a [6] Korean talked report recovery framework or secondary school math address features which utilizes the substance data. From ASR yields and address note, they constructed the general transformed list table, the substance table and the matching table. These tables are utilized for figuring the

significance score. In the study [7] by J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, they closed that the specialized dialect of scholastic addresses and absence of in-area talked information for preparing makes address interpretation a critical test that will require new systems for inferring a vocabulary and dialect model. Pong et al. [8] proposed the algorithm of segmentation in their work is focused around the differential degree of text and foundation areas. Utilizing limits they endeavor to catch the slide move. The studies depicted in Repp et al. [9] are focused around out-of-the-container business discourse different programming. Concerning such business programming, to accomplish fulfilling results for an extraordinary working space an adaption methodology is frequently needed, yet the custom expansion is once in a while conceivable.

III. IMPLEMENTATION DETAILS

A. *System Architecture*

Following Figure 1 shows proposed system architecture. The detailed description is as follows:

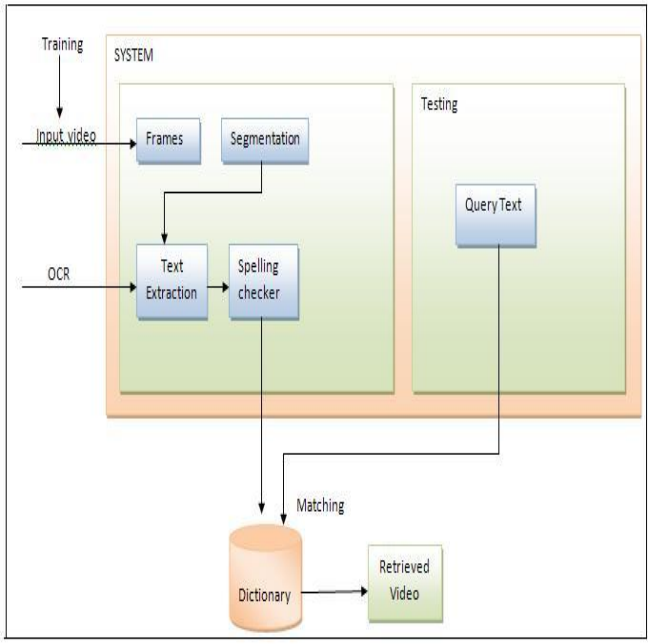

Figure1. Proposed System Architecture

Working of the system is divided into two parts:

1) Training Phase
2) Testing Phase

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. i.e. 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.

A. *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 mean t > hmean

footline if st < $s_{h\ mean}$t < $h_{y\ mean}$ = y_{max}

content text otherwise; where s_{mean} and h_{mean} denote the average stroke width and the average text line height of a slide frame, and y_{max} denotes the maximum vertical position of a text line object.

Keyword extraction and Video search

Formula for calculating TFIDF score

tf id_{seg}-internal(kw) =

1/N(tf id_{ocr}. 1/n_{type} Σn_{type} i 1wi + tf idasr .w_{asr})……(1)

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.

B. *Mathematical Model*

Let, the system S is represented as:

S = {T, F, S, T, C, R}…… (2)

1. **Training input video**

T is a set of all training input videos given to the system,

T = {t_1,t_2,t_3, ….}

Where, t_1, t_2 … are the number of input videos given.

2**. Framing**

F is a set of framing input videos which are extracted

F = {f_1,f_2,f_3,…}

Where, f_1, f_2 … are the number of different frames.

3. Segmentation Phase

Let, S is a set of segments

S = {s_1,s_2,s_3…}

Where, s_1,s_2,s_3.. are the number of different segments

4. Text Extraction

Let, T is a set of extracting text

T = {t_1,t_2,t_3,..}

Where, t_1,t_2,t_3.. are the number of extracted texts.

5. Checking for Spelling

Let, C is a set for spelling check

$C = \{c_1, c_2, c_3 \ldots\}$

Where, c_1, $c_2 \ldots$ are the number of checker for spelling.

6. Retrieved Video Output

Let, R is a set for retrieve videos

$R = \{r_1, r_2, r3 \ldots\}$

Where, r_1, $r_2 \ldots$ are the number of various retrieved videos as output.

C. *Experimental Setup*

The system is built using Java framework (version jdk 6) on Windows platform. The Netbeans (version 6.9) is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

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.

TABLE I. TABLE FOR EXISTING SYSTEM

Setup	Recall	Precision	F1 Measure
Keyframes and video	0.99	1	0.99
Keyword and video	0.99	1	0.99
All Features and video	0.96	0.99	0.97
Outline and video	0.87	0.95	0.91
Video only	0.81	0.83	0.82

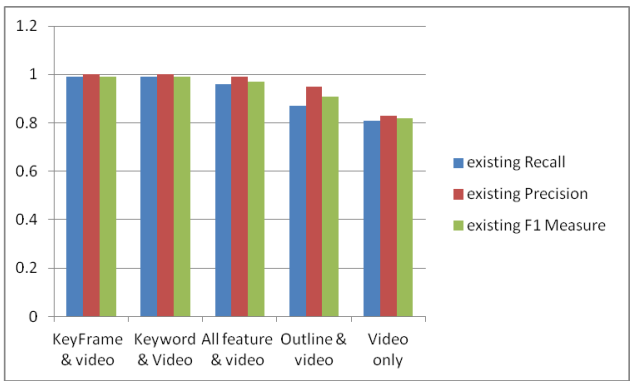

Figure 2. Graph for Existing System

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

Setup	Recall	Precision	F1 Measure
Keyframes and video	1.85	1.3	1.85
Keyword and video	1.23	1.15	1.74
All Features and video	1.05	1.54	1.57
Outline and video	1.14	1.25	1.62
Video only	1.7	1.4	1.03

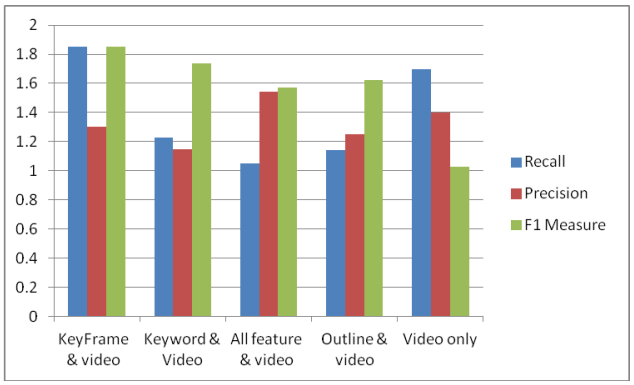

Figure 3. Graph for Time required for Training

V. CONCLUSION AND FUTURE WORK

In this paper, we displayed a methodology for content based video of lecture indexing and recovery in expansive archives of lecture video. So as to check the examination theory we apply visual and additionally sound asset of address features for concentrating content based metadata consequently. A few novel indexing features have been created in an expansive video lectures portal by utilizing those metadata and a user study has been led. In our work, we use techniques for correcting errors in OCR transcriptions. This technique first generates n-gram strings for matching OCR unedited transcriptions. The n-gram string contains all possibility for substrings having minimum 3 characters. The second method also contains dictionary of spelling check correction. For that MS word 2000 is used. It provides correction while spelling check. Our result proves that our technique reduces the number of false word candidates and also it avoids unauthenticated matching.

This technique of video retrieval using textual queries and spoken text is useful in TV broadcasting and internet video.

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 iPGCON2015 conference, organized by Sangamner 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

[1] Kwang Kwang In Kim, Keechul Jung, Jin Hyung Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm", Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:25, Issue: 12 .

[2] Datong chen,"Text Identification in complex Background Using SVM", IEEE, 2001.

[3] Mohieddin Moradi, Saeed Mozaffari, "Hybrid approach for Farsi/Arabic text detection and localisation in video frames", IET Image Process., 2013, Vol. 7, Iss. 2, pp. 154–164 doi: 10.1049/iet-ipr.2012.0441.

[4] Ali Mosleh," Automatic Inpainting Scheme for Video Text Detection and Removal", IEEE Transaction on mage Processing, Vol. 22, No. 11, November, 2013.

[5] E. Leeuwis, M. Federico, and M. Cettolo, ˆaLanguage modeling and transcription of the ted corpus lectures,ˆa in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2003, pp. 232ˆa235.

[6] D. Lee and G. G. Lee, ˆaA korean spoken document retrieval system for lecture search,ˆa in Proc. ACM Special Interest Group Inf. Retrieval Searching Spontaneous Conversational Speech Workshop, 2008.

[7] J. Glass, T. J. Hazen, L. Hetherington, and C.Wang, ˆaAnalysis and processing of lecture audio data: Preliminary investigations,ˆa in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval, 2004, pp. 9ˆa12.

[8] T.-C. Pong, F. Wang, and C.-W. Ngo, ˆaStructuring low-quality videotaped lectures for cross-reference browsing by video text analysis,ˆa J. Pattern Recog., vol. 41, no. 10, pp. 3257ˆa3269, 2008.

[9] S. Repp, A. Gross, and C. Meinel, ˆaBrowsing within lecture videos based on the chain index of speech transcription,ˆa IEEE Trans. Learn. Technol., vol. 1, no. 3, pp. 145ˆa156, Jul. 2008.

[10] Haojin Yang and Christoph Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information", IEEE Transaction on Learning Technologies, Vol. 7, No. 2, April-Junr 2014.

[11] A. Haubold and J. R. Kender, "Augmented segmentation and visualization for presentation videos," in Proc. 13th Annu. ACM Int. Conf. Multimedia, 2005, pp. 51–60.

[12] W. H€urst, T. Kreuzer, and M. Wiesenh€utter, "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web," in Proc. IADIS Int. Conf. WWW/Internet, 2002, pp. 135–143.