2015 International Conference on Information Processing (ICIP)
Vishwakarma Institute of Technology. Dec 16-19, 2015

Content Based Video Retrieval Using Integrated Feature Extraction and Personalization of Results

Mr. Pradeep Chivadshetti [1],
PG. Student,
Department of IT, Sinhgad College of Engineering, Pune, Maharashtra, India [1]
pradeepchivadshetti@gmail.com[1]

Mr. Kishor Sadafale [2]
Assistant Professor,
Department of IT, Sinhgad College of Engineering, Pune, Maharashtra, India [2]
kbsadafale.scoe@sinhgad.edu[2]

Mrs. Kalpana Thakare[3]
Associate Professor,
Department of IT, Sinhgad College of Engineering, Pune, Maharashtra, India [3]
ksthakare.scoe@sinhgad.edu[3]

***Abstract*: Traditional video retrieval methods fail to meet technical challenges due to large and rapid growth of multimedia data, demanding effective retrieval systems. In the last decade Content Based Video Retrieval (CBVR) has become more and more popular. The amount of lecture video data on the World Wide Web (WWW) is growing rapidly. Therefore, a more efficient method for video retrieval in WWW or within large video archives is urgently needed. This paper presents an implementation of automated video indexing and video search in large video database and also present personalized results. Proposed system works in three different phases, in the first phase video segmentation and key frame detection is performed to extract meaningful key frames. Secondly, OCR, HOG and ASR algorithms are applied over the keyframe to extract textual keyword. In the third phase, Color, Texture and Edge features are also extracted. Finally, search similarity measure is performed on the extracted features that are saved in SQL database and the output with personalised re-ranking results as per interest is presented to the users.**

Index Terms: CBVR, Feature Extraction, Video Retrieval, Video Segmentation, OCR, ASR tool, Re-ranking

I. INTRODUCTION

CBVR, in the application of video retrieval, is the issue of searching for digital videos in large databases with less input keywords. "Content-based" is the search which analyse the actual content of the video. The term 'Content' in this context might refer colour, texture keywords, and audios. Without the ability to examine video content, searches must rely on images provided by the user. Explosive growth of digital content including image, audio and video on internet as well as on desktop has demanded development of new technologies and methods for representation, storage and retrieval of multimedia systems. Rapid development of digital libraries and repositories are attempting to achieve the same. CBVR system works more effectively as these deals with content of video rather than video metadata [10].

Videos mainly contains Text, Audio and Images. Generally, CBVR system extracts the features by using different methods: Metadata-based, Text-based, Audio-based and Content-based. Metadata-based method which extracts the title, type, modified date, etc. Text based method which extracts the subtitles, extracted text from OCR. In the Audio-based method different speech recognition and speech to keyword extraction techniques are used. At last, Content-based method which is integration of all methods mentioned below (shown in figure 1).

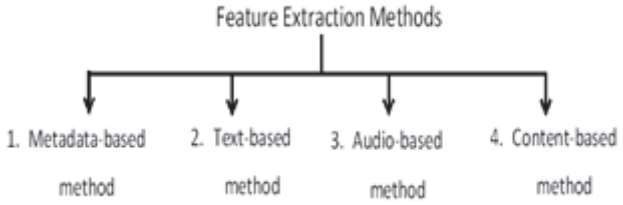

Fig. 1. Various Feature Extraction Methods

A. Video Parsing

Video processing is always performed on frames which are basic block of video. Group of frames captured together is called shot. Few minutes shot may contain hundreds of frames, that makes video large in size. Storing and processing of these individual frames are memory and computational expensive [2]. Also there is a very minute change of content information between the two consecutive frames of same shot. Selection of frames from single shot may be done to identify the key frame which represent complete shot. These key frames are then used for indexing, content processing and representing video shot/video.

B. Feature Extraction

By using the above segmented frames, we are extracting the different features: Colour, Texture, Edge, OCR, HOG, and ASR [1].

C. Video Indexing

This process is retrieving the information about the frame for indexing in a database. Video indexing is a process of tagging videos and organizing them in an effective manner for fast access and retrieval [3]. Automation of indexing can significantly reduce processing cost while eliminating tedious work. The conventional features used in most of the existing video retrieval systems are the features such as colour, texture, shape, motion, object, face, audio, etc. It is obvious that more the number of features used to represent the data, better the retrieval accuracy.

D. Video Retrieval and Browsing

Where users can access the database through queries based on text and/or visual examples or browses it through interaction with displays of meaningful icons. Users can also browse the result of query retrieval. It is meaningful that both retrieval and browsing appeal to the user's visual intuition.

978-1-4673-7758-4/15/$31.00 ©2015 IEEE

II. Related Work

Despite many research efforts, the existing low-level features are still not powerful enough to represent index frame content. Some features can achieve relatively good performance, but their feature dimensions are usually too high, or the implementation of the algorithm is difficult. Feature extraction is very crucial step in retrieval system to describe the video with minimum number of descriptors. Table I. representing the different challenges for the basic video search.

TABLE I.
BASIC VIDEO SEARCH CHALLENGES

Sr. No.	Issues	Text Search	Video Search
1	File Format	HTML/HTTP, PDF, DOC	MPEG 1,2,4, MOV, WMV, Real/HTTP, UDP
2	Summarization	Easy to extract relevant segment.	Requires video parsing first.
3	Browsing	Parallel	Serial (Linear Media)

Several CBVR systems used different features or techniques for video retrieval as follows:

A. *Metadata*:

The basic feature extraction was extraction of metadata and textual information of video. Video was retrieved by using that features like title, subtitle, properties (extension, modified date, size, etc.) [7].

B. *Colour and Texture*:

Proficient detection and segmentation of text characters from the background is necessary to fill the gap between image documents and the input of standard OCR systems [5]. The basic visual features of index frame include colour and texture. Research in content based video retrieval today is a lively disciplined, expanding in breadth Representative features extracted from index frames are stored in feature database and used for object-based video retrieval. Texture is another important property of index frames. Various texture representations have been investigated in pattern recognition and computer vision [7].

C. *Edge detection, Colour, Shape, Shot boundary etc.*:

Explosive growth of digital content including image, audio and video on internet as well as on desktop has demanded development of new technologies and methods for representation, storage and retrieval of multimedia systems. Video feature database is created using entropy feature extracted from key video frames of video database. Same feature is extracted from video frame query [7].

D. *ASR (Automatic Speech Recognition)*:

In addition to video ASR can provide speech-to-text information from different videos, which offers the chance to improve the quantity of automatically generated metadata dramatically. However, as mentioned, most video speech recognition systems cannot achieve a sufficient recognition rate [4].

E. *OCR (Optical Character Recognition)*:

An end-to-end system for text detection and recognition is important in multiple domains such as content based retrieval systems, video event detection, human computer interaction, autonomous robot or vehicle navigation and vehicle license plate recognition. Text detection in natural scenes is a challenging problem and has gained a lot of attention recently. Such texts presents low contrast with background, large variation in font, colour, scale and orientation combined with background clutter [5].

There have been numerous research efforts on Text Detection and Recognition by using OCR applications. Number of approaches for text detection in images has been proposed into the past. Automatic detection and translation of text in images done using different techniques proposed. These methods aim to detect the characters based on general properties of character pixels. The distribution of edges, colour is used in many text detection methods also for low resolution document are processed by particular method. Text detection and recognition in images and video frames, process is combination of advanced optical character recognition (OCR) and text-based searching technologies.

Unfortunately, text characters contained in images can be any grey-scale value (not always white), variable size, low-resolution and embedded in complex backgrounds. Texture is commonly used feature for text segmentation. Many researchers working on text detection and thresholding algorithm with various approaches achieved good performance depends on some constraints. Therefore, in this section, the overall comparisons of all the techniques are discussed. Different content based video retrieval techniques are discussed like feature extraction using colour, shape, texture, etc. and keyword extraction using Metadata, OCR and ASR. It is represented in the table II.

III. Proposed System

Proposed CBVR Personalized Systems implemented on six modules i.e. creation of Features and stores in database and retrieval using feature extraction with similarity measures and personalization as shown in figure 2. Firstly, a user uploads or gives a text/image/video query as input to the CBVR Personalized system. CBVR System will divide the video into frames and does selection process of relevant frames into all frames. Simultaneously ASR system will process on video input and extract the keywords by ASR tool. After frame segmentation and selection, perform OCR and extract the HOG, OCR text and Gabor Filter from selected frames and also extract the Color, Texture and Edge detector on selected frames and at last also extract the keywords and features. The same process of ASR, Frame segmentation, OCR and image processing is done on videos stored on database [1]. After pre-processing system will search for similarity in keywords and features of user query metadata and all video which are stored in database. CBVR system extracts the most matching OCR text, ASR text and keywords and features and generates relevant final video results.

TABLE II.
COMPARISON BETWEEN DIFFERENT TECHNIQUES FOR VIDEO RETRIEVAL

Sr. No.	Authors	Used Techniques	Work description	Remarks	Ref. No.
1	Kuo, T.C.T. et. al. [1996]	A Content-Based Query Language for Video Databases	Content objects are used to Extract metadata, simple keywords, user can sketch query	Users retrieve video data by specifying the state of video contents. A set of video functions is provided for describing the spatial and temporal relationships between content objects in the query predicate.	8
2	Volkmer, T. et. al. [2006]	Exploring Automatic Query Refinement For Text-Based Video Retrieval	Automatic query filtering, video speech transcripts, improvements of up to 40%.	Excellent potential for improving speech-based video retrieval.	9
3	B. V. Patel et. al. [2010]	Content Based Video Retrieval using Entropy, Edge Detection, Black and White Colour Features	Extracting Entropy, Edge detection and colour features for feature extraction	This approach can further be enhanced by integrating content features like frequency, histogram, etc. with data mining techniques.	7
5	Padmakala S. et. al. [2011]	An Effective Content Based Video Retrieval Utilizing Texture, Colour, and Optimal Key frame Features	At first, the input raw video is segmented using video object segmentation algorithm, Then, feature vectors are computed from VSR using the texture analysis and colour moments.	Only extract the keywords from features, low performance of optical key frame presentation	11
6	Kale, A. et. al. [2013]	Video Retrieval Using Automatically Extracted Audio	Used speech recognition algorithm to extract keywords	Limited to metadata and Audio, different Language may reduce performance	12
7	Hadi Yarmohammadi et. al. [2013]	Content Based Video Retrieval Using Information Theory	Proposed for video retrievals an summarization problem	Doesn't clustered automatically	6
8	Arpit Jain et. al. [2014]	Text Detection and Recognition in Natural Scenes and Consumer Videos	Proposed end to end system for text detection in natural images and videos	Extensive evaluation on a large dataset illustrates in both pixel-level text Detection and word recognition tasks.	5
9	Haojin Yang et.al. [2014]	Content Based Lecture Video Retrieval Using Speech and Video Text information	Extracting keywords by OCR, ASR, Features	Integration of all the feature extraction gives better and relevant output result	4

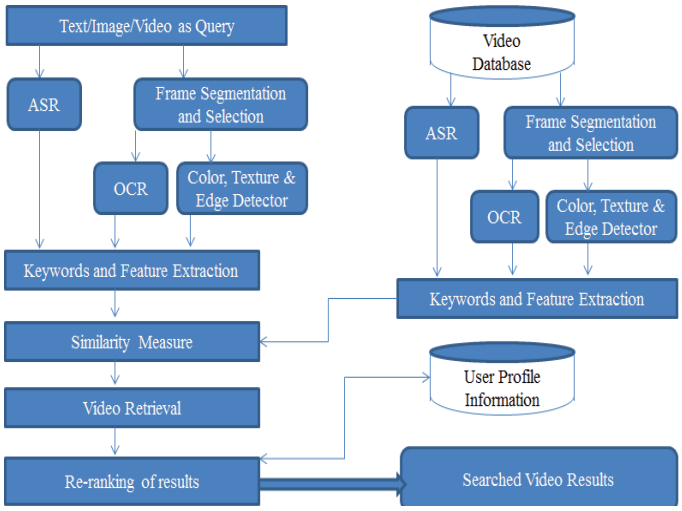

Fig. 2. Proposed Video Retrieval System

A. *Frame Segmentation and selection Technique*

If the video contains structure, i.e. several shots, then the standard techniques for video summarization involve:

1. Calculate the video Length
2. Divide the frame by particular time slot.

- However, let us assume you wish to find an interesting frame in a single continuous stream of frames taken from a single camera source. I.e. a shot.
- A mean colour histogram is computed for all frames and the key-frame is that with the closest histogram i.e. system selects the best frame in terms of its colour distribution.
- System assumes that camera stillness is an indicator of frame importance. As suggested by Beds, above. Then pick the still frames using optic-flow and use that.
- Each frame is projected into some high dimensional content space; system find those frames at the corners of the space and use them to represent the video. Frames are

evaluated for importance using their length and novelty in content space.

B. Optical Character Recognition(OCR)

Optical character recognition (OCR) is an important research area in pattern recognition. The objective of an OCR system is to recognize alphabetic letters, numbers, or other characters, which are in the form of digital images, without any human intervention. This is accomplished by searching a match between the features extracted from the given characters image and the library of image models. Ideally, they would like the features to be distinct for different character images so that the computer can extract the correct model from the library without any confusion. At the same time, the features should be robust enough so that they will not be affected by viewing transformations, noises, resolution variations and other factors. Figure 3 illustrates the basic processes of an OCR system.

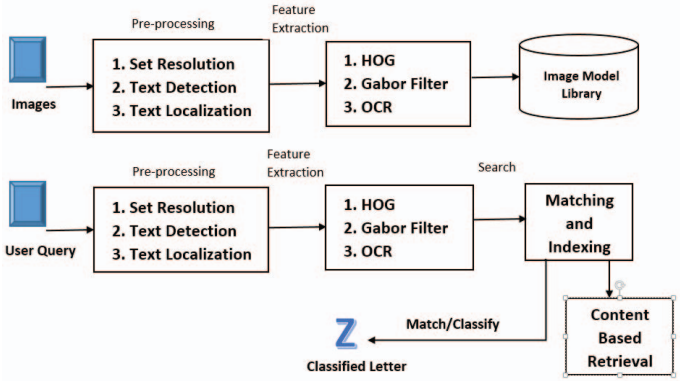

Fig. 3. Proposed OCR Architecture

OCR system propose end to end system for video text detection and recognition. OCR follows the following steps for implementation of text detection and recognition.

1. Text Localization
   - 1.1. Text Candidates using MSER
   - 1.2. Feature Extraction
     - 1.2.1. HOG features

$$G(x, y, \lambda, \phi, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{\frac{1}{2}\left(\frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2}\right)\right\} \exp\left\{i\frac{2\pi R_1}{\lambda}\right\}$$

     - 1.2.2. Gabor Filter features
   - 1.3. Dimensionality reduction using PLS
   - 1.4. SVM Classifier
2. Grouping of Localized Text Regions
3. OCR Decoding
4. OCR Engine

C. Automatic Speech Recognition

"Computer speech recognition", or just "speech to text" (STT). Some SR systems use "speaker-independent speech recognition" while others use "training" where an individual speaker reads sections of text into the ASR system. These systems analyse the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription [4].

Systems that do not use training are called "speaker-independent" systems. Systems that use training are called "speaker-dependent" systems. *Automated Speech Recognition (ASR)* In computer science and electrical engineering, Speech Recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), Speech recognition applications include voice user interfaces such as voice dialling (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), demotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input). The term voice Recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process.

D. Personalization of the Results

Re-ranking of video results with user interest profile by personalizing the result. User Interest model contains the user detail; user searched previous video URL link, related keywords and score of that video for that particular user.

IV. Results and Analysis

Video database of proposed CBVR system contains 25 videos and admin can upgrade the database any time. User sends the video queries and the system provides 9-10 re-ranked results from the database. The search result of the query video using only OCR is shown in table III and the related precision-recall graph is shown in figure 4. User can also search the video using only ASR and the corresponding results are shown in table IV. Related precision-recall graph is shown in figure 5. Finally, the proposed system can be search for the query video by integrated the OCR and ASR, and the related results are shown in table V. The precision-recall graph of integrated results are shown in figure 6. It is observe that the proposed system provides efficient and relevant results.

RECALL & PRECISION

Precision: The actual retrieval set may not perfectly match the set of relevant records.
Recall: The ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

Precision and Recall is mostly used in Information Retrieval domains such as Search Engines.

A simple Formula for Precision and Recall:

Assume the following, database contains 15 video records on each particular topic. The search was conducted on that topic and 12 records were retrieved of the 12 records retrieved, 9 were relevant.

The Precision and Recall could be calculated as follows:
Using the designations below:
A = the number of relevant records retrieved,
B = the number of relevant records not retrieved, and
C = the number of irrelevant records retrieved.

Table III.

OCR PRECISION RECALL EVALUATION

Video / Metric	Video 1	Video 2	Video 3	Video 4	Video 5
Precision	0.75	0.76	0.69	0.66	0.76
Recall	0.6	0.66	0.6	0.57	0.66
F-Measure	0.675	0.72	0.645	0.615	0.71

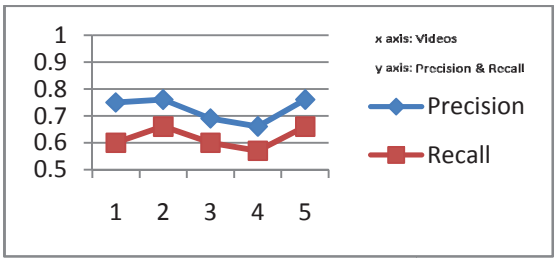

Fig. 4 OCR Precision Recall Graph

According to the table III, the OCR system gives the results to the video queries like different videos of presentation and figure 4 shows the OCR Precision Recall graph. User also requires less time to find the personal relevant snippet in search result. This is happened because this system keeps track of user interests while browsing & put the user interested snippets at the top position.

Table IV. ASR PRECISION RECALL EVALUATION

Video / Metric	Video 1	Video 2	Video 3	Video 4	Video 5
Precision	0.69	0.66	0.63	0.76	0.63
Recall	0.6	0.57	0.54	0.66	0.53
F-Measure	0.645	0.615	0.585	0.71	0.58

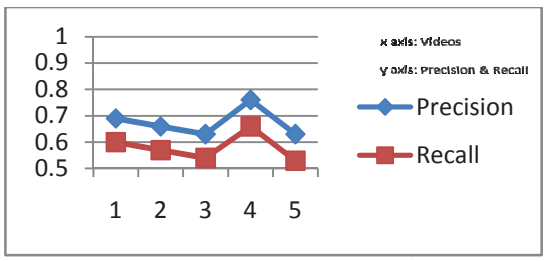

Fig. 5 ASR Precision Recall Graph

According to the table IV, the ASR system automatically converts the video to audio and at next it converts that audio to text keywords. The ASR system gives the results to the video queries like different videos of presentation and figure 5 shows the ASR Precision Recall graph. User also requires less time to find the personal relevant snippet in search result. This is happened because this system keeps track of user interests while browsing & put the user interested snippets at the top position.

Table V

COMBINATION OF OCR & ASR PRECISION RECALL EVALUATION

Video / Metric	Video 1	Video 2	Video 3	Video 4	Video 5
Precision	0.92	0.88	0.86	0.92	0.84
Recall	0.86	0.82	0.80	0.86	0.78
F- Measure	0.89	0.85	0.85	0.89	0.81

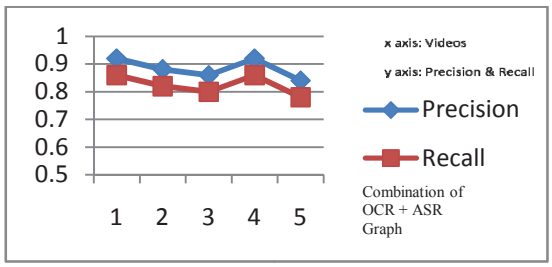

Fig. 6 Combination of the OCR & ASR Precision Recall Graph

Finally, form the Table V, it is conclude that, video retrieval with combination of OCR and ASR comparing with the separate system, our approach can do some improving by doing combination of OCR and ASR as the value of precision and recall of table V. For Personalized results video results sequence should change according to that user profile interest. After integrating the OCR and ASR methods, in the results automatically features and keywords will incremented and it will be benefit for searching the relevant and efficient results.

V. CONCLUSION

In this paper, we implement an approach for content-based video retrieval using combination of different approach in large video archives. In order to apply visual as well as audio resource of videos for extracting content-based metadata automatically. We propose an end-to-end text detection and recognition system as OCR and also applying ASR. The text detection component uses HOG based on rich shape descriptors such as HOG, Gabor and edge features for improved performance, and leverages PLS technique for dimensionality reduction, leading to SVM speed improvement. We proposed a merging scheme which overcomes the mistakes of SVM classification step and preserves word boundaries. Extensive evaluation on a large dataset illustrates the efficacy of our approach in both pixel-level text detection and word recognition tasks. At last, we integrating all the features extracted by OCR, ASR, Metadata and search each keywords with all the features and keywords extracted by OCR, ASR and metadata. At last display the personalized results to produce efficient and relevant result to users as per there interest.

ACKNOWLEDGMENT

The authors thank Haojin Yang and Christoph Meinel, Arpit Jain and Pradeep Natarajan. The research was supported by the information technology department of Sinhgad College of Engineering, Pune.

REFERENCES

[1]. Chivadshetti Pradeep, Mr Kishor Sadafale, and Mrs Kalpana Thakare.. "Content Based Video Retrieval Using Integrated Feature Extraction" in IPGCON 2015 at AVCOE, Sangmner on March 2015.

[2]. Thakre, Kalpana S., Archana M. Rajurkar, and Ramchandra Manthalkar. "An effective CBVR system based on motion, quantized color and edge density features." In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, pp. 145-149. ACM, 2010.

[3]. Thakre, Kalpana S., R. Manthalkar, A. M. Rajurkar, and Deepa Deshapande. "Video retrieval using singilar value decomposition and latent semantic indexing." In *Communication, Information & Computing Technology (ICCICT), 2012 International Conference on*, pp. 1-5. IEEE, 2012.

[4]. Haojin Yang and Christoph Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information", *in proc. IEEE transactions On Learning Technologies, Vol. 7, No. 2*, April-June 2014.

[5]. Arpit Jain‡, Xujun Peng, Xiaodan Zhuang, Pradeep Natarajan, and Huaigu Cao, "Text Detection and Recognition in Natural Scenes and Consumer Videos", *in proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) INSPEC Accession Number: 14448982*, 4-9 May 2014.

[6]. Hadi Yarmohammadi and Mohammad Rahmati, "Content Based Video Retrieval using Information Theory", *in proc. IEEE Iran Conf. ,Machine vision and Image Processing, pp. 214-218*, 2013.

[7]. B. V. Patel and A.V. Deorankar, "Content Based Video Retrieval using Entropy, Edge Detection, Black and White colour Features", *in proc. IEEE Computer Engineering and Technology (ICCET), 2nd International Conference on Vol. No. 6 Page(s): 272 –276,* 2010.

[8]. Kuo, T.C.T.; Dept. of Comput. Sci., Nat. TsingHua Univ. and Hsinchu, Taiwan ; Chen, A.L.P., "A Content-Based Query Language for Video Databases", *in proc. Third IEEE International Conference on Multimedia Computing and Systems, Page(s): 209 – 214*, 17-23 Jun 1996.

[9]. Volkmer, T. andNatsev, A., "Exploring Automatic Query Refinement For Text-Based Video Retrieval", *in proc. IEEE Multimedia and Expo*, *DOI: 10.1109/ICME.2006.262951, Page(s): 765 - 768*, 2006.

[10]. Jyothi, B., Latha, Y.M. and Reddy, V.S.K., "Relevance Feed Back Content Based Image Retrieval Using Multiple Features", *in proc. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705883, Page(s): 1 – 5*, 2010

[11]. Padmakala, S., Anandha Mala, G.S. and Shalini, M, "An Effective Content Based Video Retrieval Utilizing Texture, Color, and Optimal Key frame Features", *in IEEE International Conference on Image Information Processing (ICIIP), DOI: 10.1109/ICIIP.2011.6108864, Page(s): 1 – 6*, 2011

[12]. Kale, A. and Wakde, D.G., "Video Retrieval Using Automatically Extracted Audio", *in proc. IEEE International Conference on Cloud &* Databases", *in proc. Third IEEE International Conference on Multimedia Computing and Systems, Page(s): 209 – 214*, 17-23 Jun 1996.

[13]. B. V. Patel and B. B. Meshram., "Content Based Video Retrieval Systems*", in proc. IJU Vol.3, No.2*, April 2012.

[14]. Stephen W. Smoliar and Hongliang Zhang., "Content-Based Video Indexing and Retrieval", *in proc. IEEE 1070-986X/94*, Page(s): 62-72, 1994.