Video Retrieval Using Speech and Text InVideo

N. Radha,
Department of Information Technology,
SSN College of Engineering, Chennai, India
radhan @ssn.edu.in

***Abstract* - This paper presents an analysis process of lecture video retrieval with automated video indexing and video search for lecture video databases. The video retrieval extracts the relevant metadata from the two main parts of lecture videos, namely the visual screen and audio tracks. From the visual screen, we firstly detect the slide transition and extract each unique slide frame with its temporal transition considered as the video segment. The textual metadata from slide frames is extracted and then recognized using video OCR technique. Based on OCR results, the corresponding text in video information is saved. Secondly, the speech-to-text analysis is carried out from the audio signal. Sphinx speech recognition models are used for recognition of speech-to-text conversion process. In the proposed work, the combined text information from the visual slide frame and audio signal were used to retrieve the video from the lecture video database. The performance of this combined video retrieval system shows a significant improvement in performance when compared with an individual system built using audio and text in video systems respectively.**

Keywords -- Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), World Wide Web (WWW), Hidden Markov Model (HMM)

I. INTRODUCTION

E-lecturing based content retrieval has become more popular from a user interface point of view. The amount of video, especially lecture video based data on the WWW is also growing rapidly. Many of the organizations, universities and research oriented institutions are primarily involved in creating such e-lecturing videos. This shows that,there is a huge increase in video data on the web. The audio and visual based recorded e-lecturing videos were used more frequently and also queried by many of the user. From the end user it is very difficult to find out the relevant video from the web as well as from the video archives. Therefore, a more efficient method for the video retrieval in WWW or within large lecture video archives is desperately needed. To simplify end user's searching methodology and to retrieve the relevant video, this particular work proposes an approach for automated video indexing and video search in large e-lecture video archives.

[1] have presented a survey on content-based video indexing and retrieval. In general, a video is defined by metadata, audio and visual information. The various applications of video retrieval are,browsing of video folders, visual electronic commerce, remote instruction, digital museums, news event analysis, intelligent management of web videos, video surveillance. Table 1 shows the summary of the various video structure analysis methods and their corresponding features thatwereused for video indexing and retrieval.

Table 1. Summary of video structure analysis for video retrieval

Methods	Features	Metrics	Classification
Shots (Video structure Analysis)	Block or Colour histogram Edge change ratio Motion vectors Scale invariant feature transform Corner points Information saliency map.	Similarity measures : 1-norm cosine dissimilarity Euclidean distance Histogram intersection Chi-squared similarity and novel similarity measures etc.	Threshold-Based: comparison measure between frames with a predefined threshold Supervised Learning-Based: depending on the features, frames are classified as shot change or no shot change Ex: SVM, Adaboost, kNN, HMM Unsupervised Learning-based: Ex: K-means & fuzzy K-means
Key frames (Video structure Analysis)	Colour, edges, Texture Shapes, optical flow, Motion descriptors MPEG discrete cosine coefficient camera activity, Image variation features are caused by camera motion	Sequential comparison Global comparison Reference frame-based Clustering based Curve simplification-based Object/event-based	Gaussian mixture models (GMM) HMM
Scenes (Video structure Analysis)	Features defined for shots and key frames are used here.	key frame based, audio and visual information integration-based, merging & splitting-based, statistical model-based, and shot boundary classification-based.	GMM, HMM

[3] provides an evaluation effects of different types of information used for video retrieval from a video collection. Image similarity matching, spoken and OCR based document retrieval are the different types of information process retrieval from the video. The multimodal information for the video retrieval is already explored and TERC video retrieval track evaluation is discussed [2]. [4] have presented analysis process for retrieving metadata from video, using audio track and visual information. Textual metadata extraction from the audio tracks is done using CMU Sphinx recognition toolkit. OCR is used to extract metadata from visual slide of frames of videos. This work, proposes the performance of the integrated cue is higher in quality to that of individual acoustic (audio) and visual sub systems. Spoken text and visual slides are also used for the presentation of video retrieval [5]. Manually, 60 videos were collected and used for the evaluation. Experiments clearly state

that automatically extracted visual slide text gives higher precision video retrieval when compared with automatically recovered spoken text. Combining automatically extracted visual slide-text data and ASR (spoken text) achieves improved performance over both individual modalities.

[6] have presented multimodal information retrieval for a broadcast video using visual information and speech recognition. A lot of significant improvements have been made by the researches in improving content based video retrieval systems. The various steps involved in video retrieval systems are

- Types of videos used
- Key frames selection
- Feature extraction
- Classification
- Indexing
- Query browsing and results
- User Interface design.

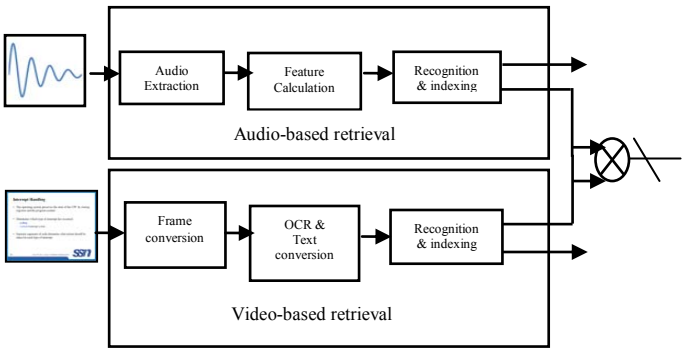

Fig1. Block diagram of proposed method

Figure 1 shows a schematic diagram of video retrieval system approach using the two modes of audio and visual/video text in frameinformation. Automatic video segmentation and key-frame detection is a primary step that offersthe visual guidelines for the video content navigation. Subsequently, textual metadata extraction is done by applying video (OCR) technology on visual frames and ASR is performed on lecture audio tracks. The OCR, ASR transcript and in addition detected slide text line types are adopted for keyword extraction, by which both video- and segment-level keywords are extracted for content-based video browsing and for search. The performance and the effectiveness of proposed indexing functionalities were proven by evaluation. The video retrieval system which involves the following steps audio extraction, visual frameconversion, text extraction from audio and video, combining index score, and the identification decision making. The system mainly contains two types of video data extraction: Audio, Visual frame separately. From the audio information module index scores are computed, based on given feature observations of audio signal. Visual frame to textconversion is the next step. OCR is used to extract text information from the video and indexing scores are computed. Finally, index scores (Audio+ Text in video) are combined using a sum rule, combined score is defined below

$$v = (1 - w)a_t + w\, v_t 0.5 \leq w < 1,\ m \in \mathbb{R} \quad (1)$$

Where$v, a_t, v_t$are the scores of the combined, speech recognition and visual system respectively.

This paper is organized as follows. In section 2, A detailed discussion is about the Audio extraction, conversion, feature extraction and recognition module. Visual slide frames to text extraction using OCR based is discussed in section 2, followed by a discussion on the combined system of audio and visual indexing scores of the system. Experiment results for video retrieval systems are discussed detail in section 3. Section 4 summarizes the work.

II. SYSTEM DESIGN

This section deals with the textual data extraction from audio and the visual signal. In the first sub section, the speech signal to text (speech recognition) conversion study was presented. Feature extraction, recognition and modeling techniques for speech recognition is also discussed. The second sub section, discusses about the visual image to text recognition and their customized modules is discussed. Combined classifier systems are explained in the sub section 3.

A. Text extracttion from Audio

The various processing steps of speech recognition system (ASR) is shown in figure 2. typically involves following signal conditioning, segmentation, feature extraction, modeling and recognition. Signal conditioning is the removal of noise from the speech signal and segmentation separates the speech from non speech regions. The feature extraction is a stable, most discriminant parametric representation and dimension reduction are also parts of the task of speech recognition system. The features extraction technique uses the source–system model of speech, and parameterization is done based on vocal tract characteristics, while the excitation source characteristics was ignored for speech recognition task. The filter bank analyzer in speech recognition is used to a measure the energy of the given speech signal with a specified frequency band. The filter bank analysis gives limited features as an example energy of the speech signal, and this is not enough to measure the performance of speech recognizers with this limited features alone, because of the non stationary property of an speech signal. Linear predictive coding and MFCC (cepstral) arethe other methods areused to extract the features from speech signalof speech recognition system.Cepstral analysis has been widely used spectral representation methods in speech recognition.

Cepstral coefficients are calculated by using the following steps: pre-emphasis, mel filter bank analysis, log energy computation, dynamic cepstral coefficient. Pre-emphasis for a given speech signal$[x^{!}(n)]$ is a first step used to boost the energy in the high frequency form. Boosting high-frequency

energy gives more information to the acoustic model and thus helpsto improve the phone recognition performance.

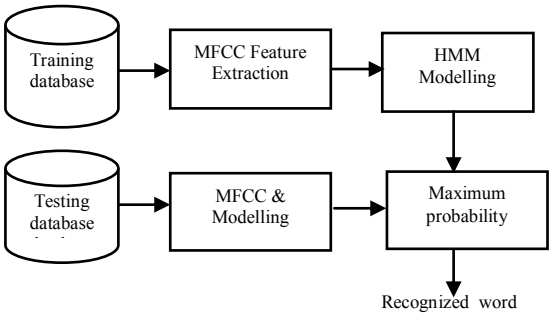

Fig 2. Block diagram of ASR system

Next, Windowing (rectangular $[w_r(n)]$ or hamming $[w_h(n)]$) performed, because of the stationary property of speech signal, and thus resultant information in a small region is also a useful cue for speech recognition. Energy (E) is an one of the feature of speech spectrum which is calculated after windowing operation. The following equation shows that the representation of rectangular, hamming and energy sources respectively.

$$w_r(n) = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & otherwise \end{cases} \quad (1)$$

$$w_h(n) = \begin{cases} 0.54 - 0.46\, cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n \le L-1 \\ 0 & otherwise \end{cases} \quad (2)$$

$$E = \sum_{t=1}^{t=n} x^2\,(\mathrm{t}) \quad (3)$$

DFT also performed over on windowing signal and ensuingin terms of magnitude and phase frequency components. The bank of filters are applied according to mel scale $mel(f)$[ie. eq. 4] to overaspeech spectrum and thus each filter output is a sum of its filtered spectral components. Then logarithm computation of the spectral magnitude of the output of Mel-filter bank gives the delta $[\Delta y(k)$ ie.eq.5] and double delta $[\Delta\Delta y(k)]$ coefficients. The MFCC features are widely used as a feature extraction method for robust speech recognition.

$$X(k) = \sum_{m=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \;//\; mel(f) = 2595\, log_{10}(1 + f/100) \;//\; Y(k) = X_t(k)\, mel(f) \quad (4)$$

$$y(k) = \sum_{m=1}^{M} log\left(|Y_t(m)|\cos(k(m-0.5)\frac{\pi}{M}\right), k = 0, \ldots, J \quad (5)$$

After feature extraction, modeling is a next step, HMM model isused to begin with the training process. The initial model can be randomly chosen or selected based on a prioriknowledge of the model parameters. The model can be re-estimated using the viterbi algorithm which maximizes the likelihood of model M for having generated the observed sequences (feature vectors). Once the HMM is trained the output result of recognition unit determines the class to which its belong in the dataset. The best state transition path calculated, and determines the likelihood value using the viterbi search algorithm for recognition. The compared likelihood value of each HMM and determines the input word as the index of HMM with the largest likelihood value. Generation of a text file with the recognized text as an output. In this work, uses the Sphinx speech recognition system. Text extraction from video using OCR is discussed in the next sub section.

B. Text Extraction from video

In this section, discuss about the different component technologies that constitute towards to an end-to-end video to text recognition system. Figure 3 contains a block diagram, that shows the processing sequence of a typical video text processing system which consist of the following steps are video text detection, localization, extraction, and recognition. The first step is the text feature extraction from video frames and thus affected by many factors like background, unknown text color information and various other features etc.

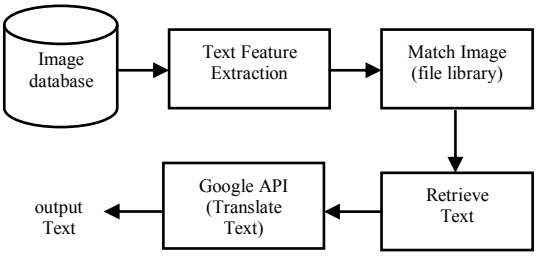

Fig 3. Block diagram of OCR system

In this work, every frame slides are processed by their header information alone. To detect a region in a frame, the following windowing operation [eq.6,7] is performed over the frame slides is shown figure 4.

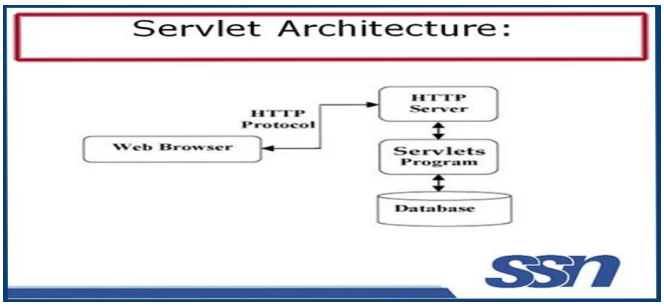

Fig 4. Region extraction from video frame

The redundancy has been reduced from the multiple copy of header information from the slide frames. In a second step, extracted image features were matched with existing file library and thus the resultant gives the text information. Extracted text from video frame slides are translated via google API function and displays the relevant text information.

$$X_{min} < X < X_{max} \quad X < X_{min} \mid X > X_{\min} \quad (6)$$

$$Y_{min} < Y < Y_{max} \quad Y < Y_{min} \mid Y > Y_{min} \qquad (7)$$

Where$X_{min}, X_{max}, Y_{min}, Y_{max}, X, Y$ are the left,right, top and bottom of selected region of video frame respectively. Figure 5 shows the conversion process of visual slide frame to text. Searching video and the video display is shown in Figure 5 and 6 respectively. Combined ASR and OCR system are discussed in the next sub section.

C. Combined ASR and OCR system

Since the speech recognition system and video-text recognition system carries complementary information about the presentation slides. These systems could be integrated to obtain a better performance. Later integration is performed using a scaled and sum rule [9]. The combined system (AV), using speech recognition (A) and visual frame to text (V) is obtained as follows:

$$AV = (1 - \alpha)\, A + \alpha\, V \qquad (8)$$

where α is the scaling factor. The video slide frame to text and audio to text recognition experimental results are presented in the next section.

III. EXPERIMENTS

The database used in this study consists of multimodal audio video data corpus which is collected from 20 speakers (12 male and 8 female speakers). Simultaneous recording audio and visual informationarecaptured. The recordings are done in a laboratory environment using a microphone and a sony camera. The SONY Handycam HDR-PJ660/B Camcorder is used for video recording (30 Minutes). All the utterances are recorded under the same lighting and with normal environmental conditions. The video consists of 25 f/swith frame width 640 and frame height 480. The horizontal distance from speaker's position to video is about 45 cm and 85 cm is the vertical measurement of a camera.

TABLE 1. Recognition of audio extraction from video

Domain Selection	Duration (s)	No of videos	Recognition Accuracy (%)
WT-D	20	75	78
DTW-D	20	75	82
IP-D	20	50	83.5

TABLE 2. Recognition of text extraction from video

Domain Selection	Duration (s)	No of videos	Frames selection	Recognition Accuracy (%)
WT-D	30	50	65	67.8
DTW-D	30	50	65	75.2
IP-D	30	50	65	70.2

Table 1 shows the recognition accuracy of audio data. Three different domain selections are considered for this work for video retrieval process. The duration of 20 sec is fixed for the all the groups of a domain in the video processing. Different experimental recognition results were provided. In that IP domain gives us higher recognition accuracy compared to the other domains.

Table 2 shows that recognition accuracy of text data from video. The duration of 30 sec is fixed for the all the groups of a domain in the video processing. Each video has beenundergone through normalization process and thus frame selectionis fixed upto 65. Experimental results were analyzed and DTW-D domain gives better performance recognition accuracy compared to the other.

TABLE 3. Recognition of video retrieval extraction from video

Domain Selection	Precision (%)	Recall (%)
Audio	78.4	60.2
Video/Text in Frame (TIV)	80.9	72.5
Audio + TIV (proposed)	88.9	76.5

The recognition accuracy for video retrieval is calculated by two measures namely precision(%) and recall(%) receptively. The precision accuracy measure for video is calculated by the no correct videos and total no of relevant videos correctly recognized. Similarly, the recall for a video is defined by the no of relevant(correct) videos and total no of videos in the database. The video retrieval using audio data information is a next step carried out for further processing. It gives the 85.4 % recognition accuracy. Text in frame-based (video) retrieval process performs with 73%. Combined audio and text-based (video) retrieval, produces the highest recognition accuracy.

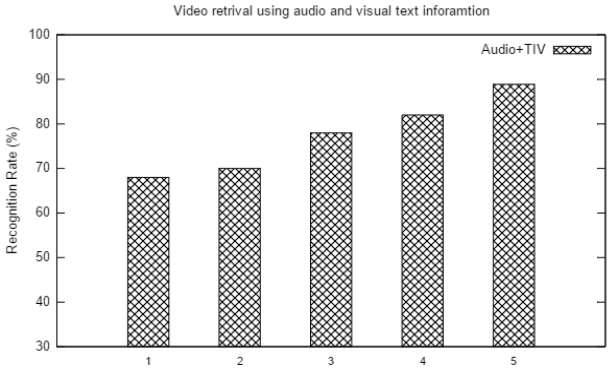

Fig 4. Recognition accuracy for combined system

Figure 4 shows the performance of the combined video retrieval system for different values of the scaling factor values (ranging from 0.5 to 0.9) are indexed to 1 to 5. The combined system with indexing value5 gives the highest recognition performance.

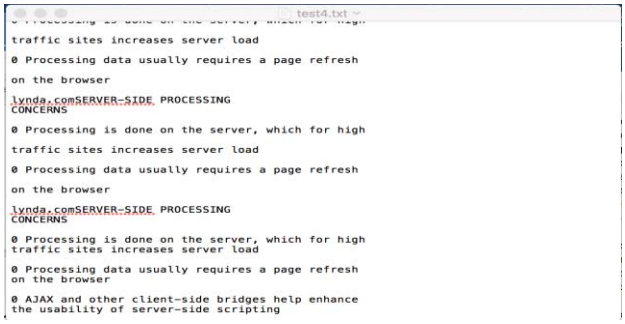

Fig 5. Coversion of Image to text

Fig 6. Searching video

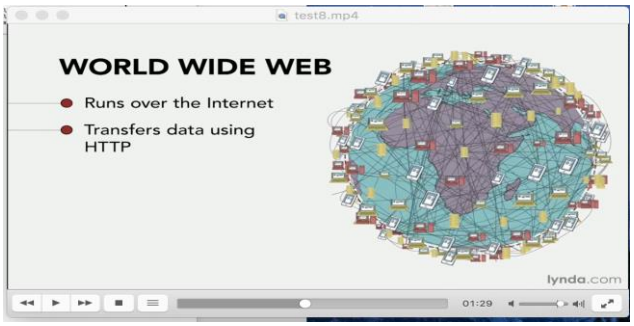

Fig 7. Video gets diplayed

IV. SUMMARY AND CONCLUSION

In order to verify the research hypothesis, audio and visual resource of lecture videos used for extracting content-based metadata automatically. An improved video retrieval system is proposed, which combines the well-known audio and visual feature level information for largelecture video datasets. A late integration using the scaled sum rule is performed here to combine the information from the two individual modes. For audio to text conversion, Sphinx speech recognition models were used. Image to text conversion is performed by OCR. The experimental evaluation is tested for the various lecture videos. This system is evaluated with the speech recognition, visual frame slide to text at the first and thus gives the proper experimental recognition results for video retrieval. The performance of this system is again tested by the combination above strategies and thus shows that reliable identification results than compared to individual one. Combined audio and visual text information gives us high recognition accuracy.

REFERENCES

[1] Hu, Weiming, et al. "A survey on visual content-based video indexing and retrieval." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41.6 (2011): 797-819.

[2] Hu, Weiming, et al. "A survey on visual content-based video indexing and retrieval." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41.6 (2011): 797-819.

[3] Voorhees E.M, and Tice, D.M., "The TREC-8 Question Answering Track Report," The Eighth Text Retrieval Conference (TREC-8), 2000

[4] Hauptmann, Alexander G., Rong Jin, and Tobun D. Ng. "Video retrieval using speech and image information." Electronic Imaging 2003. International Society for Optics and Photonics, 2003.

[5] Yang, Haojin, and Christoph Meinel. "Content based lecture video retrieval using speech and video text information." Learning Technologies, IEEE Transactions on 7.2 (2014): 142-154..

[6] Cooper, Matthew. "Presentation video retrieval using automatically recovered slide and spoken text." IS&T/SPIE Electronic Imaging (pp. 86670E--86670E). International Society for Optics and Photonics (2013).

[7] Hauptmann, Alexander G., Rong Jin, and Tobun Dorbin Ng. "Multi-modal information retrieval from broadcast video using ocr and speech recognition."Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. ACM, 2002.

[8] Wengang, Cheng, and Xu De. "Content-based video retrieval using audio and visual clues." TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. Vol. 1. IEEE, 2002.

[9] N. Radha, A. Shahina, A. Nayeemulla Khan, "A Person identification system combining recognition of face and visual lip-read passwords", in the Proc. of International Conference on Computing and Netwok Communications, pp. 882-885, Dec 2015.