Content based Video Retrieval from Gujarati News Video

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer / IT Engineering

by

Namrata Ashokbhai Dave

159997107019

under supervision of

Dr. Mehfuza S. Holia

GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD

June – 2021

DECLARATION

I declare that the thesis entitled **Content based Video Retrieval from Gujarati News Video** submitted by me for the degree of Doctor of Philosophy is the record of research work carried out by me during the period from **May 2016** to **December 2020** under the supervision of **Dr. Mehfuza S. Holia** and this has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities, if noticed in the thesis.

Signature of the Research Scholar:  Date: 22-06-2021

Name of Research Scholar: **Namrata Ashokbhai Dave**

Place: **Anand**

CERTIFICATE

I certify that the work incorporated in the thesis **Content based Video Retrieval from Gujarati News Video** submitted by **Smt. Namrata Ashokbhai Dave** was carried out by the candidate under my supervision/guidance. To the best of my knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of Supervisor: MHH Date: 22-06-2021

Name of Supervisor: **Dr. Mehfuza S. Holia**

Place: **Anand**

Course-work Completion Certificate

This is to certify that **Mrs. Namrata A. Dave** enrolment no. **159997107019** is a PhD scholar enrolled for PhD program in the branch **Computer / IT Engineering** of Gujarat Technological University, Ahmedabad.

(Please tick the relevant option(s))

☐ He/She has been exempted from the course-work (successfully completed during M.Phil Course)

☐ He/She has been exempted from Research Methodology Course only (successfully completed during M.Phil Course)

☑ ~~He~~/She has successfully completed the PhD course work for the partial requirement for the award of PhD Degree. His/ Her performance in the course work is as follows-

Grade Obtained in Research Methodology (PH001)	Grade Obtained in Self Study Course (Core Subject) (PH002)
BB	**BB**

Supervisor's Sign

(**Dr. Mehfuza S. Holia**)

Originality Report Certificate

It is certified that PhD Thesis titled **Content based Video Retrieval from Gujarati News Video** by **Namrata A. Dave** has been examined by us. We undertake the following:

a. Thesis has significant new work / knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled / analysed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using **urkund software** (copy of originality report attached) and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e. permitted similarity index <10%).

Signature of the Research Scholar: Dave Date: 22-06-2021

Name of Research Scholar: **Namrata A. Dave**

Place: **Anand**

Signature of Supervisor: MHy Date: 22-06-2021

Name of Supervisor: **Dr. Mehfuza S. Holia**

Place: **Anand**

URKUND

Document Information

Analyzed document	thesis plag check.docx (D94403793)
Submitted	2/2/2021 12:55:00 PM
Submitted by	
Submitter email	namrata.dave@gmail.com
Similarity	4%
Analysis address	namrata.dave.gtuni@analysis.urkund.com

Sources included in the report

W	URL: https://arxiv.org/pdf/1211.4683 Fetched: 2/2/2021 12:55:00 PM	1
W	URL: https://tede.ufam.edu.br/bitstream/tede/7145/12/Disserta%C3%A7%C3%A3o_BashirZeimar ... Fetched: 12/18/2019 5:40:54 PM	1
W	URL: http://www.iaeme.com/MasterAdmin/uploadfolder/50120130404055/50120130404055.pdf Fetched: 2/2/2021 12:55:00 PM	1
W	URL: https://www.ijeat.org/wp-content/uploads/papers/v9i3/C5264029320.pdf Fetched: 2/2/2021 12:55:00 PM	13
W	URL: https://www.researchgate.net/publication/339557317_Feature_Extraction_using_Convol ... Fetched: 12/2/2020 11:29:04 AM	2
W	URL: https://www.sciencedirect.com/science/article/pii/S0957417409010495 Fetched: 2/2/2021 12:55:00 PM	1
W	URL: https://www.researchgate.net/publication/271936952_A_Content_Based_Video_Retrieval ... Fetched: 2/2/2021 12:55:00 PM	1

PhD THESIS Non-Exclusive License to

GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a PhD Research Scholar at GTU and in the interests of the facilitation of research at GTU and elsewhere, I, **Namrata Ashokbhai Dave** having (Enrollment No.) **159997107019** hereby grant a non-exclusive, royalty free and perpetual license to GTU on the following terms:

a) GTU is permitted to archive, reproduce and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as the "Work") anywhere in the world, for non-commercial purposes, in all forms of media;

b) GTU is permitted to authorize, sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);

c) GTU is authorized to submit the Work at any National / International Library, under the authority of their "Thesis Non-Exclusive License";

d) The Universal Copyright Notice (©) shall appear on all copies made under the authority of this license;

e) I undertake to submit my thesis, through my University, to any Library and Archives. Any abstract submitted with the thesis will be considered to form part of the thesis.

f) I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.

g) If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.

h) I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my University in this non-exclusive license.

i) I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this non- exclusive license.

j) I am aware of and agree to accept the conditions and regulations of PhD including all policy matters related to authorship and plagiarism.

Signature of the Research Scholar: ---------- Date: 22-06-2021

Name of Research Scholar: **Namrata A. Dave**

Place: **Anand**

Signature of Supervisor: ---------- Date: 22-06-2021

Name of Supervisor: **Dr. Mehfuza S. Holia**

Place: **Anand**

Seal:

Thesis Approval Form

The viva-voce of the PhD Thesis submitted by **Smt. Namrata Ashokbhai Dave (Enrollment No. 159997107019)** entitled **Content based Video Retrieval from Gujarati News Video** was conducted on **Tuesday, 22-06-2021** at Gujarat Technological University.

(Please tick any one of the following options)

☑ The performance of the candidate was satisfactory. We recommend that he/she be awarded the PhD degree.

☐ Any further modifications in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the Supervisor or request of Independent Research Scholar after which viva-voce can be re-conducted by the same panel again.

(briefly specify the modifications suggested by the panel)

☐ The performance of the candidate was unsatisfactory. We recommend that he/she should not be awarded the PhD degree.

(The panel must give justifications for rejecting the research work)

Supervisor: **Dr. Mehfuza S. Holia**

Dr. Sumantra DUTTA ROY

External Examiner 1 : **Dr. Sumantra Dutta Roy**

DR UMESH BHADADE

External Examiner 2 : **Dr. Umesh S. Bhadade**

ABSTRACT

Recently, Video Retrieval from the vast collection of videos from the web is in demand. Video contains information in various forms such as image, text, and audio. To retrieve appropriate content quickly from the vast collection of videos is a very challenging task for researchers working in this area. To retrieve video, the 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 the English language. The model applied for English or other languages do not perfectly adaptable with the data available for Indian News Videos for retrieval tasks. As opposed to other countries, the broadcasted news does not contain any kind of transcript, closed caption details of the video, or metadata for the video. The lack of availability of data to process regional language news videos in India is the primary motivation of the proposed work.

The proposed work is divided into three key tasks, first is Key Frame Extraction from News Video. The second is to remove advertisements and extract features (text, image features) from videos available in the dataset. The third task is indexing and faster retrieval of videos based on a query (text/image). Two approaches have been proposed in the research work presented in the thesis. The first approach is text query-based Gujarati news video retrieval by extracting text from the frame of the video. The second approach is image query-based video retrieval, which uses a deep learning model for feature extraction. Text based video retrieval from Gujarati language news videos has its challenges as extraction and processing of Gujarati text data is to be done separately for retrieval of meaningful videos. The main objective of using the text feature was to simplify the searching interface for the common man of the local region who is not having the skill or knowledge of searching news in English or other languages. 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 the Gujarati language news video retrieval task. Large scale experiments performed on the dataset created with news channel videos show that the proposed approach gives faster and more memory efficient retrieval compared to proposed deep learning-based retrieval with not much loss in accuracy.

Acknowledgment

This thesis is the result of a journey of work where I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them. Firstly, I offer my adoration to “Bhagwan” (GOD) who created me, gave me the strength and courage to complete my research work. I would like to express my deep gratitude to my Guide Dr. Mehfuza S. Holia, Assistant Professor, Department of Electronics Engineering, Birla Vishwakarma Mahavidyalaya, V V Nagar for her guidance and constant support during this entire period. Her helping nature, valuable moral support, and appreciation have inspired me and will continue to inspire me for the rest of my career.

I would like to thank my doctoral progress committee: Dr. Narendra M Patel, Associate Professor, BVM Engineering College, and Dr. Vishvjit K. Thakar, Professor and Head, Department of Electronics Engineering, Hanshaba Engineering College, Shiddapur for their insightful comments and encouragement during the entire period of research work.

Also, I would like to acknowledge NVIDIA for providing us GPU TITAN Xp worth 1,25,000 INR based on our application for the research work and submitted research proposal.

I would like to thank various Gujarati News Channel Providers and media personals for the broadcast videos used in the evaluation of the system.

I would like to thank Dr. H. B. Soni, Principal, G H Patel College of Engineering & Technology (GCET), and Dr. Maulika S. Patel, Head of Department, Computer Engineering, GCET for providing all kinds of support and timely assistance for the research work carried out. I would also like to thank all teaching as well as non-teaching staff of the Department of Computer Engineering, GCET who supported, encouraged, and helped me indirectly in various ways during the entire tenure of research work.

I extend my heartfelt gratitude to my father and mother, whose hard work has raised me at this level. I thank my in-laws and brother for their love and consistent support. My sincere heartiest special thanks go to my husband for his patience, understanding, and unflinching

support. I am short of words to express my loving gratitude to my loving daughters Khanak and Kaksha, whose innocent smile and love have inspired me during the entire work.

Namrata Dave

Table of Contents

Declaration....ii
Certificate iii
Originality Report Certificate....v
Ph.D. THESIS Non-Exclusive License vii
Abstract....x
Acknowledgment....xi
Table of Contents.... xiii
List of Abbreviations xvi
List of FIGURES xviii
List of TABLES....xx
Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Basics of Content based Video Retrieval (CBVR)....2
1.2.1 Video Parsing....4
1.2.2 Structure of News Video Frame 5
1.2.3 Pre-processing and Feature Extraction 6
1.2.4 Indexing and Retrieval....6
1.2.5 Query Types....8
1.3 Applications of CBVR 9
1.4 Objective and Scope of work 10
1.5 Original contribution by the thesis 11
1.6 Work Plan....12
1.7 Organization of Thesis 12
Chapter 2: Literature Survey 14
2.1 Introduction 14
2.1.1 Pre processing....15
2.2 Video Segmentation and Key Frame Extraction 18
2.3 Advertisement detection....20
2.3.1 Classifiers 22
2.4 Feature Extraction 26
2.4.1 Text Feature Extraction 27
2.4.2 Deep learning approach for Feature Extraction 33
2.5 Existing Content based Video Retrieval System....38
2.5.1 Performance Evaluation Measurement 42
2.6 Summary of the Literature Survey 44
2.7 Challenges with CBVR 45
2.8 Definition of the Problem....49

Chapter 3: Key Frame Extraction and Advertisement Detection 50
3.1 Introduction 50
3.2 Dataset 50
3.2.1 Gujarati Language 51
3.3 Pre-Processing 53
3.3.1 Key Frame Extraction 54
3.3.2 Advertisement Detection and Removal 62
3.3.3 Pretrained Deep Learning Models 63
3.4 Implementation Details 66
3.4.1 Advertisement classification using Alexnet Model for Feature Extraction and SVM Classifier 66
3.4.2 Advertisement classification using Pretrained Model Alexnet 69
3.4.3 Advertisement Classification using Proposed Deep learning Neural Network Architecture 70
3.5 Results and Comparison 73
3.5.1 Method1: Performance using Alexnet and SVM 74
3.5.2 Method 2: Performance using Alexnet Model 74
3.5.3 Method 3: Performance using Proposed Deep Learning Neural Network Architecture 75
3.5.4 Comparison of Different Models 77
Chapter 4: Proposed Text Query based Video Retrieval Approach 79
4.1 Introduction 79
4.2 Proposed Methodology 80
4.2.1 Text Feature Extraction: 81
4.2.2 Indexing 82
4.2.3 Searching Algorithm 86
4.3 Experimental Results 88
Chapter 5: Proposed Deep Learning Approach for News Video Retrieval 93
5.1 Introduction 93
5.2 Autoencoders 94
5.3 Training and Feature Extraction using Auto Encoders 95
5.4 Experimental Details 96
5.4.1 Optimizers and Parameters used for training 100
5.5 Experimental Results 102
5.6 Comparisons 106
Chapter 6: Conclusion and Future Scope 108
6.1 Conclusion 108
6.2 Limitations and Future Scope 109
6.2.1 Limitations: 109
6.2.2 Future Scope 109

List of References 111
List of Publications 119

List of Abbreviations

SVM	Support Vector Machine
CBIR	Content based Image Retrieval
CBVR	Content based Video Retrieval
SGD	Stochastic Gradient Decent
ANN	Artificial Neural Network
CNN	Convolution Neural Network
OCR	Optical Character Recognition
MFCC	Mel Frequency Cepstral Coefficients
ASR	Automatic Speech Recognition
PLP	Perceptual Linear Prediction
LPC	Linear predictive coding
SIFT	Scale Invariant Feature Transform
SURF	Speeded-Up Robust Features
ML	Machine Learning
DL	Deep Learning
MATLAB	MATrix LABoratory
TP	True positive
TN	True negative
FP	False Positive
FN	False Negative
MAP	Mean Average Precision
AP	Average Precision
ED	Euclidian Distance
MD	Manhattan Distance
SVD	Singular Value Decomposition
RGB	Red, Green and Blue
CMY	Cyan, Magenta and Yellow
HSV	Hue, Saturation and Velocity
HSI	Hue, Saturation and Intensity
YC_BC_R	Y - luma component and CB and CR are the blue-difference and red-difference chroma components
BoW	Bag of Words

TF	Term Frequency
IDF	Inverse Document Frequency
RELU	Rectified Linear Unit
FNN	Feed Forward Neural Network
GP	Gaussian process
RNN	Recurrent Neural Network
HIST	Histogram

List of Figures

FIGURE 1.1 Block Diagram of Content based Video Retrieval 3
FIGURE 1.2 Video Structure.....3
FIGURE 1.3 Five consecutive Frames from sequence number 674 to 678 from a news video clip 4
FIGURE 1.4 News Story Format in Different Text bands of Video Frame.....6
FIGURE 2.1 Support Vectors 23
FIGURE 2.2 Biological Neuron and Basic ANN Architecture 24
FIGURE 2.3 Multi-Layer Feed Forward Neural Network Architecture 25
FIGURE 2.4 Video Retrieval based on Textual Query “Australia”.....29
FIGURE 2.5 Video Retrieval based on Textual Query “Sania” in the Hindi language 29
FIGURE 2.6 Types of Morphological operations 31
FIGURE 2.7 CNN Architecture for object category classification[108] 35
FIGURE 2.8 Convolutional Autoencoder Architecture 37
FIGURE 2.9 Snapshot of English language news channel video of non-Indian news channel (a) showing closed caption details(b) showing transcript generated with video 46
FIGURE 2.10 Snapshot of ABP NEWS Gujarati news channel (a),(b) showing no closed caption details available 47
FIGURE 2.11 Snapshot of news video of (a) News channel in India with no transcript (b) BBC News channel which shows a transcript of video.....48
FIGURE 2.12 Video Frame from Different News Story 49
FIGURE 3.1 (a)-(d) Key Frames of a news story 55
FIGURE 3.2 K Key Frame Extraction using Histogram Difference method given in algorithm 1 between frame range 150-210 of video clip 57
FIGURE 3.3 Key Frame Extraction using Edge Difference method given in algorithm 2 between frame range 150-210 of video clip 59
FIGURE 3.4 Key Frame Extraction using Rank and SVD based method given in algorithm 3 between frame range 150-210 of video clip 60
FIGURE 3.5 Performance Evaluation of Three Algorithms with three datasets 61
FIGURE 3.6 Proposed Method for Advertisement detection 67
FIGURE 3.7 (a)Training Accuracy (b) Training Loss of proposed approach using transfer learning with Alexnet pre-trained model and SVM classifier with Bayesian Optimizer 69
FIGURE 3.8 (a)Training Accuracy (b) Training Loss of proposed approach using transfer learning with Alexnet pre-trained model.....70
FIGURE 3.9 Training Accuracy and Loss during Model Training for Advertisement Classification using SGD optimizer 72
FIGURE 3.10 Training Accuracy and Loss during Model Training for Advertisement Classification using ADAM optimizer 73
FIGURE 3.11 Comparison of Performance of ADVNET model with different optimizers Adam and SGDM 77

FIGURE 3.12 Performance comparsion of different deep learning models for advertisement classification task 78
FIGURE 4.1 Block Diagram of Proposed System for Content based Video Retrieval(CBVRAPP1) 80
FIGURE 4.2 Text Extraction from the frame of input video 81
FIGURE 4.3 Text Feature Extraction 81
FIGURE 4.4 Term-Document Matrix 82
FIGURE 4.5 Dictionary File and Posting File created for indexing documents 85
FIGURE 4.6 Cosine similarity between query and document vectors in vector space model 87
FIGURE 4.7 Dataset Size used for CBVR 89
FIGURE 4.8 Number of Keyframes of different dataset 89
FIGURE 4.9 Response time for different datasets in microseconds/query 89
FIGURE 4.10 Top 6 results retrieved on query text “સલમાન” 90
FIGURE 4.11 Performance Evaluation in P@Ri of the proposed approach for query Q1 91
FIGURE 4.12 Performance of Text Query based video retrieval 92
FIGURE 5.1 Performance Evaluations for Deep Learning And Machine Learning Algorithm 93
FIGURE 5.2 Autoencoder 94
FIGURE 5.3 Block Diagram of Proposed Deep Learning approach for Image Query based Video Retrieval System 96
FIGURE 5.4 Architecture of Encoder 97
FIGURE 5.5 Conv2D parameters 98
FIGURE 5.6 3×3 kernel applied to an image with padding 99
FIGURE 5.7(a) Video frame used in training of size 128 x128x3 (b) Predicted code generated with encoder reshaped from the original shape 16x16x8 99
FIGURE 5.8 Model Accuracy and Loss for 100 epochs and 5 batches 104
FIGURE 5.9 Training Accuracy Improvement with increasing epochs 104
FIGURE 5.10 Performance Evaluation in P@Ri of the proposed approach for query Q1 105
FIGURE 5.11 Results of News Story clip Retrieval for given Query image 105
FIGURE 5.12 Performance of proposed CBVR approach 106
FIGURE 5.13 Performance of Image Query based video retrieval approach(CBVRAPP2) 107
FIGURE 5.14 Comparison of proposed approaches CBVRAPP1 and CBVRAPP2 107

List of Tables

TABLE 2.1 Literature review of Stemming Algorithm/Morph Analyser....32
TABLE 2.2 Confusion Matrix for binary classification....42
TABLE 3.1 Dataset Property....51
TABLE 3.2 Gujrati Consonants[117]....52
TABLE 3.3 Gujarati Vowels[117]....52
TABLE 3.4 Gujarati consonant ક with all matras....52
TABLE 3.5 Gujarati Digits[117]....53
TABLE 3.6 Some Gujarati conjunct consonants[117]....53
TABLE 3.7 Sample Gujarati language sentence text[117]....53
TABLE 3.8 Key frames from each dataset....61
TABLE 3.9 Architecture Details of Alexnet Model....65
TABLE 3.10 ADVNET ARCHITECTURE....70
TABLE 3.11 Confusion matrix of classification of News vs Advertisements using Alexnet and SVM....74
TABLE 3.12 Confusion Matrix for advertisement classification using pre-trained ALEXNET model....75
TABLE 3.13 Confusion Matrix for ADVNET using SGDM optimizer....75
TABLE 3.14 Confusion Matrix for ADVNET using ADAM optimizer....76
TABLE 4.1 Few examples of Stop Words in 'Gujarati' language....83
TABLE 4.2 Results using Precision-Recall and Response Time/Query for different datasets....91
TABLE 5.1 Layers, Shape and Parameters per layer of Encoder in Proposed Autoencoder Architecture....100

Chapter 1:

Introduction

1.1 Introduction

A large amount of digital content is getting generated every hour in terms of text, images, videos, etc. in day-to-day online activities all around the world. As per recent research, consumption of online video has increased to double whatever it was past year. In the last 2 years, online video usage reached 3.7 billion videos per month. This kind of tremendous growth has been driven by a significant increase in the number of online video viewers in India. In 2016, the percentage of total online video audience in India has increased by 74% i.e. 54 million viewers, with an average Internet user watching 18% more videos[1]. Online video audience in India is expected to reach 500 million in 2020, reveals "Technology, Media, and Telecommunications (TMT) Predictions 2020" released by Deloitte Touche Tohmatsu India[2].

To process a large amount of video cost-effectively, it is important to have techniques that serve the purpose of extracting meaningful information quickly. Researchers have built several technologies for intelligent video management which include shot transition detection, key frame extraction, video summarization, content-based video retrieval, and more.

Content-based video retrieval is one of the challenging tasks in the domain of information retrieval. The system helps the users of the system in the retrieval of preferred video stories or clips from the collection of videos efficiently based on the video contents. Mostly, the content-based video retrieval system is divided into two main tasks. The first task is the extraction of features from video clips or segments and the second task is to provide a retrieval model to position similar video clips from the video database.

A large number of approaches employed a wide variety of features to symbolize a video sequence of which color histogram [3], shape information [4], motion activity [5], and text analysis [6] are renowned few approaches. Also, many approaches utilized the said features to improve the retrieval performance [7].

A video clip is a sequence of frames and each frame is similar to one image. Indexing each of the frames as a still image causes extremely high redundancy due to the similarity of frames in nearby regions and is difficult for the number of frames in a video for even one hour. Video is a structured collection of frames in which actions and events, in a time and space domain, comprise stories or carry particular visual information. Due to this fact, a video program should be viewed as a document rather than a non-structured sequence of frames [3]. It is required to find the structure of the video and divide the video into basic components. Index can be built based on the structural information and image frames of the video.

1.2 Basics of Content based Video Retrieval (CBVR)

CBVR is one of the many applications of the computer vision domain which are useful for searching digital videos in large databases. The term "Content-based" means that the contents of the video document are analyzed for searching the desired video from the collection based on the user's query. In this context, the term "content" might refer to a different type of features such as colors, shapes, textures, or any other information that can be derived from the video itself [4].

CBVR has many possible solutions out of which mainly explored solutions are based on the content of a video extracted from the video itself and another solution is annotations-based video retrieval. CBVR based content of the video is desirable because searches that rely purely on metadata are dependent on annotation quality and completeness. It can be time-consuming and may not capture the keywords chosen to label the video when annotation of video is done manually by entering keywords or metadata in a huge database. The evaluation of the effectiveness of keyword video search is subjective and has not been well-defined. The increased availability and usage of digital video have created a need for automated video content analysis and multimedia database management techniques. The vast amount of content information brings a great need for efficient techniques of finding, accessing, filtering, and managing video data.

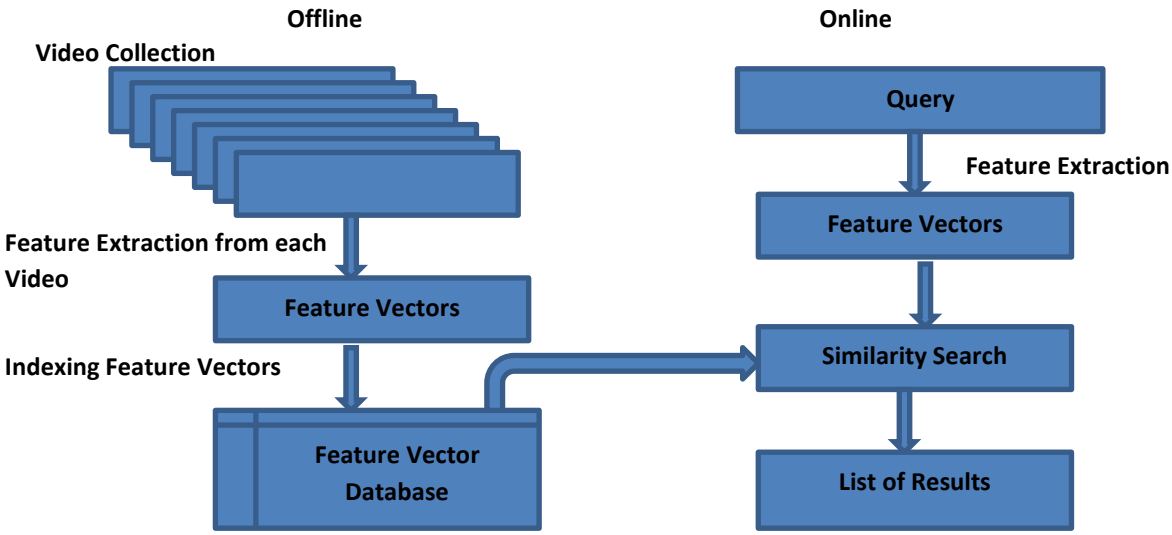


FIGURE 1.1 Block Diagram of Content-based Video Retrieval

The content-based video retrieval system is explained in the block diagram in Figure1.1. For the indexing task, a feature extractor is applied to the database of video to extract feature vectors. Features are extracted from each video document and stored in a database of feature vectors. The next task is the indexing of feature vectors for faster and accurate retrieval. The user of the CBVR system gives input query in any form either text, image, or video clip. Feature extraction is done similarly from the query. For the retrieval task, the similarity between the query vector and database vectors is measured and final results are displayed based on matching between them.

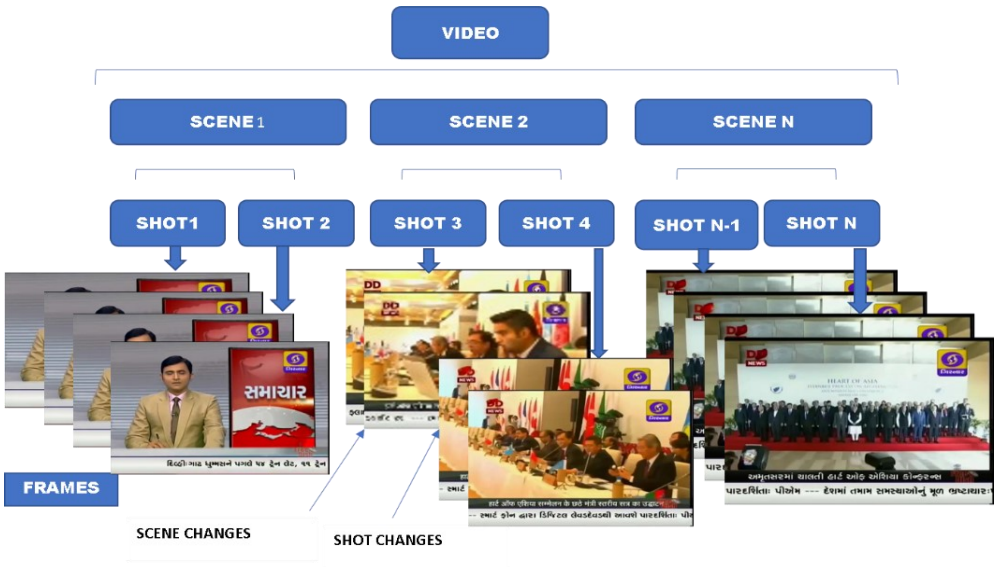


FIGURE 1.2 Video Structure

1.2.1 Video Parsing

A video can be perceived as a document. Indexing of the video can be similar to the indexing of text documents. The structural analysis is performed on the text document to divide it into paragraphs, sentences, and words followed by generating indices using them.

Altogether, Video is a collection of various stories or scenes as shown in the structure of the video in Figure 1.2. Each scene is further divided into shots. Shots are made up of multiple frames containing almost similar features. There is a 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 the time taken to search through video. A key frame is a frame representing the content of each shot of the scene in the video.

FIGURE 1.3 Five consecutive Frames from sequence number 674 to 678 from a news video clip

Many frames of shot can have similar information. Normally frame rate for News videos of regional languages like Gujarati is 25 frames per second. We can select one of the frames

representing more than 60 frames or more. Any video can be seen as a sequence of similar frames with very little change in visual information as shown in Figure 1.3.

Normally when a book-writer writes a book, a table of contents is generated for the content's order, and also the index of keywords and phrases is generated to enable searching the book by the content. In the same manner, to ease the searching of video data with speed and accuracy, a video document gets segmented into shots and scenes to compose a table of contents, and key frames or key sequences are extracted and utilized to generate index entries for stories or scenes.

Though, generating an index for video from the vast collection of videos is relatively more complex than generating a book index for books. The format of the index generated for books is fixed whereas generating an index for videos depends largely on the viewer's interests and it may cover a wide range[4].

Due to this fact, the majority of researchers are contributing to content-based video retrieval tasks are developing technologies for parsing video documents automatically, using audio, visual features, and/or text to retrieve meaningful composition structure and also to extract and represent content features of the video documents.

Proposed content-based video retrieval is implemented on the news video dataset as described in chapter 3 of the thesis.

1.2.2 Structure of News Video Frame

Video is a collection of frames containing useful information for video processing. As shown in Figure 1.4, the frame of the news video contains various visual features as well as news headlines as overlaid text in the frame of the video. In different news channels, the format is different to show headline news and main news story.

In some of the news channels, the main story runs in the upper part of the video while headline or advertisements appears in the last text band in most of the news channels videos. Also, the logo of the news channel appears at different places in the frame of the news video of the various news channel. The common text band for a current news story is most of the news channel video is either the first band in two or three-band appear in the bottom part of the frame as shown in Figure 1.4 or it may appear in the header part in the frame of video.

FIGURE 1.4 News Story Format in Different Text bands of Video Frame

1.2.3 Pre-processing and Feature Extraction

Pre-processing steps for the video like video segmentation, key frame extraction, and feature extraction are necessarily performed on input video before content-based retrieval is implemented on the data.

Because of the importance of keyframe extraction in video retrieval, many researchers are working in this area. Many approaches are used for keyframe extraction. One of them is shot-based structuring of video, in which the first shot is detected then each shot is represented by a fixed or variable number of frames.

Features used for further processing of indexing tasks are extracted from the input video. Most features in the video can be classified into two broad categories: spatial and temporal features as well as saptio-temporal features. Mostly used visual features are color, histogram, texture, edge information, edge change ratio, and motion features. SIFT and SURF are also used due to the benefits of these features like robustness, rotation invariance, etc.

1.2.4 Indexing and Retrieval

Indexing of data elements plays crucial role in overall system performance of any CBVR system. Documents in any form whether it is text documents, images or videos can be retrieved based on the user query from the indexed data with speed and ease only if the indexing is done properly.

The history of manual indexing is very old and also automatic or semi-automatic indexing concept was explored in 1948-1951. In 1948, the machine was invented called 'Univac' which was able to search for text references associated with the given subject code. The machine was able to process 120 words per minute[5][6].

Indexing can be related to a process of information extraction rather than information analysis. Various indexing methods are found in the literature. A few of the popular indexing methods are inverted files, suffix trees, and signature files[7]–[14]. Each of the indexing methods has its own merits and demerits. Indexing required storage of files with all necessary descriptors such as keywords, concepts, metadata about files, etc. Also, the knowledge, as well as expectations of the searcher, needs to be taken care of while generating an index for the given data based on the information domain.

Later on, schemes based on term weighting and inverse document frequency emerged as a new concept for indexing documents. With the new concepts and variations in the tf-idf scheme, the information retrieval models were extended further. Advances in the basic vector space model were also developed for retrieval tasks and LSI-Latent Semantic Indexing[15] is the most well-known model. Indexing methods evolve with time and also the concept of semantic indexing emerged as one of the ways of indexing[16][17].

Retrieval approaches also evolve with time. The approach based on ranking to search which documents were sorted with a query was explored by researchers working in information retrieval. The approach of ranked retrieval is refined and revised over the time by researcher community. The superiority of this approach over the traditional Boolean search approach was demonstrated in many experiments over the years.

The next task after generating feature vectors from data and query is to measure similarity between them. There exist many methods to find similarity between feature vectors and can be applied based on the nature of data to avail better performance. Similarity measures are normally used to rank the documents based on the similarity scores between the feature vector and query vector generated from the user query. To find similarity methods such as Euclidian Distance, Cosine Similarity Measure, Jaccard Similarity Measure, Dice Coefficient Measure, Pearson Correlation Coefficient (PCC), etc.

The proposed research work is for news video retrieval tasks based on the user query. News stories can be retrieved using either of the query-by text, image, audio, video clip, or

multimodal paradigm to respond to user queries. In the proposed approach, both types of query text and image are used for news story retrieval tasks using two separate methods.

1.2.5 Query Types

Retrieval is performed based on queries from the user in the CBVR system. User query can be of multiple forms such as text, image, sketch, video clip, etc. Queries using objects, sketches or example images do not utilize semantic information. Some of the widely known query types are discussed below.

In the system with 'Query by Object', the object image is provided. The occurrences of objects in the video database are detected and the locations of the object determine the success of the query [18].

Where as in text query base approach text is the user input as a query. It is a widely used and well-known method of query in any information retrieval task including CBVR.

Another well-known way is 'Query by Image'. It is popular for content-based image retrieval, example images can be used as queries to retrieve relevant videos in a database of videos (query by example) but it has a limitation that motion information of the video being searched is not utilized. It relies only on appearance information. Also, finding a video clip for the interesting concept may become too complex using an example image. Textual query offers a more natural interface and claims to be a better approach for querying in video databases [10].

Query by Example: Query by example is better if visual features of the query are used for content-based video retrieval [2]. Low-level features are obtained from key frames [9] of the query video and then they are compared to separate the similar videos using their key frame's visual features.

Query by shot: Some systems utilize the entire video shot as the query instead of key frames [5]. This can be a better option but with a higher computational cost.

Query by a clip: A clip can be used for better performance of video retrieval as compared to the technique when a shot is used because a shot does not represent sufficient information about the whole context. All the clips which possess a significant similarity or relevancy with the query clip are retrieved [2].

Query by Faces and Texts: Faces and Texts can also be used as a query to retrieve a video segment containing frames labeled for a specific type according to faces and texts [23]. A suitable algorithm can be used to search the video enquired by the query clip using information.

1.3 Applications of CBVR

Innovations in technologies lead to a large amount of video data on the web as well as storage centers all over the world. Due to the richness of content of video data, utilization of video is now at an all-time high. Emerging demands of videos and the technologies to index and retrieve them are the need of the day. Content based video retrieval applications are found in numerous fields such as movie videos[18], sports videos[19], broadcasted news videos[20], lecture videos, surveillance videos[16] etc. content-based video retrieval as well as summarization of Lecture videos[21][22][23], movie videos[18], sports videos, news videos are of high demand now a days.

Some of the applications of CBVR are listed below:

- Video Lectures: At present entire world is facing a COVID-19 pandemic situation which makes distance learning compulsory for students all over the world. For many years, distance learning was already used for many courses all over the world. Videos created for Lectures can be indexed and retrieved using audio and visual features which makes searching for relevant topics easy for the viewer[21].
- Medical Video indexing and retrieval: Videos generated in the medical field are different than the normal videos. Preprocessing, feature extraction, etc. required domain knowledge for indexing as well as retrieval of medical videos. Taken as input a stream of video through the camera monitoring the ongoing surgery, in real-time similar video subsequences are retrieved by the system from video datastore created for the same task [24].
- News Broadcasting: The task of retrieving news stories from the videos based on specific events or topics or personalities makes it a very attractive and useful application to the end-users. Textual content either available as a transcript, closed captions, or appearing on screen as well as visual features plays a vital role in retrieving news stories from the vast collection of videos.
- Advertising Retrieval: a visual-based approach applies to the task of advertisements retrieval from the collection of videos rather than the text-based approach of retrieval.

- Film and Television video: In this digital era, the working pattern of professional media production has been changed. Labeling of sequences of images as well as video footage based on the content has become necessary for successive stages of film and tv production. Various approaches are explored to support professional media production using a tool for effective searching and retrieval of video data [18][25].
- Music video clips: The characteristics of such documents may be difficult to express, based on given textual annotations stored on a database. Retrieval of music video based on audio is a straightforward way to search the music contents from video. Studies of the literature show the cross-modal music video retrieval in terms of emotion similarity[26].

1.4 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 the Gujarati news channel video dataset. Text processing for the Gujarati language is challenging due to a lack of resource availability for the Gujarati language. Gujarati language text processing is still not fully explored by researchers. Due to this reason, neither the benchmark dataset nor the efficient tools for OCR, ASR, Stemming, and Lemmatization available for 'Gujarati' language processing.

Generation of the tremendous amount of digital content daily and challenge in the retrieval of the required content from it 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 the dataset. The proposed work is focused on the retrieval of news stories from the collection of news video datasets using the 'Gujarati' language text query. Also, an alternative approach is proposed which is mainly to explore the unsupervised deep learning for the CBVR task based on image query to compare the retrieval performance. 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 an efficient and cost-effective model for further reducing the amount of data to be processed by removing the advertisement from the dataset of keyframes using the transfer learning approach.

- ‘Gujarati’ Text Feature Extraction from video frames and text processing using various NLP fundamentals as well as indexing them for faster retrieval.
- Exploring efficient Unsupervised Deep Learning Architecture for image query-based video retrieval approach.

The 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 of video, closed caption details etc. required to process text-based video retrieval tasks.

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

1.5 Original contribution by the thesis

In this thesis, we have proposed content-based video retrieval for Gujarati news videos. The retrieval task is evaluated using two different approaches based on query type and content type for retrieval.

A keyframe extraction algorithm is developed to reduce the time for processing input video. Further to remove unwanted information from the dataset created for processing, it is necessary to remove advertisement frames too. So, a very efficient approach based on Deep Learning is proposed to classify frames as advertisements and remove them from the data to be processed further for indexing and retrieval tasks.

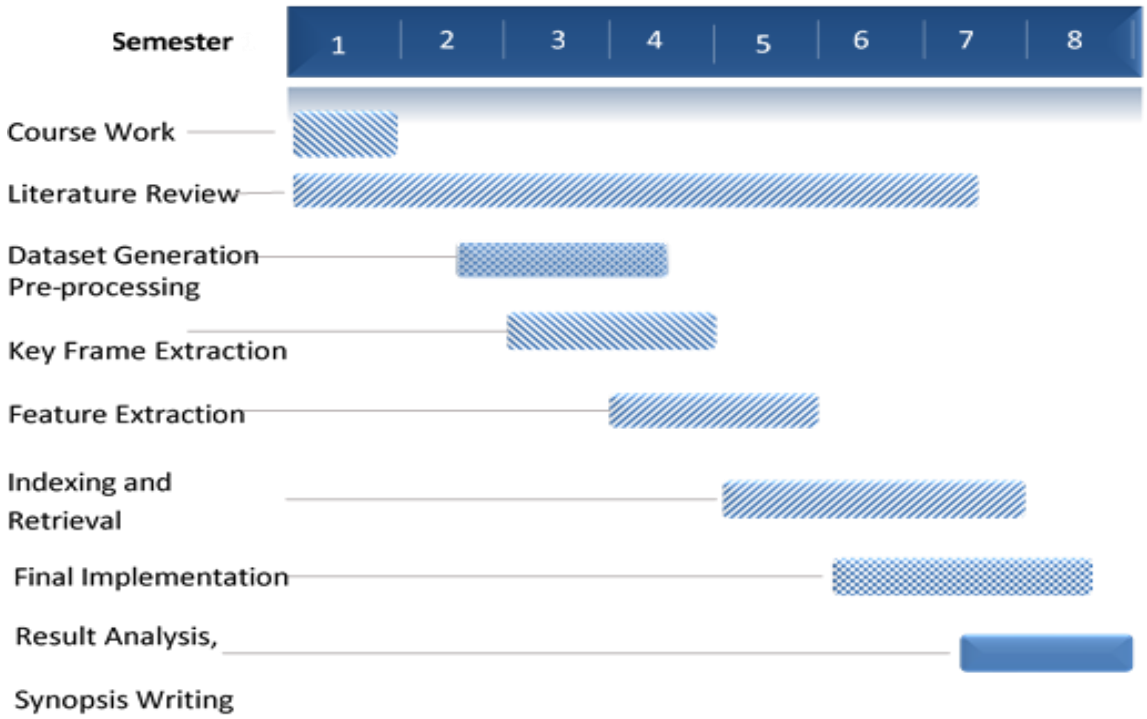
The first retrieval approach is based on a ‘Gujarati’ language text query to retrieve news stories from the dataset of Gujarati language News videos. This kind of work is not explored for regional language like ‘Gujarati’ which is spoken by more than 55 million people worldwide and is the 26^{th} most widely spoken language in the world[27].

The second approach is based on Deep Learning architecture-Auto Encoders, which is used for unsupervised learning to retrieve news stories based on image query. This approach is proposed to test the system for optimized performance. With both proposed approaches the system gives good results.

Based on all the proposed methods, total of four research papers are published in different reputed peer-reviewed international journals and conference as mentioned at the end of the

thesis. One more research paper is already submitted to a reputed Scopus indexed journal for further processing.

1.6 Work Plan



1.7 Organization of Thesis

The thesis is organized as given below:

Chapter 2 presents a literature review of content-based video retrieval systems. It covers the fundamentals of image preprocessing which is also part of video preprocessing tasks. For the proposed work, the literature review is given for different modules like key frame extraction, Advertisement detection using deep learning, text-based video retrieval, Gujarati language text processing, and deep learning approaches for video retrieval. Also, methods for indexing and similarity measurements are discussed in brief. At the end of the literature review, a summary of the literature survey is given followed by challenges in the implementation of the proposed work. Finally, the problem definition is presented at the end of the chapter.

Chapter 3 gives details about the dataset used for the retrieval of the news videos task. Also, the preprocessing on dataset videos and properties of video are described in the chapter. The

chapter also discusses implementation details and results of key frame extraction and Advertisement Classification approaches which are fundamental steps in our proposed approaches of content-based video retrieval.

Chapter 4 describes the “Gujarati” language text query based news story retrieval from Gujarati news video approach and its implementation details.

Chapter 5 of the thesis presents the Deep learning concept is used in the implementation of the second approach for News Story Retrieval using Image as a Query from the news video collection of ‘Gujarati’ news channels. Experimental results are discussed with comparisons with the first approach at the end of the chapter.

In Chapter 6, the conclusions for the research work mentioned in the thesis based on the objectives achieved are mentioned. The limitation of the proposed work and the future scope of the research is also discussed in the chapter.

Chapter 2

Literature Survey

In this chapter, a detailed description of different methods, algorithms, and architectures used in various phases of the entire Content based video retrieval (CBVR) task is provided. To express the content briefly and clearly, a detailed description of supporting theory is avoided as much as possible. A brief introduction is provided for highlighting basic concepts that contributed to the understanding of this thesis.

Mainly the literature survey is conducted for CBVR based on text features, image features, and deep learning concepts. Also, existing content-based video retrieval for news videos with Hindi, Bangla, and other Indian languages is explained in brief for reference. At last, all the literature listed helped us in designing the proposed system for research work.

2.1 Introduction

In Content based Video Retrieval task, various algorithms at a different stage of the entire process deal with the processing of data in visual, audio, or text form. During the data acquisition process, data in any of this mode might get affected by some noise which makes it difficult to process data directly.

Video data is generally not processed directly and is usually converted to visual frames or audio frames for further processing. To build an efficient CBVR system, it is essential to improve the quality of images by removing any unwanted information from the video frames.

Image enhancement can be considered the very first step in digital image processing. Image enhancement in the spatial domain can help improve image quality based on the type of application for which the image needs to be enhanced. Image enhancement using neighborhood processing is one of the spatial domain methods used frequently in many applications of the image processing domain. In neighborhood processing, the pixel as well as its immediate neighbors are considered for processing.

2.1.1 Pre processing

Various image pre-processing techniques for enhancement of image like the elimination of noise from the input image, histogram modeling, color space conversion, morphological operations, binarization, Image Segmentation are discussed in the following paragraph briefly.

2.1.1.1 Elimination of noise:

The first method used for image preprocessing is the removal of noise. Noise is nothing but a high-frequency signal and can be eliminated with low pass filtering techniques in the spatial domain as well as in the frequency domain. The main source of noise in a digital image or video arises during the acquisition as well as the transmission process. Also, the noise comes from the sensors or during the analog to the digital conversion process. Broadcasted videos also contain the problem of video quality due to the transmission process. Also, shooting certain news stories at public places degrades the quality of visual information of the shot of the video.

Various types of noise that may occur in images or videos are salt-and-paper noise, shot noise, quantization noise, gaussian noise, Poisson noise, Gamma noise, and speckle noise[28]. Shot noise and quantization noise are found in the video. Quantization is the process of reducing a large set of normally continuously changing values to a smaller representative set of values in the output. Quantization noise in the video can occur when the image clips far beyond the extreme ends and becomes worse if the image already contains another strong type of noise. Also, it is difficult to find quantization noise in the video. Also, the camera capturing the image generated noise due to various reasons like heat, sensor illumination levels, and electricity[14]–[17].

Depending on the type of noise present in the image, different filters can be applied to improve the image. Generally used filters in the image processing domain for noise removal and image enhancement are mean, median, adaptive, average, wavelet transforms, etc. Most of the filter's aim is to remove the noise without loss of useful information of the images[29][32].

2.1.1.2 Histogram Modelling:

A histogram of the image is used to provide global details or descriptions of the image. The techniques used for histogram modeling modifies the image globally by providing the

desired shape to the histogram of that image. One of the popular histogram modeling methods which provide uniform distribution of all the possible intensities in the image is histogram equalization. Images that are very light or dark have a narrow histogram. Histogram equalization increases the dynamic range of the image and hence improves the image quality. The equalized histogram provides images with better quality that in turn helps to generate better feature vectors during the processing of images for feature extraction tasks in a video retrieval system.

2.1.1.3 Morphological operations:

Morphological operations are used for extracting image components which are useful to represent regions and shapes. It is performed on binary images. Operations are performed using a template called structuring elements of different sizes and shapes. Erosion, Dilation, Opening, and closing are the main basic operations used in the morphological processing of images. In the dilation operation, valleys between spiky edges are filled and objects in binary image thicken based on the structuring element chosen. As opposed to it, the Erosion operation deletes spiky edges in the binary image, and objects shrink or even are eliminated based on the structuring element applied[29].

Particular information can be simply retrieved from the frame of the video with the help of morphological operators discussed above. Final results can be improved by a significant amount with the help of video frames obtained after successfully applying morphological operators. Morphological operators can be applied to remove unwanted content[29].

2.1.1.4 Color space conversion:

The Color model or Color space is the standard specification of colors accepted universally. The Color space is represented with a coordinate system and each color is considered as a single point in a three-dimensional coordinate system of color space.

In the field of digital image processing, the most commonly practiced models are the RGB (red, green, and blue) model, CMY(cyan, magenta, and yellow) color model, CMYK (cyan, magenta, yellow and black), HSI (hue, saturation, and intensity), YCbCr color model, etc.[29].There are numerous color models in practice nowadays. Models which are popular due to their inherent properties are discussed due to their importance in the field of image processing applications.

RGB Color Model: Each point in a color image is represented by a point in the 3D coordinate system of R, G, and B. The origin of the RGB space with Red=0, Green=0, and Blue=0 represents black color while the opposite corner with Red= max, Green=max, and Blue=max represent the white color. Image in RGB color space is represented as R x C x 3 where R x C is the physical size of the image and 3 represents the R, G, B planes of that image. It represents the image as red, green, and blue color. The RGB model is generally used for color monitors as well as color video cameras.

HSV Color model: In the HSV Color model, H stands for Hue, S for Saturation, and V for value. It is one of the popular cylindrical shape color models. Unlike other color models, HSV is closer to the human visual system and is more related to how a human perceives colors. Generally for specific applications converting RGB image into HSV color space gives better results as color and luminance components can easily be separated in it[33].

YCbCr Color model: YCbCr is also one of the very popular color models used in a few of the algorithms, where the images from RGB space were converted to images in YCbCr color space. Human eyes are more sensitive to changes in brightness or luminance information in comparison to color changes. YCbCr color model generally considers only the luminance component of the image and Y indicates luminance which means light intensity. Cb is chroma component blue-difference and Cr is chroma component red-difference. An image represented in one type of color space can be converted to another type for better results in specific applications. This is called color space conversion.

2.1.1.5 Binarization: The binarization process is the pre-processing technique that uses image intensity distribution to create the binary image of the original colored image. Binarization is a process of converting multi-tone or gray scale images into black & white binary images. Binarization is the basic step for segmentation in many applications[22][34]–[36]. Binarization is a necessary step to separate text and non-text in the frame of news video.

2.1.1.6 Normalization:

In the digital image processing field, normalization is the primary step in numerous image processing applications. Basically, in normalization range of pixel intensities changes. Normalization transforms a grayscale image with an intensity range (max, min) to a new image with the intensity range (new max, new min). Normalization provides a global description of the appearance of an image. It is also referred to as contrast stretching. Images

with normalized intensity values as well as with uniform illumination improve the quality of images[29].

1.1.1.7 Image Segmentation:

Segmentation is the process of dividing an image into multiple segments which pertain to specific information and can be used for further processing. Many applications require segmentation as a basic and most important step also. Applications like object identification, tracking, counting of objects in the image, etc. require good segmentation techniques to generate accurate results[29][37][38].

In text-based video retrieval, it works with optical characters, handwritten characters, scanned text, annotated text, and many more types of text information. News video retrieval using text present in the scenes of the video involves detection and processing of text from the scene of the videos. Various pre-processing techniques need to apply to text to get better results for retrieval. Normalization, edge contrasting, contour extraction, morphological operations applied on the text band for the news [7][25][26].

2.2 Video Segmentation and Key Frame Extraction

Video can be viewed as a collection of meaningful scenes, shots, and frames. The scene is further divided into shots. The shot is a collection of frames captured during a single camera motion. The frame is the most basic unit of video to consider for processing. In Content based video retrieval system, the first task is to segment video into shots and select key frames representing each shot uniquely. Keyframes are the representative frames that are used to provide a suitable abstraction and framework that will help for indexing, browsing, and retrieval of video. By selecting Key Frame, the task of processing video is reduced by a large amount as all frames are not required to be processed 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[23][25][40]–[42].

Segmentation of news video is one of the fundamental steps for news video retrieval or summarization approaches[43]. The segmentation can be performed by identifying the shot boundary of the video based on the different features present in an audio stream or video stream. Combined features are also helpful in segmenting videos efficiently.

Because of the importance of key frame extraction in video retrieval, many researchers are working in this area. Many approaches are used for key frame extraction or shot boundary detection. In many applications, first shots are detected from the video and key frame extraction is applied to each shot based on different methods. As each shot contains multiple frames with different kinds of features, key frame extraction based on the features can be done.

Automatic shot boundary detection from the input video is explored using various techniques such as block based techniques, transform based techniques, Pixel-based [44] techniques, histogram based techniques, feature based techniques, etc. [46][47]. In the shot boundary detection using pixel-based technique utilize pixel difference as the key parameter for shot detection. These techniques are highly sensitive to noise. The technique based on block processing work on the fundamentals of pixel processing itself but the method operates on the image at a time in blocks of pixels which makes it faster compared to pixel processing techniques. The most popular approach is color histogram-based approaches, but it does not require any position of the pixel. The methods of shot boundary detection using entropy measure[47] are also found effective [48][49].

Ni et al. [50] proposed and analyzed 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 function 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[25].

Ji et al. 2019[48] 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 color 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.

Conditional random fields have been used by many researchers for the segmentation of video[51]–[53]. Kannao et al.[53] proposed segmentation of TV news broadcast into semantically meaningful stories. A 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. Wang et al. [52]proposed multimodal features using conditional random fields (CRFs) for the segmentation of broadcast news stories. With the use of the multimodal features, a linear-chain CRF was used to identify each candidate as boundary/non-boundary tags.

2.3 Advertisement detection

Multimedia information broadcasted on television always contains visual content such as news stories, sports videos, movies, or daily shows which is of interest to users watching television and also commercials in between different shows or stories. The length of advertisements or commercials is limited and rules are defined for the length and frequency of advertisements in different countries. Although, each country in the world has its laws for multimedia content broadcasted on television.

In the context of news videos 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 events in Indian news recordings. Due to this fact, it is required to identify advertisements from the videos and separate them is a very necessary step for news or any other multimedia video processing. Researchers have worked in the area of identifying and extracting advertisements or commercials from the videos. Advertisement detection is also useful to the advertisement industry for different commercial reasons.

Commercial or advertisement detection or recognition, as well as removal, can be achieved for different categories of broadcast videos. Researchers have worked for advertisement detection from broadcast videos of movies, news, sports, etc. using various techniques [40] [50] [55].Mostly visual features are used for advertisement detection from video. Also, the combination of visual features, as well as acoustic features, are used for commercial detection from news videos [40] [56].

Classification of advertisements from the video has been achieved using various visual features such as Edge Change Ratio ECR, the difference of frames, length of video shot [56] [57], text location in the frame, and availability of specific text bands in news video frames [40] [59] [60]. Also, audio features such as spectral centroid, flux, roll-off frequency [60], short-time energy, zero-crossing rate (ZCR), etc. are used in different applications to detect commercials[17][61]. Audio features such as MFCC (Mel Frequency Cepstral Coefficients), PLP, LPC [62][63] which is very much used in speech processing applications, is used to differentiate advertisements in news channel effectively [61]. It is observed that when an advertisement starts the audio loudness increases dramatically. This fact is used by many researchers for advertisement detection in news videos as well as in sports videos.

Vyas et al. presented features for the task of automatically identifying as well as extracting 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 a dataset made up of 54 hours of video recorded with three English Indian news channels and obtained around 97% F-measure[61].

Almgren et al.[64] analyzed 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 applications. They employed convolutional neural networks to classify scanned images as either advertisements or non-advertisements (i.e., articles).

B. Zhang et al.[49] 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.[65] 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 data is divided into shots and classification is done in commercial class and non-commercial class using ANN and SVM. With their approach, they can 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 the user experience of watching advertisements 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 a 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 an algorithm to recognize the scene changes with the visual significance map, through which the system can find stable areas in distinct scenes for advertising[66].

Li et al. proposed an automatic commercial detection system for TV broadcasting. This system works at a 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 the deep convolutional neural network and train a support vector machine classifier to complete shot classification for TV programs[54].

2.3.1 Classifiers

Different classifiers like Support Vector Machine[43][50][65], Artificial Neural Network [65], CNN, etc. are used to classify advertisements from other frames. Features extracted by the deep network are recently being used efficiently for feature extraction in commercial detection by researchers[54]. Convolutional Neural Network[54] also being used by many researchers for the advertisement detection task.

2.3.1.1 Support Vector Machine

Support Vector Machine (SVM) algorithm is robust and most commonly used classification and regression algorithms in various fields of data processing applications. The SVM has been playing an important role in pattern recognition in the digital image processing field as well. Research in some fields where SVMs do not perform well in earlier days has encouraged the development of other variants such as SVM for large data sets, SVM for multi-class classification, and SVM for unbalanced datasets. Further, SVM has been integrated with different optimization algorithms based on the application to improve the performance of classification by optimizing various parameters.

The main objective of the support vector machine is to distinctly classify data points in N-dimensional space using an N-dimensional hyperplane. As shown in Figure 2.1, data points might be separated linearly by small margins or large margins. Support vectors are the data

points closer to the hyperplane that separates data points with the maximum possible margin. The data points in support vectors influence the position and orientation of the hyperplane.

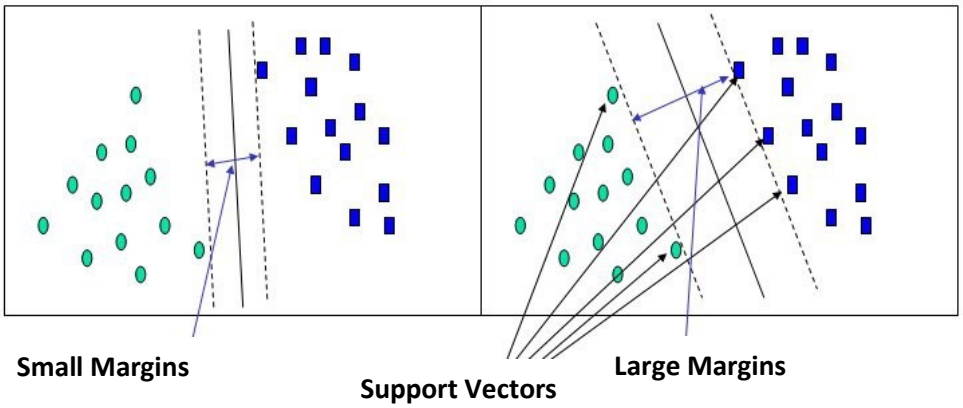


FIGURE 2.1 Support Vectors

If there are n training data points $x_i \epsilon R_d$, $i = 1,2, \ldots$, n, which can be divided into two different classes and their respective class labels are given by $y_i \epsilon \{-1, 1\}$, i = 1, . . ., n. Generally, a linear support vector machine which creates a hyperplane described by equation 2.1 to separate data points in the d-dimensional space R_d which takes the largest distance to the adjacent training data points of each class. Multiple such hyperplanes are possible and an optimal hyperplane is chosen out of all the possible ones.

$$\boldsymbol{wz + b = 0} \qquad (2.1)$$

Cortes and Vapnik [67] shown in that research that the weights w for the optimal hyperplane given by equation 2.1 in the feature space can be written as some linear combination of support vectors as given in equation 2.2.

$$\boldsymbol{w = \sum_{support\ vectors} \alpha_i z_i} \qquad (2.2)$$

The linear decision function I in the feature space will accordingly be of the form given by equation 2.3,

$$\boldsymbol{I(z) = sign\left(\sum_{support\ vectors} \alpha_i z_i \cdot z + b\right)} \qquad (2.3)$$

here equation 2.3 represents the dot-product of support vectors denoted by zi and vector z in feature space[67]. The decision function is fully specified by a subset of training samples called support vectors. SVM can be applied to non-linearly separable data also. The data points are mapped to new space where it is possible to separate them with help of a hyperplane. Finding optimal hyperplane is very important in SVM.

SVMs can be extended to n- class multiclass classification problems. One of the ways to do it is by doing one v/s many classifications, where n different SVMs are trained based on the set of points that belongs to the class and set of points that do not belong to the class. These decision boundaries are then collected to decide overall boundaries.

In SVM a set of mathematical functions are used which are defined as the kernel. The kernel takes data as input and transforms it into the required form. Different combinations of SVM algorithms and kernel functions are explored by researchers. Kernel functions are of many types such as linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

Support vector machine classifier is very efficient in the larger dimension data. SVM is proven to be successful when the number of features is more than training samples in the experiment. SVM is also considered an excellent approach for easily separable classes. The support vectors affect the hyperplane which makes the system less sensitive to outliers' data. In binary classification as well as multiclass classification, SVM is providing great results.

The major disadvantage of the support vector machine is the time taken for training is more when the amount of data is large. SVM doesn't well perform in the case of overlapping classes. Also, it is very crucial to select optimal values of hyperparameters for efficient performance. To find a suitable kernel function for the application is a very difficult task many times.

2.3.1.2 Artificial Neural Network

Artificial Neural Networks are the foundation of DL technologies. ANNs resemble the brain since they are obtained by the interconnection of many simple units, called neurons as shown in Figure 2.2. The brain receives the stimulus from the outside world, does the processing

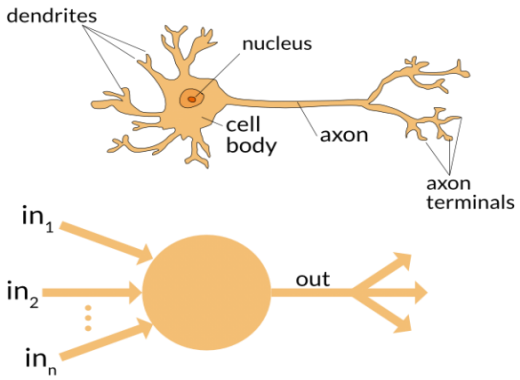


FIGURE 2.2 Biological Neuron and Basic ANN Architecture

on the input, and then generates the output. As the task gets complicated, multiple neurons form a complex network, passing information among themselves.

Various ANN architectures have been developed during the years: the simplest one is the so-called Feedforward Neural Network (FNN) where each neuron is connected to all the neurons in the previous layer, defining a directed graph without any loops. Figure 2.3 shows multilayer FNN where there is more than one hidden layer in between the input layer and output layer.

In recent years, more complex networks, called Convolutional Neural Networks (CNNs), have gained more and more popularity thanks to their achievements in Computer Vision. CNN's exploit a multilayer structure similar to FNNs but where different kind of hidden layers is alternated. In particular, we can distinguish three kinds of hidden layers: (i) convolutional, (ii) pooling, (iii) fully connected.

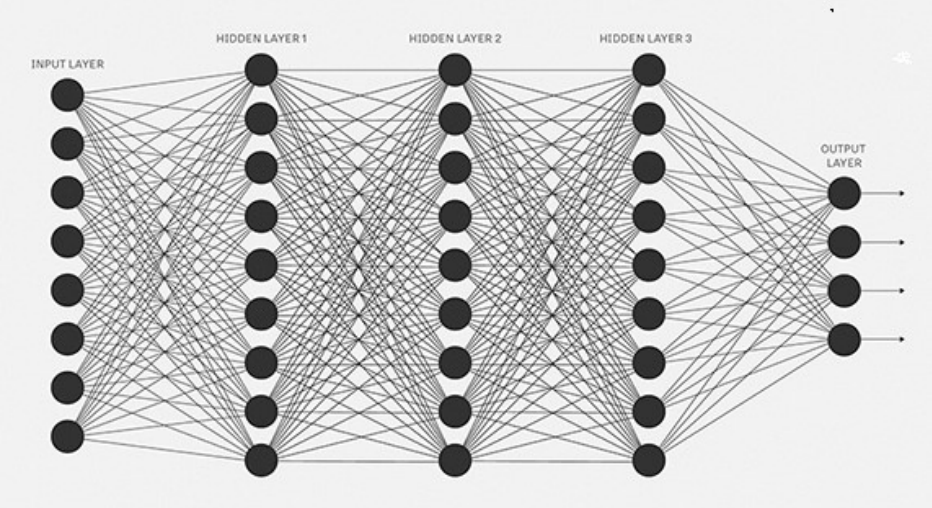


FIGURE 2.3 Multi-Layer Feed Forward Neural Network Architecture

ANNs provide an approximation function given by equation 2.4 of an arbitrary complex continuous function f, where x and y correspond to input and output data respectively.

$$y = f(x; \theta) \qquad (2.4)$$

function f parametrized by a set of coefficients θ (matrices and biases in FNNs, kernel/matrices, and biases for CNNs) [68]. The creation of the predictive model thus requires the estimation of the parameters θ that better approximate the desired output; this is achieved by minimizing a cost function defined according to the output layer properties; common choices are MSE for regression and cross-entropy for classification. Mostly, an algorithm based on gradient-descent is used, based on backpropagation [69].

In the past few years, CNN and other deep learning models have become popular and are used in every field of research like image and video processing, Natural Language Processing (NLP), etc. CNN is explained in detail in the section of classification using deep learning of the same chapter in the thesis.

2.4 Feature Extraction

The first step for managing any video content is video parsing. There are five main components in designing CBVR systems: (1) To divide the video into segments according to its organizational structure; (2) To find appropriate algorithms for feature extraction of the low-level feature vectors; (3) similarity-based searching technique to compare feature vectors of video segments with query features; (4) to answer queries over vast video sequences; and (5) to present the result or list of results.

Video content can be represented by spatial, temporal, or spatial-temporal characteristics. The features of video data in the spatial domain can be extracted from the frames of video, which is based on pixel information in that region and their relationship can be taken as a descriptor. The temporal domain features can be used to partition the video into frames, shots, scenes, and video segments. Video data normally contains audio and visual features such as color, texture, edge information, motion vectors, loudness, pitch, etc.

In the case of textual information present in the video clip, the text data which are continuously being displayed for a certain time gives some important information about what is currently being viewed. This type of information is normally present in broadcasted news videos. Some of the shots in news videos are having text regions that are being displayed for a long duration to give an idea about the current topic of news, place, event, or personality in news, etc. Some of the broadcasted videos contain closed caption (CC) information which is very useful for text query-based video retrieval. The close caption track is having texts to be displayed to viewers in synchronization with videos. Videos which do not contain such text captions which makes retrieval task difficult. Along with video retrieval, computationally efficient indexing of video collection is a very important task to be done for the management of video documents. Video indexing can be done similarly to document indexing. Traditional methods of indexing are not much useful for vast video databases.

Chang et al.[79] proposed text detection mechanism for street view images in their research. To deal with the 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.

2.4.1 Text Feature Extraction

Extraction of features from the video frames is an important part of the 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[86] and recognition [80] for improving accuracy in text extraction from video frames [39][58].

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, advertisements, 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. Literature reports limitations of this approach, producing the alternative Paragraph Vector method. 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. [81] 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.

C. Jawahar et al.[82] proposed a text query based video retrieval approach for English, Hindi, and Telugu Languages for the Indian news video database they created. Regions of the frame of the video are identified where textual information was present to search text in the frame. Frames of video were annotated with the text content identified from the frame. An approach for image-level matching using DTW (Dynamic Time warping) algorithm was proposed for video retrieval. As shown in Figures 2.4 and 2.5 query text-based videos were retrieved from a database of videos which were sorted based on the relevance. Results are shown from video collections in English, Hindi, and Telugu.

Figure 2.4 shows results for Video retrieval based on textual query "Australia" from the video collection. Video retrieval based on textual query "सानिया" in the Hindi language is shown in Figure 2.5 where the news channel video of Hindi news channel was used. The image of the word is matched with the frames using the DTW method for searching the relevant video frame[82].

Tesseract OCR [83] is one of the efficient text OCR used in text retrieval from document images as well as scene text from video [35][82][84]. Tesseract is an efficient and well known optical character recognition engine which is works on various platforms. Tesseract is open source which is released under Apache licence. Tesseract provides support to unicode UTF-8 text format. Tesseract is able to recognize many languages all over the world. Many times to obtain good OCR outcomes, the quality of the image provided as input to Tesseract engine is required to be improved.

Tesseract uses Leptonica library for opening input images. Tesseract 4 which is the newer version of the engine available to be used, adds a new neural net called LSTM based OCR engine. The new Tesseract engine focused on line recognition. The newer version also provides support to the legacy Tesseract OCR engine of Tesseract 3 which works by recognizing character patterns. Tessdata files are available to be used with tesseract ocr engine in languages like Arabic, Bengali, Canadian, Devnagari, Greek, Gujarati, Gurumukhi, Hindi, Kannada, Malayalam, Tamil, Telugu and other popular and widely spoken languges. Tesseact engine can also be trained using custom dataset for specific language to be used for OCR purpose.

FIGURE 2.4 Video Retrieval based on Textual Query "Australia"[82]

FIGURE 2.5 Video Retrieval based on Textual Query "Sania" in the Hindi language[82]

2.4.1.1 Challenges with Text Extraction in Indian Language Video

Often the quality of the video data affects the text detection algorithm performance. Researchers have explored the text detection approach for Hindi, Bangla, Telugu languages from the videos but there is no detailed information or study found on for challenges in detecting in natural images and videos with Indian language text. Languages like Hindi, Gujarati required to be treated differently than the English language as the average length of connected components for them is considerably different from that of the English language.

In poor quality videos, it is very difficult to detect words and a lot of extra characters can be generated along with textual contents. Also, the video with poor resolution creates problems while detecting words from the scene text. Either the words are partially detected or wrong words are interpreted due to missing components. Also, to select the correct word, it is required to determine which frames to consider to extract the word from the sequence of frames of a particular shot containing the intended word. This type of issue is more critical with broadcasted videos [82].

2.4.1.2 Text pre processing

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. The next step is Term weighting which assigns weights to terms according to their importance in the document, in the collection, or some combination of both.

In natural language processing, sometimes it becomes necessary to recognize that “asked” and “ask” are two different tenses of the same word. So the basic idea is to derive core or root words from the different forms of the word. Stemming is the process of finding the root word of the given word, for example, walking, walked, etc. stem to walk. Stemming is normally used to reduce dictionary size and improve the searching performance of words in the document collection. Porter stemmer[84]–[86], Lovins[87] are some of the popular stemming algorithms for the English language, Dutch language, Farsi language, and other foreign language stemming. Stemmers for other popular foreign languages also exist.

2.4.1.2.1 Morphological Analysis and Stemming

A morpheme is the basic meaningful morphological unit in a language that cannot be divided further to get the root of the word. A word and a morpheme are different in the way that morpheme sometimes does not stand alone whereas a word does. A word is normally made up of one or more basic units of language i.e., morpheme.

To reduce the dictionary size for the document indexing, words are truncated. There are various methods of truncating words like Rule-based methods, statistical methods, and hybrid methods.

Rule-based methods are based on truncating a word at the n^{th} symbol i.e., keep n letters and remove the rest. Words with the length shorter than n are kept as it is. In this method, chances of over stemming increase when the word length is small. In statistical methods of truncating words, most methods work based on the idea of removing the affixes but after implementing some statistical procedure. While the hybrid methods are mostly combined approach with inflectional and derivational stemmer as shown in Figure 2.6. Derivational stemmer may change the grammatical category of the word. For e.g., જવાબદારી *javābdārī* 'responsibility' derived from જવાબદાર *javābdār* 'responsible')

Whereas in Inflectional Stemmer, Inflection may occur by adding a suffix or affix to the terms. For example, the suffix ી (*-ī*) should be stripped off for verbs (as in case of કરી *karī* 'did'), but not for nouns (as in case of ઈમાનદારી *īmāndārī* 'honesty').

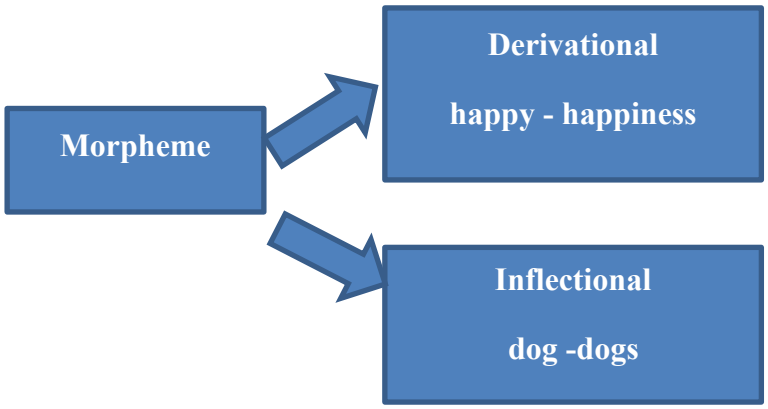


FIGURE 2.6 Types of Morphological operations

Along with other regional languages, in Gujarati language processing's major contribution is done by TDIL [88] program through the government of India. Researchers have worked on different stemming techniques for the Gujarati language on either EMILLE [89] text corpus or their dataset [90]–[93]. There is a lot of scope of research in this field for Gujarati language text processing. Literature review of stemming algorithm as well as other morphological analyzers of mainly 'Gujarati', 'Hindi' and 'Assamese' language was done extensively as few notable works are listed in Table 2.1. As Gujarati language standard benchmark datasets to work for scene text recognition or image/video retrieval is not available, it was a challenging task to generate a dataset as well as getting good performance with it.

TABLE 2.1 Literature review of Stemming Algorithm/Morph Analyzer

Proposed Method	Year	Author	Language	Dataset	Size	Accuracy
Inflectional and Derivational	2019	H. Patel & B. Patel[94]	Gujarati	Gujarati words	2197	98.33
Rule-based Suffix stripping	2016	C. D PATEL[93] et al.	Gujarati	EMILEE corpus	1099	-
Knowledge-based hybrid method & statistical hybrid method	2015	J. Baxi et.al	Gujrati	Gold dataset	200 nouns, 200 verbs 100 adj.	92.34% and 82.84%
Longest matched, Rule-based	2012	J. Ameta, N.Joshi, I. Mathur [90]	Gujarati	EMILLE corpus	3000	91.5%
Brute Force technique Suffix Stripping	2012	Upendra Mishra, Chandra Prakash [95]	Hindi	Hindi Words	15000	91.5%
Lookup method, Suffix stripping	2012	Navanath Saharia [96]	Assamese	EMILEE corpus	123753	82%
Hybrid Inflectional and Derivational Stemmer	2011	K. Suba, D. Jiandani, P. Bhattacharyya [91]	Gujarati	EMILEE corpus	8,525,649	90.7%

2.4.2 Deep learning approach for Feature Extraction

Deep learning is one of the parts of machine learning. In deep learning techniques, a model learns to perform classification or feature extraction kind of tasks straight from input images, text, or sound. Deep learning is typically realized using a neural network architecture. The word “deep” in the deep learning architectures is used to refer to the greater number of layers in the network. More layers in the network mean the deeper the network. Traditional neural networks contain only 2 or 3 layers, while deep networks can have hundreds.

Image retrieval using deep learning with image query is a widely explored approach so far[98] [99]. Nowadays, researchers are paying attention to large-scale retrieval of video or images using images as queries [18][100]. Whenever it is required to deal with a large collection of images or video frames, a normal system cannot give good performance with state-of-the-art methods of retrieval. Due to 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.

Feature extraction task is tremendously time-consuming for large-scale retrieval from video data. In contrast, methods using automatic feature extraction are plagued by information loss, and every so often leads to poor prediction capabilities. Comparatively sophisticated methods of feature extraction have been proposed in the past few years to overcome the aforementioned problems.

An autoencoder is a particular Artificial Neural Network (ANN) that is trained to reconstruct its input. Usually, the hidden layers of the network perform dimensionality reduction on the input, learning relevant features that allow a good reconstruction. Moreover, deep autoencoders exploit multiple non-linear representational layers that learn complex hierarchical features from the data with high informative content.

In particular, the convolution operation is highly effective at extracting local features from images: actually, Convolutional Neural Networks (CNNs) are extensively employed for problems like object localization and recognition [11], face recognition [12], and text recognition [28]. For this reason, the proposed model will be based on CNN's

Mühling et al.[18] 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. [40] proposed DELE local feature descriptor for image retrieval tasks at a 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. The 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, [100] 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. [101] 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.

Object detection and tracking algorithms have applications in image and video security where features extracted from images or videos are used. Various classifiers are applied for the image classification based on identified objects. Also, object counting can be applied once the objects are identified in the video. YOLO[102][103] is the popular algorithm for object detection tasks and also used in combination with the GMM model for feature extraction and classification tasks [104][105].

2.4.2.1 Convolutional Neural Networks

Convolutional neural network (CNN) is one of the variants of traditional artificial neural network mostly used feature extraction task from the image or video data as well as for the classification of input data of higher dimension. The convolutional neural network is also called ConvNets in many works of literature. The basic idea of convolutional neural networks was first proposed in 1989s by Yann LeCun [106] and was further improved by various researchers.

The initial version of CNN was named LeNet from the name of the researcher LeCun. An earlier version of CNN was designed to recognize handwritten digits [106]. Due to the ability of CNN to read zip codes written by hand on postal envelopes and handwritten digits on checks in banks, the utilization of CNNs was limited to applications in banking and postal services. The network was also facing the problem of scalability due to the limitation of computational power and the small size of the training dataset.

The convolutional neural network suddenly reappeared in 2012 and become a very famous architecture with high computation power and a large dataset for image recognition tasks with the invention of the Alexnet model [107]. The Alexnet model is a very successful model due to the flexible design of its layers and also with more layers of neurons being added, it ultimately boosts the learning capacity of the network. Graphical Processing Units (GPU) played a vital

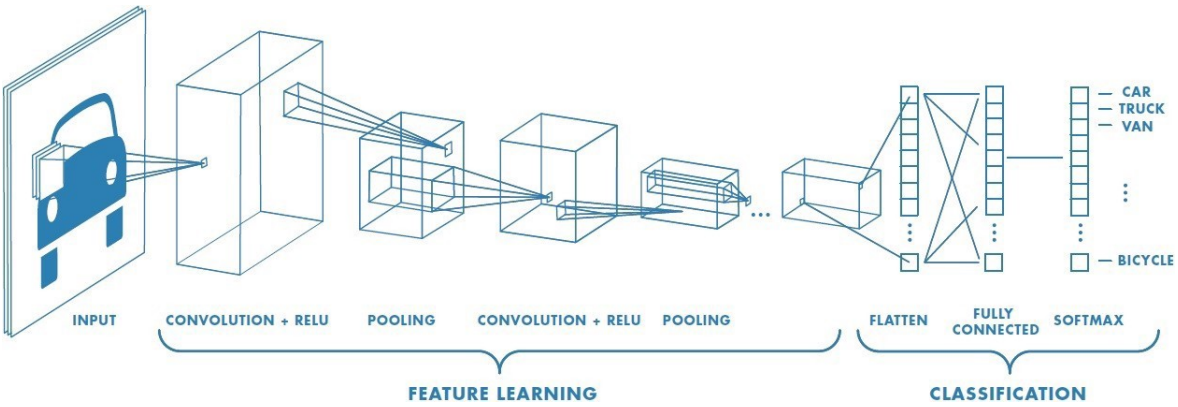


FIGURE 2.7 CNN Architecture for object category classification[108]

role in the success of deep learning architectures designed so far[109][110]. A convolutional neural network (CNN) is specifically designed to identify patterns in the input dataset and make sense of them. Due to the usefulness of CNN in pattern detection and recognition, it has become an important tool for image and video analysis.

One of the well-known architectures of CNN for object recognition tasks is shown in Figure 2.7. As shown in the Figure 2.7, architecture is divided into main two parts based on the task such as feature learning and classification. Layers used in feature learning are repeated combinations of the convolutional layer, activation layer, and pooling layer. In the classification layer, mostly output of the previous layers is flattened and feed into the fully connected layer which is mostly the last layer of the network. A fully connected layer is

having the number of neurons equal to the number of classes or categories into which data is classified at the end.

CNN can be visualized as a sequence of layers of artificial neurons where every layer is accomplished unique functions on the input data given to the network. The main layers used in designing the CNN architecture are as follows:

- The input layer is designed to input the raw data fed to the model,
- Convolutional layer – as the name suggests, this type of layer computes a dot product between the input image patch and the filters, and gives the output volume based on input,
- Activation function layer – the output of the convolution layer is fed to this layer which applies activation function on elements given as input,
- Pooling layer- this layer is designed to make the output of the previous layer memory-efficient which in turn reduces the cost incurred in computation,
- The fully-Connected layer is mostly used as the last layer which gives computed 1-D array class scores from the input received.

From experiments mentioned in an article by Hubel et. al., [111] it is understood that the visual cortex of the brain of animals where the cerebral cortex resides and processes the visual information is a complex network made of cells. Each of the cells of the network is sensitive to the receptive field which is a tiny sub-region of the visual fields. Cells are of two types termed Simple and Complex cells. Simple cells mainly extract features and cells which combine several such local features are called complex cells [112]. This biological structure is used to design CNN, which works similarly by extracting the features from the given input and performs the classification. The reason for the popularity of deep CNN architecture is the automatic feature extraction and classification as opposed to the state-of-the-art methods where manual features are extracted and fed to the classifier used in the classification model.

2.4.2.2 Autoencoders for Unsupervised Learning

An autoencoder is a special kind of artificial neural network which is used to learn data encodings in an unsupervised manner efficiently. An autoencoder learns a representation for a set of data by training the network to ignore signal “noise” and also reduces dimensionality. An autoencoder architecture is known for its applications in the field of data compression, information retrieval, etc. Autoencoder is the type of neural network which is used for

discovering structure within input data which is useful in the development of a compressed representation of the given input. Variants of the autoencoder models exist and can be distinguished based on their ability to extract meaningful information in different applications and the nature of data it is used for.

The architecture of a basic autoencoder comprises three main components. First, the encoding architecture comprises of series of layers with the number of nodes in decreasing order and eventually reduces to a latent view representation. The second component is Latent view representation which is the lowest level space in which the inputs are reduced and information is preserved. The third component is decoding architecture, which is the mirror image of the encoding architecture but in which the number of nodes increases in every layer and ultimately outputs a similar input. typically, the biggest challenge when working with autoencoders is getting your model to learn a meaningful and generalizable latent space representation. The training then involves using back propagation to minimize the network's reconstruction loss.

Denoising or noise reduction is the process of removing noise from a signal. This can be an image, audio, or document. You can train an Autoencoder network to learn how to remove noise from pictures.

Convolutional Autoencoder (CAE) is a variant of the CNN model which are used as the tool for unsupervised learning. They are generally applied in the task of image reconstruction to minimize reconstruction errors by learning the optimal filters. Once they are trained in this

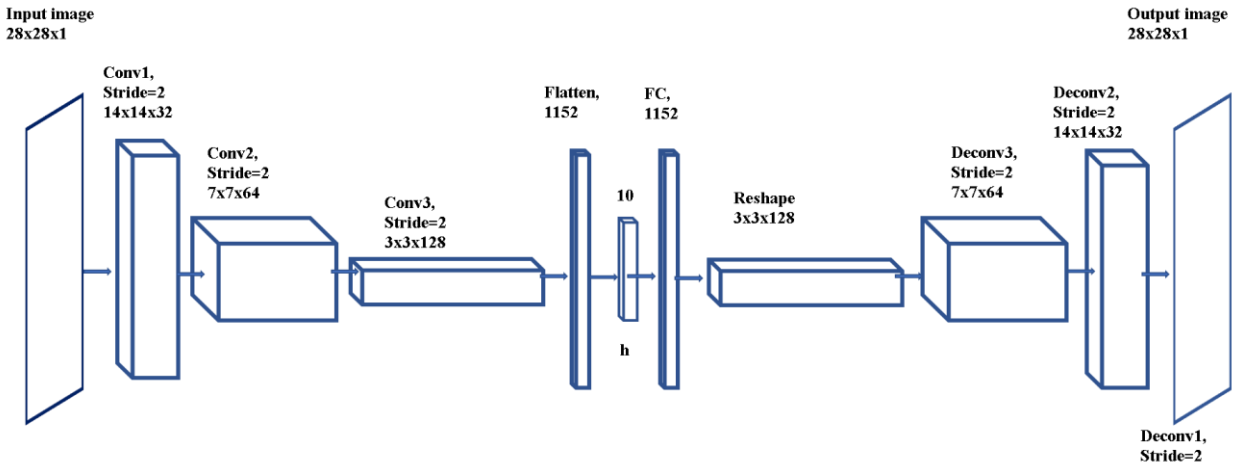


FIGURE 2.8 Convolutional Autoencoder Architecture

task, they can be applied to any input to extract features. Convolutional Autoencoders are general-purpose feature extractors differently from general autoencoders that completely ignore the 2D image structure. In autoencoders, the image must be unrolled into a single vector and the network must be built following the constraint on the number of inputs.

The architecture of the convolutional autoencoder for input images with size 28x28x1 is shown in Figure 2.8. The input image is reduced to size 14x14x32 after convolution filters from the conv1 layer is applied on it. The convolution filters are basically extracts useful features as well as compress the input image. The size of stride is kept 2 for this layer. Further, in conv2 layer image reduces to 7x7x64 with stride 2. In the last layer conv3 with stride 2, the image is further reduced to 3x3x128 size.

Finally, the image is flattened to 1152 neurons followed by the last layer of the encoder which generates an intermediate representation of the input image. Similar steps in reverse order are applied with deconv layers to reshape the image at each layer by reconstructing the image at each layer. In the decoding part of the architecture, the Upsampling layer is used for 2D inputs. In the upsampling layer the rows and columns of the data are repeated based on upsampling factor which are given by number of rows and number of columns to repeat respectively. In the final step, the output image is reconstructed to original size 28 x 28 x 1 from the input size 14x14x32 which is not exactly as input but retains the most useful information of the original image.

In one of the variants of CAE[113], convolutional layers are stacked on the input images to extract hierarchical features. In the next step, all units in the last convolutional layer are flattened to generate a vector which is generally followed by a fully connected layer with n units which is called the embedded layer. The input 2D image is thus transformed into n-dimensional feature space. To train it in the unsupervised manner a fully connected layer can be used along with convolutional transpose layers to transform the embedded feature back to the original image. The parameters of the encoder and decoder are updated by minimizing the reconstruction error.

2.5 Existing Content based Video Retrieval System

It is very clear from the Literature reviewed that content-based video retrieval is the active field of research and there exist many research problems that can be explored for further

research. Alan F. Smeaton (2007) divides the traditional VR approaches into five categories. Each approach has its advantages and disadvantages.

The first category is using Metadata and browsing Keyframes. In this technique, metadata is used to search the video. Metadata includes characteristics such as video title, date, actor(s), video genre, running time and file size, video format, reviews by users and user ratings, copyright, and ownership information. Each of these metadata fields is searchable and most systems are coupled with a keyframe that allows users to preview the video content itself visually.

The second category is using text for video searching. This type of system uses spoken dialogue in the video for assistance. If the user is searching for a video that consists of spoken commentary, for example, this can be things like nature documentaries or TV news broadcasts, then this method can be used for retrieval. Here the spoken dialog may reflect the contents of the video itself. In such cases, a search through the video can be considered as a good video search. Spoken dialog from a video can be obtained from Automatic Speech Recognition (ASR) and text from the video can be obtained from Optical Character Recognition (OCR). The main disadvantage of text-based video retrieval is that all video contents have no associated text. All information needs cannot be expressed as a set of text queries.

The third category is Key frame matching, in which the representative frame of the shot known as keyframe is used for retrieval. This type of video retrieval can be called Content-Based Image Retrieval (CBIR). The image which is to be searched is used as a query. This que ry is compared against the video key frames from the video library. The key frame-based matching is good for video searching but the user needs to be very precise about the visual component.

The fourth category is based on semantic features for video retrieval. This type of system uses semantic features for search. Semantic feature refers to high-level or mid-level features that convey semantic contents. This can include indoor, outdoor, moving car and extracting these kinds of features itself is a challenging task. The automatic extraction of low-level features like color and texture is not so difficult, compared to it.

The fifth category is using object-based video retrieval. Here the retrieval is based on objects. Retrieval of video based on an object is a straightforward task theoretically but practically it

is not so simple. If an object is viewed at different angles then it could be seen in different shapes, colors, and textures due to different lighting conditions and shadows. Some examples of such queries are boat and car. Research in the field of object-based shot retrieval has been very little as well. So, there is no experimental result to support it. This technique has come up very recently and so more work can be done in this field.

Kannao et al.[81], 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 approach proposed is claimed as superior on news videos acquired from three Indian English television news channels along with benchmark datasets.

Chang et al. [25] proposed a text detection mechanism for street view images in their research. To deal with the 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. They claimed that their experimental results of the proposed scheme are good in processing complex streetscape also.

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 leads to rich, dense, and meaningful semantic associations between the text and video and improves the accuracy of the 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. The proposed algorithm was demonstrated on a publicly available tennis dataset comprising of tennis match videos from Olympics games.

A. Mishra et al. [27] proposed an approach for text to image retrieval tasks using the text available in images. A 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 broadcasts which helps users of pervasive devices to browse quickly into news archives; their solution integrating many aspects such as summarizing, indexing textual content, and on online recognition of the handwriting. Firstly, the summarizing process is to accelerate

the video content browsing based on a genetic algorithm. Secondly, the indexing process, which operates on video summaries based on text recognition.

Researchers have contributed many methods for semiautomatic or automatic indexing of video documents. Methods for video indexing are either based on one feature such as visual, audio, textual, etc., or multimodal. In the multimodal indexing method generally, more than one representative feature of the video is combined. Kulkarni et. al., [114] proposed a method using video clips as input queries for discovering the temporal patterns in the video contents. The discovered temporal patterns are applied to achieve efficient indexing along with the sequence matching technique to increase the retrieval accuracy with the reduction in the computation cost.

Padmakala et al.[115] have proposed an algorithm for retrieving video for a given query, the raw video data is represented by two different representation schemes, Video Segment Representation (VSR) and Optimal key Frame Representation (OFR) based on the visual contents. All the features are used collectively and stored in the feature library. For the query video clip, the aforesaid features are extracted and compared with the features in the feature library. The comparison is achieved via the feature weighted distance measure and similar videos are retrieved from the collection of videos.

Kong Juan and Han Cuiying[116] have investigated the significant video retrieval-based technologies. Based on their studies, a content-based video retrieval system has been constructed according to the system design requirements. Superior video analysis and retrieval capabilities have been achieved by splitting the system into video preprocessing and video query subsystems. Patel et al. [117]proposed key frame extraction using entropy and edge features of video for developing proposed video retrieval systems.

Gao Haolin and Li Bicheng[118] have proposed a novel video signature based on data count fluctuation independent upon video frame content, just relevant about frame data count for the compressed domain of MPEG video, which first calculates the data count for each frame on the compressed domain as signature, then slide the window in the step of Group Of Picture (GOP) length after the alignment of I frame, and then compare the difference of data count fluctuation between the query clip and the target video, at last, the similarity result is given. This signature could be used both in video database retrieval and online video matching.

2.5.1 Performance Evaluation Measurement

Performance of classification task can be measured mostly by confusion matrix, accuracy and also using the measures like precision and recall. While the information retrieval system can be evaluated mainly by two different ways of retrieval, the first is the evaluation measure for unranked retrieval and the second is the evaluation measure for ranked retrieval. Each of them is discussed in the following sections.

The confusion matrix is one of the useful and widely used metrics for the performance evaluation of a classifier. The confusion matrix generally gives the idea about the capability of the classifier to correctly recognize data samples of different classes. The confusion matrix for the binary classification problem is given in Table 2.2. The meaning of the term True Positive is that the tested sample data is actually part of positive class and also predicated as positive, while the term False Positive means that the sample is taken from the positive class, but it is classified as negative. Similarly, the term False Negative means the sample is taken from negative class but classified as positive and the term True Negative indicates the sample is taken from negative class and classified as negative.

TABLE 2. 2 Confusion Matrix for binary classification

	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

As the number of classes is increasing, it is very difficult to visualize the whole confusion matrix, but the details provided by the confusion matrix have been used for another performance parameter that better represents the performance of the classifier model. These parameters are accuracy, precision, recall, and F1 score.

2.5.1.1 Accuracy

The term accuracy is defined as “the ratio of correctly recognized samples to the total number of tested samples”. This parameter works well when the numbers of samples in each class are balanced. For example, for one class A, the number of samples are 95% and another class B has only 5%, for this, the training accuracy is very good, but the testing time, if class A

has 60% data and class B has a 40% than accuracy is falling drastically. The Accuracy measurement formula is given by equation 2.5.

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP}+\boldsymbol{TN})}{(\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN})} \quad (2.5)$$

Where TP = True Positive,

TN = True Negative,

FP = False Positive,

FN = False Negative

2.5.1.2 Evaluation of information retrieval system

The usual method for the evaluation of information retrieval systems relies mostly on the concept of relevant and non-relevant documents. Concerning the user input, a document can be viewed as a relevant document or a non-relevant document. The test document collection and suite of information needs have to be of a reasonable size for the evaluation. Relevance can reasonably be thought of as a scale, with some documents highly relevant and others marginally. For the simplicity of evaluation, one can consider the document as relevant if the relevancy is more than the specified threshold on a scale of say one to five.

For the unraked retrieval, Precision and Recall can be used as measures of evaluation. In the context of information retrieval, Precision can be defined by the given equation:

$$Precision(P) = \frac{\text{Number of Relevant items retrieved}}{\text{Retrieved items}} \quad (2.6)$$

Recall can be defined similarly in terms of relevant documents retrieved for unranked retrieval as given by the equation:

$$Recall\ (R) = \frac{\text{Number of Relevant items retrieved}}{\text{Relevant items}} \quad (2.7)$$

Precision, Recall, and F1-score are measures used mostly in unranked retrieval tasks and also not much used for ranked retrieval tasks. For ranked retrieval of items or documents, Mean Average Precision (MAP) is a widely used measure of evaluation.

2.5.1.3 Precision@K(P@k)

Only Recall is no longer a useful metric for many modern information retrieval systems. As many queries have lots of relevant documents and not all documents are of interest to the user. Precision@k is a useful metric in IR systems. Using P@ke.g. P@5 or Precision@5

corresponds to the number of relevant items among the top 5 items retrieved). If the query has fewer relevant documents than k, even a perfect system can have a score of less than 1. So it is very important to choose a value of k. This metric is useful in the sense that only top k documents are needed to be examined whether they are relevant or not.

2.5.1.4 Average precision

Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. Let consider the rank position of each relevant documents as $K_1, K_2, \dots, K_R$ and next computer Precision@K for each $K_1, K_2, \dots, K_R$ [119]. By computing, a precision@k average precision can be calculated as,

$$\text{Average precision} = \text{average of P@K} \tag{2.8}$$

2.5.1.5 Mean Average Precision (MAP)

Among all evaluation measures used in ranked information retrieval, MAP has been shown to have especially good discrimination and stability. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is if the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \dots, d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then MAP can be defined by equation 2.9:

$$MAP(Q) = \frac{1}{Q}\sum_{J=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.9}$$

$$MAP(Q) = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{2.10}$$

Mean Average Precision can also be determined using Average Precision averaged over a set of queries Q as given by the equation 2.10.

2.6 Summary of the Literature Survey

As per the analysis of the literature survey done so far, it is observed that very less or negligible work for content-based video retrieval is done using a text query-based approach

for any of the languages spoken or written in India. Due to this survey and research gap found, research work is proposed using ‘Gujarati’ text query-based video retrieval for news videos in chapter 4 of the thesis.

Also, the kind of NLP library which is explored long ago for English language, similar support of the library is not available for very popular Indian regional languages like ‘Gujarati’ or if available it is not publicly accessible or produced anywhere.Also, content-based video retrieval approaches mostly process all frames of video or process key frames but do not try to reduce overall training time by reducing the number of frames to process. For example, finding and ignoring advertisement frames while processing videos can reduce overall processing as proposed by research work produced in this thesis chapter 3.

Deep learning-based video retrieval is used recently for large-scale video retrieval based on image query. But unsupervised learning based CBVR is not proposed so far where the labeled data is unavailable and it is tedious to manually label them for supervised learning. Due to this finding and to compare the performance of propose CBVR text-based approach with the deep learning-based approach, in chapter 5 of the thesis deep learning based CBVR is proposed where unsupervised learning is employed through autoencoder architecture proposed.

2.7 Challenges with CBVR

As the result of literature work carried out, the following challenges are found in implementing CBVR with ‘Gujarati’ language video retrieval with a text query-based approach.

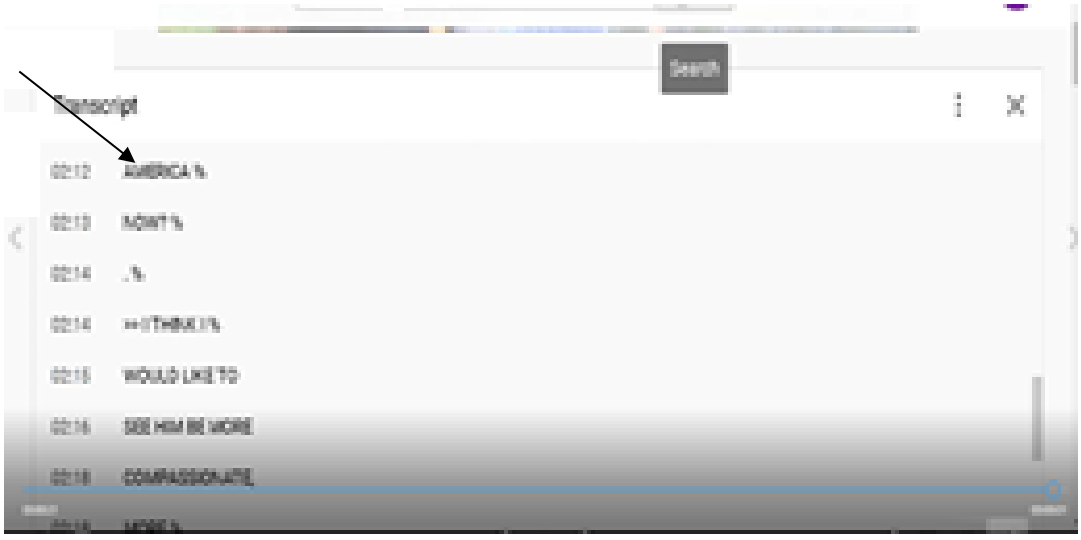
- In India, there are no specific laws for multimedia content used in news video broadcasted or any other multimedia videos.
- No Closed Caption details available[120]- which is mandatory in many countries. Closed caption details are very useful for people with hearing aids. In Figure 2.9(a), it can be seen that the closed caption details are available with English language news channel which is not available in Figure 2.10(a) of video from the news channel in India.
- No Transcriptions available with the videos in India while it is mandatory with multimedia videos in other countries. In Figure 2.11 (b), an English news channel video snapshot is shown with transcription details available. As opposed to it, in Figure 2.11(a) transcription is not available with news channels in India.

- No fixed format of displaying different news details like running news stories, headlines, advertisements or other information, logo, etc. in various news channels video in India. Also, Font size, style, orientation, etc are very much different in different channels.

(a)

CC/SUBTITLES

TRANSCRIPT OF NEWS STORY



(b)

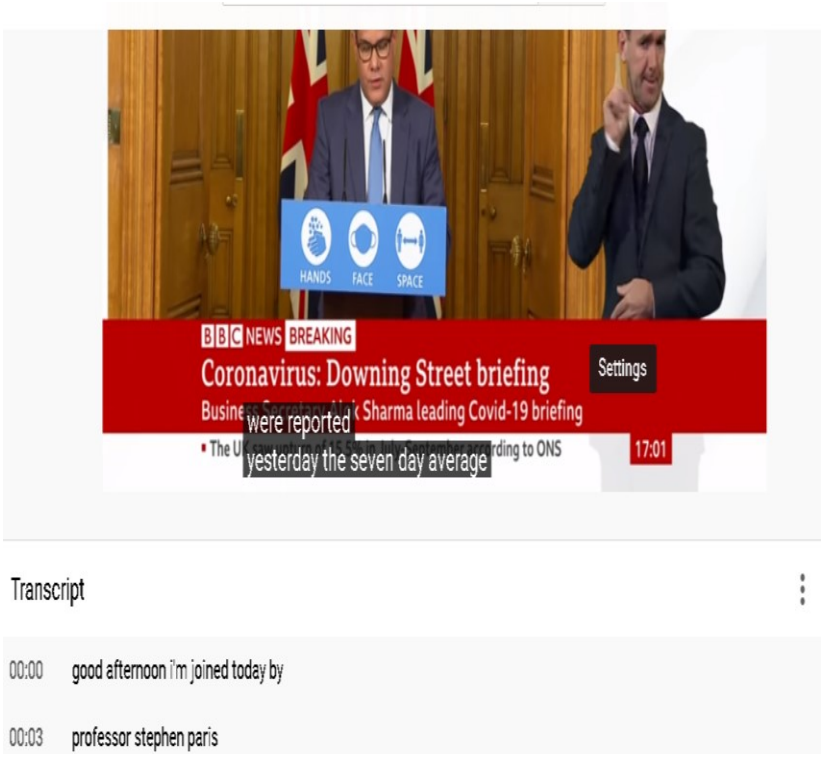
FIGURE 2.9 Snapshot of English language news channel video of non-Indian news channel (a) showing closed caption details(b) showing transcript generated with video

(a)

(b)

FIGURE 2.10 Snapshot of ABP NEWS Gujarati news channel (a), (b) showing no closed caption details available

(a)



(b)

FIGURE 2.11 Snapshot of news video of (a) News channel in India with no transcript (b) BBC News channel which shows a transcript of the video

- The news may not contain the face of the person for which the new story is running as shown in Figure 2.12(a). Sometimes faces are not recognized which can be seen clearly in Figure 2.12(b).
- As opposed to face-based retrieval, text-based retrieval can be more accurate as text information appears as an overlay which is a meaningful feature for retrieving news stories.

The above-mentioned challenges present in the retrieval task motivated us to take up the proposed research definition.

(a) (b)

FIGURE 2.12 Video Frame from Different News Story

2.8 Definition of the Problem

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

The main objective of using the text feature was to simplify the searching interface for the common man of the local region who is not having the skill or knowledge of any other language other than the mother tongue or local region language. Also, an alternative approach for CBVR is proposed with the use of unsupervised deep learning based autoencoder architecture to explore the possibility of getting faster retrieval with more accuracy on the same dataset.

Chapter 3

Key Frame Extraction and Advertisement Detection

3.1 Introduction

Content based video retrieval using a text-based query is explored by researchers all over the world. The literature confirms that most of the work has been carried out for English language text query based retrieval tasks. Other than English, video scene text spotting, as well as recognition task, are performed for the Arabic language also. Lecture video retrieval based on video text and audio information for Chinese and Korean languages has been found in the literature. For popular Indian languages like Hindi, Bengali, and Malayalam, very few research articles are found where information retrieval from the video has been carried out.

The proposed Content-based video retrieval task is divided into main three important tasks. First is Key frame extraction from the input video. The second task proposed here is advertisement classification and removal as well as feature extraction. The third task in the proposed approach is indexing and retrieval of documents.

In this chapter of the thesis, the first two important tasks are described with algorithm and experimental details as well as results. The dataset used for the proposed task and properties of the dataset is described first followed by key frame extraction and advertisement classification task.

3.2 Dataset

Our proposed approach for Content based video retrieval for ‘Gujarati’ language video using text query is discussed in chapter 4 of the thesis. For this task, there is no benchmark dataset for news videos in any of the Indian languages including ‘Gujarati’ is available for research

work. For the Gujarati language, any application using information retrieval with text query is not found so far in any of the literature, and also no dataset is available for the same. Due to this fact, the dataset of News Videos of Gujarati News Channels is created from different news channels ' broadcasted videos of 'Gujarati' language for the task of news story retrieval based on 'Gujarati' query text. News channel video of continuous two days is recorded for three different channels ETV Gujarati, VTV NEWS, Sandesh News channels for research. Dataset properties are listed in Table 3.1.

TABLE 3.1 Dataset Property

Property Name	Value
Size	90 GB total
Width	490/480/450/640
Height	360
Frame Rate	25fps
Bits Per Pixel	24
Video Format	'RGB24'
Extension	.mp4

3.2.1 Gujarati Language

In India, the Gujarati is one of the popular local languages spoken mostly in the Gujarat state that is named after the Gujar or Gurjar people. Gujarati is also the official language in the union territory of Diu, Daman and Dadra, and Nagar Haveli. The language is a part of the Indo-Aryan family, which are members of the Indo-European dialects. It is the 6th most widely spoken language of India as of 2011. The Government of India declares 23 different languages officially and the Gujarati is one of them. The Gujarati language remains at the 26th rank in the most widely spoken local language all over the world and about 55.5 million people speak in Gujarati all through the world. Gujarati is a more than 700 years old language.

The greater part of the Gujarati word is obtained from the Sanskrit language. There is a very small vocabulary of Sanskrit-derived words are available in the Gujarati language. Gujarati was the native language of the "Father of India", Mohandas K. Gandhi, the "Iron man of India", Sardar Vallabhbhai Patel, and many other famous personalities in and outside India.

Gujarati script is written and read from left to right. It is an abugida, that is, each consonant letter contains an intrinsic vowel. There are forty-seven total characters used in the language of which eleven are vowels and the remaining are consonants. Ten of these vowels have two

structures, a free structure is used when the vowel isn't gone before by a consonant, and a diacritic structure, which is composed above, underneath, or close by a first consonant to adjust the characteristic vowel. One vowel, 'A', just has an autonomous structure; this is the

TABLE 3. 2 Gujarati Consonants

ક	ખ	ગ	ઘ	ઙ	ચ	છ	જ	ઝ	ઞ
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña
[kə]	[kʰə]	[gə]	[gʱə]	[ŋə]	[ʧə]	[ʧʰə]	[ʤə]	[ʤʱə]	[nə]
ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન
ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na
[ʈə]	[ʈʰə]	[ɖə]	[ɖʱə]	[ɳə]	[tə]	[tʰə]	[də]	[dʱə]	[nə]
પ	ફ	બ	ભ	મ	ય	ર	લ	વ	
pa	pha	ba	bha	ma	ya	ra	la	va	
[pə]	[fə]	[bə]	[bʱə]	[mə]	[jə]	[rə]	[lə]	[ʋə]	
શ	ષ	સ	હ	ળ	ક્ષ	જ્ઞ			
śa	ṣa	sa	ha	ḷa	kṣa	gña			
[ʃə]	[ʃə]	[sə]	[ɦə]	[ɭə]	[kʃə]	[gnə]			

the vowel that is inalienable in a consonant letter, so it shouldn't be composed independently when following a consonant. Table 3.2 shows 36 consonants of Gujarati script called '*Vyanjan'* and Table 3.3 shows 13 vowels called '*Swar'*.

TABLE 3. 3 Gujarati Vowels

અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ
a	ā	i	ī	u	ū	ṛ
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]
એ	ઐ	ઓ	ઔ	અં	અઃ	
e	ai	o	au	aṃ	ah	
[e/ɛ]	[əy]	[o/ɔ]	[əʊ]	[əŋ]	[əh]	

TABLE 3.4 Gujarati consonant 'ક' with all 'Maatras'

ક કા કિ કી કુ કૂ કે કૈ કો કૌ કં કઃ

The vowels are also used as a modifier for consonants called '*Maatras*'. Different frequently used modifiers generally applied to consonants in the Gujarati language are shown in Table

3.4 with the use of consonant ક along with modifiers. Table 3.5 presents Gujarati language Numerals or digits called '*Ank'* from 0 to 9.

TABLE 3.5 Gujarati Digits

૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
મીંડું	એકડો	બગડો	ત્રગડો	ચોગડો	પાંચડો	છગડો	સાતડો	આઠડો	નવડો
mīṃḍuṃ	ekaḍo	bagaḍo	tragaḍo	cogaḍo	pāṃcaḍo	chagaḍo	sātaḍo	āṭhaḍo	navaḍo
0	1	2	3	4	5	6	7	8	9

TABLE 3.6 Some Gujarati conjunct consonants

ખ્ખ	ગ્ક	ઘ્ક	ચ્ક	ઞ્ક	ણ્ક	ત્ક	ધ્ક	ન્ક	પ્ક
khkha	gka	ghka	cka	ñka	ṇka	tka	dhka	nka	pka
બ્ક	ભ્ક	મ્ક	ય્ક	શ્મ	શ્લ	ષ્ટ	શ્ચ	ઙ્ક	ક્ર
bka	bhka	mka	yka	śma	śla	ṣṭa	śca	ṅka	kra
ખ્ર	ટ્ર	ર્ક	શ્ર	ત્ર	દ્ર	હ્ર	હ્ય	હ્મ	દ્વ
khra	ṭra	rka	śra	tra	dra	hra	hya	hma	dva
દ્ધ	દ્મ	દ્ય	ટ્ટ	ડ્ડ	ઠ્ઠ	ઢ્ઢ	ત્ત	દ્દ	
ddha	dma	dya	ṭṭa	ḍḍa	ṭṭha	ḍhḍha	tta	dda	

TABLE 3.7 Sample Gujarati language sentence text

સરકારે કોરોના વાયરસ સંક્રમણ જેવી મહામારીથી લોકોને

બચાવવા માટે કોરોના વેક્સીન માટે જોગવાઇ કરી છે.

The conjunct consonants are shown in Table 3.6. Consonants are requested by the standards of articulatory phonetics, that is, they are written in the way they are pronounced. There are numerous other conjunct symbols utilized in Gujarati composing. Mostly the characters which have a vertical line in their composition are used to generate the conjunct consonants. The general principle for composing conjuncts is that those letters lose the vertical line at the beginning or average position, and just the last character in the group holds it. In Table 3.7, a few Gujarati sentence text is displayed for reference.

3.3 Pre-Processing

A video contains much useful information in different forms such as visual information, meta data information, scene text, temporal information, etc. Feature extraction from the

video data can be done after successful preprocessing to reduce the complexity of processing and also to reduce the overall time taken to process video.

Video preprocessing for the CBVR task is divided into two major tasks.

- Key Frame Extraction
- Advertisement Detection and Removal

3.3.1 Key Frame Extraction

In literature, it is found that various methods for shot boundary detection and key frame extraction are explored in different domains by researchers working on video data. Major steps in key frame extraction are feature extraction from a base frame and similarity measurement between frames. Features like histogram, edge, motion vectors, scale-invariant feature transform (SIFT), Speeded-Up Robust Features (SURF) Features, corner points, information saliency map, etc. are used in Key Frame Extraction. Sample key frames generated from a news story are given in Figure 3.1. Each key frame represents a different news story as shown in the Figure 3.1.

(a)

(b)

(c)

(d)

FIGURE 3.1 (a)-(d) Key Frames of a news story

Histograms are useful graphing tools. In image processing, the histogram of an image refers to a histogram of the pixel intensity values. This histogram is a graph showing the number of pixels in an image at each different intensity value found in that image.

The presence of camera motion or object motion during smooth shot boundaries increases the difficulty of shot detection using color histogram features. Color histograms can be used for small camera motions. On the other hand, they are sensitive to gradual changes in scenes and cannot differentiate between the shots of a scene. So, the features using color histograms are not more useful to large camera motions. Compared to color histograms, edge features are invariant to changes in illumination as well as motion. Motion features can be very effective to handle the situation where video content is affected by the motion of the camera or object. However, in general, color histograms can perform well on average compared to other types of features [10].

Many of the methods for shot boundary detection extracts visual features from each frame. The next step is to find frame similarities using the extracted feature vectors. Based on the similarity measures shot boundaries can be detected between frames that are not similar. The Euclidean distance, the histogram intersection, etc. can be used for finding similarity for extracted feature vectors of frames.

The threshold-based methods to detect short boundaries are also useful in practice. The method based on threshold uses a predefined value of the threshold. The threshold value used can be either adaptive or globally defined. Also, the combination of both adaptive and global thresholds serves a purpose many times.

Algorithms experimented with are as follows:

Algorithm 1: Key Frame Extraction using Histogram Difference

Algorithm 2: Key Frame Extraction using Edge Difference

Algorithm 3: Key Frame Extraction using Matrix Factorization and rank of a matrix

Key Frame Extraction using Algorithms 1 is carried out on a dataset created using news videos of different news channels. Figure 3.2 shows experimental results of the keyframes generated in the range of frames (150-210) of the video input using the histogram difference

Algorithm 3.1: Key Frame Extraction using Histogram Difference

Input: Video

Output: Key Frames

Step 1: Extract Frames $\boldsymbol{F_{i,1}, F_{i,2}, F_{i,3}, \ldots, F_{i,n}}$ from Video Clip $\boldsymbol{V_i}$ where i=1, 2,…, m.

m = Total Collection of Videos in dataset, n= Total video frames.

Step 2: Find Histogram $\boldsymbol{H_{i,j}}$ for all frames j= 1 to n of video i from the collection of m videos from the dataset.

Step 3: Find Difference of Histograms of consecutive frames using Euclidian Distance

$$\boldsymbol{D_{i,j}} = \sqrt{\sum_{i=1,j=1}^{m,n} (\boldsymbol{H_{i,j}} - \boldsymbol{H_{i,j+1}})^2}$$

Step 4: If $\boldsymbol{D_{i,j} > T}$ then select frame j as a key frame. Let threshold T = μ + c*σ,

Where μ is mean and σ is standard deviation, c=constant

Step 5: Repeat steps 2 to 4 for all videos in the dataset.

method. As described in algorithm 3.1, key frames are generated by taking the difference of histogram and if the difference is greater than the threshold. In Figure 3.2, keyframes generated are marked with the frame number and histogram difference values. Dashed lines

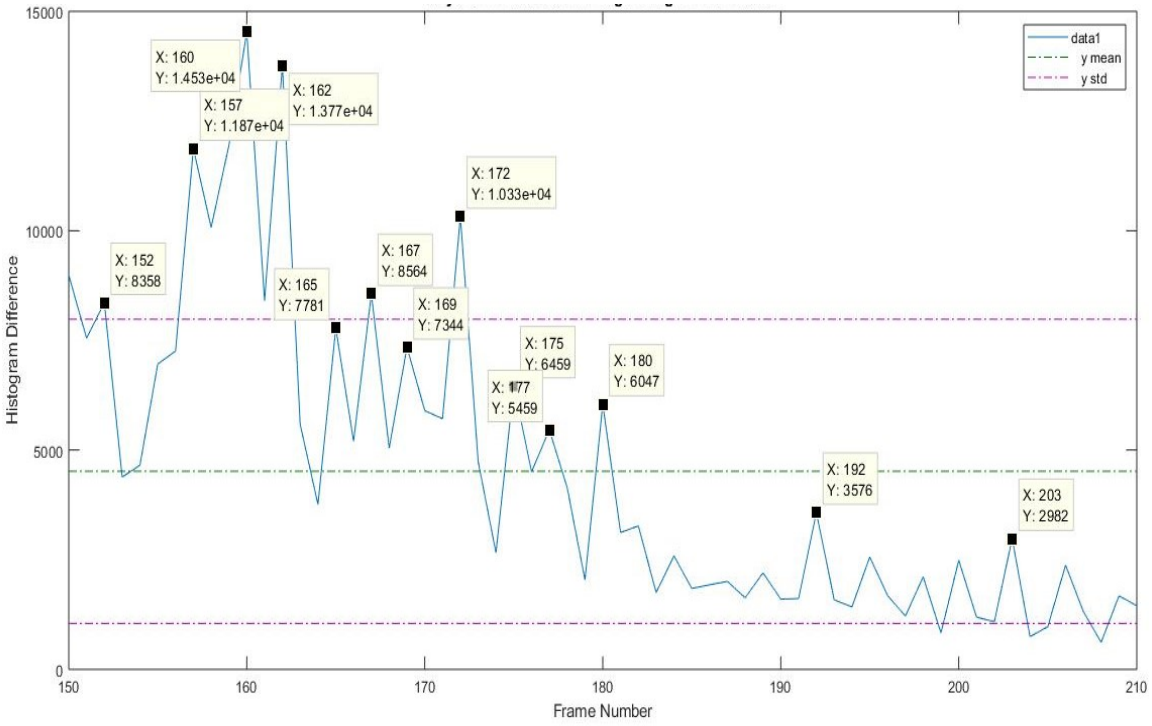


Figure 3.2 Key Frame Extraction using Histogram Difference method given in algorithm 1 between frame range 150-210 of video clip

showing Mean and Standard deviation are also presented in Figure 3.2 which is used for a threshold to limit the number of keyframes per shot.

Edge features are also very important for finding keyframes as given by algorithm 3.2. 'Canny' edge detection method finds edges by looking for local maxima of the gradient of the image. The gradient is calculated using the derivative of a Gaussian filter. This method uses two thresholds to detect strong and weak edges, including weak edges in the output if they are connected to strong edges. By using two thresholds, the Canny method is less likely than the other methods to be fooled by noise, and more likely to detect true weak edges.

In the canny edge detection algorithm, the first image is smoothed using the Gaussian function. In two dimensions, a Gaussian is given by equation 3.1 where $\sigma^2$is variance.

$$G(x.y) = \sigma^2 e^{-\frac{x^{2+}y^2}{2\sigma^2}} \qquad (3.1)$$

and G has derivatives in both the x and y directions. The approximation to Canny's optimal filter for edge detection is G', and so by convolving the input image with G', image E is

obtained with enhanced edges, even in the presence of noise, which has been incorporated into the model of the edge image.

Gradients at each pixel in the smoothed image are determined by applying what is known as the Sobel operator. The first step is to approximate the gradient in the x and y direction respectively by applying the kernels.

The gradient magnitudes can then be determined as a Euclidean distance measure by applying the law of Pythagoras as shown in Equation 3.2

$$|G| = \sqrt{{G_x}^2 + {G_y}^2} \qquad (3.2)$$

where $G_x$and G_y are the gradients in the x and y directions respectively. The direction of the edges is determined by Equation 3.3.

$$\theta = \tan^{-1}\frac{|G_y|}{|G_x|} \qquad (3.3)$$

The next step is to use a double threshold for identifying three kinds of pixels: strong, weak, and non-relevant. Strong pixels are pixels that have an intensity so high that we are sure they contribute to the final edge. Weak pixels are pixels that have an intensity value that is not enough to be considered as strong ones, but yet not small enough to be considered as non-relevant for the edge detection. Other pixels are considered as non-relevant for the edge.

Algorithm 3.2: Key Frame Extraction using Edge Difference

Step 1: Extract Frames$F_{i,1}, F_{i,2}, F_{i,3}, \ldots, F_{i,n}$from Video Clip V_i and convert it to a Gray scale image $G_{i,1}, G_{i,2}, G_{i,3}, \ldots, G_{i,n}$ where i=1,2,..,m, m=Total Collection of Videos in dataset n= Total video frames.

Step 2: find edges $E_{1,i}, E_{2,i}, E_{3,i}, \ldots, E_{n,i}$ using the canny method for all frames j= 1 to n of video V_i from the collection of m videos.

Step 3: Find Difference of Histograms of consecutive frames using

$$D_{i,j} = \sum_{i=1,j=1}^{m,n}(E_{i,j} - E_{i,j+1}) \qquad (3.2)$$

Step 4: If $D_{i,j} > T$, select frame j as a keyframe. Threshold T = μ + c*σ,

Step 5: Repeat steps 2 to 4 for all videos in the dataset.

Key Frame Extraction using Algorithms 3.2 is carried out on dataset video. Figure 3.3 shows the keyframes generated from the range of frames (150-210) of the video input using edge differences.

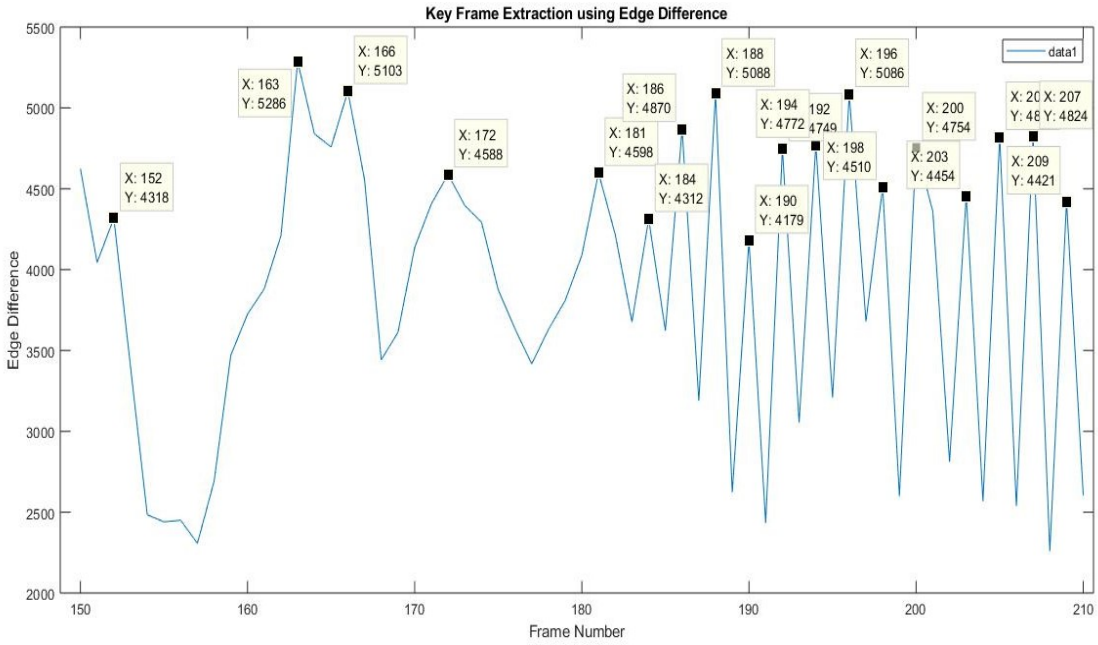


FIGURE 3.3 Key Frame Extraction using Edge Difference method given in algorithm 2 between frame range 150-210 of video clip

Algorithm 3.3: Key Frame Extraction

Step 1: Extract Frames $F_{i,1}, F_{i,2}, F_{i,3}, \ldots, F_{i,n}$ from Video Clip V_i where i=1,2,..,m, m is the total Collection of Videos in dataset n is total video frames.

Step 2: Find $H_{F_{i,j}}, S_{F_{i,j}}, V_{F_{i,j}}$ plane for each frame F_i , i=1,2,…,n into HSV color space.

Step 3: Concatenate Histograms each plane

$hist(F_{i,j}) = [hist\left(H_{F_{i,j}}\right)' \; hist\left(S_{F_{i,j}}\right)' \; hist(V_{F_i,j})' \;]$ frame $F_{i,j}$,i = 1, …,n

Step 4: Repeat Steps 2 to 4 for all frames of Video V_i

Step 5: Apply Matrix Factorization M using SVD method on N number of frames using equation 3.3

$$M = ADV^T \qquad (3.3)$$

A and V are orthogonal matrices of size p x p and n x n, diagonal matrix D with dimension p x n.

Step 6: Find the rank of the matrix using equation 3.4

$$C_R(M) = L(\frac{M}{M(1)} > T), \quad (3.4)$$

Where Threshold T = μ + c*σ, C_R is current rank,

Step 7: Select the current frame as given in equation 3.5,

$$K_{f_{i,j}}(\mathrm{k}) = F_{i,j} \text{ if } C_R < P_R \quad (3.5)$$

Where k=1..K number of key frames stored in the structure of kframes, P_R is previous rank

Step 8: Repeat steps 5 to 7 to generate a collection of keyframes of input video clip V_i where i=1,2,..,m.

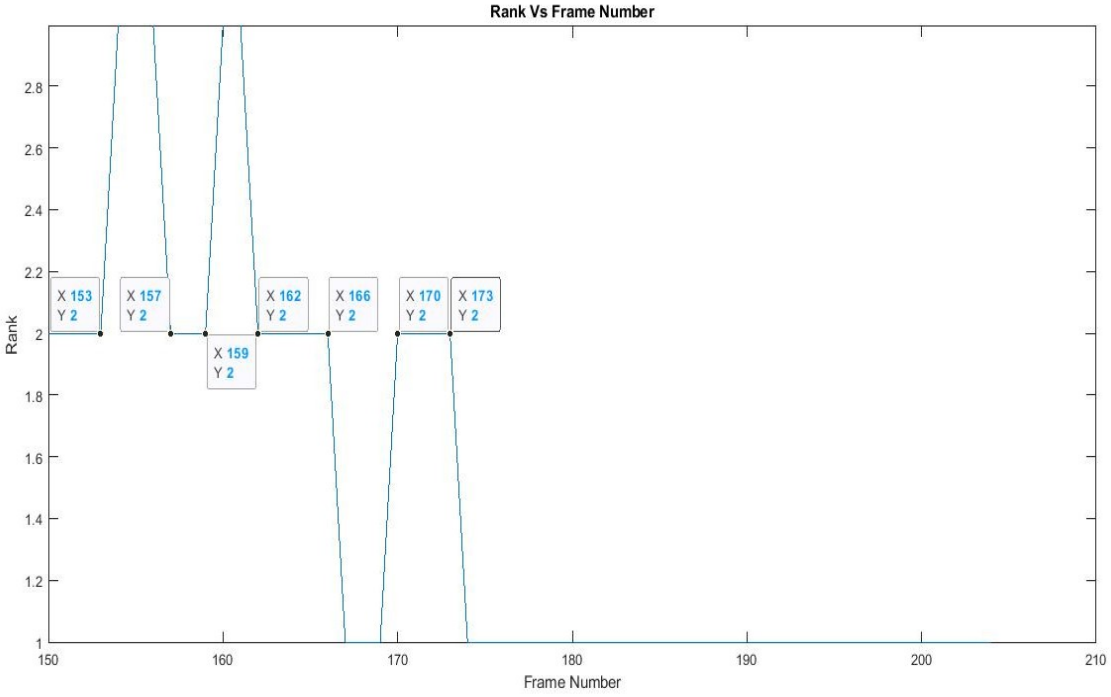


FIGURE 3.4 Key Frame Extraction using Rank and SVD based method given in algorithm 3 between frame range 150-210 of video clip

Key Frame Extraction using Algorithms 3 is carried out on dataset video. Figure 3.4 shows the key frames generated from the range of frames (150-210) of the video input using Singular Value Decomposition and rank of the matrix approach. The SVD has become one of the popular algebraic transforms in image processing applications due to its inherent property of generating singular values important for representing the image. Also, the rank of the matrix gives information about the number of linearly independent rows of columns of the matrix. With the SVD applied on frames of video, unique keyframes are generated as described by algorithm 3.3.

To compare the performance of all three key frame extraction methods, experiments were performed on the same news video clip. As a result, for a particular range of frames i.e., 150 – 210 experiments show that algorithm 3 generated required keyframes whereas the other two algorithms generated more redundant frames for the new story taken into consideration.

3.3.1.1 Results

In this part of the thesis, the experimental results of the methods for extracting key frames to represent a shot in video input are given. Proposed algorithm 3 works well on the dataset of Gujarati Language News Videos. Key frame extraction methods mentioned in the previous section are applied to three different sets of news videos ETVNG, DD11NG, and SANNG datasets of different news channels. Details of the size of the video in hours in each dataset along with total frames and key frames are given in Table 3.8.

TABLE 3.8 Key Frames from each Dataset

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

$$\text{Compression Ratio (CR)} = 1 - KF/TF \quad (3.1)$$

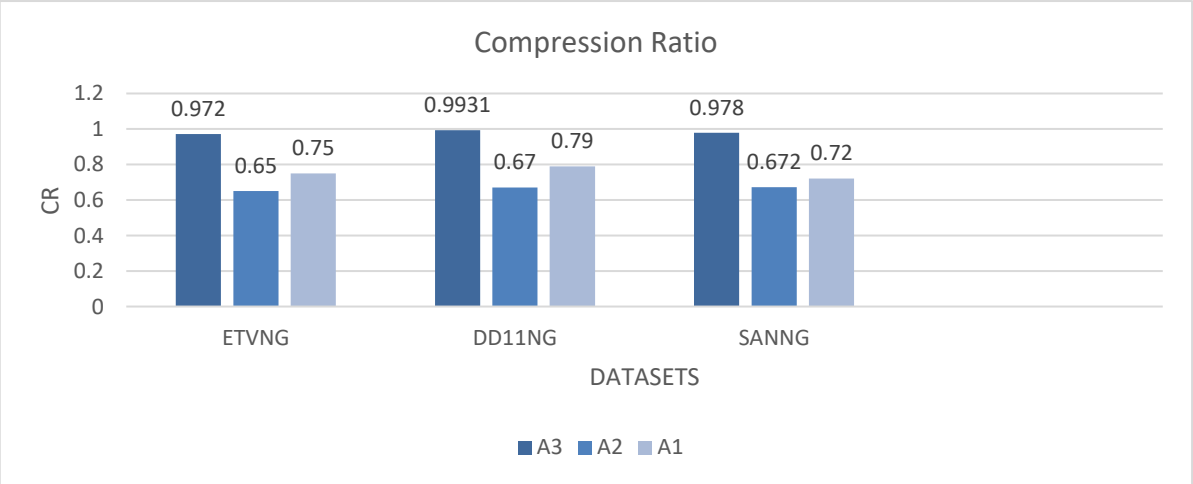


FIGURE 3.5 Performance Evaluation of Three Algorithms with three datasets

To evaluate the performance of all three algorithms 3.1, 3.2, and 3.3 on three different datasets ETVNG, DD11NG, and SANNG compression ratio is used as the evaluation measure the compression ratio can be calculated as:

For example, for dataset ETVNG the compression ratio is calculated using equation 3.1 gives the value 0.972. The comparison of all three algorithms on three news channel video datasets is given in terms of compression ratio in Figure 3.5.

It can be seen from the comparison given by Figure 3.5, algorithm 3 performs better in comparison with the other two algorithms for all three datasets used in the experiments performed. Also, algorithm2 performs poorly among all three algorithms overall on all the datasets used in the experiments.

The main benefit of obtaining a good compression ratio is to reduce overall frames to process further in the following steps of the proposed CBVR system.

3.3.2 Advertisement Detection and Removal

In a day-to-day video of news, sports, etc. major shares are comprising of advertisements which are unwanted information many times[4]. Advertisement detection is having a useful application in multimedia processing. Processing video to cut portions of advertisement is a tricky part of the advertisement detection task. With help of a deep learning approach, it can be achieved with a small dataset of images for training and testing the model.

The concept of Deep transfer learning is widely used in many models. Deep transfer learning is analogous to the concept of fine-tuning well know training methods in the deep neural network. The choice of retrained layer plays a significant role in the deep transfer learning concept as compared to conventional methods. In the conventional method, the whole network is trained again for brand sparking new categories.

Although it can achieve good results in many cases, it is not the best one because it is hard to know how much the previous training process helps. Besides, only retraining the last layer cannot guarantee the best results either, because categories for source training data are usually different and the semantic representations in the higher layer are quite specific to the training categories. Therefore, it is not that suitable to use the highly specific semantic representation in the new recognition task. The reason for this phenomenon lies in the semantic representation for recognition.

In a convolutional neural network, higher layers of the network represent high-level semantic features such as objects or parts of objects whereas lower layers mostly represent features such as edges, textures, colors [3], etc. considered as low-level features. Low-level features represent similar information most of the time in many systems. Compared to low-level features, the high-level features are not the same for different categories of objects. Due to this fact, any good scheme used for transfer learning should have the capability of learning features that are common and tune the high-level features according to the categories of the target. To achieve the best results using transfer learning, we have carried out a variety of experiments to get the right choice for already trained layers.

In this part of the proposed work, we have proposed advertisement detection using a transfer learning scheme that used pre-trained Alexnet model, SVM classifier, Bayesian optimizer to achieve better results compared to state of art work

3.3.3 Pretrained Deep Learning Models

In the computer vision field, the concept of transfer learning and the use of Pretrained models have attracted researchers and application developers to explore new dimensions using limited resources and also try out their work with an existing deep learning model. Due to constraints like limited time, limited dataset, and computational complexity, it is difficult to construct a network model from scratch and train them for a very large dataset with diversity in the dataset. The concept of a pre-trained network is used to test a small dataset of a similar type or improve existing models for a new or similar task.

Also, the pre-trained models using transfer learning methods became popular in other fields such as NLP. In the context of the Natural Language Processing task, the transfer learning concept can be seen as the ability to train one model for one dataset and later on adapt that model on a different dataset to do different NLP tasks.

Also, the pre-trained networks are used for generating features from a dataset. Pretrained Networks are trained on larger datasets like the ImageNet database[107][121], which are consisted of more than a million images and are classified into a large number of classes. The most popular pre-trained networks for large scale image retrieval tasks are Googlenet[122], Squeezenet[123][124], Resnet18[125], Resnet50[126], Alexnet[107][110], Vgg16[110] [127], Vgg19[127], etc.

3.3.3.1 Alexnet Model

AlexNet is designed by Alex Krizhevsky and won the 2012 ImageNet LSVRC-2012 competition. The AlexNet architecture has depth value 8, depth means "the number of convolutional and fully connected layers from input to output layer path in sequence". Originally, Alexnet is used for classifying images into 1000 different classes of objects like pencil, mug, mouse, keyboard, and many animals. The architecture of Alexnet is shown in Table 3.4 where the input size of the image is 227x227x3. Alexnet has a total of 25 layers that contain five convolution layers and three fully connected layers. After every convolution and fully connected layer, the ReLU layer is applied. Dropout is used before the second fully connected year.[123]

A pre-trained AlexNet convolutional neural network is fine-tuned to perform feature extraction as well as classification on a new comparatively smaller collection of images. AlexNet has been trained on over a million images and can classify images into 1000 object categories like keyboard, coffee mug, pencil, animals, etc. The network has learned rich feature representations for a wide range of images. The network takes an image as input and outputs a label for the object in the image together with the probabilities for each of the object categories.

3.3.3.2 Alexnet as a feature extractor

In this approach, some layers of the Alexnet are used for feature extraction, and then those features are dedicated to traditional classifiers for classification. To extract the features using this approach from keyframes of the database, the first input images were resized into 227x227x3 and the activation function is applied at 'fc7' fully connected layer where 4096 features set of the image was generated. The extracted feature set was given to the SVM classifier. Here 80% of the images were used for training and 20 % images for testing.

3.3.3.3 Transfer Learning Approach

The transfer learning approach using a pre-trained network is a very faster and easy way of training. The weights of the data are taken and transferred to new layers that work as another neural network. To carry out the transfer learning approach, all the pre-trained Alexnet network layers were used apart from the last three layers. The last three layers are replaced by the following layers in the transfer learning network:

1. Fully connected layer

2. SoftMax layer
3. Classification output layer

An image datastore enables you to store large image data, including data that does not fit in memory, and efficiently read batches of images during training of a convolutional neural network. Divide the data into training and validation data sets. Use 70% of the images for training and 30% for validation.

TABLE 3.9 Architecture Details of Alexnet Model

Sr No	Layer	Operation	Details
1.	'data'	Image Input 227x227x3	Images with 'zerocenter' normalization
2.	'conv1'	Convolution 96 11x11x3	convolutions with stride [4 4] and padding [0 0 0 0]
3.	'relu1'	ReLU	ReLU
4.	'norm1'	Cross Channel Normalization	cross channel normalization with 5 channels per element
5.	'pool1'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0 0 0]
6.	'conv2'	Grouped Convolution	2 groups of 128 5x5x48 convolutions with stride [1 1] and padding [2 2 2 2]
7.	'relu2'	ReLU	ReLU
8.	'norm2'	Cross Channel Normalization	cross channel normalization with 5 channels per element
9.	'pool2'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0 0 0]
10.	'conv3'	Convolution	384 3x3x256 convolutions with stride [1 1] and padding [1 1 1 1]
11.	'relu3'	ReLU	ReLU
12.	'conv4'	Grouped Convolution	2 groups of 192 3x3x192 convolutions with stride [1 1] and padding [1 1 1 1]
13.	'relu4'	ReLU	ReLU
14.	'conv5'	Grouped Convolution	2 groups of 128 3x3x192 convolutions with stride [1 1] and padding [1 1 1 1]
15.	'relu5'	ReLU	ReLU
16.	'pool5'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0 0 0]
17.	'fc6'	Fully Connected	4096 fully connected layer
18.	'relu6'	ReLU	ReLU
19.	'drop6'	Dropout	50% dropout
20.	'fc7'	Fully Connected	4096 fully connected layer
21.	'relu7'	ReLU	ReLU
22.	'drop7'	Dropout	50% dropout
23.	'fc8'	Fully Connected	1000 fully connected layer

24.	'prob'	SoftMax	SoftMax
25.	'output'	Classification Output	1000 classes

The network constructs a hierarchical representation of input images as explained in Table 3.9. Deeper layers contain higher-level features, constructed using the lower-level features of earlier layers. To get the feature representations of the training and test images, use activations on the fully connected layer 'fc7'. To get a lower-level representation of the images, use an earlier layer in the network.

3.4 Implementation Details

Advertisement Classification is implemented using the following approaches:

1. Advertisement classification using Alexnet Model and SVM Classifier
2. Advertisement classification using Pretrained Model Alexnet
3. Advertisement Classification using Proposed Deep learning Neural Network Architecture

3.4.1 Advertisement classification using Alexnet Model for Feature Extraction and SVM Classifier

The proposed method given in algorithm 3.4 exploits the concept of transfer learning with the Alexnet model for advertisement detection from the news video dataset. Dataset is created with the collection of news video data of DD Girnar, ETV Gujarati, tv9 news, VTV News and Sandesh broadcasted news channel of Gujarati language.

The transfer learning approach is applied with the Alexnet network. Changes applied in the last layers of the network to train the model for the dataset of news video frames for the advertisement detection task. Although the Main idea of Alexnet is object detection, the model fits perfectly to the task of advertisement classification. The architecture of the 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.

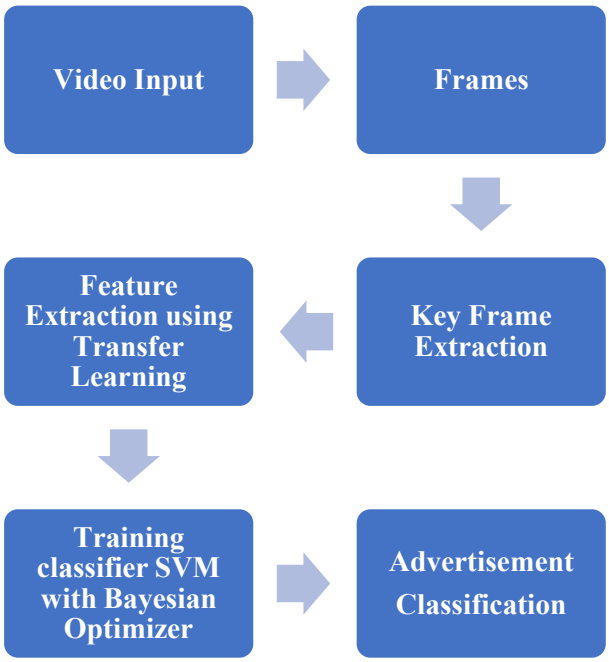


FIGURE 3.6 Proposed Method for Advertisement detection

The proposed method for advertisement detection is explained in Figure 3.6. As shown in Figure 3.6, Frames are extracted from video input. The next task of extracting key frames is performed using singular value decomposition SVD and ranking-based algorithm [6]. Feature extraction from key frames is performed with the Alexnet model for the small size datasets followed by binary classification with SVM to obtain desired advertisement detection task.

Algorithm 3.4: Advertisement Classification using Pretrained Model and SVM Classifier

Step 1: Generate Key Frames $KF_{i,1}, KF_{i,2}, KF_{i,3}, \dots, KF_{i,n}$ using Algorithm for Key Frame Extraction from 3.4: Video Clip V_i for two different categories i.e., Advertisements and News

Step 2: Divide the Dataset in ratio X: Y for training and testing collection

Step 3: Augment the dataset by pre-processing the frames with operations resizing and normalizing for training and testing

Step 4: Alexnet Pretrained model is modified for the dataset. The model can be seen as a function E_{CNN} described by equation 3.6,

$$F_k = E_{CNN}(F_{i,k}),\ \forall F_{i,k} \in V_i\ ,\ i = \quad 1, 2, \dots, m\ , k = ,2, .., n \tag{3.6}$$

model generates f=4096 features vectors $F_{i,k}$ for video V_i from the dataset of videos. Feature Vector can be characterized as a collection of features given by equation 2.

$$F_k = \left((f_k^{(1)}, f_k^{(2)}, \dots, f_k^{(f)})^T\right) \tag{3.7}$$

Step 5: Train the SVM classifier with a Bayesian Optimizer for the training set of feature vectors generated.

Step 6: classify the samples of the test set using the predictors and label data generated from the SVM model trained in step 4.

The network requires an input video frame of size 128-by-128-by-3, but the frames of video have a different resolution for different news channels. To resize the training and test images before they are input to the network, augmented datastores of the input frames need to be created. For this purpose, specify the desired input frame size, and use these datastores as input arguments to activations.

Feature vectors F_k are given as input to the SVM multiclass classifier for the classification task where Bayesian optimization is used for tuning the hyperparameters. For the classification purpose, data is mainly divided into two classes labeled as advertisement and news respectively. Key frames are divided into two classes advertisement and news for training and testing purposes. Dataset has been divided into 75:25 proportions for training and testing of the system.

In binary classification, task SVM has performed well compared to other classifiers due to its kernel trick to handle nonlinear input spaces. SVM finds an optimal hyperplane which helps in classifying new data points. Due to this fact, we have applied classifier SVM with a Bayesian optimizer to boost the classification performance. Bayesian optimizer attempts to minimize a scalar objective function f(x) for x in a bounded domain. The function can be deterministic or stochastic, meaning it can return different results when evaluated at the same point x.

The Gaussian process model of function f(x), Bayesian update procedure for modifying the Gaussian process model at each new evaluation of function f(x) and acquisition function a(x) (based on the Gaussian process model of f) that you maximize to determine the next point x for evaluation are the key elements of Bayesian Optimization.[18] the objective function f will be sampled at $x_t = argmax_x\, u(x|D_{1:t-1}))$ where u is the acquisition function and $D_{1:t-1} = (x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ are the t−1 samples drawn from the function f.

The Bayesian optimization procedure is as follows. For t=1, 2,… repeat:

- Find the next sampling point x_t by optimizing the acquisition function over the GP:

$$x_t = argmax_x\, u(x|D_{1:t-1}) \qquad (3.8)$$

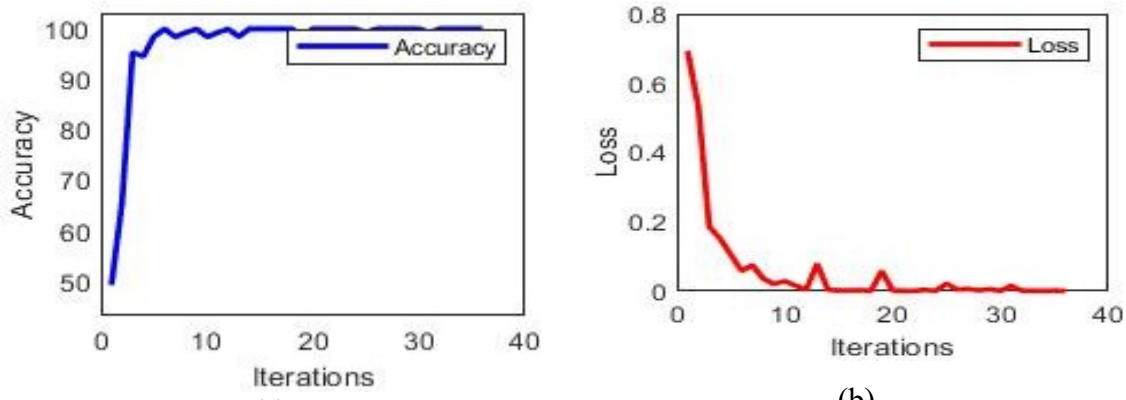
- Obtain a possibly noisy sample from the objective function f.

$$y_t = f(x_t) + \epsilon_t \quad (3.9)$$

- Add the sample to previous samples using equ.3.10

$$D_{1:t} = D_{1:t-1}, (x_t, y_t) \quad (3.10)$$

and update the GP. Bayesian optimization is used to tune the hyperparameters of a model.



(a) (b)

FIGURE 3.7 (a)Training Accuracy (b) Training Loss of proposed approach using transfer learning with Alexnet pre-trained model and SVM classifier with Bayesian Optimizer

The plot of training accuracy of experiments performed with Alexnet pre-trained model with SVM classifier and Bayesian optimizer is shown in Figure 3.7 (a). In Figure 3.7 (b) plot of training loss is given. The base learning rate is taken as 1.0000e-04 with 5 epochs and 50 iterations per epochs used in training.

3.4.2 Advertisement classification using Pretrained Model Alexnet

Algorithm 3.5: Advertisement Classification using Pretrained Model and Transfer Learning Approach

Step 1: Generate Key Frames $KF_{i,1}, KF_{i,2}, KF_{i,3}, \ldots, KF_{i,n}$using Algorithm 3 for Key Frame Extraction from Video Clip V_i for two different categories i.e., Advertisements and News

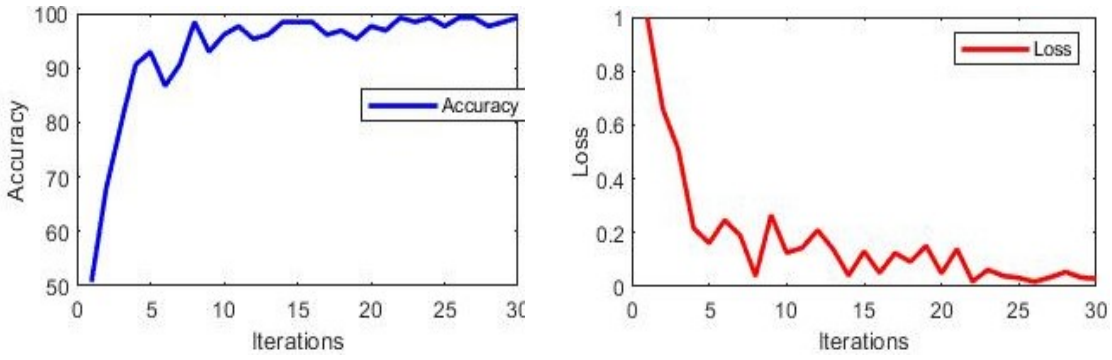
Step 2: Divide the Dataset in ratio X: Y for training and testing collection

Step 3: Augment the dataset by pre-processing the frames with operations resizing and normalizing for training and testing

Step 4: To use the Alexnet Pretrained model, Modify the Last layers to train the data and train the network.

Step 5: Classify the data into two classes ADV and NEWS.

In the experiments performed for transfer learning using Alexnet, the model is trained using a dataset created for advertisement and news classification tasks. The plot of training accuracy of experiments performed with the Alexnet pre-trained model is shown in Figure 3.8 (a). In Figure 3.8 (b) plot of training is given. The base learning rate is taken as 1.0000e-04 with 5 epochs and 50 iterations per epochs used in training.



(a) (b)

FIGURE 3.8 (a)Training Accuracy (b) Training Loss of proposed approach using transfer learning with Alexnet pre-trained model

3.4.3 Advertisement Classification using Proposed Deep learning Neural Network Architecture

Pretrained model and transfer learning concepts were explored first for the advertisement classification task. A neural network model with deep learning is proposed for advertisement classification to compare the performance with the pre-trained model.

TABLE 3.10 ADVNET ARCHITECTURE

Sr No	Layer Name	Input/operation	Description
1	'imageinput'	Image Input	28x28x3 images with 'zerocenter' normalization
2	'conv_1'	Convolution	8 3x3x3 convolutions with stride [1 1] and padding 'same'
3	'batchnorm_1'	Batch Normalization	Batch normalization with 8 channels
4	'relu_1'	ReLU activation function	ReLU activation function applied
5	'maxpool_1'	Max Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]
6	'conv_2'	Convolution	16 3x3x8 convolutions with stride [1 1] and padding 'same'

7	'batchnorm_2'	Batch Normalization	Batch normalization with 16 channels
8	'relu_2'	ReLU activation function	ReLU activation function applied
9	'maxpool_2'	Max Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]
10	'conv_3'	Convolution	32 3x3x16 convolutions with stride [1 1] and padding 'same'
11	'batchnorm_2'	Batch Normalization	Batch normalization with 32 channels
12	'relu_3'	ReLU activation function	ReLU activation function applied
13	'fc'	Fully Connected	2 fully connected layer
14	'softmax'	Softmax	Softmax
15	'classoutput'	Classification Output	crossentropyex with classes 'ADV' and 'NEWS'

In the proposed deep learning neural network architecture, 15 layers are used which is described in Table 3.10. The proposed architecture is named ADVNET which is trained using different parameters to achieve better performance. The main two different notable experiments with optimizers SGD and Adam are described in the following sections.

3.4.3.1 ADVNET using SGDM Optimizer

The stochastic gradient descent with momentum algorithm might oscillate along the path of steepest descent towards the optimum. Adding a momentum term to the parameter update is one way to reduce this oscillation[128]. The stochastic gradient descent with momentum update is

$$\theta_{\ell+1}=\theta_{\ell}-\alpha\nabla E(\theta_{\ell})+\gamma(\theta_{\ell}-\theta_{\ell-1}) \qquad (3.11)$$

where ℓ is the iteration number, $\alpha>0$ is the learning rate, θ is the parameter vector, and $E(\theta)$ is the loss function. In the standard gradient descent algorithm, the gradient of the loss function, $\nabla E(\theta)$, is evaluated using the entire training set, and the standard gradient descent algorithm uses the entire data set at once. Where γ determines the contribution of the previous gradient step to the current iteration[129].

With the proposed model, classification accuracy achieved is 99.74 which is slightly better compared to previous methods using the pre-trained model and transfer learning-based model. The model is trained for 6 epochs with two different optimizers Adam and SGD. The

mentioned results with 99.74 percent of accuracy are obtained with a SGD optimizer. Figure 3.9 shows the training accuracy and loss during the training of the model with 6 epochs and 6 iterations per epochs i.e., a total of 36 iterations. Time taken for training is 15 sec.

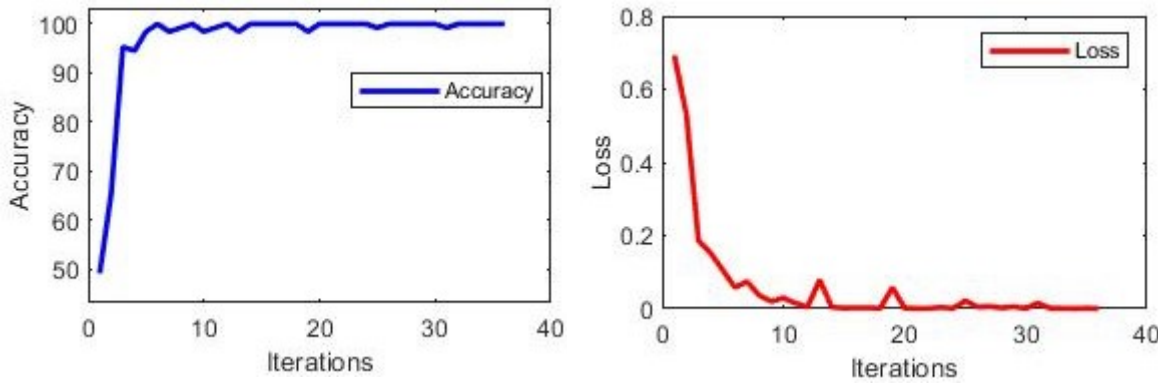


FIGURE 3.9 Training Accuracy and Loss during Model Training for Advertisement Classification using SGD optimizer

3.4.3.2 ADVNET using Adam Optimizer

Adam (derived from *adaptive moment estimation*) optimization algorithm is the alternative of SGD optimization algorithm used for training in deep learning models. Adam uses a parameter update that is similar to RMSProp (Root Mean Square Propagation) with momentum. Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can handle sparse gradients on noisy problems.

In practice, Adam is currently recommended as the default algorithm to use and often works slightly better than RMSProp. However, it is often also worth trying SGD+Nesterov Momentum as an alternative. Adam keeps an element-wise moving average of both the parameter gradients and their squared values as given by the following equations.

$$m_\ell = \beta_1 m_{\ell-1} + (1-\beta_1)\, g_t \qquad (3.12)$$

$$v_\ell = \beta_2 v_{\ell-1} + (1-\beta_2)\, [g_t]^2 \qquad (3.13)$$

Where m and v are moving averages, g is gradient on current mini-batch, β1 and β2 are hyper-parameters of the algorithm. β1 is used for decaying the running average of the gradient (0.9 is the default value). β2 is used for decaying the running average of the square of gradient (0.999). Almost no one ever changes these values. The vectors of moving averages are initialized with zeros at the first iteration.[124], [125] Specify the β1 and β2 decay rates using the 'Gradient Decay Factor' and 'Squared Gradient Decay Factor' name-

value pair arguments, respectively. Adam uses these averages to update the network parameters as

$$\theta_{\ell+1} = \theta_{\ell} - \frac{\alpha\, m_{\ell}}{\sqrt{v_{\ell}} + \varepsilon} \qquad (3.14)$$

If gradients over many iterations are similar, then using a moving average of the gradient lets the parameter updates pick up momentum in a certain direction. If the gradients contain mostly noise, then the moving average of the gradient becomes smaller, and so also the parameter updates become smaller. One can specify ε by using the 'Epsilon' name-value pair argument. The default value usually works well, but for certain problems, a value as large as 1 works better.

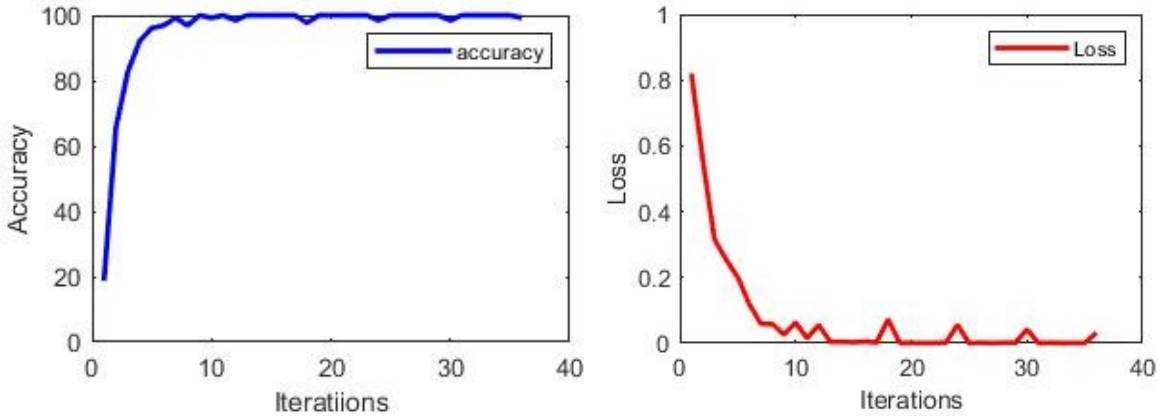


FIGURE 3.10 Training Accuracy and Loss during Model Training for Advertisement Classification using ADAM optimizer

With the proposed model, classification accuracy achieved is 99.47 percent which is slightly less compared to the previous model with an SGDM optimizer. The model is trained for 6 epochs with Adam optimizer. Figure 3.10 shows the training accuracy and loss during the training of the model with 6 epochs and 6 iterations per epochs i.e., a total of 36 iterations. Time taken for training is 21 sec which is also more compared to a model with a SGDM optimizer.

3.5 Results and Comparison

In the proposed approach, the experiments are mainly performed using a machine with an i7-7500 u 2.7 GHz processor, 16 GB RAM, and MATLAB library for image, video, and deep learning architecture. The experiments are performed with MATLAB video processing and a deep learning toolbox. Pretrained model Alexnet is used here for experiments.

Experiments were performed on a small dataset of keyframes extracted from various news channel videos and divided into two classes advertisement and news.

3.5.1 Method1: Performance using Alexnet and SVM

The results of classification are shown in Table 3.11 with the use of a confusion matrix. Out of total frames 32.7 percentage of advertisements frames correctly classified and 0.8 percentage of

TABLE 3.11 Confusion matrix of classification of News vs Advertisements using Alexnet and SVM

OUTPUT CLASS \ TARGET CLASS	ADV	NEWS	
ADV	**32.7 %**	0.0 %	100 % 0.0 %
NEWS	0.8 %	**66.5 %**	98.8 % 1.2 %
	97.6 % 2.4 %	100 % 0.0 %	**99.2 %** 0.8 %

misclassification for advertisement class can be seen in Table 3.11. Whereas, 66.5 percent of total frames are news frames that are correctly classified as news with no misclassification. The proposed method performs very well with an accuracy of 99.2 percent reported. Frames with news stories are correctly classified and labeled as News and also Advertisement is classified correctly and labeled as ADV.

3.5.2 Method 2: Performance using Alexnet Model

The results of classification are shown in Table 3.12 with the use of a confusion matrix. Out of total frames, 32.4 percentage of advertisements frames correctly classified, and 1.1 percentage of misclassification for advertisement class can be seen in Table 3.12. Whereas, 66.6 percent of total frames are news frames that are correctly classified as news with no misclassification.

The proposed approach using the pre-trained Alexnet model performs very well on the dataset with categories news and advertisement with an accuracy of 98.9 percent as per Table 3.12 reported along with training time of 5 min 12 sec taken by the model. Training accuracy and loss are given in Figure 3.11.

TABLE 3.12 Confusion Matrix for advertisement classification using pre-trained ALEXNET model

OUTPUT CLASS \ TARGET CLASS	ADV	NEWS	
ADV	**32.4 %**	0.0 %	100 % 0.0 %
NEWS	0.8 %	**66.6 %**	98.4 % 1.6 %
	96.9 % 3.1 %	100 % 0.0 %	**98.9 %** **1.1 %**

3.5.3 Method 3: Performance using Proposed Deep Learning Neural Network Architecture

TABLE 3.13 Confusion Matrix for ADVNET using SGDM optimizer

OUTPUT CLASS \ TARGET CLASS	ADV	NEWS	
ADV	**33.4 %**	0.3 %	99.2 % 0.8 %
NEWS	0.0 %	**66.3 %**	100 % 0.0 %
	100 % 0.0 %	99.6 % 0.4 %	**99.7 %** **0.3 %**

Results of classification with proposed Deep Learning architectureADVNET using SGDM optimizer are shown in Table 3.13 with the use of confusion matrix. The proposed approach performs well with 99.74 percent accuracy achieved. Out of total frames, 33.4 percentage of advertisements frames correctly classified, and 0.3 percentage of misclassification for advertisement as news class as given in table. Whereas, 66.3 percent of total frames are news frames that are correctly classified as news.

Experiments performed for classification were based on the 8:2 ratio for training and testing set from the dataset taken for advertisement classification where 843 frames were taken in the news category and 435 frames belong to the advertisement category.

TABLE 3.14 Confusion Matrix for ADVNET using ADAM optimizer

OUTPUT CLASS \ TARGET CLASS	ADV	NEWS	
ADV	**32.9 %**	0.0 %	100 % 0.0 %
NEWS	0.5 %	**66.6 %**	99.2 % 0.8 %
	98.4 % 1.6 %	100 % 0.0 %	**99.47 %** **0.53 %**

Results of classification with proposed Deep Learning architecture ADVNET using ADAM optimizer are shown in Table 3.14 with the use of confusion matrix. In this variant with changed optimizer accuracy achieved is 99.5 which is slightly less than that of with earlier optimizer SGDM. Out of total frames, 32.9 percentage of advertisements frames correctly classified, and 0.3 percentage of misclassification for news as advertisement class can be seen in Table. Whereas, 66.6 percent of total frames are news frames that are correctly classified as news.

3.5.4 Comparison of Different Models

The advertisement classification task is mainly performed to detect and remove extra keyframes containing advertisements from the dataset to reduce overall processing time for the retrieval task. The proposed architecture of the deep learning model named 'ADVNET' has given better performance for advertisement classification tasks as compared to other models experimented with our dataset.

As shown in Figure 3.11, the Performance of the ADVNET model with different optimizers Adam and SGDM is given. The experiments performed with ADVNET+SGDM have given accuracy 99.74 which is slightly better than the accuracy achieved with ADVNET+ADAM which is 99.47. Not only accuracy but training time is taken by ADVNET+SGDM was also 15 sec which less compared to the time taken by ADVNET+ADAM which was 21 sec.

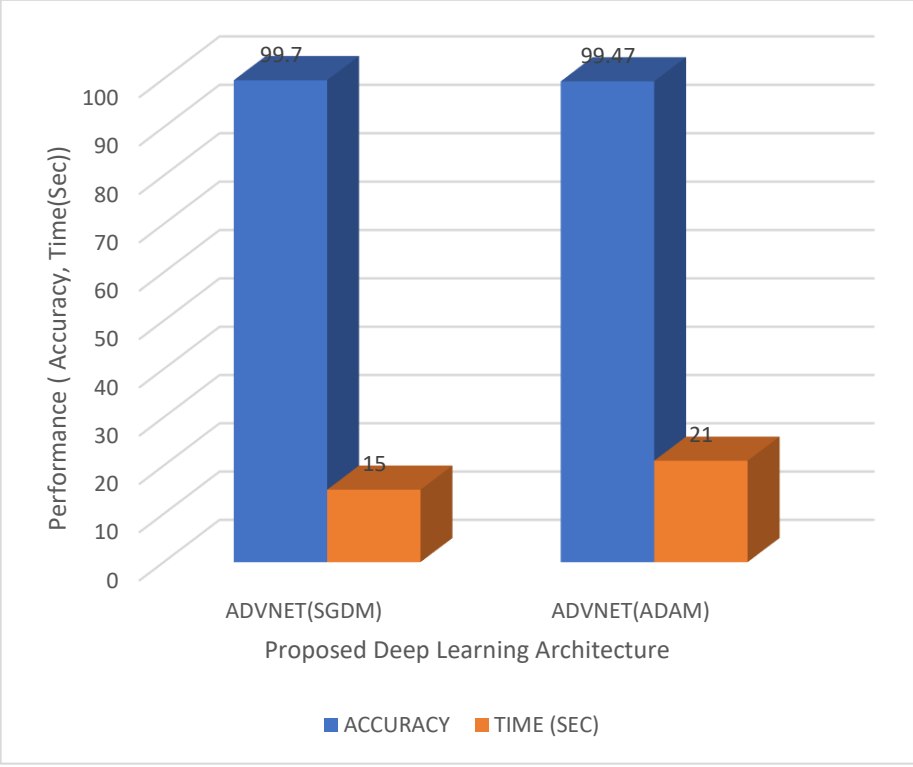


FIGURE 3.11 Comparison of Performance of ADVNET model with different optimizers Adam and SGDM

Finally, in Figure 3.12 comparison of all the different models experimented with the advertisement classification task is given. As per Figure 3.13, it can be seen that the proposed 'ADVNET+SGDM' model performs better in terms of accuracy i.e., 99.74 compared to all the models tried which are ADVNET+ADM with accuracy 99.47, ALEXNET with the

accuracy obtained 98.9 and ALEXNET as feature extractor with SVM classifier with accuracy 99.2.

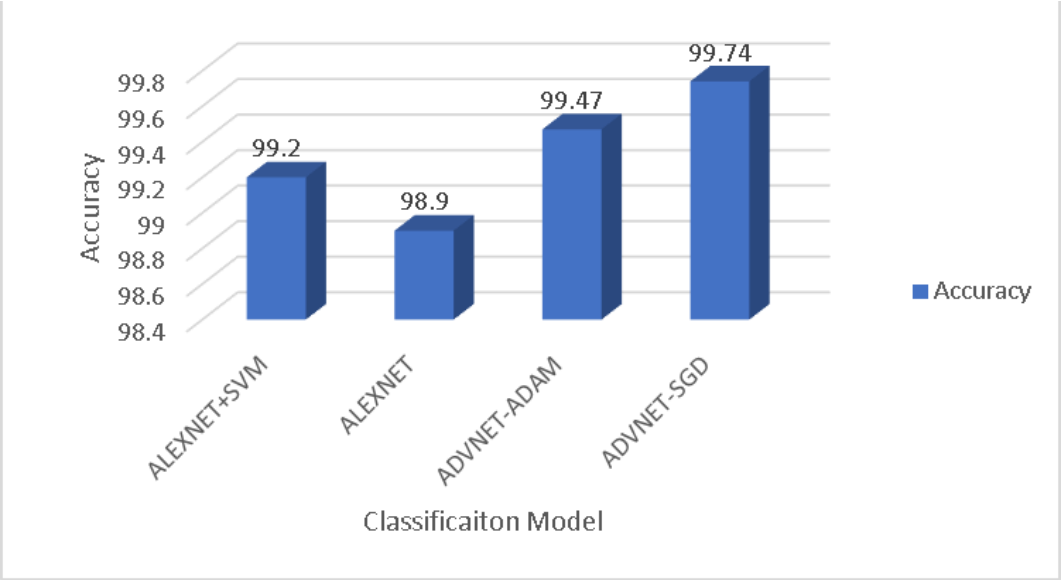


FIGURE 3.12 Performance comparison of different deep learning models for advertisement classification task

Chapter 4

Proposed Text Query based Video Retrieval Approach

4.1 Introduction

In the era of Digital content, it is required to do an intelligent analysis of digital information to make the content beneficial for the applications which are useful to the common man in society. Text content-based video retrieval is an area which is explored for many years in developed countries for the English language and other popular foreign languages. Many applications exist today where English language text information is processed and useful results are generated from various digital contents in various forms such as text documents, images, videos as well as audio. Similar work if searched in the context of regional languages of India, then it is easy to find the gap between technologies and applications that exist for the local languages as compared to the English language.

As discussed in the literature survey chapter, content-based video retrieval can be done for video data very similar to the task to CBIR i.e., Content-based Image Retrieval. As opposed to CBIR, in the CBVR task, the video is parsed and divided into meaningful shots which are processed further to get more meaningful information which is called Keyframes. Key Frame can be seen as a representative frame of the shot that gives necessary information about the shot and features extracted from it can represent a shot. One or more keyframes are used generally to get the required features of the shot. The features extracted are further processed and used for indexing video data. Various indexing techniques exist and the one suitable for the task used to efficiently create an index for the input video dataset. For retrieval, a query image or query text or any other form of query can be used to retrieve relevant video clips from the vast collection of videos of the dataset.

Since Broadcasted Video in India is lacking in metadata information such as closed captioning, transcriptions, etc., retrieval of videos based on text data is a trivial task for most of the Indian language video. To retrieve a specific story based on text query in a regional

language is the key idea behind the proposed approach. Broadcast video is segmented to get shots representing small news stories.

4.2 Proposed Methodology

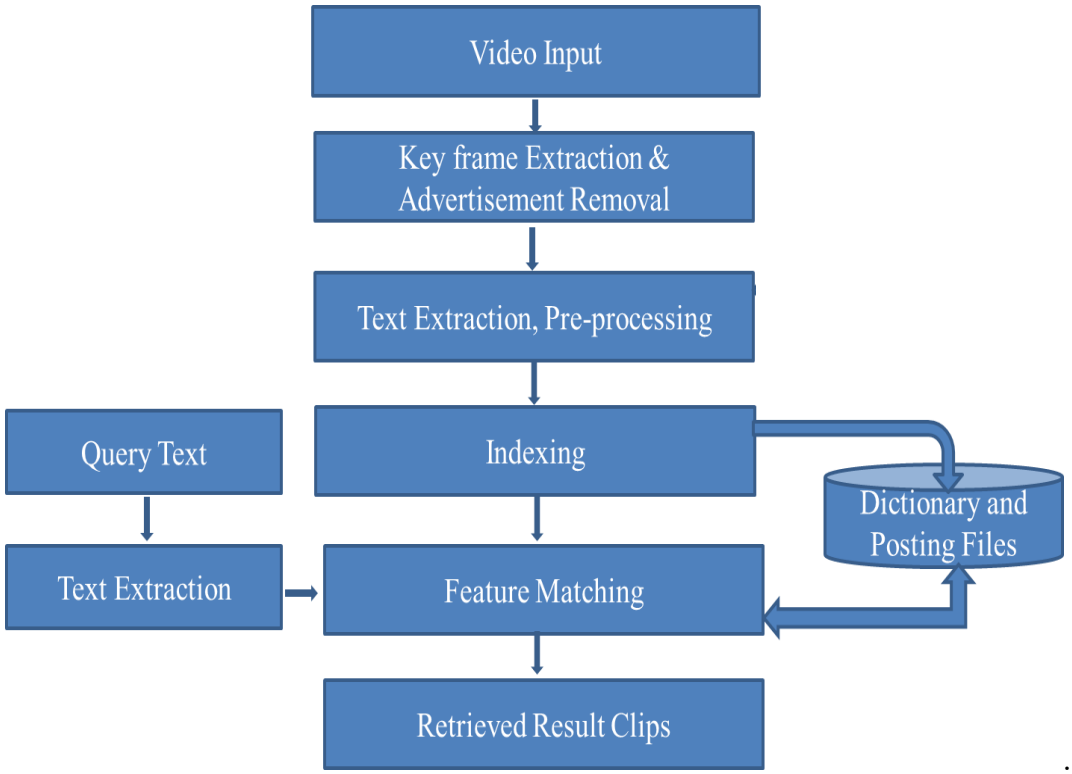


FIGURE 4.1 Block Diagram of Proposed System for Content based Video Retrieval (CBVRAPP1)

The proposed system for content-based video retrieval for news videos is explained in the block diagram of Figure 4.1. As shown in the diagram, the first step is to process the video taken as input to generate a keyframe or set of keyframes to represent each shot or story in the news video. To represent each shot efficiently, keyframe extraction using singular value decomposition and rank of a matrix is proposed as explained in Algorithm 3.3 in chapter 3 of the thesis. Keyframes extracted are processed further for advertisement removal. Advertisement is separated using the proposed advertisement classifier as explained in chapter 3 of the thesis. As advertisements are not useful frames for news story retrieval tasks so they are ignored for further processing in the system.

As shown in the block diagram, the third important task is text feature extraction from the frame of the video. Text is extracted from keyframes using Tesseract OCR and text features are used for indexing which is explained in the following sections of the chapter.

4.2.1 Text Feature Extraction:

દિલ્હીઃગાઢ ધુમ્મસને પગલે ૫૪ ટ્રેન લેટ, ૧૧ ।

FIGURE 4.2 Text Extraction from the frame of input video

Tesseract[83] OCR library is used for text extraction. Figure 4.2 shows the output of Tesseract OCR on the frame of the news video. As the frame is clean and without any noise, the extracted text is much clear as compared to the noisy frame with low resolution.

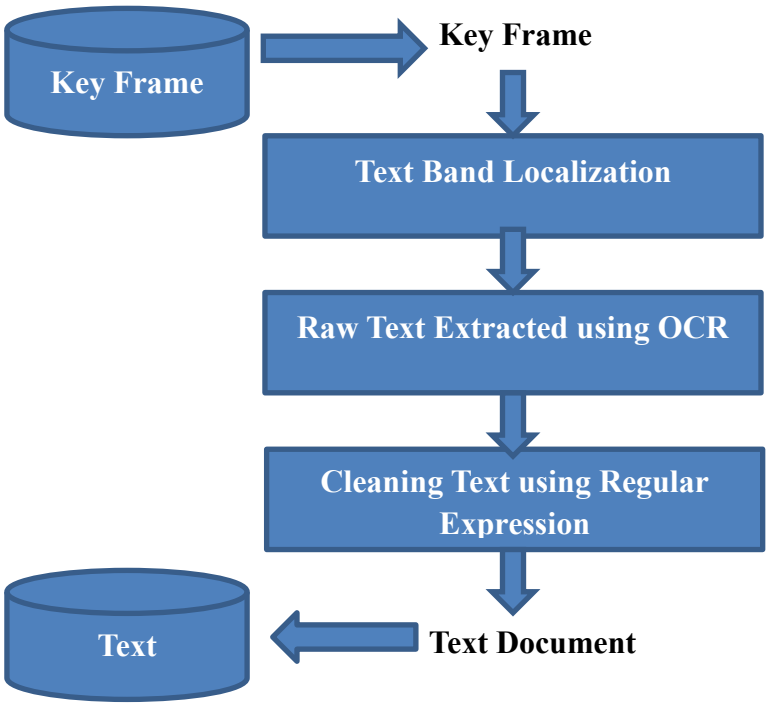
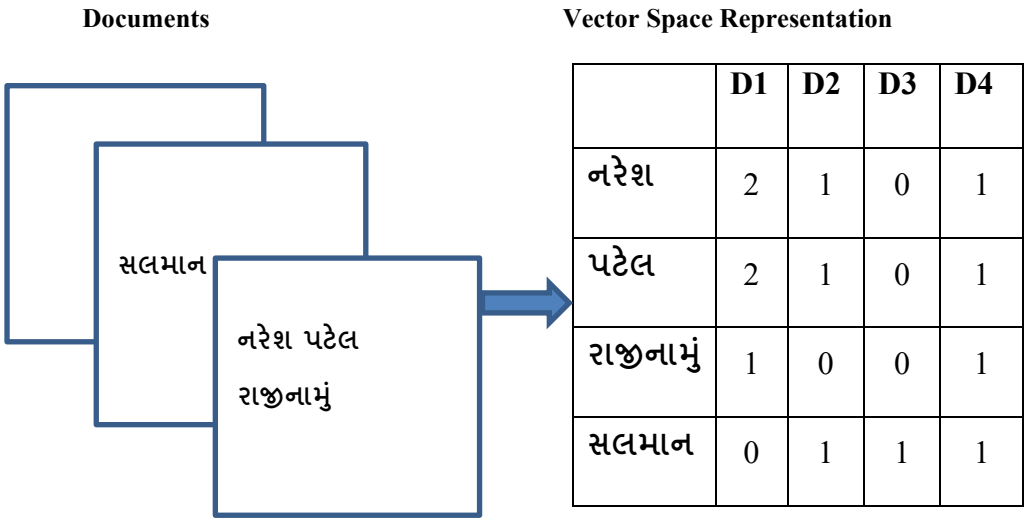


FIGURE 4.3 Text Feature Extraction

Text features are extracted from the keyframe as shown in Figure 4.3. The text band in the video frame is located followed by optical character recognition to extract raw text. Text in raw format contains extra symbols which are cleaned using regular expression formed to process the Gujarati language text and extra symbols.

In parallel to text extraction, another important task is to process document text with steps like tokenization, punctuation, and stemming which are important for indexing.

Normally documents are represented as a vector. In Figure 4.4, an example is shown where documents are represented in vector space where each term and its frequency in documents D1, D2, etc. is shown.



	D1	D2	D3	D4
નરેશ	2	1	0	1
પટેલ	2	1	0	1
રાજીનામું	1	0	0	1
સલમાન	0	1	1	1

FIGURE 4.4 Term-Document Matrix

4.2.2 Indexing

Indexing is a significant method of information retrieval systems. It is the core functionality and also the first step in the information retrieval task. Indexing makes it efficient to retrieve information by reducing documents in the informative terms available within them. It provides an effective mapping of terms to the documents containing those terms. To achieve faster retrieval, it is necessary to build effective indexing for the documents. Basic steps to index documents for retrieval are explained as follows:

Steps for Indexing:

Step 1: Tokenization and Removal of Extra Symbols

For example, a sentence નરેશપટેલ રાજીનામું આપશે is tokenized into four tokens:

નરેશ, પટેલ, રાજીનામું, આપશે

Step 2: Stop word Removal for Gujarati Text:

As part of indexing, stop words are identified and created a list of 90 stop words. After generating tokens from the documents, each token is verified against the list of stop words and removed if found to match with the list. Few stop words are listed in Table 4.1.

Step 3: Stemming

Stemming is the process of finding the root word. e.g., યુનિવર્સિટીમાં is stem to યુનિવર્સિટી. Stemming is explained in detail in section 4.2.2.1.

Step 4: Term weighting

TABLE 4.1 Few examples of Stop Words in "Gujarati" language

ન	નો	છો	જી	લેવા
મૂકી	નહી	બધું	હા	તું
નો	છો	જી	એ	છીએ
હોવા	નું	તેથી	જેવી	હશે
એવા	એની	થતાં	હતાં	તેવી
થયો	એવી	થી	થયું	ત્યાં
માં	ની	આપી	રહે	તેઓ
પાસે	તેમ	ને	તેને	હું

4.2.2.1 Stemming for Gujarati Language

Stemming means finding the root word of a given word. For ex. stemming of ‘walking’, ‘walked’ is ‘walk’, in the Gujarati language for example ‘બોપલની’ stem to ‘બોપલ’.

Stemming is used to reduce dictionary size and improve searching performance. All possible word partitions for the given word are searched and the optimal split position is determined to find the root word. The optimal split position using the given equation is determined to stem the words from the input.

Stemming Example:

- word તપાસમાં root word તપાસ suffix માં
- word ખોટી root word ખોટ suffix ી
- word આવ્યું root word આવ્ય suffix ું
- word તપાસમાં root word તપાસમ suffix ાં

4.2.2.2 Inverted indexing

Inverted indexing is one of the popular indexing methods for text document indexing for faster retrieval. The main components of an inverted index are Dictionary and Postings Lists. For each term in a document from the collection of documents, there is a posting list associated which contains information about the occurrence of the term in the provided collection.

1.1.1.7.1 Dictionary

The dictionary works as a lookup data structure on top of the posting lists. Given an inverted index and a query, our first task is to determine whether each query term exists in the vocabulary. As a first step, it is required to identify if the word is used for searching is available in the vocabulary i.e., the inverted index and if so, identify the corresponding postings. This lookup operation uses a data structure called the dictionary.

1.1.1.7.2 Posting List

The actual index data is stored in the posting list. It is accessed through the dictionary. Each term has its postings list assigned to it. Since the size of the posting list can be large and therefore it’s better to keep this stored over the disk to reduce the cost.

Dictionary and posting files are shown in Figure 4.5 with a snapshot of few entries in the actual dictionary file and posting files generated from the system. For example, દલિતોના 8 120240 entry in dictionary file represents the token દલિત appeared 8 times and last 120240 number represents document id. Each document is uniquely identified using its document id. Indexing of approximately 1.5 lacs documents has been done for the news videos of the dataset.

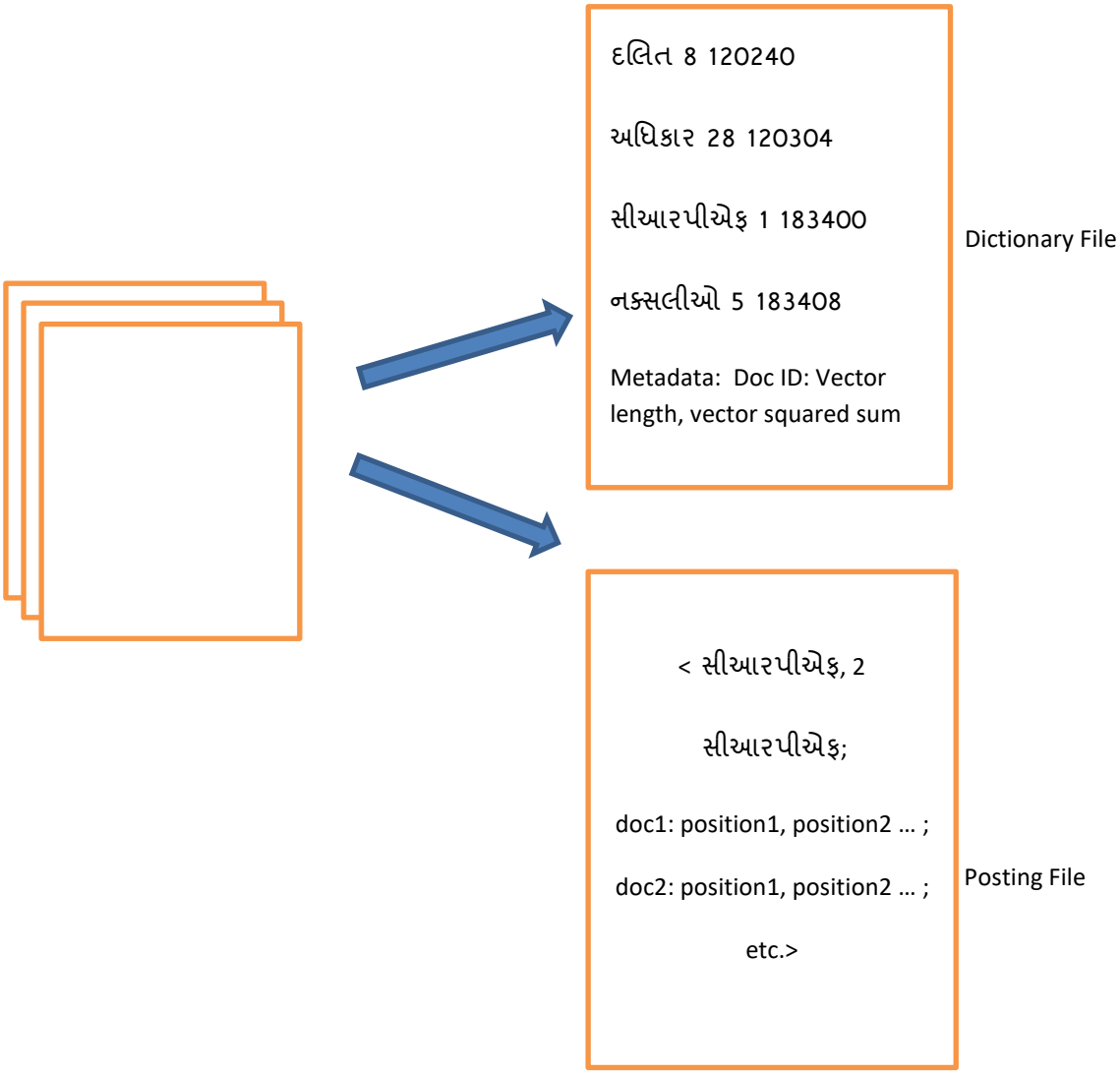


FIGURE 4.5 Dictionary File and Posting File created for indexing documents

4.2.2.3 Score using TF IDF

The mechanism for determining the score of a document includes how good a matching document is for a query. The document which mentions term from query more often should

receive high score while matching. To implement a scoring mechanism based on ranking a score is computed based on term weight.

For the indexing in the proposed approach, documents for each keyframe are generated with Gujarati text extracted after all preprocessing steps to remove extra symbols. Following steps are performed for indexing the documents using Term Frequency and Inverse Document Frequency.

Each document d_n is represented as a vector of terms $t_1, t_2, \dots, t_m$ from the Collection of documents $D = \{d_1, d_2, \dots, d_n\}$. The document vector is given by equation 4.2.

$$\overrightarrow{v_{d_n}} = \left(tf(t_1, d_n),\ tf(t_2, d_n), \dots, tf(t_m, d_n)\right) \quad (4.2)$$

Find the frequency of a term in the document by following equation4.3.

$$tf(t,d) = \sum_{x \in d} f(x,t), \text{ where } f(x,t) = \begin{cases} 1, & if\ x == t \\ 0, & otherwise \end{cases} \quad (4.3)$$

Find Inverse Document Frequency using equation 4.4.

$$idf(t) = \log \frac{|D|}{1+|\{d \in D : t \in d\}|} \quad (4.4)$$

where $|D|$ = Total Number of Documents

4.2.3 Searching Algorithm

- Returns the k most relevant docIDs in the result for the given query
    - params:
        - query: the query string
        - dictionary: the dictionary in memory
        - indexed_docIDs: the list of all docIDs indexed

1. Represent each document in the dataset as a weighted tf-idf vector $\vec{d}$.
2. First finds term **T** and term weights **TWq** from query string using term frequency and inverse document frequency to represent the query as weighted tf-idf vector $\overrightarrow{q}$.
3. Compute cosine score using equation 4.5 as:

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\circ\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sum_{i=1}^{|V|} {q_i}^2 \sum_{i=1}^{|V|} {d_i}^2} \quad (4.5)$$

Here $|\vec{q}| = length\ of\ \vec{q},$ and$|\vec{d}| = length\ of\ \vec{d}$

The equation gives the cosine of the angle between the vectors $\vec{q}$ and $\vec{d}$

4. Calculate score for each doc in posting list using equation 4.6 and 4.7 as given by,

$$TW_d = 1 + \log(tf(t,d)) * \log\frac{|D|}{1+|\{d \in D : t \in d\}|} \quad (4.6),$$

$$S(D) = \cos(\vec{q},\vec{d}) \quad (4.7)$$

TW- Term Weight, TF- Term Frequency in Document

5. Accumulate all scores using equation 4.8 as

$$SS = \sum_1^N S(D) \quad (4.8)$$

Where N is total documents in the dataset, S-Score, SS-Accumulated Score, D-Documents

4. Normalize all scores by the length of the Document.

5. Retrieve top k relevant videos.

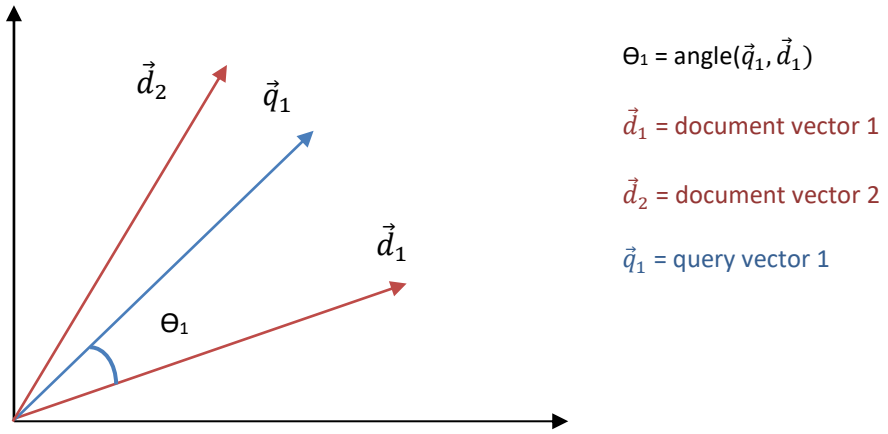


Figure 4.6 Cosine similarity between query and document vectors in vector space model

In the experiments done on the dataset prepared, retrieval of relevant video clips is done based on the query text. For the set of documents of size n, retrieved results that are relevant are taken as true positive, and retrieved video clips which are not relevant to the query are

taken as false positive. Video clips that are pertinent to the query but not retrieved by the system are called false negatives.

Query set is prepared from the dataset. Query set is used to retrieve relevant documents from the dataset. To retrieve top k relevant documents from dataset similarity between test and indexed data is measured using Cosine similarity as described by Figure 4.6. As the cosine of the angle between document vector and query vector is small the more relevant documents are found, as cosine is a monotonically decreasing function of the angle for the interval [0°, 180°]

4.3 Experimental Results

In the proposed approach, the experiments are mainly performed using a machine with an i7 3.3 GHz processor, 16 GB RAM for training the model. The experiments are done with python 3.6 with OpenCV library for video processing and basic text processing library for preprocessing of text features.

The size of different datasets in hours used for experiments is shown in Figure 4.7. Also, the count of keyframes from each dataset for which text feature extraction is done and collection of the text document is generated is given in Figure 4.8.

Outcomes of the proposed approach are acquired by utilizing a query set of size ten. The query set for the task is designed with a variable length of the query text. The total dataset contains approximately 1.5 lacs documents records out of which ten most pertinent archives are retrieved. Assessment of framework is finished utilizing precision and recall metric as well as mean average precision. Precision and Recall are calculated based on retrieved results as following for documents retrieved for each query. The mean average precision value obtained is 91.5. The maximum number of documents retrieved is ten for each query. Text query set is designed to test the performance of the system.

In Figure 4.9, response time for datasets ETVNG, DD11NG, SANNG in microseconds/query is shown. As compared to ETVNG whose response time is 0.5 microseconds per query, response time with the other two datasets was better which 0.39 microseconds per query for SANNG and 0.4 microseconds per query for the DD11NG dataset. In the proposed system, a single query takes an average of 10.53 microseconds for k=10. Once a query is submitted,

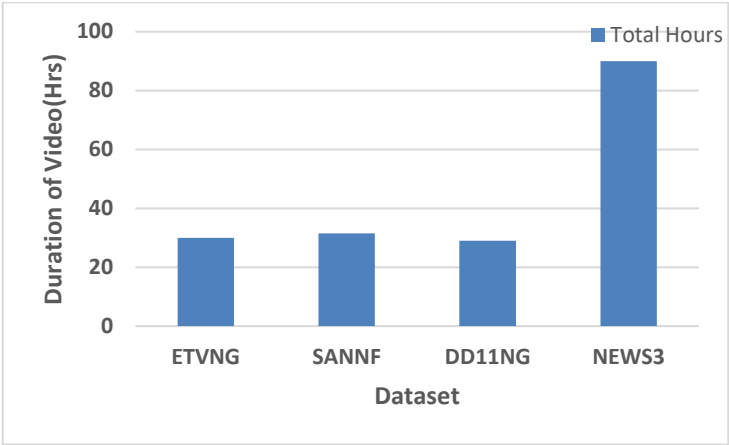


Figure 4.7 Dataset Size used for CBVR

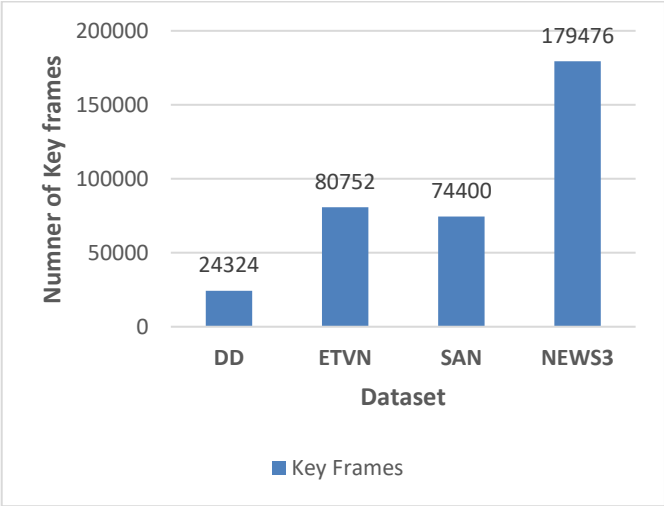


Figure 4.8 Number of Keyframes of different dataset

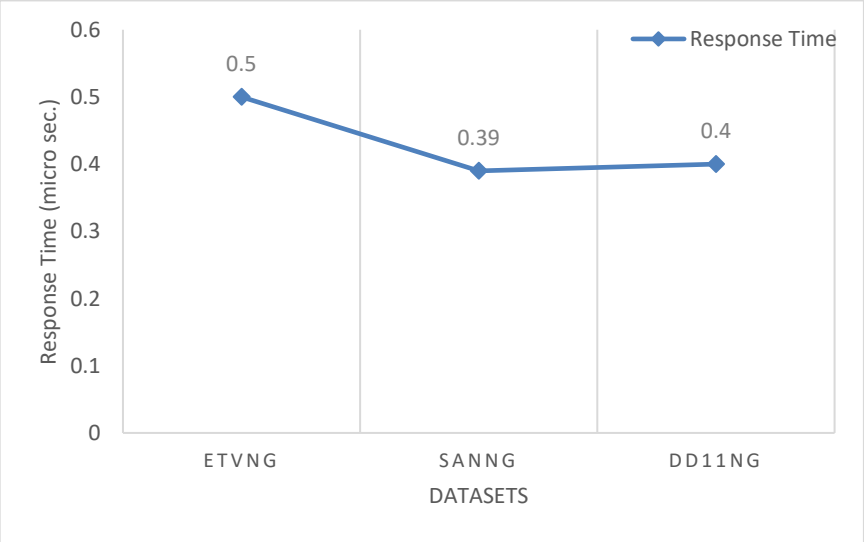


FIGURE 4.9 Response time for different datasets in microseconds/query

results are retrieved and ranking of retrieved results is done based on similarity with query vector. Results of top-six retrieved video clip for query term “સલમાન” is shown in Figure 4.10.

FIGURE 4.10 Top 6 results retrieved on query text “સલમાન”

Performance evaluation of the proposed CBVR using the text query-based approach is done using precision-recall as well as other measures like P@Ri and MAP. Table 4.2 lists precision and recall values calculated using the equation described in chapter 2. The highest precision and recall are achieved precision 1 and recall 0.8 with SANNG dataset, whereas the dataset DD11NG which is small in size compared to SANNG gave comparable recall of 0.78 with precision value 1. ETVNG is the dataset that gets a lower recall value as compared to the other two datasets i.e., 0.6 with the precision value 1. Also, the precision-recall is calculated with the combined large dataset NEWS3 with all videos, which gives recall value 0.73 and precision value 1.

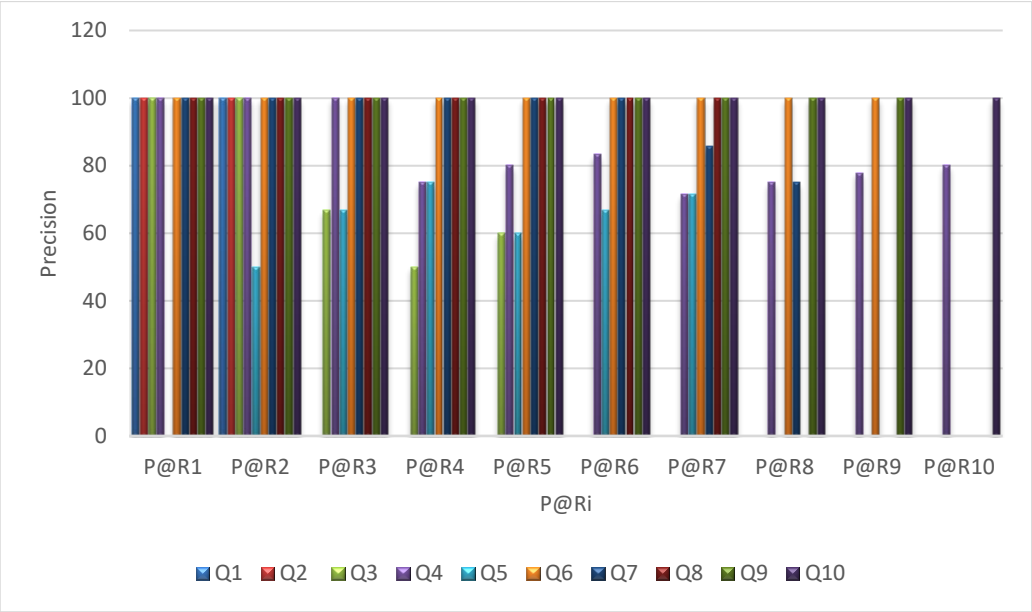


FIGURE 4.11. Performance Evaluation in P@Ri of the proposed approach for query set Q

TABLE 4.2 Results using Precision-Recall and Response Time/Query for different datasets

Dataset	Precision & Recall
ETVNG	P=1, R=0.6
SANNG	P=1, R=0.8
DD11NG	P=1, R=0.78
NEWS3	P=1, R=0.73

To measure the performance of the proposed system using P@Ri performance measure, values for P@R1, P@R2,..P@R10 are calculated for the query set with k=10. The plot of Figure 4.11 shows the results for queries from the set Q={Q_1,…, Q_{10}}. For the higher value of k value for P@R is decreasing for different queries as for few queries in the set, the number of relevant documents is less than k. With performance measures for each query, finally, precision and average precisionis calculated as given in equation 4.9. The mean average precision MAP is calculated from the average precision value obtained to evaluate the performance of the system over the query set.

$$AP = 1/n \sum_{k=1}^{n} P@Rk \quad (4.9)$$

$$MAP(Q) = \frac{1}{Q} \sum_{J=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \quad (4.10)$$

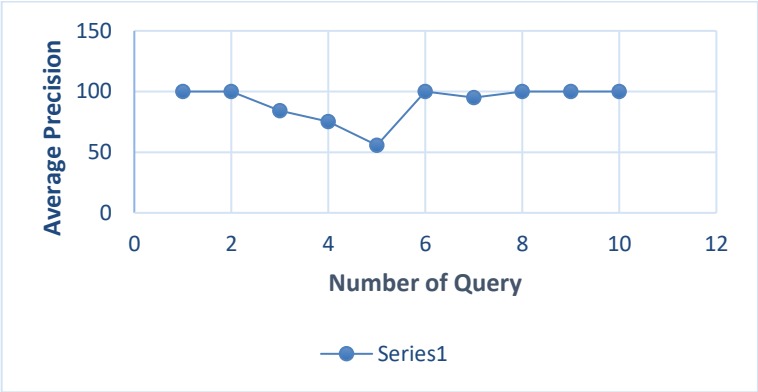


FIGURE 4.12 Performance of Text Query-based video retrieval

The Performance of text query-based retrieval in terms of average precision is shown in Figure 4.12. Finally proposed text query-based CBVR approach is evaluated with Mean Average Precision from the average precision value calculated as given in Figure 4.12. The MAP value obtained is 91.5 using equation 4.10. For the large dataset, experiments are performed which provides faster retrieval with good MAP value along with less memory requirement for storing index due to preprocessing steps for video as well as text preprocessing before indexing.

Chapter 5

Proposed Deep Learning Approach for News Video Retrieval

5.1 Introduction

Deep learning is part of artificial intelligence (AI) that working likes human brains. It learns from examples and creating the pattern that is used at the time of decision making. It is a subgroup of machine learning that has an efficient network, which learns in an unsupervised manner from a large amount of labeled data. Deep learning used a similar kind of architecture used in Neural Network, so it is also well-known as a deep neural network or deep neural learning.

The word “deep” typically points out the number of hidden layers in the neural network. Usually, the traditional neural networks have only two-three hidden layers since deep neural networks have a large number of layers like 150 layers. The traditional neural network requires manual feature extraction, while deep neural learning models are trained automatically, with no need for manual feature extraction from data, but it required a large amount of labeled data to learn itself [117].

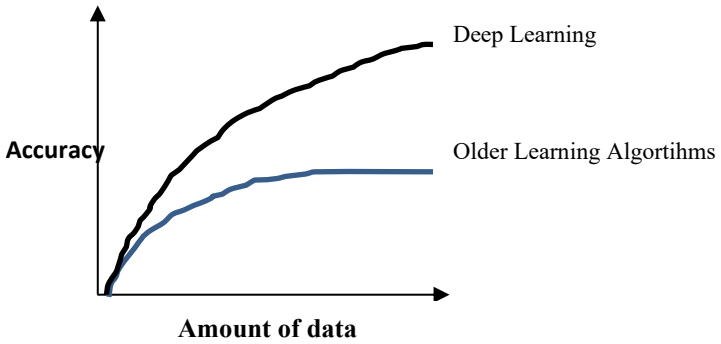


FIGURE 5.1 Performance Evaluations for Deep Learning and Machine Learning Algorithm

Figure 5.1 shows the performance of deep learning vs. older machine learning algorithms, which indicates for a large amount of data, deep learning gives very higher accuracy than

other feature extraction-based classifiers. So, as datasets size is increasing, the performance of a deep neural network is also increased.

5.2 Autoencoders

A typical autoencoder architecture comprises three main components as shown in Figure 5.2. First, an encoding architecture comprises of series of layers with the number of nodes in decreasing order and eventually reduces to a latent view representation. The second component is Latent view representation which is the lowest level space in which the inputs are reduced and information is preserved. The third component is decoding architecture, which is the mirror image of the encoding architecture but in which the number of nodes increases in every layer and ultimately outputs a similar input.

A highly fine-tuned autoencoder model should be able to reconstruct the same in-put which was passed in the first layer. Autoencoders are widely used with image data and in the applications, such as Feature Extraction, Dimensionality Reduction, Image Compression, Image Denoising, and Image Generation.

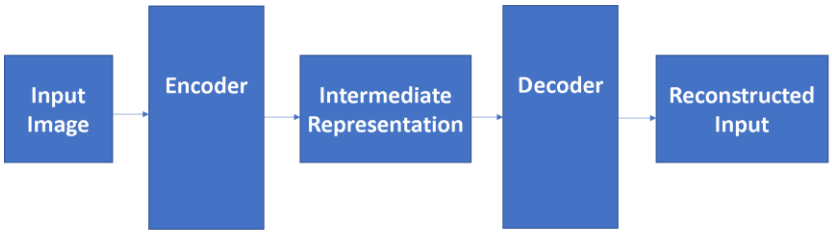


FIGURE 5.2 Autoencoder

Variants of autoencoders are denoising autoencoder, convolutional autoencoder, stacked autoencoders, etc. A special kind of deep learning neural network architecture autoencoders can use for the task of unsupervised learning. The main benefit of using autoencoders is it is used for dimensionality reduction as well as data compression and retrieval. Denoising autoencoders don't just copy inputs as noise is added to the input image before feeding it to the network. In denoising autoencoder, the first step is to corrupt the initial input x into x˜ using a stochastic mapping as shown in equation 5.1.

$$\boldsymbol{x^{\sim} \sim qD(x^{\sim}|x)} \tag{5.1}$$

The intermediate representation is given by y in equation 5.2 where W is weight and b is bias.

$$\boldsymbol{y = f\theta(x^{\sim}) = s(Wx^{\sim} + b)} \tag{5.2}$$

The output image is reconstructed using equation 5.3.

$$z = g\theta'(y) \tag{5.3}$$

Parameters θ and θ′ are trained to minimize the average reconstruction error over a training set. Reconstruction Error or Loss can be defined as given in equation 5.4 where x_k is a training example, z_k is the predicted value.

$$L_H(X, Z) = -\sum_{K=1}^{d}[x_k \log z_k + (1 - x_k)\log(1 - z_k)\,] \tag{5.4}$$

The main objective is to minimize reconstruction error which amounts to maximizing a lower bound on the mutual information between input X and learned representation Y. Good reconstruction of its input means that it has retained much of the information that was present in that input.

5.3 Training and Feature Extraction using Auto Encoders

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

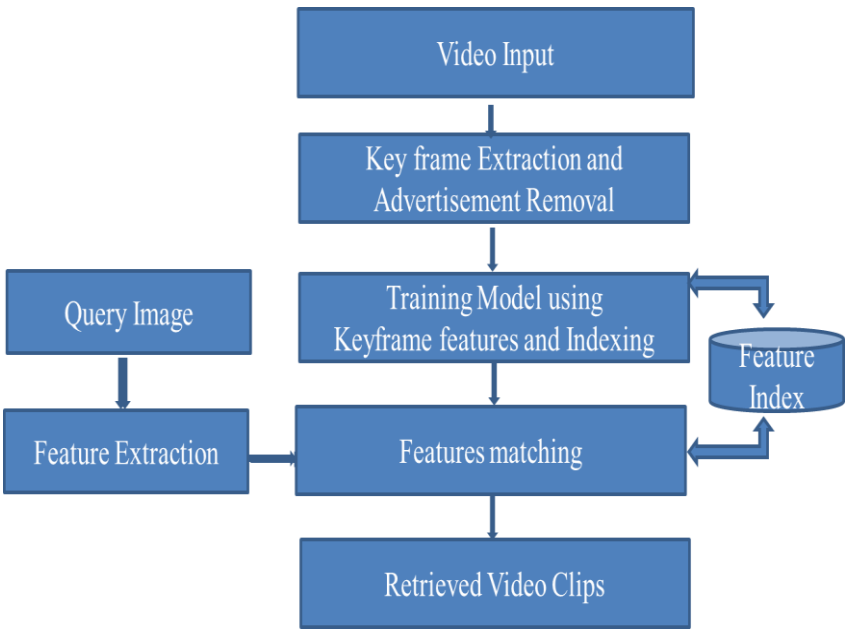


FIGURE 5.3 Block Diagram of Proposed Deep Learning approach for Image Query based Video Retrieval System

As shown in Figure 5.3, keyframes collected after removing advertisements are fed into the autoencoder architecture to train the model. The encoder model is trained with different

epochs, batch size, optimization function to achieve better performance. Features are extracted using a well-trained encoder model for the news video dataset and stored separately to match with query features.

5.4 Experimental Details

Experiments are mainly performed using a machine with GPU NVIDIA Titan Xp with i7 processor, 16 GB RAM for training the autoencoder model. Keras is the python library used for deep learning applications which are used in the implementation task. Keras framework is set up on top of TensorFlow for the proposed work using the autoencoder model. Keras is a neural network library while TensorFlow is the open-source library for several various

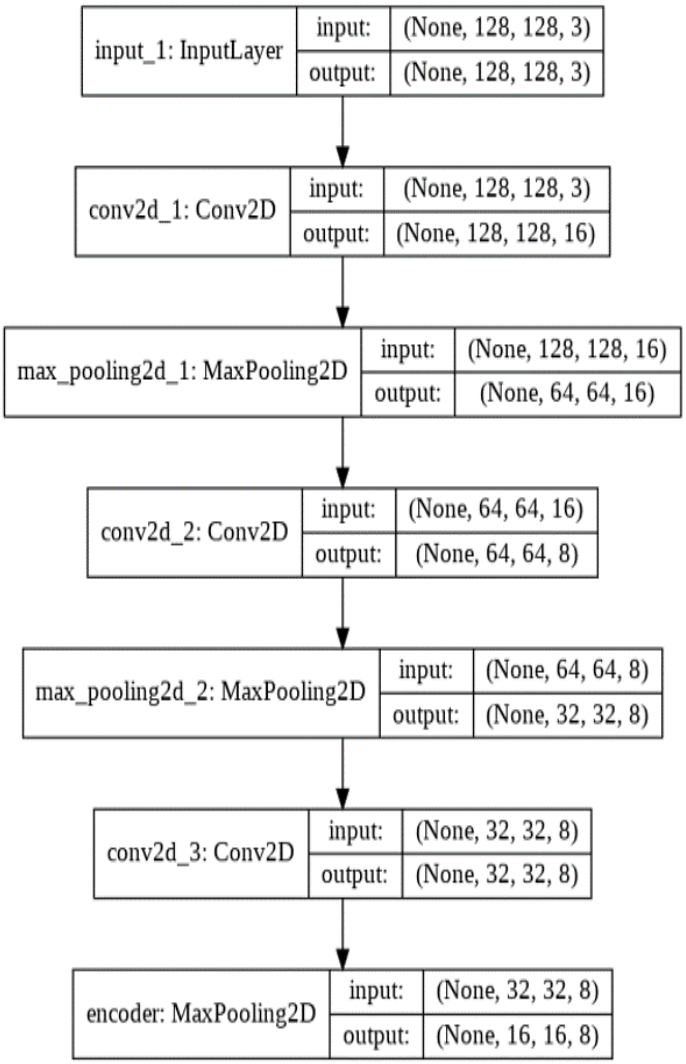


FIGURE 5.4 Architecture of Encoder

tasks in machine learning. The encoder is used to generate the compact representation of input data and the decoder is used to do a reverse task which the encoder does. In the

proposed approach, the encoder is used to extract the features from the input video frame taken as input.

The architecture of the encoder is shown in Figure 5.4. Total seven layers are taken in the encoder model proposed to process the input frame. The first layer, the input layer, takes a frame with dimension 128x128x3 as input which is further given in the second layer as input.

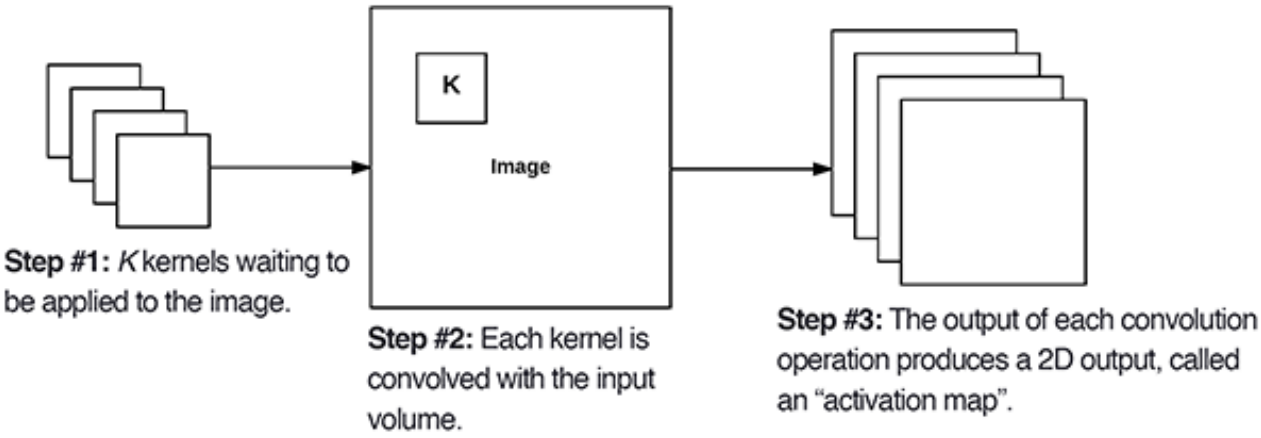


FIGURE 5.5 Conv2D parameters[131]

As shown in Figure 5.5, filters determine the number of kernels to convolve with the input volume. Each of these operations produces a 2D activation map as shown in Figure 5.5. Figure 5.6 shows the results of the filter applied to the image with padding.

Convolution is performed by convolving the input image with the kernel as shown in fig 5.3 which can be explained with equation 5.5.

$$y(i,j) = w * x(i,j) = \sum_{k=1}^{m} \sum_{l=1}^{n} w(k,l)\, x(i-k, j-l) \qquad (5.5)$$

Here $x(i,j)$ is the original image, $y(i,j)$ is the image obtained after convolution of the input image with kernel w of size m x n, $1 \leq k \leq m, 1 \leq l \leq n$.

In the third layer, max pooling with size 2x2 is applied with 'same' padding which generates output dimensions 64x64x16. As the main objective of the max-pooling layer is to downsample an input representation. Max pooling works by calculating the maximum value for each patch of the feature map.

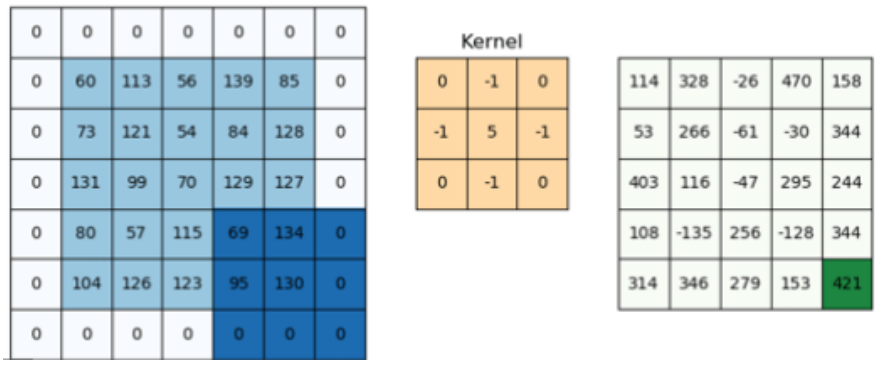


FIGURE 5.6 3×3 kernel applied to an image with padding

In layer fourth, again 8 convolution filters of kernel 3x3 and activations 'relu' is applied on the output generated from the third layer which generated output dimensions of 64x64x8 where padding is used to generate similar size output. ReLU stands for Rectified Linear Unit. The main advantage of using the ReLU over other activation functions is that RELU does not activate all the neurons at the same time. The neurons will only be deactivated if the output of the linear transformation is less than 0. Activation function relu is defined by equation 5.6.

$$f(x) = \max(0, x) \qquad (5.6)$$

In the fifth layer, max pooling with size 2x2 was applied which generated 32x32x8 dimensions of feature maps.

In the following layers ie layer sixth and seventh, 8 convolution filters of kernel 3x3 with activation function 'relu' and max-pooling of size 2x2 are applied respectively. Padding is applied in the sixth layer to generate an output size of 32x32x8.

(a) **(b)**

FIGURE 5.7 (a) Video frame used in training of size 128 x128x3 (b) Predicted code generated with encoder reshaped from the original shape 16x16x8

The output of the seventh layer is 16x16x8 which is named as encoder layer. The feature map generated for the image shown in Figure 5.7(a) using the encoder layer is shown in Figure 5.7(b).

TABLE 5.1Layers, Shape and Parameters per layer of Encoder in Proposed Autoencoder Architecture

Sr No	Layer	Output Shape	Parameters
1.	input_1 (InputLayer)	(None, 128, 128, 3)	0
2.	conv2d_1 (Conv2D)	(None, 128, 128, 16)	448
3.	max_pooling2d_1	(MaxPooling2 (None, 64, 64, 16)	0
4.	conv2d_2 (Conv2D)	(None, 64, 64, 8)	1160
5.	max_pooling2d_2	(MaxPooling2 (None, 32, 32, 8)	0
6.	conv2d_3 (Conv2D)	(None, 32, 32, 8)	584
7.	encoder (MaxPooling2D)	(None, 16, 16, 8)	0

For the model with encoder, total params are 2192 and Trainable params are also 2192 which includes parameters from layer2 conv2d_1, layer 4 conv2d_2, and layer 6 conv2d_3 as shown in Table 5.1. For the model with encoder and decoder both, the total parameters are 4,963 and trainable parameters are also 4,963.

The objective function for the convolutional autoencoder is binary cross-entropy which is also considered as a cost function for the model which is given by equation 5.7. The stochastic gradient descent (SGD) method to train the model as given in the following section and randomly initialize the weight parameters of each layer.

$$L_H = -\sum_{k=1}^{d}[x_k log z_k + (1 - x_k)\log(1 - z_k)] \quad (5.7)$$

5.4.1 Optimizers and Parameters used for training

The basic task of optimizers is to optimize the performance by adjusting various parameters.

5.4.1.1 Adadelta optimizer

Adadelta is a more robust extension of AdaGrad that seeks to reduce its aggressive, monotonically decreasing learning rate based on a fixed moving window of gradient updates, instead of accumulating all past gradients.

Adadelta functions as a stochastic gradient descent method. It has a float known as the Adadelta decay factor (p). For the Adadelta optimizer, you can adjust the parameters.

5.4.1.2 Stochastic gradient descent (SGD):

The stochastic gradient descent algorithm can oscillate along the path of steepest descent towards the optimum. Adding a momentum term to the parameter update is one way to reduce this oscillation [2]. The stochastic gradient descent with momentum (SGDM) update is given by equation 5.8,

$$\theta_{\ell_{+1}} = \theta_{\ell} - \alpha \nabla E(\theta_{\ell}) + \gamma(\theta_{\ell} - \theta_{\ell-1}) \quad (5.8)$$

where γ determines the contribution of the previous gradient step to the current iteration. where ℓ is the iteration number, $\alpha>0$ is the learning rate, θ is the parameter vector, and $E(\theta)$ is the loss function. In the standard gradient descent algorithm, the gradient of the loss function, $\nabla E(\theta)$, is evaluated using the entire training set, and the standard gradient descent algorithm uses the entire data set at once.

Batch gradient descent performs redundant computations for large datasets, as it recomputes gradients for similar examples before each parameter update. SGD does away with this redundancy by performing one update at a time. It is therefore usually much faster and can also be used to learn online. SGD performs frequent updates with a high variance that cause the objective function to fluctuate heavily.

While batch gradient descent converges to the minimum of the basin the parameters are placed in, SGD's fluctuation, on the one hand, enables it to jump to new and potentially better local minima. On the other hand, this ultimately complicates convergence to the exact minimum, as SGD will keep overshooting. However, it has been shown that when we slowly

decrease the learning rate, SGD shows the same convergence behavior as batch gradient descent, almost certainly converging to a local or the global minimum for non-convex and convex optimization respectively. Mini-batch gradient descent is typically the algorithm of choice when training a neural network and the term SGD usually is employed also when mini-batches are used.

5.4.1.3 Momentum

SGD has trouble navigating ravines, i.e., areas where the surface curves much more steeply in one dimension than in another, which are common around local optima. In these scenarios, SGD oscillates across the slopes of the ravine while only making hesitant progress along the bottom towards the local optimum. Momentum is a method that helps accelerate SGD in the relevant direction and dampens.

5.4.1.4 Learning rate

The Learning rate is an important parameter that is used for controlling the size of the update steps along the gradient. Normally, the learning rate sets how much of the gradient should be updated with every step, for example, 1 = 100% but normally much smaller learning rate, e.g., 0.001 is used for training purposes.

It is analogous to a rolling ball, where the calculation is made to find where the ball should roll next in discrete steps. How long these discrete steps are is the learning rate. While training a neural network, it is very important to choose a good learning rate parameter to obtain good results.

A small learning rate is related to small progress and may get stuck in local minima and not reaching the global minima. Whereas, larger steps mean that the weights are changed more every iteration, so that they may reach their optimal value faster, but may also miss the exact optimum. Smaller steps mean that the weights are changed less every iteration, so it may take more epochs to reach their optimal value, but they are less likely to miss optima of the loss function. Learning rate schedule allows you to use large steps during the first few epochs, then progressively reduce the step size as the weights come closer to their optimal value.

5.4.1.5 Optimizer and other parameters used in experiments:

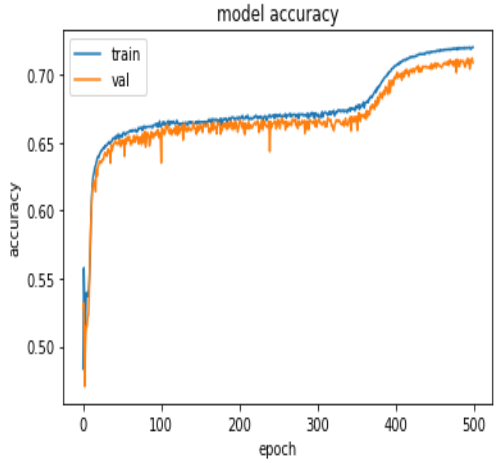
In the proposed method experiments are performed with different combinations of parameters and optimizers to test the performance of the system. The optimizer algorithm and parameters are:

- Adadelta optimizer
    - learning_rate=1.0
    - Adapts learning rates based on a moving window of gradient updates, instead of accumulating all past gradients.
    - Adadelta continues learning.
    - Slow learning, training accuracy not improved more than 60%.
- Stochastic gradient descent(SGD)
    - Learning rate = 0.01, momentum = 0.9.
    - Escape from local minima, speed is better than batch gradient decent.
    - Good results in terms of training accuracy and loss.

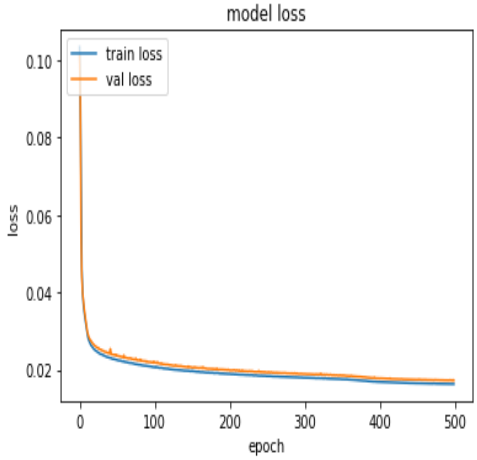
5.5 Experimental Results

In the proposed approach, the experiments are mainly performed using a machine with GPU NVIDIA Titan Xp with an i7 processor, 16 GB RAM for training the autoencoder model. The experiments are done with python 3.6 with OpenCV library for video processing and deep learning architecture.

The activation function used here is rectified linear unit (Relu) and the optimizer is a stochastic gradient descent method with a learning rate=0.01, momentum=0.9 in training. The model has experimented well with different combinations of optimizers, epochs, and other hyperparameters to optimize performance. Model performance in terms of Accuracy and loss is for the initial experiment with epochs 100 and batch size 5 is evaluated shown in Figure 5.8.



(a)



(b)

FIGURE 5.8 Model Accuracy and Loss for 100 epochs and 5 batches

FIGURE 5.9 Training Accuracy Improvement with increasing epochs

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 for the dataset used as displayed in Figure 5.9.

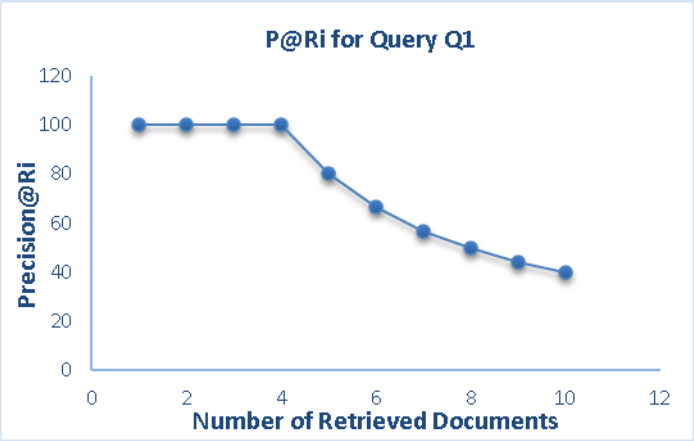


FIGURE 5.10 Performance Evaluation in P@Ri of the proposed approach for query Q1

The query image is processed similarly for feature extraction tasks as training frames of video. The similarity between query features and features stored in the dataset is done for retrieving similar clips from the dataset. Figure 5.10 shows the performance of the proposed approach to CBVR for query 1 in terms of P@Ri values obtained from top k results retrieved. The results of a query and the first four retrieved video clip results are shown in Figure 5.11. To calculate mean average precision, precision $P@R_i$ is calculated for i=1..k for queries in the set Q = (Q1, Q2, ..., Q_k) as shown in Table in one of the experiments on query set of size k=10.

Major Benefits

With this model input video frames with noise and poor resolution also can be matched with queries of a similar type to retrieve relevant video stories from the dataset. Image features extracted with convolution layers of the autoencoder architecture performs better with noise and low-resolution videos.

To evaluate the performance of the proposed CBVR system. P@Ri, i=1..k measure is used for the set of queries for k=10 as shown in Figure 5.12. Finally, the mean average precision MAP is calculated from the average precision values obtained from the query set. In proposed approach, MAP 91.5 is obtained. Figure 5.13 shows the average precision value obtained for the query set of size 10. The proposed system is tested on a machine configured with NVIDIA GPU titan Xp received as a grant for research work.

FIGURE 5.11 Results of News Story clip Retrieval for the given Query image

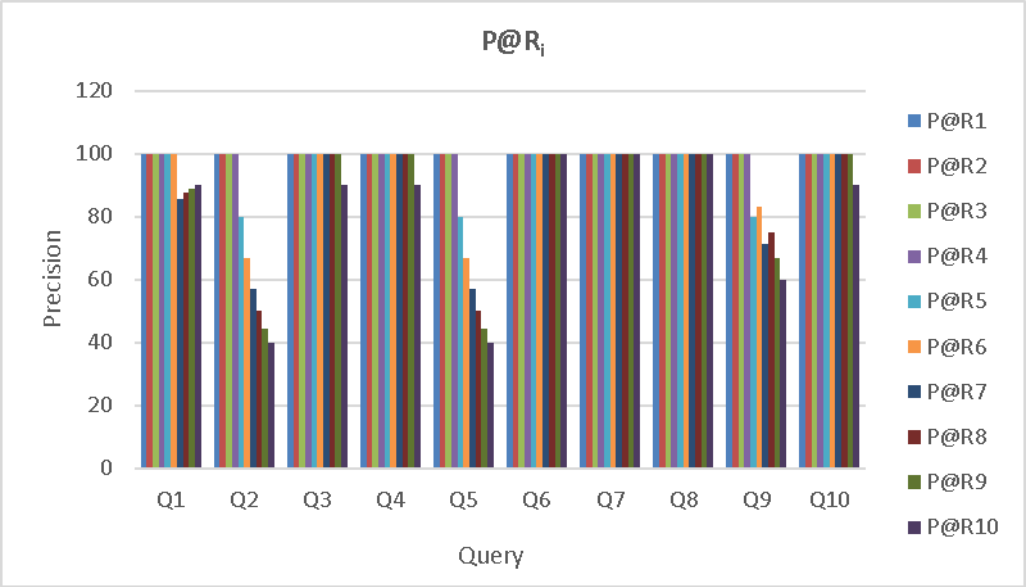


FIGURE 5.12 Performance of proposed CBVR approach

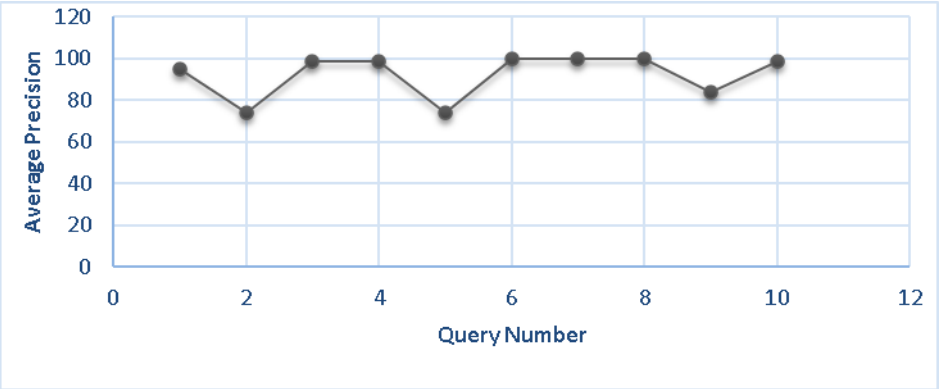


FIGURE 5.13 Performance of Image Query based video retrieval approach (CBVRAPP2)

5.6 Comparisons

As shown in Figure 5.14, the proposed model outperforms the first approach for the retrieval task. The proposed system is mainly divided into three important phases: the first is keyframe extraction and advertisement removal, the second phase is feature extraction and the third phase is indexing and retrieval task. The query of two types is used in two different approaches. In the first approach, text query in the Gujarati language is used for the retrieval task, and text extracted from the news stories displayed on the screen is the feature used for indexing documents. This kind of text-based retrieval on large scale is not found in the literature for Gujarati or any regional language including the Hindi language. So, the comparison is mainly done with the existing system based on the size of data and response time for the data. Scene text retrieval based on Hindi, Bangla, etc. language text is found in the literature but retrieval of video clips and their final performance outcome is not found in the literature for comparison purposes

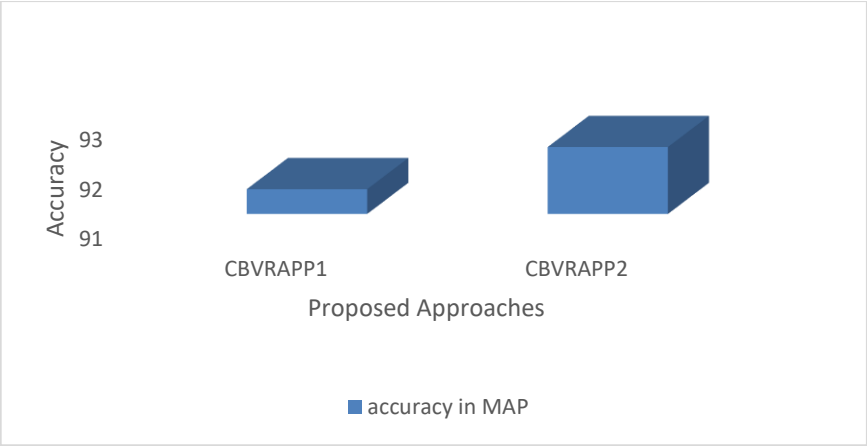


FIGURE 5.14 Comparison of proposed approaches CBVRAPP1 and CBVRAPP2

.

The mean average precision obtained is 91.5 percent with CBVRAPP1. Also, the proposed approach CBVRAPP2 is better than the first approach. With deep learning autoencoder architecture as feature extractor and image query for large scale video retrieval from large collection video input dataset, CBVRAPP2 yields a mean average precision of 92.35 percent which is better in comparison to CBVRAPP1.

Chapter 6

Conclusion and Future Scope

This chapter describes the summary of the thesis and points out the probable expansion for future work. The chapter is separated into two sections. The first section 6.1 concludes the thesis by mentioning the major contributions and also the limitations of the system proposed while the second section 6.2 presents future directions for further research.

6.1 Conclusion

The content-based video retrieval system presented in the thesis primarily involves three steps. First is the extraction of frames and finding keyframes for each shot from input video as well as reducing total dataset size by eliminating advertisements using the proposed advertisement classification model. The second task is feature extraction. The third task is indexing and retrieval.

One of the key contributions of research work is the keyframe extraction algorithm. Another key contribution is advertisement classification which is used for the detection and removal of advertisement tasks. The proposed keyframe extraction algorithm gives a comparatively good compression ratio of 0.9931 which compressed from 29,70,000 frames to 80752 keyframes which reduced overall processing time for all subsequent steps to a great extent. Keyframe extraction followed by advertisement detection and removal further eliminated all frames with advertisements from the collection. Classification accuracy with different models with deep learning architecture is ranging from 98.9 to 99.74 with different parameters and combinations of feature extractor models as well as classification models. Best accuracy is obtained with the proposed deep learning architecture which also takes less response time too.

Another key contribution is the “Gujarati” text query-based video retrieval approach. With this approach text features are extracted and processed with a natural language library

implemented for the retrieval task to extract meaningful words for indexing. Retrieval performance is 91.5 MAP.

To achieve better performance of retrieval tasks, deep learning-based CBVR is also implemented. The final key contribution is CBVR implemented using an unsupervised learning deep autoencoder network for image query-based retrieval tasks. The best accuracy measured is a 92.35 MAP among all the experiments performed. The accuracy achieved with the deep learning-based approach is slightly better compared to the text-based retrieval task but deep learning-based implementation required more time for training as well as requires high-performance computing and GPU for faster processing.

6.2 Limitations and Future Scope

6.2.1 Limitations:

The major problem faced in implementing the model is the time required to train the large data and the infrastructure to support such training. With existing resources and data, experiments are performed to achieve better results which are mentioned in the thesis. The work can be further analyzed for a very large dataset on a specific cluster of deep learning GPUs to reduce the overall training time for the model.

6.2.2 Future Scope

The methods presented in this thesis have experimented with Gujarati Language Query on the News Story Retrieval task. Future work that needs to be carried out includes the following:

- This work opens an opportunity for many interesting problems in this domain like News story retrieval based on a query for famous personalities.
- Also, the work can be adapted easily for other types of videos like lecture videos, sports videos, movies, and video songs where the Gujarati text is used.
- The work can also be used to explore other regional language scene text retrieval or similar task.
- Also, Lot of scopes exist in the NLP domain for Gujarati language processing where the input data is available or extracted in very raw form with a lot of extraneous symbols.

- The deep learning approach for image query-based retrieval can also be further explored with changes in the model or other parameters for the retrieval or feature extraction task.

List of References

[1] E. Bureau, "'Digital Videos: India to Overtake the US on Time Spent on Digital Videos.,'" *The Economic Times, Economic Times*, Sep. 2019.

[2] E. Martinec, "Deloitte Insights, Deloitte's Technology, Media & Telecommunications Group, 8 Dec.," May 2020. Accessed: Dec. 08, 2020. [Online]. Available: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions.html.

[3] H. Zhang, "Content-Based Video Analysis, Retrieval and Browsing," Springer, Berlin, Heidelberg, 2003, pp. 27–56.

[4] Z. Xiong, Y. Rui, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "9.2 - A Unified Framework for Video Summarization, Browsing, and Retrieval," in *Handbook of Image and Video Processing (Second Edition)*, Second Edi., A. L. BOVIK, Ed. Burlington: Academic Press, 2005, pp. 1013–1029.

[5] R. V. Head, "Univac: A Philadelphia Story," *IEEE Ann. Hist. Comput.*, 2001, doi: 10.1109/85.948906.

[6] M. Sanderson and W. B. Croft, "The History of Information Retrieval Research," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, May 2012, doi: 10.1109/JPROC.2012.2189916.

[7] 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.

[8] M. Fekihal, I. Jaluta, and D. K. Saini, "TB±tree: Index structure for Information Retrieval Systems," in *2015 2nd International Conference on Computer Science, Computer Engineering, and Social Media, CSCESM 2015*, 2015, doi: 10.1109/CSCESM.2015.7331890.

[9] F. Kounelis and C. Makris, "Space Efficient Data Structures for N-gram Retrieval," *AIMS Med. Sci.*, 2017, doi: 10.3934/medsci.2017.4.426.

[10] L. M. Boitsov, "Using signature hashing for approximate string matching," *Comput. Math. Model.*, 2002, doi: 10.1023/A:1016014301288.

[11] D. Belazzougui and G. Navarro, "Alphabet-independent compressed text indexing," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, doi: 10.1007/978-3-642-23719-5_63.

[12] R. Grossi and J. S. Vitter, "Compressed suffix arrays and suffix trees with applications to text indexing and string matching," *SIAM J. Comput.*, 2006, doi: 10.1137/S0097539702402354.

[13] Y. Chen, "On the general signature trees," in *Lecture Notes in Computer Science*, 2005, doi: 10.1007/11546924_21.

[14] Y. Chen, "Building signature file hierarchies into object-oriented databases," in *Proceedings of the 6th IASTED International Conference on Software Engineering and Applications, SEA 2002*, 2012.

[15] 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.

[16] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1168–1181, 2007, doi: 10.1109/TIP.2006.891352.

[17] R. Kannao and P. Guha, "Segmenting with style: detecting program and story boundaries in TV news broadcast videos," *Multimed. Tools Appl.*, 2019, doi: 10.1007/s11042-019-7699-9.

[18] 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.

[19] V. Naik and S. Savalagi, “Textual Query Based Sports Video Retrieval By Embedded Text Recognition,” *Int. J.*, 2013.

[20] M. D. A. Asif, U. U. Tariq, M. N. Baig, and W. Ahmad, “A novel hybrid method for text detection and extraction from news videos,” *Middle - East J. Sci. Res.*, 2014, doi: 10.5829/idosi.mejsr.2014.19.5.21019.

[21] T. Tuna, J. Subhlok, L. Barker, S. Shah, O. Johnson, and C. Hovey, “Indexed Captioned Searchable Videos: A Learning Companion for STEM Coursework,” *J. Sci. Educ. Technol.*, 2017, doi: 10.1007/s10956-016-9653-1.

[22] K. Davila and R. Zanibbi, “Whiteboard Video Summarization via Spatio-Temporal Conflict Minimization,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2017, doi: 10.1109/ICDAR.2017.66.

[23] H. Yang and C. Meinel, “Content based lecture video retrieval using speech and video text information,” *IEEE Trans. Learn. Technol.*, 2014, doi: 10.1109/TLT.2014.2307305.

[24] G. Quellec, M. Lamard, G. Cazuguel, Z. Droueche, C. Roux, and B. Cochener, “Real-time retrieval of similar videos with application to computer-aided retinal surgery,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2011, pp. 4465–4468, doi: 10.1109/IEMBS.2011.6091107.

[25] Z. Rasheed and M. Shah, “Detection and representation of scenes in videos,” *IEEE Trans. Multimed.*, 2005, doi: 10.1109/TMM.2005.858392.

[26] D. Zeng, Y. Yu, and K. Oyama, “Audio-Visual Embedding for Cross-Modal MusicVideo Retrieval through Supervised Deep {CCA},” *CoRR*, vol. abs/1908.0, 2019, [Online]. Available: http://arxiv.org/abs/1908.03744.

[27] W. Contributors, “Gujarati Language (Wikipedia),” *Wikipedia, The Free Encyclopedia.*, 2020. https://en.wikipedia.org/wiki/Gujarati_language (accessed Dec. 08, 2020).

[28] C. Stephens, “Understanding Image Noise in Your Film and Video Projects,” *Premium Beat*, 2018. https://www.premiumbeat.com/blog/understanding-film-video-image-noise/ (accessed Dec. 08, 2020).

[29] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Prenctice Hall, 2007.

[30] W. Contributors, “Image noise - Wikipedia.” https://en.wikipedia.org/wiki/Image_noise (accessed Dec. 08, 2020).

[31] E. Martinec, “Noise, Dynamic Range and Bit Depth in Digital SLRs (S/N and Exposure Decisions),” May 2008. Accessed: Dec. 08, 2020. [Online]. Available: https://theory.uchicago.edu/~ejm/pix/20d/tests/noise/noise-p3.html#ETTR.

[32] H. Chudasama, N. Patel, “Event based video summarization for cricket’’, PhD Thesis, Charotar University of Science and Technology, 2020.

[33] N. Mohd Ali, N. K. A. Md Rashid, and Y. M. Mustafah, “Performance comparison between RGB and HSV color segmentations for road signs detection,” in *Applied Mechanics and Materials*, 2013, doi: 10.4028/www.scientific.net/AMM.393.550.

[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] P. Liu, J. M. Guo, K. Chamnongthai, and H. Prasetyo, “Fusion of color histogram and LBP-based features for texture image retrieval and classification,” *Inf. Sci. (Ny).*, 2017, doi: 10.1016/j.ins.2017.01.025.

[36] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, “OCR binarization and image pre-processing for searching historical documents,” *Pattern Recognit.*, 2007, doi: 10.1016/j.patcog.2006.04.043.

[37] D. Saravanan, Vaithyasubramanian, and K. N. J. Vengatesh, “Video content reterival using historgram clustering technique,” in *Procedia Computer Science*, 2015, doi: 10.1016/j.procs.2015.04.084.

[38] 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.

[39] 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.

[40] 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.

[41] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A survey on visual content-based video indexing and retrieval,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. 2011, doi: 10.1109/TSMCC.2011.2109710.

[42] X. Qi, C. Liu, and S. Schuckers, “IoT edge device based key frame extraction for face in video recognition,” in *Proceedings - 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2018*, 2018, doi: 10.1109/CCGRID.2018.00087.

[43] T. Zlitni, B. Bouaziz, and W. Mahdi, “Automatic topics segmentation for TV news video using prior knowledge,” *Multimed. Tools Appl.*, 2016, doi: 10.1007/s11042-015-2531-7.

[44] U. Patel, P. Shah, and P. Panchal, “Shot Detection Using Pixel wise Difference with Adaptive Threshold and Color Histogram Method in Compressed and Uncompressed Video,” *Int. J. Comput. Appl.*, 2013, doi: 10.5120/10625-5347.

[45] N. J. Janwe and K. K. Bhoyar, “Video shot boundary detection based on JND color histogram,” in *2013 IEEE 2nd International Conference on Image Information Processing, IEEE ICIIP 2013*, 2013, doi: 10.1109/ICIIP.2013.6707637.

[46] S. Meng and H. Jiang, “The video shot boundary cut detection based on color histogram,” *J. Comput. Theor. Nanosci.*, 2016, doi: 10.1166/jctn.2016.4975.

[47] Z. Cerneková, C. Nikou, and I. Pitas, “Shot detection in video sequences using entropy-based metrics,” in *IEEE International Conference on Image Processing*, 2002, doi: 10.1109/icip.2002.1038995.

[48] 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.

[49] B. Zhang, T. Li, P. Ding, and B. Xu, “TV commercial segmentation 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.2012.6324579.

[50] 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.

[51] S. Xu, B. Feng, and B. Xu, “Multi-modal topic unit segmentation in videos using conditional random fields,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, doi: 10.1109/ICASSP.2013.6638062.

[52] X. Wang, L. Xie, M. Lu, B. Ma, E. S. Chng, and H. Li, “Broadcast news story segmentation using conditional random fields and multimodal features,” in *IEICE Transactions on Information and Systems*, 2012, doi: 10.1587/transinf.E95.D.1206.

[53] 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.

[54] 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.

[55] 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.

[56] X. S. Hua, L. Lu, and H. J. Zhang, "Robust learning-based TV commercial detection," in *IEEE International Conference on Multimedia and Expo, ICME 2005*, 2005, doi: 10.1109/ICME.2005.1521382.

[57] N. Liu, Y. Zhao, Z. Zhu, and H. Lu, "Exploiting visual-audio-textual characteristics for automatic TV commercial block detection and segmentation," *IEEE Trans. Multimed.*, 2011, doi: 10.1109/TMM.2011.2160334.

[58] 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.

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

[60] M. Kos, Z. Kačič, and D. Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features," *Digit. Signal Process. A Rev. J.*, 2013, doi: 10.1016/j.dsp.2012.10.008.

[61] 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.

[62] M. Goyani, N. Dave, and N. M. Patel, "Performance analysis of lip synchronization using LPC, MFCC and PLP speech parameters," in *Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010*, 2010, doi: 10.1109/CICN.2010.115.

[63] N. Dave, "Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition," *Int. J. Adv. Res. Eng. Technol.*, 2013.

[64] 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.

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

[66] 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.

[67] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, 1995, doi: 10.1023/A:1022627411411.

[68] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1, no. 2. MIT press Cambridge, 2016.

[69] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, "A Convolutional Autoencoder Approach for Feature Extraction in Virtual Metrology," *Procedia Manuf.*, vol. 17, pp. 126–133, 2018, doi: https://doi.org/10.1016/j.promfg.2018.10.023.

[70] 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.

[71] 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.

[72] 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.

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

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] C. Jawahar, B. Chennupati, B. Paluri, and N. Jammalamadaka, “‘Video retrieval based on textual queries.,’” in *Proceedings of the thirteenth international conference on advanced computing and communications, Coimbatore. 2005.*, 2005.

[83] 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.

[84] M. Porter, “The Porter Stemming Algorithm,” Program, Vol. 14 No.3, pp. 130-137, 1980.

[85] P. Willett, “The Porter stemming algorithm: Then and now,” *Program*, 2006, doi: 10.1108/00330330610681295.

[86] W. Kraai, “Porter’s stemming algorithm for Dutch,” *Informatiewetenschap*, 1994.

[87] J. B. Lovins, “Development of a stemming algorithm,” *Mech. Transl. Comput. Linguist.*, 1968.

[88] I. L. T. P. and D. Centre, “TDIL,” *Indian Language Technology Proliferation and Deployment Centre,- Home, Ministry of Electronics & Information Technology , MeitY, Govt.* http://tdil-dc.in/index.php?lang=en.

[89] T. E. P. (Enabling M. L. E. The EMILLE Corpus, “https://www.lancaster.ac.uk/fass/projects/corpus/emille/,” *Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. Distributed by the European Language Resources Association.,* 2003. www.lancaster.ac.uk/fass/projects/corpus/emille/ (accessed Dec. 18, 2018).

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

[91] 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.

[92] 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.

[93] 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.

[94] H. Patel and B. Patel, "Stemmatizer—Stemmer-based Lemmatizer for Gujarati Text," in *Emerging Trends in Expert Applications and Security*, Springer, 2019, pp. 667–674.

[95] U. Mishra and C. Prakash, "MAULIK : An Effective Stemmer for Hindi Language," *Ijcse*, 2012.

[96] N. Saharia, U. Sharma, and J. Kalita, "Stemming resource-poor Indian languages," *ACM Trans. Asian Lang. Inf. Process.*, 2014, doi: 10.1145/2629670.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

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

[102] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767, 2018.

[103] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, doi: 10.1109/CVPR.2016.91.

[104] A. Mukhtar, M. J. Cree, J. B. Scott, and L. Streeter, "Gait Analysis of Pedestrians with the Aim of Detecting Disabled People," *Appl. Mech. Mater.*, 2018, doi: 10.4028/www.scientific.net/amm.884.105.

[105] Q. Peng *et al.*, "Pedestrian detection for transformer substation based on Gaussian mixture model and YOLO," in *Proceedings - 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2016*, 2016, doi: 10.1109/IHMSC.2016.130.

[106] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," *Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 396–404, 1989.

[107] 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.

[108] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science," *Towards data Science*. pp. 1–12, 2018, Accessed: Feb. 01, 2021. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[109] Hegde, Vishakh, and Sheema Usmani. "Parallel and distributed deep learning." Tech. report, Stanford University., 2016.

[110] M. Rhu, N. Gimelshein, J. Clemons, A. Zulfiqar, and S. W. Keckler, "VDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design," in *Proceedings of the Annual International Symposium on Microarchitecture, MICRO*, 2016, doi: 10.1109/MICRO.2016.7783721.

[111] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *J. Physiol.*, vol. 148, no. 3, p. 574, 1959.

[112] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 8614–8618.

[113] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep Clustering with Convolutional Autoencoders," 2017, pp. 373–382, doi: 10.1007/978-3-319-70096-0_39.

[114] P. Kulkarni, B. Patil, and B. Joglekar, "An effective content based video analysis and retrieval using pattern indexing techniques," in *2015 International Conference on Industrial Instrumentation and Control, ICIC 2015*, 2015, doi: 10.1109/IIC.2015.7150717.

[115] S. Padmakala, G. S. AnandhaMala, and M. Shalini, "An effective content based video retrieval utilizing texture, color and optimal key frame features," in *ICIIP 2011 - Proceedings: 2011 International Conference on Image Information Processing*, 2011, doi: 10.1109/ICIIP.2011.6108864.

[116] J. Kong and C. Han, "Content-based video retrieval system research," in *Proceedings - 2010 3rd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2010*, 2010, doi: 10.1109/ICCSIT.2010.5565041.

[117] B. V. Patel, A. V. Deorankar, and B. B. Meshram, "Content based video retrieval using entropy, edge detection, black and white color features," in *ICCET 2010 - 2010 International Conference on Computer Engineering and Technology, Proceedings*, 2010, doi: 10.1109/ICCET.2010.5486262.

[118] G. Haolin and L. Bicheng, "A novel signature for fast video retrieval," in *Proceedings - 4th International Congress on Image and Signal Processing, CISP 2011*, 2011, doi: 10.1109/CISP.2011.6099932.

[119] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[120] GM, "'why-is-there-no-closed-captioning,'" *Blog*, 2006. Glad-to-hear.blogspot.in/2006/03/why-is-there-no-closed-captioning-in.html. (accessed Dec. 08, 2018).

[121] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, 2015, doi: 10.1007/s11263-015-0816-y.

[122] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[123] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Feb. 2016, Accessed: Jan. 16, 2021. [Online]. Available: http://arxiv.org/abs/1602.07360.

[124] G. Abhinav, "Deep Learning Reading Group: SqueezeNet," 2016. https://www.kdnuggets.com/2016/09/deep-learning-reading-group-squeezenet.html (accessed Jan. 16, 2021).

[125] C. Crawford, "ResNet-18 | , ResNet-18 Pre-trained Model for PyTorch," *Kaggle*, 2018. https://www.kaggle.com/pytorch/resnet18 (accessed Jan. 16, 2021).

[126] Keras, "ResNet-50 | Kaggle," 2018. https://www.kaggle.com/keras/resnet50 (accessed Jan. 16, 2021).

[127] keras.io, "VGG16 and VGG19." https://keras.io/api/applications/vgg/ (accessed Jan. 16, 2021).

[128] K. Murphy, "Machine Learning: A Probabilistic Perspective.," *The MIT Press, Cambridge,*

Massachusetts., 2012. https://mitpress.mit.edu/books/machine-learning-1 (accessed Jan. 21, 2021).

[129] MathWorks, "Options for training deep learning neural network - MATLAB training Options." https://in.mathworks.com/help/deeplearning/ref/trainingoptions.html (accessed Jan. 21, 2021).

[130] N. Dave, M. Holia, "Advertisement Detection in broadcasted videos using Transfer Learning and Support Vector Machine," in *International Conference on Research and Innovations in Science, Engineering & Technology*, 2020.

[131] A. Rosebroc, "keras_conv2d_num_filters." https://pyimagesearch.com/wp-content/uploads/2018/12/keras_conv2d_num_filters.png.

List of Publications

[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”, ICRISET2020 (International Conference on “Research and Innovations in Science, Engineering& Technology”, September 2020.

[5] Namrata Dave, Mehfuza Holia, “Advertisement Detection Using Transfer Learning and Support Vector Machine”, Solid State Technology Journal. (Scopus Indexed Journal - Selected)