Content based Video Retrieval from Gujarati News Video

Ph.D. Synopsis

Submitted To

Gujarat Technological University

For the Degree

of

Doctor of Philosophy

in

Computer/ IT Engineering

By

Mrs. Namrata Ashokbhai Dave

Enrollment No: 159997107019

Supervisor:

Dr. Mehfuza S. Holia

Assistant Professor, Electronics Engineering Department,

Birla Vishwakarma Mahavidyala, V V Nagar, Anand.

(August 2020)

Table of Contents

Content based News Video Retrieval from Gujarati News Videos

1 Abstract

In recent times, Video Retrieval from vast collection of videos from web is in demand. Video contains information in various forms such as image, text and audio. To retrieve appropriate content quickly from vast collection of videos is a very challenging task for researchers working in this area. To retrieve video, a user can use text, image as well as small video clip as input query to the system. Most work found in literature is appropriate to videos with closed captions and meta data information in English language. The model applied for English or other languages do not perfectly fit with the data available for Indian News Videos for retrieval task. As opposed to other countries, the broadcasted news does not contain any kind of transcript, closed caption details of video or metadata for the video. Lack of availability of data to process regional language news video in India is the primary motivation of the proposed work.

Proposed work is divided in three key tasks. First is Key Frame Extraction from News Video. Second is to remove advertisement and extract features (text, image features) from videos available in dataset. Third task is indexing and faster retrieval of videos based on query (text/image). Two approaches have been proposed in the research work. First approach is text query-based Gujarati news video retrieval by extracting text from key frames. Second approach is image query-based video retrieval, which uses deep learning model for feature extraction. Text based video retrieval from news videos of Gujarati language has its own challenges as extraction and processing of Gujarati text data is to be done separately for retrieval of meaningful videos. Main objective of using text feature was to simplify searching interface for common man of local region who is not having skill or knowledge of searching news with English or other language. The experiments performed and comparisons done on the dataset created with gujarati news video have shown the effectiveness and preeminence of the proposed approaches for Gujarati language news video retrieval task.

2 Brief description on the state of the art of the research topic

In recent times, the area of content-based video retrieval is being explored in view of large collection of data. Many aspects of video retrieval task are still capturing interest of researchers in the field of computer vision. Retrieval of video from large collection of videos can be achieved using multiple modalities such as image features, text, speech, motion features etc.

The problems associated with automatic analysis of news telecasts are more severe in a country like India, where there are many national and regional language channels, besides English. H. Ghosh et al.,[1] presented a framework for multimodal analysis of multilingual news telecasts. They focused on a set of techniques for automatic indexing of the news stories based on keywords spotted in speech as well as on the visual features. English keywords are derived from RSS feed and converted to Indian language equivalents for detection in speech and on ticker texts. They Restricted the keyword list which resulted in improvement in indexing performance.

T. Jain et al., [2] explored techniques which can support IR applications on large scale video databases. Almost 25 times faster approximation of a widely used color representation is proposed by using information from the MPEG compressed domain. A scheme is presented for real time matching of videos in online video feeds. LSH based indexing is performed for handling efficient near neighbour queries and support fast clustering. A video clip can be searched in more than 100 hours of video data in few seconds.

In this research, two approaches are proposed. Literature reviewed is also given based on major tasks done in both approaches. Key Frame Extraction, Advertisement Detection and removal are common tasks in both approaches. Text extraction, Indexing and Retrieval is part of proposed approach I and Image Query based Retrieval approach using Deep learning is proposed approach II.

2.1 Key Frame Extraction

Video can be viewed as collection of meaningful scenes, shots and frames. Scene is further divided into shots. Shot is a collection of frames captured during single camera motion. Frame is the most basic unit of video to consider for processing. In Content based video retrieval system, first task is to find shot boundary and select key frame representing each shot uniquely. Keyframes are the representative frames which are used to provide a suitable abstraction and framework that will help for indexing, browsing and retrieval of video [1]. As

by selecting Key Frame, the task of processing video is reduced by large amount as all frames are not required to be processed in order to retrieve meaningful information. As video contains multiple frames with almost similar contents, only one or two representing frames are selected out of all frames comprising the shot.

Pixel based, block based, transform based, feature based and histogram-based techniques are mainly used for automatic shot boundary detection from raw video stream. Pixel-based methods approach uses pixel difference as key parameter for detecting shot. These techniques are highly sensitive to noise. Block Based techniques work on fundamentals of pixel processing but they operate on image at a time due to which these methods are faster compared to pixel processing techniques. Color histogram-based approaches are the most popular, but it does not specify any position of pixel. [3] are using entropy measure which is one of the texture descriptors.

Because of importance of keyframe extraction in video retrieval many researchers are working in this area. There are many approaches which are used for keyframe extraction. One of them is shot based structuring of video, in which shot is detected then each shot is represented by one or more frames.

Ni et al.,[3] proposed and analysed a nonparametric region-based active contour model for segmenting cluttered scenes. The proposed model is unsupervised and assumes pixel intensity is independently identically distributed. The proposed energy functional consists of a geometric regularization term that penalizes the length of the partition boundaries and a region-based image term that uses histograms of pixel intensity to distinguish different regions. Wasserstein distance is used to determine the dissimilarity between histograms. Rasheed et al., have proposed the color histogram-based method of UCF. This algorithm uses the color histogram for measuring the intersection similarity to extract key frames. [4]

Ji et al., 2019 proposed a Deep-learning Semantic-based Scene-segmentation model (called DeepSSS) that considers image captioning to segment a video into scenes semantically. The system performs compares colour histograms to get shot boundary detection followed by maximum-entropy-applied keyframe extraction. The task of semantic analysis is performed using image captioning from deep learning. DeepSSS approach considers low as well as high level features of videos to achieve a scene segmentation.[5]

Kannao et. al [7] proposed segmentation of TV news broadcast into semantically meaningful stories. Hybrid approach using conditional random fields (CRFs) is proposed for

news story segmentation. The story boundary detection problem is converted into a shot classification problem by classifying video shots into either of the four categories. Features introduced for the task are overlay text based semantic similarity and grid-wise edge orientation histogram. They have demonstrated Experimental results on approximately 50 hours of news videos and achieved good efficiency.

2.2 Advertisement Detection

Almgren et al. [8], analysed visual features of images to detect advertising images from scanned images of various magazines. The aim is to identify key features of advertising images and to apply them to real-world application. They employed convolutional neural networks to classify scanned images as either advertisements or non-advertisements (i.e., articles).

In the context of news video of Indian channels telecasted, several channel specific norms are not followed compared to well-known foreign English news channels and frequently news and advertisements have equivalent frequencies of event in Indian news recordings. Vyas et al. presented features for the task of automatically identifying as well as extracting of commercial blocks in Indian news videos telecasted. They used features like acoustic features MFCC bag of words (BoW) and overlaid text distribution from video shot on dataset made up of 54 hours of video from recorded with three English Indian news channels and obtained around 97%. F-measure [9]

Zhang et al.[10], proposed a novel method to fuse audio and visual features to detect commercial blocks. Contextual features for each shot are generated with time expansion from TV channels in China has shown good results.

Zafar et al. [11] proposed a system that works on TV broadcast/Online videos to automatically detect commercials. In their system they detect commercials with the information of shot-level. Video's data is divided into shots and classification is done in commercial class and non-commercial class using ANN and SVM. With their approach they are able to handle a variety of program types, unclear commercials, and give good precision and recall.

It is a difficult task to insert advertisements at suitable positions while making user experience of watching advertisement contented. Earlier approaches insert advertisements at the static positions and did not pay attention to the variation of scenes, which can reduce the appeal of videos. Y. Liang et al., proposed method to produce and embed textual advertisements for online videos automatically. They estimated the visual significance of the

main elements in the video frames via human face localization as well as detecting saliency features. They proposed algorithm to recognize the scene changes with the visual significance map, through which the system can find stable areas in distinct scenes for advertising. [12]

Li et al., [13] proposed an automatic commercial detection system for TV broadcasting. This system works at shot level and detects commercials in streaming videos, including TV broadcasting and online videos. It consists of two modules, the shot boundary detection module and the shot classification module. Then, they extract shot features with deep convolutional neural network, and train a support vector machine classifier to complete shot classification for TV programs.

2.3 Text extraction, Indexing and Retrieval

Retrieval of video can be done using variety of features like SIFT, SURF, Edge, Histogram[7][14][15], color features[15][16][5], texture features etc. image-based features as well as audio-based features[17][18][19][20]. Also, the retrieval task has been done by combining features.

Text in frames will exhibit many variations according to their properties such as Geometry (size, alignment, inter-character distance), Color (monochrome, polychrome), Motion (static, linear moment), Edge (text boundaries, strong edges), Compression, etc.

Short texts are present in many computer systems. Examples include social media messages, advertisement, Q&A websites, and an increasing number of other applications. They are characterized by little context words and a large vocabulary. As a consequence, traditional short text representations, such as TF and TF-IDF[21][22][23], have high dimensionality and are very sparse. The research field of word vectors has produced interesting word representations that are discriminative regarding semantics, which can be algebraically composed to create vector representations for paragraphs and documents. Pita et al., [22] proposed a novel representation method based on the PSO meta-heuristic. Results in a document classification task are competitive with TF-IDF and show significant improvement over Paragraph Vector, with the advantage of dense and compact document vector representation.

Kannao et al. presented a contrast enhancement based pre-processing stage for overlay text detection and a parameter free edge density-based scheme for efficient text band detection. They also proposed a novel approach for multiple text region tracking. Their adopted Tesseract OCR for the specific task of overlay text recognition using web news articles. The proposed

approach is tested and found superior on news videos acquired from three Indian English television news channels along with benchmark datasets. [24]

Chang et al.,[25] proposed text detection mechanism for street view images in their research. To deal with relatively complicated content of street views in urban areas, the proposed scheme consists of a Fully Convolutional Network employed to locate street signs and Region Proposal Network to extract text lines in the identified traffic/shop signs.

Sukhwani et al., [26] presented a method to generate frame level fine grained annotations for a given video clip. Access to the frame level fine grained annotations lead to rich, dense and meaningful semantic associations between the text and video and improves the accuracy of system. They demonstrated the use of probabilistic label consistent sparse coding and dictionary learning with a K-SVD algorithm to generate 'fine grained' annotations for a class of videos - lawn tennis on a publicly available tennis dataset.

A. Mishra et al., [27] proposed approach for text to image retrieval task using the text available in images. Query-driven search approach is used to approximately locate characters in the text query, and impose spatial constraints to produce a ranked collection of images. They have evaluated their approach on public scene text datasets, IIIT scene text retrieval, Sports-10K and TV series-1M datasets.

H. Karray et al., [28] proposed a framework for multimodal analysis of Arabic news broadcast which helps users of pervasive devices to browse quickly into news archive; their solution integrating many aspects such as summarizing, indexing textual content and on on-line recognition of the handwriting. Firstly, the summarizing process is to accelerate the video content browsing based on genetic algorithm. Secondly, the indexing process, which operates on video summaries based on text recognition.

2.3.1 Gujarati text processing

Extraction of features from the video frames is important part of retrieval process. As opposed to other visual features, text features from news video frame are not easy to extract as it is for Indian languages. Researchers are working on scene text extraction and recognition[29] for improving accuracy in text extraction from video frames[30][31].

Tesseract OCR[32] is one of the efficient text OCR used in text retrieval from document images as well as scene text from video[33][24][34]. The Gujarati text data retrieved is processed further using natural language processing techniques such as tokenization, removal

of extra symbols including punctuation marks, stemming of words wherever necessary to reduce dictionary size etc.

In Gujarati language processing major contribution is done by TDIL[35] program through government of India. There are no publicly available tool of Gujarati language stemming and processing for raw gujarati text. Researchers have worked on different stemming techniques for Gujarati language on either EMILLE[36] text corpus or their own dataset[37][38][39][40]. There is a lot of scope of research in this field for gujarati text processing. As Gujarati language standard benchmark datasets to work for scene text recognition or image/video retrieval are not available, it remains a challenging task to generate dataset as well as getting good performance with it.

2.4 Image Query based Retrieval approach using Deep learning

Image retrieval using deep learning with image query is widely explored approach so far [41][42]. Nowadays, researchers are paying attention on large scale retrieval of video or image using image as query [43][44]. Whenever it is required to deal with large collection of images or video frames, normal system cannot give good performance with state-of-the-art methods of retrieval. Due this fact, CBVR approaches using deep learning architectures are being explored by different researchers in various fields such as news videos, sports videos, movie videos etc.

Mühling et al.[44] proposed approaches using deep learning for effective video inspection and retrieval. They proposed efficient algorithms for media production as well as introduced components for novel visualization and achieved average precision of approximately 90% on the top-100 video shots using concept detection. They have used pre-trained CNN models based on visual recognition tasks.

Noh et al., [45] proposed DELE local feature descriptor for image retrieval task at large-scale. The new feature is based on convolutional neural networks, which are trained only with image-level annotations. They proposed an attention mechanism for key point selection, which shares most network layers with the descriptor. System produces reliable confidence scores to reject false positives-in particular, it is robust against queries that have no correct match in the Google-Landmarks dataset.

S. Lange et. al, [46] discusses the effectiveness of deep auto-encoder neural networks in visual reinforcement learning (RL) tasks. They have proposed a framework for combining the

training of deep auto-encoders (for learning compact feature spaces) with recently-proposed batch-mode RL algorithms (for learning policies). They have used synthesized and real images.

Y. Wang et al, [47] investigated the dimensionality reduction ability of auto-encoder. Their experiments were conducted both on the synthesized data for an intuitive understanding of the method, mainly on two and three-dimensional spaces for better visualization, and on some real datasets, including MNIST[48] and Olivetti face datasets.

3 Definition of the Problem

Content based video retrieval can be simply described as retrieval of relevant news clips based on "Gujarati" language text query or with image query from the Gujarati News Video Collection. Major challenge for research work is absence of metadata information such as transcripts or closed caption details for videos in dataset which is available normally with English language videos in US or other countries.

Main objective of using text feature was to simplify searching interface for common man of local region who is not having skill or knowledge of any other language other than mother tongue or local region language.

4 Objective and Scope of work

Searching content using text available on screen is the basic idea to propose the 'Gujarati' language text query-based video retrieval from Guajarati news channel video dataset. Text processing for Gujarati language is difficult due to lack of resource availability for the Gujarati language. Gujarati language text processing is still not fully explored by researchers and due to this reason, no benchmark dataset or efficient tools for OCR, ASR, no Stemming and Lemmatization tools are available to use for language processing.

Generation of tremendous amount of digital content on daily basis is a major source of inspiration to work on content-based retrieval from video data using regional language textual information as well as exploring the concepts of deep learning for faster retrieval of contents from dataset. Proposed work is focused on retrieval of news stories from the collection of news video dataset using text query or image query. Objectives of the thesis are as follows:

- To reduce the processing time of feature extraction from video data by selecting representative frames from shots of each video of the dataset.

- To propose efficient and cost-effective model for further reducing the amount of data to be processed by removing advertisement from the dataset of keyframes using transfer learning approach.
- Text Feature Extraction and processing from the dataset as well as indexing them for faster retrieval.
- Exploring efficient Unsupervised Deep Learning Architecture for image query-based video retrieval approach.

Scope of the research work is as follows:

The research work carried out on mainly videos collected from three Gujarati Language News Video Channels which is lacking meta data information such as transcript, closed caption details etc. required to process text-based retrieval task.

The second approach proposed to explore unsupervised deep learning architecture for retrieval of news story using image query as input.

5 Methodology of Research, Results / Comparisons

Research work carried out is mainly divided into two approaches. First approach is text query-based news story retrieval for Guajarati language text query and task is performed on the dataset collected from Guajarati language news channels. Second approach is retrieval of news story based on image query. Proposed work for Content based Video Retrieval from Gujarati News Video is show in figure 1.

The research work carried out is mainly focused on three aspects of video story retrieval that are, Key Frame Extraction with Advertisement Removal, Feature Extraction and Indexing documents for retrieval.

Experiments are performed on dataset created with continuous recording for two days of three gujarati language channels of ETV News, DD 11 and Sandesh News. Also, few video recordings from channel TV9 and DD Girnar are taken in dataset. Overall size of the generated dataset is 90 GB and the format used for processing is mp4. As shown in figure 1, features are extracted from the dataset and indexed to match with query features.

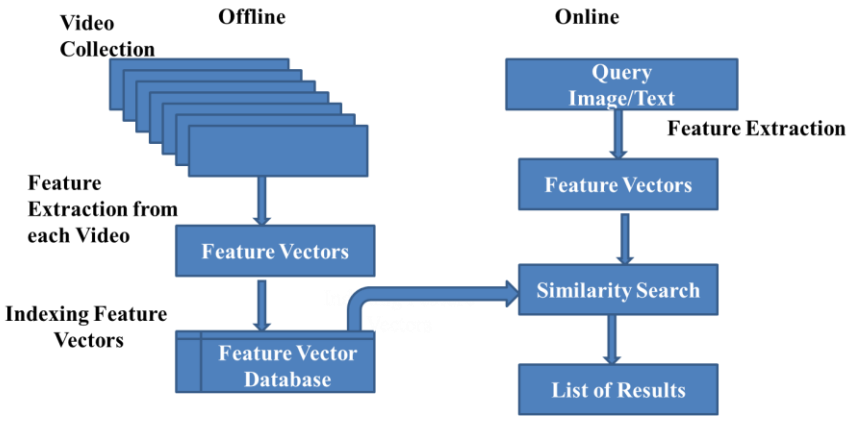

Figure 1 Block Diagram of Proposed System for Content based Video Retrieval from Gujarati News Video.

5.1 Key Frame Extraction and Advertisement Removal

Video is collection of frames which are very similar for particular shot of part of shot. One or more frames are chosen as key frame which represents the shot or part of shot. In this way, the processing time of video reduces by large amount. Also, in the recordings of news videos unnecessary information i.e. advertisements for retrieval task are found. To remove advertisements, advertisement classification approach is proposed to further reduce the processing time of video data.

5.1.1 Key Frame Extraction Algorithm

Main focus of extracting key frame is to distinguish the frames of one shot from another efficiently. Key Frame Extraction algorithm A1 is proposed based on histogram difference of video frames converted into Gray scale for finding key frames. As the histogram gives frequency of pixels for entire frame and less sensitive to motion in similar frames, it can be utilized to find major difference based on threshold between consecutive frames. With this characteristic of histogram in mind, proposed algorithm finds key frames based on difference of consecutive frames in Gray scale.

Key frame extraction approach A2 is based on edge information of frames. Differences in consecutive frames can be highlighted easily using edge information retrieved using various operators. Canny edge detector is applied to get edge information which extract key frames based on absolute difference between edges of consecutive frames.

Proposed Key Frame Extraction approach A3 is using HSV color model to find frame differences of successive frames of video. In the HSV color model (hue-saturation-value), hue used to represent pure colors. Saturation represents the measure of the degree where white color dilutes the pure color. HSV model is much closer to people's perception of color than RGB

color model, so HSV model is used in color histogram. In the proposed algorithm, separate histogram is created for h-plane, s-plane and v-plane of color image frame extracted from input video. All three histograms are combined to represent each frame. Next step is to find histogram of all consecutive frames. At a time, algorithm will process N fix number of frames to find key frames representing the shot.

Table 1: Details of Results obtained with Key Frame Extraction algorithm using approach 3 on three datasets of news channels ETV, DD11, Sandesh

Datasets	Total Hours (TH)	Total Frames (TF)	Key Frames (KF)	CR
ETVNG	30 hours	29,70,000	80,752	0.972
DD11NG	29 hours	26,10,000	17,832	**0.9931**
SANNG	31.5 hours	32,85,000	74,400	0.978

In proposed algorithm, Matrix Factorization is implemented as first step on Histogram vectors obtained from the three planes H, S and V of HSV color model from frames under consideration. Matrix Factorization will help in determining unique values from consecutive frames with help of rank of a matrix. The Singular Value Decomposition[49] theory shows that the linear transformation used is independent of scaling in each coordinate direction.

The rank of a matrix can be determined as the number of linearly independent rows, which is the same as the number of linearly independent columns. So, for the diagonal matrix rank can be calculated as the number of nonzero diagonals elements. Orthogonal transforms preserve linear independence. With low Ranks of matrix, non-similar frames can be separated out, which ultimately gives us sharp cuts in determining shot boundary of input video. Proposed algorithm achieves good performance in terms of compression ratio 0.972 for ETVNG dataset, 0.99 for DD11NG dataset and 0.978 for SANNG dataset using this algorithm. Comparison of three algorithms on three datasets are given in figure 2. It can be observed that approach with algorithm A3 performs better compared other two approaches for given dataset.

Figure 2 performance comparison of Key frame extraction algorithms

5.1.2 Transfer Learning Approach for advertisement detection

Deep learning using convolutional neural networks is emerged as the best approach for object detection. Performance of convolutional neural network depends on its architecture as well as large dataset for training. To train the model with small dataset and get most accurate results transfer learning approach is adopted.

Using learned weights of well-defined network Alexnet trained on very large dataset, very good results are achieved by adding new dataset in training. It has been experimented well on different models with different layers to generate final results. To classify images, support vector machine[13][50][51][52] classifier along with Bayesian optimizer[53][54] is applied to improve classification performance. Alexnet model[55] is used for training the dataset using transfer learning method. Dataset used is mainly divided into two classes advertisement and news.

Table 2 confusion matrix for advertisement detection

	ADV	NEWS	
ADV	83 32.7 %	0 0.0 %	100 % 0.0 %
NEWS	2 0.8 %	169 66.5 %	98.8 % 1.2 %
	97.6 % 2.4 %	100 % 0.0 %	**99.2** % 0.8 %

Confusion matrix for advertisement classification is given in table 2. Accuracy of 99.2 percent is achieved with proposed classification approach for 1268 key frames of different news channels such as DD Girnar, ETV Gujarati, tv9. We have divided dataset in 75:25 ration

for training and testing of system. Architecture of Alexnet model has 5 convolutional layers and 3 fully connected layers. Activation Relu is applied after every layer. Dropout is applied before the first and the second fully connected year.

5.2 Feature Extraction, Indexing and Retrieval for Text Query based Approach

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. Block Diagram of Proposed System for Content based Video Retrieval from Gujarati News video using text query is shown in figure 3. To retrieve specific story based on text query in regional language is the key idea behind the proposed 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. Key frames extracted are processed further for advertisement removal followed by text band localization with some morphological operations. Text as a feature extracted from keyframe and the extracted text features are used for indexing documents.

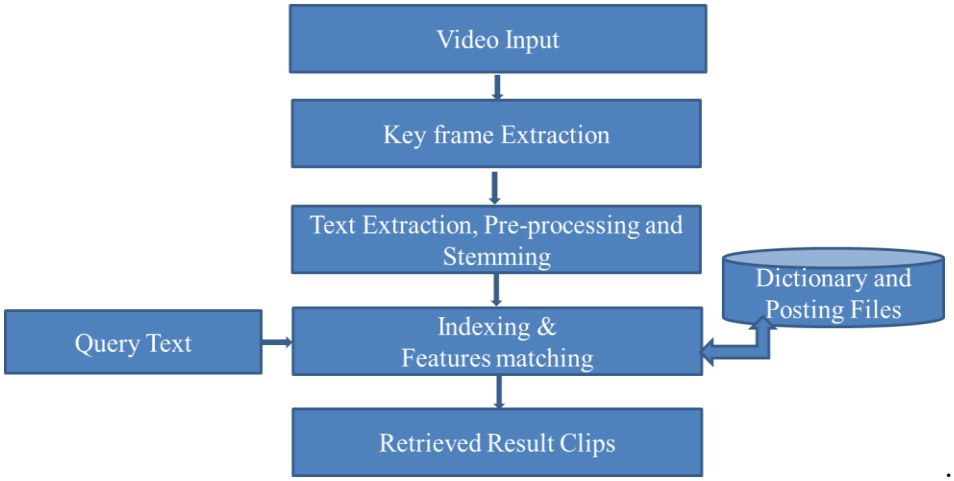

Figure 3 Block Diagram of Proposed System for Content based Video Retrieval from Gujarati News video using text query

In parallel to text extraction, another important task is to process document 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 pre-processing of text in Guajarati language, stemming technique for gujarati word is proposed to reduce dictionary size for efficient indexing of text data. To maintain records of words and position of words in each document dictionary file and posting files are created as shown in block diagram of figure 3.

A query set for the top-k image to video query consists of 20 queries. The query response time is employed to evaluate the performance of the system. The size of video dataset size is 90 hours; the number of the query words changes in size for different query. To retrieve top k relevant documents from dataset similarity between test and indexed data is measured using Cosine similarity. Performance of text query-based retrieval in terms of average precision is shown in figure 4.

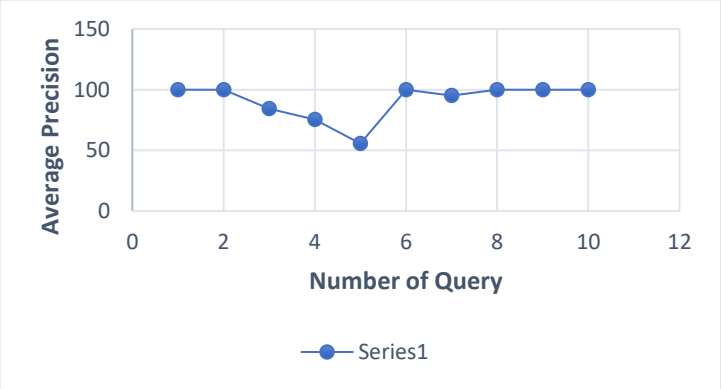

Figure 4. Performance of Text Query based video retrieval

In proposed system, a single query takes average 10.53 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. Indexing of approximately 1.5 lacs documents has been done for the news channels ETV Gujarati, DD 11 and Sandesh news out of which ten most pertinent archives are retrieved. Results of top six retrieved video clip for query term "સલમાન" is shown in figure 5.

Assessment of framework is finished utilizing precision and recall metric as well as mean average precision. Precision and Recall is calculated based on retrieved results for documents retrieved for each query. Mean average precision value obtained is 91.5. Maximum number of documents retrieved is ten for each query. Results are obtained on machine configured with i7 Intel processor having 3.3 GHz processing frequency, 8 GB RAM.

Figure 5 Top 6 results retrieved on query text "સલમાન"

Time taken for retrieval was 10.53 micro seconds time on average which is better than text-based video retrieval task done by V. Naik[56] et al. in which they claimed 1.55 seconds time for retrieval.

5.3 Image Query based retrieval using Deep Learning

Autoencoder is a mostly used unsupervised deep learning architecture for image compression, image reconstruction, image retrieval task, etc. In proposed approach with image query-based video retrieval, a model using autoencoder is used to extract image features in compact form and used it for matching with query and retrieval.

As shown in figure 6, key frames collected after removing advertisements are fed into the autoencoder architecture to train model. The autoencoder model used here is denoising convolutional autoencoder with three convolutional layers with different kernel size and three max pooling layers. The benefit of using denoising encoder is it do not just copy features while training and extract meaningful information which can be utilized to reconstruct image efficiently without much loss of information.

By defining the image to video query problem as a visual-aggregation problem, it is preferable to devise an indexing structure which is dependent on the inverted index rather than complex hybrid methods to solve this problem efficiently.

Activation function used here is rectified linear unit (Relu) and optimizer is stochastic gradient descent method with learning rate=0.01, momentum=0.9 in training. The model is

experimented well with different combination of optimizers, epochs and other hyper parameters to optimize performance. From the extensive experiments performed with encoder, it is observed that stochastic gradient descent optimizer yields better performance with epochs size more than 500 and batch size 32 while training encoder model.

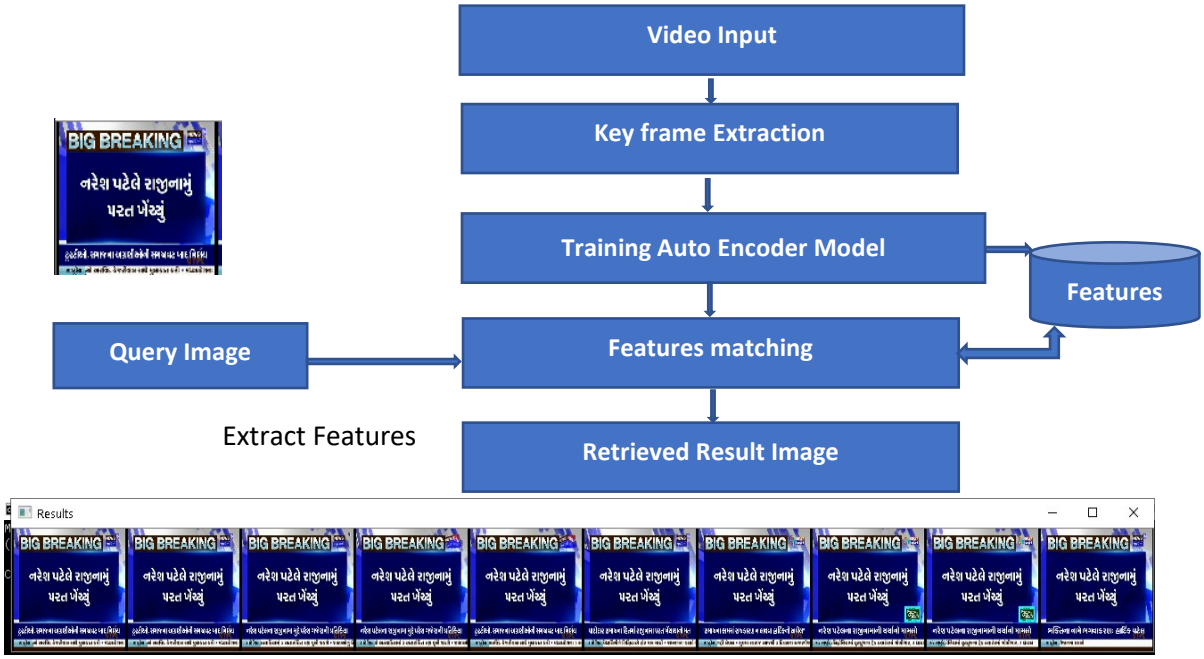

Figure 6. Image Query based video retrieval using autoencoder approach

Query image is processed in similar manner for feature extraction task as training frames of video. Similarity between query features and features stored in dataset is done for retrieving similar clips from dataset. Figure 7 shows average precision value obtained for query set of size 10. With the proposed approach good performance in terms of retrieval as 92.35 mean average precision has been achieved. Results are good compared to the state of art methods. Proposed system is tested on a machine configured with NVIDIA gpu titan Xp received as grant for research work.

Second approach using autoencoders gives 92.35 map which is better compared to work done by Zhang et al. with deep learning architecture CNN is used and map of 80 percent is achieved with large datasets of 500 hours of video.[10] Mean average precision (map) decreases with increase in size of video realized in our work with size of video increasing from 50 hours to 90 hours.

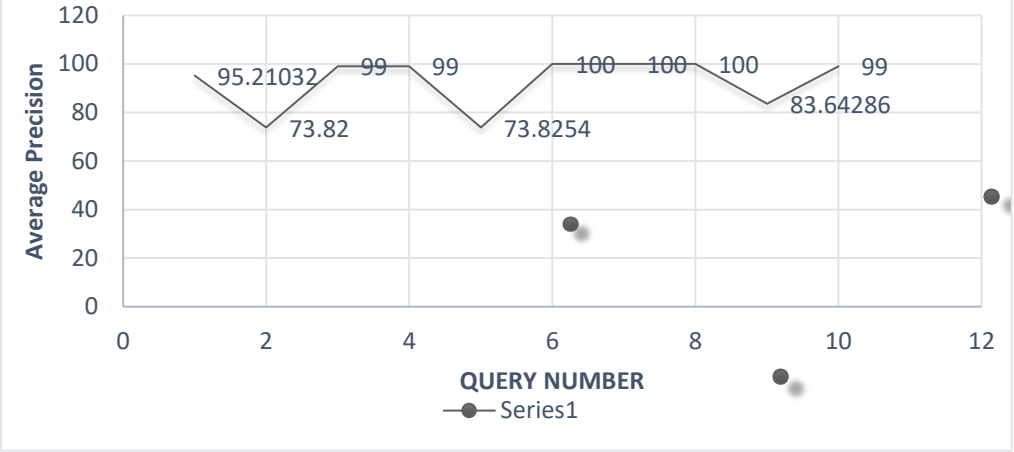

Figure 7 Performance of Image Query based video retrieval

6 Achievements with respect to objectives

Main objective of the proposed work is to implement efficient content-based news video retrieval (CBVR) system to handle Guajarati language news video without any metadata such as closed caption, transcriptions of video.

Algorithm for efficient key frame extraction is proposed followed by advertisement detection and removal using transfer learning approach to reduce the overall computation time for video processing in retrieval task.

Approaches to retrieve video clips from dataset of Gujarati language news videos are proposed. In the first approach, Gujarati language text query is used to retrieve relevant information in form of video clips from dataset. In this approach, Gujarati texts are extracted and used for matching and retrieval of relevant videos based on text query. In second approach of video retrieval, unsupervised learning based deep learning architecture using denoising convolutional auto encoders is proposed to retrieve features followed by matching algorithm using Euclidian distance metric.

Proposed approaches of key frame extraction, advertisement detection, processing of Gujarati text and searching for relevant information based on Gujarati language text, denoising convolutional autoencoder based unsupervised learning for image query-based retrieval gives good performance on the dataset.

7 Conclusion

Main challenge for developing the proposed system was to extract scene text and process it for efficient retrieval of news videos as metadata like transcription and closed captions are

unavailable with Gujarati news channel videos. Proposed approach using video scene text as a feature representing frame content for searching in a 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 are important and challenging task as they are hardly explored in Gujarati language for the raw text. With the proposed approach 91.5 map have been achieved.

Deep learning-based approach using denoising convolutional auto encoders are giving good performance compared to text-based approach. Second approach using autoencoders gives 92.35 map for the same dataset and query set used for approach 1. Better accuracy achieved at the cost of training model for the dataset.

Also, the advertisement classification model proposed with transfer learning approach outperforms state of the art work and yields 99.2 percent accuracy. Key frame extraction algorithm proposed also performs well with compression ratio of 0.98 on average for all three-news channel dataset.

8 Copies of papers published and a list of all publications arising from the thesis

1) N. Dave, M. Holia, "Content based video retrieval", Indian Journal of Technology and Education (IJTE) special Issue for ICRISET 2017, pp 155-160.
2) N. Dave, M. Holia, "Shot Boundary Detection for Gujarati News Video", International Journal for Research in Applied Science and Engineering Technology, 2018. 6. 3477-3480. doi 10.22214/ijraset.2018.3730. (UGC CARE LIST APRIL 2018)
3) N. Dave, M. Holia, "News Story Retrieval Based on Textual Query", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020. DOI:10.35940/ijeat.C5264.029320, https://www.scopus.com/sourceid/21100899502.
4) Namrata Dave, Mehfuza Holia, "Advertisement Detection Using Transfer Learning and Support Vector Machine", Selected in ICRISET2020 (International Conference on "Research and Innovations in Science, Engineering& Technology", September 2020.

Acknowledgements:

Thanks to NVIDIA for providing us GPU Titan Xp as a grant from to work for deep learning project proposed as approach2 here along with other projects supervised by Dr. Mehfuza S. Holia.

9 References

[1] H. Ghosh *et al.*, "Multimodal indexing of multilingual news video," *Int. J. Digit. Multimed. Broadcast.*, 2010, doi: 10.1155/2010/486487.

[2] C. V. J. Tarun Jain, "Compressed Domain Techniques to Support Information Retrieval Applications for Broadcast Videos," in *Proceedings of National Conference on Computer Vision Pattern Recognition Image Processing and Graphics (NCVPRIPG'08)*, 2008, pp. 154–159.

[3] P. M.Kamde, S. Shiravale, and S. P. Algur, "Entropy Supported Video Indexing for Content based Video Retrieval," *Int. J. Comput. Appl.*, vol. 62, no. 17, pp. 1–6, 2013, doi: 10.5120/10169-9974.

[4] K. Ni, X. Bresson, T. Chan, and S. Esedoglu, "Local histogram based segmentation using the wasserstein distance," *Int. J. Comput. Vis.*, vol. 84, no. 1, pp. 97–111, 2009, doi: 10.1007/s11263-009-0234-0.

[5] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, 2005, doi: 10.1109/TCSVT.2004.839993.

[6] H. Ji, D. Hooshyar, K. Kim, and H. Lim, "A semantic-based video scene segmentation using a deep neural network," *J. Inf. Sci.*, vol. 45, no. 6, pp. 833–844, 2019, doi: 10.1177/0165551518819964.

[7] R. Kannao and P. Guha, "Story segmentation in TV news broadcast," in *Proceedings - International Conference on Pattern Recognition*, 2016, doi: 10.1109/ICPR.2016.7900085.

[8] K. Almgren, M. Krishnan, F. Aljanobi, and J. Lee, "AD or Non-AD: A Deep Learning Approach to Detect Advertisements from Magazines," *Entropy*, vol. 20, no. 12, 2018, doi: 10.3390/e20120982.

[9] A. Vyas, R. Kannao, V. Bhargava, and P. Guha, "Commercial block detection in broadcast news videos," in *ACM International Conference Proceeding Series*, 2014, doi: 10.1145/2683483.2683546.

[10] B. Zhang, T. Li, P. Ding, and B. Xu, "TV commercial detection using audiovisual features and support vector machine," in *Proceedings - 2012 International Symposium on Instrumentation and Measurement, Sensor Network and Automation, IMSNA 2012*, 2012, doi: 10.1109/MSNA.2012.6324578.

[11] H. Zafar, U. Shabbir, and S. Muntaha, "ARTIFICIAL NEURAL NETWORK BASED ON APPROACH FOR COMMERCIAL DETECTION," *Int. J. Inf. Technol. Secur.*, 2019.

[12] Y. Liang, W. Liu, K. Liu, and H. Ma, "Automatic Generation of Textual Advertisement for Video Advertising," in *2018 IEEE 4th International Conference on Multimedia Big Data, BigMM 2018*, 2018, doi: 10.1109/BigMM.2018.8499465.

[13] M. Li, Y. Guo, and Y. Chen, "CNN-based commercial detection in TV broadcasting," in *ACM International Conference Proceeding Series*, 2017, doi: 10.1145/3171592.3171619.

[14] F. Garcia-Lamont, J. Cervantes, A. López, and L. Rodriguez, "Segmentation of images by color features: A survey," *Neurocomputing*, 2018, doi: 10.1016/j.neucom.2018.01.091.

[15] G. H. Liu and J. Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, 2013, doi: 10.1016/j.patcog.2012.06.001.

[16] P. Duygulu, M. Y. Chen, and A. Hauptmann, "Comparison and combination of two novel commercial detection methods," in *2004 IEEE International Conference on Multimedia and Expo (ICME)*, 2004, doi: 10.1109/icme.2004.1394454.

[17] D. Mistry and A. Banerjee, "Comparison of Feature Detection and Matching Approaches: SIFT and SURF," *GRD Journals- Glob. Res. Dev. J. Eng.*, 2017.

[18] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," *IEEE Trans. Multimed.*, 2010, doi: 10.1109/TMM.2010.2052027.

[19] J. A. Vanegas, J. Arevalo, and F. A. Gonzalez, "Unsupervised feature learning for content-based histopathology image retrieval," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, 2014, doi: 10.1109/CBMI.2014.6849815.

[20] H. Kandil and A. Atwan, "A Comparative Study between SIFT-Particle and SURF-Particle Video Tracking Algorithms," *Int. J. Signal Process. Image Process. Pattern Recognit.*, 2012.

[21] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," in *Procedia Engineering*, 2014, doi: 10.1016/j.proeng.2014.03.129.

[22] M. Pita and G. L. Pappa, "Strategies for short text representation in the word vector space," in *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, 2018, doi: 10.1109/BRACIS.2018.00053.

[23] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, 2016, doi: 10.1109/ICEEOT.2016.7754750.

[24] R. Kannao and P. Guha, "Overlay text extraction from TV news broadcast," in *12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015*, 2016, doi: 10.1109/INDICON.2015.7443440.

[25] P. W. Chang, G. X. Zeng, and P. C. Su, "Text Detection in Street View Images by Cascaded Convolutional Neural Networks," in *International Conference on Digital Signal Processing, DSP*, 2019, doi: 10.1109/ICDSP.2018.8631678.

[26] M. Sukhwani and C. V. Jawahar, "Frame level annotations for tennis videos," in *Proceedings - International Conference on Pattern Recognition*, 2016, doi: 10.1109/ICPR.2016.7899740.

[27] A. Mishra, K. Alahari, and C. V. Jawahar, "Image retrieval using textual cues," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, doi: 10.1109/ICCV.2013.378.

[28] H. Karray, M. Kherallah, M. Ben Halima, and A. M. Alimi, "An interactive device for quick Arabic news story browsing," *Int. J. Mob. Comput. Multimed. Commun.*, 2012, doi: 10.4018/jmcmc.2012100104.

[29] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained scene text and video text recognition for Arabic script," 2017, doi: 10.1109/asar.2017.8067754.

[30] A. Jindal, A. Tiwari, and H. Ghosh, "Efficient and language independent news story segmentation for telecast news videos," in *Proceedings - 2011 IEEE InternationalSymposium on Multimedia, ISM 2011*, 2011, doi: 10.1109/ISM.2011.81.

[31] A. Tiwari and H. Ghosh, "Ticker text extraction from Bangla news videos," in *Proceedings of the 2010 Annual IEEE India Conference: Green Energy, Computing and Communication, INDICON 2010*, 2010, doi: 10.1109/INDCON.2010.5712595.

[32] M. K. Audichya, "A Study to Recognize Printed Gujarati Characters Using Tesseract OCR," *Int. J. Res. Appl. Sci. Eng. Technol.*, 2017, doi: 10.22214/ijraset.2017.9219.

[33] R. Smith, "An overview of the tesseract OCR engine," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2007, doi: 10.1109/ICDAR.2007.4376991.

[34] M. Koistinen, K. Kettunen, and T. Pääkkönen, "Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing," *Proc. 21st Nord. Conf. Comput. Linguist.*, 2017.

[35] "http://tdil-dc.in/index.php?lang=en." .

[36] "https://www.lancaster.ac.uk/fass/projects/corpus/emille/." .

[37] J. Ameta, N. Joshi, and I. Mathur, "A Lightweight Stemmer for Gujarati," *Proc. 46th Annu. Natl. Conv. Comput. Soc. India.*, 2011.

[38] K. Suba, D. Jiandani, and P. Bhattacharyya, "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati," in *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, 2011.

[39] J. Sheth and B. Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language," in *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*, 2014, doi: 10.1109/ICICICT.2014.6781269.

[40] C. D and J. M. Patel, "Improving a Lightweight Stemmer for Gujarati Language," *Int. J. Inf. Sci. Tech.*, 2016, doi: 10.5121/ijist.2016.6214.

[41] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *ESANN 2011 - 19th European Symposium on Artificial Neural Networks*, 2011.

[42] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2014, doi: 10.1109/CVPRW.2014.131.

[43] A. Singhal, P. Sinha, and R. Pant, "Use of Deep Learning in Modern Recommendation System: A Summary of Recent Works," *Int. J. Comput. Appl.*, 2017, doi: 10.5120/ijca2017916055.

[44] M. Mühling *et al.*, "Deep learning for content-based video retrieval in film and television production," *Multimed. Tools Appl.*, 2017, doi: 10.1007/s11042-017-4962-9.

[45] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-Scale Image Retrieval with Attentive Deep Local Features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, doi: 10.1109/ICCV.2017.374.

[46] S. Lange and M. Riedmiller, "Deep auto-encoder neural networks in reinforcement learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2010, doi: 10.1109/IJCNN.2010.5596468.

[47] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, 2016, doi: 10.1016/j.neucom.2015.08.104.

[48] Y. LeCun and C. Cortes, "MNIST handwritten digit database," *AT&T Labs [Online]. Available http//yann. lecun. com/exdb/mnist*, 2010.

[49] K. Baker, "Singular value decomposition tutorial," *Ohio State Univ.*, 2005, doi: 10.1021/jo0008901.

[50] R. Kannao and P. Guha, "TV advertisement detection for news channels using Local Success Weighted SVM Ensemble," in *12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015*, 2016, doi: 10.1109/INDICON.2015.7443801.

[51] Z. Heng, M. Dipu, and K. H. Yap, "Hybrid Supervised Deep Learning for Ethnicity Classification using Face Images," in *Proceedings - IEEE International Symposium on Circuits and Systems*, 2018, doi: 10.1109/ISCAS.2018.8351370.

[52] "Tutorial on Support Vector Machine," *Appl. Comput. Math.*, 2016, doi: 10.11648/j.acm.s.2017060401.11.

[53] K. Yang, M. Emmerich, A. Deutz, and T. Bäck, "Multi-Objective Bayesian Global Optimization using expected hypervolume improvement gradient," *Swarm Evol. Comput.*, 2019, doi: 10.1016/j.swevo.2018.10.007.

[54] S. Ahn, A. Korattikara, and M. Welling, "Bayesian posterior sampling via stochastic gradient fisher scoring," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "2012 AlexNet," *Adv. Neural Inf. Process. Syst.*, 2012, doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[56] C. Zhang, Y. Lin, L. Zhu, A. Liu, Z. Zhang, and F. Huang, "CNN-VWII: An efficient approach for large-scale video retrieval by image queries," *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2019.03.015.