International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

News Story Retrieval Based on Textual Query

Namarata Dave, Mehfuza S. Holia

Abstract: ***This paper presents news video retrieval using text query for Gujarati language news videos. Due to the fact that Broadcasted Video in India is lacking in metadata information such as closed captioning, transcriptions etc., retrieval of videos based on text data is trivial task for most of the Indian language video. To retrieve specific story based on text query in regional language is the key idea behind our approach. Broadcast video is segmented to get shots representing small news stories. To represent each shot efficiently, key frame extraction using singular value decomposition and rank of matrix is proposed. Text is extracted from keyframes for further indexing data. Next task is to process text using natural language processing steps like tokenization, punctuation and extra symbols removal as well as stemming of words to root words etc. Due to unavailability of stemming and other methods of preprocessing of text in Guajarati language, we have given basic stemming technique to reduce dictionary size for efficient indexing of text data. With proposed system 82.5 percent accuracy is achieved on Gujarati news video dataset ETV.***

Keywords: Key frame extraction, Guajarati OCR, stemming, video retrieval, text query

I. INTRODUCTION

In this era of Digital information, it is required to have intelligent analysis of digital information present in multimedia data. Active research work has been carried out in the field of digital information processing for multimedia data. Video management and processing is having wide applications in areas such as detection of shot boundary, summarization of video, large scale retrieval and indexing of videos of different domains etc., [1,2]. Video clip or shots can be considered as a document for indexing and retrieval task. So, indexing of video clips is much similar to what we do while indexing documents or images. Generally, features of images, objects of images or content of document are used for indexing the data. To create similar index for searching precise content from video quickly, we need to parse a video document and divide it into scenes and further scenes can be divided to short stories to create index. The video can be categorized in entertainment, sports, news, multimedia messages, tutorials, lectures, e-learning videos, etc. Almost all type of videos consists of information such as text, objects, shapes, textures [3, 23, 29, 30], etc. which can serve as low level features for retrieval systems. To access any information from digital video involves indexing, retrieval, querying and browsing with help of user. To do this, we require automated methods to extract content of video. In the area of content-based video retrieval task, mostly spatial domain features as well as time domain features are used to represent the content of video. Features vectors in spatial domain are calculated from various blocks of frame. Association between these features vectors is encoded as a descriptor. The time domain feature vectors of video are calculated generally by dividing video sequence into segments like shots, scenes, frames etc. Once shot boundary is decided, next task is to extract features like texture, histograms, moments, motion vectors, text, etc. [25, 26, 27] to represent shots. We have proposed text-based approach by providing query as text and searching through the dataset of videos based on textual content available as scene text or frame text. In general, Video is comprised of multiple stories or scenes as show in Figure 1. Each scene is further divided into shots. Shots are made up of multiple frames containing almost similar features. There is lot of redundant data in consecutive frames of each video shot. We can take advantage of this idea of redundancy to process video to minimize time taken to search through video. To represent each shot, one key frame from collection of sequential frames is selected.

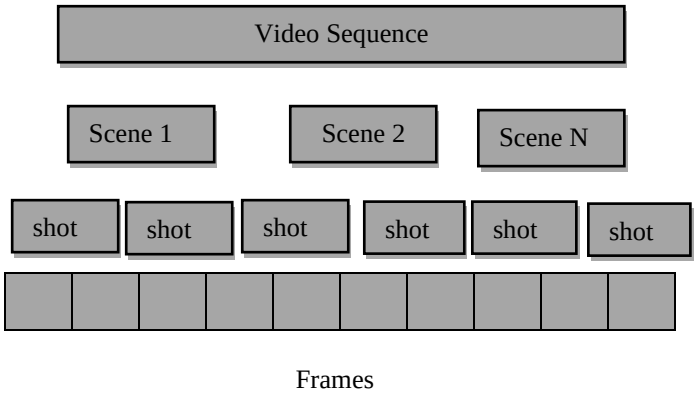

Figure 1 Video Structure

The approach of text query-based video retrieval is mainly divided in three submodules. First, we have implemented efficient Key Frame Extraction method by selecting accurate shot boundary. Second, we extracted text data using OCR on key frames extracted. Third and most important module is preprocessing of raw text extracted, indexing and retrieval based on the text. Our key contributions are Key Frame Extraction method and Stemming of Gujarati Text and retrieval based on it. Our paper is divided into four sections. First section is about literature referred so far as well as give details of dataset used in our experiment, third section describes proposed framework. Fourth section of the paper discusses results obtained so far using our dataset.

Revised Manuscript Received on February 05, 2020.
* Correspondence Author
Namrata Dave*, Computer Engineering Department, Gujarat Technological University, India. Email: namrata.dave@gmail.com
Dr. Mehfuza S. Holia, Assistant Professor, Electronics Department, Birla Vishwakarma Mahavidyalaya, India. Email: msholia@bvmengineering.ac.in

II. LITERATURE REVIEW

The system of retrieval of video based on content is basically divided into three main tasks. First, Shot boundary detection and key frame extraction. As shown in Figure 1, we can find a representative frame for each shot to simplify processing and indexing of video data. Second task is to extract features from key frames extracted and indexing it. Last task is to implement searching and retrieval based on query type.

We can simply define shot as collection of frames recorded in single camera action. During recording of shot, the camera device may be stationary or in motion. Examples of camera motions are zooming, panning, tracking, tilting, etc. There exist a lot of content similarities between frames in a shot of the scene. Shot is considered as a basic unit of video scene.

Extraction of shot from scene needs the appropriate procedure to determine the dissimilarity between successive frames of video. Also, it is important to determine a threshold to correctly define a shot boundary [4, 5, 6]. Shot changes can be categorized into sudden changes and slow changes [7]. It is easy to detect Sudden or Abrupt change in successive frames of video. On the other hand, it is comparatively difficult to determine the gradual change in video sequence. So, one should be very careful while choosing the value of threshold as it may sometime results in false boundary if not chosen properly. There are sorts of features available for detecting shot boundary. Among available options, color features are robust in situations like complex background, orientations issues, image size variations, etc. [8, 9]. Cut detection in case of sudden changes is easier than gradual transition detection in videos.[10]

Methods studied for shot boundary detection from literature so far mostly extracts visual features from each of the frame of video clip. Next step is to seek out frame similarities using the extracted features vectors. supported the similarity measures shot boundaries are often detected between frames which aren't similar. The Euclidean distance, the histogram intersection, etc. are often used for locating similarity for extracted feature vectors of frames. [11-14, 29, 30].

Many researchers have combined audio content along with visual features to improve accuracy of shot detection [16]. The audio which is synchronized with visual part is can very useful in determining shot boundary. Approach based on perceptual pause is highly related to segmentation of temporal audio at different levels of semantic [9, 15]. Also finding local correlation minima is used for segmentation of audio information from given video clip [31]. Methods based on multiple features such as anchor-person shot, caption information, voice feature and silence to segmentation of news stories from news video [32, 34].

In the key frame-based approach, shot similarities can be identified using key frames. Sometimes more than one frame can serve as key frame. Key frame is a frame chosen to represent content of entire shot. Shots of the scene are related by frame similarities [28].

Features can be visual, audio or text. Extracted features from key frames can be indexed for faster retrieval. To reduce the dimensionality of index, many recent literatures mentions used of LSH locality sensitive hashing technique for faster retrieval [17].

Query can be divided into many categories. In the example-based query, features of key frame extracted from example video or image are matched with the indexed features of key frames from database [18, 22]. In sketch-based query approach, users are required to draw sketches to find the interested videos from the database. Retrieval of desired videos is done by matching features of stored video and features generated from the user sketch. In this method trajectories of sketch given as user query are matched to trajectories of videos of database [21, 22]. Query by objects is also one of the popular ways of searching the video. In this approach, user inputs object's picture or image as query [33]. In Query by objects method, the system searches the collection of videos based on similarity measures and list out all occurrences of the object of interest from the video collection [20]. Generally, the results of searching with object as query approach are the positions of the object of interest from the video database [22].

Keyword based query approach represent the user's query using keywords. This approach is very easy and simplest one. In this approach, semantics of videos can be extracted to certain extent. Keywords can refer to transcripts of video, visual concepts, video metadata, etc [22]. Textual query for video retrieval approach is based on the text content of the video. Text content can be extracted from frames of video. Annotation of video using the textual content can be done for better retrieval. Recent approaches of text-based video retrieval could localize the text present in frames of video followed by recognition of the text. Finally, it measures the similarity between the query text and indexed feature vectors of database of the video. To get better results, this approach largely depends availability of efficient ocr algorithm for accurately recognizing the text from frames of video. Also, the quality of the input video matters a lot to get successful retrieval of video [13, 29]. Query by video clip is also popular approach now a days [18].

A. Dataset

Our main objective is to work on Guajarati language video retrieval. As the dataset for Gujarati language videos is not available to work on, we have created our own dataset for three Gujarati language news video channels named DD 11, ETV Gujarati and Sandesh. We have collected mpeg format recording for 48 hours for all three news channels which is total 70 GB of data by summing up all the videos. The resolution of original video is different for all three channels. Also, the place of overlay text which is displayed on screen is different in each channel. There is no transcript or any metadata information is available for any news video. Width for video is varying i.e. 490/480/450, height is 360, and Frame Rate is 25fps, video Format: 'RGB24'.

III. PROPOSED METHODOLOGY

We have proposed text query-based video retrieval from Guajarati language news video dataset. We cannot apply models used for text query-based retrieval of English and other foreign languages as we don't have metadata in form of text available with our Gujarati dataset that is the challenging part of dataset that we are using.

As show in figure2, our proposed approach takes video data as input. Video is then processed to work on frames to extract key frames representing each shot [19]. Features in form of text is extracted from key frames and indexed. The searching is done using tf-idf method where the query is text and retrieved results will be text documents index which in turn is mapped to the video clip in our original video dataset. Our main focus is on two algorithms, first is efficient key frame extraction method and second is Guajarati language stemmer to reduce size of dictionary for faster retrieval.

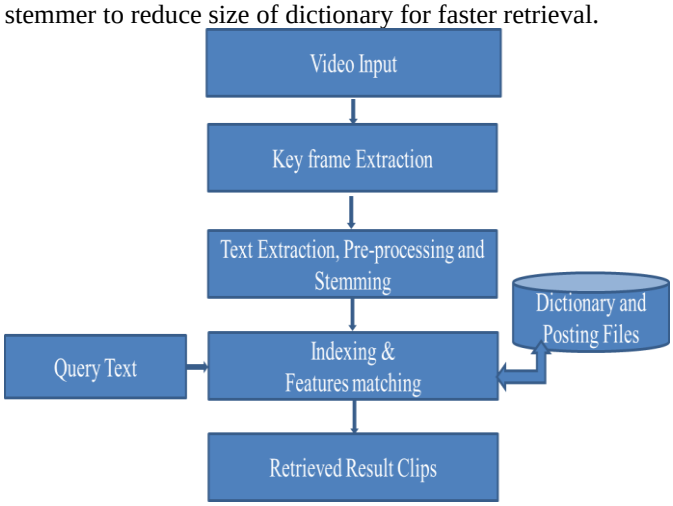

Figure 2 Proposed Text query-based Video Retrieval Approach

Key frame extraction is the key part before feature extraction in the retrieval system. To extract key frame, we have used rank-based matrix normalization approach using singular value decomposition. We can compute every matrix as product of matrices as show in figure 1 using the concept of singular value decomposition in linear algebra.

$$M = A\ D\ V^T \quad (1)$$

Here matrix M is calculated by taking product of the matrices A, D and transpose of V. matrices A and V are orthogonal matrices of size p x p and n x n, the middle matrix is diagonal matrix D with dimension p x n. Generally diagonal elements are ordered to satisfy the criteria $D_{ii} \geq D_{jj}$ for all $i < j$. we have taken matrix transpose V^T. If we consider only the n largest singular values of a matrix, then re-computing $U\Sigma'V^T$ gives good n-rank approximation to the matrix. It is important to understand that if we take the total of the first s singular values, further divide it by the summation of all the singular values. The result of this process gives the percentage of information that those singular values contain. If we want to retain x % of the information from any matrix or image, we can do so by calculating sums of singular values till we reach x percent of the total obtained from summation. Rest of the singular values can be removed which ultimately compress the data in given domain.

We have applied this linear algebra factorization approach to set of video frames which is represented as matrices internally. We applied factorization and ranking iteratively to find unique frames out of the set of frames. The matrix rank determines the total linearly independent rows which in fact is also true for columns. To find rank of any diagonal matrix one need to find the nonzero diagonal elements count. These diagonal entries are termed as singular values which are used to determine rank of given matrix. By comparing the rank with redefined threshold, we can separate out non-similar frames. This ultimately gives us sharp cuts in determining shot boundary of input video.

Text is displayed on each frame of video in form of overlay text. This text data is extracted using optical character recognition after localizing text in the frame correctly. We have extracted text from each frame which is taken as feature to represent the frame. Preprocessing is required to clean text from extra symbols which is tokenized further in words.

Porter stemmer, Lovins, Dawson [35] are some of the popular stemming algorithm for English language. Since for Gujarati language no stemmer is available to use, we have developed our own stemmer. Main objective function use to find optimal split position to decide root word of the token. Eq.2 shows the function used to determine split position i which varies between min length 1 to max value L i.e. length of word.

$$S(split_i) = S(stem = w_{1,i}) * S(suffix = w_{i+1,L}) \quad (1)$$

The main task of stemmer is to remove suffixes from the words. Stemming is done on the collection of text documents we have generated for our dataset. Feature-space has been created and normalized by using the Term Frequency – Inverse Document Frequency method. It typically measures how important a term is. The main purpose of doing a search is to find out relevant documents matching the query. Since tf considers all terms equally important, thus, we can't only use term frequencies to calculate the weight of a term in the document. We need to weigh down the frequent terms while scaling up the rare ones to do find important information. Term which appears more frequently in documents is not important most of the time. So, we assign less weight to them. Logarithms help us to solve this problem.

$$\vec{v}_{d_n} = \left(tf(t_1, d_n), tf(t_2, d_n), \ldots, tf(t_n, d_n)\right) \quad (3)$$

Each document is represented as vector of the terms of vocabulary as shown in eq. 3. $tf(t_1, d_n)$ represents frequency of term t1 in document represented by dn.

$$f(x,t) = \begin{cases} 1 \text{ if } x == t \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

$$tf(t,d) = \sum_{x \in d} f(x,t) \quad (5)$$

The function in equation 4 and 5 simply returns the number of times term t found in document denoted by d.

$$idf(t) = log\ \frac{|D|}{1+|\{d:t \in d\}|} \quad (6)$$

Inverse document frequency of term t can be found using equation 6. In given equation denominator term simply means the number of documents where the term t is available. In the numerator we are taking cardinality of set of documents. Where we take log of the value computed which compresses the value further.

Term weight can be found using given eqn7

$$w(t,doc) = \begin{cases} 1 + log\left(t * f(t,doc)\right) \\ \quad if\ f(t,doc) > 0 \\ 0\ otherwise \end{cases} \quad (7)$$

Score can be computed by taking product of term weights of query term and document terms as given in equation 8.

Figure 3. Top 6 results retrieved on query text "સલમાન"

$$S(doc) = w(term, qu) * w(term, doc) \quad (8)$$

Finally, accumulate all scores for number of query n using formula given by equation 9.

$$SS = \sum_1^n S(doc) \quad (9)$$

We follow the mentioned procedure to generate index of all documents in our dataset. We have generated test set of queries to test the performance of system. Query set is used retrieve relevant documents from the dataset. To retrieve relevant documents from dataset similarity between test and indexed data is measured using Cosine similarity given by equation 10. For our system, a single query takes 10.5312714 micro seconds time for k=10. Once query is submitted, results are retrieved and ranking of retrieved results is done based on similarity with query vector.

$$sm(p, q) = \frac{\sum_{i=1}^{k} p_i q_i}{\sqrt{\sum_{i=1}^{k} p_i^2} \sqrt{\sum_{i=1}^{k} q_i^2}} \quad (10)$$

We have collection of 12228 documents indexed for one news channel ETV Gujarati. We have designed text query set of 10. Results of top six retrieved video clip for query term "સલમાન" is show in figure 3.

IV. RESULT ANALYSIS

In the experiments done on our dataset, Precision and Recall are used to evaluate performance of our system. We are retrieving video clips related the query text. For set of documents of size n, retrieved results which are relevant are taken as true positive and retrieved video clips which are not relevant to query are taken as false positive. Video clips which are relevant to query but not retrieved by system is called false negative. Our results are obtained using a query set of size ten. Our query set is designed with different length of query text. We have taken names and incidence as query text to retrieve query clips related to person as well as incident. The total dataset contains 12228 documents out of which ten most relevant documents are retrieved. Evaluation of system is done using precision and recall metric given by equation 7 and 8. Precision and Recall is calculated based on retrieved results as following for documents retrieved for each query. Maximum number of documents retrieved is ten

Precision = True Positive / True Positive + False Positive (11)

Recall = True Positive / True Positive + False Positive (12)

for each query. We have obtained these results on machine with i5 Intel processor with 3.3 GHz processing frequency, 8 GB RAM. Results of query set of size 10 are shown in table1.

Table 1 evaluation metric for query set 10 on ETV channel news dataset

	Actual		
Predicted		Positive	Negative
	Positive	True Positive 70	False Positive 18
	Negative	False Negative 14	True Negative 0

V. CONCLUSION AND FUTURE WORK

Main challenge for developing proposed system was to extract scene text and process it for efficient retrieval of news videos.as metadata like transcription, closed captions are unavailable with Gujarati news channel videos. Proposed approach using video scene text as feature representing frame content for searching in dataset is different than existing approaches and not explored specifically in Gujarati Language Video domain. Also, Natural Language Processing methods like stemming, removal of frequent words, etc. from raw text an important and challenging task as it is not much explored in Guajarati language for raw text. With our proposed approach we have achieved 82.5 percent accuracy on ETV news dataset. We are trying to improve the accuracy by preprocessing on reducing number of features to speed up searching. Also, there is a scope of multimodal retrieval to improve performance of retrieval.

REFERENCES

1. Yang, Haojin, and Christoph Meinel. "Content based lecture video retrieval using speech and video text information." IEEE Transactions on Learning Technologies 7.2 (2014): 142-154.
2. A. Araujo and B. Girod, "Large-Scale Video Retrieval Using Image Queries," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 6, pp. 1406-1420, June 2018.doi: 10.1109/TCSVT.2017.2667710
3. Liu, Ying, et al. "A survey of content-based image retrieval with high-level semantics." Pattern recognition 40.1 (2007): 262-282.
4. A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" Circuits and Systems for Video Technology, IEEE Transactions on, vol. 12, pp. 90- 105, 2002.
5. J. Yu and M. D. Srinath, "An efficient method for scene cut detection," Pattern Recognition Letters, vol. 22, pp. 1379-1391, 2001.
6. H. J. Zhang, "Content-based video browsing and retrieval," in Handbook of Internet and multimedia systems and applications, B. Furht, Ed.: CRC press LLC, 1999.
7. W.-Y. Ma and Z. Hong-Jiang, "Content-based Image Indexing and Retrieval," in Handbook of Multimedia Computing, B. Fuhrt, Ed. Boca Raton: CRC press, 1999.
8. N. Dimitrova, Y. Rui, and I. Sethi, "Media Content Management," in Design Management of Multimedia Information Systems: Opportunities Challenges, S. M. Rahman, Ed.: Idea group publishing, 2001.
9. S. Pfeiffer, "Pause concepts for audio segmentation at different semantic levels," presented at ACM International Conference on Multimedia, Ottawa, Canada, 2001.
10. Mas, Jordi, and Gabriel Fernandez. "Video shot boundary detection based on color histogram." Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, 2003.
11. Tjondronegoro, D. W.,Content-based Video Indexing for Sports Applications using Multi-modal approach", PhD Thesis,(Doctoral dissertation, Deakin University, Melbourne, Australia)., 2005.
12. C.G.M. Snoek and M. Worring. "A State-of-the-art Review on Multimodal Video Indexing", Proceedings of the 8th Annual Conference of the Advanced School for Computing and Imaging pages 194--202, Lochem, The Netherlands, 2002.
13. C V. Jawahar, BalaKrishna Chennupati, Balamanohar Paluri and Nataraj Jammalamadaka, "Video Retrieval Based on Textual Queries", in Proceedings of the Thirteenth International Conference on Advanced Computing and Communications, Coimbatore, December 2005.
14. O'Toole, C., A. Smeaton, N. Murphy and S. Marlow, "Evaluation of automatic shot boundary detection on a large video suite". In: 2nd U.K. Conference Image Retrieval: The Challenge f Image Retrieval, Feb. 25-26, Newcastle, U.K. , 1999.
15. Pfeiffer, Silvia, Rainer Lienhart, and Wolfgang Efflsberg. "Scene determination based on video and audio features." Multimedia Tools and Applications 15.1 (2001): 59-81.
16. N Dave, M Holia, "Content based Video Retrieval", Indian Journal Of Technology And Education (IJTE), special issue of ICRISET 2017, pp 156-160.
17. Chafik, Sanaa & Daoudi, Imane & Elouardi, Hamid & El Yacoubi, Mounim & Dorizzi, Bernadette. (2014). Locality sensitive hashing for content based image retrieval: A comparative experimental study. International Conference on Next Generation Networks and Services, NGNS. 10.1109/NGNS.2014.6990224.
18. Anil Jain, Aditya Vailaya, Wei Xiong., "Query by video Clip," In Proceedings of Fourteenth International Conference on Pattern Recognition, vol.1. pp: 909-911, 16- 20 Aug 1998.
19. Namrata Dave, Dr. Mehfuza Holia, "Shot Boundary Detection for Gujarati News Video", International Journal for Research in Applied Science and Engineering Technology (IJRASET) , Volume 6, Issue III, pp: 3477-3480,March 2018.
20. R. Kanagavalli, Dr. K. Duraiswamy, "Object Based Video Retrievals" International Journal of Communications and Engineering Volume 06– No.6, Issue: 01 March2012.
21. Chen, Tao, et al. "Sketch2photo: Internet image montage." ACM Transactions on Graphics (TOG) 28.5 (2009): 124.
22. Hu, Rui, and John Collomosse. "A performance evaluation of gradient field hog descriptor for sketch based image retrieval." Computer Vision and Image Understanding 117.7 (2013): 790-806.
23. Hu, Weiming, et al. "A survey on visual content-based video indexing and retrieval." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41.6 (2011): 797-819.
24. Günsel, Bilge, A. Müfit Ferman, and A. Murat Tekalp. "Temporal video segmentation using unsupervised clustering and semantic object tracking." J. Electronic Imaging 7.3 (1998): 592-604.
25. Ngo, Chong-Wah, Ting-Chuen Pong, and Hong-Jiang Zhang. "Motion analysis and segmentation through spatio-temporal slices processing." IEEE Transactions on Image Processing 12.3 (2003): 341-355.
26. Jun Yue, Zhenbo Li, Lu Liu, Zetian Fu, "Content-based image retrieval using color and texture fused features", Mathematical and Computer Modelling, Volume 54, Issues 3–4, August 2011, Pages 1121-1127.
27. Tarun Jain, C.V. Jawahar, "Compressed Domain Techniques to Support Information Retrieval Applications for Broadcast Videos", Proceedings of National Conference on Computer Vision Pattern Recognition Image Processing and Graphics (NCVPRIPG'08), pp.154-159, Jan 11-13, 2008, DA-IICT, Gandhinagar, India.
28. C. T. Dang, M. Kumar and H. Radha, "Key frame extraction from consumer videos using epitome," 2012 19th IEEE International Conference on Image Processing, Orlando, FL, 2012, pp. 93-96.
29. Mr. Sumit R. Dhobale, Prof. Akhilesh A. Tayade , " A survey on Text Retrieval from Video " , International Journal of Application or Innovation in Engineering & Management (IJAIEM) , Volume 3, Issue 11, November 2014 , pp. 079-085 , ISSN 2319 - 4847.
30. LYU, Rung Tsong Michael, and Chu Hong HOI. "A Study of Content-Based Video Classification, Indexing and Retrieval." (2002).
31. Sundaram, Hari, and Shih-Fu Chang. "Video scene segmentation using video and audio features." 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532). Vol. 2. IEEE, 2000.
32. Song, Yu, Wenhong Wang, and Fengjuan Guo. "News story segmentation based on audio-visual features fusion." Computer Science & Education, 2009. ICCSE'09. 4th International Conference on. IEEE, 2009.
33. Mohamadzadeh, Sajad, and Hassan Farsi. "Content Based Video Retrieval Based on Hdwt And Sparse Representation." Image Analysis & Stereology 35.2 (2016): 67-80.
34. Namrata Dave, Mehfuza Holia, "Content based Video Retrieval", Indian Journal of Technology and Education, Special issue of ICRISET 2017,155-160.
35. Ismailov, Alisher & Jalil, Masita & Abdullah, Zailani & Abd Rahim, Noor Hafhizah. (2016). A Comparative Study of Stemming Algorithms for use with the Uzbek Language. 10.1109/ICCOINS.2016.7783180.

AUTHORS PROFILE

Namrata Dave, is research scholar pursuing her PhD from Gujarat Technological University. She is working as Assistant Professor in Computer Engineering Department of G. H. Patel College of Engineering & Technology. She is having teaching experience of 14 years.

Dr. Mehfuza Holia, is working as Assistant Professor in Electronics Engineering Department of Birla Vishwakarma Mahavidyalaya Engineering College. She has completed her PhD from S.P. University, Gujarat in 2013. She is having more than 18 year of experience of teaching. She has received many awards and grants so far.