Video Indexing and Retrieval Taking Human Face as a Cue

Thesis Submitted

by

Sanjoy Ghatak

DOCTOR OF PHILOSOPHY(Engineering)

Department of Computer Science and Engineering

Faculty Council of Engineering & Technology

Jadavpur University,

Kolkata-700032, India

2025

JADAVPUR UNIVERSITY

KOLKATA-700032, INDIA

INDEX NO: 139/22/E

REF. NO.: D-7/E/421/22

1. TITLE OF THE THESIS:

Video Indexing and Retrieval Taking Human Face as a Cue

2. NAME, DESIGNATION & INSTITUTION OF THE SUPERVISORS:

Dr. Debotosh Bhattacharjee
PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JADAVPUR UNIVERSITY, KOLKATA-700032

3. LIST OF PUBLICATIONS:

a) JOURNAL:

i) **Ghatak, S**., Bhattacharjee, D. Video indexing through human face images using LGFA and window technique. *Multimedia Tools & Appl* **81**, 31509–31527 (2022). https://doi.org/10.1007/s11042-022-12965-2 (**SCI Journal) (IF -3.0**)

ii) **Ghatak, S.,** Bhattacharjee, D. "Video Indexing Through Human Faces by Combined Deep Learning Neural Networks", Journal of Information Systems Engineering and Management Vol. 10, 653-670 (2025). https://doi.org/10.52783/jisem.v10i16s.2653 (Scopus Journal) (Impact Score 1.26)

iii) **Ghatak, S**., Bhattacharjee, D. "Using Deep Learning algorithms, video indexing through the human faces represented as EAN-8 Linear bar code." Journal of Signal Processing: Image Communication (Submitted in SCI Journal)

b) PATENT:

i) **S. Ghatak, D. Bhattacharjee,** "A System and Method for Barcode representation of face images", **Australian Patent, Status: Granted, Year 2022.**

c) CONFERENCE:

i) **Ghatak, S.** (2018). Facial Representation Using Linear Barcode. In: Bhattacharyya, S., Chaki, N., Konar, D., Chakraborty, U., Singh, C. (eds) Advanced Computational and Communication Paradigms. Advances in Intelligent Systems and Computing, vol. 706. Springer, Singapore. https://doi.org/10.1007/978-981-10-8237-5_76(Scopus Index)

ii) **Ghatak, S**., Bhattacharjee, D. (2020). Barcode Representation of Face Image Combining LGFA and Windowing Technique. In: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds) Human Interaction and Emerging Technologies. IHIET 2019. Advances in Intelligent Systems and Computing, vol. 1018. Springer, Cham. https://doi.org/10.1007/978-3-030-25629-6_73(Scopus Index)

iii) **Ghatak, S**., Bhattacharjee, D. (2021). Video Indexing Through Human Face. In: Sabut, S.K., Ray, A.K., Pati, B., Acharya, U.R. (eds) Proceedings of International Conference on Communication, Circuits, and Systems. Lecture Notes in Electrical Engineering, vol. 728. Springer, Singapore. https://doi.org/10.1007/978-981-33-4866-0_13.(Scopus Index)

iv) **Ghatak, S**., Kollman, C., Bhattacharjee, D. (2024). Video Indexing Through QR Code of Human Faces Using MTCNN Algorithm. In: Das, N., Khan, A.K., Mandal, S., Krejcar, O., Bhattacharjee, D. (eds) Proceedings of International Conference on Data, Electronics and Computing. ICDEC 2023. Lecture Notes in Networks and Systems, vol. 1103. Springer, Singapore. https://doi.org/10.1007/978-981-97-6489-1_1.

v) **Ghatak, S**., Battacharjee, D. "**Video indexing and retrieval through human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8."** Paper Presented in 2nd International Conference on Advances in Communication Networks & Systems (COACONS-2025) 26-27th March, 2025 organized by Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur. (Waiting for Published in Scopus Index AIP Book)

d) BOOK CHAPTER

Ghatak, S., Bhattacharjee, D. "A review of state-of-the-art in video indexing and retrieval using human faces as cues from video surveillance systems." Book chapter Published in Book "Advances in Computer Science" Volume - 25", Page No. 123-152, Paperback ISBN:978-93-6135-605-6,2025. https://doi.org/10.22271/ed.book.3088

4. **LIST OF PUBLICATION OUTSIDE Ph. D. WORK:**

a) **JOURNAL:**

i) **Ghatak, S.,** "Key frame extraction Using Threshold Technique", International Journal of Engineering Applied Sciences and Technology, 2016, Vol. 1, Issue 8, ISSN No. 2455-2143, Page 51-56.

ii) **Ghatak, S.,** "Extraction of Key Frame from News Video Using Face Recognition", International Journal of Advanced Technology in Engineering and Science, Vol. No. 3, Special Issue No. 01, November 2015.

iii) **Ghatak, S.,** Bhattacharjee, D. "Extraction of Key Frames from News Video Using EDF, MDF, and HI method for News Video Summarization", International Journal of Engineering and Innovative Technology, 2013

iv) Banerjee, S., Singh, K, A., **Ghatak, S**., Kumar, S., "News Video Indexing System using Inserted –Caption Detection and its Retrieval", International Journal of Digital Image Processing, 2011, Print: ISSN 0974 – 9691 & Online: ISSN 0974 – 9586

v) **Ghatak, S.**, Pradhan, A., Khandelwal, D., Tamang, P.L., "News Video Indexing and Retrieval Using Combination of S.A.D and E.C.R. Scoring Techniques" International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 11, November 2012, ISSN 2250-2459

vi) Rai, P., **Ghatak, S**., "Bag of Visual Words", International Journal of Computer Science and Engineering, 2016, Vol. 1, Issue 5, Page 19-26.

b) CONFERENCE:

Sunuwar, J., Borah, S., Agarwal, **S., Ghatak**, S. (2022). A Study on Hand Gesture Segmentation Approaches. In: Gandhi, T.K., Konar, D., Sen, B., Sharma, K. (eds) Advanced Computational Paradigms and Hybrid Intelligent Computing. Advances in Intelligent Systems and Computing, vol. 1373. Springer, Singapore. https://doi.org/10.1007/978-981-16-4369-9_61(scopus)

c) BOOK PUBLISHED:

Ghatak, S., Kumar, S., "News Video Indexing and Retrieval Method", Lambert Academic Publishing, German, Page 76, (Online Book), ISBN-10: 3659384755, ISBN-13 : .3659384752-978

PROFROMA-1

Statement of Originality

I .... SANJOY GHATAK..... registered on 01/06/2022.... do hereby declare that this thesis entitled "Video Indexing and Retrieval taking Human Faces as Cue" contains literature survey and original research work done by the undersigned candidate as a part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 7%.

Sanjoy Ghatak

Signature of Candidate:

Date: 30/06/25

30.06.2025

Dr. Debotosh Bhattacharjee
Professor
Dept. of Computer Sc. & Engg.
Jadavpur University
Kolkata - 700032

Certified by Supervisor(s):

(Signature with date, seal)

PROFORMA -2

CERTIFICATE FROM THE SUPERVISOR/S

This is to certify that the thesis entitled "Video Indexing and Retrieval taking Human Face as a Cue", submitted by Sri. Sanjoy Ghatak, who got his name registered on /01/06/22 for the award of Ph. D. (Engg.) degree of Jadavpur University is absolutely based upon his own work under the supervision of Prof. Debotosh Bhattacharjee (Dept. of C.S.E., JU) and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

30/06/2025

Signature of the Supervisor

and date with official Seal

Dedicated to my Parents and Family Member

Acknowledgements

I am deeply indebted to many people who have contributed to making the completion of this thesis work possible. Firstly, I would like to express my earnest gratitude and respect to my supervisor **Prof. Debotosh Bhattacharjee** for his relentless guidance, tremendous support, endless care, invaluable suggestion, and constant monitoring during this work.

I would also like to express my sincere respect to Prof. Mita Nasipuri, Former Coordinator of Centre for Microprocessor Application for Training, Education and Research (CMATER) Laboratory, Jadavpur University(JU), Prof. Mahantapas Kundu, Coordinator CMATER Laboratory, JU, Prof. Ram Sarkar, Prof. Subhadip Basu and Prof. Nibaran Das of Department of Computer Science and Engineering, Jadavpur University, for providing me constant inspiration, support and useful suggestions during the course of this thesis work.

I would like to thank Prof. Nirmalya Chowdhury, HOD, Department of Computer Science and Engineering, Jadavpur University for making available all the departmental facilities and for her valuable comments.

I am also grateful to the Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, for providing me No Objection Certificate for completing my Ph. D. work in Jadavpur University.

I am also thankful to my co-authors and co- researchers Prof. Christian Kollmann, Medical University of Vienna for their valuable suggestions and constant feedback that have made a remarkable contribution to my Ph. D. work.

Finally, I owe my encompassing debt to my parents, my family member and some very close friends whose love, support, encouragement, good wishes and inspiration enabled me to complete this thesis.

Date: 25/08/25

Sanjoy Ghatak

Place: Jadavpur University

(Sanjoy Ghatak)

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

A video index that describes the video content is required to browse, search, and work with video documents. It serves as the basis for applications such as multimedia-rich digital libraries and filtering algorithms [1], which automatically search for relevant video documents based on user profiles. The indexes need to be as complete and extensive as feasible to accommodate these diverse applications. Until now, documentarists have mostly created indexes by hand, assigning a limited number of keywords to each video. Due to the specialized nature of the work, manually indexing video documents is time-consuming and expensive. Automatic classification of video content is therefore required. This process is known as "video indexing." This method automatically identifies video documents with content-based metadata [2].

1.1 . Video Representation

The definition, analysis, and summarisation of videos will all be covered in this section.

1.1.1 Video

The word "video" (from the Latin verb "videre," meaning "to see") is used to describe a variety of moving picture storage formats, including analog video tapes such as VHS and Betamax, as well as digital video formats like Blu-ray Disc, DVD, QuickTime, and MPEG-4. Various physical media can be used to capture and transmit video, including MPEG-4 or DV digital media, when recorded by digital cameras or magnetic tape, such as PAL or NTSC electric signals by video cameras. To put it another way, any video saved in a certain format is simply a collection of numerous picture shots recorded one after the other with very little time between each frame. In essence, a video consists of sequences that can be further divided into

consecutive shots, each composed of a frame. (Figure 1.1)

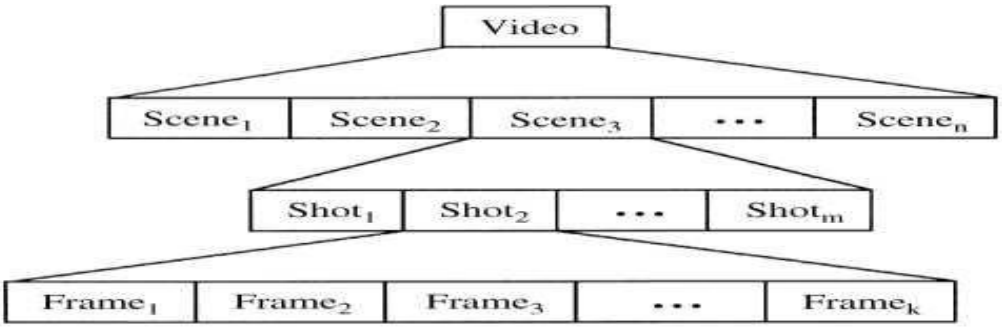


Figure 1.1: The hierarchical structure of a Video [3]

1.1.2 Video Analysis

Given that, a video can be thought of as a collection of image frames, examining a video can be thought of as examining each image separately and then integrating the findings. As covered in [4], this analytical idea is known as Shot Boundaries Detection. Video analysis can use some of the techniques used for image and audio analysis. For instance, specific representative frames taken from video clips can be subjected to image analysis and retrieval techniques. Video can be searched and analyzed in various ways. There are difficulties with each of these approaches.

1.1.3 Video Summarization

The practice of extracting abstract representation from a video to meaningfully condense its essence is known as video summarization. Pictorial summarisation, which is constructed from a few chosen video frames, is the most basic type of video summarisation [5]. In other words, a subset of the original video's keyframes or highlights must be extracted, which serve as entries for shots, scenes, or narratives. The outcome of the abstraction process serves as the foundation for content-based video indexing and browsing, as well as the representation of video content. Using data from several sources, such as speech, sound, video transcript, and image analysis, is an effective strategy.

1.2 . Indexing videos from video documents using a variety of modalities

When creating a video document, the author of the study [6] considers the following three modalities or information channels:

• **Visual Approach:** incorporates the scene setting, or all of it, whether it is produced organically or artificially in the video record;

• **Acoustic mechanism:** encompasses voice, music, and ambient sounds that are heard in the document video;

• **Documented communication mode:** contains textual materials that describe the content of the video document;

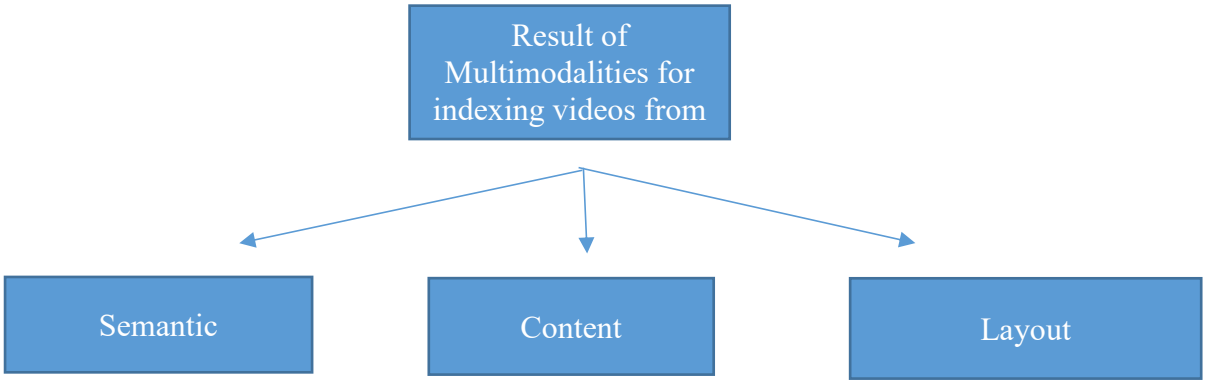


Figure 1.2 illustrates how multi-modality invariably leads to three views for each modality: semantic, content, and layout. [8]

According to Paper [6], multimodality invariably leads to three viewpoints for each modality: layout, content, and semantics. This thesis explores the significance of visual content, including human faces, as an indexing method for footage from video surveillance systems.

1.2.1 Index of Semantics

Granularity can differ between defined segments and is defined as a meaningful unit of the descriptive coarseness of multimodal information [7]. To model this level of information, the authors define segments in a five-level semantic index structure. The entire video document is pertinent to the initial three levels. The conclusion that a writer creates a video with a certain goal is the foundation for the greatest level. The next two tiers define sections based on

consistent display or arrangement of content components. The next level of our semantic index hierarchy is linked to content segments and is determined by logical units and name events. Please be aware that named events cannot have a temporal duration of zero.

1.2.2 Content

From a content standpoint, segments relate to the elements a writer uses when creating a video document. The following components can be distinguished [9].

• **Setting:** the time and place where the video's story takes place, which may also emphasize the tone or ambiance;

• **Objects:** noticeable moving or stationary parts in the video clip;

• **People:** people portrayed in the document video;

Logical units and settings are usually connected. People and objects are the main constituents of named events. The author of the video document has control over the display of various content elements thanks to modality-specific style elements. For the visual modality, a writer may use certain colors, lighting, camera angles, distances, and movements. Auditory style includes the qualities of loudness, rhythm, and melody. Its phraseology and writing style determine the text's look. The author's meaning is communicated through each of these stylistic elements.

1.2.3 Layout

The author's selection of the video document's syntactic structure is considered from the layout standpoint. The syntactic structure of each modality lacks a temporal dimension and is merely a chronological list of basic units. The main feature that distinguishes the different modalities is the nature of these units. The visual modality of a video document consists of a sequence of framed images. Individual picture frames are, hence, the fundamental building blocks.

1.3 . Video document segmentation for indexing purposes

The authoring process should be reversed to segment video materials for analysis and review. A video document should be divided into sections based on its structure and content. The results can be used to index certain segments. Many people consider segmentation to be a

form of classification. The video indexing literature suggests several heuristic methods. The most advanced techniques explicitly utilize pattern recognition. Consequently, we will begin by discussing the different categorization methods used in video indexing. Next, we'll discuss rebuilding the layout for each modality. Finally, we will focus on content segmentation.

1.3.1 Identification of pattern

For video indexing, the layout and content categories must be determined by identifying patterns of interest. Sub-images, samples, or characteristics derived from layout and content elements are a few examples of these patterns. According to [10], statistical classification, neural networks, syntactic or structural matching, and template matching are the four most successful techniques for pattern recognition.

Statistical classification: The distribution of patterns in the space covered by pattern features is used to classify the pattern that has to be identified.

Neural networks: A network that learns nonlinear input-output interactions uses the pattern to recognize input.

Syntactic or structural matching: the pattern that has to be found is compared to a small set of taught schema and grammatical rules for combining primitives;

Template matching: By comparing the pattern to be recognized with a learned template, template matching enables scale and posture modifications;

1.3.2 Rebuilding the layouts

Finding sensor shots and transition edits in the video data is known as layout reconstruction. For analysis, the layout must be recreated. The arrangement should guide analysis as it guides the viewer's experience with the video content. Several methods exist for recreating the visual layout when determining the shot boundary, such as segmenting a video document at the camera shot level. Several techniques for identifying video segments have been proposed in the literature on video indexing. These techniques use a fixed or dynamic threshold to compare successive frames at the pixel, edge, block, or frame level. The time required for decompression is decreased when coding in MPEG since block-level properties are extracted directly from the image channel using motion vectors.

1.3.3 Segmenting content

Subsection 1.2.2 introduced the content elements. This thesis will discuss how to automatically discover them using several detection algorithms that utilize visual data sources, such as human faces.

1.3.3.1 . Recognizing People

People can be identified in video records using various methods. In the visual modality, they can be recognized by their faces or other body parts. More details on techniques specific to the visual modality based on the human face are provided in the following sections. We refer to the specified references for a detailed explanation of the visual techniques. Most visual modality approaches make it easier for people to be identified when they recognize a human face. Regardless of their three-dimensional position, orientation, or lighting conditions, face detection approaches aim to identify all image regions that contain faces. If a face is found, they also offer the image's location and extent [11]. This detection is not straightforward due to changes in position, orientation, scale, and attitude. This detection is not straightforward due to changes in position, orientation, scale, and attitude.

Over the years, numerous methods for identifying faces in pictures and image sequences have been published; see [11] for a thorough and critical review of these methods. More significantly, the neural network-based approach detects about 90% of frontal and upright faces, only infrequently mistaking non-face regions for faces. Face detection is currently commonly used in deep learning and machine learning algorithms. Suggested in the chapter of the book [8]. In the research, the author proposes a novel face detection network that addresses three key features of face identification: anchor-based data augmentation, progressive loss design, and improved feature learning [12]. The research author claims that Blaze Face is a high-performing, portable face detector built for mobile GPU inference [13]. It runs at 2000–1000 frames per second or faster on flagship smartphones.

Due to its rapid performance, it can be utilized with any augmented reality workflow that requires a specific facial region of interest as input for task-specific models, such as 2D/3D facial keypoint or geometry estimates, facial feature or emotion categorization, and face

region segmentation. According to the paper [14], YOLO v3 can perform face detection tasks quickly, as it requires only one neural network to predict bounding boxes and class probabilities. YOLO v3 (You Only Look Once, version three) is an end-to-end neural network approach that simultaneously predicts bounding boxes and class probabilities. This contrasts with the conventional approach taken by previous object detection systems, which modified classifiers to satisfy detection specifications.

The research authors [15] propose a deep, cascaded multi-task architecture that leverages the inherent connection between tasks to enhance face detection efficiency. Specifically, this approach employs a three-stage cascaded structure with carefully designed deep convolution networks that make coarse-to-fine predictions about landmark location and face. The computationally efficient CNN architecture ShuffleNet was developed by Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun [16], specifically for portable devices with limited processing power, such as those with 10–50 million floating-point operations per second (MFLOPs). Point-wise group convolution and channel shuffle are two novel operations in the new architecture that drastically lower computing costs without compromising object detection (facial) accuracy. Among all the well-known designs for small devices, Shuffle Net is ranked as one of the best in this article. However, some important frames, such as faces, are lost during real-time face recognition. As a result, facial recognition accuracy is decreased when a shuffle net is the only tool used.
Facial recognition algorithms attempt to identify a person whose face appears in a video. Several face recognition algorithms, such as traditional feature-based algorithms, require human-designed features, while others utilize deep learning-based algorithms. The following algorithms are some of the most popular.

Eigenfaces: One popular method for face recognition is matching using Eigenfaces [17]. Principal component analysis is the method used by this algorithm to extract information from the facial image. In this instance, a single image is used for matching, and the process can recognize faces in various positions. It is still used in some situations and was one of the early facial recognition algorithms.

Fisher faces: An expansion of Eigenfaces; this approach considers the class labels of the face pictures. The authors of [18] demonstrate that using Fisher faces lowers the error rates in tests

on particular face databases. Additionally, the Fisher face approach is most effective when changes in expression and lighting occur simultaneously.

Deep Face: This algorithm extracts features from facial images using a deep convolutional neural network. It was one of the first algorithms to achieve human-level performance.

FaceNet: This approach utilizes a triplet loss function to learn a mapping from face photos to a high-dimensional feature space. Several visage recognition standards have achieved best-of-class results, including the LFW, Age DB, CFP-FP, and IJB-C datasets.
Deep learning methods, such as Convolutional Neural Networks (CNNs), are the most accurate for face recognition, whereas lightweight models like MobileNets are optimized for speed, particularly on mobile devices. However, the requirements of a particular application often determine the best course of action. One drawback of face recognition as a video indexing technique is its limited generalisability [19]. Results have shown that face recognition is effective in confined spaces, particularly when a face is positioned directly in front of the camera [17, 18, 19]. This limited applicability should be considered when applying facial recognition techniques to video indexing.

1.4. Background

The technologies that enable people to record and distribute digital video data easily are advancing rapidly and becoming increasingly accessible today. While high-speed and dependable networking has shifted towards mobile and wireless access, personal computers continue to become faster, smaller, and less expensive. The days of watching and editing video, which required only a television and a ridiculous number of cords, are long gone. The Internet and portable devices are now widely used for creating and sharing video documents. Because video contains a wealth of content for numerous important applications, its use as one of the most popular media formats has increased significantly. Further improvements are needed on existing Content-Based Video Retrieval (CBVR) systems to support and sustain the expansion of video content.

The capacity of a video document to provide a rich semantic presentation through synchronized text, visual, and audio presentations over time is its most distinctive feature. The

human face is one of the most crucial visual elements in a content-based video retrieval system. Since face detection and tracking from video automatically identify and locate the face region in input frames, they are essential parts of face recognition systems. Typically, the face recognizer receives the necessary features from the located face region. A brief overview of research on video indexing and retrieval using human faces as cues is presented in this section.

The literature on face identification and retrieval in video is reviewed by Caiffeng Shan [20]. The aforementioned practices undergo a comprehensive analysis and resolution. Using the frame difference methodology, the paper [21] extracts key frames from a video clip. The MTCNN method is then utilized for face detection and alignment, and the Face Net model is used for face recognition to recover all the keyframes in which the person appears in the video. Essential staff can easily find and modify video frames, saving time and effort. However, this method has problems with both erroneous and absent detection. The entire condensed binary code is generated using CNN's deep video code (DVC) framework, which takes facial videos as input. The trained DVC outperforms state-of-the-art hashing techniques on two generic image and video datasets, as well as three challenging face video datasets, for various image and video retrieval tasks. They attribute three things to their success: First, the three phases are compatible by integrating hash coding, video-level modeling, and frame-level feature learning into a unified framework. Second, optimizing a smooth upper bound on the triplet loss function for hash learning prevents the model from slipping into a suboptimal local optimum or converging. Third, using the designed video modeling schemes, the intricately crafted attention mechanism aggregates video information. For example, the weighted temporal average pooling and max-pooling can mine complementary information from various frames. Currently, only the first-order statistic of the image collection is used for video modeling in this method. The deeper integration of higher-order statistics, such as second-order pooling, is a topic for future study, as discussed in this paper. According to this analysis, they also believe a strong need exists to create bigger and more challenging face video datasets to study and evaluate complex end-to-end frameworks.

A method for fast video retrieval based on the output of face detectors and recognizers was introduced in the publication [22]. The proposed technique is quick and reliable, based on a convolution-like video content similarity computation and fully utilizing database indexing. A system that uses face detection and recognition to index real videos has been employed to

evaluate the retrieval performance of the proposed technique. The technique demonstrates that face-related data possess sufficient discriminant power to index and retrieve videos. It may be

possible to modify the suggested face-based method to index video by using individuals' appearances from various modalities or even other classes of objects with distinct identities.

1.5. Motivation

As discussed in subsection 1.2 of this thesis, the following three information channels or modalities—visual approach, acoustic mechanism, and documented communication mode—were seen in a video document and served as the impetus for this thesis. With applications ranging from biometric identification to video search and retrieval, visual surveillance, and human-computer interaction, face recognition in video has generated significant interest over the last decade. Despite significant advancements, the issue remains challenging to resolve for videos captured in unrestricted settings. A further research study discusses the need for "Video indexing and retrieval taking human faces as cues" to address various problems. This work improves the following aspects: (1) increases the size of the training data set to improve recognition accuracy; (2) uses new deep and machine learning algorithms to improve the accuracy of face detection;(4) aims to address significant challenges such changing position, having a limited amount of storage space, losing crucial video frames, and the intricacy of time and place; and (3) look into ways to use a faster and better face recognition technique to increase retrieval speed and accuracy. This effort will address the primary challenges outlined in several studies, including the use of databases, the quality of video data, and computational costs. According to many publications, there is also a great need to build a large and challenging face video collection to study and evaluate more complex end-to-end frameworks. This attempt is, therefore, more significant.

1.6 . The significance of employing human faces as cues in video indexing

Video usage for many significant purposes has increased dramatically due to technological improvements. Video will be one of the biggest issues facing education and information

technology in the future. To boost the effectiveness of their content on the web and generate new revenue streams, future content owners, publishers, and educators will need to deliver video to users in ways that leverage established web economic models and validate the demand, both qualitatively and quantitatively, for future internet usage patterns. Video processing technology is receiving more attention to reduce network transfer stress, analyze video data efficiently, and extract reliable information from videos based on human faces. Key frame extraction and video segmentation significantly reduce the data used in video processing. "Video indexing and retrieval using the human face as a cue" has consequently gained more attention in research. This research aims to identify unlawful activity in military or armed forces utilizing video surveillance systems. Aside from that, this discovery is crucial for various video face identification applications. The aforementioned study is necessary to enhance effective communication and communication speed in a communication channel. This thesis's primary goal is to develop a video indexing and retrieval technique that utilizes a person's face as a cue to aid in identifying them across various video clips. The work enhances the following features:

(a) Major issues, including posture changes, storage capacity limitations, the inability to save crucial video frames, and the complexity of time and space, are all addressed in this thesis.
(b) To improve recognition accuracy, the size of the training data collection must be increased.
(c) To increase the precision of facial detection.
(d) Examine how to use a faster and more effective face recognition technique to increase retrieval speed and accuracy.

1.7. Various cutting-edge techniques for indexing and retrieving videos from video surveillance systems that use human faces as signal

As digital technology advances and social networking and web streaming become increasingly widespread, more individuals can alter video objects and desire to use them for an increasing number of reasons. Researchers are particularly interested in camera faces since they resemble human faces daily. As a result, a face is now regarded as an essential object for video indexing. Researchers have attempted to develop new and improved video indexing

methods that utilize human faces as cues on multiple occasions over the last few decades. Each of these methods will be explained in this section.

1.7.1 Applying Face Recognition and Detection Methods to Video Indexing

The author of the paper [23] discussed face detection and identification techniques for video indexing in great depth. This research proposes a face detection and identification-based video indexing method. The figure-1.3 shows the block diagram of this method.

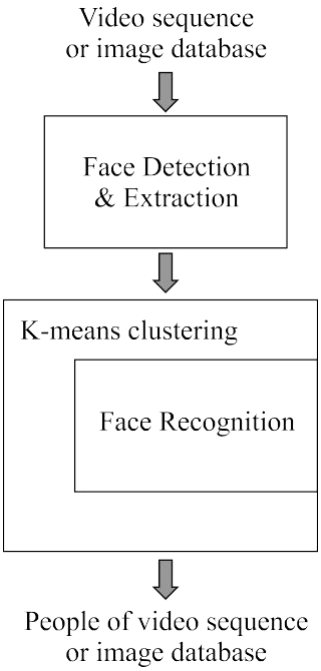


Figure 1.3 Applying Face Recognition and Detection Methods to Video Indexing [23]

The initial step of this method involves searching for faces in video clips using a face detector built on neural networks. The faces are retrieved from the sequence when they have been located. Then, the k-means clustering approach is applied in conjunction with a facial recognition method that clusters faces using pseudo-two-dimensional Hidden Markov Models. Images of a single person's face make up each generated cluster. The recognized faces in the video sequence are then categorized as belonging to different individuals, enabling an evaluation of their frequency of occurrence. Results of the proposed method on a TV news

segment are presented. The system might identify three different newsreaders and the interviewee. The experiments were conducted using a test corpus of representative broadcast news.

1.7.2 Face Recognition and Clustering for Video Indexing Applications

The author details video indexing, including face detection and clustering approaches [24]. This paper presents a method for automatically recognizing human faces in random video footage. To give a certain level of certainty about whether faces are included or excluded in video footage, the creator of this work used an iterative algorithm. Skin color filtering is applied to a fixed number of frames in each video capture, followed by shape and size heuristics. The remaining candidate regions are assessed following normalization and projection into an Eigenspace, where the reconstruction error serves as a measure of confidence for the existence or absence of a face. The confidence score of the complete video clip is then determined. They use an incremental procedure that combines a PCA-based dissimilarity measure and spatiotemporal correlation to classify extracted faces into a set of face classes.

1.7.3 Face Recognition and Shot Transition in video indexing

The author of the paper [25] explained a method for video indexing that recognizes facial features and shot changes. The proposed method operates as follows: first, the video is processed through frame analysis, which utilizes the Kullback-Leibler (KL) divergence to identify the various shots. In particular, every shot's first frame, last frame, and duration, both in frames and seconds, are recorded. An architecture of cascaded convolutional neural networks (CNNs) is utilized to track and extract facial images within the same video scanning procedure. A pre-trained ResNet-50 model, trained on the VGGFace2 dataset [26], is used after extracting face features. The distance cosine between characteristic vectors is computed and used as a metric to assess the probability that two faces belong to the same person. When faces in various camera shots match, a stricter threshold is applied to reduce errors; if faces are in the same shot, a more lenient threshold is used. The proposed pipeline recognizes and re-identifies people in each image, automatically indexing subject-dependent video. There are no limitations on appearance, including eyeglasses, beards, or haircuts, when processing faces. The addition and use of a shot transition detector have confirmed the results.

1.8. Scope of the thesis

Video data usage has permeated every aspect of our daily lives, from traditional radio broadcasting and entertainment to camera systems for enhancing intelligent urban settings, wearable technology, and applications in medical and transportation. In addition to the vast video collections, video indexing enables systems to efficiently and effectively handle,

organize, and preserve video data, allowing users to quickly satisfy their basic information needs. There are special accessibility issues with video content because indexing and searching videos have always been exceedingly time-consuming. Images also include much redundant content, but videos preserve rich substance. Furthermore, for effective use, video processing and analytics require significant computer complexity, including browsing and recovery. In many respects, retrieving related films from a large collection is more challenging than retrieving images. Manually recovering videos is a tedious and time-consuming task for humans.

Furthermore, as people are prone to errors, there is a chance that the results will be unsuitable. While system efficiency improves in a highly competitive network environment and the load of retrieving video features is reduced, the system also requires a significant amount of bandwidth and resources. One pressing problem that needs to be addressed is how to quickly and effectively process and understand videos. One simple technique for searching and recording large amounts of video data is to remove video keyframes. Keyframes can be extracted from videos using various methods. The key frame of a movie can be extracted from it utilizing text, audio, and visual material (such as pictures or faces).

The primary topic of this thesis is the use of human faces as keyframes for video indexing. Face detection in videos is a challenging task, as evidenced by the analysis of existing techniques for indexing videos using human faces. Angular variations of the face, changes in posture, illumination-invariant aspects, time and spatial complexity, and the inability to save key frames from videos are the primary sources of this issue. This thesis presents a novel approach to video indexing that utilizes various machine learning and deep learning algorithms, along with human faces as cues, to address these challenges. Using a variety of

machine learning and deep learning techniques along with human faces as cues, this thesis explored a novel approach to video indexing that addresses these issues.

A general overview of a common video indexing method, which utilizes human faces as cues, is displayed in Figure 1.4.

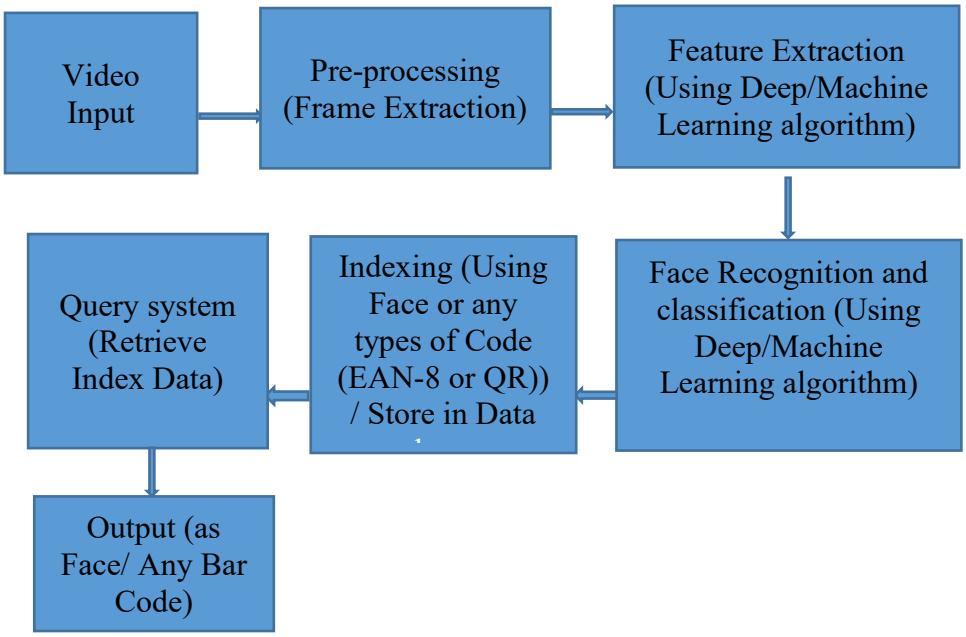


Figure 1.4 A General overview of a common video indexing method using human faces as cues. [8]

A block diagram for each stage of a popular video indexing technique that utilizes human faces as cues is shown in Figure 1.4.

(a) The initial stage in indexing a video using a human face as a cue is pre-processing the incoming video by removing the frame. Faces are recognized as frames from the input video.

(b) Faces identified by machine learning or deep learning algorithms are subjected to features (key frames).

(c) Machine learning or deep learning algorithms are used to recognize or classify faces.

(d) Faces are then indexed as standard faces or unique barcodes and stored in a database.

(e) Then, when required, obtain the index data from the Query System.

(f) The output can be bar codes of any type or photos of faces.

This dissertation's primary goal is to utilize machine learning and deep learning methods to develop automated video indexing and retrieval systems that leverage human faces as cues. They employ Viola Jones, MTCNN, DSFD, Blaze-face, YOLO v3, YOLOv5s, and YOLOv8n, among others, for face detection and frame extraction from the input video. The color histogram approach, combined with various machine learning and deep learning algorithms,

extracts key frames from a sequence of images. These essential frames are then stored in a database as a standard human face and a linear EAN-8 bar code or QR code to index and retrieve videos. A diagrammatic illustration of the primary focus of this thesis can be seen in Figure 1.5.

The automated system for video indexing using face Images and Barcodes is covered in Chapter 3 of this dissertation. An EAN-8 linear bar code represents the human face image used for video indexing in the study. This dissertation uses EAN-8 linear bar code to construct an automated system for video indexing based on human facial cues, as illustrated in Figure 1.5. An automated framework for video indexing using human faces as cues via QR codes is created in Chapter 4 of this dissertation. The concept of creating facial image QR bar codes from video keyframes was considered in this study. As seen in Figure 1.5, this dissertation builds an automated system for video indexing based on human face cues using QR codes.

Chapter 5 of this dissertation covers an automated system for video indexing that utilizes human faces as signals, employing various machine learning and deep learning techniques. Using several machines and deep learning algorithms, this study used human faces as keyframes for video indexing from input videos. Figure 1.5 illustrates how this dissertation

uses various machine learning and deep learning algorithms to create an automated system for video indexing based on human facial cues.

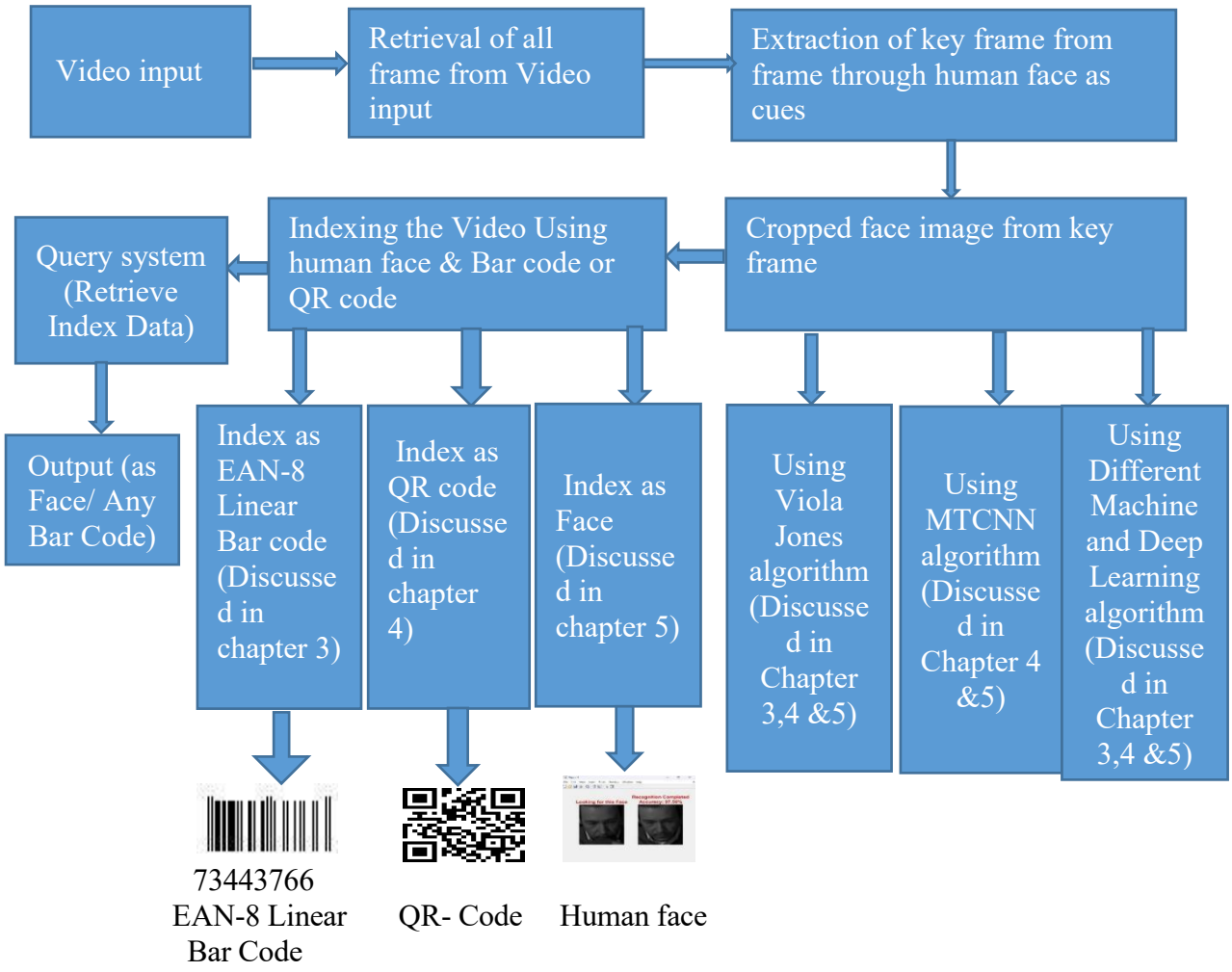


Figure 1.5 Representing Block diagram showing the scope of the thesis.

Therefore, the primary contribution of this dissertation can be summarized as follows.

- Pre-processing and frame extraction from the input video Dynamic video combines the frame and the recorded scene. Thus, extracting still images from input videos, represented as scenes, shots, and pictures, is the first stage. A sequence of filmed and shot frames make up the scene. The frame is a still image containing extraneous information in a video clip. Redundant frames are extracted from the incoming video

using pre-processing. Instead of extracting frames from the input video for the experiment, direct frames are taken from several video datasets.

- Key frame extraction from the frame- The key frame captures the essential details of each shot. The keyframes in this study feature human faces with unique expressions, postures, lighting conditions, and illumination. The consequences of the curve saliency motion capture, the likelihood scale, a few parallels, and additional techniques. However, in this dissertation, the keyframe is extracted from each frame of a particular video using the Colour Histogram approach. Suppose the observed difference exceeds the threshold's magnitude. Once the color histogram discrepancy is set to the threshold, the frame will be chosen as the following keyframe.

- Face image cropped from the key frame - Several machine learning and deep learning algorithms (Viola-Jones, MTCNN, ShuffleNet, Combined MTCNN and ShuffleNet, DSFD, BlazeFace, YOLOv3, YOLOv5s, and YOLOv8n) are used to detect faces from the extracted keyframes. Human faces with distinctive expressions, age, postures, lighting circumstances, and illumination directional changes can be clipped from key frames using these machine and deep learning algorithms. The "Holly Wood Video Data Set, YouTube Face Video Dataset, TV Series Video Database, Face Video Dataset FDDB, Face Video Dataset WIDER, Face Video Dataset LFW, Face94," "YaleB Face Database," "Face database FERET," and FG-NET dataset, among others, are used to test this approach. To determine the gradient of the face image, the faces from keyframes must be converted to greyscale. Image gradients are computed from this greyscale image (faces) using distinct techniques, including LGFA and the sliding window approach.

- Using human faces as cues to index the incoming video, a unique QR code, a linear EAN-8 bar code, and identification from the key frame of the input video are created and saved as an index for each face. The benefit of human face barcode indexing is that it takes less time and storage space to index these barcodes. The facial picture gradient is scanned horizontally by linear bar codes and vertically and horizontally by QR codes. As a result, the QR code has fewer facial features missing than the linear

bar code. However, compared to linear barcodes, QR codes require greater storage space.

- Remove the index data from the query system. Next, retrieve the index data from the Query System as needed. In addition to enabling consumers to readily satisfy their fundamental information needs, it allows systems to efficiently and successfully handle, organize, and store video data.

1.9. Thesis Organization

The issue of video indexing using human faces as cues has been thoroughly examined in this thesis. As mentioned previously, studying all existing techniques for indexing videos using human faces reveals that face detection in videos is a challenging task. Angular variations of the face, changes in posture, illumination-invariant aspects, time and spatial complexity, and the inability to save key frames from videos are the primary sources of this issue. This dissertation proposes a novel approach to video indexing that leverages human faces as cues and utilizes a variety of machine learning and deep learning algorithms to identify faces in input videos, thereby overcoming these difficulties. For video indexing, all faces are represented as linear bar codes and QR codes after facial image detection from the input video. All important frames are arranged in the database as faces, linear bar codes, and QR codes. Numerous methods for key frame extraction from videos, face detection from keyframes, facial image gradient computation, linear barcode production, and QR code generation have been developed during the research. The thesis is organized as follows.

Chapter 2: Literature Survey - This chapter thoroughly examines the recent state-of-the-art video indexing and retrieval method, which utilizes human faces as cues. These methods include text-based video indexing, content-based video indexing, key frame extraction based on human faces, gradient calculation of face images, key frame selection, and video indexing using Deep Learning and Machine learning techniques.

Chapter 3- Video Indexing through the Face Images Using Barcode-There are numerous face recognition systems for video indexing, but none of them can simplify time and space. A new method for handling distinct data from video, utilizing a person's face, is presented in this

chapter. It can be ordered and recovered at any crucial moment. The following developments are the main focus of this work: key frame extraction from the information video, face recognition from keyframes, face distinguishing proof using standardized identification, and barcode-based video sorting. In this work, we first identify a video as a source of information and then extract each edge from it. The keyframes are selected from each frame in the video using the color histogram difference. A color histogram is utilized to determine the difference between two frames. If the threshold of the color histogram is met, the keyframe is returned as a face from the input video. Faces are recognized from the main frames extracted using the Viola-Jones, MTCNN, Shuffle Net, Combined MTCNN and Shuffle Net, DSFD, Blaze-face, and YOLO v3. The human face retrieved from the video is identified using a unique EAN-8 barcode for video indexing reasons, and the image gradient is produced from the face image using the sliding window technique and the combined LGFA and sliding window approach. This work aims to recognize human faces in videos and convert them into EAN-8 linear video indexing barcodes. This method works well for security, video surveillance networks, and communication channel video descriptions, among other applications. It is helpful for indexing and retrieving videos. This reduces the storage space, time complexity, and communication bandwidth. After characterizing each unique human face in the movie, this method indexes the video as a linear EAN-8 barcode depending on the human face in the video.

Chapter 4 - Video Indexing through Face Images using QR Codes suggests a method for creating facial image QR barcodes from a video's keyframes. Various methods exist for using a person's face as a cue for video indexing. Nevertheless, due to factors such as changes in the face's direction, brightness, and illumination, none of the techniques effectively identify faces in videos. Additionally, there is no compact representation of faces, even if an effective detection method were to be achieved. A linear bar code that loses much information when scanned horizontally is a compact representation.

Additionally, the accuracy of facial detection affects the system's overall accuracy. This chapter features human faces with a QR code that can be scanned to bypass it. This technique extracts key frames from videos that contain faces using Multi-Task Cascaded Convolutional Neural Networks (MTCNN). A QR bar code is created from the key frame once it has been detected for video indexing and recognition. This technology is helpful for security purposes,

Human activity recognition, video surveillance, and video description of communication channels, among other applications. This approach has outperformed the Viola-Jones method for low-brightness face, illumination-invariant face, and directional change face recognition when indexing videos using a QR barcode with MTCNN.

Chapter 5- Video Indexing through the Face Images using Deep Learning Models- A technique for indexing a video that uses a face image as a keyframe is proposed in this chapter. Numerous methods for indexing videos using human faces as cues have already been mentioned; however, none can recognize faces in videos due to factors such as changes in the face's direction, variations in image brightness, illumination, and small size in the scene, among others. Real-world scenarios often present small target faces with various challenges, including intricate backdrops, occlusion, and scale shifts. This leads to the problem of face detection results being missed or misidentified. To address all of these problems, a novel method called "Video indexing through the human face as a cue using Deep Learning model" is employed in this study and covered in this chapter. This problem is resolved by extracting keyframes, or the human face, from video frames using Shuffle Net, Combined MTCNN and Shuffle Net, YOLOv3, YOLOv5, and v8. Videos use this keyframe for indexing and identification. The primary use case for this research is the identification of small human faces in videos and their subsequent application in video indexing and retrieval. This technique is useful for various applications, including communication channel description, human activity recognition, security, and video surveillance.

Chapter 6-Conclusion- In the last chapter, the key points of this thesis are outlined, and the outcome is discussed. Additionally, the scope of future research on the works presented in this study is also highlighted in this chapter.

Chapter 2

Literature Review

2.1. Introduction

Video indexing and retrieval are essential for managing, storing, and retrieving multimedia data, allowing users to access resources quickly and efficiently [27]. One important and dynamic type of multimedia information is video. It typically has the following traits: (1) a vast amount of raw data, (2) more vital individual image information, and (3) very little preceding organization [28]. Due to this characteristic, indexing and retrieving videos can be challenging. The video database was comparatively smaller in prior decades, and manual keyword annotations were used for extraction and indexing [29]. These databases have gained popularity recently, and as a foundation, they require content-based video retrieval to automatically analyze videos with minimal human involvement. Large amounts of data can now be accessed more quickly and easily due to recent advancements in information and communication technologies. In March 2020, COVID-19 led to a 700% increase in daily YouTube video uploads and a 210% increase in views at home compared to the equivalent levels before the pandemic [30]. It is estimated that there will be approximately 5.3 billion internet users and 891 million 4K TV connections by 2023 [31], indicating an increase in both the volume of video traffic and its quality. Additionally, there is a wide range of applications for content-based video retrieval, including faster video folder browsing, an examination of visual electronic commerce (such as user ordering and selection analysis, correlation analysis [39] among influences and advertisements), digital museums, remote instructions, intelligent web video management, news event analysis (video search and video tracing), and video surveillance. As a result, there has been active research on video applications [32]– [36].

Since video data is inherently complex, developing management strategies requires a deep understanding of its particular features [37, 38]. Video is distinguished from other data classes by a few key features. First, video has a higher resolution, a larger data volume, and a larger

set of data that can be interpreted. However, it also presents more interpretation ambiguity and requires more interpretation work than alphabetic data because it is stored as binary. Second, although text is non-spatial and static, images are spatial and static, while videos have both temporal and spatial dimensions. Furthermore, video semantics typically contains complex relationships and is unstructured.

Semantic content and audio-visual (AV) presentation are the two primary elements of a video document. Since the semantic information in a video can be expressed explicitly or implicitly, it is the most complex component of the video data. After watching or listening to the audio-visual presentation, viewers must use their prior knowledge to comprehend the implicit semantics, but they ought to be able to understand the explicit semantics more naturally. The language that is shown to viewers in a film-type video to introduce the cast is an example of explicit semantics. Viewers can extract perceptual audio and visual information, including color, texture, pitch, volume, object motions, and object relationships, from AV records.

Additionally, some text displays are included in visual presentations that are long enough to provide users with specific information about what the audio-visual presentation is currently showing. For instance, in certain news video shots, text is displayed to tell viewers about the current topic, the participants, the location of the event, and other details. A closed caption (CC) track, which includes sequential text that must be displayed in sync with the audiovisual track, is another feature of some broadcast videos. Therefore, the CC track is referred to as the spoken word script or subtitles.

Low-level (perceptual) characteristics and high-level (semantic) annotation are the two primary layers of video content that can be used to classify video indexing techniques [38, 40, 41, 42]. Low-level feature-based indexing algorithms have the following primary advantages:

- With the use of feature extraction methods, such as image and sound analysis, they can be completely automated.
- By utilizing specific feature qualities, such as the color and shape of items within a frame or the volume level of the soundtrack, users can employ a similarity search.

Nevertheless, feature-based indexing often overlooks semantic content, even when users prefer to search for videos using semantics rather than low-level attributes. Features that are above the range of perception might make feature-based indexing extremely time-consuming and imprecise. For instance, users may not be able to specify the properties of specific objects they wish to inspect for every query.

High-level semantic-based indexing's primary advantage is its ability to allow more robust, flexible, and natural query formats. For instance, users can search for specific videos using keywords and explore videos using semantic hierarchy concepts, such as topical classification. Due to the importance and popularity of video retrieval and indexing, numerous surveys have been carried out. The survey indicates that there are three main indexing methods: 1) Techniques for feature-based video indexing, such as shot-, object-, and event-based indexing; 2) Video indexing based on annotations; and 3) Indexing strategies that try to close semantic gaps. Feature-based video indexing methods can be divided into groups according to the segments and features that are taken out. Video can also be broken down into a hierarchy equivalent to video storyboards using segment-based indexing techniques [43, 44, 45]. Therefore, according to this Segment-based Indexing framework, a frame is a single image or picture, a shot is a group of frames that share similar attributes, a scene is a group of shots that share a common semantic content, and a story is a group of scenes that tell a single, cohesive semantic story.

A multi-level abstraction is employed in a hierarchical video browser to help users efficiently locate specific video frames or segments. This type of browsing design is frequently referred to as a storyboard, as it includes a series of frames that symbolize the key ideas or moments in the video. By storing keyframes instead of the entire video document, bandwidth and latency requirements for delivering the video contents over a network for viewing and review are decreased [40].

One of the most popular techniques for depicting video segments is to use a series of keyframes to represent each segment, such as a shot, with the hopes that one of the frames will encapsulate the shot's core ideas. This approach is especially useful for viewing video content, as it provides viewers with visual details about each indexed video segment.

Object-based video indexing aims to identify specific objects within a video clip, enabling the tracking of content changes over time. A complex collection of objects, their locations, physical characteristics, and their relationships specifically constitute a video scene. The method of extracting objects is more intricate than that of obtaining low-level attributes, such

as color, texture, and volume. However, when a frame contains a large number of small, moving objects that cause blur, object extraction is typically quite challenging. Certain

circumstances make object detection challenging, such as complex backgrounds, small objects, and numerous overlapping objects, like a human face.

Event-based video indexing tracks the actions of objects to identify events in video segments. The goal of event-based video indexing is to automatically identify noteworthy occurrences from unprocessed video tracks [46]. An event is commonly defined as the relationships between items that exist during a specific period, which may occur before or after another event [47]. In general, event detection is a challenging task.

Keywords or free texts are used to annotate the semantics of video segments in annotation-based video indexing, enabling the management of video content. The following are some significant disadvantages of annotation-based indexing that are already to be anticipated: 1) The choice of keywords and free text is arbitrary and frequently based on the needs of the application and domain 2) Frequently, a picture is worth a thousand words. It is, therefore, assumed that words will be incredibly inadequate to describe a video segment because words are frequently unable to express a single frame adequately. Additionally, in situations where users are unable to articulate their needs through words, they frequently choose to ask using a similar visual or sound. Similar to this, people frequently discover that visual key frame representations are more engaging and beneficial than text-only text when perusing a video document. Notwithstanding these drawbacks, further research into this strategy is necessary because annotations may still be the most accurate representation of the semantic content of videos.

Bridging the Semantic Gap Indexing techniques contribute to bridging the semantic divide between low-level features and high-level concepts. The main focus of this thesis is on features-based video indexing, including shot and object-based indexing, which utilize a person's face as the keyframe in a video shot and treat each face as an object within this keyframe. To understand video indexing using a human face as a cue from any video, this chapter provides a thorough overview of the existing techniques for Video Indexing and retrieval, key frame extraction from the input video, face detection from keyframes, face

recognition from detecting faces, image gradient calculation from detecting faces, linear bar code, and QR code generation from human faces.

2.2. Existing Techniques for Video Indexing and Retrieval

Researchers are interested in the many applications of video retrieval and indexing. Over the past decade, one of the most significant research challenges in the domains of artificial intelligence, computer vision, digital image processing, and natural language understanding

has been content-based video retrieval. The rapid growth of video retrieval has given rise to several research issues, as well as opportunities to design and build a wide range of practical applications, including face video retrieval and recognition, video tagging, video annotation, crime investigation, Web albums, and healthcare. The primary focus of this research is developing effective content-based video retrieval methods for massive video databases. Although there are many more video retrieval methods available, none of them are widely recognized. To establish large-scale content-based video retrieval systems, it is necessary to efficiently solve the high-dimensional indexing problem. Researchers worldwide have proposed numerous state-of-the-art techniques for video indexing. These methods include image video shot boundary detection, key frame extraction, structure analysis, and scene segmentation. They also include the extraction of key frame features, static features, motion features, and object features, as well as video data mining, video annotation, and video retrieval, including similarity measures, query interfaces, relevance feedback, and video browsing. Table 2.1 provides a detailed summary of each of these approaches.

Stefan Clicker et al.'s study [23] provided a detailed examination of face detection and identification techniques for video indexing. This research proposes a face detection and identification-based video indexing method. Using a neural network-based face detector to search for faces in the video clips is the first stage in this process. After faces are located, they are removed from the sequence. Then, a face recognition method is applied that groups faces into clusters using pseudo-two-dimensional Hidden Markov Models in combination with the k-means clustering algorithm. Each cluster is composed of a single image of a person's face. The faces recognized in the video sequence are then categorized as belonging to different individuals, allowing for an evaluation of their frequency of occurrence. Results of the proposed method on a TV news segment are presented. The system might identify three

different newsreaders and the interviewee. The trials were conducted using a sample broadcast news test corpus.

Csaba Czirjek et al. [24] described the use of facial identification and clustering approaches for video indexing. This research proposes an automatic method for recognizing human faces in random video footage. A degree of confidence regarding the presence or exclusion of faces in video footage is provided by the author of this work using an iterative algorithm. After applying skin color filtering to a predefined number of frames in each video capture, shape and size heuristics are used. The reconstruction error serves as a measure of confidence in the existence or absence of a face, and the remaining candidate regions are assessed after normalization and projection into an eigenspace. The confidence score for the full video clip is then determined. They employ a PCA-based dissimilarity measure in conjunction with spatiotemporal correlation, utilizing an incremental approach to classify extracted faces into a set of face classes.

D. Cazzato et al. [25] presented a method for video indexing that recognizes facial features and shot changes. Using the Kullback-Leibler (KL) divergence, the proposed method first processes the video through a frame analysis to identify the various shots. To be more precise, the first, last, and duration in frames and seconds are recorded for every shot. Facial pictures are tracked and extracted using a cascaded convolutional neural network (CNN) framework throughout the same video scanning procedure. A ResNet-50 that has already been trained on the VGGFace2 dataset [26] is used following the extraction of face features. Following that, the distance cosine between the characteristic vectors is computed and used as a metric to evaluate the probability that two faces belong to the same person. To reduce errors, a softer threshold is applied when faces are in the same shot, while a more robust threshold is used when faces match across distinct camera shots. By recognizing and re-identifying people in each image, the proposed pipeline automatically indexes subject-dependent video. Processing faces does not impose any constraints on appearance, spectacles, beards, or haircuts. Utilizing a shot transition detector, the results have been verified.

The authors of the research [47], Noboru Babaguchi et al., suggest event-based video indexing, a type of indexing based on the semantic content of the video. Because video data comprises multiple modal information streams, including textual (closed captioning, CC), visual, and auditory streams, the research presents an intermodal collaboration strategy that involves

collaborative processing, taking into consideration the semantic dependencies between these streams. Its goal is to increase the dependability and effectiveness of video content analysis.

By extracting keywords from the CC stream and indexing shots in the visual stream, the suggested method aims to find periods during which events are expected to occur, with a focus on temporal congruence between the visual and CC streams. Intermodal collaboration is successful for video indexing, as evidenced by events such as touchdowns (TD) and field goals (FG), according to experimental results on broadcast sports footage of American football games.

The Spatio-Temporal Histogram of the Randon Projections pattern, created by Mohamed Mansoor Roomi et al. [48], is based on concatenated HRP descriptors from three orthogonal

planes. Temporal motion attributes and appearance will be captured by the suggested descriptor using a single descriptor. The STHRP pattern is described by Fisher's Linear Discriminant Analysis (LDA), which reduces dimensionality by utilizing discriminant characteristics. Using the bagged tree model, the images were recognized.
A foreground-driven concept co-occurrence matrix and a novel ranked intersection filtering technique have been proposed by Nitin Janwe et al. [49] to enhance semantic concept-based video retrieval. The classifier was built using a combination of asymmetrically trained deep CNNs, FDCCM, and RIF techniques to address the problem of data imbalance. This assessment uses the Precision parameter and the TRECVID dataset.

E. Asha et al. [50] developed a multivariate feature extraction approach for video retrieval. Motion vectors, texture properties, and color were the main subjects of the investigation. First, scene change detection and key frame extraction are done using the color histogram method. The following features are extracted, and the feature vectors are compared with database features using Euclidean distance. Pertinent videos are then retrieved. Additionally, color strings are extracted for color features, texture features are extracted using the LBP operator, and SADs are computed as motion features for a query video. Utilizing numerous features improves retrieval efficiency, and the suggested strategy is superior in capturing both spatial and temporal aspects.

Meinel et al. [51] expected a model to index videos from a large database. Their speech is recognized by the segmentation method, and their texture features are recovered using the OCR approach. Using this technique, keywords are extracted from a content-based search. However, due to the presence of video noise, accuracy is lower.

Li et al. [52] expected an automated system to automatically record, identify, and follow the presenter and projection screen with visible video. To provide a list of representative images that reconstruct the primary presentation structure, this system analyzes the visual content of the monitored screen region to detect slide progressions and extract a high-quality, non-occluded, geometrically compensated image for each slide. The algorithm then identifies the text content and extracts keywords from the slides, enabling keyword-based browsing and video retrieval. Experiments are conducted using keywords and text content. The study of multiple videos suggests that accuracy is inadequate.

Nguyen et al. [53] proposed a method for conducting document analysis. Subject bagging, a technique for text recognition and localization, is applied here to extract text and images from video slides. The set of subjects in this paper's setting is a lecture video, where each subject is represented by a bag of mixed words—both textual and visual—derived from speech recognition and optical character recognition (OCR) engines. On the other hand, this work deals with cross-model and multi-modal video retrieval. There are certain mistakes found in these techniques.

According to Gayathri et al. [54], several drawbacks arise when pre-processing video frames before accessing private video collections. To overcome the pre-processing problems, feature extraction and classification techniques are considered. Here, video indexing with numerous extraction capabilities and the input video frame's dominating frame formation have been anticipated. Frame structures are divided into dominating structures using a fuzzy-based SVM classifier. Color attribute extraction and the multidimensional histogram of directed gradients (HOGs) are used to eliminate texture information from a video clip. However, with this approach, storage capacity is constrained, and the classifiers are unable to focus on video processing applications to identify specific signals.

According to Lin et al. [55], deep neural networks—particularly for facial recognition—have been the subject of extensive research, and profound learning models are frequently employed to identify artifacts. Therefore, this research proposes a deep learning cloud-based video recovery method. After that, it extracts and pre-processes a dataset, removing distorted images and matching the remaining images to produce a suitable dataset for CNN templates. Following development, the final dataset is utilized to pre-train the CNN face recognition models (VGGFace, ArcFace, and FaceNet). However, using this approach, efficiency is not increased, and the system is not upgraded to obtain more datasets.

Li et al. [56] focused on addressing the issues with smart town protection video retrieval in the link management program through single-packet processing. To start, they proposed a traffic location quantization index based on backbone traffic characteristics to assess the traffic region characteristics in the backend messaging quantitatively. The proposal for deep learning-based keyframe abstraction and video retrieval aims to improve the efficiency and precision of video retrieval. To this end, an adaptive key frame selection algorithm is created, key frame features are extracted using the current convolutional neural network architecture, and supervised, semi-supervised, and unsupervised retraining models are constructed. Nevertheless, keyframes are not stored; therefore, this approach does not maintain space and temporal complexity.

Jacob et al. [57] propose a novel approach to video content analysis that utilizes indexing and video storytelling techniques to extract relevant video clips from lengthy videos. Video footage is analyzed, and a video explanation is produced using the video storytelling technique. To ensure that a keyword of set length L can be identified in the shortest time, the wormhole approach is used to build an index using the video description. Video searching algorithms may use this video index to find the pertinent part of the video because the term appears frequently in the keyword search. Users have the option to download and transfer just the necessary video clip rather than downloading and uploading the full video. Thus, time complexity may arise in this approach.

Priya et al. [58] proposed a novel automatic shot-based keyframe extraction method for video indexing and retrieval applications. The frames are first clustered into shots in a sequential manner utilizing the shot frame clustering technique, feature extraction, and the continuity value creation steps of the shot border detection procedure. To extract the keyframes based on

sub-shots, the cluster with the highest dispersion rate is selected for inter-cluster similarity analysis (ICSA). Video indexing and retrieval are performed using extracted keyframes.

In Markos Zampoglou et al. [59], a feature set for content-based video shot classification based on color and motion was provided. A significant number of different classes are demonstrated in the applications of the expected feature set based on TV station archives, where features are obtained from the properties of essential training samples and achieve higher generalization levels. The focus will be placed on both feature refinements and the integration of spatial features for shape and texture, providing a workable method for improving classification strategies for large classes while creating a comprehensive system for classifying TV station archives.

Wali et al. [60] presented a novel high-level feature extraction approach for retrieval based on video surveillance. The results are rudimentary. It may help human operators use context inquiries and react more quickly, which is the main advantage. Applications such as multimedia data mining utilize metadata to help extract motion and object descriptions from keyframes and convert them to XML. Finally, complementary tools extend initial concepts and events to systems by improving system functionalities.

Dutta et al. [61] In addition to creating a phrase for video frames using the same model used for image captioning, the objective is to construct a sentence for an image by identifying features using deep learning techniques. During the creation of the movie, keyframes are extracted from the video by passing it through the keyframe extraction framework built into

the application. The key frames extracted from the video are then entered into the same image captioning model used to generate captions for the pictures. Using the frames taken from the film, captions are generated by feeding the images into a pre-trained model. Nevertheless, one challenge was producing appropriate, meaningful sentences from a vast vocabulary, and another was converting video images into a coherent frame sequence.

According to Krishnaraj et al. [62], despite cloud services offering effective picture indexing, the semantic gap between user queries and the diverse semantics of large databases remains a significant issue. This article will present a visual semantic indexing-based RTI paradigm for cloud-based photography. First, an interactive optimization model is used to create the

standard semantic and visual descriptor space. To determine the best way to search for bigger data sets, the semantic visual space-sharing model is then combined with an RTI architecture. The distributed Spark model is completed with the addition of an online picture retrieval service. Holidays 1M and Oxford 5K are two benchmark datasets that verify the effectiveness of the suggested system in terms of average precision (mAP) and processing time across a range of dataset sizes. Nevertheless, neither machine learning nor computing speed is improved by this approach.

Lorenzo Baraldi and colleagues [63] put out a multimedia prototype system for video indexing and reuse. The video's narrative framework is subsequently broken down into more manageable chunks by AI techniques, allowing users to quickly gain an overview of the content and retrieve a specific segment of the video. Using images, music, discourse, and literary content, a deep learning architecture was created to divide the video into narratives and provide commentary on related major themes, utilizing labels and discourse exchanges extracted from the visual content.

B. -C. Chen et al. [64] suggested a technique for indexing and retrieving video using a sparse representation of Bag-of-Faces. By encoding face tracking as a sparse representation of a single bag of faces, this technique enables the handling of massive amounts of face data with an effective indexing mechanism.

Mingtao Pei [65] developed a face video retrieval system that focuses on the deep learning of binary hash representations. In this work, a deep convolution neural network (deep CNN) was created by the researcher to learn from compact binary representations and discriminative face-to-face video retrieval.

Zouhair Mbarki et al. [66] present a face detection and recognition method for real-time applications. Combining the local binary pattern (LBP) and convolutional neural network (CNN), the proposed algorithm is divided into two stages: the initial phase involves face recognition from the video sequence, followed by face identification after feature extraction using the LBP method. The reason for this is that deep learning methods perform remarkably well on a variety of identification and recognition tasks.

Table 2.1. State-of-the-art techniques for video indexing

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Stefan Clicker et al.'s [23]	The K-Means clustering approach is employed for face identification in this paper, while neural networks are used for face detection from input video clips.	For the experiment, the TV news program was recorded at a frame rate of 5 frames per second and a resolution of 384 x 288 pixels.	The three newsreaders may be accurately assigned the video segments using this procedure.	Face tracking and the identification of cuts and edit effects can further enhance the method provided.
Csaba Czirjek et al. [24]	Videos are analyzed using an iterative color-based face detection and clustering technique.	Using a video collection of Irish news programs from the Físchlár.	The average precision value is 71.50%.	This approach cannot handle large changes in ambient settings, as the underlying PCA is sensitive to scale, rotation, and changes in illumination conditions.
D. Cazzato et. al. [25]	Face Appearance and Shot Transition Detection for Video Indexing	The first dataset comprises several videos recorded during the Spanish Parliament's debating sessions in the Canary Islands. The second dataset (Dataset #2) comprises two videos recorded during therapy sessions at a healthcare institution in Alessano, Italy.	Precession-0.91, Recall - 0.95	The system was unable to distinguish certain angular and occlusions faces.
Noboru Babaguchi et.	Event-based indexing of	Recorded footage of an American	The average precision and	The fundamental

al.[47]	Broadcasted Sports Video by Intermodal Collaboration using Key word chain method.	football game from a television program	recall rates were 74% and 81%, respectively.	concept of this method applies to most sports videos.
Mohamed Mansoor Roomi et al. [48]	Video classification and retrieval through spatio-temporal Radon features	UCF-101data set (for classification)	Classification rates of 96.15%	Using a single descriptor, the suggested descriptor can simultaneously capture the appearance and temporal motion features.
		HMDB51dataset(for classification)	Classification rates of 71.7%	
		10contexts dataset (for classification and retrieval)	Classification rates - 93.24%	
		TRECVID 2005 datasets(for classification and retrieval)	Classification rates - 97.3%	
		JHMDB dataset (for retrieval)	Good precision rate.	
Nitin Janwe et al. [49]	Semantic concept-based video retrieval using convolutional neural network	TRECVID2007 dataset (Dataset consists of 110 video clips) Training dataset 90 Testing dataset 20	Precision (MAP) 0.544	This research primarily contributes a framework for video retrieval based on idea co-occurrence information, as well as the proposed Ranked Intersection Filtering (RIF) method for effective video retrieval.
D. Asha et al. [50]	Content-based video Retrieval System using Multiple Features	40 online video dataset used for experiment.	80% recall rate and accuracy for online data retrieval	Superior in collecting temporal and spatial features for retrieval of videos.
Meinel et al. [51]	Content-Based Lecture Video Retrieval Using Speech and Video Text Information	The lecture video test data set is used	Accuracy is lower	Keywords are derived from a search based on content.

Li et al. [52]	Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis	To verify the system, five lecture videos from Yale University courses and a University of Washington machine learning seminar were used on YouTube.	Tracking the Projection Screen Region Precision, Recall, and F-score are 0.985, 0.977, and 0.981, respectively. The precision, recall, and F-score for extracting semantic keyframes are 0.989, 0.971, and 0.980, respectively.	Text content and keywords are used in the experiments.
Nguyen et al. [53]	Multi-modal and cross-modal for lecture video retrieval	They examined a total of 65 hours of movies from 47 French lecture recordings of their lab.	Accuracy is 67.38%	The focus of this work is multi-modal and cross-model video retrieval.
Gayathri et al. [54]	Improved Fuzzy-Based SVM Classification System Using Feature Extraction for Video Indexing and Retrieval	Any real-time Video(YouTube video)	About 95.4% accuracy, 100% precision, 100% F1 score, and 100% recall are achieved using this method.	Storage space is constrained, and the classifiers are unable to focus on video processing applications to specify signals.
Lin et al. [55]	A cloud-based face video retrieval system with deep learning	Caltech faces, Yale face B database, Facial images database, MIT-CBCL face recognition database, and YouTube faces database	Accuracy 99.96% and Computational time (millisecond per image) 4.6	With this approach, efficiency is not increased, and the system is not upgraded to obtain more datasets.
Li et al. [56]	Fast key- frame image retrieval of intelligent city security video based on deep feature	Five publicly available datasets: UQ_VIDEO, Holidays, Corel-5K, UKbench, and Landmarks in San	Average accuracy 0.95278	Since keyframes are not stored, this method does not preserve temporal and

	coding in a high concurrent network environment	Francisco (SFL)		spatial complexity.
Jacob et al. [57]	Video content analysis and retrieval system using video storytelling and indexing techniques.	Video story dataset	86% accuracy	Time complexity could arise with this approach.
Priya et al. [58]	Shot-based key frame extraction for ecological video indexing and retrieval	Publicly available ecological video datasets. (Test video set contains 4274 shot breaks)	94.2% F1-score for shot boundary identification	Treats video as static images and removes any motion or temporal information.
Markos Zampoglou et al. [59]	Integrating Motion and Colour for Content-Based Video Classification	Video on the Omega TV channel in Thessaloniki, Greece (the database was manually cut into 1,074 single-shot videos of varying content, including newscasts, sports, talk shows, and theatrical plays).	The classification results are as follows: 89.1% for football, 96.7% for interviews, 82.7% for speakers, and 99.36% for newscasts.	Avoided the temporal segmentation problem and concentrated on classification instead.
Wali et al. [60]	Multimodal Approach for Video Surveillance Indexing and Retrieval	Numerous video surveillance sequences from the TRECVid 2009 database, as well as other road traffic sequences, are used in the experiments.	The competitive minimum DCR (Detection Cost Rate) results for the following events are 0.961, 0.953, 0.952, 0.923, 0.956, and 0.322, respectively: embrace, people split up, object put, opposing flow, people run, and especially lift no entry.	The metadata of the video is used for indexing.
Dutta et al. [61]	Using a deep learning model, create a caption	For experiments, datasets such as Flickr8k, Flickr30k,	The model's accuracy rate was	The more epochs there are, the greater

	by extracting features from images and videos.	and MSCOCO are utilized.	approximately 80%.	the model loss.
Krishna Raj et al. [62]	An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in a cloud environment.	Holidays 1 M and Oxford 5 K datasets were used	mAP value of 0.83 during experimentation. It is seen that the solitary server takes up to 118 minutes to finish the process,whereas the spark cluster takes at least 19 minutes.	This approach does not minimize computation time and does not improve machine learning.
Lorenzo Baraldi and colleagues [63]	An Interactive Multimedia System for Video Reuse and Indexing: Neural Story	The video series datasets from BBC Planet Earth and BBC Romans, featuring Mary Beard, are used. The total number of video clips in this prototype database is 19, and they average fifty minutes in duration.	The evaluation metrics for every video in the database are the number of shots, scenes, and searchable unigrams.	The idea links automatic annotation techniques with temporal video segmentation along with a visualization and re-use interface,.
B.-C. Chen et al. [64]	Retrieving Scalable Face Tracks in Video Archives with Bag-of-Faces Sparse Representation	TRECVid News Dataset	MAP 86.4%	The creator of this technique utilizes temporal information to extract faces from consecutive video frames, thereby generating face tracks.
		NHKnews7 News Dataset	MAP 73.6%	
Mingtao Pei et al.[65]	Retrieving Face Videos Using Deep Learning of Binary Hash Representations	TV-Series dataset(the Big Bang Theory)	Precision 0.9412	This approach combines hash learning and feature extraction into a single network.
		TV-Series Dataset(Prison Break)	Precision 0.3261	
Zouhair Mbarki et al. [66]	Using the LBP technique with a convolutional	For the experiment, a video clip captured with a	97% of faces can be recognized, and	By increasing the number of epochs, the

	neural network, faces can be detected and identified in real-time from video sequences.	Huawei Y7 smartphone's HD (1920 x 1080) quality is used.	98% of faces can be identified.	CNN approach improves performance and stabilizes, ultimately achieving the target learning rate.

2.3. Key frame extraction from input video

Many videos are uploaded to the internet every minute, a result of the surge in multimedia content creation brought on by smartphones and increased internet access. Users can search for hundreds of recommended videos online when they wish to learn more about a recent event or topic. Users then find it much more difficult to find the appropriate video. Since watching each video takes much time, end users find it quite difficult. Therefore, it becomes necessary to extract the most important information from a video while removing any redundant or overlapping content. Video summarization, also known as keyframe extraction, is one such method. Around the world, researchers have developed several automated systems for extracting key frames from input videos. Table 2.2 provides a detailed summary of all these techniques.

Li et al. [67] proposed a new technique for key frame extraction from input video, which they named key frame extraction in the summary space, based on the idea that the video data are near a subspace of a high-dimensional space. The goal of the suggested method is to identify the video's representative frames and remove similar frames from the representative frame collection. The video data are first transferred to a high-dimensional space called the summary space. Then, by examining the summary space's inherent structure, a new representation is discovered for every frame. In particular, the learned representation is used to select representative frames and can reflect the representativeness of each frame. The similarity of example frames is then assessed using the perceptual hash technique. Therefore, after removing frames that are similar to the representative frame set, the key frame set is obtained.

Sahbi Bahroun et al. [68] proposed a key frame extraction method for video surveillance systems based on the quality of facial images. By discarding frames without faces, the amount of data is decreased. Images of faces are then grouped by identification. A selection of

candidate frames is then chosen to move forward. Four criteria are used to evaluate face quality: position estimation, sharpness, brightness, and resolution. The key frame is the frame with the best face quality.

Ujwalla Gawande et al. [69] presented a successful approach that is mostly based on the combination of histogram and deep features. By removing the ambiguity of keyframe selection, it retrieves the greatest number of pertinent keyframes. It reduces the computational and temporal complexity of the video sequence by being applied both concurrently and in parallel. In terms of extracting pertinent key frames from videos, the algorithm's performance demonstrates its effectiveness.

Chuhong Li et al. [70] conducted a quantitative analysis of the traffic localization characteristics in the backbone connection by proposing the traffic locality quantization index based on the traffic characteristics of the backbone link. The efficiency and accuracy of video retrieval are then improved by a deep learning-based key-frame abstraction and retrieval method. An adaptive key-frame selection algorithm is developed, key-frame features are extracted using the current convolutional neural network framework, and unsupervised, semi-supervised, and supervised retraining models are created to enhance the effectiveness of the convolutional neural network's feature extraction and improve the accuracy of video retrieval.

Chao et al. [71] proposed an "augmented 3-D keyframe," which is a more compact and meaningful representation of a keyframe. Representative objects, important materials (such as license plates and faces), motion data (including trajectories and simple movement situations), and certain indications of moving objects from a surveillance video clip captured by a stationary camera are added to it. To create this augmented 3D keyframe without reconstructing an entire 3D scene, they innovate by utilizing basic 3D geometric information.

Kumar et al. [72] suggested a method for video summarisation based on the Eratosthenes Sieve. The best group of clusters was found using the Davies-Bouldin Index (DBI). The technique can be transformed in three different ways. In the first transformation, the entire video is grouped into keyframe groups. In the second transformation, the video is divided into equal-sized frames and then clustered into keyframe groups. Using the Eratosthenes Sieve Theory, the video is divided into frames of varying sizes, utilizing prime and non-prime numbers in the third transformation. These frames are then grouped into keyframes.

Determining the optimal number of clusters for each variety is achieved using the Davies-Bouldin Index.

Sanjay K. Kuanar et al. [73] introduced an automated method for video keyframe extraction that utilizes dynamic Delaunay graph clustering with an iterative edge pruning procedure. Additionally, the video summary is enhanced by a structural restriction that takes the shape of a lower limit on the deviation ratio of the graph vertices. To capture more interesting frames, they also apply an information-theoretic pre-sampling technique, which utilizes notable dips in the mutual information profile of subsequent frames in a video. A variety of keyframe visualization approaches for videos are included to facilitate effective navigation and browsing.

Usman Saeed et al. [74] proposed a key frame selection technique based on lip motions to mitigate the fluctuations caused by visual speech. The suggested approach is divided into two modules. The first module proposes a key frame selection technique that analyzes lip motion in a single video and selects key frames based on a significance criterion (optical flow) after a set of videos showing a person repeating the same phrase in each video. After comparing the motion of these keyframes to that of the other videos, the next module chooses frames that have motions comparable to those of the keyframes. Later, the classical Eigenface technique will be used to compare these frames with random frames to assess the improvement in face recognition.

Dang et al. [75] presented a novel paradigm for key frame extraction using robust principal component analysis (RPCA). The idea behind the suggested framework stems from the observation that the RPCA breaks down an input data set into two parts: 1) a low-rank component that shows the systematic information across the data set's elements and 2) a collection of sparse components, each of which has unique information about every element in the same data set. A single 1-norm-based non-convex optimization problem combining the two types of information is used to extract the required number of keyframes. To address this optimization issue, they also create a brand-new iterative algorithm. The detection of shots, segmentation, and semantic comprehension of the underlying video are not necessary for the suggested RPCA-based architecture.

Mentzelopoulos et al. [76] introduced the Entropy-Difference algorithm, which is capable of segmenting spatial frames. The author of the proposed model assumes that if entropy is distributed throughout the image, the regions with the most important items in a video sequence will be described by the greater entropy distributions. Thus, any alteration to the object's appearance in one of these prominent areas will affect the story sequence's pertinent semantic information. This approach utilizes entropy as a local operator rather than a global characteristic for the entire image. Due to its dependence on the measurement context, entropy is a useful tool for expressing the impurity or unpredictability of a dataset.

Rasheed et al. [77] offered a technique for high-level video segmentation into scenes. A scene is a section of a play that exhibits continuous action in a single location or when the setting is fixed. The author of this study leverages this fact and transforms the assignment into a graph partitioning problem, thereby suggesting a novel method for grouping shots into scenes. To achieve this, a shot similarity graph (SSG), a weighted, undirected graph, is constructed. Each node in the SSG represents a shot, and the edges connecting the shots are weighted according to their similarity based on motion and color information. Then, using the normalized cuts for graph partitioning, the SSG is divided into subgraphs. The resulting partitions serve as representations of distinct video scenes. Instead of considering individual shot pairs, they cluster the images based on their global commonalities. Additionally, they suggest using a single, representative image from the video as a scene key-frame to explain the topic of each scene.

Wu et al. [78] presented a unique clustering-based approach for static video summarisation, framing it as a clustering problem. They integrated several key aspects of video summarization into their model to offer the VRHDPS method based on insights from the HDPS approach. First, they use pre-sampling to obtain candidate frames and lessen the video's redundancy. Next, they display the candidate frames' visual content using the BoW model. To create static video summarization results, a thoughtful technique is also applied, and candidate frames are grouped into clusters using the suggested VRHDPS clustering approach.

Mademlis et al. [79] deal with the static summarization of movies that feature actions with particular recurring characteristics. An activity video summary is defined in this context as the collection of keyframes that can both rebuild the original, full-length video and concurrently

highlight its most important elements. It is possible to optimize both goals simultaneously across various input modalities. To provide a rigorous definition of the video summarization problem, the two criteria are combined into a learning task called the "salient dictionary," which encompasses numerous existing techniques. This description gives rise to three distinct, innovative approaches to video summarisation: The Numerical, Greedy, and Genetic Algorithms. While the saliency term is modeled as an outlier detection issue, a low-rank approximation problem, or a summary dispersion maximization problem, the reconstruction term is algebraically modeled as a Column Subset Selection issue (CSSP) in all formulations.

Azra Nasreen et al. [80] employed mean squared error and k-means clustering to develop and apply a novel, robust key frame extraction and foreground isolation technique for films with varying frame rates. Along with removing the noise produced during recording, they also separated the foreground elements in the video. Using this technique, the flickering of frames in a recorded video, which results from a fluctuating frame rate, is significantly reduced. To speed up the computation's outcomes, k-means clustering is also carried out on Apache's Hadoop infrastructure. After using this technique, they were able to obtain findings that were sufficiently clear to extract important details from the frames. The approach's findings have been demonstrated to be superior to comparable results produced by well-known methods, such as the Gaussian Mixture Model.

K. Wu [81] conducted research on the subject of video structure analysis. It has features including video browsing, key frame extraction, and shot boundary identification. Using movies with both sudden cuts and smooth shot transitions, five distinct shot boundary recognition methods are implemented and compared. Pixel-to-pixel algorithms and histogram-based algorithms are the two primary categories into which these techniques fall. It is demonstrated that all five techniques function flawlessly in videos with only hard cuts. The algorithms' performance considerably declines in videos with slow motion or gradual transitions. A key frame is retrieved for each detected shot, allowing the video content to be easily browsed.

Qiang Zhang [82] proposed a novel key-frame extraction method for use with motion capture data. The foundation of this approach is an unsupervised cluster algorithm. The motion sequence is first divided into two groups based on the similarity of the distance between

neighboring frames. This enables the adaptive determination of the thresholds required in the subsequent step. The frames closest to the center of each class are automatically extracted as key frames of a sequence after all the frames have been clustered using a dynamic cluster algorithm called ISODATA. In contrast to many earlier clustering methods, the current enhanced cluster algorithm does not require user-specified parameters to handle various motion types automatically.

L. Asim et al. [83] introduced a technique for video summarization that utilizes a mixture of color features extracted from sections of a video frame rather than the entire frame to identify shot boundaries. Precisely identifying the video shots enhances the suggested method's resistance to various types of video transitions. To extract a keyframe from the most representative sub-shot of a single video shot, each shot is further subdivided into sub-shots based on the structural similarity between the frames. Lastly, the key frames extracted from each video shot's sub-shot are compared one at a time to eliminate any duplicated frames.

C. Huang et al. [84] have developed a new paradigm for effective video content and motion summarisation. Capsules Net is first trained as a spatiotemporal information extractor; then, using those spatiotemporal features, an inter-frame motion curve is produced. Then, to automatically divide the video streams into shots, a technique for detecting transition effects is suggested. Ultimately, key-frame sequences within the shots are selected using a self-attention model; as a result, optical flows can be computed for video motion summarization, and key static images are chosen for video content summarization.

V. Benni et al. [85] introduced a novel method for detecting shot transitions and selecting representative images using eigenvalues. Using this method, a data matrix was initially made for each frame that followed one another in the original video. The differences in intensity levels between subsequent photos were then ascertained by computing the covariance matrix. A modified method for computing the covariance matrix was employed to recalculate the entire matrix each time a new image was introduced to the data matrix, thereby reducing the computational load. Then, the Eigenvalues were found using the computed covariance matrix. Variations between the frames were calculated using the chosen minimum eigenvalue. A comparison was made between the lowest eigenvalue and a predetermined threshold. The

current image is selected as the representative frame if the Eigenvalue is greater than the threshold, and the prior image is regarded as a transition point.

Zheng R et al. [86] proposed a novel method utilizing GPUs (graphics processing units) to extract key frames from traffic surveillance videos, ensuring high accuracy and efficiency. Motion, particularly in surveillance recordings, is a more noticeable element when showing movements or events to identify significant frames. The GPU extracts the motion feature to reduce execution time. The final keyframes are chosen from the frames with local maxima of motion information after they have been smoothed to minimize noise.

Table 2.2. State-of-the-art techniques for video key frames extraction

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Li et al. [67]	Summary space, Perceptual hash algorithm, weight matrix learning	The Open Video dataset(contains 50 videos) The YouTube dataset(contains 50 videos)	Consistency, Accuracy Rate and Error Rate, Precision, recall, F-measure	Deep features are considered in this method
Sahbi Bahroun et al. [68]	Key frame selection based on face quality assessment	YTF dataset FRI CVL dataset	Accuracy rates based on Pose estimation, sharpness, brightness, and resolution	By employing only frontal faces instead of all the video frames, this technique has demonstrated its effectiveness.
Ujwalla Gawande et al. [69]	Deep features and histograms are the foundation of the method.	Publicly available database	Recall-0.95, Precision-0.92 CPU time (ms)-0.50	By selecting only a few key frames from a video sequence, a high level of precision can be achieved.
Chuhong Li et al. [70]	The BVLC Reference Cafe Net model is utilized as a deep neural network model, with model training	UK bench, Corel-5 K, Holidays, UQ_VIDEO, and San Francisco Landmarks Data sets	The average accuracy is 0.95278, The processing time of large-scale data sets is 0.061	The deep neural network is retrained using this key-frame description as the target for the interest layer.

	conducted using the softmax loss function.			
Chao et al. [71]	3D augmentation method	Five clips of surveillance footage are utilized as datasets.	Number of frames, number of objects, and time required to detect the frame.	Highly time complex
Kumar et al. [72]	AVS, EVS, and ESVS Method	Selected 50 videos from the Open Video Project (OVP) as the initial dataset.	Precision - 70.9% in AVS, 65.8% in EVS and 69.4% in ESVS Recall – 59.6% in AVS, 60.7% in EVS and 61.8% in ESVS F-measure- 64.8% in AVS, 63.1% in EVS and 65.4% in ESVS	Using the Davies-Bouldin Index (DBI) to determine the ideal collection of clusters is a challenging task.
		The second dataset comprises fifty videos, categorized into various sections, including "sports," "newsflash," "drawings," "advertising," "TV shows," and "home videos," among others.	Precision- 55.6% in AVS, 53.0% in EVS and 58.5% in ESVS Recall (%) - 49.4% in AVS, 49.7% in EVS and 50.0% in ESVS F-measure(%)- 52.3% in AVS, 51.3% in EVS, and 53.9% in ESVS	
Kuanar et.al. [73]	Dynamic Delaunay Clustering, PCA, Delaunay Grap	Open video project dataset, YouTube dataset	Fidelity, shot reconstruction degree, compression ratio, clarity, conciseness, and overall quality.	Effective edge trimming in a Delaunay graph is the foundation of the clustering approach.
Usman Saeed et al. [74]	Classical Eigen face algorithm	A subset of the Valid database, which has 106	Lip Feature, Number of KeyFrames, and	Based on lip features

		participants, was used for the tests.	Identification Rates	
Dang et al.[75]	Robust principal component analysis method	Consumer video dataset	Accuracy rate and Number of Key Frame	Assumptions do not always reflect better results
		Videos from the Open Video Project.	Accuracy rate and Number of Key Frame	
Mentzelopoulos et al.[76]	Entropy method	Three categories can be used to group the video clips they have chosen: 1. High-action clips (themes from Star Wars, Lord of the Rings, and James Bond), 2. Face-to-face dialogue clips (from Greek films and Star Wars), and 3. Simple motion clips (from Greek films, Star Wars, Lord of the Rings, and James Bond). The duration of the video clip varies between 30 and 4 minutes.	The total number of Keyframes, the total number of redundant frames, and the number of missing frame	External factors, such as lighting conditions, affect performance.
Rasheed et al.[77]	Histogram method	Five Hollywood movies: "Golden Eye (GE)," "Top Gun (TG)," "Terminator II (T-II)," "A Beautiful Mind (BM)," and "Gone in 60 Seconds (G-60)."	Scenes with ground truth, Scenes Detected, Correct detection, False Positive, False Negative, Recall and Precision	Cannot consider the local similarities
Wu et al.[78]	High-density peaks search clustering algorithm based clustering algorithm	VSUMM, Open Video project dataset	Precision- 0.68 Recall- 0.63 F-measure- 0.63	Cannot considered the dynamic video summarization
Mademlis et	Greedy and the	IMPART video	Execution time,	Cannot

al.[79]	Genetic Algorithm, Column Subset Selection Problem (CSSP), Local Outlier detection, Regularized SVD-based low-rank approximation, Maximum-dispersion global summary saliency	dataset, IXMAS dataset,i3DPOST dataset	IR (Independence Ratio) score	considered the dynamic video summarization
Azra Nasreen et al.[80]	K-means clustering and mean squared error method are used	The MPEG-4 format and an AVI format video with 3-channel images are considered part of the dataset.	Accuracy and Time	Using the information and characteristics of video sequences, key frame clusters are produced in this technique.
K. Wu [81]	Global pixel-to-pixel, cumulative pixel-to-pixel, Simple histogram, Maximum histogram, and Weighted histogram methods are used.	Chosen movie clips and security camera footage as the main data set.	Precision and Recall	Identifying slow shot transitions, user-defined thresholds, and frame-by-frame calculations are the main issues with this method.
Qiang Zhang [82]	Here, ISODATA—a modified method-based cluster algorithm—is suggested for key frame extraction from motion capture videos.	At a frame rate of 120 Hz, over 100 actual human motion sequences of various motion kinds were recorded from CMU.	Measurements of the mean absolute error value and reconstructed motion.	When used with highly regular movements, its benefits are not immediately apparent.
M. Asim et al. [83]	To identify sets of video frames with related	Open Video Dataset	Recall-0.73 Precision-0.62 F-Measure-0.67	An effective key frame extraction

	content, the method compares color attributes extracted from frames in batches.			method is represented to generate static video summaries based on color features.
C. Huang et al. [84]	Caps Net, Transition effects detection (TED), and self-attention model are used.	VSUMM Data Set	F-Score-0.90	Determining the number of clusters in a specific video file before performing the clustering operation is a challenging task.
		TvSum Data Set	F-score-0.87	
		SumMe Data Set	F Score-0.89	
		RAI Data Set	F-Score-0.91	
V. Benni et al. [85]	Eigenvalues are used in the suggested method to quantify the difference between successive frames.	TRECVID 2002 Video Dataset is used	Recall for Cuts-0.83 Precision for Cuts-1	Slowly changing frames, camera movements (such as zooming, tilting, or panning), and abrupt lighting (such as flashlights) in the video sequence continue to present difficulties for key frame extraction.
		House Tour	Recall for Cuts-1 Precision for Cuts-1	
		Hasselt	Recall for Cuts-1 Precision for Cuts-1	
		Mechelen Belgium	Recall for Cuts-0.97 Precision for Cuts-1	
Zheng R. et al. [86]	Based on GPU processing	Publicly available benchmark dataset(Highway)	Recall-0.61 Precision-0.8	Colors, edges, and events are not combined with motion information for keyframe selection.
		The Campus road surveillance video dataset(Campus-road)	Recall-0.65 Precision-0.72	

2.4. Existing Techniques for Face Detection from Key Frame

The introduction of low-cost video cameras and great processing power has made video-based facial recognition techniques easily superior to image-based methods. For this reason, video-based facial recognition has garnered considerable attention from researchers lately. Face detection offers a high level of accuracy in a regulated setting. Due to head motions, lighting conditions, facial expressions, the occlusion of other objects or clothing, resolution issues, and blurring caused by people's movements in front of the cameras—such as sunglasses and scarves—this function remains the most challenging in a crowded environment. On the one hand, face recognition works better in images than in a single shot. Second, processing this massive amount of data for every video is challenging due to the time required to handle each frame.

Additionally, faces in these images will be out of date or irrelevant because they are not accurate. It is necessary to select specific frames to detect the problem with a large number of data frames in a video. Don't concentrate on keyframe extraction by purchasing facial images. As a result, the ordering of pictures of faces only considers the correctness of the face in these frames, which leads to reduced storage but increased complexity in terms of time and space. It is evident from a careful review and analysis of this study that individual recognition is challenging when video indexing is performed using low-level attributes. Time and space are utilized in all these methods for indexing and storing data. When using a human face for video indexing, the image's posture, mood, illumination changes, occlusions, and facial expressions are all crucial components to consider. It has also been demonstrated that inherent ambiguities, such as inadequate resolution sensitivity, postural changes, and partial occlusion of the facial cavity, hinder video-based detection. Several automatic frameworks for face detection from key frames of the input video have been proposed by researchers worldwide. Table 2.3 contains a detailed summary of each of these approaches.

J. Li and colleagues [12] proposed a novel face recognition network that addresses three key aspects of face identification: enhanced feature learning, progressive loss design, and anchor-based data augmentation. They began by proposing a Feature Enhance Module (FEM) to improve the initial feature maps and convert single-shot detectors into dual-shot detectors. To properly enable the features, they also used Progressive Anchor Loss (PAL), which is

calculated by two distinct sets of anchors. To improve the regressor's initialization, they employed an Improved Anchor Matching (IAM) technique by incorporating a novel anchor assignment mechanism into the data augmentation process. They called the suggested network Dual Shot Face Detector (DSFD) because these methods are all associated with the two-stream design.

D. Bazarevsky et al. [13] presented Blaze Face, a lightweight and efficient face detector made for mobile GPU inference. It operates on flagship devices at 200–1000 FPS or more. Any augmented reality pipeline that requires a precise facial region of interest as input for task-specific models, such as 2D/3D facial keypoint or geometry estimation, facial feature or expression classification, and face region segmentation, can utilize it thanks to its super-real-time performance. A GPU-friendly anchor technique, adapted from the Single Shot MultiBox Detector (SSD), a lightweight feature extraction network modeled after but distinct from MobileNetV1/V2, and an enhanced tie resolution strategy that substitutes for non-maximum suppression are among their contributions.

Redmon J. et al. [14] made several modifications to YOLO. They improved it by making numerous minor design adjustments. They also trained this new, rather good network. The name of the network is YOLOv3. It's more accurate but a little larger than the last time. It's still quick, so don't worry. YOLOv3 achieves 28.2 mAP at 320 × 320 in 22 ms, which is three times faster than SSD while maintaining accuracy. They found that YOLOv3 is a rather good detection metric when using the outdated—5 IOU mAP metric.

K. Zhang et al. [16] presented Shuffle Net, a CNN architecture that is incredibly computationally efficient and specifically made for mobile devices with relatively little processing power (such as 10–150 MFLOPs). Two new operations—point-wise group convolution and channel shuffle—are employed in the new architecture to significantly reduce computing costs without compromising accuracy. The majority of the suggested network consists of three stages, each comprising a stack of Shuffle Net units. Stride = 2 is used to apply the initial building block in each step. The output channels for the subsequent stage are doubled, while other hyper-parameters within a stage remain unchanged.

Viola P. et al. [87] proposed a face detection framework that utilizes very rapid image processing and achieves high detection rates. Three significant contributions are included. First, a novel image representation known as the "Integral Image" was introduced, enabling the fast computation of features used by our detector. In the second approach, a straightforward and effective classifier is constructed by employing the AdaBoost learning algorithm (Freund and Schapire, 1995) to select a limited number of key visual characteristics from an extensive collection of possible features. The third contribution is a method for "cascade" classifiers, which enables the rapid removal of background areas in the image while allocating more processing power to face-like regions that show promise.

Qiang Zhu et al. [88] developed a quick and precise human detection system by combining the cascade-of-rejectors technique with Histograms of Oriented Gradients (HOG) characteristics. This method uses HoGs of variable-size chunks, which automatically identify key human characteristics. From a wide range of potential blocks, they choose the right set of blocks using AdaBoost feature selection. This system utilizes a rejection cascade and integral image representation, which significantly accelerates computation.

L. -T. Pham et al. [89] proposed an expansion of the integral image to quickly integrate into the interior of any polygon that isn't necessarily rectilinear. The method's integration time is quick, unaffected by image resolution, and only proportional to the number of vertices in the polygon. Applying the technique to the object detection framework developed by Viola and Jones, they suggest enhancing conventional Haar-like features with polygonal Haar-like features. To evaluate the surface terrain of Mars, experiments are conducted in three domains: rock detection, fixed-pose hand detection, and frontal face detection.

X. Zhu et al. [90] presented a unified model for face detection, posture estimation, and landmark estimation in real-world, crowded images. Every face landmark is modeled as a part of this model, which is based on mixtures of trees with a common pool of parts. Global mixtures are used to account for topological changes resulting from viewpoint. In contrast to dense graph topologies, they demonstrate that tree-structured models are simple to optimize and surprisingly good at representing global elastic deformation. Their technique appears to improve the state-of-the-art, sometimes significantly, for all three tasks, as evidenced by their extensive findings on standard face benchmarks and a new "in the wild" annotated dataset.

B. Yang et al. [91] extended the image channel to various types, such as gradient magnitude and oriented gradient histograms, thereby straightforwardly encoding rich information by applying the concept of channel characteristics to the face detection domain. A multi-scale version of features with improved performance is found when they fully explore feature design and implement a novel variant known as aggregate channel features. To address facial positions in the wild, they suggest a multi-view detection method that includes detection adjustment and score re-ranking.

M. Mathias et al. [92] presented two striking new findings in top performance for face detection. To start, they demonstrate that a well-trained vanilla DPM outperforms both commercial and research systems. Secondly, they demonstrate that a detector that is built on stiff templates—which are structurally similar to the Viola and Jones detector—can achieve comparable top performance on this task.

J. Yan et al. [93] were able to overcome the performance bottleneck of the deformable part model (DPM) on challenging datasets while maintaining detection accuracy. Cascade part pruning, HOG feature extraction, and 2D correlation between the root filter and feature map are the three prohibitive processes in the cascade version of DPM that are expedited. In the case of 2D correlation, since the root filter must be low-rank, a more effective linear combination of 1D correlations can be used to calculate the 2D correlation. To learn the low-rank filter discriminatively and gradually, a proximal gradient algorithm is employed. It is suggested that neighborhood-aware cascade be used for cascade component pruning to capture the reliance on aggressive pruning in neighborhood regions. Instead of explicitly calculating part scores, hypotheses can be refined using neighborhood scores with the first-order approximation. Look-up tables are designed to replace costly orientation partition and magnitude calculations with more straightforward matrix index operations for HOG feature extraction.

According to research by Y. Sun et al. [94], deep learning can effectively solve the face recognition problem when both face identification and verification signals are used as supervision. Using meticulously crafted deep convolutional networks, the Deep Identification-verification features (DeepID2) are learned. DeepID2, extracted from distinct identities, is used in the face identification task to increase inter-personal variations, while DeepID2,

extracted from the same identity, is used in the face verification task to decrease intra-personal variations. Both tasks are critical to face recognition. It is possible to generalize the DeepID2 features learned to new identities that were not present in the training data.

C. Chen et al. [95] introduced a brand-new, cutting-edge method for facial recognition. Since aligned face shapes offer superior characteristics for face classification, the main concept is to integrate face alignment with detection. By leveraging recent developments in face alignment, their method learns both tasks simultaneously within the same cascade framework, thereby enhancing the effectiveness of this combination. While maintaining real-time speed, such cooperative learning significantly improves cascade detection capabilities.

S. Yang and colleagues [96] presented a new deep convolutional network (DCN) that performs exceptionally well on FDDB, PASCAL Face, and AFW. Crucially, they employ a novel approach to face recognition by evaluating facial part responses based on their spatial arrangement and structure. The grading system was meticulously designed with difficult situations involving partially obscured faces in mind. This factor enables their network to identify faces in situations with extreme occlusion and unrestricted position variation, which are the primary challenges and bottlenecks of most face detection techniques currently in use.

H. Li et al. [97] suggested a cascade architecture based on convolutional neural networks (CNNs) that maintains excellent performance while having a very strong discriminative capability. In the fast, low-resolution stages, the suggested CNN cascade rapidly eliminates the background regions, and in the last, high-resolution stage, it carefully assesses a limited number of difficult candidates. A CNN-based calibration stage is added after each detection stage in the cascade to increase localization efficacy and lower the number of candidates at later stages. The output of each calibration stage is used to adjust the location of the detection window for the subsequent stage.

Z. Zhang et al. [98] explored the idea of enhancing detection robustness through multi-task learning rather than approaching the detection task as a single, independent problem. In particular, their goal is to maximize facial landmark identification in conjunction with diverse yet subtly connected tasks, such as facial attribute inference and head posture estimation. The fact that different activities have varying learning challenges and rates of convergence makes

this a non-trivial issue. With task-wise early stopping to promote learning convergence, they develop a unique tasks-constrained deep model to solve this issue. In comparison to the state-of-the-art approach based on a cascaded deep model, the proposed task-constrained learning significantly reduces model complexity and (i) outperforms current methods, particularly when handling faces with severe occlusion and position variation.

D. Zhang et al. [99] improved the detection performance by presenting a multi-task deep learning technique. In particular, they construct a deep convolutional neural network capable of simultaneously learning the face/non-face choice, face position estimation, and facial landmark localization problems. They demonstrated how a multi-task learning approach like this can further increase the classifier's accuracy.

K. Zhang et al. [15] proposed a deep cascaded multitask architecture that enhances performance by leveraging the natural connection between alignment and detection. Specifically, their methodology employs a cascaded architecture comprising three levels of meticulously crafted deep convolutional networks to provide coarse-to-fine predictions for face and landmark locations. They also propose a novel online hard sample mining technique that further enhances performance in real-world applications.

Sign Modou Bah et al. [100] introduced a novel approach that addresses some of the problems impeding face recognition accuracy by combining the Local Binary Pattern (LBP) algorithm with sophisticated image processing techniques, including Contrast Adjustment, Bilateral Filter, Histogram Equalization, and Image Blending. This enhances the LBP codes, which in turn improves the overall accuracy of the face recognition system.

Xu Tang et al. [101] proposed the Pyramid Box, a unique context-assisted single-shot face detector, to address the challenging face detection problem. They enhanced their use of contextual information in the following three areas after realizing its importance. They begin by creating a brand-new context anchor, which they refer to as a Pyramid Anchor, to oversee high-level contextual feature learning using a semi-supervised approach. Second, they suggest using the Low-level Feature Pyramid Network to integrate low-level facial features with sufficient high-level semantic context features. This enables the Pyramid Box to predict faces of all scales in a single shot. Thirdly, a context-sensitive structure is introduced to enhance the

prediction network's capacity and ultimately improve output accuracy. They also boost the diversity of training data for smaller faces by augmenting the training samples across various scales using the data-anchor-sampling technique.

Jialiang Zhang et al.'s [102] creative, simple, and effective "Feature Agglomeration Networks" (FANet) framework was utilized to develop a new one-stage face detector that not only delivers state-of-the-art performance but also operates efficiently. The main concept of their framework, which was inspired by Feature Pyramid Networks (FPN) [11], is to use the multi-scale features that are inherent in a single convolutional neural network by combining higher-level .Semantic feature maps of various scales as contextual cues to enhance lower-level feature maps in a hierarchical agglomeration fashion at a slight additional computational cost. To properly train the FANet model, they also suggested a Hierarchical Loss.

Wei J et al. [103]. presented a tiny object detection network based on YOLOv4. It enhances the detection of small objects in complex background situations, such as those found in drone aerial survey images. It overcomes factors such as occlusion by large objects, image noise, and a lack of effective features that hinder the performance of traditional methods in small object detection tasks within complex road environments. The cross-stage partial network (CSPNet) structure is incorporated into the spatial pyramid pooling (SPP) structure of the YOLOv4 network, followed by convolutional layers and a concatenation operation, which reduces the network's computational requirements and GPU memory usage. Second, by eliminating the large object detection head and replacing it with a more suitable tiny object detection head, the model's accuracy improves for small item identification tasks. After that, a new branch is added to the backbone section to extract feature information at a shallow location. The feature information from this branch is then fused in the neck section to enhance the data the model has extracted about the location of small objects. When combining feature information from various backbone levels, a weighting mechanism is introduced to increase the fusion weight of useful information, thereby enhancing detection performance at each scale. To improve feature representation capability and enable the model to focus on spatial location linkages and inter-channel relationships, a coordinated attention (CA) module is finally placed in the neck section at an appropriate location.

Qi, D. et al. [104] developed a face detector called YOLO5Face that is based on the YOLOv5 object detector. The head, neck, and backbone make up this structure. YOLOv5 utilizes a newly developed backbone known as CSPNet. The features in the neck are aggregated using a PAN and an SPP. Both regression and classification are employed in the head. The Wing loss function is utilized, along with a five-point landmark regression head. They created detectors with varying model sizes, ranging from a large model for optimal performance to a very small model for real-time detection on a mobile or embedded device.

Sirisha U et al. [105] are an excellent resource for anyone interested in object detection using YOLO, as they provide a thorough explanation of the YOLO architecture and its variations, along with their advantages and disadvantages. This research study thoroughly examines the most recent developments in object detection using YOLO and its variations. The development of the YOLO architecture and the advancements gained with each iteration are covered in the paper. Additionally, it covers several methods for enhancing the performance of YOLO and its variations, including attention mechanisms, feature pyramid networks, and multi-scale training. Furthermore, using a variety of benchmark datasets, the study assesses the performance of YOLO and its variations in comparison to other cutting-edge object detection methods. The study concludes that YOLO and its variations have achieved cutting-edge results in terms of accuracy, speed, and memory usage across various test datasets.

Table 2.3. State-of-the-art techniques for Face Detection from Key frames

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
J. Li and colleagues [12]	Feature Enhance The Module (FEM), Progressive Anchor Loss (PAL), and Improved Anchor Matching (IAM) methods are employed.	WIDER FACE Dataset (Contains 393 703 annotated faces with large variations in scale, pose, and occlusion in total 32 203 images. For each of the 60 event classes, 40%, 10%, and 50% images of the database are randomly selected	On the validation set, the average precision (AP) was 96.6% (Easy), 95.7% (Medium), and 90.4% (Hard); on the test set, it was 96.0% (Easy), 95.3% (Medium), and 90.0% (Hard).	The model is, in fact, quite large and intricate.

		as training, validation, and testing sets.)		
		FDDB Dataset (Consists of 2 845 images featuring 5 171 faces selected from the wild data collection.)	When the number of false positives equals 1,000, DSFD obtained state-of-the-art performance on both discontinuous and continuous ROC curves, i.e., 99.1% and 86.2%.	
V. Bazarevsky et al. [13]	Used Blaze Face	Dataset of 66K images	Average precision (AP)-98.61% The mobile GPU inference time(ms)-0.6(ms)	Major AR self-expression apps and AR developer APIs for smartphones are powered by the technology outlined in this paper.
Redmon, J. et al. [14]	Yolov3	COCO dataset	Mean Average Precision -50-57.9 Inference Time-51(ms)	Problem with Anchor box x, y offset predictions, Linear x, y predictions instead of logistic and Focal loss.
X. Zhang et al. [16]	Shuffle Net	ImageNet 2012 classification dataset	Map [.5, .95] (300× image)-18.7% Map [.5, .95] (600× image)-25%	Speed up Better than Alex's net. Good for Mobile devices
Viola, P. et al. [87]	An integral image, the AdaBoost learning algorithm, and combining classifiers in a	MIT + CMU frontal face Data set(MIT + CMU test set containing 130 images and 507 faces.)	Detection rates-94.1%	Used for frontal face detection

	"cascade" are used.			
Qiang Zhu et al. [88]	Combine the Histograms of Oriented Gradients (HOG) features with the cascade-of-rejectors technique to create a fast and accurate human detection system.	MIT pedestrian database(Containing over 1800 annotated human images with a large range of pose variations and backgrounds.)	Detection rate-88%	Used Variable size blocks
M.-T. Pham et al. [89]	Used the technique to enhance conventional Haar-like features with polygonal Haar-like features in Viola and Jones' object identification framework.	Used the MIT+CMU standard test set(Consisting of 130 grayscale images with 507 frontal faces)	Average detection speeds for frontal faces(in fps)-10.5	Used for frontal face detection, fixed-pose hand detection, and rock detection for Mars' surface terrain assessment
X. Zhu et al. [90]	This model is built on a variety of trees that share a common set of components.	CMU MultiPIE face dataset(Contains around 750,000 images of 337 people under multiple viewpoints, different expressions, and illumination conditions.)	Pose estimation error (in degrees)-99.9% Average localization error as a fraction of face size-99.8% Average localization error as a fraction of face size(for Frontal face)-100%	The multiview model can run as fast as a single-view model
		Annotated face-in-the-wild (AFW) Data set(This produced a 205-image dataset with 468 faces.)	Average Precision for all faces-88.7% Average Precision for large faces-92.9%	

			Pose estimation error (in degrees)-81% Average localization error as a fraction of face size-76.7%	
B. Yang et al. [91]	For multi-view face detection, aggregate channel features were used.	AFW Databases	Average precision - 96.8%	In real-world applications where there is greater visual fluctuation in human faces, this detector may perform significantly worse.
		FDDB databases	Average precision- 83.7%	
M. Mathias et al. [92]	Used deformable part models (DPM)	AFLW Datasets(Contains 26000 annotated faces)	Average Precision -95%	High computational cost and frequently need pricey annotation during training.
		Pascal Faces Datasets (851 Pascal VOC images with bounding boxes)	Average Precision - 90.29%	
		AFW Datasets(205 images with bounding boxes)	Average Precision - 97.21%	
		FDDB Datasets(2845 images with ellipses annotations)	Average Precision - 0.864	
J. Yan et al. [93]	Used deformable part models (DPM) for object detection	AFW Faces Database	Average Precision - 93.7%	High computational cost and frequently need pricey annotation during training.
Y. Sun et al. [94]	Deep convolutional networks, which have been meticulously developed, are used to learn the Deep	LFW dataset (13,233 images of 5749 identities were gathered from the Internet and included in it)	The accuracy of face verification has been reached at 99.15%, and the error rate has been drastically decreased by	Taken more processing time

	IDentification-verification features (DeepID2).		67%.	
D. Chen et al. [95]	Used random forests to perform alignment and detection simultaneously using the properties of pixel value difference.	Publicly Available FDDB Face Datasets	Recall- 80.07% Accuracy	A subpar face detector limits the accuracy by creating initial detection windows.
		Publicly Available AFW Face Datasets	Recall Accuracy	
		Publicly Available CMU-MIT Face Datasets	Recall Accuracy	
S. Yang and colleagues [96]	Used Faceness-Net	FDDB Data set	Recall- 90.99%	Due to the complex CNN structure, this approach requires a significant amount of practice time.
		PASCAL faces Data set	Average precision- 92.11%	
		AFW Data Sets	Average precision- 97.20%	
H. Li et al. [97]	Used convolutional neural networks(CNN)	FDDB Data set(Contains 5, 171 annotated faces in 2, 845 images)	Recall- 95.9%	By requiring bounding box calibration from face detection at an extra computational expense, this approach disregards the inherent connection between bounding box regression and facial landmark localization.
		AFW data Set(Contains 205 images dataset)	Average precision- 87.48%	
Z. Zhang et al. [98]	Used Deep Multi-task learning	AFLW Data Set	Mean error and Failure rate	The first detection windows that a subpar face detector creates limit the accuracy.
C. Zhang et al. [99]	Used a post filter for a	Publicly available FDDB data set(Data	Detection Rate -77.32%	This technique does not

	boosting-based multiview face detector using multi-task DCNN.	set contains 5171 faces in 2845 images)		account for facial characteristics such as gender, age, lighting, or facial expression.
K. Zhang et al. [15]	Used MTCNN algorithm for face detection and alignment	FDDB Dataset(Contains the annotations for 5171 faces in a set of 2845 images)	True Positive Rate-0.9504	In terms of mouth corner localization, this approach is less effective. It may be caused by slight variations in expression in their training data, which have a significant impact on the location of the mouth corner.
		WIDER FACE dataset(Consists of 393 703 labeled face bounding boxes in 32 203 images, where 50% of them are for testing (divided into three subsets according to the difficulty of images), 40% for training, and the remaining for validation)	Recall for Easy set-0.851 Recall for Medium set-0.820 Recall for Hard set-0.607	
		AFLW Data set(Contains the facial landmarks annotations for 24 386 faces)	Mean Error rate-6.9%	
Sign Modou Bah et al. [100]	Utilized the Local Binary Pattern (LBP) algorithm in conjunction with sophisticated image processing methods like picture blending, contrast adjustment, bilateral filtering, and histogram equalization.	Three distinct datasets were created, each comprising 181 x 181 pixel faces in various orientations and conditions: Dataset [I], Dataset [II], and Dataset [III].	Face recognition accuracy-99%	This study doesn't address the issue of mask faces and occlusion in facial recognition.

Xu Tang et al. [101]	Used Pyramid Box	FDDB Data Set (Contains 5171 faces in 2845 images)	Pyramid Box obtained state-of-the-art performance on both discontinuous and continuous ROC curves, i.e., 99.0% and 86.0%.	Useful for Hard-face
		WIDER FACE Data Set (Contains 32,203 images and 393703 annotated faces. h a high degree of variability in scale, pose, and occlusion. The database is split into training (40%), validation (10%), and testing (50%) sets, where both validation and test sets are divided into "easy," "medium" an)	mAP values for the validation set were 0.961 (easy), 0.950 (medium), and 0.889 (hard), whereas the testing set had mAP values of 0.956 (easy), 0.946 (medium), and 0.887 (hard).	
Jialiang Zhang et al. [102]	Used FANet	WIDER FACE Datasets (Contains 32,203 images with about 400k faces for a large range of scales. There are three subgroups in all: 50% for testing, 10% for validation, and 40% for training.)	For the Validation set, the mAP values were 89.5% (hard), 94.7 (middle), and 95.6% (easy). In contrast, the mAP values for the testing set were 94.9% (easy), 93.9% (medium), and 88.7% (hard).	Manages scale variance in face detection.
		FDDB Face Data Set(Contains 5,171 faces in 2,845 images)	With 99.0% and 86.0% performance on both discontinuous and continuous ROC curves, respectively,	

			FANet achieved state-of-the-art results.	
		PASCAL FACE dataset(Contains 1,335 labelled faces in 851 images)	mAP 98.78%	
Wei J et al.[103]	YOLOv4's neck's CSP structure and SiLU activation function were used	VisDrone2019 dataset	mAP-52.76%	YOLOv4 has a large number of parameters and is not suitable for mobile devices.
		S2TLD dataset	mAP-96.98%	
Qi, D. et al. [104]	The YOLO5Face method is used	Wider Face dataset (32,203 images and 393,703 faces are included)	For large models with more than 3 million parameters and over 5G flops, the corresponding mAPs are 86.55% for the Hard subsets, 95.08% for the Medium subsets, and 96.67% for the Easy subsets.	The model is very small and used in mobile devices.
		FDDB dataset (contains 2845 images with annotations for 5171 faces.)	mAP -0.9880	
Sirisha, U et al. [105]	YOLOv1 to YOLOv8 Methods are used to detect objects.	MSCOCO dataset, Voc 2007 Dataset, Voc 2012 Data set, COCO Dataset, Open Images Data set, KITTI Data Set, and Visual Genome Data set are Used	Average precision, Intersection over Union (IoU), Mean Average Precision (mAP), False-Positive Rate (FPR), and Recall	The drawbacks of YOLO and its variations are also highlighted in the research, including its sensitivity to object aspect ratios and its incapacity to identify small objects (in cases of

				variability in object appearance, occlusion, scale variation, illumination changes, limited training data, and computational complexity).

2.5. Existing Techniques for face recognition from detecting faces

Over the past decade, numerous efficient face recognition techniques have been developed, and the accuracy of face recognition has also steadily improved. It is challenging to develop and train an end-to-end face recognition model, as face recognition systems typically comprise distinct components, including face detection, face alignment, and facial feature extraction. Numerous approaches based on image processing, machine learning, and deep learning techniques have been proposed by researchers worldwide for recognizing faces. Table 2.4 has a detailed summary of each of these approaches.

Lenc L. et al. [106] focused on the well-known Local Binary Patterns (LBP), which are widely used in this industry and have a high degree of recognition. They do this by using a set of Gabor filters. After identifying local extrema in the filter responses, these are utilized as feature points. A clustering method significantly reduces the number of points. This study investigated the importance of fiducial point locations in automatic facial recognition. Two novel approaches have been outlined and contrasted with one that was previously suggested. The performance of our matching algorithm is demonstrated using the first method, Grid Position, which utilizes points in a standard grid. While the second approach, Face Specific Position, identifies the feature locations individually for each face, the third method, Global Position, derives the feature positions based on a representative subset of the gallery. The last two methods use Gabor wavelets to identify critical locations, which are subsequently clustered using the k-means algorithm.

Fenggao Tang et al. [107] suggest a spatial transformation layer-based end-to-end face recognition technique. In particular, the face recognition network's spatial transformation layer is positioned in front of the feature extraction layer, and alignment learning—which does not require pre-existing knowledge or artificially determined geometric transformations—is used to align the face region. By utilizing knowledge about the face identity category, the convolutional neural network can automatically select the optimal face alignment.

Masi et al. [108] focused on the issue of excessive pose changes and presented a technique to advance unconstrained face recognition in the wild. Their approach explicitly addresses pose variation by utilizing multiple pose-specific models and rendered face images, in contrast to existing methods that either normalize images to a single frontal pose or expect a single model to learn pose invariance through massive amounts of training data. To create discriminative representations known as Pose-Aware Models (PAMs), they use deep Convolution Neural Networks (CNNs).

S. Liao et al. [109] suggested a universal partial face recognition method that does not necessitate face alignment using any other fiducial points or eye coordinates. They develop a face representation technique that does not require alignment, based on Multi-Key Point Descriptors (MKD), in which the image's actual content determines the size of the face's descriptor. This enables the sparse representation of any probing face image, whether partial or holistic, using a vast lexicon of gallery descriptions. The Gabor Ternary Pattern (GTP), a novel keypoint descriptor, has also been developed for reliable and discriminative face recognition.

Chen Y.C. et al. [110] presented a brand-new multivariate sparse representation technique for recognizing faces in video-to-video. Both correlations and coupling information between the video frames are simultaneously taken into consideration by their method. A sparse linear combination of training data is used in their process to represent all of the video data jointly. Additionally, they make adjustments to their model to make it resilient to occlusion and noise. To address the non-linearities present in video data, they also kernelize the approach.

Shaohua Zhou et al. [111] proposed a tracking-and-recognition technique that simultaneously addresses tracking and recognition uncertainty within a single probabilistic framework. To describe the changing kinematics and identity in the probe video, a time-series state-space model is used to fuse temporal information. The model comprises three fundamental components: a formula for motion that governs the kinematics of the tracking motion vector, an identity equation that regulates the temporal evolution of the identity variable, and a relationship between the identity variable and the motion vector, as established by an observation equation.

P. Forczmanski et al. [112] addressed the challenge of facial recognition for images with lighting issues, including flashes, shadows, and extremely low brightness levels. To address the aforementioned issues, the presented technique utilizes 2DDCT (two-dimensional discrete cosine transform), which is enhanced by reducing brightness gradients, minimizing spatial low-frequency spectral components, and fusing spectral features based on average intensities. Roy H. et al. [113] presented a new technique for illumination-invariant and heterogeneous facial recognition (HFR) termed Local Gravity-facial (LG-face). The Local Gravitational Force Angle is a concept that LG-face uses (LGFA). The direction in which the center pixel pulls the other pixels in its local neighborhood is known as the Local Group Field Alignment (LGFA). The LGFA is an illumination-invariant feature, according to theoretical research, which takes into account the reflectance component of the local texture impact of the nearby pixels. Additionally, it maintains edge information.

Table 2.4. State-of-the-art techniques for Face Recognition

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Lenc L., et al. [106]	Local Binary Patterns, Gabor wavelets, and k-means algorithm are used.	AT&T database of faces	The recognition rate for the grid (to show the matching importance) is - 65.67%	Current LBP-based techniques have the disadvantage of having set fiducial point locations and numbers. This technique eliminates this
		FERET dataset	The recognition rate for the grid (to show the matching importance) is -	

			98.49%	disadvantage.
		AR face database	The recognition rate for the grid (to show the matching importance) is - 97.29%	
		Czech News Agency (CTK) database	The recognition rate for the grid (to show the matching importance) is - 51.72%	
Fenggao Tang et al. [107]	Used Convolution Neural Network,	CASIA-Web Face dataset(contains 10575 people with total 494,414 face images)	Loss(MSE)-0.004635(Network Structure 3 Conv. 2 FC)	These algorithms often have to extract facial features after completing a face alignment process based on prior knowledge of facial structure.
		LFW face dataset(contains 5749 people including total 13233 face images)	Accuracy-99.02%	
		YTF dataset(contains 3425 video sequences with 1595 different identities)	Accuracy-94.6%	
I. Masi et al. [108]	Used deep Convolution Neural Networks (CNNs).	IARPA JANUS Benchmark-A (IJB-A) dataset	Recognition Rate at Rank-10 is reported for the identification	Learn to deal with pose variations. For frontal, half-profile, and full-profile positions, Pose-Aware Models (PAMs)
		People In Photo Albums (PIPA) dataset		
		CASIA Web Face Dataset		
S. Liao et al. [109]	Multi-Key Point Descriptors(MKD) and Gabor Ternary Pattern (GTP) are used	FRGCv2.0 dataset	Detection and identification rate	Recognized faces that were partial or holistic without needing to align.
		AR dataset	Detection and identification rate	
		LFW dataset	Genuine Accept rate	
		Pub Fig dataset	Genuine Accept rate	

Chen Y.C. et al. [110]	Used multivariate sparse representation method	UMD dataset	Average Identification rates-95.37%	This research focused on the urgent need to save time and space, as well as significant intra-class face variability.
		Multiple Biometric Grand Challenge (MBGC) dataset	Average Identification rates-88.04%	
		Honda/UCSD dataset	Average Identification rates-96.58%	
Shaohua Zhou et al. [111]	The video is captured using a time-series state-space model, and the numerical solutions to the model are obtained using SIS techniques.	Database-0(No. of subjects -12)	Recognition rate within top 1 match-100%	Robustness, Resampling, and computational load are the issues with this method.
		Database-1(No. of subjects -30)	Tracking accuracy(case-4)-100% Recognition w/in top 1 match(case-4)-93% Recognition w/in top 3 match(case-4)-100%	
		Database-2(No. of subjects -25)	Recognition rate-56%(When top 1 match is considered) Recognition rate-88%(When top 3 match is considered)	
P. Forczmanski et al. [112]	Used 2DDCT (two-dimensional discrete cosine transform)	Yale B dataset	Recognition rate-94.4%	Address the issue of facial recognition in images with lighting issues, such as flashes, shadows, and extremely low brightness levels.
		Yale B+ dataset	Recognition rate-98.7%	
Roy, H. et al. [113]	Used Local Gravitational Force Angle (LGFA).	CMU-PIE database	Rank 1 recognition rates of 97.78%	Evaluate the effectiveness of the proposed approach in various face recognition
		Extended Yale B	Rank 1 recognition rates of 97.31%	
		CUFS database	Rank 1	

			recognition rates of 99.96%	scenarios, including those with different lighting conditions, postures, noise levels, and changes in modality.
		CUFSF database	Rank 1 recognition rates of 98.67%	
		CASIA-HFB database	Rank 1 recognition rates of 99.78%	

2.6. Image gradient calculation from detecting faces

An essential component of a facial image is the gradient. After face detection and recognition from keyframes, the next crucial step is to calculate the image gradient based on the detected face. The QR code and linear bar code for video indexing using a human face as a cue are generated from this facial image gradient. Several methods and models based on the utilization of gradient field characteristics have been proposed to study the structure and attributes of facial images. Table 2.5 provides a detailed description of each technique, accompanied by rich imagery.

Kukharev, G. et al. [114] tackled the issue of facial image linear bar code production. The two approaches offered are based on intensity gradients computed across images, utilizing their original characteristics and intensity histograms. Following a certain number of intervals, these attributes are averaged and then quantized in the decimal digit range of 0 to 9, transformed into a standard barcode. The blocks of the barcode generation system are described, and their structure is suggested.

Carcagnì P et al. [115] presented a thorough investigation into the use of the histogram of oriented gradients (HOG) descriptor in the face recognition (FER) problem, emphasizing how this potent method might be successfully applied for this objective. This research emphasizes, in particular, that this descriptor can be one of the best for characterizing differences in face expressions if the HOG parameters are adjusted correctly.

Table 2.5. State-of-the-art techniques for image gradient calculation

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Kukharev, G. et al. [114]	Used intensity gradients and intensity histogram	Face 94 Dataset and Face Sketch FERET Database	The stability of the generated barcodes.	Ensure the stability of the barcodes generated when the source image is mirrored, when there are changes in scale, position, and facial expression, and when local lighting casts shadows on the faces.
Carcagnì P et al. [115]	Used histogram of oriented gradients	Chon-Kanade (CK+6) data set(a subset of 347 images was obtained with the following distribution among the considered classes of expressions: anger, disgust, fear, happiness, sadness, and surprise	Recall-95.9 %, Precision-95.8 %, Accuracy-98.9 % and F-score-95.8 %.	The detector can accurately register the face at angles between -30 and 30 degrees, and the classifier is sufficiently robust to ensure classification performance comparable to that of the frontal face case.
		Radboud Faces Database (contains images of 67 subjects performing eight facial expressions (anger, disgust, fear, happiness, contemptuous, sadness,	Recall-92.9 %, Precision-93 %, Accuracy-98.2% and F-score-92.9 %.	

		surprise, and neutral) with three gaze directions and five different face orientations)		

2.7. Linear facial Bar code generation from human faces

The term "facial barcode" refers to the concept of encoding facial features into a barcode for identification or verification purposes; this technique is frequently employed in security or access control systems. This technology produces a strong identification system by combining facial recognition with barcode scanning. Researchers worldwide have developed various approaches to representing facial barcodes. Table 2.6 provides a detailed summary of all these techniques.

According to empirical research by Dakin, S. C. et al. [116], processes tuned to horizontal visual organization play a major role in communicating facial identification information. In particular, when compared to other orientation bands, observers significantly outperform them in recognizing faces that have been filtered to include only horizontal details. Then, using computational methods, they demonstrate that, in contrast to images of real scenes, horizontal structures within faces have a peculiar tendency to cluster vertically. They suggest that these clusters have significant computational features and refer to them as "bar codes." They suggest that faces are "special" visual stimuli because of this characteristic, which allows them to convey information as a dependable spatial sequence—a highly limited one-dimensional code. They demonstrate that while this structure offers computational benefits for face detection and decoding, such as resilience to typical environmental image deterioration, it also leaves faces vulnerable to specific transformation classes that alter the bar sequence, including spatial inversion or contrast-polarity reversal.

Matveev Y. et al. [117] suggested a straightforward technique for creating linear barcodes of standard type from images of faces. Different image brightness gradients are used in the process. It entails quantizing the results into the decimal range of 0 to 9, averaging the gradients into a finite number of intervals, and converting the table into the final barcode.

S.Ghatak [118] proposed a novel method for producing high-quality linear barcodes from facial images. The process computed the difference in the image shine gradients; then, it needed to normalize the gradients into a finite number of intervals, quantize the results into decimal digit limits between 0 and 9, and then translate the table into a final barcode. The stability of representing its features is ensured by a theoretical analysis that demonstrates that the upper portion of the physiognomy is not affected by a remark (such as a change in physiognomy utterance, change in face range size, or change in eyes range (gaping and occlude eyes) after mirror rotation of the input image changes with human age). The top 70 and 75% of the facial image is thus encoded with a typical type linear EAN-8 barcode using this procedure. Nonetheless, a suggested technique is constructed in this study based on the notion that a person's nose, lips, eyes, and other facial features are constantly at constant distances from one another. Human faces may, therefore, be recognized and matched by comparing their feature edges, which has been accomplished by utilizing the window technique to determine the image gradient. For each face's histogram to be represented, the calculated gradient data is further condensed into a smaller data set. The barcode is created for every facial image based on this histogram.

Ghatak S. et al. [119] reported a novel work for standard-type rectilinear bar codes from illumination invariant face images. To determine the qualification in slopes of image sparkle, this approach utilizes the LGFA and Windowing methodology. Next, it uses standardization to determine the normal of the angles into a small number of intervals. This is followed by the conversion of the chart into an extremely standardized format, and the quantization effect is represented by the breaking points of decimal digits, ranging from 0 to 9. The current approach undoubtedly produces the best-quality face pictures, directly compatible with the EAN-8 standard bar code. Using the windowing technique, LGFA scans the input image horizontally and extracts the top 75% and 70% of the gradients, assisting us in creating gradient information from illumination-invariant face images.

Table 2.6. State-of-the-art techniques for linear facial Bar code generation from human faces

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Dakin, S. C. et	Filtering and	Used images of	Correctly	No proper

al. [116]	Normalization technique is used	celebrities collected from the Internet as a dataset	identified faces(Horizontal)- 56% Correctly identified faces(Vertical)- 35%	dataset was used; the bar code sequence was altered by inverting the horizontal components of the image but not by reversing the vertical ones.
Matveev, Y. et al. [117]	Various image brightness gradients are employed.	The Faces94 dataset and a database of composite faces at different ages are used as the dataset.	Stable EAN-8 barcode	It ensures that the generated barcodes remain stable in the face of changes in scale, position, and mirroring of facial images, as well as variations in facial expressions and local lighting shadows on faces.
S. Ghatak [118]	Used distinction in gradients of image shine using the Window technique, normalization, quantization, and EAN-8 linear bar code.	Face94 database(contains 100 classes with 11 images)	Accuracy- 78% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 80% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	Stable barcodes are hard to obtain for illumination-invariant faces.
		FERET database Face(contains 14051 images	Accuracy- 65% (75 percent of the upper portion of	

		with 10465 different subject)	the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 70% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		YaleB face database(contains 585 images with 10 different subject)	Accuracy- 10% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 15% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		FG-NET Aging Database(contains 1002 face images of 82 subjects)	Accuracy- 75% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 80% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the	

			barcode's stability)	
		Composite face dataset(contains 20 Composite faces)	Accuracy- 80% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 83% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
Ghatak, S. et al. [119]	Used LGFA and Window Technique	FACE94 dataset(contains 100 classes with 11 images)	Accuracy- 80% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 85% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	It is possible to create a stable barcode from an illumination-invariant face image.
		FERET database(contains 14051 images with 10465 different subject)	Accuracy- 75% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 80% (70 percent of the upper portion of the face image,	

			except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		Face dataset YaleB(contains 585 images with 10 different subject)	Accuracy- 70% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 75% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		NIR-VIS cropped dataset (includes images of identical subjects taken using both visible light (VIS) and near-infrared (NIR) techniques)	Accuracy- 70% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 72% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		IFW dataset(contains 13,233 images of 5,749 people)	Accuracy- 70% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	

			Accuracy- 72% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		Aging dataset FG-NET(contains 1002 face images of 82 subjects)	Accuracy- 80% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 85% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	
		Aging dataset of composite faces(contains 20 Composite faces)	Accuracy- 81% (75 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability) Accuracy- 85% (70 percent of the upper portion of the face image, except the area beneath the middle of the mouth or nose, to verify the barcode's stability)	

2.8. QR code generation from human faces

According to S. Tiwari [120], a QR code, also known as a "Quick Response" code, is a 2D matrix code that was created with two requirements in mind: it must be able to hold a lot more data than 1D barcodes and decode quickly using any handheld device, such as phones. High data storage capacity, quick scanning, omnidirectional reading, and numerous other benefits, such as error correction (which enables the proper reading of broken codes) and multiple

version support, are all provided by QR codes. QR codes are being utilized in various fields today, including marketing, security, academia, and more, and their popularity is growing rapidly. This technology, which combines QR code scanning with facial recognition, creates a robust identification system. Researchers from around the world have developed various methods for displaying facial QR codes. Table 2.7 provides a comprehensive overview of each of these methods.

Linglong Tan et al. [121] employed the PCA technique to analyze the dimensionality reduction process in face image analysis for their discussion of two-dimensional face code identification technology. The creation of two-dimensional face codes is made possible by examining QR encoding and decoding techniques. PCA can be utilized in the experiment to encode and decode the two-dimensional code, thereby reducing dimensionality to retrieve the face's effective information.

Table 2.7. State-of-the-art techniques for QR code generation from human faces

Author	**Method**	**Datasets**(No of samples)	**Evaluation Metrics**	**Remarks**
Ling long Tan et al. [121]	Used PCA algorithm	ORL face database,	Efficiency of face two-dimensional code recognition operation.	Used a straightforward classification approach to identify the two-dimensional code's face.

2.9. Discussion and future direction

We'll delve into great detail regarding the current research gap in state-of-the-art methods for video indexing and retrieval, focusing on human faces as cues, in this section. Low-level attributes, such as texture and color, are commonly used by the majority of image indexing techniques, as previously discussed in this chapter. However, it is impossible to index persons based on low-level image and video indexing techniques, and it is challenging to identify those who use them. It is a difficult task to analyze a video and identify each component that is present.

The video's content could include text, audio, and images. The human face is among the most significant elements of an image that appears in a video. The volume of videos is growing daily, making it difficult to store the data on our storage devices. More time is needed to deliver the video over a communication channel, and more space is required to store the information contained in the video. To reduce the complexity of time and space, video summarization and indexing are necessary. Numerous automated frameworks for facial recognition-based video indexing have been presented by researchers worldwide.

Nevertheless, it is concluded from all these frameworks that person detection is challenging when video indexing is performed using low-level features. From a storage and indexing perspective, all these approaches require a significant amount of time and space. An essential component of video indexing through the human face is the facial expression, position, emotion, lighting change, and occlusions of the face image. Additionally, the inherent ambiguities of video-based recognition were noted to be impacted, including position fluctuations, sensitivity to low resolution, and partial occlusion of facial cavities. In these challenging scenarios, the current state-of-the-art techniques have failed to yield desirable results. Therefore, a novel framework was proposed, which is covered in detail in chapter 3 of this dissertation, to address the difficulties of video indexing through human face images, including the detection of faces from videos with varying facial expressions, poses, emotions, illumination changes, and occlusions of the video's face image, as well as the time and space complexity of video indexing through human faces. Using face images with different facial expressions, positions, moods, illumination changes, and occlusions, the framework

effectively indexes the video. It also takes into account the time and space complexity of indexing human faces from input videos using linear barcodes.

Several new frameworks have been proposed by researchers worldwide for video indexing and retrieval, utilizing a person's face as a cue. Nevertheless, these frameworks have the following two shortcomings. Firstly, all methods are ineffective in identifying faces in videos due to factors such as the face's direction, changes in brightness, and variations in illumination. Consequently, the accuracy of video indexing using human faces will be reliant on the accuracy of face detection from videos in various lighting conditions, including variations in the direction of the human face, brightness, illumination, etc. The overall indexing and retrieval methods will suffer if the keyframes (human faces) of the video are not correctly detected.

Second, the faces detected as keyframes in the input video are not compactly represented. A linear bar code that loses much information when scanned horizontally is a compact representation. Additionally, the person's barcode cannot be recognized if any part of the linear barcode is damaged after scanning. The inability to scan the input image vertically is another issue with linear bar codes. The state-of-the-art methods have yet to thoroughly investigate the compact representation of face images.

To address the above two drawbacks, we have proposed an efficient framework for video indexing through face Images using QR codes in Chapter 4. Instead of using a face image, this framework utilizes a QR code that can be found in various video formats to index the human face. It then utilizes the MTCNN algorithm to detect faces in the input video. The advantage of a QR code is that it will still function properly even if a portion of it is damaged or incomplete. The input image (a face) is scanned both vertically and horizontally to create a QR code. In this approach, the MTCNN algorithm's primary task is to detect and align faces in angular keyframes.

An additional crucial component of a video indexing and retrieval system that uses a human face as a cue is face detection from video in portable devices. Researchers from around the world have developed several innovative frameworks for face detection in portable devices using input video. However, these frameworks have the following main drawbacks.

Firstly, the majority of frameworks are unable to recognize faces on portable devices. It took longer for most frameworks to detect faces in input video from lightweight devices. Due to factors such as changes in the face's direction, variations in image brightness, and changes in illumination, the accuracy of face detection rates is lower in crowded videos using current state-of-the-art frameworks.

Second, tiny faces play a significant role in input videos. For detecting small faces in input videos for video indexing and retrieval purposes, researchers worldwide have proposed various novel frameworks. The majority of frameworks struggle to detect small faces in

cluttered input videos. One of the most popular study areas is creating automated systems that can identify and locate small faces in congested input videos for video indexing and retrieval. In this thesis, Chapter 5 presents an effective framework for small-scale face detection from crowded input videos on lightweight devices for video indexing and retrieval, utilizing human faces as cues to address the aforementioned issues. Using a combination of deep learning techniques, the framework eliminates the need for face detection processes in input videos on lightweight devices. The framework for video indexing through face images using deep learning models will outline the small face detection processes from input video in lightweight devices.

Chapter 3

Video indexing through the Face Images using a Barcode

3.1. Introduction

As we have already covered in Chapter 1, the need for methods to automatically access visual data based on content is growing in tandem with the daily increase in the number of videos and images.

Our daily lives now revolve around the usage of video data in everything from traditional entertainment and radio broadcasting to camera systems for the advancement of intelligent urban settings, wearable technology, medicine, and the transportation industry. As a result, numerous strategies for video indexing and retrieval have been developed. In addition to its vast video collections, it enables systems to efficiently and successfully handle, organize, and preserve video data, allowing users to quickly satisfy a fundamental information need [122, 123, 124]. Due to its methodology, the description of images or videos on existing structures utilizes file name searching instead of the structure itself [125, 126, 127].

Deep learning is a subcategory of soft computing that utilizes millions of separate images to extract data. Multimedia scientists have studied feature representation and similarity computation in great detail for decades [128,129]. These two aspects are critical to the effectiveness of an image retrieval system based on the material. The accessibility of video content presents special difficulties because it has always been exceedingly time-consuming to index and search video. Video indexing and search can be automated with new software that leverages deep learning techniques, increasing the discoverability and utility of video content [130]. These deep learning resources include modern and intriguing methods for managing, decoding, and transcribing video. The concept of deep learning is explored in this column, along with several novel applications that can be achieved with video deep learning methods [131].

Through a snap without a tag, consumers can locate a movie's video on the Internet. The ability to search for an online lecture recording with a specific slide is another useful feature.

The scientific team has worked hard on the video recovery problem. The simplest solution to this issue is to refer to any video frame as a single image. This makes it difficult to determine whether an image is part of a video when the background is identical to that of the image [132]. Deep neural networks represent a significant development in machine learning, having been applied to a variety of real-world systems, including autonomous vehicles, sports, research, and even the arts. In various fields, including computer vision, natural language processing (NLP), and voice processing, deep learning has driven significant advancements. Rather than focusing on multimodal learning challenges, single-mode deep learning is the primary focus of most research. However, for various data types, the final identification and recovery output in a multimedia system can be significantly enhanced, especially when there are faults or missing values in one or more modes [133].

Images also include much redundant content, but videos preserve rich substance. Furthermore, for effective application, video processing and analytics require a significant amount of computational complexity, including browsing and recovery [69,134]. In many respects, retrieving related videos from a large library is, therefore, more challenging than retrieving images. Manually recovering videos is a tedious and time-consuming task for humans. Manually recovering videos is a tedious and time-consuming task for humans.

Furthermore, there is a chance of inaccurate results because humans are prone to errors [135]. One simple method for searching and capturing large amounts of video data is video keyframe removal. Additionally, it serves as the foundation for video applications [136].The image is most likely aware of the visual formation and will be destroyed in its later parts. Low-level features, such as texture and color, are used by the majority of image indexing techniques.

Low-level image and video indexing methods make it difficult to identify individuals and are not suitable for indexing people. One of the most significant elements in an image and video scenario is a human. A method for merging the identification and detection of individuals in pictures and video sequences is covered in the work [23]. The technique used in the paper [23] involves indexing the video based on a person's face, which requires more time and space.

Nowadays, the most popular source of entertainment and fun for Internet users is video. Additionally, it serves as a commercial, business, and personal inspiration via a communication channel. Therefore, when transmitting data from one device to another, the bandwidth of the communication channel is also crucial. It will require more bandwidth, time, and space if we transmit the human face as data across a communication channel. This whole issue could not be covered in Papers [23] and [63].

Face recognition research has historically concentrated mostly on recognizing faces in still images. However, familiar problems like changing illumination, pose variety, facial expression, and occlusion make recognition from a single image challenging. These elements frequently result in differences in facial image that outweigh those brought on by identity changes. The recording, storing, and analysis of face videos is now feasible thanks to the development of low-cost video cameras and increased computational power. Multiple-frame video inputs give redundant and expensive data. It is generally accepted that by precisely capturing additional information, video-based recognition can resolve the inherent ambiguities of image-based recognition, including pose variations, low-resolution sensitivity, and partial occlusion, leading to more accurate and dependable face recognition. Furthermore, facial dynamics that aid in facial identification can be captured through video inputs.

The contributions to video keyframe applications include insufficient continuity of the user-extracted keyframes and poor representation and redundancy in the retrieved keyframes. Keyframes that can generalize the original video material and are more compatible with the human visual system are to be removed [56]. Conventional frame extraction procedures eliminate redundant and identical key frames from a video without affecting the visual quality [69]. To overcome the complexity of the video picture key frames fusion problem with multiple modalities, earlier researchers have employed numerous or cross-mode hashing [137]. Conventional approaches are used to draw video evidence, which is based on frames, such as the original, middle, and end frames. The earlier method is simple to implement, but it could save time and computations by avoiding some of the videos in these mainframes [138]. Avoid focussing on the facial image ordering for key frame extraction. In this manner, the consistency of the face in these frames is taken into account, rather than when sorting the face images. It demonstrates that obtaining frames with high facial content improves face recognition accuracy [68].

Finding faces in images and video scenes, as well as indexing and retrieving those required for information retrieval based on the face image, are the goals of video indexing using the Human face approach. Using the human face method, video indexing identifies faces in images and video scenes, indexes them, and retrieves those necessary for information retrieval based on the face image. A video used for this task consists of numerous frames, not all of which are required for facial recognition. To deduce the major frames from the extracted frames, all the frames from the provided video must first be extracted. For video indexing and retrieval, keyframes are then identified from each extracted frame based on the presence of a human face.

For video indexing and retrieval using bar codes as faces, accurate and effective face detection from key frames is essential. To automatically detect faces in key frames of input videos, several artificial intelligence (AI) techniques—especially machine learning and deep learning—have been investigated in recent years. In addition to face detection, numerous methods have been studied for face recognition, image gradient computation, and the creation of bar codes for faces in key frames of input video. Chapter 2 provides a detailed review of these techniques. However, the facial expressions, positions, moods, lighting variations, occlusions of the face in pictures, and directional changes of faces in videos and images make it difficult to distinguish faces from keyframes and videos. Face recognition, gradient calculation, and consistent linear bar code generation for facial features are also difficult tasks after faces have been detected from input video and images. Therefore, it is crucial to accurately detect and recognize faces from input video and keyframes to generate barcodes from facial images.

To address the shortcomings of certain current methods [112, 117, 118, 139] for bar code tagging of human face information, we have developed a new invention called "A System and Method for Bar code representation of face image." This barcode enables us to index the video using a human face as a keyframe and retrieve it throughout the indexing process. This suggested framework uses LGFA to determine the qualification in image sparkling slopes. Additionally, the windowing technique then calls for standardization to determine the normal of the angles into a small number of intervals. This is followed by the conversion of the chart into an extremely standardized format, and the quantization effect is represented by the breaking points of decimal digits, ranging from 0 to 9. The current approach undoubtedly

produces the best-quality face pictures using a direct EAN-8 standardized tag. LGFA helps us generate gradient information from illumination-invariant face images. By using the windowing technique, we scan the input image horizontally and extract the upper 75% and 70% of the gradients. The barcodes for these two cases are extracted from face images of five datasets, and it is found that the present technique is useful for accounting, discovery, acknowledgment, and identification of people in a group.

Moreover, we first identify a video as information before extracting each edge from it during the video indexing process, which utilizes barcodes and face images. The color histogram difference is used to choose the key frames from each frame in the video. The key frame from the input video is returned as the face of the color histogram's threshold is satisfied. The Viola-Jones, MTCNN, ShuffleNet, Combined MTCNN and ShuffleNet, DSFD, BlazeFace, and YOLO v3 are used to identify faces from the extracted mainframes. For video indexing purposes, a unique EAN-8 barcode is used to identify the human face extracted from the video. The sliding window technique and the combined LGFA and sliding window approach are then used to create the image gradient from the face image. Some of the significant contributions of the work mentioned in this chapter are as follows: -

- We have proposed a video indexing through the human face mechanism in video indexing through the face images using bar code to focus the key frame extraction using the color histogram method, the face detection from various facial expressions, pose, emotion, illumination change, age, and occlusions of the face image of the video and keyframe, image gradient calculation using window technique and a linear bar code generation from image gradient, as elaborately described in subsection 3.2.1.

- In our proposed video indexing mechanism using human face recognition, the color histogram approach [140] is employed for key frame extraction, which is a color-based frame difference technique. This method's premise is that there will be little to no difference in the matching histograms of two frames with identical backgrounds and identical (although moving) items.

- Key frames of the input video are used to detect faces using the Viola-Jones object detection procedure. Information about this approach is covered in the paper [141].

- The Sliding Window approach is used to determine image gradients from the grey-scale image after the face image has been converted to grey-scale. A thorough discussion of this approach was covered in papers [114], [117], and [118].

- From the resultant face image gradient, a facial linear bar code is generated using the EAN-8 linear bar code. Information about the bar code generation technique is covered in papers [114], [117], [118], and [142].

- As detailed in our experimental results, discussions, and failure instances, we have conducted numerous experiments to develop a facial linear bar code for face images, utilizing an EAN-8 linear bar code.

- Our new invention, "A System and Method for Bar Code Representation of Face Image," uses a linear facial bar code to address the issue of illumination variation in human face images. The image gradient is calculated using the window and LGFA techniques, and a linear bar code is generated from the image gradient, as detailed in sub-section 3.2.2.

- After the face image has been converted to grayscale, our invention utilizes the LGFA and Sliding Window technique to extract image gradients from the grayscale image. This method was discussed in detail in articles [113] and [143].

- The systematized tag is created based on 70% and 75% of the top portion of the facial image, excluding the area hacked down, to assemble the nose and mouth, verifying the scanner label concept using the LGFA and Windowing Technique.

- Comparative results between windowing and combining LGFA and windowing techniques demonstrate that the combination of these two techniques can provide a

more reliable bar code from illumination-invariant facial images. The results have been elaborately reported in section sub 3.3.2

- To focus the key frame extraction using the color histogram method, the face detection from various facial expressions, pose, emotion, illumination change, and occlusions of the face image of the video and keyframe, the image gradient calculation using the LGFA and window technique, and the linear bar code generation from image gradient, we have proposed a video indexing through human face images using LGFA and window technique. This is explained in detail in sub-section 3.2.3.

- Once again, our proposed method utilizes the Viola-Jones object detection algorithm to identify faces in key frames of the input video. The publication [141] provides a detailed discussion of this method.

- Once the face image has been converted to greyscale, image gradients are extracted from the greyscale image using the LGFA and Sliding Window technique. This method was thoroughly discussed in the following papers: [113], [143], [114], [117], and [118].

- Prior technologies, including the Cloud-based Face Video Retrieval System (CBFVRS), Video Retrieval Model on Convolutional Neural Network (VRMCNN), and Fuzzy-based Support Vector Machine classifier (FBSVM), are compared to the suggested method's performance. Our method obtained an F1-score of 99.85%. Subsection 3.3.3 provides a detailed report of the results.

- To determine the best face detection algorithm for various facial expressions, poses, emotions, illumination changes, age, and occlusions in face images from videos and keyframes, we have proposed a method called "Using Viola Jones, MTCNN, DSFD, Blaze face and YOLOv3 algorithms, video indexing through the human faces represented as EAN-8 Linear bar code." This method is detailed in section 3.2.4 and involves calculating the image gradient using the window technique and generating a linear bar code from the face image.

- Our suggested approach detects faces in key frames of the input video using Viola Jones, DSFD, Blaze face, YOLO V3, and MTCNN algorithms. The papers [141], [12], [13], [14], and [15] provide an in-depth analysis of each of these approaches.

- When comparing the outcomes of the suggested method, it is evident that DSFD and YOLOv3 identify more faces from the input video, even if the MTCNN algorithm operates more quickly. The little face may be recognized from input via DSFD and YOLOv3. Following the use of portable devices, YOLOv3 and Blaze Face can identify faces. MTCNN provides faster face detection. Results are presented in detail in sub-section 3.3.4

The remainder of the chapter is organized as follows. Section 3.2 provides a detailed description of the proposed framework for video indexing through face images using barcodes. Section 3.3 contains detailed experimental evaluations, discussions, and failure cases. Concluding remarks and the future direction of this work are presented in Section 3.4.

3.2. Proposed methodology for the video indexing through the face images using bar code

This section of the chapter will primarily focus on the sequence of steps involved in developing our proposed video indexing framework, which utilizes face images and barcodes for indexing. The workflow of our proposed technique is presented in Figure 3.1. (A) The first step in

The input video is frame extraction. (B) Keyframe (based on human face) extraction from the detected frame. (C) Face detection from the key frame (D) Image gradient calculation from these detected faces. (E) Finally, linear facial barcode generation is performed from the image gradient of the detected faces in the key frame for indexing purposes.

A. Frame extraction from an input video

A dynamic video combines the frame, shot, and scene. As a result, the first step is to extract the still images, which are represented as scenes, shots, and frames in the input videos. A shot is a group of frames, while the scene is a collection of shots. Frame extraction is one of the most popular and straightforward methods for detecting human faces from keyframes in videos for video indexing, utilizing face images and barcodes.

B. Keyframe (based on human face) extraction from extracted frames

A key frame is a frame that contains important information about each shot. Human faces in various poses, expressions, and lighting circumstances are regarded as keyframes in this suggested work. Publications [74], [144], [145], [146], and [147] describe various key frame extraction techniques.

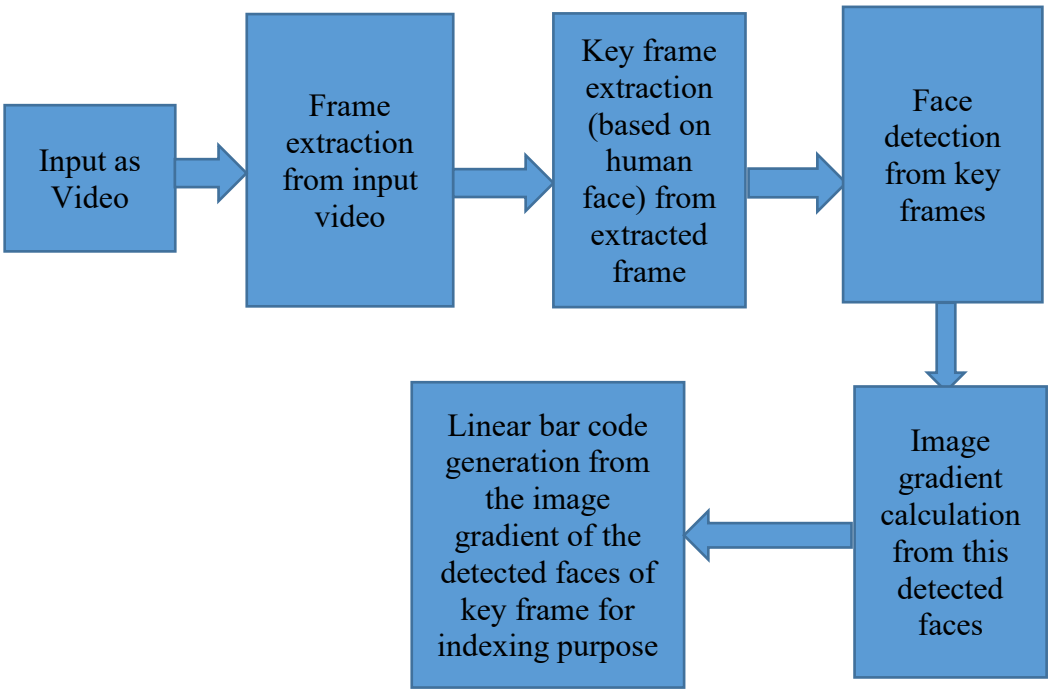


Fig 3.1: Flow diagram of the proposed method

C. Face detection from keyframe

Faces are detected from the collected key frames using various state–of–the–art methods for face detection purposes. This thesis's second chapter provides an in-depth examination of this approach.

D. Image gradient calculation from these detected faces

Following the conversion of the detected face image to a grey-scale face image, image gradients are computed using several cutting-edge methods. The following studies provided in-depth discussions of this approach: [113], [143], [114], [117], and [118].

E. Finally, linear facial bar code generation from the image gradient of the detected faces of the key frame for indexing purposes

We'll talk about how to generate an EAN 8 barcode using the received gradient value. The barcode in our approach is a linear representation of the video's recognized face image. Since these barcodes are a linear representation of face images, they can be used to index human faces found in any video.
The methods suggested for video indexing using face images using bar codes in this chapter are as follows:

3.2.1. Proposed Video indexing through human Face Images

Video indexing using low-level features is difficult for person recognition, according to a survey and analysis of various publications in Chapter 2. Video indexing through the human face also heavily relies on the facial expression, position, mood, lighting changes, and occlusions of the face image. Additionally, the inherent ambiguities of video-based recognition, including position variations, sensitivity to low resolution, and partial occlusion of facial cavities, were noted. The system must meet the following requirements to address the problems with the way face detection is currently implemented:

- The system will produce sufficiently accurate results with only slight changes in illumination, facial emotions, and facial posture.
- Results from the procedure should be sufficiently accurate, regardless of a person's age or race.
- Achieve these outcomes in almost real-time.

A novel method (video indexing through the human face) is suggested to address the aforementioned problems. This method indexes human faces found in various video formats using a linear EAN 8 bar code of face images rather than the face images themselves. The several steps of the approach suggested in this chapter, as shown in Figure 3.2 with a block diagram, are as follows: (a) The initial step involves extracting frames from the input video. (b) Using the color histogram difference, keyframes are extracted from the input video frames. (c) From the keyframe, detect faces using the Viola-Jones method. (d) Every face image has been transformed into a grey-scale face image. (e) A sliding window technique is used to compute the gradient of a grayscale face image. (f) The EAN-8 sequence table is used to create an EAN-8 barcode from image gradients.

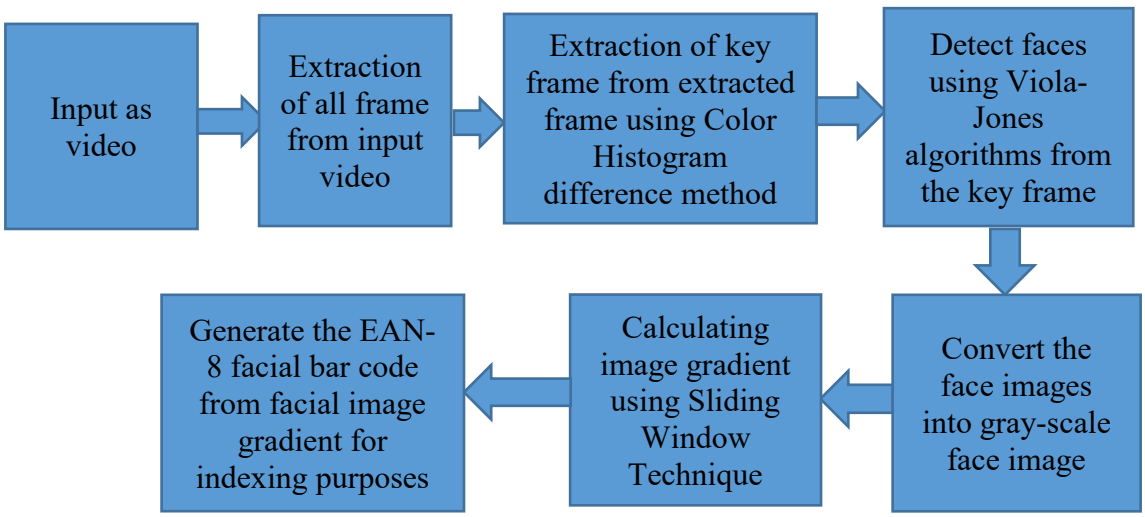


Fig 3.2: Steps for the proposed method (with block diagram)

3.2.1.1. Extracting Frames

The scenario, shot, and frame all work together to create a dynamic video. As a result, the first step is to extract the still images, which are represented as scenes, shots, and frames in the input videos. Each shot is a group of frames, and the scene is a collection of shots. Traditional video contains 20 to 30 frames per second and is packed with content. A still image that appears in a video and has extraneous information is called a frame—the frame from the Hollywood movie sequence Victory Bad is depicted in Figure 3.3.

Fig 3.3: The frame of Bad Victory Video

3.2.1.2. Extracting Keyframes

The term "key frame" refers to the frame that contains the most important information for each shot. Human faces with various expressions, poses, and lighting conditions are considered keyframes in this proposed process. Papers [74], [146], [147], [148], and [149] cover various critical frame extraction techniques. Curve saliency motion capture data, probability ratios, pair-by-pair comparisons, and many other methods are among the methods. However, in this approach, the keyframe is extracted from every frame of a given video using the Colour Histogram technique. The difference in the color histogram is used to extract key frames from the frames. The frame will be chosen as the following keyframe if the difference that is obtained is higher than the threshold value. The level of fluctuation in the color histogram is known as the threshold. Some key frames based on a human face from a Hollywood movie (Victory Bad) are displayed in Figure 3.4.

A. The Colour Histogram Method

Key frame extraction is a color-based frame difference technique that utilizes a color histogram to identify key frames. This is because the color is one of the most crucial visual elements in describing an image. When the camera moves, a color histogram remains stable and is easy to calculate. This method is based on the theory that two frames with identical backgrounds and identical (but moving) objects will have similar histograms. The choice of threshold is a crucial component of this approach. The following equation (3.1) can be used to determine the color histogram difference between two successive frames:

$$D\left(F_I, F_{I+1}\right) = \sum_{J=1}^{n} \frac{\left|h_I\left(J\right) - h_{I+1}\left(J\right)\right|^2}{h_{I+1}\left(J\right)} \qquad (3.1)$$

In this case, h_I and h_{I+1} stand for the histograms of the two consecutive frames, F_I and F_{I+1}, respectively. A shot transition takes place when D (F_I, F_{I+1}) exceeds the specified threshold.

The frame's number is "n." Below is an explanation of Algorithms 3.1 and 3.2, which determine the color histogram difference between two images and generate keyframes from the extracted frames, respectively.

Algorithm 3.1: An algorithm for determining the difference in color histograms between two images.

Input: The two distinct images (Image 1 and Image 2).
Output: The two images' histogram differences (images 1 and 2)
Steps

1. Start
2. For each of Images 1 and 2, compute the image histograms along each RGB channel (1, 2, and 3).
3. Use the formula histogram_norm=histogram/max (histogram) to normalize each histogram for each of the three channels in each image.
4. Determine the histogram error, or the Euclidean distance difference, between two images for every channel as follows:
   heR= $(histogramR1 - histogramR2)^2$ // for red color, heR means histogram error for red color.
   heG= $(histgramG1 - histogramG2)2$ // for Green color, heG means histogram error.
   heB =$(histgramB1 - histogramB2)2$ // for the blue color, heB means histogram error.
5. Determine the difference in the color histogram by taking into account each channel's contribution as follows:
   Histogram Difference = 0.2989*heR+0.5870*heG +0.1140*heB
6. End

Algorithm 3.2: An algorithm that uses a video input's extracted frames to create keyframes.

Input: A video file
Output: Key Frame from a video input frame that has been extracted
Steps:

1. Start
2. Read the user input video
3. Record the number of frames in the video.
4. From all of the frames in the video, select the keyframe by comparing the two corresponding frames.
5. Determine the differences in color histograms between the two corresponding frames.
6. After that, compare it to the threshold value. // The color histogram's threshold is the degree of variation.
7. If the threshold is exceeded or equivalent to the color difference of the chosen frame.
8. This frame is, therefore, regarded as the keyframe.
9. Else, proceed with steps 4 through 7.
10. The process is terminated once all of the key frames have been detected through a comparison of every frame in the video.
11. Stop

Fig 3.4: Key Frame of Victory Bad Video (based on the face as key content)

3.2.1.3. Apply the Viola-Jones method to detect faces from the keyframe

Faces are detected from the retrieved key frames using the Viola-Jones object detection procedure. The details of this method are covered in the paper [141]. Although training is slow, the Viola-Jones algorithm has the advantage of quick detection. Figure 3.5 shows a few screenshots of Key frames' face detection.

Fig 3.5 Cropped faces from the keyframes of the Holly Hood Movie's Victory terrible scene.

3.2.1.4. Grey scale face image after face image to greyscale face image conversion

The gradient of the face image must be calculated once the faces from keyframes have been converted to greyscale.

3.2.1.5. Calculating an image's gradient with the Sliding Window Technique

Using the Sliding Window approach, image gradients are computed from this greyscale image (faces). A thorough discussion of this approach was covered in papers [114], [117], and [118]. This technique calculates the image's gradient values by considering the top 70% of the face

image. The top 70% of the face image is used to create a stable bar code. A stable barcode can be made from this image gradient if the sliding window is positioned at the top 70% of the facial image, according to a paper [118].

3.2.1.6. Using the EAN 8 sequence table, a barcode is used as an index

This section will cover how to generate an EAN-8 barcode using the gradient value obtained. Our approach utilizes a barcode that represents the face image linearly found in the video. Since these barcodes are a linear representation of face images, they can be used to index human faces found in any video. One benefit of indexing using barcodes from human faces is that they require less storage space and indexing time. An algorithm for the bar code-generating process is also presented in this publication, along with specifics of the bar code-generation technique covered in papers [114], [117], [118], and [142]. The barcode examples in Figure 3.6 were found in a variety of Hollywood (victory bad scene) facial video datasets. A description of Algorithm 3.3, which utilizes the gradient values of facial images to generate a linear EAN-8 bar code, is provided below.

Fig 3.6: An EAN 8 barcode is displayed alongside the cropped face from the Victory Bad Video scene.

Algorithm 3.3: An algorithm for creating EAN 8 bar codes based on the gradient values of gathered facial images.

Input: The face image gradient that was extracted from the input video.
Output: EAN 8 Face image linear barcode.
Steps:

1. Start
2. Determine the gradients' maximum value.
3. Set the maximum gradient value as the initial value for max_gradient.
4. for i= 1 to T-(windowsize+1) do
   gradient (i) = gradient (i)/max_gradient // NORMALIZATION
   end
5. for i=1 to T do
   gradient(i)=floor(gradient(i)*10) // QUANTIZATION
   end
6. Build a barcode zeros matrix.
7. Initialize scale=9.5
8. for i=1 to 7 do
   initialize num with value 0
   for j=1 to m do
   num=num+gradient((i-1) *m+j)
   end
   barcode(i)=round((scale*num)/m)
   end
9. S_odd=barcode (1) +barcode (3) +barcode (5) +barcode (7) // S_ odd means summation of odd position number present in EAN 8 barcode
10. S_even= barcode (2) +barcode (4) +barcode (6) // S_even means summation of even position number present in EAN 8 barcode
11. barcode (8) = mod(10-mod((3*S_odd+S_even),10)10 // Calculate Checksum Value
12. for i= 1 to odd, do
    Store values of the barcode(i) into barcode
    End
13. Use the EAN 8 sequence table and gradient values to generate graphical storage and store it in a barcode image.
14. Stop

3.2.1.7. Barcode determination accuracy

Based on the eight-digit EAN-8 barcode of two identical or dissimilar facial images, the accuracy of the human face barcode is calculated. Since the checksum digit is an error correction digit, it is not taken into consideration here. Seven more digits are taken into

consideration. After comparing the numbers in two EAN-8 barcodes at the same location on two human faces from the same image or a different image, accuracy is determined.

In the same location, two distinct barcode numbers are compared. Two distinct barcodes with the same position number are regarded as the same face if their images are identical. To determine a bar code's correctness, compare all of the bar codes created using the input video's facial image. If the accuracy of the two bar codes is more than 80% after comparison, they are regarded as the same faces; if it is less than 60%, the bar code of the image is considered to be different faces. These face bar code types aren't considered indexing, though, if the accuracy is less than 60% and the faces in the images match the bar code. However, it can be indexed to maintain records. It may lead to ambiguity in this situation. Bar code accuracy can be determined using the following equation (3.2).

$$\text{Accuracy} = ((F-I)/F) * 100 \qquad (3.2)$$

Where “I” is the initial value, or zero, and “F” is the barcode's size.

3.2.2. Invented System and Method for Bar code representation of face image

This subsection provides a comprehensive explanation of our latest invention, "A System and Method for Bar Code Representation of Face Image," which addresses the issue of illumination variance in human face images by utilizing a linear facial bar code. The window and LGFA approaches are used to determine the image gradient, which is then used to create a linear bar code. This barcode enables us to index the video using a human face as a keyframe and retrieve it throughout the indexing process.

Using the LGFA and Windowing techniques, this method determines the qualification in image sparkle slopes. Next, it uses standardization to categorize the angles into a small number of intervals. The chart is then transformed into an extremely straight standardized identification, and the quantization effect is transferred into decimal digit breaking points ranging from 0 to 9. Undoubtedly, the current approach generates the best face picture quality for direct EAN-8 standardized tags.

Regardless, the current study proposes a system that uses the window approach and LGFA to retrieve edge-based information. Using the windowing technique, we scan the input image horizontally and extract the top 75% and 70% of the gradients. LGFA helps us create gradient

information from illumination-invariant face images. It has been discovered that the current technique helps account for, locate, and identify individuals within a group. The barcodes for these two scenarios are taken from images of the faces of five datasets. The four phases that make up the method that is suggested in this section and illustrated in Figure 3.7 are (a)

Preliminary image processing, (b) Feature extraction with the window and LGFA techniques, (c) Feature coding, and (d) Barcode generator.

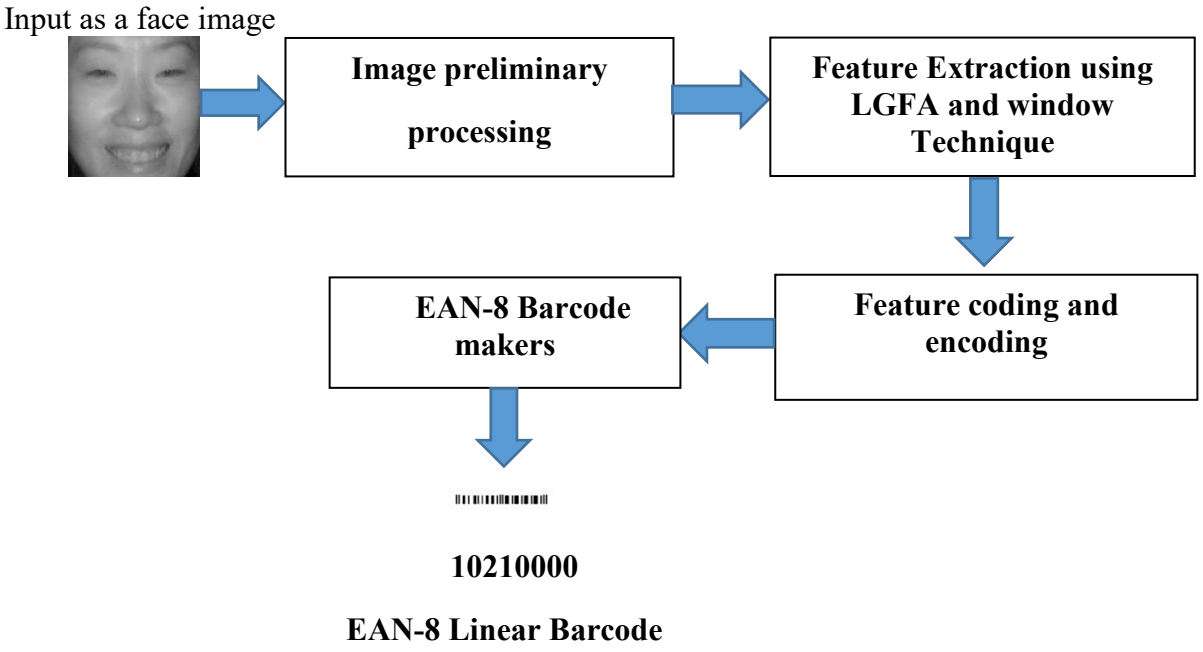


Fig. 3 .7: EAN-8 scanner tag generation ventures.

3.2.2.1. Initial processing of images

After cropping the facial area from the background, noise is eliminated using median filtering.

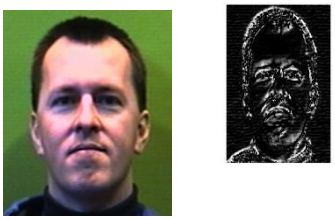
3.2.2.2. Utilising the LGFA and window technique for feature extraction

LGFA is used to extract features. According to the research [113], the illumination reflectance model serves as the foundation for LGFA. The Division Method was used to remove the light portion of the image by setting up sets of connected pixels. The law of global gravitation, as discussed in papers [148] and [113], and local gravity face (LG-face), as described in papers

[113] and [143], are combined in LGFA. The paper [113] discusses the formulas used to calculate local gravity faced. Equation 3.3 uses the following formula to determine the local gravity face angle.

$$\alpha = \arctan\left(\frac{-\frac{R_2}{2\sqrt{2}} - R_3 - \frac{R_4}{2\sqrt{2}} + \frac{R_6}{2\sqrt{2}} + R_7 + \frac{R_8}{2\sqrt{2}}}{-R_1 - \frac{R_2}{2\sqrt{2}} + \frac{R_4}{2\sqrt{2}} + R_5 + \frac{R_6}{2\sqrt{2}} - \frac{R_8}{2\sqrt{2}}}\right) \quad (3.3)$$

Gradients are calculated using the formula in Equation 3.3, and the resulting image is displayed in Figure 3.8(b).



(a) (b)

Fig 3.8: (a) Real-life facial image (b) The gradient of the real face image following the LGFA application

The window technique is used in the image gradient after the gradient is obtained using LGFA; this is covered in the papers [114], [117], and [118]. The windowing approach has been utilized to create a stable barcode.

3.2.2.3. Encoding and feature coding

Encoding is carried out to a significant number of decimal places. The paper [118] discusses the subsequent errands that are carried out for features code. The angle's vector components are standardized to an extreme inclination estimate (max_gradient).

3.2.2.4. Barcode generator in EAN-8

The final eight-digit EAN-8 standardized tag is now produced. The final information picture produced, known as an EAN-8 direct standardized identification, is shown in Figure 3.9. The total number of digits in this scanning tag is eight. Out of the eight digits in the yield, the

eighth is thought of as the checksum digit for the first seven. The publication [118] provides a detailed description of the EAN-8 barcode and its checksum digit. The checksum digit in Figure (3.9) is 2. The purpose of the checksum digit is to identify and fix mistakes.

11001102
EAN-8 Linear Barcode

Fig 3.9: A linear barcode in EAN-8

3.2.3. Proposed Video indexing through human face images using LGFA and window technique

We have proposed a "video indexing through human face images using LGFA and window technique" in this section to focus the key frame extraction using the color histogram method, the face detection from various facial expressions, pose, emotion, illumination change, and occlusions of the face image of the video and keyframe, the image gradient calculation using the LGFA and window technique, and the linear bar code generation from image gradient. In addition to addressing time and spatial complexity, this work aims to tackle significant challenges, including changes in posture, storage capacity limitations, and the inability to preserve key video frames.

When using a human face for video indexing, the image's posture, mood, illumination changes, occlusions, and facial expressions are all crucial components to consider. It has also been demonstrated that intrinsic ambiguities, including inadequate resolution sensitivity, posture variations, and partial facial cavity blocking, hinder video-based recognition. A method must be created to address the following needs and overcome the problems with face detection implementation. The system will enable sufficiently accurate performance with minor variations in lighting conditions, facial expressions, and face position. Invariant facial image lighting should yield reasonably detailed results regardless of the person's age or skin tone.

Significant problems with the current methods include limited storage capacity, inability to save key frames from videos, complexity in time and space, changes in posture, and partial facial cavity blockage. Therefore, a color histogram is employed to identify the differences between two images and eliminate the problem of posture change, thereby addressing these issues. To extract key frames from the face, the Viola-Jones algorithm has been used to crop images of the face.

A novel hybrid sliding windowing technique and LGFA have been introduced to determine image gradients, thereby reducing time and space complexity and increasing storage capacity. The problem of storing keyframes is resolved by proposing the use of barcodes as an index, utilizing the EAN-8 sequence table, which enables key frames from videos to be stored using EAN-8 barcodes.

Figure 3.10, which includes a block diagram, illustrates all of the stages of the suggested approach. (a) Recovering every video input frame is the initial stage. (b) From the frames acquired from the video input, keyframes are extracted using the color histogram discrepancy.

(c) The Viola-Jones method can be used to identify faces in the primary image. (d) Every facial image is transformed into greyscale. (e) Using LGFA and the sliding window technique, the gradient of the face image is computed. (f) Using the EAN-8 sequence table, the image's gradients generate the EAN-8 linear barcode

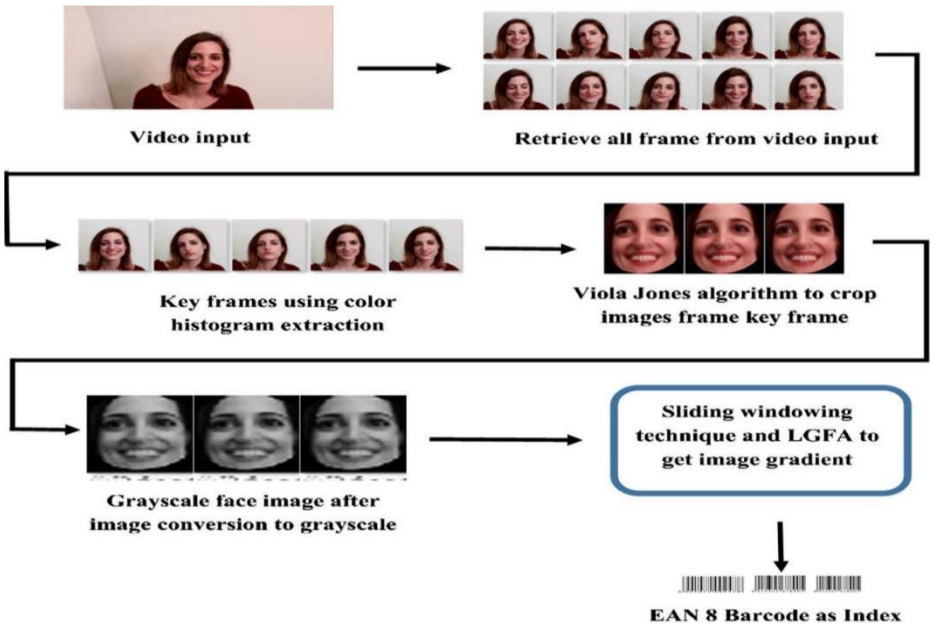


Fig 3.10: Block Diagram for the Suggested System

3.2.3.1. Frame Extraction

Since the scene, shot, and frame together make up a dynamic video, the first step is to extract the still images from the input videos that are represented as a scene, shot, and frame. The frame is a still image that contains extraneous information found in a video. The frame seen in the "Good as It Gets" (plate number 01766) is shown in Figure 3.11. A scene from the feature film Holly Wood

Fig 3.11: The frame of As Good as It Gets-01766 Hollywood movie scene

3.2.3.2. Keyframe Extraction

The key frame is the frame that contains the most important information about each shot. In this suggested approach, the human face, with its unique expressions, posture, lighting, and illumination, is regarded as a keyframe. The color histogram method is used to extract the keyframe from all frames of a given video. The specifics of the Colour-Histogram method are covered in subsection 3.2.1.2. When the threshold for the color histogram disagreement is reached, the frame will be chosen as the subsequent keyframe. Figure 3.12 shows a few of the main shots of a human face in a Hollywood film (As Good as it Gets-01766). Subsection 3.2.1.2 of this chapter covers the Colour-Histogram and key frame extraction algorithms.

Fig 3.12: Key Frame of as Good as It Gets-01766 (based on the face as key content)

3.2.3.3. Face images cropped from key frames using the Viola-Jones algorithm

The Viola-Jones object detection method is used to detect faces in the extracted keyframes. The advantage of the Viola-Jones method is that while training is slow, detection is quick. When a grey-scale algorithm is applied to a picture, it recognizes multiple smaller sub-regions and looks for certain features in each sub-region to try to find a face. It entails checking multiple sizes because a picture will have numerous sides of varied sizes. Viola-Jones was created for frontal faces rather than sideways, backward, or upward faces. To speed up encoding and reduce the amount of information available, images are converted to grayscale until a face is detected. The Viola-Jones algorithm first detects the face in the grey picture before determining the location of the colored picture. Figure 3.13 displays a few Key Frame Face Detection screenshots.

Fig 3.13: Faces cropped from keyframes in the Hollywood film As Good as It Gets-01766

3.2.3.4. After converting an image to greyscale, a greyscale face image

After obtaining the faces from keyframes, the face picture must be converted to greyscale to determine its gradient.

3.2.3.5. Using the LGFA and sliding window techniques, the gradient of the facial image is calculated from greyscale

From this greyscale image (faces), image gradients are extracted using the Sliding Window approach and LGFA. An image gradient from an illumination-invariant facial image is calculated using the Local Gradient Feature Alignment (LGFA) method. The sliding window technique is shown to be ineffective in generating stable bar codes from illumination-invariant face images in the paper [130]. As a result, a stable barcode is created by combining the LGFA and sliding window approaches.

The input image is scanned using the sliding window approach from the forehead to the upper lip area. This approach considers the top 70% of the facial image when calculating the image gradient values.For the development of a secure barcode, up to 70% of the face image is captured. The stable barcode can be created from the image's gradient if the sliding window moves over the top 70% of the facial image.

3.2.3.6. Barcode as an index utilizing the EAN- 8 sequence table

This section will cover the process of generating an EAN-8 bar code using the obtained gradient value. The barcode in this methodology is a linear representation of the face image that was recognized from the video. Since these barcodes are a linear representation of face images, they can be used to index human faces that are identified from any video. The benefit of human face barcode indexing is that it takes less time and storage space to index these barcodes. Figure 3.14 displays instances of certain LG faces and their associated barcodes, which were identified from a variety of Hollywood face video datasets (As Good as It Gets-01766).

Fig 3.14 Faces in the as Good as It Gets-01766 video scene and the associated EAN 8 barcode are cropped.

In this chapter, section 3.2.1.6, the bar code generation technique based on the obtained gradient values of the facial images is discussed.

3.2.3.7. Accuracy of barcode determination

Using two comparable or dissimilar images of the face and an eight-digit EAN-8 bar code number, the correctness of the human face bar code is determined. Section 3.2.1.7 of this chapter discusses the specifics of bar code determination accuracy and formula.

3.2.4. Proposed method "Using Viola Jones, MTCNN, DSFD, Blaze face, and YOLOv3 algorithms, video indexing through the human faces represented as EAN-8 linear bar code."

Our proposed method uses Viola Jones, DSFD, Blaze face, YOLO V3, and MTCNN algorithms to detect faces in key frames of the input video. Major issues, including posture shift, lighting invariant characteristics, and facial angular changes, are not accurately addressed by the Viola-Jones algorithm or paper [130, 149]. However, in real-time lightweight devices that employ the Viola-Jones algorithm, face detection takes longer. Paper [150] does not sufficiently address important aspects, such as changes in posture, lighting-invariant characteristics, and facial angular alterations. MTCNN cannot reliably identify small faces in input video after employing lightweight devices.

"Video indexing through the human face as an EAN-8 linear bar code using Machine learning and Deep learning algorithm," the suggested approach, provides a fresh approach to these problems. This study addresses several significant issues, including the complexity of time and space, storage space limitations, facial angular variations, illumination-invariant features, shifting posture, and the difficulty in saving crucial video frames. The main contribution of the research is video indexing using human faces as an EAN-8 linear barcode, achieved through machine learning and deep learning methods. This method is proposed to solve these issues.

- To extract the key frames from the frame and solve the keyframe storage difficulties, cropped face portraits were created using Viola Jones, DSFD, MTCNN, Blaze Face, and YOLO v3 algorithms.

- The key frame extraction and face detection methods of Viola Jones, DSFD, MTCNN, Blaze Face, and YOLO v3 are then used for comparison.
- Each person will receive a unique linear EAN-8 barcode, as described in the papers [117, 130, 149], that they can use to access a database. A linear EAN-8 barcode will also be produced. It is necessary to use the appropriate type of barcode because not all linear EAN-8 barcodes are compatible with human faces. To prevent such issues, choosing the right linear EAN-8 barcode size is also essential.
- This method can also be used to recognize small faces in input video.

Therefore, the suggested approach of indexing movies addresses the problems of changing posture, storage capacity limitations, storing video keyframes, and time and spatial complexity. Compared to the original face image, the EAN 8 linear bar code requires less time and storage space. Sending this bar code over the communication channel, however, takes less bandwidth than sending the actual facial image. Still, one drawback of the barcode representation of facial images is its instability. As previously mentioned in studies [117, 118], the sliding window technique is employed in this investigation to generate a reliable barcode. In this method, faces in angular keyframes (faces) are mostly identified and aligned by the MTCNN algorithm.

Security cameras and facial recognition systems are just two of the many applications that benefit from DSFD's high-speed optimization and real-time face detection capabilities. To improve accuracy and handle complicated situations, a batch complex mining strategy and a multi-scale approach are employed. Blaze Face's design is optimized for mobile and embedded devices, allowing it to operate with little power and processing resources. A single neural network predicts bounding boxes for each class, enabling YOLO v3 to complete face detection tasks from the input video efficiently. YOLO v3 (You Only Look Once version 3) is a neural network approach that simultaneously predicts bounding boxes and class probabilities. Figure 3.15 illustrates a block diagram for each of the processes outlined in the strategy presented in this section. (a) The first phase in the input video is frame extraction. (b) The

color histogram difference is used to separate the keyframes from the frames extracted from the input video. (c) Faces in the key frame are recognized by Viola Jones, DSFD, Blaze face, YOLO V3, and MTCNN algorithms. (d) A linear EAN-8 bar code is generated by this key frame from a human face or a face that a person has detected.

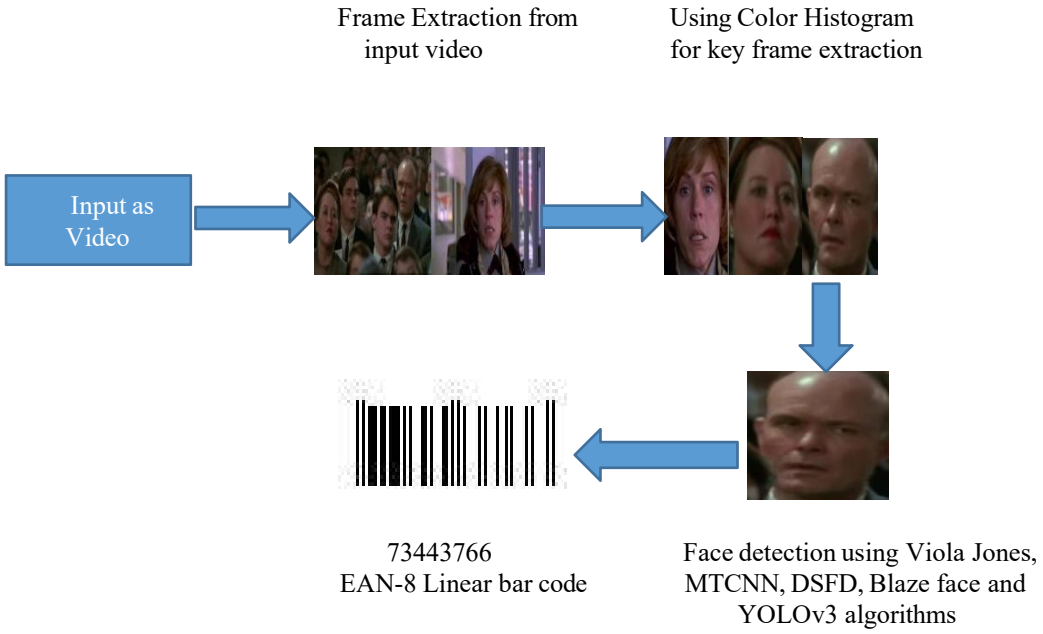


Fig 3.15: System Block Diagram Concept

3.2.4.1. Accessing the frame

The initial phase involves extracting still images, which represent the scenes, shots, and frames contained in the input videos. The still image in the frame is full of extraneous details and appears in a video. A frame from the Hollywood film Fargo is shown in Figure 3.16.

Figure 3.16: The frame from the Fargo video

3.2.4.2. Using a Colour Histogram to extract the keyframe

The key frame contains the most important information about each image. Human faces with distinctive expressions, positions, lighting, and illumination define the keyframes in this work. The Difference in Colour Histogram can be used to retrieve key frames from the frames, which was already discussed in Section 3.2.1.2. The frame is selected as the subsequent keyframe if the measured difference exceeds the threshold magnitude when the color histogram disagreement threshold is met. A few key frames from the Hollywood film Fargo, starring Holly Wood, are shown in Figure 3.17.

Figure 3.17: Fargo's Key Frame, where the face is the main component

Section 3.2.1.2 of this chapter explains the formula and algorithm for separating the two successive frames in the Colour Histogram.

3.2.4.3. Viola Jones, MTCNN, DSFD, Blaze-face, and YOLO v3 algorithms were used to crop facial images from the keyframes

The best face detection technique from the input video's key frame is determined by comparing the results of the Viola-Jones, MTCNN, DSFD, BlazeFace, and YOLOv3 algorithms. Over time, several methods have been developed to help computers recognize faces. The Viola-Jones method, which utilizes Haar-Cascades to accurately pinpoint faces, is the first technique employed in this work to identify faces from the keyframe. The Haar cascade technique can detect objects in images regardless of their size or position. Each cascade window's features are calculated to determine whether or not it could be an item. It

works like this. Haar features are evaluated and matched using a window scan of the sample size.

Deep learning has led to the development of improved methods that can detect the human face at various viewpoints, illumination conditions, and clarity levels. Among these the best methods is MTCNN. The Multi-Task Cascaded Convolution Neural Network, or MTCNN, is a deep learning-based face detection system. Even if faces are partially obscured or have varied sizes, orientations, or positions, it does a great job of recognizing them in pictures. Faces are reconstructed from the recovered key frames using the MTCNN object detection technique. The system uses a cascaded combination of three different neural networks to recognize faces. In the first network, the face is located in the image; in the second, its contours are defined; and in the third, facial features such as the eyes, nose, and mouth are identified to enhance detection. MTCNN is widely used in various applications, including facial recognition, facial expression analysis, and facial attribute detection.

The Dual Shot Face Detector (DSFD) algorithm is a face detection system that recognizes faces in images by utilizing two distinct neural networks. After the first network generates a set of candidate face regions, the second network refines these regions to improve accuracy and lower false positives. To handle difficult scenarios and increase accuracy, a batch hard mining strategy and a multi-scale approach are employed.

The popularity of smartphones and other low-end devices has fueled a well-known desire for improved models that are compatible with them. Blaze Face is an advanced, portable, and highly successful deep-learning face detection method. Its primary objective is to provide fast and accurate real-time face recognition, making it ideal for use cases that demand rapid processing. A single neural network at the heart of the Blaze Face algorithm uses a feature pyramid to identify faces of different sizes and resolutions. Designed for mobile and embedded devices, Blaze Face can operate with minimal power and processing resources. The Blaze Face model architecture is built upon four fundamental design principles: (a) increasing the receptive field sizes, (b) feature extraction, (c) anchor scheme, and (d) post-processing.

A real-time face detection technique called "You Only Look Once version 3," or "YOLO v3," utilizes deep learning to swiftly and accurately recognize faces in images and videos. To identify faces, YOLO v3 divides the input image into a grid, creating bounding boxes and class probabilities for each cell. The method predicts multiple bounding boxes per cell, enabling it to recognize multiple faces within a single image. Because bounding box and class probability prediction can be accomplished with a single neural network, YOLO v3 can efficiently complete face identification tasks. YOLO v3 (You Only Look Once version 3) is a comprehensive neural network method that simultaneously predicts bounding boxes and class probabilities. This differs from the traditional strategy of earlier face detection techniques, which involved modifying classifiers to meet detection requirements. Figure 3.18 displays some screenshots of the keyframe face detection function. This demonstrates that compared to the Viola-Jones, DSFD, Blaze Face, and YOLO v3 algorithms, the MTCNN method has a greater face detection ratio.

Fig 3.18: Faces clipped in Dead Poets Society and Fargo keyframes (from the film Holly Hood)

3.2.4.4. Converting an image to greyscale and then creating a greyscale facial image

To establish the gradient of the facial image, the visages from the important frames must be obtained, and the image must then be converted to greyscale. In this way, faces are transformed into greyscale images.

3.2.4.5. Sliding Window Method for Calculating Image Gradients

The Sliding Window method is used to calculate image gradients from this greyscale image (faces). The details of this approach were discussed in papers [118]. The process above uses the top 70% of the facial image for calculating the image gradient values. A stable bar code is

created by capturing the top 70% of the face image. In the article [118], it was mentioned that a stable bar code might be produced from this gradient image if the sliding window moved up to 70% of the face image.

3.2.4.6. Using a linear EAN-8 barcode and a human face index

From each face, a linear EAN-8 bar code is created, recognized by the key frame of the input video, and saved as an index. Our method uses the barcode as a linear representation of the video-identified face image. Since these bar codes are a linear representation of face images, they can index human faces in any video. It takes less indexing time and storage space to use bar codes that are based on human faces. The algorithm for the bar code generation process is described in this publication, along with the specifics of the bar code generation technique discussed in studies [117] and [118]. Figure 3.19 illustrates several examples of barcodes.

Several Hollywood face video datasets (e.g., Dead Poets Society and Fargo) were used to identify them.

Face	Barcode Values	EAN-8 Barcode
	99343435	
	73443766	

Fig 3.19: The cropped faces and corresponding EAN 8 barcodes from the Fargo and Dead Poets Society video clips.

The algorithms for generating barcodes from the acquired facial image gradient values are discussed in Sections 3.2.1.6 and 3.2.1.7, respectively, along with an assessment of their accuracy.

3.3. Experimental Result and Discussion

In this section, we evaluate the performance of video indexing using face images and barcodes. A thorough explanation of the datasets used in this investigation is given in subsection -3.3.1. A result of our proposed approach, "Video indexing through human Face Images," will be discussed in subsection -3.3.2. The results and explanation of our proposed invention, "A System and Method for Bar Code Representation of Face Image," are detailed in subsection -3.3.3. The detailed results and explanation of our proposed approach for creating linear EAN-8 bar codes from image faces for video indexing applications will be presented in subsection 3.3.4, "Video indexing through human face images using LGFA and window technique." "Using Deep Learning algorithms, video indexing through the human faces represented as EAN-8 Linear bar code" is the title of the subsection -3.3.5, which contains the detailed results and discussion of our suggested method for linear EAN-8 bar code synthesis from face images for video indexing applications.

3.3.1. Dataset Description

In the "Video indexing through human Face Images" method we suggested, the key frame was first taken from the video dataset that had human faces. The barcode was then generated from this cropped face. The barcode of the person's face that appears in the video is then used to index the content. Some video data sets, on the other hand, contain the key frame of the facial picture directly; therefore, we generate the bar code from these datasets without removing the keyframe. (A dataset of YouTube facial videos).

We evaluate our procedure using three distinct video datasets. First, we offer a realistic TV series video dataset [151] consisting of 27 episodes from six well-known TV shows. The 27 episodes included Sons of Anarchy (3), Mad Man (3), Modern Family (6), How I Met with

Your Mother (8), Breaking Bad (3), and 24 (4). These videos have a combined duration of 16 hours. In this film, there are 6231 acts and 30 actions in all.

Hollywood video dataset [152] is then used for the experiment. This contains clips of video from thirty-two human action moves. To mark the sample, one or more eight groupings must

be present. The 20-film data set is divided into two practice sets, each comprising 12 films and a test set that is distinct from both. The 233 video samples in the automated learning set were gathered through automatic script-based action labeling, and the classifications achieved a roughly 60% accuracy rate. A clean training set of Hollywood results contains 211 video samples with manually confirmed labels and 219 video samples with manually checked labels.

Lastly, the tests utilize YouTube face datasets [153]. This collection comprises 3,425 videos featuring 1,595 distinct individuals. Every video was collected from YouTube. Each subject can access a recording at a normal of 2.15. The longest clasp has 6070 edges, the shortest has 48 outlines, and the typical length of a video clasp is 181.3 edges.

In the "A System and Method for Bar code representation of face image" approach that we proposed, tests are conducted to assess the robustness of the systematized face picture tag under different facial affirmation scenarios. The following datasets are used in this process: "Face94" [154], "Face Dataset YaleB" [155], "FERET face dataset" [156], "FG-NET creating

face dataset" [157], "NIR-VIS edited dataset" [158], "LWF Dataset" [159], and "a composite appearances face dataset at changed ages."

The "Face94" dataset includes 153 distinct individuals, each with 20 photos (133 males and 20 Females) and 100 classes with 11 photographs. Each image is 180 by 200 pixels in size. Between images, there are significant variations in expression. The location of the head has changed slightly. The data set image format is JPEG.

The Yale Face Database includes 5,850 photos of 10 distinct persons. There are 9 x 64 illumination conditions in the database. For every subject in a specific position, a picture with ambient (background) illumination was also taken. JPEG and BMP are the file formats. There are 14051 images with 10465 distinct subjects in the FERET Face database. The expressive

pose, lighting, and available illumination in the database vary significantly. This database is in the JPEG file format.

FG-Net is a dataset used for cross-age face recognition and age estimation. It consists of 1,002 pictures of 82 individuals aged 0 to 69, with an age difference of up to 45 years. Four

recording sessions were conducted to gather the images for the NIR-VIS 2.0 database: spring 2007, summer 2009, autumn 2009, and summer 2010. The first session is identical to the one used in the HFB database. The NIR-VIS 2.0 database contains a total of 725 subjects. Each participant has 5–50 NIR and 1–22 VIS facial images.

The following information is contained in the NIR-VIS 2.0 database: (1) VIS images in JPEG format and NIR images in BMP format are examples of raw images. They both have resolutions of 640 x 480. (2) The VIS and NIR pictures' eye coordinates. An eye detector automatically labels them [160], and several incorrect coordinates are manually fixed. (3) Versions of the raw VIS and NIR images that have been cropped. The procedure is carried out using the ocular coordinates, and the resolution is 128 x 128.

To investigate the issue of unconstrained face recognition, a database of images of faces called Labelled Faces in the Wild (LFW) was created. Researchers at the University of Massachusetts, Amherst, produced and are responsible for maintaining this database. The Viola-Jones face detector identified and cantered 13,233 images of 5,749 persons that were

gathered from the internet. The dataset includes two or more different photographs of 1,680 individuals. Four distinct sets of LFW images and three distinct types of "aligned" images are available in the original database.

The Composite Face dataset contains 20 composite faces of various ages. The database's representation varies with human age in this database. Both PNG and JPEG formats are available in the database.

The key frame was initially extracted from the video dataset containing human faces in the "Video Indexing through Human Face Images using LGFA and the sliding window technique." method that we proposed. The bar code was then generated using this cropped

face. A human face bar code that appears in the video is then used to index the content. The face image key frame is explicitly present in some video datasets; thus, the barcode is generated directly from the dataset without the key frame being extracted (Face video dataset for YouTube). To verify the approach, three different video data sets will be used. The Hollywood video dataset [152] is first used for the experiment. Next, show a realistic TV

series video dataset [151] with 27 episodes from six well-known TV shows. Lastly, YouTube facial data sets [153] are used for the tests.

The bar code for this experiment was generated using the cropped face of the keyframe, which was collected from the human face-based video dataset, according to the "Using Deep Learning algorithms, video indexing through the human faces represented as EAN-8 Linear bar code." approach we proposed. Then, indexing is done using the EAN-8 linear bar code of the human face in the video. However, some video datasets (such as the Face video dataset for FDDB, WIDER, and LFW) specifically contain a facial image keyframe; hence, the EAN-8 linear bar code is generated from these datasets without the key frame being removed. Four separate video data sets were used to validate the approach.

The initial step of the investigation uses the Hollywood video dataset [152]. There are additional clips from thirty-two human action flicks. The instance must be assigned a label from one of the eight categories. To construct the 20-film data set, the test set is divided into two 12-film practice sets. The data were collected via an automatic script-based action labeling process, and approximately 60% of the 233 video recordings in the automated

learning set had accurate labels. A clean training collection of Hollywood outcomes consists of 291 video samples with manually confirmed labels and 211 videos with manually tested labels.

The Face Detection Dataset and Benchmark (FDDB) dataset consists of labeled faces from the Faces in the Wild dataset [164]. The images range in size from 229x410 to 363x450, and a total of 5,171 faces have been annotated. This dataset presents several challenges, including low resolution, out-of-focus faces, and problematic stance angles. There are images in both color and grey scale.

The benchmark face detection dataset, called WIDER FACE, is derived from the publically available WIDER dataset [161]. This dataset contains 32,203 images that identify 393,703 faces, and the sample images reveal considerable differences in size, position, and low contrast. The 61 event classes were used to plan the WIDER FACE dataset. For each event class, 40%, 10%, and 50% of the data are randomly selected to serve as training, verification,

and test samples, respectively. Similar to the Caltech and MALF datasets, the WIDER FACE dataset uses the same assessment metric as the PASCAL VOC dataset. Finally, LFW [159] data sets are collected for the experiments. The original database contains four sets of LFW images as well as three types of "aligned" images.

3.3.2. The result of our suggested method, "Video indexing through human Face Images."

To confirm that the bar code generated by our suggested "Video indexing through human Face Images" is accurate, we conducted several experiments in this subsection. The following tables display the results of various key frame extraction, face detection, and valid bar code generation techniques on different datasets. The bar code generated from a portion of the Hollywood Movies Dataset is displayed in Table 3.1. This data set's videos are all in AVI format. Table 3.1's results show that while the number of faces found in the Butterfly Effect -00696, Forrest Gump -01277, and Gandhi -02262 datasets is 15, 35, and 5, respectively, the number of valid bar codes for individual faces obtained from these datasets is 3, 16, and 4. The reason for this is that the bar code of the facial images is the same for the same faces, and

we found that the accuracy of the bar code is less than 60%. Therefore, none of these barcodes for face images are considered indexed.

Table 3.1: Shows the statistics for the number of faces found, keyframes created, and valid bar codes generated from a Hollywood movie dataset.

Video Name	Size(KB)	No. of Key frames	No. of faces detected	No. of valid barcodes
American Beauty	1420	2	1	1

- 00170				
As Good As It Gets – 01766	6150	18	5	5
Big Fish - 00664	1640	6	1	1
Butterfly Effect, The - 00696	1940	15	15	3
Casablanca– 00250	628	7	0	0
Crying Game, The- 01482	3210	18	1	1
Forrest Gump – 01277	22300	424	35	16
Gandhi-02262	18000	18	5	4

Table 3.2 displays the bar code results computed from the TV series video collection. Every video in this collection is in MP4 format. Here, each face detected from key frames is given a different barcode number. It will be decreased if the bar code number of the same image of the face is taken into account.

Table 3.2: The number of validated bar codes, extracted keyframes, and detected faces from a few TV series video dataset episodes.

Video name	Size(KB)	No. of Key Frames	No. of faces detected	No. of the valid barcode.
24_ep2	848000	4471	1970	1970
Breaking_bad_ep1	1130000	2473	567	567
How_I_Met_Your_Mother_ep 1	435000	553	532	532
Mad_men_ep1	980000	3092	1983	1983
Modern_Family_ep1	465000	2356	1415	1415
Sons_of_Anarchy_ep1	1110000	3964	1648	1648

For the YouTube face video dataset, the results of our effort are shown in Table 3.3.

Table 3.3: Keyframe no. statistics, faces no. Identified, and barcode produced from a few video files in the YouTube database

Video Name	Size(KB)	No. of Key Frame detected	No. of faces detected	No. of valid bar code generated
Aligned_video_0	556	84	84	84
Aligned_video_5	819	166	166	173
Aligned_video_3	1171	173	173	173
Aligned_video_1	1942	307	307	307
Aligned_video_2	3158	468	468	468
Aligned_video_4	664	119	119	119

3.3.3. The final result of our "A System and Method for Bar Code representation of face image" invention

We carried out several tests in this subsection to verify the accuracy of the bar code produced by our newly developed "A System and Method for Bar code representation of face image." The systematized tag in this invention is made using 70% and 75% of the top portion of the facial image, except the section that hacks down the nose/mouth assembly to test the scanner label concept using the LGFA and Windowing Technique.

Tables 3.4 and 3.5 present the comparative results of the windowing technique and its combination with LGFA. These tables demonstrate that the proposed method—a combination of the LGFA and windowing techniques—can yield a more stable barcode from illumination-invariant facial images.

Table 3.4: Bar code generation accuracy rate for images of faces with varying facial expressions across databases

Name of Dataset	Windowing Technique		Combining LGFA and windowing Technique	
	75% of Upper face image	70% of the upper face image	75% of the upper face image	70% of the upper face image

FACE94 dataset	78	80	80	85
FERET database	65	70	75	80
Face Dataset YaleB	10	15	70	75
NIR-VIS Cropped dataset	15	20	70	73
IFW dataset	20	25	70	72

Table 3.5: Verifying the generated bar codes' resilience to facial aging

Name of Dataset	**Window Technique**		**Combining LGFA and windowing Technique**	
	75% of Upper face image	**70% of Upper face image**	**75% of Upper face image**	**70% of Upper face image**
Aging dataset FG-NET	75	80	80	85
Aging dataset of composite faces.	80	83	81	85

3.3.4. Our "Video indexing through human face images using LGFA and window technique" method's results

In this section, we compare the performance metrics of our proposed "Video indexing through human face images using LGFA and window technique" with the current method to verify the accuracy of the bar code it generates. Using the recommended strategies, the following tables present the results of different key frame extraction, face detection, and valid barcode production on various datasets.

Table 3.6 shows the bar code produced from a subset of the Hollywood Movies Dataset. From this table, it is shown that the number of faces detected from the Butterfly Effect 00696, Forrest Gump-01277, and Gandhi-02262 datasets is 15, 35, and 5, respectively. However, the number of faces detected from these datasets is 3, 16, and 4, respectively. The number of valid bar codes for each face is created using the window technique, but no bar code is increased (8, 20, and 5) after combining the LGFA and window techniques. This can be explained by the fact that the bar code accuracy is less than 60%, and the bar code of the facial images is identical. Therefore, it is not considered that any of these bar codes for face images are indexed.

Table 3.6 presents the statistics for the number of Keyframes extracted, faces detected, and barcodes generated from several Hollywood movie datasets.

Video Name	Size(KB)	No. of Keyframes	No. of faces detected	No. of valid barcodes generated using the window technique	No. of valid barcodes generated using combining LGFA and window technique
American Beauty - 00170	1420	2	1	1	1
As Good As It Gets – 01766	6150	18	5	5	5
Big Fish - 00664	1640	6	1	1	1
Butterfly Effect, The - 00696	1940	15	15	3	8
Casablanca – 00250	628	7	0	0	0
Crying Game, The - 01482	3210	18	1	1	1
Forrest Gump – 01277	22300	424	35	16	20
Gandhi - 02262	18000	18	5	4	5

Table 3.7 displays the result of the bar code generated from the TV series' video dataset. Every video in this collection is accessible in MP4 format. Every face that has been identified from keyframes is not barcoded in this list. This will be reduced if the bar code number of the identical picture faces is taken into consideration.

Table 3.7: Number of valid bar codes, number of faces recognized, and number of detected key frames from a few TV series video dataset episodes

Video name	Size (KB)	No. of Key Frames	No. of faces detected	Number of valid barcodes using window Technique.	No. of valid barcodes using combining LGFA and window technique
24_ep2	848000	4471	1970	1970	1970
Breaking_bad_ep1	1130000	2473	567	567	567
How_I_Met_Your_Mother_ep1	435000	553	532	532	532
Mad_men_ep1	980000	3092	1983	1983	1983
Modern_Family_ep1	465000	2356	1415	1415	1415
Sons _of _Anarchy_ep1	1110000	3964	1648	1648	1648

Table 3.8 presents the outcome of the bar code created using the face video dataset from YouTube.

Table 3.8: Face number detected, barcode number generated, and keyframe no statistics produced from a few YouTube Face video database video files

Video Name	Size(KB)	No. of Key Frame detected	No. of faces detected	No. of valid barcode generated using Window Technique	No. of valid barcodes generated using combining LGFA and window technique

Aligned_video_0	556	84	84	84	84
Aligned_video_5	819	166	166	166	166
Aligned_video_3	1171	173	173	173	173
Aligned_video_1	1942	307	307	307	307
Aligned_video_2	3158	468	468	468	468
Aligned_video_4	664	119	119	119	119

3.3.4.1. Metrics for performance

Accuracy: The accuracy of face recognition is the total number of accurate predictions that are returned. The following equation, 3.4, is the equation of accuracy.

$$\text{Accuracy}=\frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ positive+False\ Negative} \quad 3.4$$

Precision: The ratio of the number of pertinent returned images to the total number of returned images is known as precision. The precision equation is 3.5, which is as follows:

$$\text{Precision}=\frac{TP}{TP+FP} \quad 3.5$$

The system allocates similar image pairs to the same cluster, denoted by TP, which stands for true positive pairs. FP indicates the false positive pairs. To put it another way, putting two distinct images in the same class.

Recall refers to the ratio of successfully clustered image pairs to the total number of images in the same cluster. This indicates that the percentage of the retrieved pertinent images among all returned ones is the recall measurement. The following is the Recall equation, which is 3.6

$$\text{Recall} = \frac{TP}{TP+FN} \quad 3.6$$

FN is the process of assigning two similar image pairings to distinct clusters. Referred to as false negative pairs.

F1-Score: An approach to integrating the model's precision and recall is the F1-score, which is the harmonic mean of the two metrics. The following equation (3.7) is used to determine the F1-Score.

$$\text{F1-Score}=2\frac{Precision\ X\ Recall}{Precision+Recall} \qquad 3.7$$

3.3.4.2. Comparison strategies

This section compares the performance of the proposed method with other technologies, including the Fuzzy-based Support Vector Machine classifier (FBSVM), the Video Retrieval Model on Convolutional Neural Network (VRMCNN), and the Cloud-based Face Video Retrieval System (CBFVRS). The resulting figure is displayed in Figure 3.20 for comparison with other approaches, such as FBSVM, VRMCNN, and CBFVRS, in terms of accuracy. Comparing the suggested framework to earlier methods, the figure shows that it achieved great accuracy.

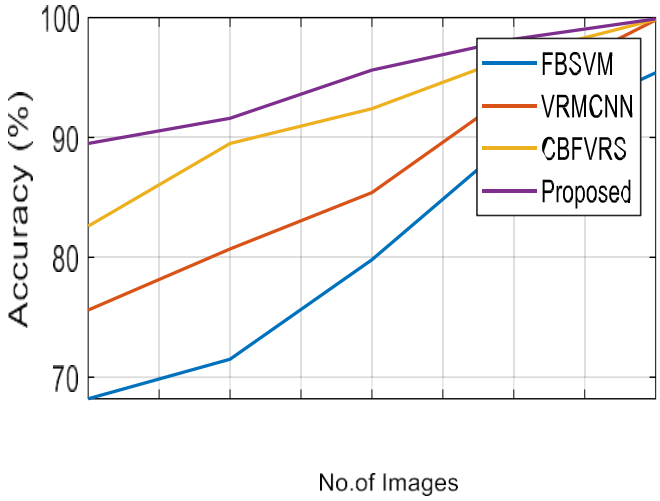


Figure 3.20: Accuracy Comparison

The resulting plot for the Precision comparison with earlier approaches, such as FBSVM, VRMCNN, and CBFVRS, is displayed in Figure 3.21. When compared to earlier methods, the figure shows that the suggested framework achieved great precision.

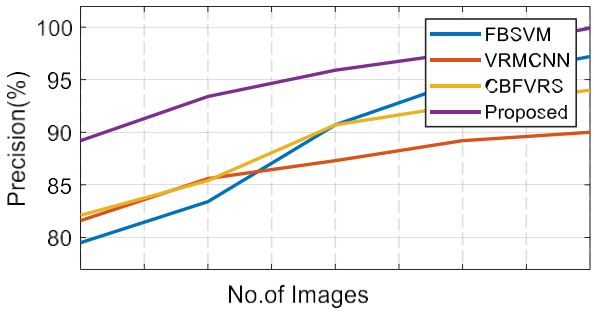


Fig 3.21: Precision Comparison

The resulting plot is displayed for comparison with previous approaches, such as FBSVM, VRMCNN, and CBFVRS, in Figure 3.22. The figure shows that, in contrast to earlier methods, the proposed framework achieved a significant improvement in recall.

The resulting plot for the F1-score comparison with earlier approaches, such as FBSVM, VRMCNN, and CBFVRS, is displayed in Figure 3.23. Comparing the suggested framework to earlier methods, the figure shows that it achieved a high F1 score.

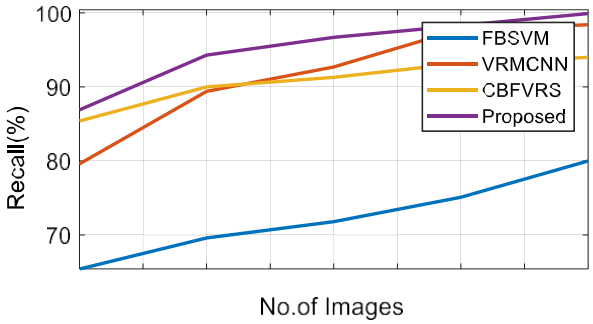


Fig 3.22: Recall Comparison

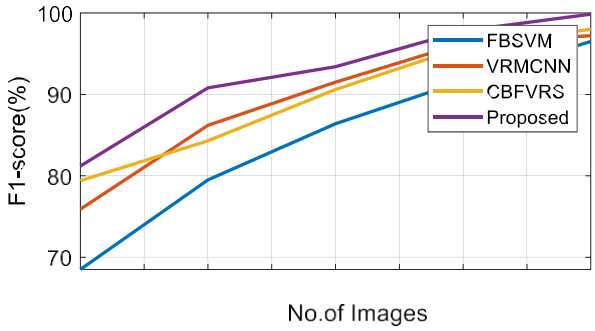


Fig 3.23 F1-Score Comparison

Table 3.9 presents a comparison of different approaches with different parameters. There are 200, 400, 600, 800, and 1000 images, among others.

Table 3.9: Presents a comparison of several approaches for various parameters.

Methodologies	Number of images	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FBSVM	200	68.2	79.5	65.4	68.5
	400	71.5	83.4	696	79.5
	600	79	90.7	71.8	86.4
	800	89.9	94.9	75.1	91.65
	1000	95.4	97.2	80	96.5
VRMCNN	200	75.6	81.6	79.6	75.9
	400	80.7	85.6	89.4	86.2
	600	85.4	87.3	92.7	91.5
	800	93.8	89.2	97.8	96.3
	1000	99.8	90	98.4	97.2
CBFVRS	200	82.6	82.1	85.4	79.4
	400	89.5	85.4	90	84.3
	600	92.4	90.7	91.3	90.6
	800	96.8	92.6	93.2	95.8
	1000	99.84	94	94	98
Proposed	200	89.5	89.2	86.9	81.2
	400	91.6	93.4	94.3	90.8
	600	95.62	95.9	96.7	93.4
	800	98.2	97.6	98.3	97.8
	1000	99.89	99.9	99.91	99.85

When compared to the previous approach, the proposed method achieves 99.89% accuracy, whereas FBSVM, VRMCNN, and CBFVRS attain 95.4%, 99.8%, and 99.84% accuracy,

respectively. Comparing the suggested method to the existing technique, FBSVM achieved a precision of 97.2%, VRMCNN achieved 90%, and CBFVRS achieved 94%. The proposed method achieved an accuracy of 99.9%. The recall achieved with the previous method was 80% for FBSVM, 98.4% for VRMCNN, 94% for CBFVRS, and 99.91% for the suggested method. The proposed technique achieved 99.85%, while the F1-Scores obtained using the known methods were 96.5% for FBSVM, 97.2% for VRMCNN, and 98% for CBFVRS. Therefore, in comparison to the prior method, the suggested method provides improved efficiency in video indexing and retrieval.

3.3.5. Our approach "Using Viola-Jones, MTCNN, DSFD, Blaze Face, and YOLOv3 algorithms, video indexing through the human faces represented as EAN-8 Linear bar code." The results of our approach

This section examines the accuracy of the linear bar code generated by the approach "Using Viola Jones,MTCNN,DSFD,Blaze Face and YOLOv3 algorithms, video indexing through the human faces represented as EAN-8 Linear bar code" and compares the face detection ratio. The following tables display the face detection results on several datasets using the Viola-Jones (Haar Cascade), MTCNN, DSFD, Blaze Face, and YOLO v3 algorithms. The number of frames and faces detected, the ratio of face detection (in faces per millisecond), the number of EAN-8 linear bar codes, and the time needed in milliseconds for several video clips from the Hollywood Data set (taking into account 10 Video Data set of Hollywood movie) are all displayed in Table 3.10.

The FDDB, LFW, and WIDER datasets' face detection times, numbers, ratios, and EAN-8 linear bar code numbers are all covered in Tables 3.11, 3.12, and 3.13.

Table 3.10: This displays the number of faces and frames found, the face detection ratio (in faces per millisecond), the number of EAN-8 linear bar codes, and the time needed in milliseconds for several video clips from the Hollywood Data set (taking into account 10 video clips of Hollywood movies).

Method Name	Frames Detected	Time Taken (Millisecond)	Faces Detected	No. of EAN-8 Linear Barcode	Ratio (Face /Millisecond)
Haar-Cascade	39883	3715.27	32452	32452	8.73
MTCNN	78717	2513.52	53478	53478	21.28
DSFD	114809	4937.64	97292	97292	19.70
Blaze face	97292	929.20	26980	97292	29.04
YOLOv3	50652	13247.74	39248	97292	2.96

Table 3.11: The number of face detections, face detection time (in milliseconds), number of linear EAN-8 bar codes, and face detection ratio (face per millisecond) on FDDB data sets are displayed in this table.

Method Name	Faces Detected	No. of EAN-8 Linear Barcode	Time Taken (Millisecond)	Ratio (Face/Millisecond)
Haar-Cascade	18697	18697	14061.44	1.33
MTCNN	19923	19923	8111.16	2.46
DSFD	20812	20812	14728.18	1.41
Blaze face	14632	14632	14835.40	0.99
YOLOv3	17736	17736	16226.31	1.09

Table 3.12 Presents the number of face detections on LFW data sets, together with the face detection ratio (face per millisecond), number of linear EAN-8 bar codes, and face detection time (in milliseconds).

Method Name	Faces Detected	No. of EAN-8 Linear Barcode	Time Taken (Millisecond)	Ratio(Face/Millisecond)

Haar-Cascade	14759	14759	1042.85	14.15
MTCNN	15573	15573	581.22	26.79
DSFD	16426	16426	3171.77	5.18
Blaze face	13798	13798	3292.38	4.19
YOLOv3	15069	15069	5647.88	2.66

Table 3.13: Illustrates the number of face detections on WIDER Face data sets, together with the face detection ratio (face per millisecond), number of linear EAN-8 bar codes, and face detection time (in milliseconds).

Method Name	Faces Detected	No. of EAN-8 Linear Barcode	Time Taken (Millisecond)	Ratio (Face/Millisecond)
Haar-Cascade	18697	18697	14061.44	1.33
MTCNN	19923	19923	8111.16	2.46
DSFD	20812	20812	14728.18	1.41
Blaze face	14632	14632	14835.40	0.97
YOLOv3	17736	17736	16226.31	1.09

3.3.5.1. Comparative strategies

The performance comparison of the proposed method with the algorithms YOLOv3, MTCNN, DSFD, Blaze Face, and Viola-Jones (Haar Cascade) is presented in this section. When the face detection process is complete, it is clear that MTCNN outperforms other machine learning and deep learning algorithms in terms of the face detection accuracy ratio from the input video, as covered in our study article. Comparing the outcomes, it is evident that DSFD and YOLO v3 identify more faces in the input video, while the MTCNN algorithm operates more quickly. DSFD and YOLOv3 can recognize a small face from the input. Blaze Face and YOLOv3 can recognize faces after using lightweight gadgets. MTCNN has a quicker face detection rate. As a result, the face detection ratio of the MTCNN algorithm is superior to that of the DSFD, Viola-Jones, Blaze-Face, and YOLOv3 methods.

Figure 3.24 displays a bar graph that highlights the key distinctions between the DSFD, YOLO v3, MTCNN, DSFD, and Haar-cascade algorithms. The LFW face dataset, FDDB face dataset, WIDER face dataset, and Hollywood movie video dataset are used in this bar graph to assess the performance of the aforementioned methods. Following the tests, the graph illustrates the number of faces discovered using the previously mentioned algorithm.

The bar graph in Figure 3.25 compares the performance of the aforementioned algorithms using the Hollywood movie video dataset, the FDDB face dataset, the WIDER face dataset, and the LFW face dataset. After the testing, the graph displays the amount of time required to use the aforementioned technique to determine the number of faces.

The MTCNN algorithm is more accurate and less prone to errors in face detection compared to the Viola-Jones, DSFD, Blaze Face, and YOLOv3 algorithms. Additionally, MTCNN performs better at identifying a wide variety of faces.

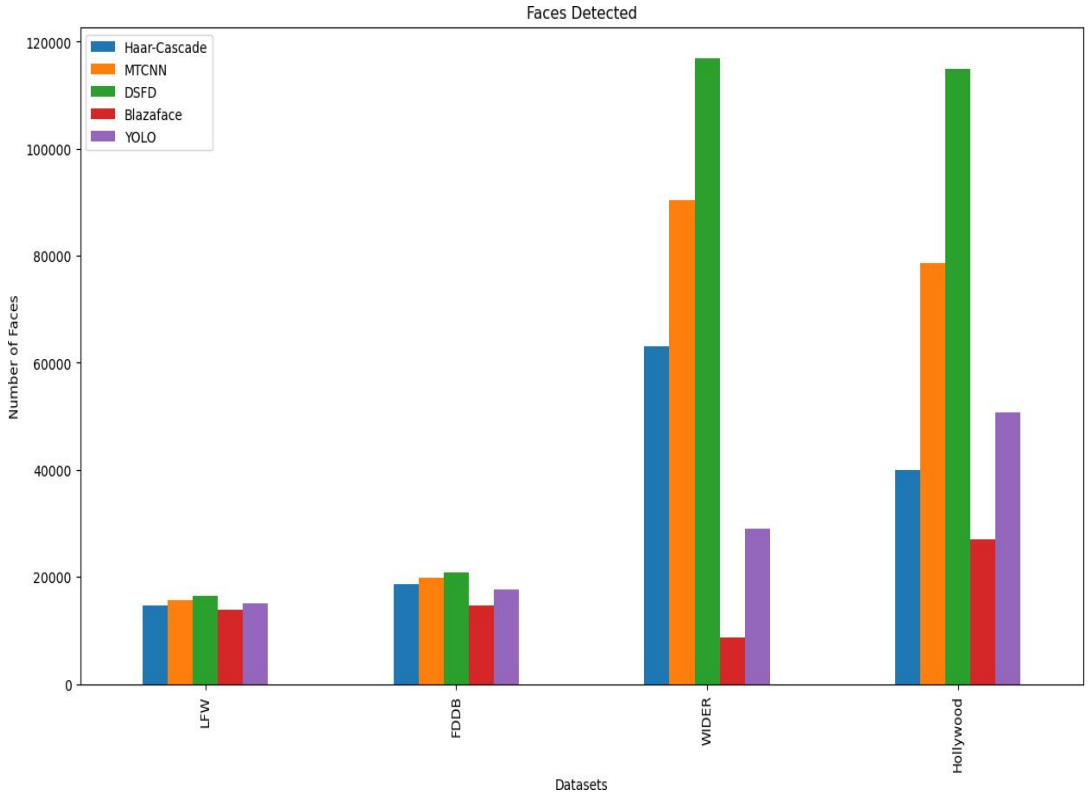


Fig 3.24: Using the Haar-cascade, MTCNN, DSFD, Blaze-Face, and YOLO v3 algorithms, the number of faces in the LFW face dataset, FDDB face dataset, WIDER face dataset, and Hollywood movie video datasets is compared.

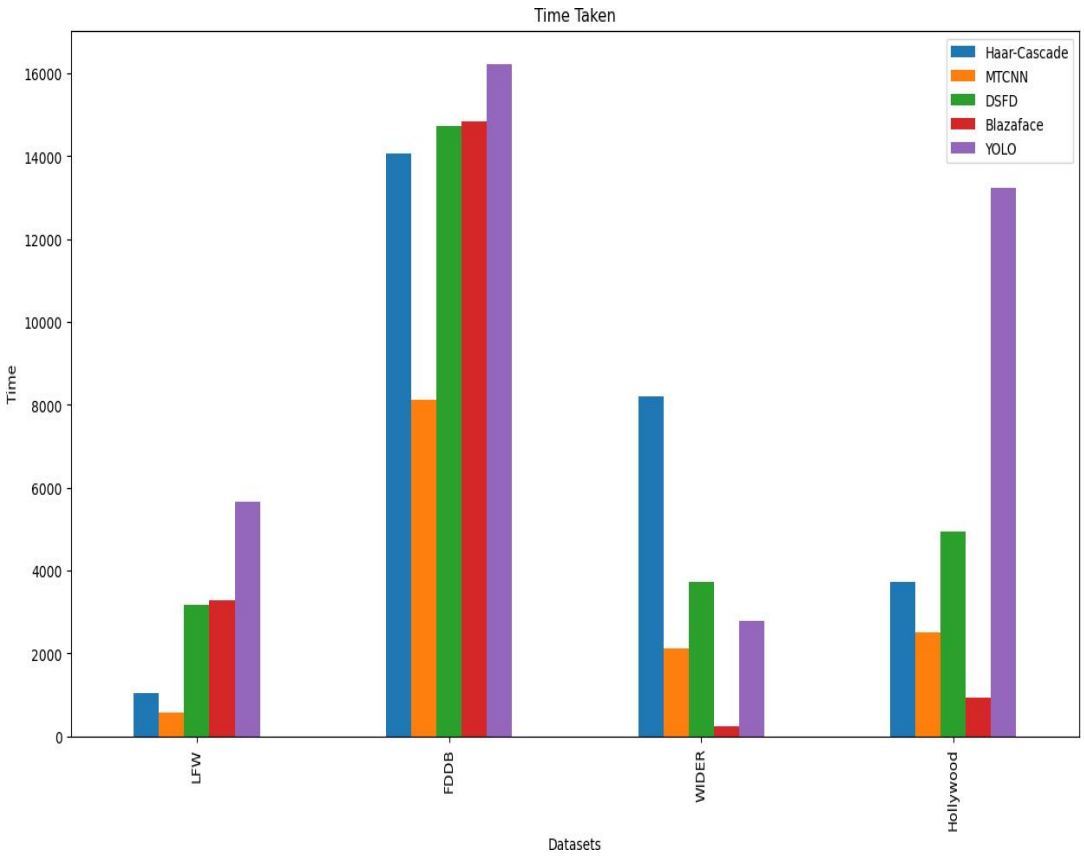


Fig 3.25: The LFW face dataset, FDDB face dataset, WIDER face dataset, and Hollywood movie video datasets were compared based on the time required for face detection utilizing the Haar-cascade, MTCNN, DSFD, Blaze-Face, and YOLO v3 algorithms.

There is a discernible drop in frame rate when comparing MTCNN to YOLOv3, DSFD, Blaze Face, and Viola Jones. A linear EAN-8 bar code is generated from each distinct face image for indexing purposes after facial recognition utilizing all of the machine learning and deep learning algorithms covered in our study. The fact that human-readable linear EAN-8 barcodes consume less storage space and index more quickly is one advantage of adopting them. The human face can also be scanned after scanning this Linear EAN-8 bar code with a barcode reader.

3.3.6. Failure cases

When the faces in the input video were more angular and illumination-invariant, the suggested video indexing through face images utilizing barcodes failed. Because of this, creating a consistent barcode from these types of input videos is challenging. We noted this in Table 3.1, which shows that for the Casablanca–00250 Hollywood data, zero faces are detected, and zero barcodes are generated for the outcome of our recommended method, as covered in Section 3.3.2. The accuracy of bar code generation from the Yale B, NIR-VIS Cropped, and IFW

datasets is 10% for the upper 75% of face images and 15% for the upper 75% of face images using window technique, 15% for the upper 75% of face images and 20% for the upper 75% of face images using window technique, 20% for the upper 75% of face images and 25% for the upper 75% of face images using window technique, according to table 3.4 of our results and discussion section 3.3.3. Face images that are illumination invariant are the cause of this.

Stable bar code creation varies with changes in human age. In the FG-NET aging dataset, we noticed this. Frontal faces can be detected using the Viola-Jones algorithm. Therefore, it is not possible to use the Viola-Jones algorithm to detect an angular face image from an input video. Detecting a small face from input footage is another challenge.

3.4. Conclusion

This proposed approach to video indexing using face recognition employs an EAN 8 linear bar code for face representation, as identified from the video, which is discussed in Section 3.2.1. This method utilizes the Viola-Jones object detector as a face detector and the color histogram method for keyframe detection. The face is indexed as a bar code using the EAN 8 bar code, and the image gradient is calculated using the sliding window approach. When creating an EAN 8 barcode from human faces in videos, this method also takes into account occlusion, lighting variance, facial expressions, and slight changes in face direction.

Stable barcode generation will encounter issues with illumination, invariance face images, and notable variations in face image postures. It is a time and space-efficient approach. The test is conducted on the TV series video set, YouTube face video data collection, and Hollywood

video dataset. The primary challenge with this technology is generating linear barcodes from angular and illumination-invariant face images. These problems will be addressed in the next assessment.

The combined window and LGFA approach, gradient calculations, gradient directions, normalization, and quantization form the foundations of the proposed process for generating linear EAN-8 barcodes from face images, as outlined in Section 3.2.2. Finally, EAN-8 linear standardized identifications are used to convert input faces to tags. The test is conducted using the following datasets: Face94, YaleB Face dataset, FERET dataset, FG-NET Ageing dataset, a composite face database (comprising a total of 20 faces), NIR-VIS cropped dataset, and LWF dataset. The test performed as anticipated, demonstrating that the computation provided validates the concept of standardized names by a slight reflection of the main image when the plane, point, and apparent look change.

Deep learning and image processing are utilized in video indexing and retrieval to reduce the time and space complexity of obtaining angular faces from videos and increase storage capacity for storing the keyframes from the video. Therefore, to address these problems, a new hybrid sliding window and LGFA technique, the Viola-Jones Algorithm for use as a face detector, and the EAN 8 bar code for use as a barcode to face index—all of which are covered in Section 3.2.3—have been introduced. Furthermore, this form is used to improvise the illumination case of the invariant facial image, which is computationally simple. Additionally, the suggested approach reduced the complexity of time and space while increasing storage capacity. As a result, the proposed approach effectively improves video indexing and retrieval methods. The Hollywood video dataset, YouTube video dataset, and TV show video dataset were all used in the test. The recommended approach needs to be enhanced for small faces and angular features.

To improve storage capacity for important frames from movies and reduce the time and space required to acquire angular faces from videos, deep learning, and machine learning are utilized in image processing for video indexing and retrieval. The linear EAN-8 bar code is used for face indexing, as described in Section 3.2.4, and the MTCNN Algorithm has been

proposed for use as a face detector to address these issues. Additionally, this format is used to create the illusion of illumination on the invariant facial picture, and the calculation is straightforward. Lastly, EAN-8 linear barcodes are used to indicate input faces. The suggested

method increased storage capacity while reducing time and space complexity. This effectively improves the video indexing and retrieval techniques in the recommended way. The test utilized the Hollywood video, FDDB, LFW, and WIDER face datasets. Authentication, affirmation, and individual search can be defined for an account using the video indexing technique. In the future, angular images of faces will be used to generate a linear barcode.
This chapter explores various methods for utilizing a person's face as a cue for video indexing. Nevertheless, due to factors such as the face's orientation shifts, brightness, and illumination, none of the algorithms are effective at identifying faces in videos. Furthermore, if an effective

detection system exists, it lacks a concise representation of faces. A linear bar code that loses much information when scanning face images horizontally is a compact representation. Additionally, the accuracy of facial detection affects the system's overall accuracy. All of these chapter 3 problems are resolved in the following chapter, chapter 4. In this chapter, face images are represented by a QR bar code.

Chapter 4

Video indexing through the Face Images using a QR code

4.1. Introduction

As we covered in Chapter 1, a video index that explains the video content is necessary for viewing, searching, and working with video documents. Chapter 3 introduced an automated system for video indexing utilizing barcodes and face images. All of the techniques covered in Chapter 3 remain ineffective at identifying faces in videos due to factors such as the face's directional shifts, brightness, and lighting. Furthermore, if an effective detection mechanism exists, there isn't a compact representation of faces. When face images are scanned horizontally, a linear bar code represents a concise representation with significant information loss. Additionally, the total accuracy of the system is affected by the accuracy of face detection. To overcome this, we will introduce an automated framework in this chapter for video indexing using face images and QR codes.

Face recognition and facial expression analysis are only two of the many face applications that depend on precise face detection and alignment. However, the numerous visual variations of faces, such as occlusions, fluctuations, and harsh lighting, in real-world applications present challenges and variations for these tasks.AdaBoost and Haar-Like features are used to train cascaded classifiers that exhibit good real-time performance in the cascade face detector created by Viola and Jones [87]. According to several articles [88,89,91], this detector may degrade significantly in real-world applications where human faces exhibit greater visual variability, even with more advanced features and classifiers. In addition to the cascade structure, deformable part models (DPMs) are introduced for face detection [90, 92, 93], achieving exceptional performance. Nevertheless, they are computationally expensive and

often require costly annotation during training. Convolutional neural networks (CNNs) have recently demonstrated impressive advancements in various computer vision applications, including face recognition and image classification [94, 163].

Since CNNs have been highly successful in computer vision tasks, several CNN-based face identification techniques have been introduced in recent years. To recognize facial attributes, Yang et al. [96] develop deep learning neural networks with strong responsiveness in face areas, which then generate candidate windows of faces. The complex CNN structure of this approach, however, makes it time-consuming to practice. Li et al. [97] employ cascaded CNNs for face detection; however, this approach requires bounding box calibration from face detection, incurring extra computational expense, and overlooks the inherent connection between facial landmark localization and bounding box regression. Face alignment is also quite important.

One frequent type is regression-based approaches [166–168], and another is template-fitting methods [90, 164, 165]. Recent work by Zhang et al. [98] suggested using facial attribute identification as a supplementary task to enhance face alignment performance using deep convolutional neural networks. The majority of face alignment and identification methods often overlook the inherent connection between these two objectives. Even when several works attempt to tackle them together, they remain limited in their scope. Hen et al. [94], for example, utilize pixel value difference features to perform alignment and detection using random forests jointly.

Nevertheless, a subpar face detector's initial detection windows limit the accuracy. However, to increase the detector's strength, vigorous sample mining is necessary during training. On the other hand, traditional hard sample mining typically occurs offline, resulting in a considerable increase in manual operations. It is desirable to design an online hard sample mining method for face alignment and detection that automatically adjusts to the current training procedure.

In Paper [15], it is suggested that these two tasks be integrated using multi-task learning and unified cascaded convolutional neural networks (CNNs). There are three stages to the proposed CNNs.It utilizes a shallow CNN to rapidly generate candidate windows in the first stage. A more sophisticated CNN then refines the windows to reject a significant portion of

non-face windows. Finally, it refines the results and outputs the facial landmark positions using a more powerful convolutional neural network (CNN).

"Video Indexing through human face" is covered in Paper [130]. The Viola-Jones technique, which utilizes AdaBoost and Haar-like features, is employed in Papers [130, 149] to detect faces. All individual faces are represented as linear EAN-8 bar codes for indexing purposes when the key frame (human face) from the video has been identified. To generate linear barcodes, however, not all information is included because the linear barcode input image is scanned from left to right horizontally. Additionally, the person's barcode cannot be recognized if any part of the linear barcode is damaged after scanning. The inability of linear bar codes to scan the input image vertically is another issue.

This chapter discusses a novel method called "Video indexing through the Face Images using QR code" to address all of these issues. This research addresses important issues, including changing posture, illumination-invariant features, facial angular variations, storage space limitations, the inability to store important video frames, and time and spatial complexity. To solve these problems, the research's primary contribution is "Video indexing through the human face as a QR code utilizing the MTCNN algorithm."

- The Viola-Jones and MTCNN algorithms were used to construct cropped facial portraits, enabling the extraction of keyframes and overcoming keyframe storage issues.
- Following this comparison, the Viola-Jones method and MTCNN are employed for face detection and keyframe extraction.
- A QR code will be generated, and each individual will receive a unique QR code that can be used to access a database. Not all QR codes are compatible with human faces; thus, it's important to use the correct one. Additionally, the correct QR code size must be selected; otherwise, problems may occur.

The recommended approach to video indexing thus tackles issues with changing posture, storage capacity constraints, storing video keyframes, and time and spatial complexity.

The chapter's remaining sections are arranged as follows. Section 4.2 provides a thorough explanation of the suggested framework for video indexing using QR code-based facial images. In-depth experimental analyses, discussions, and cases of failure are included in

Section 4.3. Section 4.4 provides concluding remarks and discusses the future direction of the work.

4.2. Proposed methodology for the video indexing through the face images using QR code

According to a survey and analysis of numerous publications in Chapter 2, it is challenging to recognize people in videos that are indexed using low-level attributes. In response to problems with video indexing and retrieval, Paper [149] proposes a novel technique called Video Indexing utilizing Human Face Images. The aforementioned problems are resolved by using the Sliding Window Technique and LGFA.In this research, the EAN-8 linear bar code is used to construct the bar code, and the Viola-Jones algorithm is used to detect the Key Frame (Human face). The issue with the Viola-Jones algorithm is that, despite being quicker than MTCNN, it is unable to identify angular faces in videos. The linear form of the EAN-8 bar code representation of the face image is another issue in the study [149]. Not all facial features can be extracted during barcode synthesis, as the face is scanned horizontally rather than vertically in linear barcodes. Because of this, the barcode's accuracy is decreased, and it won't work if any part of the EAN-8 barcode is broken. Additionally, it noted that intrinsic ambiguities, including position shifts, partial occlusion of facial cavities, and susceptibility to low-resolution impacts, affect video-based recognition.

To address all of these problems, this section suggests a revolutionary method (video indexing through the human face as a QR code using the MTCNN algorithm). This approach utilizes the MTCNN algorithm to detect faces in the input video and indexes the human face using a QR code, which can be found in various video formats rather than a face image. A QR code's advantage is that it will still work properly even if a portion of it is broken. After scanning the input image (a face) both vertically and horizontally, a QR code is created. In this approach, the MTCNN algorithm's primary task is to identify and align faces in angular keyframes.

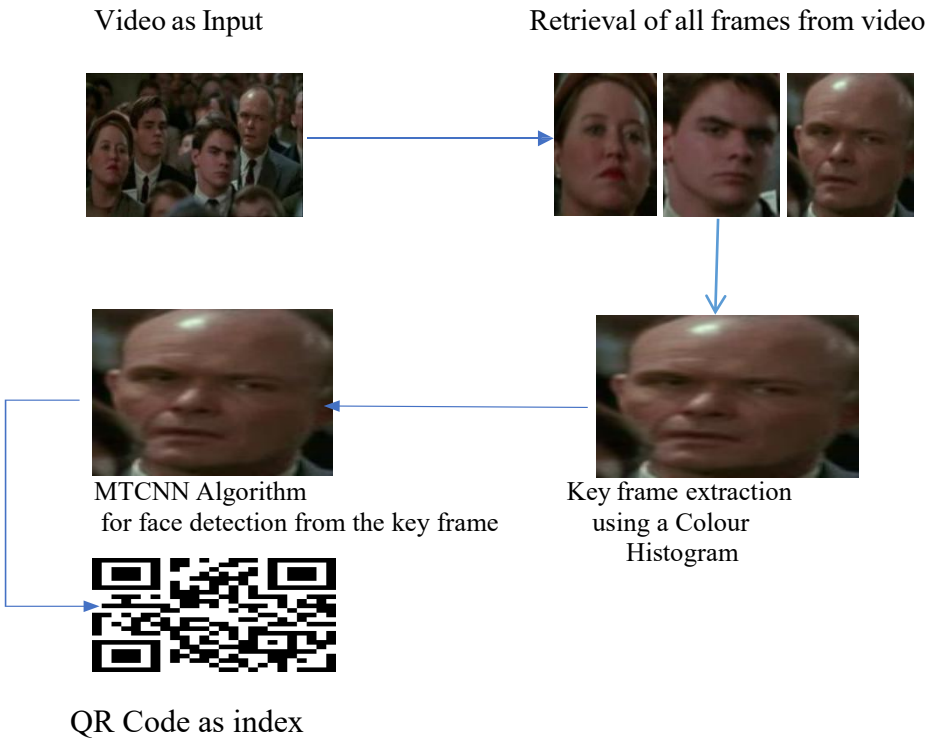


Figure 4.1: Proposed System Block Diagram

Each of the various phases in the approach suggested in this study is depicted by a block diagram in Figure 4.1. (a) In the input video, frame extraction is the first stage. (b) The keyframes are extracted from the input video frames using variations in the color histogram. (c) The MTCNN algorithm is used to detect faces from the keyframe. (d) A QR code is generated using the detected faces of the keyframe.

4.2.1. Frame extraction from input video

The first stage is to extract the still images from the input videos, which are represented as scenes, shots, and frames since these elements combined form a dynamic video. A scene is a collection of shots, and a shot is a collection of frames. Conventional video has a frame rate of

20 to 30 frames per second, conveying a significant amount of information. There are unnecessary details in the frame, which is a still image that is a portion of a video. The frame from the Hollywood movie Dead Poets Society is shown in Figure 4.2.

Figure 4.2: The Dead Poet Society Video's frame

4.2.2. Extracting the Key Frame

The frame that best captures each image's primary features is called the keyframe. Keyframes, according to this work, are human faces with characteristic expressions, poses, lighting, and illumination. In this phase, however, the Colour Histogram method is used to extract the key frame from each frame of a specific video. Keyframes can be extracted from the frames using the Colour Histogram difference if the observed difference is greater than the threshold. The frame is selected as the next keyframe when the threshold for color histogram disagreement is met. Figure 4.3 displays a few key frames from the Holly Wood-starring Hollywood movie Dead Poets Society.

Figure 4.3: The Key Frame of the Society of Dead Poets video clip from the Hollywood movie Dataset (based on the face as the major information)

The algorithm and formula for splitting the color histogram's two consecutive frames are discussed in a study [149].

4.2.3. The MTCNN algorithm was used to crop the face images from the keyframes

Faces are recovered from the obtained key frames using the MTCNN object detection technique. The MTCNN algorithm consists of three steps. It utilizes a shallow CNN to rapidly generate candidate windows in the first stage. Finally, it refines the output and facial landmark placements using a more powerful convolutional neural network (CNN).

Stage 1 (P-Net): In this stage, a fully convolutional network known as the Proposal Network (P-Net) is used to generate candidate windows and their corresponding bounding box regression vectors. The candidates are calibrated using the calculated bounding box regression vectors. Then, significantly overlapping candidates are merged using non-maximum suppression (NMS).

Stage 2 (R-Net): Every candidate is sent to the Refine Network (R-Net), a separate CNN that performs non-maximum suppression (NMS) candidate merge, calibration using bounding box regression, and rejects a significant portion of incorrect candidates.

Stage 3(O-Net): The primary goal of stage 3 (O-Net) is to give a more thorough description of the face, even if it is comparable to stage 2. The network will explicitly output five face marker sites.

The following is the MTCNN algorithm used to crop the face portraits from the keyframes in algorithm 4.1:

Algorithm 4.1: The MTCNN algorithm was used to crop the facial portraits from the keyframes.

Input: Key frames are input.
Output: The face detected
Steps:

1. Start
2. Before processing, the input image is resized and converted to greyscale.

3. Three deep neural networks are defined based on the three steps of the MTCNN algorithm.

4. The first neural network constructs candidate bounding boxes for the faces in the image by applying a set of convolutional filters at different scales. Weak bounding boxes should be eliminated by non-maximum suppression.

5. The second neural network enhances the candidate bounding boxes produced in Step 4 by applying a bounding box regression technique. The coordinates of prominent facial features, like the corners of the mouth, nose, and eyes, are also predicted by the second network.

6. The third neural network, which predicts face landmark points more accurately, is used to enhance the candidate bounding boxes further.

7. Provide the last set of bounding boxes and facial landmark points for the identified faces.

8. A post-processing step is used to eliminate overlapping bounding boxes and false positives.

9. The detected faces are provided along with their bounding boxes and corresponding facial landmarks.

10. Stop

While slower than the Viola-Jones approach, the MTCNN algorithm has the advantage of detecting more faces and detecting angular faces in input videos. Screenshots of the face detection feature of the Keyframes are shown in Figure 4.4.

Figure 4.4: Faces cut from the keyframes of the Hollywood movie Dead Poets Society

4.2.4. Face detection from a keyframe generates a QR code, which is then utilized as an index

A QR code is generated from a single face detected in the key frame of the input video and saved as an index. A QR code is a square grid of white and black squares on a white

background. The information included in the square pattern that constitutes the QR code can be decoded using a QR code reader or the camera on a smartphone. The information can be in

the form of text, contact details, a website URL, or any other type of data. In this chapter, the information encoded in the QR bar code can be derived from a specific human face in the input video.

Figure 4.5 displays the American Beauty-00222 video clip with a cropped face and the corresponding QR code.

This section covers the procedure for creating QR codes with recognized faces. As per our technology, the face image extracted from the video is represented by a two-dimensional QR code. As a two-dimensional representation of face images, these QR codes can be used to index human faces found in various videos and remain functional even if a section is broken. One benefit of indexing with human-readable QR codes is that they require less storage space and index more quickly.

The QR barcode-generation approach described in the work [169] is also described in this study, along with a QR code-generating algorithm. Figure 4.5 displays examples of various bar codes found in different Hollywood (American Beauty) facial video datasets. Recognized images of faces are converted into QR codes using the following algorithms. The following algorithm 4.2 is used to generate QR codes from images of faces taken from input videos:

Algorithm 4.2: QR code generation algorithm using the detected face

Input: Faces detected from key frames of input video

Output: QR code of the detected A face image from the keyframe

Steps:

1. Start
2. To ascertain the range of characters that require encoding, the input data stream of the facial image should be analyzed.
3. The input data is converted into a bit stream with one or more segments per segment in a separate mode. It is necessary to separate the resultant bit stream into 8-bit code words. To meet the version's data code word count, add padding characters as needed.
4. Split the code word sequence into the required number of blocks so that the error correction process can be processed. Add the error repair code words to the end of the data code words list for every block. Reed-Solomon error control coding is used in QR codes to identify and correct errors.
5. Any extra bits should be added as needed, and the data and error-correction code words should be inserted in between each block.
6. Add the finder pattern, separators, timing pattern, and alignment patterns (if needed) to the matrix along with the code word modules.
7. Each data masking pattern should be applied to the encoding region of the symbol. Examine the data and select the design that minimizes the occurrence of undesirable patterns while achieving the optimal balance between the dark and light modules.
8. Complete the QR bar code symbol and, if required, create the format and version information.
9. Stop

4.3. Experimental Result and Discussion

This section assesses the performance of video indexing utilizing QR code-based facial images. The datasets used in this study are explained in detail in subsection -4.3.1. In subsection 4.3.2, we will discuss the results of our proposed method, "Video indexing through QR code of human faces using MTCNN algorithm."

4.3.1. Dataset Description

The keyframe for this study was first extracted from the video collection based on human faces, and the QR code was generated from this cropped face. The human face QR code

visible in the video is then used for indexing. However, some video files (such as the Face video dataset for FDDB, WIDER, and LFW) expressly contain the face image keyframe.

Consequently, when generating the QR code from this dataset, the keyframe is preserved. Four different video datasets were used to validate the approach.

The Hollywood video dataset [152] is used to start the research. This also contains snippets of videos from thirty-two human action films. It is necessary to label the sample using one or more of the eight groups that are accessible. The 20-film data set is created by combining two 12-film practice sets from the test set. Automatic script-based action labeling produced accurate labels for about 60% of the 233 video recordings in the automated learning set. The Hollywood results clean training collection comprises 211 video samples with manually tested labels and 219 video samples with manually checked labels.

Collected from the Faces in the Wild collection and the Face Detection collection and Benchmark (FDDB) dataset [162], this dataset consists of designated faces. The images range in size from 229x410 to 363x450, and there are 5171 facial annotations. Among the dataset's issues include low resolution, faces that are out of focus, and troublesome stance angles. There are images in both color and greyscale.

WIDER FACE [161] is a benchmark dataset for detecting faces derived from the publicly available WIDER dataset. The sample images demonstrate the broad range of scale, location, and occlusion variance in the data collection, which comprises 32,203 images and labels 393,703 faces. The 61 event classes are used to organize the WIDER FACE dataset. 40%, 10%, and 50% of the data are randomly selected for the training, validation, and testing sets, respectively, for each event class. The assessment metric from the PASCAL VOC dataset is used in the WIDER FACE data collection, just like it is in the Caltech and MALF datasets.

Finally, LFW [159] data sets are collected for the experiments. To investigate the problem of unrestricted face identification, the Labelled Faces in the Wild (LFW) database was established. Researchers at the University of Massachusetts, Amherst, established and maintained this database (particular references are cited in the Acknowledgements section). Out of 13,233 images downloaded from the internet, 5,749 people were recognized by the Viola-Jones face detector. 1,680 of the individuals depicted in the collection have two or more

distinct images. The original database contains three types of "aligned" images and four sets of LFW images.

4.3.2. The results of our "Video indexing through QR code of human faces using MTCNN algorithm" technique

To verify the accuracy of the QR code produced by our proposed "Video indexing through QR code of human faces using MTCNN algorithm," we carried out several tests in the following section. It is clear from the experiment that MTCNN outperforms Haar Cascade as an algorithm. The average accuracy of the MTCNN algorithm is 89.6% when in use, while the

average accuracy of the Haar Cascade technique is 59.2%. As it revealed no false positive detections during the testing phase, MTCNN is also less likely to display any false positives. One drawback of the Haar Cascade approach is that it showed six false positive detections at once. However, the average frame rate of the video was 17 fps, whereas MTCNN only managed 8 fps; therefore, the Haar Cascade was a much faster method.

The results of face detection using the MTCNN and Viola-Jones methods on various datasets are presented in the following tables. Table 4.1 presents the time required, the number of frames detected, and the number of false positive detections for various video clips from the Hollywood dataset. Tables 4.2, 4.3, and 4.4 discuss the face detection ratio, number of faces detected, and face detection time on the FDDB, LFW, and WIDER data sets, respectively.

Table 4.1 shows the time, the number of detected frames, and the number of false positive detections for different video clips in the Hollywood Data set.

Name of Data set	Time Taken (MTCNN) (Sec.)	Time Taken (Haar Cascade) (Sec.)	No. of Frames (Haar Cascade)	No.of Frames (MTCNN)	No.of False Positives (Haar Cascade)	No.of False Positives (MTCNN)
American Beauty-1	42	7.45	24	51	6	0
American Beauty-2	46	7.96	26	53	7	1

Being John Malkovich1	47	8.01	27	59	4	0
Being John Malkovich2	51	7.67	31	61	3	0
Big Fish-1	53	7.79	23	57	7	0
Big Fish -2	49	8.05	25	53	6	0
Big Fish-3	42	7.61	26	54	5	1
As Good as It Gets	57	7.74	26	52	4	0
Casablanca 1	47	8.16	24	56	7	1
Casablanca 1	49	7.49	27	59	6	1

Table 4.2: Displays the number of faces detected, the face detection ratio, and the time required for face detection using FDDB data sets.

Method Name	**Number of Faces Detected**	**Time Taken for face Detection(Second)**	**Face detection Ratio(Frame/Second)**
Viola Jones	18697	14061.44	1.33
MTCNN	19923	8111.16	2.46

Table 4.3: Represents the face detection ratio, number of faces detected, and face detection time using LFW data sets.

Method Name	**Number of Faces Detected**	**Time Taken for face Detection(Second)**	**Face detection Ratio(Frame/Second)**
Viola Jones	14759	1042.85	14.15
MTCNN	15573	581.22	26.79

Table 4.4 illustrates the face detection ratio, number of faces detected, and face detection time on the WIDER datasets.

Method Name	Number of Faces Detected	Time Taken for face detection(Second)	Face detection Ratio(Frame/Second)
Viola Jones	63134	8213.76	7.68

4.3.2.1. Comparison strategies

A performance comparison between Viola Jones' strategy and the suggested methodology is shown in this section. It is evident from the completion of the face detection procedure that MTCNN is a more accurate algorithm than the Viola-Jones algorithm. The average accuracy of the Viola-Jones algorithm is 59.2%, while the average accuracy of the MTCNN algorithm, when applied, is 89.6%. It is also less likely to do so because MTCNN found no false positives throughout the testing phase. Six false positive detections were also found by the Viola-Jones algorithm, which is concerning. Overall, Viola Jones is a faster algorithm than MTCNN, as the video played at an average of 17 frames per second, compared to 8 frames per second.

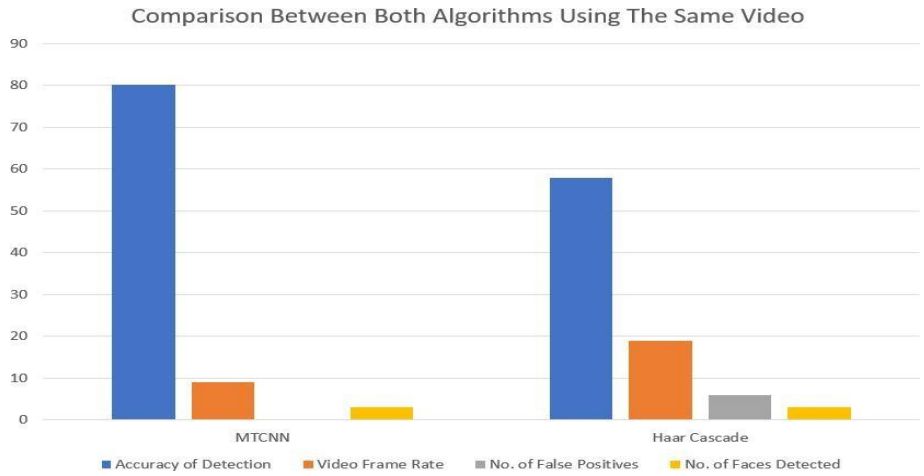


Figure 4.6: Bar graph comparison between MTCNN and the Viola-Jones algorithm

Figure 4.6 illustrates the main differences between the two algorithms using a bar graph. In this bar graph, the performances of the two algorithms are contrasted using the same scene from "The American Beauty." The graph displays the average accuracy, video frame rate, false positives, and faces discovered after the experiments.

Figure 4.7 shows a bar graph with ten Hollywood movie trailers. This bar graph displays the time, the number of frames found, and the face detection ratio for the input 10 video clips of the highlighted Hollywood movies. Compared to Viola Jones, the MTCNN algorithm is less likely to provide false positive results and is more accurate. MTCNN also performs better in recognizing a wider range of faces.

While the Viola-Jones method is substantially faster, the Frame Rate in MTCNN is noticeably lower than in Viola-Jones. As a result, MTCNN can be considered superior even though it requires more time than the alternative method. Thus, it may be claimed that MTCNN is superior even if it requires more time than the alternative approach. Because it can scan the input image both horizontally and vertically, QR codes are employed. While the linear bar code (EAN-8) cannot recognize a human face, the QR bar code may be readable if any part of it is intact. One can search for a person's face after using a QR bar code reader to scan this code.

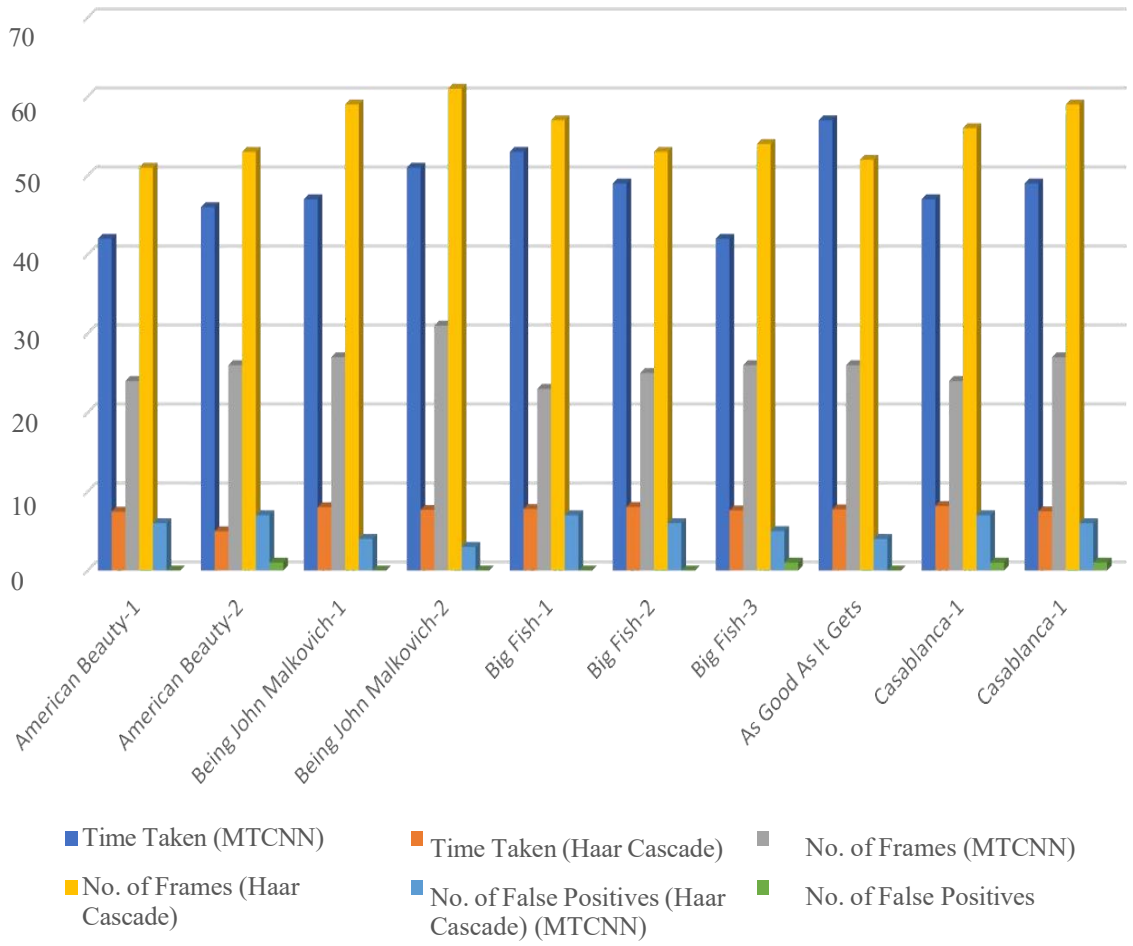


Figure 4.7: Compares ten video clips from a Hollywood movie data set based on the MTCNN and Viola Jones algorithms for face detection time, face detection number, and false positive detection number.

4.3.3. Failure Cases

The proposed video indexing using face images and QR codes did not yield 100% accurate results when the faces in the input video were more angular and invariant to lighting. The MTCNN algorithm is often slower because the video was played back at an average speed of 8 frames per second. MTCNN takes more processing time and computing power. Face detection, however, takes longer with lightweight real-time devices that utilize the Viola-Jones and MTCNN methods. Furthermore, it has been observed that Viola Jones and MTCNN

are not good at identifying small faces in videos. Scanning QR codes on damaged surfaces or in poor light can be challenging.

4.4. Conclusion

The idea presented in this chapter is to use a video's keyframes to create QR codes of facial images. Deep learning reduces the time and space complexity of obtaining angular faces from videos and increases storage capacity for key frames from videos in image processing for video indexing and retrieval. The MTCNN Algorithm has been proposed as a face detector to address these issues, and the QR bar code is utilized as a barcode for face indexing. Furthermore, the case of lighting of the invariant facial picture is created using this form, and it is computationally straightforward. The proposed method decreased time and space complexity while simultaneously increasing storage capacity.

A QR code's advantage is that it will still work properly even if a portion of it is broken. After scanning the input image (a face) both vertically and horizontally, a QR code is generated. In this approach, the MTCNN algorithm's primary task is to identify and align faces in angular keyframes. As a result, the recommended approach successfully enhances video indexing and retrieval.

The Hollywood video, FDDB, LFW, and WIDER face datasets were used in the test. An account's personal search, authentication, and verification can be facilitated using video indexing technology. Security, human activity detection, video surveillance, communication channel description, and other applications benefit from this technology.

This chapter covers the techniques for using a person's face as a cue for video indexing with a QR code. However, when the faces in the input video were more angular and lighting invariant, the suggested video indexing through the face photos using a QR code did not yield a 100% accurate result. Since the average framerate of the video was 8 frames per second, the MTCNN algorithm is frequently slower. MTCNN requires more processing power and time. However, lightweight real-time systems that employ the Viola-Jones and MTCNN techniques require more time to detect faces.

Additionally, it has been noted that MTCNN and Viola Jones struggle to recognize small faces in videos. It can be difficult to scan QR codes in low light or on damaged surfaces. More complex facial images will be used in future QR code generation. The next chapter, Chapter 5, addresses all the issues from Chapter 4. This chapter discusses the use of facial images for video indexing.

Chapter 5

Video indexing through the Face Images using Deep Learning Models

5.1. Introduction

As we covered in Chapter 1, the growth of digital technology, web streaming, and social networking has enabled more people to modify video objects and utilize them for a greater variety of purposes. Since camera expressions are so prevalent in daily life, experts are particularly interested in them. As a result, the human face is currently regarded as an essential component for video indexing. In the previous two chapters, Chapters 3 and 4, we covered video indexing using face images through a linear barcode and a QR code, respectively. This chapter discusses the use of deep learning models for video indexing using face images.

Biometrics is a very complex and fascinating field of research. By employing sophisticated mathematical techniques, it is possible to differentiate between individuals, making the use of biometrics necessary in highly diversified settings. This diversity is also reflected in the vast array of facial recognition algorithms that have been created. A person's expressions, feelings, and unique facial features are all part of their complex, multivariate human face, which can provide important information about them [100].

Face recognition has been extensively studied over the last few decades, and analyzing facial data has become a difficult and time-consuming task [170]. These technologies are highly valued in robotic production [172], clinical psychology [173], multimedia [174], intelligent security [171], and automobile security [175], as well as in applications such as face recognition and identification. Video can now be processed in real-time for computer vision jobs due to recent improvements in memory and CPU speed. The development of convolutional neural networks (CNN), one of the most well-known face detection and

identification techniques, has significantly enhanced the performance of computer vision tasks [107].

Despite the advantages of CNN, face recognition algorithms continue to struggle with changes in facial posture. To mitigate this issue, further research is recommended. Masi et al. [108] described how to create two CNN models that compare the frontal and profile faces and calculate the posture distribution of the training data. Liao et al. also used multi-keypoint descriptors to represent align-free faces in a partial face recognition localization approach [109], where the content of the picture and the face image determined the size of the descriptors. This study [66] aims to thoroughly examine the application of the local binary pattern (LBP) in conjunction with the convolutional neural network (CNN) for real-time human recognition and face detection.

This is a result of the remarkable performance of deep learning methods on a range of identification and recognition tests. The CNN method stabilizes until it achieves the target learning rate and performs better as the number of epochs increases. Using AdaBoost and Haar-Like features, Viola and Jones [87] developed a cascade face detector that trains cascaded classifiers with good real-time performance. In real-world applications with greater visual diversity in human faces, this detector may suffer from multiple degradations, even with more advanced features and classifiers, as noted in several articles [88, 89, 91]. Particularly good results are obtained when deformable part models (DPMs) are integrated into the cascade structure for face detection [90, 92, 93].

However, they are computationally expensive, and they typically require costly annotation during the training phase. Recent years have witnessed significant advancements in convolutional neural networks (CNNs) in various computer vision applications, including face recognition and image categorization [94, 163]. Several CNN-based face detection algorithms have been released in recent years, largely due to the efficiency of CNNs in computer vision applications. For facial attribute recognition, Yang et al. [96] create candidate face windows using deep learning neural networks that have strong responsiveness in face areas. However, due to the complex CNN structure, this method is computationally intensive and requires a significant amount of time.

Li et al. [97] employ cascaded convolutional neural networks (CNNs) to recognize faces. However, their approach requires bounding box calibration from face identification, incurring an extra computational expense while ignoring the intrinsic connection between bounding box regression and facial landmark localization. Additionally, the alignment of the face is crucial.

Template fitting methods [98, 164, 165] and regression-based methods [166–168] are widely employed. Zhang et al. [98] proposed employing facial attribute detection as an additional task using deep convolutional neural networks to enhance face alignment performance. However, the clear connection between these two objectives is often overlooked by most face alignment and identification methods. Although numerous works attempt to address them collectively, they have limitations. For instance, Chen et al. [95] carry out alignment and detection concurrently using random forests by utilizing the characteristics of pixel value differences. However, the performance is limited by the usage of handcrafted parts.

Using multitasking, Zhang et al. [99] enhance the accuracy of multi-view face recognition with CNN. The accuracy is still constrained by the early detection windows produced by a subpar face detector. On the other hand, mining complicated samples is necessary to improve the detector's performance throughout training. However, conventional hard sample mining typically occurs offline, which significantly raises the number of manual procedures. The ideal hard sample mining technique would be an online approach that automatically adapts to the existing face alignment and detection training process. The integration of these two tasks is suggested in Paper [15] using unified cascaded CNNs and multitask learning. The suggested CNNs have three stages. It quickly creates candidate windows in the first stage using a shallow convolutional neural network (CNN). A large percentage of the non-face windows are subsequently rejected by refining the windows using a more advanced convolutional neural network (CNN).

After using a more powerful CNN to refine the results, it then outputs the locations of the facial landmarks. A discussion is held regarding a study titled "Video Indexing through Human Face" [130]. The Viola-Jones algorithm, which utilizes AdaBoost and Haar-like features, is employed in Papers [130, 149] to detect faces. The drawback of this method is that faces that are angular or have changed direction in images or videos cannot be accurately detected. To enhance the quality of the EAN-8 linear bar code in illumination-invariant face

images, the researcher in [149] combined the Local Gabor Filter Approach (LGFA) with the Viola-Jones technique for face detection from input video. The window approach creates a stable bar code. However, the study does not address whether faces are rotated in the input video, whether faces are angled, or whether small face detection is used. This study, therefore, investigates an indexing method that utilizes a human face to overcome the aforementioned challenges.

The highly computationally efficient CNN architecture known as ShuffleNet was developed by Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun [16], specifically for mobile devices with extremely limited processing power (such as 10-150 MFLOPs). Pointwise group convolution and channel shuffle are two new processes used in the new architecture, which significantly lowers computation costs without sacrificing object detection (facial) accuracy. According to this article, Shuffle Net is among the best designs for small devices among all the popular and well-known ones. In real-time video face recognition, however, certain keyframes (i.e., faces) are often omitted. Therefore, relying solely on a shuffle net lowers the accuracy of face detection.

Chapter 1 also notes that the widespread use of video recording technology and the Internet's rapid expansion over the past few decades have led to the production, storage, sharing, and transmission of enormous amounts of video data every day. Meanwhile, this information has a wide range of potential applications in the fields of pattern recognition and computer vision, such as semantic video indexing and retrieval [176], [177], [178], video action recognition [179], person re-identification, retrieval of similar or nearly identical videos, and movie character identification [180]. The vast majority of the videos include people. For this reason, it may be a frequent occurrence in our daily lives to acquire images that depict a certain person based on a single question about that person [181].

Despite the great interest in face video indexing and retrieval, many challenges remain. Several methods for indexing and retrieving videos have been developed. Most image indexing techniques leverage low-level features such as color and texture. It is not easy to find and index people with subpar image and video indexing technology. In the scene seen in the picture and video, a human is one of the most important components. In the study [23], a method for integrating person detection with image and video sequence recognition is

described. Indexing the video using the face-based approach described in the paper [23] takes more time and space.

Nowadays, the majority of Internet users get their enjoyment primarily from videos. It inspires individuals, companies, and commerce through communication channels. Therefore, the bandwidth of the communication connection is crucial for data transfer between devices. Transferring a human face as data via a communication channel will need more time, space, and bandwidth. The problem could not be fully addressed in Papers [23] and [63]. Most previous face recognition research has focused on facial identification in still images. Nonetheless, it is challenging to recognize individuals from a single image due to common

problems such as changes in illumination, occlusions, angular faces, directional changes in faces, and facial expressions. These factors frequently cause more changes in face image than changes in identity. It is now feasible to capture, store, and analyze face films due to improvements in computational power and the development of reasonably priced video cameras.

Using multiple-frame video inputs results in duplicate and expensive data. More accurate and dependable face recognition is believed to be possible by carefully recording additional information, which is considered capable of mitigating the inherent uncertainties in recognition based on images, such as poor resolution, sensitivity to changes in posture, and occlusion. Additionally, video inputs can be used to capture facial dynamics useful for facial recognition.

The paper [12] proposes a new face detection network called the DSFD (Dual Shot Face Detector), which is based on three innovative ideas: enhanced feature acquisition, progressive loss design, and anchor assignment using enriched data. The paper [13] proposes a novel face detection system called Blaze-Face. The Single Shot Multi-Box Detector (SSD) framework serves as its foundation, and it is tailored for inference on mobile GPUs. Blaze Face has consistently outperformed other face identification algorithms while maintaining real-time performance across various benchmarks. Consequently, it has been widely adopted in multiple industries, including social networking, augmented reality, and video conferencing. Nevertheless, small devices are the primary application for this neural network.

In a study, Redmon Joseph and Farhadi Ali discuss the YOLO v3, a lightweight face detector model [14]. An improvement over earlier YOLO detection networks is YOLO v3. It has improved multi-scale detection, the feature extractor network's strength, and various loss function modifications compared to previous iterations. Many more targets, both large and small, may now be detected by this network. Like previous single-shot detectors, YOLO v3 is, of course, quick and allows for real-time inference on GPU devices. However, this neural network is primarily used in lightweight gadgets.

Building upon YOLO v3, Bochkovskiy et al. introduced YOLO v4, the fourth iteration of the YOLO object recognition system, in 2020. The publication discusses this advancement [182]. To better fit the size and shape of the objects that are recognized, YOLO versions three and four employ anchor boxes with different sizes and aspect ratios. "K-means clustering," a new technique for building anchor boxes, is introduced in YOLOv4. A clustering technique is used to first group the ground truth bounding boxes into clusters, and then the anchor boxes are

constructed from the cluster centroids. This allows the anchor boxes to reflect the size and shape of the identified items more accurately. In YOLO v3 and v4, the models are trained using comparable loss functions; however, YOLO v4 introduces a new concept known as "GHM (Gradient Harmonised Mechanism) loss." This particular version of the loss function is designed to enhance the model's performance on imbalanced datasets.

Furthermore, YOLO v4 builds upon YOLO v3's Feature Pyramid Network (FPN) design. For small object detection, the author's research [103] employs a unique YOLOv4-based method. Although the aforementioned techniques can effectively improve the model's detection accuracy, the original YOLOv4's detection accuracy in small object detection tasks is limited by the complex background, low resolution of the object to be detected, and limited amount of information available. Moreover, YOLOv4 includes a large number of parameters and is not suitable for mobile devices.

According to the study [105], YOLOv4 serves as the foundation for YOLOv5, one of the most popular object detection models among researchers. To better balance detection speed and accuracy, YOLOv5 has enhanced and optimized YOLOv4. Due to its high computational

speed and low device performance requirements, the YOLOv5 model is more suitable for deployment on intelligent terminal devices than the YOLOv7 and YOLOv8 models.
The YOLOv5 algorithm has four distinct size models: YOLOv5s, YOLOv5l, YOLOv5x, and YOLOv5m. The YOLOv5s model is the smallest in the series, with the fewest layers and the least computational complexity. It performs well on low-processing-power devices. As a result, the YOLOv5s method is used in this study to compare the YOLOv8n algorithm for face detection from input video. The publication [105,183] claims that Yolov5s is slower than Yolov8n in terms of speed (FPS) and mean average precision.

When it comes to object detection, the YOLOv5s model typically achieves around 108 frames per second. However, depending on the technology and other variables, the actual frame rate may differ. Moreover, some studies indicate that YOLOv5s can achieve a frame rate of 182.4 frames per second on a CPU without a GPU. The hardware and particular setup determine YOLOv8n's Frames Per Second (FPS). Nonetheless, it is designed to be incredibly fast and can reach high frame rates, particularly on GPUs. For instance, it can achieve 83 frames per second (FPS) on an NVIDIA T4 GPU. Even at much lower rates, a CPU can still be helpful.
Additionally, YOLOv8n is optimized for devices with limited resources, such as mobile devices and drones. The very small amount of FLOPs (Floating-point Operations Per Second) and parameters of YOLOv8 made it an efficient and scalable approach, according to this

research. However, the performance of the models has been observed to vary depending on the specific dataset and task. The authors of these papers emphasize the importance of considering the particular requirements of each job when selecting a model, and they also suggest that YOLOv8 is a promising model for real-time object recognition tasks. The authors of these studies not only indicated that YOLOv8 is a promising model for real-time object recognition tasks, but they also emphasized the importance of considering each job's specific requirements when selecting a model.

The objectives of video indexing using the human face approach are to identify faces in images and video scenes, index and retrieve those needed for information retrieval based on the face image, and perform other related tasks. Video indexing recognizes faces in pictures and video scenes using the human face approach, then uses the face image to index and retrieve the required faces for information retrieval. Not all of the frames in a video used for

this task are necessary for facial recognition. All of the frames from the given video must be extracted before the significant frames can be inferred from the extracted frames. The important frames are then selected from each extracted frame using a human face for video indexing and retrieval.

Face detection from keyframes must be precise and efficient to index and retrieve videos utilizing human faces. Recent years have seen the investigation of several artificial intelligence (AI) approaches, particularly machine learning and deep learning, for automatically distinguishing faces in key frames of input videos. These methods are thoroughly reviewed in Chapter 2. Nevertheless, it is challenging to identify faces from keyframes and videos due to factors including facial expressions, positions, moods, illumination variations, occlusions of the face image, and directional shifts of faces in images as well as videos. Additionally, it is challenging to detect small faces in densely populated input footage. Accurately detecting and recognizing faces from incoming video and keyframes is, therefore, essential.

This chapter presents an innovative approach, "Video indexing through the Face Images using Deep Learning Models," to address the issues mentioned above. Furthermore, by using face images for video indexing, we first classify a video as information before extracting each edge from it. From each frame in the video, the keyframes are selected based on the color histogram difference. The face is returned as the key frame from the input video if the threshold of the color histogram is met. The extracted key frames are subjected to face recognition using MTCNN, ShuffleNet, a Combined MTCNN and ShuffleNet, YOLOv5s, and YOLOv8n. The key frame of the input video is used to extract an identified human face for video indexing purposes. The following are some major contributions made by the research mentioned in this chapter:

- Section 5.2.1 provides a detailed description of our proposed "Video Indexing through Human Faces by Combined Deep Learning Neural Networks" method, which uses Deep Learning Models to focus key frame extraction using the color histogram method and face detection from various facial expressions, pose, emotion, illumination change, age, and occlusions of the face image of the video and keyframe.

- The color histogram approach [140], a color-based frame difference technique, is employed for key frame extraction in our proposed video indexing mechanism, which utilizes human faces. The underlying assumption of this method is that two frames with identical backgrounds and identical (but moving) objects will have matching histograms that are similar in nature.

- Using the MTCNN, Shuffle Net, and Combined MTCNN and Shuffle Net algorithms, faces are detected in key frames of the input video. Papers [15] and [16] provide details on these methods, respectively.

- We have conducted several experiments to utilize a deep learning model for video indexing, aiming to detect face images from key frames of input videos. Our experimental results, remarks, and failure instances are presented below.

- Reduce the face detection time and increase the number of faces found in the video by combining the MTCNN and Shuffle Net methods. This will help index the input video by using the faces as cues.

- Eigen face recognition, a facial identification technique that employs principal component analysis (PCA) for security (including identifying criminals and preventing crimes, locating missing children, and expediting investigations), then uses the input video to determine the face to monitor attendance, healthcare, the retail sector, and finance.

- Results comparing MTCNN, ShuffleNet, and the combined MTCNN and ShuffleNet techniques show that the combination of these two methods can recognize more faces and perform face detection more efficiently. The findings are described in detail in sub-section 5.3.1

- We have proposed a technique called "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8" to address several important problems, including the intricacy of time and space, the

difficulty of saving important video frames, the lack of storage capacity, angular changes in the face, lighting invariant features, and changing posture. In section 5.2.2, this approach is described in depth.

- When there are multiple persons in a video, use YOLOv5s and YOLOv8n to identify the smaller faces.

- The key frame storage problems were resolved, and the keyframes were extracted from the frames by creating cropped face portraits using the Lightweight Deep Learning algorithm (YOLO v5s and YOLO v8n).

- The results are then compared using the key frame extraction and face detection algorithms YOLOv5s and YOLO v8n. A detailed description of the results can be found in subsection 5.3.2.

The rest of the chapter is organized as follows. The suggested methodology for video indexing using face images using deep learning models is described in depth in Section 5.2. Detailed experimental evaluations, discussions, and failure instances are included in Section 5.3. Section 5.4 provides concluding remarks and outlines the future course of the work.

5.2. Proposed methodology for the Video indexing through the Face Images using Deep Learning Models

This chapter's section will primarily focus on the processes involved in creating our proposed video indexing system, which utilizes face images and deep learning models for video indexing. Figure 5.1 illustrates the workflow of the proposed method. (A) Frame extraction from the input video is the initial stage. (B) Extracting the keyframe (based on a human face) from the identified frame. (C) Identifying faces using the key frame (D) Lastly, store these keyframe faces that were detected and use them for indexing.

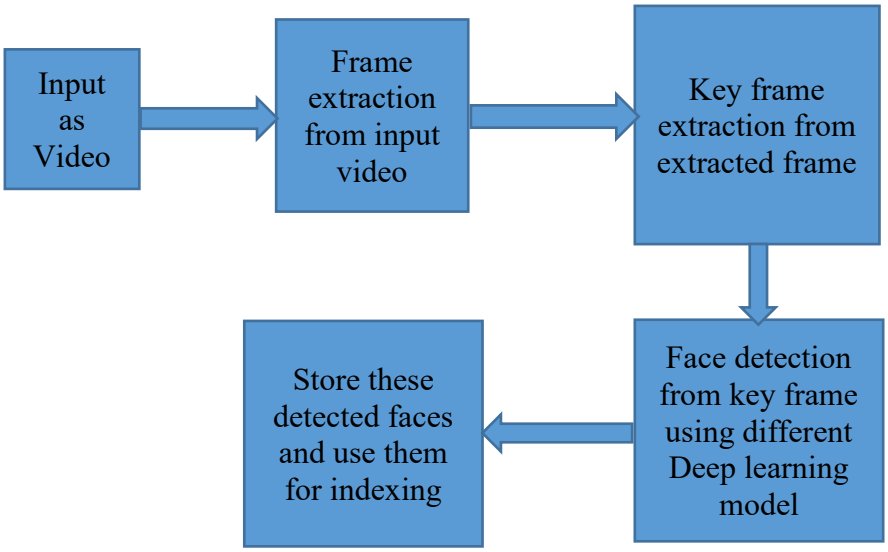


Fig 5.1: Flow diagram of the proposed method

A. Extracting frames from the input video

The shot, frame, and scene all combine to produce a vibrant video. Therefore, extracting the still images—which are represented in the input videos as scenes, shots, and frames—is the first step.

B. Extracting key frames from extracted frames based on a human face using the Colour Histogram method

A keyframe has essential information about each image. In this proposed work, human faces in different positions, expressions, and lighting conditions are considered important frames. A range of essential frame extraction methods are described in publications [74], [144], [145], [146], and [147].

C. Face detection from key frames using various deep-learning models

Faces are identified from the gathered key frames using a variety of deep-learning techniques. This method is discussed in full in the second chapter of this thesis.

D. These detected faces should be saved and used for indexing.

Once these distinct faces have been identified from key frames of various input videos using multiple deep-learning models, they will be saved and used for indexing.The following are the techniques recommended in this chapter for video indexing utilizing face images and deep learning models:

5.2.1. "Video Indexing through Human Faces by Combined Deep Learning Neural Networks" is the suggested approach

A survey and analysis of several studies in Chapter 2 indicate that person detection is challenging when video indexing is performed using low-level attributes. However, several characteristics, such as facial expressions, position, mood, illumination fluctuations, occlusions in face pictures, and directional shifts of faces in images and videos, make it challenging to detect faces from keyframes and videos. Detecting small faces in input footage with a high population density is also difficult. Shuffle Net is one of the best designs for small devices among all the popular and well-known ones, according to the paper [16]. Nevertheless, while identifying faces in real-time video, some crucial frames—faces—are omitted. Therefore, using just a shuffle net lowers the accuracy of face detection. This section, "Video indexing through human face as a cue using Combined Shuffle net and MTCNN algorithm and recognition of face image," offers a creative solution to these issues.

Fundamental issues, such as posture shift, lighting-invariant features, facial angular changes, enhanced face detection, storage capacity restrictions, the inability to store important video frames, and time and spatial complexity, are all addressed in this work. By integrating the MTCNN and ShuffleNet algorithms, the research primarily contributes to face image recognition, as well as video indexing that utilizes the human face as a cue.

- Cropped facial portraits were generated using the MTCNN, ShuffleNet, and Combined ShuffleNet algorithms to extract key frames from the video and address keyframe storage issues.
- Reduce the face detection time and increase the number of faces found in the video by combining the MTCNN and Shuffle Net methods. This will help index the input video by using the faces as cues.

- Eigen face recognition, a facial identification technique that employs principal component analysis (PCA) for security (including identifying criminals and preventing crimes, locating missing children, and expediting investigations), then uses the input video to determine the face to monitor attendance, healthcare, the retail sector, and finance.

Thus, the proposed approach of indexing videos addresses issues with posture change, storage capacity restrictions, storing video keyframes, and time and space complexity. A block diagram in Figure 5.2 depicts each step in the method described in this chapter. (a) The first stage involves extracting frames from the video input. (b) A difference in the color histogram is used to differentiate keyframes from frames extracted from the input video. (c) Detecting faces within the key frame using MTCNN and the Shuffle Net approach. (d)The next step is to identify a face using Eigen face recognition, a PCA-based facial identification technique. (e) Video indexing relies on face detection and recognition.

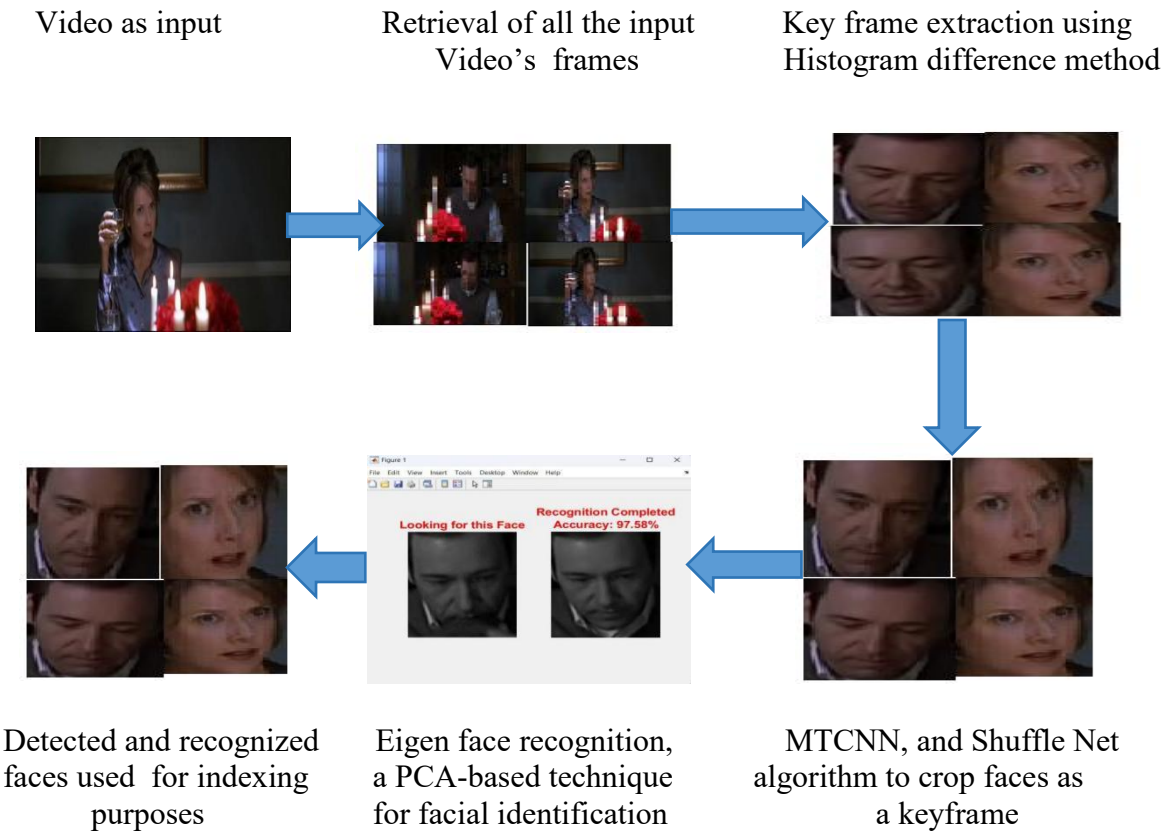


Fig 5.2: Block diagram of the proposed system

5.2.1.1. Extraction of frames

Dynamic video is the result of the scene, shot, and frame working together. Thus, extracting the still images—which are represented as a scene, shot, and frame—from the input videos is the first stage. A scene is a collection of shots, and a shot is a series of frames. With a frame rate of 20 to 30 frames per second, the vintage video is especially comprehensive. There is unnecessary information in the frame, which is a still image that is a component of a video. The frame from Holly Wood's movie American Beauty-00222 is shown in Figure 5.3.

Fig 5.3: The frame of American Beauty -00222 Video

5.2.1.2. The Key Frame's Extraction

The keyframe includes the most crucial elements of each shot. Human features in various positions, lighting settings, and illuminations are essential elements for this work. The curve saliency motion capture findings, the likelihood scale, and a few more strategies are examples of key frame extraction strategies. Nevertheless, the Colour Histogram technique is used at this stage to extract the keyframe from every frame of a particular video. Keyframes can be extracted from the frames using the Colour Histogram difference if the detected change exceeds the threshold. The next Keyframe is selected when the color histogram disagreement threshold is met. Figure 5.4 displays some of the most important frames from the Holly Wood-starring Hollywood movie American Beauty -00222.

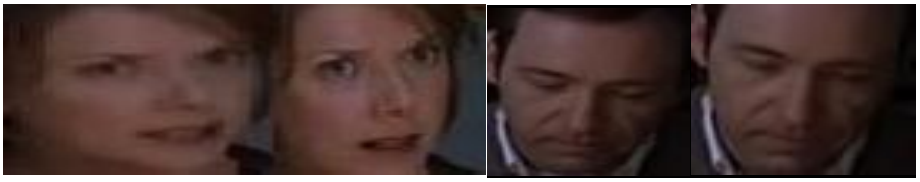


Fig 5.4: The keyframe for "American Beauty"-00222, with the face as the primary component.

The algorithm and formula for separating the two consecutive frames of the color histogram are described in a paper [130, 149].

5.2.1.3. Face detection from key frame utilizing the Shuffle Net and MTCNN combination technique

MTCNN, ShuffleNet, and a combination of the two algorithms are used to detect faces from keyframes. The specific methods used by these algorithms are listed below.

A. MTCNN Algorithm

Faces are recovered using the MTCNN object detection approach from the keyframes that were retrieved. There are three steps in the MTCNN algorithm. In the initial phase, candidate windows are rapidly generated using a shallow CNN. Then, to reject a large percentage of the non-face windows, the windows are refined using a more advanced CNN. Finally, it uses a more powerful CNN to improve the accuracy of face landmark placement and output. The architectural diagram of MTCNN is displayed in Figure 5.5. As discussed in the publication [15], the accompanying graphic deconstructs the MTCNN approach into its several phases.

The Proposal Network (P-Net), using a fully convolutional network known as a proposal network, stage 1 yields the candidate windows and their bounding box regression vectors. The computed bounding box regression vectors are used to calibrate the candidates. Once heavily overlapped candidates are combined, non-maximum suppression (NMS) is applied.

The Refine Network (R-Net), a separate CNN, receives all the candidates after combining NMS candidates, excluding many inefficient candidates, and calibrating using bounding box regression.

Stage 3 (O-Net): Although it is comparable to Stage 2, this stage aims to provide a more detailed description of the face. In particular, the network will output the five facial landmarks' locations.

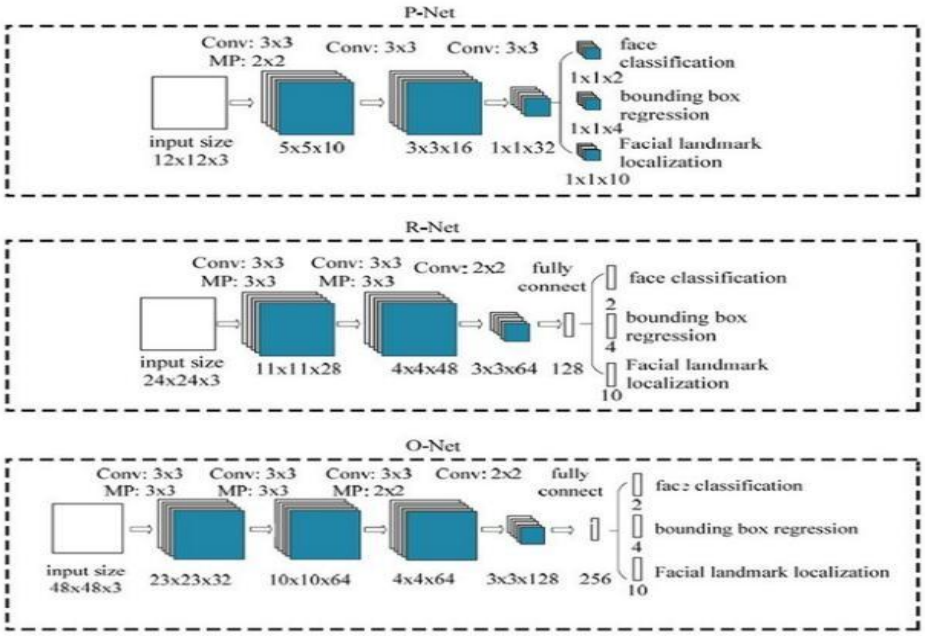


Fig 5.5: Architectural diagram of MTCNN [101]

B. Shuffle Net Algorithm

Shuffle Net [16], also known as Face++, is a CNN architecture developed by Megvii Inc. for mobile devices with a processing power of 10–15 million floating-point operations per second (MFLOPs). The Shuffle Net utilizes channel shuffle and pointwise group convolution to reduce computational costs without compromising accuracy. It can classify images from ImageNet with fewer top-1 errors than the Mobile Net system and outperforms Alex Net by more than 13 times in actual speed. Low computing costs and a small number of parameters are needed to achieve great precision. The basic blocks of Xception and Res Net strike an exceptional balance between computational expense and representational ability by integrating depth-wise separable or group convolutions.

Nevertheless, 1x1 convolutions, sometimes referred to as pointwise convolutions, must be taken into consideration to some extent. Pointwise convolutions are expensive and can drastically lower the accuracy of small networks, limiting the number of channels that can be used to satisfy the complexity criterion. Shuffle Net's architectural diagram is shown in Figure 5.6.

In this study [16], we demonstrate the potential of group convolutions to reduce computing costs through examples. Two stacked group convolution layers, as shown in Figure 5.6(a), hinder information transfer between channel groups and compromise representations. The group convolution can receive input data from many groups, as illustrated in Figure 5.6(b).

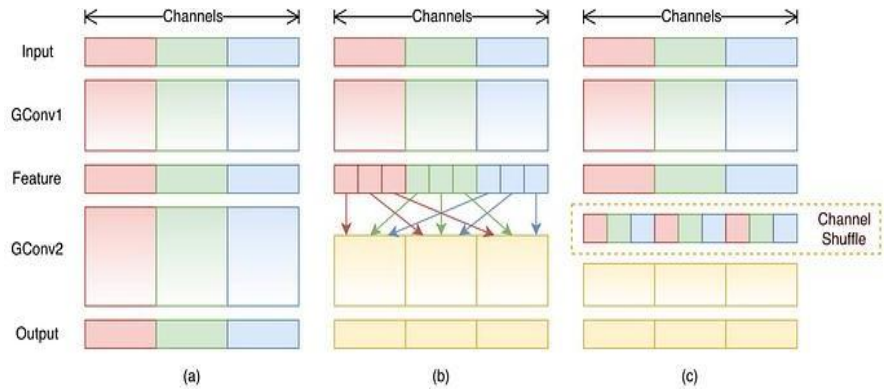


Fig 5.6: Two stacked group convolutions and channel shuffling. The abbreviation for group convolution is GConv.a) Two convolution layers with the same number of groups laid between them. Only the input channels in the group are connected to each output channel. when GConv2 receives data from various groups; b) input and output channels are fully connected following GConv1; c) a channel shuffle implementation equivalent to b).[16]

Examine the degree of correlation between the input and output channels. A channel shuffle operation is used to configure and implement the feature map from the previous group layer, as shown in Figure 5.6(c). More muscular architectures can be produced by numerous group convolutional layers using the channel shuffle algorithm. Since the channel shuffle operation

still works in the stacked layers, the architecture in Figure 5.6(c) is favored even though the convolutions in Figure 5.6(b) have different groups. Additionally, because it is differentiable, network topologies can incorporate it.

C. Shuffle Net and MTCNN Combined Algorithm

This technique combines the Shuffle Net and MTCNN algorithms to achieve high accuracy at a cheap computational cost. Using Shuffle Net to preprocess the input image speeds up

MTCNN. After the image has been preprocessed, MTCNN is used to detect faces precisely. The following algorithm, 5.1, discusses a hybrid Shuffle Net and MTCNN algorithm for face detection from input video.

Algorithm 5.1: Face detection with a combination of MTCNN and Shuffle Net algorithms.

Input: The input is video.
Output: Face as the result of the video input
Steps:
1. Start
2. Read the video that the user entered.
3. Extract the frame from the input video.
4. The total number of frames in the video should be stored.
5. Extract the Keyframe (face) by applying the color histogram difference to the retrieved frame.
6. Use the Shuffle Net technique in conjunction with MTCNN to detect faces in the keyframes.
7. After processing every keyframe in the input video and locating every face using MTCNN and ShuffleNet, the operation is complete.
8. If not, repeat the process.
9. Store every face that appears in the video.
10. Stop

The face detection feature in Key frames, utilizing a combined Shuffle Net and MTCNN algorithm, is depicted in screenshots in Fig. 5 .7.

Fig 5.7: Important sequences from "Holly Wood's American Beauty-00222" with faces detected

5.2.1.4. A PCA-based technique for facial identification called Eigen face recognition is utilized to recognize a face

The benefits and drawbacks of facial recognition technology are a topic of discussion. Although many participants highlight the benefits, critics typically focus on the drawbacks. Facial recognition technology raises concerns about privacy invasion, power abuse, and

potential misuse by rogue government officials. The media is paying more attention to facial recognition than ever before. Recent historical occurrences have caused a sharp increase in facial recognition spending. An surge in investment in biometric technologies, such as facial recognition, could result from the global COVID-19 epidemic. The highly contagious nature of COVID-19 makes contactless interactions extremely valuable.

The primary application of facial recognition technology remains in security measures. Facial recognition is acknowledged as one of the most simple and reliable methods for identifying people in a variety of fields, including enhanced public safety, aviation and transportation, retail, access and authentication, quicker processing, seamless integration, and more. Eigen face recognition, a facial recognition method that analyses human faces to identify objects, is based on Principal Component Analysis (PCA). The Eigen facial recognition method of PCA is used to recover global features. Two factors that affect the accuracy of the recognition system are facial recognition and face perspectives. The challenge is that, to guarantee precise feature extraction, every image must have precisely the same size and color depth.

Since the system requires a large number of training images to achieve high recognition accuracy, the quantity of images is CNN's primary issue. As a result, the Eigen facial recognition approach is used in this work. The premise behind this method is that faces can be represented as a linear combination of "Eigenfaces" made from a collection of training images. These Eigenfaces are arranged in ascending order by eigenvalues in the covariance matrix of the training set. The primary features of facial images, including head tilt, illumination, and facial expressions, are captured by the Eigenfaces. When Eigenface recognition is used, the system must first be trained using facial image examples.

The system generates feature vectors for each face by computing the Eigenfaces and projecting each training image into the Eigenface space. To identify the closest match during the recognition phase, the system compares the feature vectors of the input face with those of the training faces. Selecting the most important eigenvectors and incorporating additional training images can improve recognition performance. Eigen facial recognition is simple,

effective, and precise, among other benefits. It can adapt to changes in lighting, head tilt, and facial expressions and function with low-resolution images. It is sensitive to changes in the size and form of the face, and it cannot detect partial faces, occlusions, or disguises, among

other limitations. An Eigen face recognition technique that employs principal component analysis (PCA) to recognize faces is covered in algorithm 5.2 below:

Algorithm 5.2: Eigen face recognition algorithm, which recognizes faces using principal component analysis (PCA)

Input: A human face dataset that was extracted from the input video.
Output: Faces were recognized from the given video.
Steps:

1. Start
2. From the input video, open the database of human faces that were detected.
3. Select a face image at random from the face database.
4. Once the chosen face has been recognized, remove the selected image from the database.
5. A set of Eigenfaces is generated by applying Principal Component Analysis (PCA) to the remaining images.
6. Find the signature for each of the remaining facial images.
7. Compute the signature of the image.
8. Calculate the distance between the signatures.
9. Find the nearest image.
10. Evaluate the recognition accuracy. To get the percentage, the recognition accuracy is calculated as (1 - z(i) / (norm (s, 2)) * 100.
11. Stop

Figure 5.8 displays a snapshot showing the accuracy of face recognition from the movie "Holly Wood's American Beauty-00222" with faces got rid of.

5.2.1.5. Face detection and recognition are utilized for video indexing

The faces are used for video indexing after they have been recognized and detected in the input video.

Figure 5.8: Eigen Face identification, which employs principal component analysis, removes faces from a screen grab showing the accuracy of face identification from the movie Holly Wood's American Beauty-00222.

5.2.2. The proposed technique is "Video indexing and retrieval using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8 through the human face as a cue."

To concentrate on key frame extraction using the color histogram method, we have proposed a "Video indexing and retrieval using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8 through the human face as a cue," which includes face detection from different facial expressions, pose, emotion, illumination change, and occlusions of the face image of the video and keyframe. Apart from the complexities of time and space, this work addresses significant issues, including changing posture, storage capacity constraints, the inability to recognize small faces in video, and the inability to save key video frames.

The paper's topic is "Video Indexing through Human Face" [130]. "The Viola-Jones approach, which makes use of AdaBoost and Haar-Like features, is used in studies [130,149] to detect faces from video and face datasets. Although the Viola-Jones technique employs an outdated framework, it remains a reliable, fast, and robust face detection algorithm (not recognition). This algorithm's limitation is that it can only detect faces that are fully frontal and upright.

Major problems, including illumination-invariant characteristics, facial angular variations, and posture shifts, are not addressed by the Viola-Jones algorithm or study [130, 149]. Face detection takes longer, however, in lightweight, real-time devices that utilize the Viola-Jones method. Furthermore, it has been observed that Viola Jones and MTCNN are not good at identifying small faces in videos.

In the article "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8," an innovative solution to all of these issues is demonstrated. Numerous critical problems are addressed in this study, including the intricacies of time and space, the difficulty of saving crucial video frames, the lack of storage capacity, facial angular variations, illumination-invariant features, and shifting posture. The Lightweight Deep Learning algorithm (YOLO v5 and v8) is used for video indexing, utilizing human faces as a cue, which is the main contribution of the research. This approach is offered as a solution to these problems.

- In a video scenario with multiple people, use Yolov5s and Yolov8n to detect small faces.
- To address the keyframe storage problems and extract key frames from the frame, cropped face portraits were created using the Lightweight Deep Learning method (YOLO v5s and YOLO v8n).
- The results are then compared using the key frame extraction and face detection algorithms YOLOv5s and YOLO v8n.

As a result, the suggested method for indexing videos addresses the complexities of location and time, storage capacity limitations, moving postures, and saving important video frames. The method presented in this work is divided into several parts, each of which is illustrated by the blocks shown in Fig. 5.9 (a). The first step is to extract the frame from the input video. (b) Keyframes are selected from the frames extracted from the video input using a variation of the color histogram. (c) Faces in the key frame are detected using two lightweight deep-learning algorithms, Yolov5s and Yolov8n. (d) Faces detected from keyframes are used for indexing.

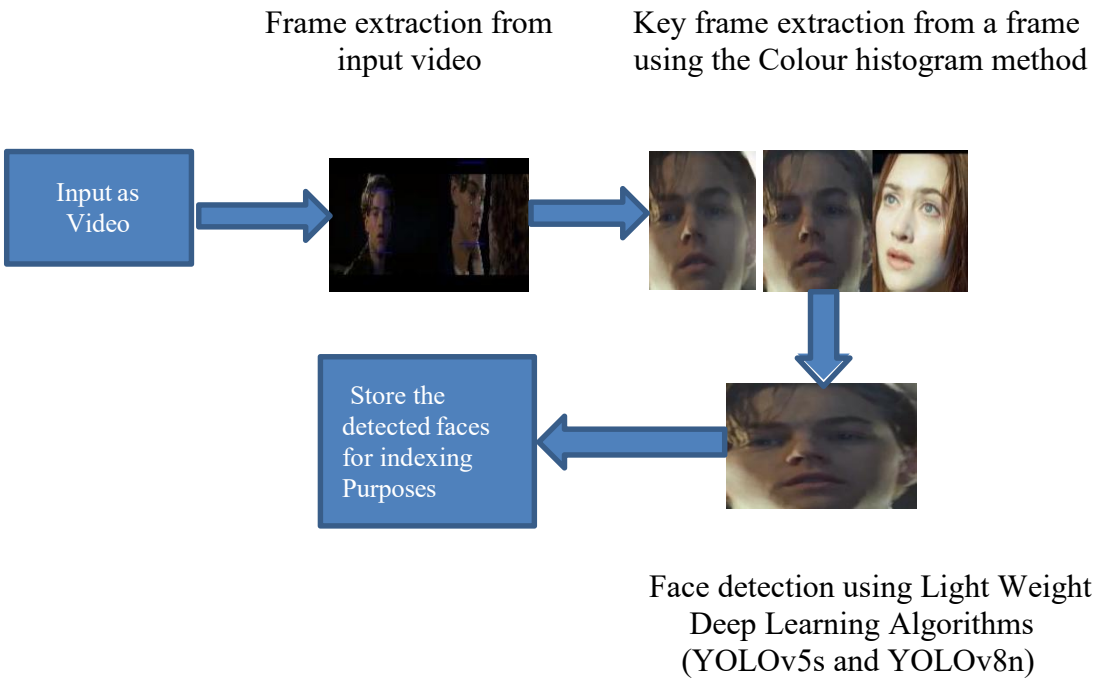


Figure 5.9: Conceptual System Block Diagram

5.2.2.1. Extract a frame from the input video

The scene, shot, and frame all work together to create a dynamic video. As a result, the first step is to extract the still images, or the scenes, shots, and images that were displayed in the input videos. The frame contains a still image from a video that includes much unnecessary information. Figure 5.10 shows a frame taken from the Titanic trailer_2 trailer video collected.

Figure 5.10: The video frame from Titanic_Trailer_2

5.2.2.2. To extract the key frame from the frame, use a color histogram

The most crucial information is contained in the key frame of any image. Keyframes in the current work are defined by human faces with distinct lighting, illumination, locations, and expressions. The similarities include the likelihood scale, the curve saliency motion capture results, and several more methods. However, the Colour Histogram Technique is now used to retrieve the key frame of a specific video from each frame. It is possible to extract significant frames from the frames using the Difference in Colour Histogram. The frame is selected as the next keyframe if the observed difference exceeds the threshold magnitude and the color histogram disagreement criterion is met. Figure 5.11 displays a selection of key frames from the Titanic trailer trailer video collection.

Figure 5.11: The Face is the Primary Focal Point in the Titanic_Trailer_2 Key Frame.

A paper [130] describes the formula and algorithm for splitting the two subsequent color histogram frames.

5.2.2.3. Lightweight Deep Learning Algorithms were used to trim the face portraits in the key frames (YOLO v8n and v5s)

The keyframes of the provided video are used to identify faces by the YOLOv5s and YOLOv8n algorithms. By comparing results, the most effective method for face detection from key frames is determined. Numerous methods have been developed over time to aid computers with facial recognition. The first method for identifying faces from keyframes in this work is the YOLOv5s method. To enhance the fitness of the anchor box and the actual object, this algorithm first determines the priority anchor box size using the K-means

algorithm. Then, it modifies the anchor box's size. Second, the SE (Squeeze-and-Excitation) attention mechanism is incorporated into the fundamental network structure of YOLOv5s. It could enhance the network's capacity to extract features. Finally, this study utilizes four-scale feature detection to enhance the network's performance in detecting small-face targets.

"You Only Look Once version 8 Nano," or "YOLO v8n," is an instantaneous face detection method that uses deep learning to quickly and accurately identify faces in images and videos. An advancement over earlier YOLO models, YOLOv8n uses a convolutional neural network with two main components: the head and the backbone. The head is composed of multiple convolutional layers, followed by fully connected layers that predict class probabilities, bounding boxes, and objectness scores. Simply put, YOLOv8n incorporates a self-attention system into the network's brain and employs a feature pyramid network for multi-scaled object detection. This enables it to recognize objects of various sizes and scales and concentrate on particular regions within an image. Keyframe facial detection feature screenshots are shown in Figure 5.12. This suggests that the face detection ratio of the YOLOv8n algorithm is higher than that of the YOLOv5s algorithms.

Figure 5.12: Key Fames from the Titanic_Trailer_2 video data collection, with faces cropped

5.2.2.4 Utilising a Human Face Index Check Video to Index the Input

After being detected and saved, faces from the key frame are used for indexing.

5.3. Results and Discussion of the Experiment

In this section, we utilize deep learning models to examine the effectiveness of video indexing using face images. The datasets employed in this study are explained in detail in subsection - 5.3.1. We will discuss the final results of our proposed method, "Video Indexing through

Human Faces by Combined Deep Learning Neural Networks," in subsection 5.3.2. The results and discussion of our proposed method, "Video indexing and retrieval using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8 through human face as a cue," are explained in detail in subsection 5.3.3.

5.3.1. Description of the Dataset

In the human face-based video collection, this cropped face served as the initial source of the Keyframe used in this investigation. Next, indexing is done using the human face that can be seen in the video. To validate the method, the following video data sets were employed.
The initiative begins with the Hollywood video dataset [152]. Video snippets from thirty-two human action films are also included. The sample must fall into at least one of the eight categories. The test set is divided into two 12-film practice sets to form a 20-film data set. Two hundred thirty-three video recordings comprised the automated learning set, which was assembled using automatic script-based action labeling. Approximately 60% of the labels

were correct. The training set of Hollywood results is clean, with 219 video samples with manually checked labels and 211 video examples with manually tested labels.

Next, we used the Movie Trailer Face Dataset [184], which included 101 YouTube movie trailers from the 2010 release year that featured celebrities from the PublicFig+10 dataset supplement. In these videos, face tracks were created using the previously described methodology. Of the 3,585 face tracks in the final dataset, 514 have known identities, and 63% are unknown (not shown in PubFig+10).

The trials then collect the dataset of trailer videos. The Trailer Video dataset is a large-scale, multimodal video-language dataset comprising over 20 million trailer clips and high-quality multimodal captions that include context, visual frames, and background music. Its objective is to enhance fine-grained multimodal-language model training and cross-modality investigations. In summary, the dataset included 27.1k hours of trailer videos with 2M+ LLaVA Video captions, 2M+ Music captions, and 60M+ Coca frame captions.

Lastly, there is a realistic TV series video dataset [151] that includes 27 episodes of six popular TV series. The 27 episodes included Sons of Anarchy (3), Modern Family (6), Mad Man (3), How I Met Your Mother (8), Breaking Bad (3), and 24 (4). The total duration of these videos is sixteen hours. In total, there are 6231 and 30 acts and activities in this movie. The dataset contains metadata for each action instance (e.g., single person, occluded, part of the action missing) that can be used to assess how well a method performs in specific challenging scenarios.

5.3.2. The result of the "Video Indexing through Human Faces by Combined Deep Learning Neural Networks" technique that we proposed

The result of the "Video Indexing through Human Faces by Combined Deep Learning Neural Networks" method, which we suggested, is covered in this subsection. A comparison of various algorithms demonstrates that MTCNN is the most specialized and accurate face detection method. However, it is not suitable for embedded and mobile devices because of its high computational cost. Shuffle Net may not be as precise as MTCNN despite being portable

and useful. The combined approach offers a great balance between accuracy and computational efficiency, making it suitable for mobile and embedded devices.

Additionally, it is noted that Shuffle Net, a thin neural network architecture, was developed for embedded and mobile devices. It uses less memory and processing power while maintaining acceptable accuracy because of a group convolutional operation. Shuffle Net can be trained to identify faces, although it might not be as accurate as other, more specialized algorithms. MTCNN is a specialized detection method that uses a multi-stage neural network to locate and identify faces in an image. It is well known for processing faces of various sizes, orientations, and lighting conditions with remarkable precision and adaptability. The Shuffle Net and MTCNN combination algorithm combines the two algorithms to achieve high accuracy at a cheap computational cost. MTCNN operates more efficiently when the input image is pre-processed using ShuffleNet. MTCNN is then used to precisely recognize faces in the pre-processed image.

The following tables display the face detection results and face-finding time for different video datasets utilizing the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN algorithms. Table 1 presents the number of faces detected in various video clips from the Hollywood dataset using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods. Table 2 presents the times for face detection using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods on several video clips from the Hollywood dataset.

Table 5.1: Comparison of the Shuffle Net and MTCNN Combined Algorithms for the Number of Faces Detected

Name of the video	**Shuffle Net**	**MTCNN**	**Shuffle Net+MTCNN**
	The number of faces detected	**The number of faces detected**	**The number of faces detected**
American Beauty 00170	111	222	222
American Beauty 00222	58	244	244
American Beauty 00443	164	350	350

American Beauty 00951	300	248	248
American Beauty 01597	562	1342	1342
As Good As It Gets -01766	279	830	830
As Good As It Gets -01935	149	454	454
Big Fish - 00674	439	1621	1621
Big Lebowski, The -00818	128	378	378
Casablanca - 03025	62	202	202

Table 5.2: Shuffle Net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm Execution time (in seconds) table

Name of the video	**Shuffle Net**	**MTCNN**	**Shuffle Net+MTCNN**
	Execution time (in seconds)	**Execution time (in seconds)**	**Execution time (in seconds)**
American Beauty-00170	12.083	14.501	13.823
American Beauty-00222	9.534	12.689	11.56
American Beauty-00443	71.818	72.739	73.026
American Beauty-00951	164.069	166.422	165.968
American Beauty-01597	136.176	137.285	137.185
As Good As It Gets -01766	47.77	53.272	51.357
As Good As It Gets -01935	23.334	25.554	54.445
Big Fish - 00674	37.99	52.685	51.164
Big Lebowski, The -00818	23.307	24.953	23.307
Casablanca - 03025	12.931	14.865	13.396

There are TV series and videos in the movie trailers. The number of faces detected in the Trailer Face Dataset and TV series video dataset, as well as the time required to detect faces using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN algorithms, are covered in Tables 5.3 and 5.4.

Table 5.3: Shows the number of faces extracted from the movie trailer face dataset and television series video dataset using the Shuffle Net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm.

Name of the video data set	Shuffle Net	MTCNN	Shuffle Net+MTCNN
	The number of faces detected	The number of faces detected	The number of faces detected
Movie Trailer face video Data set	1271	3050	3050
TV series Video Data set	630	1972	1972

Table 5.4: Shuffle Net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm Execution time (in seconds) table

Name of the video data set	Shuffle Net	MTCNN	Shuffle Net+MTCNN
	Execution time (in seconds)	Execution time (in seconds)	Execution time (in seconds)
Movie Trailer face video Data set	662.502	509.666	461.024
TV series Video Data set	165.912	195.783	194.545

Eigenface recognition utilizes principal component analysis (PCA) to identify faces following face detection. The 99.35% face detection accuracy is the most remarkable result in this testing. But it varies according to the quality, size, and form of the faces.

5.3.2.1. Strategies for comparison

This subsection presents a comparison of the performance of the MTCNN and ShuffleNet algorithms with the proposed method, which is the combined ShuffleNet and MTCNN algorithms. After face detection is finished, it is clear that combining Shuffle Net and MTCNN yields a more accurate method than either Shuffle Net or MTCNN alone. Compared to MTCNN and Shuffle Net, the proposed technique finds more faces and requires less

processing time. A graph in Figure 5.13 illustrates the primary distinctions between the Shuffle Net and MTCNN combined MTCNN and Shuffle Net algorithms.

This graph uses the same scene from "Holly Wood Movie" to compare the algorithmic performance. The number of faces detected by combining the Shuffle Net and MTCNN algorithms is shown in the graph after the experiments.

The graph in Figure 5.14 illustrates the key differences between the combined Shuffle Net and MTCNN algorithms, as well as the Shuffle Net and MTCNN algorithms. This graph uses the same "Holly Wood Movie video clip" to compare algorithmic performance. The testing is

followed by a graph that displays the execution time required for face detection using a combination of the Shuffle Net and MTCNN algorithms.

A comparison of the Shuffle Net algorithm, MTCNN, and the combined Shuffle Net and MTCNN algorithm for face detection in a movie trailer and TV video dataset, based on the number of faces found, is shown in Figure 5.15. The Shuffle Net algorithm, MTCNN, and the Shuffle Net and MTCNN combined algorithm for execution time needed for face detection in the movie trailer and television series video data set are compared in Fig. 5.16.

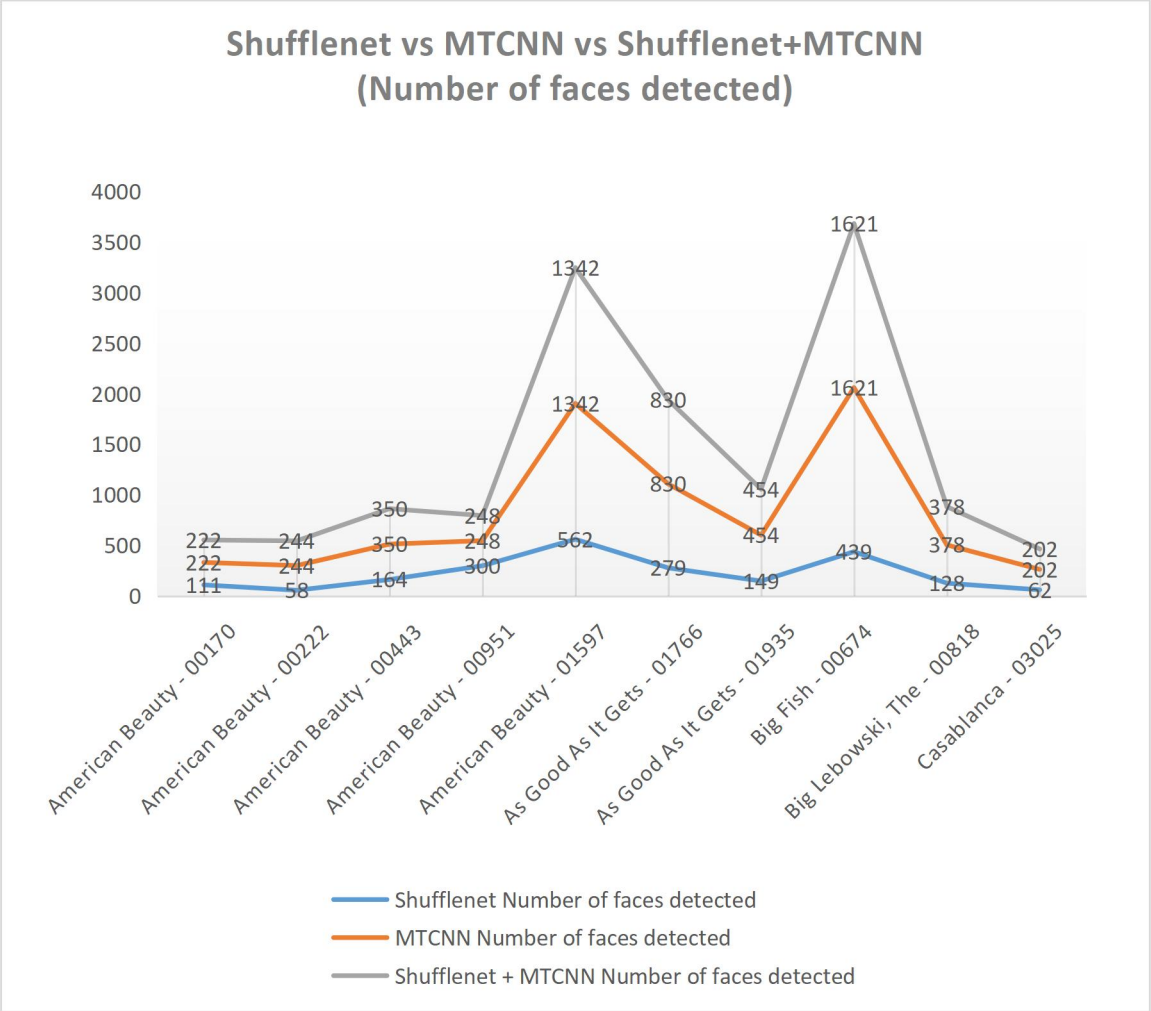


Fig 5.13: Graph for Comparative Analysis of Shuffle Net Algorithm, MTCNN, and ShuffleNet and MTCNN Combined Algorithm for Face Detection in 10 distinct Holly Wood video clips on multiple faces detected.

A comparison of different face detection algorithms shows that MTCNN is the most specialized and accurate approach. Its high computational cost, however, makes it unsuitable for mobile and embedded systems. Despite being portable and useful, Shuffle Net was unable to match MTCNN's accuracy. The combined approach offers a great balance between accuracy and computing economy, which makes it suitable for mobile and embedded devices.

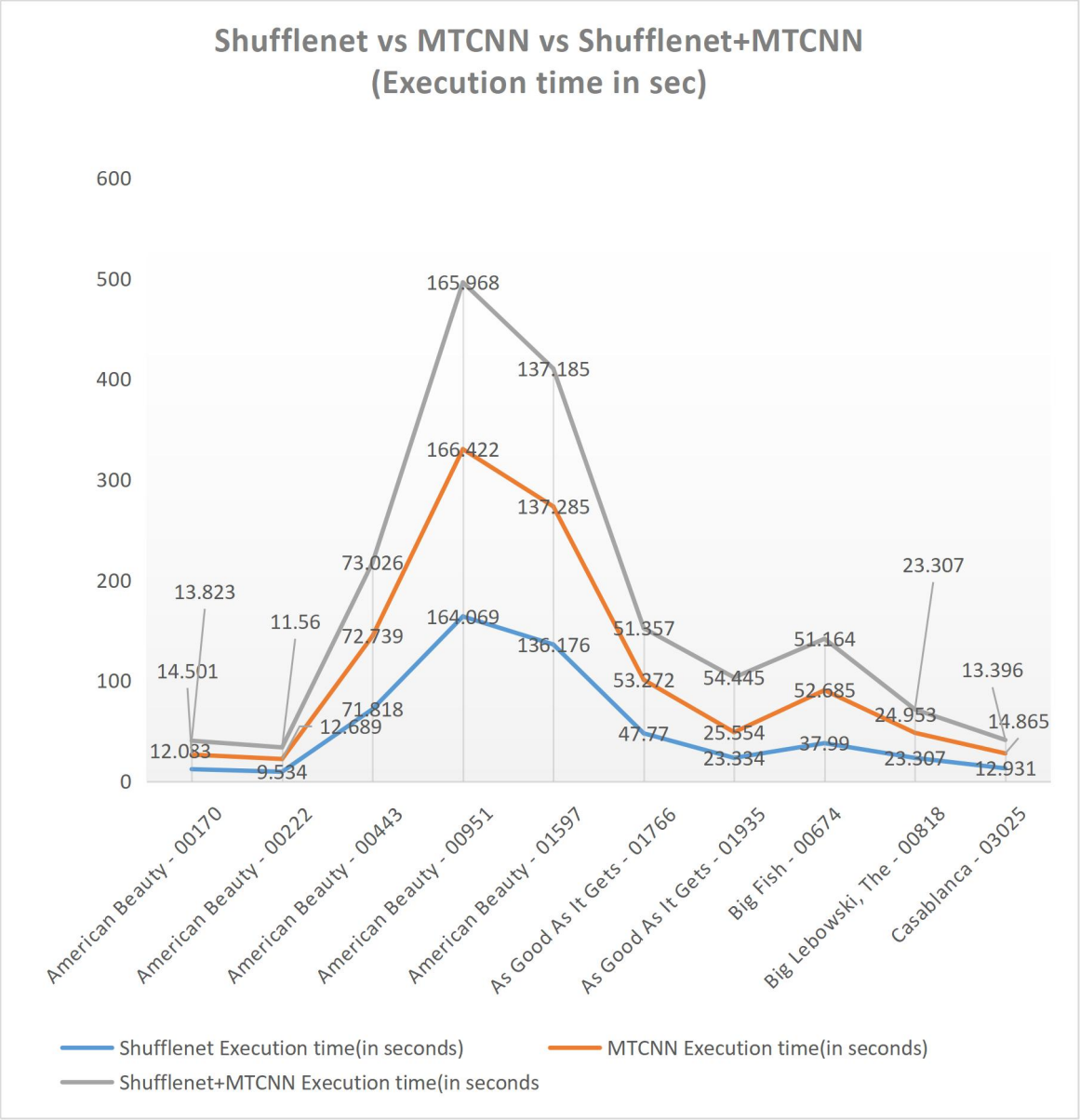


Figure 5.14: Shows a comparison of the execution times for face detection in ten distinct video clips from the Holly Wood video data set using the Shuffle Net, MTCNN, MTCCN, and ShuffleNet combination algorithms.

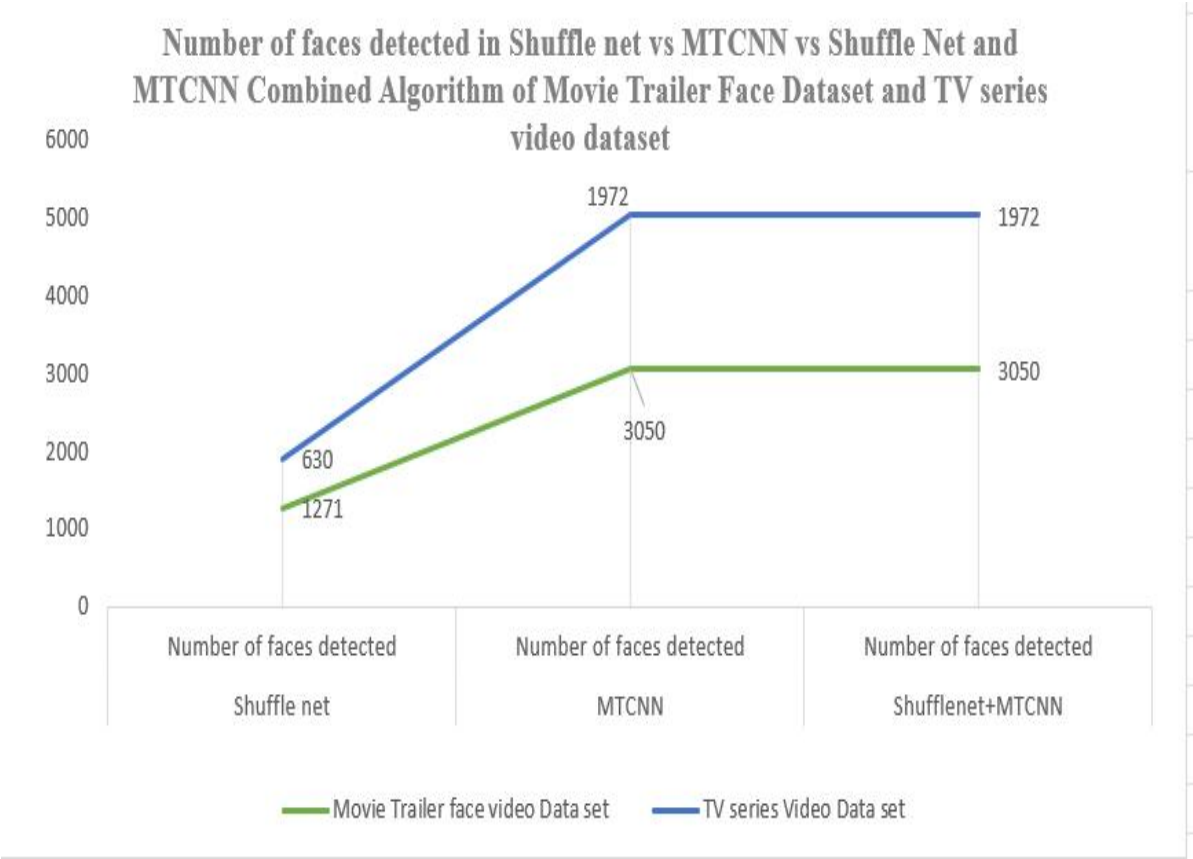


Figure 5.15: Shows a comparison graph between the Shuffle Net algorithm, MTCNN, and the Shuffle Net and MTCNN combined algorithm for face detection in a movie trailer and television series video data set based on the quantity of faces found.

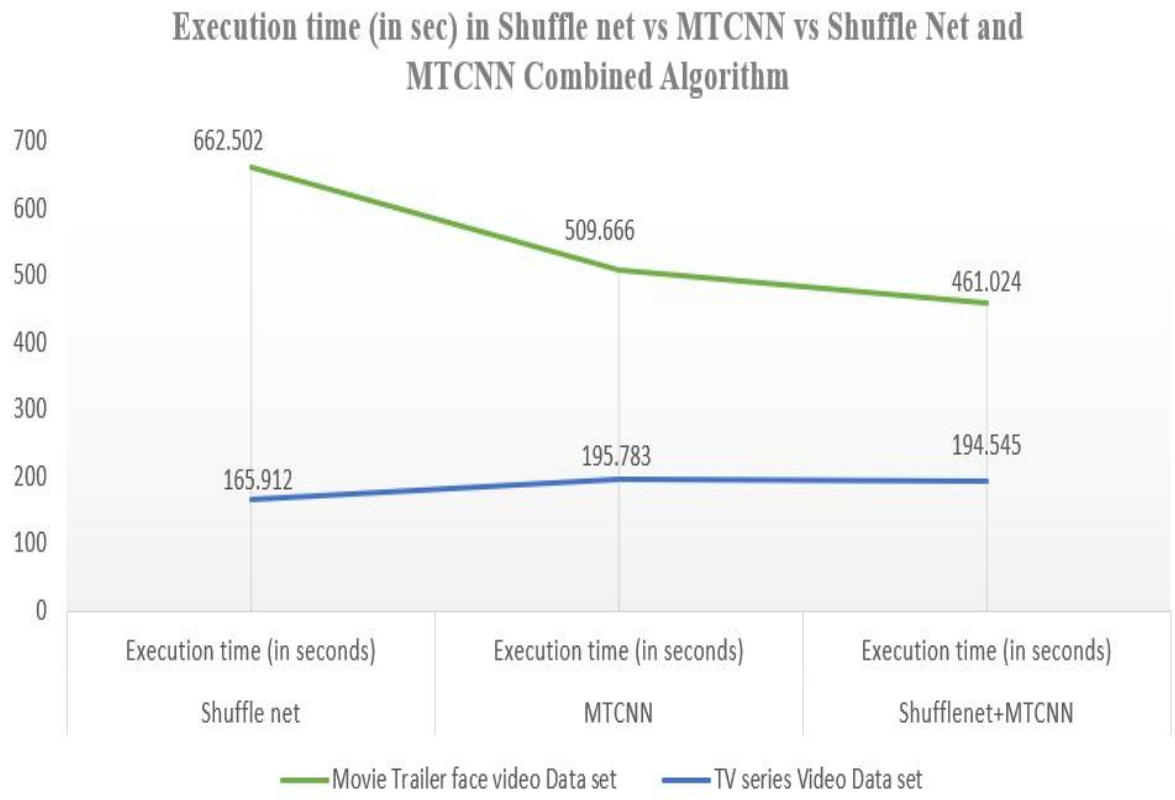


Figure 5.16 shows a graph comparing the execution times of the Shuffle Net algorithm, MTCNN, and the combined Shuffle Net and MTCNN algorithm for face detection in the movie trailer and TV video datasets.

Finally, the approach will be determined by the needs of the application. If accuracy is your primary concern, MTCNN is the best choice. Shuffle Net or the combination approach may be better if face detection depends more on computational efficiency. This comparison makes it clear that the optimal technique for face detection from the input video is the combination of the shuffle Net and MTCNN algorithms; that is why the proposed method is used. After face detection, facial recognition is performed using the Eigenface recognition technique, and the identified faces are then utilized for video indexing from the input videos.

5.3.3. Results of the method we suggested, "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8."

This subsection discussed the results of our proposed "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8" method. It is clear by the end of the experiment that YOLOv8n outperforms YOLO v5s in terms of accuracy. Metrics such as accuracy, execution time, the number of frames detected, and the number of essential frames produced are listed for several video clips from the Hollywood dataset (considering the 10 video datasets of Hollywood movies) in Table 5.5.

Table 5.5 presents the metrics for several videos from the Hollywood dataset, considering the Hollywood Movie 10 video dataset. Precision, execution time (second), number of frames discovered, and number of keyframes generated.

Video Name	YOLOv5s				YOLOv8n			
	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)
Being John Malkovich - 00719	87.2	127	102	89	97	27.2	102	99
As Good As It Gets - 01311	81.6	117	87	71	94.3	21	87	82
American Beauty - 02217	91.4	134.7	117	107	100	19.2	117	117
Big Fish - 01297	66.7	97	120	80	93	34	120	112
Big Lebowski, The - 00818	83	96.45	100	83	94	21	100	94
Casablanca - 00100	82.08	101	134	110	94.1	25.2	134	126
Crying Game, The – 00952	91.67	129.1	127	117	97.6	29	127	124
Dead Poets Society – 02590	87.67	137.8	130	114	96.2	22	130	125
Erin Brockovich – 00816	88.16	197	212	187	98	41.7	212	208
Fargo – 01189	94	139	129	121	100	23.2	129	129

Tables 5.6 and 5.7 discuss the accuracy, execution time, number of frames discovered, and number of keyframes detected for the Trailer and TV Series video data sets, respectively.

Table 5.6: Displays the trailer video data set's accuracy, execution time (second), number of frames found, and number of keyframes produced.

Video Name	YOLOv5s				YOLOv8n			
	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)
Titanic_trailer_02	19.3	4088.41	2808	542	88.89	509.83	2808	2496
Titanic_trailer_03	20.2	8479.11	4103	829	64.22	657.82	4103	2635
Titanic_trailer_04	34.22	12300.27	3498	1197	59.97	639.80	3498	2098

Table 5.7: Table 5.7 displays the TV Series Video Data Set's accuracy, execution time (second), number of frames found, and number of keyframes produced.

Video Name	YOLOv5s				YOLOv8n			
	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)	Acc. (%)	Exec. Time (S)	Frame Detected (No.)	Key Frame Generate (No)
TV Series Video Data Set	42.31	1173.8	780	330	52.7	194.7	1480	780

5.3.3.1. Strategies of Comparison

This subsection compares the YOLOv5s and YOLOv8n algorithms within the proposed approach. After finishing the face detection procedure, it is evident that YOLOv8n outperforms the other lightweight deep learning method covered in this research study, namely YOLOv5s, in terms of accuracy for the face detection ratio from the input video. It is clear from comparing the results that the YOLOv8n method produces the keyframe from the input video faster than the YOLOv5s approach. The little face may be identified by both algorithms from the input video. After using lightweight devices, YOLOv5s and YOLOv8n can recognize faces. YOLOv8n is faster at detecting small faces.

Figures 5.17, 5.18, and 5.19 show a bar graph that illustrates the main differences between the YOLO v5s and YOLOv8n algorithms. These bar graphs use the Holly Wood (Figure 5.17), Trailer (Figure 5.18), and TV Series (Figure 5.19) video files to evaluate the performance of

the aforementioned algorithms. The graph after testing displays the accuracy, execution time (in seconds), number of frames detected, and number of keyframes generated using the previously indicated algorithm.

In comparison to YOLOv5s, YOLOv8n is more accurate and less likely to detect faces. Additionally, both algorithms work well on lightweight devices and are more effective at recognizing a limited number of faces from input video. However, the mAP and FPS of YOLOv8n are better than those of the YOLOv5s algorithm. As a result, YOLOv8n has higher accuracy and recognizes more frames than YOLOv5s. Indexing utilizing human faces has the benefit of being faster and requiring less storage space. The human face can also be searched after using the input video's human face index.

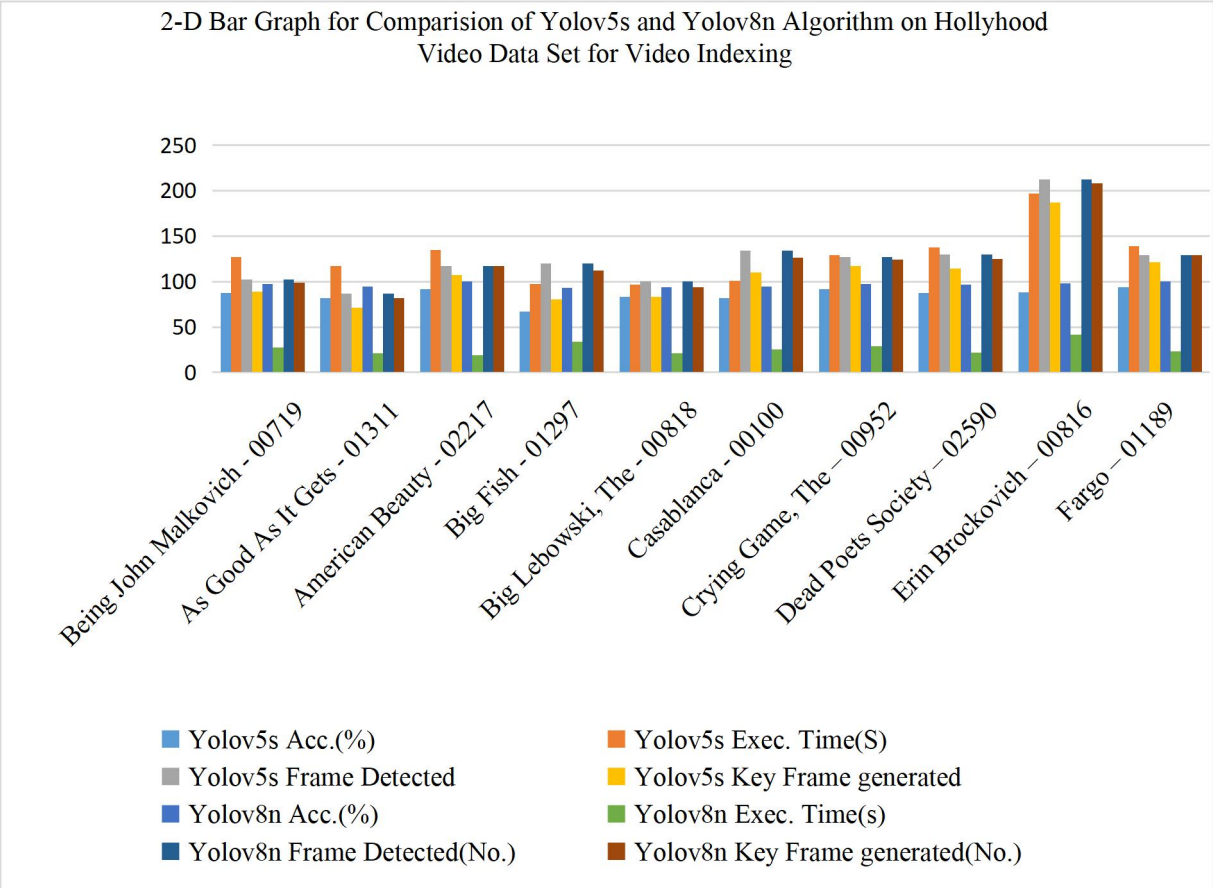


Figure 5.17: Shows a comparison of the accuracy (%), execution time, the total number of frames detected, and the total number of keyframes generated from Hollywood movie video datasets for the YOLOv5s and YOLOv8n algorithms.

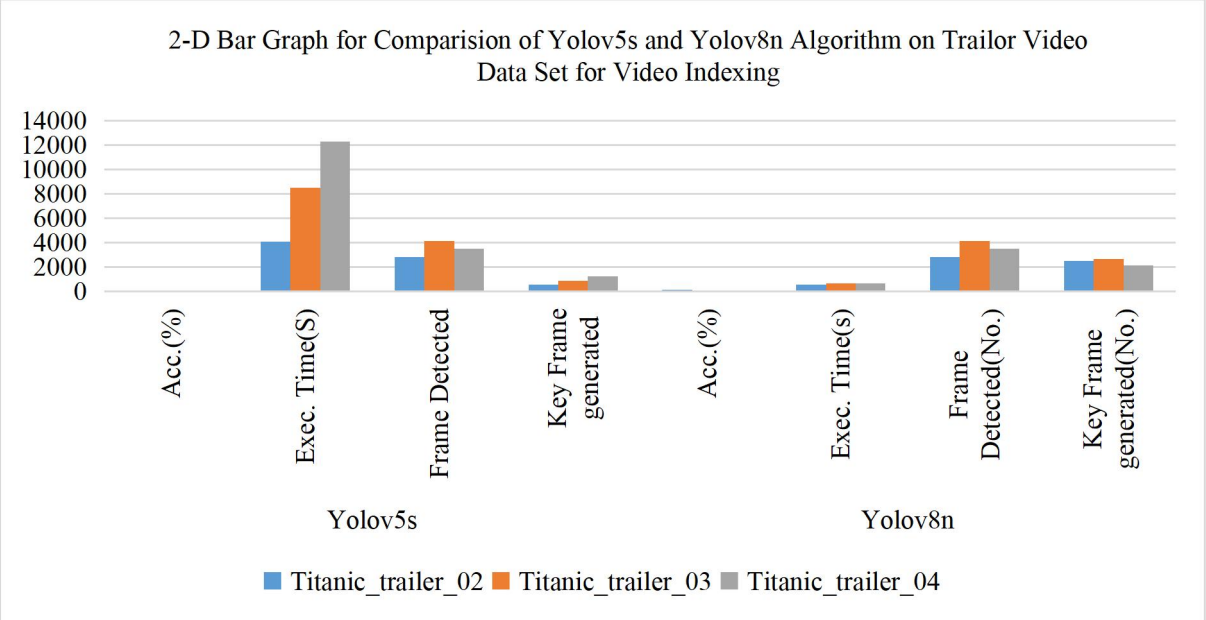


Figure 5.18: Shows a comparison of the accuracy (%), execution time, the total number of frames detected, and the total number of keyframes generated from trailer movie video datasets for the YOLOv5s and YOLOv8n algorithms.

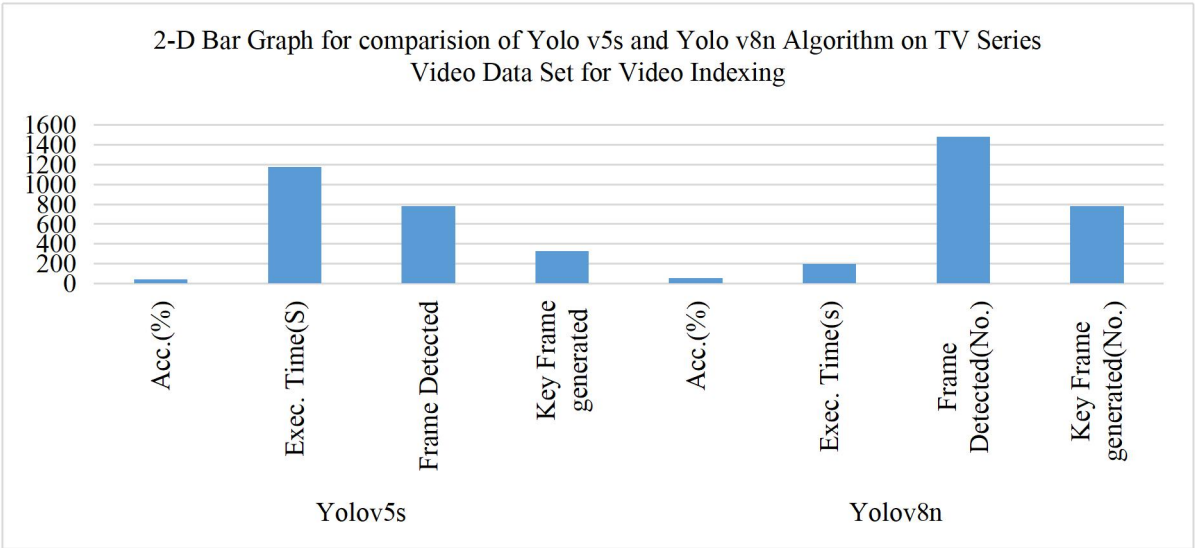


Figure 5.19: Comparison of the Accuracy (%), Execution Time, Total Number of Frames Detected, and Total Number of Key Frames Generated from TV Series Video Datasets of the YOLOv5s and YOLOv8n Algorithms.

5.3.4. Failure cases

With more angular and lighting-invariant faces in the input video, the suggested "Video indexing through the Face Images using Deep Learning Models" could not yield a 100% accurate result. This model offers low accuracy if the face size is very small and the facial image resolution is very low.MTCNN is the best option if accuracy is your top concern. If face detection relies more on computational efficiency, Shuffle Net or the combination approach can be preferable. The combined method provides a good compromise between processing economy and precision, making it appropriate for embedded and mobile systems.

However, the exclusion of small objects in the original Yolov5s object detection method results in low object detection accuracy.YOLO v8n might struggle to identify small items in face images. Minimal pixel-sized objects pose a problem since the model's receptive field may not be able to capture sufficient information, which could impact accuracy. The study's disadvantage is that these two approaches (YOLOv5s and YOLOv8n) cannot extract the human face from the input video if there is a significant directional change in the face. An additional problem in video indexing and retrieval using human faces as cues is the issue of overlapping face detection in the input video. Future study directions will be determined with consideration for this problem.

5.4. Conclusion

For video indexing by face identification, the suggested method, "Video Indexing through Human Faces by Combined Deep Learning Neural Networks," combines the Shuffle Net and MTCNN algorithms for face detection from the input video, as described in Section 5.2.1. In this section, Eigenfaces are used to enhance face recognition accuracy, while ShuffleNet and MTCNN are combined to improve face detection accuracy. Before face identification using the Eigenface approach and principal component analysis (PCA), faces are first detected using a combination of the Shuffle Net and MTCNN approaches. As a result, the work demonstrates how the two elements can be combined to create a comprehensive, trustworthy, and real-time face detection application. With a face recognition learning rate of 99.35%, the obtained results demonstrate that the proposed system can accurately identify individuals in real time and across various scenarios. The results are promising, as the quick processing time enables

the suggested method to be used on a device. The test utilizes a collection of videos from television shows, the Hollywood video dataset, and the Movie Trailer Face video dataset.

The proposed approach, "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8." uses the YOLOv5s and YOLOv8n algorithms for face detection from the input video, which is discussed in section 5.2.2, to index videos by face identification. Lightweight deep learning is utilized in image processing for video indexing and retrieval, reducing the time and space required to capture angular faces from films and increasing the storage capacity for key video frames. In lightweight gadgets, it helps in face detection. To overcome these issues, the YOLOv8n algorithm has been proposed as a face detector. The human face is used for video indexing.

Furthermore, the computation is straightforward, and this format is used to generate the instance of the invariant's illumination facial image. The suggested method increased storage capacity while simplifying time and space. Consequently, the proposed approach successfully enhances the techniques for video indexing and retrieval. The results of an experiment utilizing multiple datasets show that the YOLOv8n model performs better than the YOLOv5s model in terms of accuracy, execution time, number of frame detections, and number of key frame creations from input video. The test uses the Trailer Video, Hollywood Video, and TV Series Video databases. The video indexing technology can be used to facilitate personal search, authentication, and account verification. The disadvantage of the work is that these two techniques cannot extract a human face from an input video if there is a bigger directional change in the face. Another difficulty in video indexing and retrieval using human faces as cues is overlapping face detection from input footage. Future study directions will be determined with consideration for this problem.

Low object detection accuracy is the result of the original Yolov5s object detection method's exclusion of small objects. It may be difficult for YOLO v8n to recognize small objects in face images. The issue with minimal pixel-sized objects is that their accuracy may be impacted by the model's receptive field's inability to gather sufficient information. This issue will be taken into account while deciding on future research directions.

Chapter 6

Conclusion

This dissertation presents an evaluation and discussion of various methods for video indexing and retrieval, utilizing human faces as a cue. Numerous and varied applications of multimedia information systems have been made in both practical and research settings. As a result, multimedia content—especially videos—is also widely present in our daily lives. As an illustration, consider multimedia broadcasting, video conferencing, and distance learning. Nevertheless, gathering relevant video data and managing it with human effort is getting more challenging. When one individual's face video is provided, an application called Face Video Indexing and Retrieval can search across a video database for the video of that person. In recent years, face video retrieval has been a popular research topic and is applied in various fields. Video surveillance systems that can identify and track suspects from surveillance footage are one example.

Video sharing and searching have grown increasingly popular due to the quick development of social networking and semantic websites. Additionally, the entertainment sector, particularly the film industry, has seen a rise in demand for technologies that enable viewers to select specific actors or actresses in a film and directors to find the cast that best suits their productions. Consequently, the need for a technology that can effectively and automatically analyze video footage is critical and especially urgent. The human face has always been a fascinating subject in video content analysis. Video face recognition can be utilized to develop multimedia applications and tools that provide content-based access.

Image series processing can serve as a foundation for enhancing video processing. Keywords or descriptions are used in the majority of conventional image retrieval techniques, and the annotation words carry out the extraction process. However, this manual approach is costly and time-consuming. Accuracy has increased significantly recently, particularly in the area of

face recognition, due to the rapid development of deep learning and the widespread application of convolutional neural network (CNN) models in object recognition. Face

recognition remains a challenging issue due to the wide range of position variations, lighting conditions, occlusions, and facial expressions, even though many face recognition algorithms can accurately distinguish frontal faces with varying sizes, locations, and background images. Nowadays, the most popular source of entertainment and fun for Internet users is video. In a communication channel, it is also utilized as an inspiration for personal, commercial, and business purposes. Therefore, the bandwidth of the communication channel is also crucial for transmitting data from one device to another. More bandwidth, time, and space will be required if we transmit the human face as data across a communication channel. However, smaller faces with locations, overlapping, and background images are not detectable by many face detection systems. Multiple-frame video inputs offer redundant and expensive data.

The generation of accurate barcodes, QR codes, and image gradients from facial images is also a challenging task in video indexing and retrieval using human faces as cues. Using human faces as cues, EAN-8 linear barcodes, and QR codes, this dissertation primarily focuses on creating novel and effective methods for video indexing and retrieval from input videos. Finally, we have developed a system for video indexing and retrieval, taking the human face as a cue. Security, video surveillance networks, video channel descriptions, and other applications can benefit from this method. For the indexing and retrieval of video content, it is helpful. In the section to follow, we will in detail the summary of this dissertation and their future perspective.

6.1. Summary of the Dissertation

The dissertation begins with a brief overview of video representation, advanced methods for indexing and retrieving videos from video surveillance systems that use human faces as signals, segmenting video documents for indexing purposes, indexing videos from video documents using a variety of modalities, the importance of using human faces as cues in video indexing, and other topics (Chapter 1). This chapter has addressed the definition, analysis, and summary of videos in this video representation section. Three different modalities, or

information channels, have been demonstrated in this chapter for indexing videos from video documents. The first modality that uses the scene setting—whether it is created naturally or artificially—in the video recording is the visual approach. The second modality, acoustic mechanism, includes the ambient sounds, music, and voice that are audible in the document

video. Textual resources that explain the content of the video document are included in the third mode, known as documented communication. This chapter discusses various state-of-the-art methods for indexing and retrieving videos from video surveillance systems that utilize human faces as signals.

To analyze the video resources, video document segmentation is necessary for indexing purposes. We have also explored the importance of using human faces as cues in video indexing in great depth. In this thesis chapter, several state-of-the-art methods for indexing and retrieving videos from video surveillance systems that employ human faces as a signal are also covered. Lastly, chapter 1 provides a detailed description of the scope of this dissertation. An extensive review of several methods for video indexing and retrieval, utilizing human faces as cues, was conducted in Chapter 2. This chapter is broken up into seven smaller sections. The current methods for indexing and retrieving videos are covered in the first section of Chapter 2. The literature review on key frame extraction from input video is the second section of this chapter. The existing technique for face detection from key frames is the primary subject of the third section of this chapter. A survey of current methods for recognizing faces, including those for detecting faces, is included in the fourth section. The literature review on image gradient computation for face detection is the fifth section of this chapter. A study of current methods for creating linear facial bar codes from human faces is covered in the sixth section. The last section (seventh) of this chapter primarily discusses the methods currently used to generate QR codes from human faces. Each paper covered in this chapter has been examined in general terms concerning the dataset used, the number of samples in each dataset, performance evaluation, and critical comments on each article that address its main shortcomings or successes.

In Chapter 3 of this dissertation, a novel framework has been presented for video indexing through face images using Barcodes. A survey and analysis of several studies in Chapter 2 indicate that person detection is challenging when video indexing is performed using low-

level attributes. A person's facial expression, posture, mood, lighting variations, and occlusions are all important factors in video indexing using the human face. Moreover, the intrinsic uncertainties of video-based identification were observed, including changes in location, sensitivity to low resolution, and partial occlusion of facial features. From the perspective of indexing and storage space, all of these techniques are time- and space-intensive. We created a video indexing method called "Video indexing through human Face Images" to get over all of these problems. To identify faces in the video, this suggested method of video indexing by face recognition utilizes an EAN 8 linear bar code, as described in subsection 3.2.1.

This technique employs the color histogram method for keyframe detection and the Viola-Jones object detector for face detection. The EAN 8 bar code is used to index the face as a bar code, and the sliding window method is used to determine the picture gradient. This approach also considers occlusion, illumination variation, facial expressions, and minor shifts in face direction while generating an EAN 8 barcode from human faces in videos. It is a method that saves both time and space. The Hollywood video dataset, YouTube face video data collection, and TV series video set are all used in the test.

The proposed procedure for producing linear EAN-8 barcodes from images of faces is discussed in Section 3.2.2 and is based on the combined window and LGFA technique, gradient calculations, gradient directions, normalization, and quantization. The final step is to convert input faces to tags using EAN-8 linear standardized identifications. The facial datasets used in the test include Face94, YaleB, FERET, FG-NET, a composite face database (comprising a total of 20 faces), the NIR-VIS cropped dataset, and the LWF dataset. As expected, the test showed that the computation supplied confirms the idea of standardized names by virtue of a slight reflection of the main image when the plane, point, and apparent look vary.

In video indexing and retrieval, deep learning, and image processing are utilized to increase storage capacity by saving important frames from the movie and to decrease the time and space complexity of extracting angular faces from videos. Thus, to tackle these issues, a novel hybrid sliding window and LGFA technique, the EAN 8 barcode for use as a face index, and the Viola-Jones Algorithm for use as a face detector have been introduced (see Section 3.2.3).

This form is also used to improvise the illumination case of the computationally simple invariant face image. The recommended strategy also increased storage capacity while simplifying time and space. Thus, the proposed strategy effectively enhances video indexing and retrieval techniques. The test was conducted using one of three video datasets: Hollywood, YouTube, and TV series.

In image processing for video indexing and retrieval, deep learning, and machine learning are utilized to increase storage capacity for key movie frames and reduce the time and space required to learn angular faces from videos. Section 3.2.4 discusses the use of the linear EAN-8 bar code as a face index. To get over these issues, the MTCNN Algorithm has been proposed for use as a face detector. The scenario of illumination of the invariant facial picture

is also created using this format, and the computation is straightforward. EAN-8 linear barcodes finally indicate input faces. While decreasing the complexity of time and space, the proposed strategy enhanced storage capacity. This improves the video indexing and retrieval methods in the recommended manner. The test utilized the FDDB, LFW, WIDER FACE, and Hollywood Video datasets.

In Chapter 4, a novel method is proposed for Video indexing through face images using QR codes. This chapter introduces the concept of creating QR codes for facial images using the key frames of a video. In image processing for video indexing and retrieval, deep learning enhances the storage capacity for key frames from videos and reduces the time and space complexity of extracting angular faces from videos. To address these issues, the MTCNN Algorithm has been proposed as a face detector, and the QR code is utilized as a barcode to facilitate face indexing. Additionally, this form is used to generate the lighting situation of the invariant face picture, which is computationally simple. The suggested approach increased storage capacity while reducing time and space complexity. One benefit of a QR code is that it will continue to function correctly even if part of it is damaged or incomplete. A QR code is created once the input image (a face) has been scanned both horizontally and vertically.

The primary function of the MTCNN algorithm in this method is to recognize and align faces in angular keyframes. The suggested method significantly enhances video indexing and retrieval techniques. The test utilized the FDDB, LFW, WIDER FACE, and Hollywood Video

datasets. Using video indexing technology, an account's personal search, authentication, and affirmation can be specified. Applications for this technology include communication channel description, video surveillance, security, and human activity detection.

In Chapter 5, we presented two innovative techniques for video indexing using deep learning models and face images. As discussed in Section 5.2.1, the first method for video indexing by face identification is "Video Indexing through Human Faces by Combined Deep Learning Neural Networks," which combines the ShuffleNet and MTCNN algorithms for face detection from the input video. In this section, the accuracy of face detection is enhanced by combining ShuffleNet and MTCNN, while the accuracy of face recognition is improved using Eigenfaces. Initially, faces are detected by combining the Shuffle Net and MTCNN techniques, followed by face identification using the Eigenface approach and principal component analysis (PCA). Consequently, the work demonstrates how the two components can be integrated to produce a thorough, reliable, real-time face detection solution. The results indicate that the proposed system can recognize individuals in various situations and in real-time, with a face recognition

learning rate of 99.35%. The results are encouraging because the proposed method can be implemented on a device due to its short processing time. The Hollywood video dataset, the Movie Trailer face video dataset, and a selection of TV series video datasets were used in the test.

The YOLOv5s and YOLOv8n algorithms for face detection from the input video, which are covered in section 5.2.2, are used in the second suggested method, "Video indexing and retrieval through the human face as a cue using Light Weight Deep Learning algorithm YOLO v5 and YOLO v8." to index videos by face identification. In image processing, lightweight deep learning is utilized for video indexing and retrieval to enhance storage capacity for important video frames and reduce the time and space required to capture angular faces from videos. It facilitates facial recognition in portable devices. A face detector based on the YOLOv8n algorithm has been developed to address these problems. For video indexing, the human face is utilized.

Additionally, this format is utilized to construct the instance of the invariant's illumination face image, and the computation is simple. The proposed approach simplified time and space

while increasing storage capacity. As a result, the proposed method effectively improves video indexing and retrieval methods. The YOLOv8n model outperforms the YOLOv5s model in terms of accuracy, execution time, number of frame detections, and number of keyframe creations from input video, according to the findings of an experiment using multiple datasets. The Trailer Video, Hollywood Video, and TV Series Video databases are used in the exam.

The following is a summary of the thesis's most significant achievements to date.

- It is feasible to create a stable EAN-8 linear bar code from a face image in an input video using video indexing, where the human face serves as the bar code. With only slight changes in lighting, facial expressions, and facial posture, the device can operate with adequate accuracy.
- We can create a stable barcode with the aid of the sliding window technique.
- The technique should provide sufficiently accurate results, without making any distinctions based on an individual's age or race, after using the Viola-Jones face detection algorithm and the sliding window technique for stable barcode generation.
- The accuracy of angular face detection is improved by using the MTCNN algorithm.
- We can decrease the time and space complexity of video indexing by using an EAN-8 linear bar code, as it requires a smaller storage device than the original face image.
- For the linear EAN-8 bar code and QR code to be sent across a communication channel, a narrower bandwidth is needed.
- We can avoid the problem of loss of human face characteristics by employing QR codes for face representation in video indexing and retrieval, which use the human face as a cue. QR codes can scan input images in both vertical and horizontal orientations.
- We can identify smaller faces in an input video by combining the MTCNN and ShuffleNet algorithms. This technique improves processing power and facial detection accuracy.
- We can detect small faces in input videos from lightweight devices by combining the MTCNN and ShuffleNet algorithms.

- YOLOv5s and YOLOv8n, two lightweight deep learning algorithms, are used to detect small faces in input videos that have overlapping faces.
- The cost of computation is not much.
- Less power consumption is required using lightweight deep learning models.

6.2. Limitation

This section contains a detailed discussion of the limitations of the work offered in this dissertation.

- The automated method that uses barcodes to index videos based on face images is described in Chapter 3. The method's limitations include difficulties in detecting and creating a stable bar code from more angular, overlapping, small, and illumination-invariant face images of input videos. Another issue with this method is that human age affects the ability to create stable bar codes from face images.

- However, a number of factors, including variations in face position, brightness, and illumination, make it impossible for any of the algorithms to identify faces in movies. Furthermore, even if a detection system is effective, it lacks a clear image of faces.

- A linear bar code offers a succinct representation when horizontally scanning photographs of the face, but it loses a lot of information.

- An automated system for video indexing using a human face based on QR codes was discussed in Chapter 4, and it was tested on only three face video datasets (FDDB, LFW, and WIDER) and one video dataset (the Hollywood movie video dataset). However, the system should be tested on other complex video datasets, like the 3D face dataset.The proposed video indexing through the face images using a QR code might not produce a 100% accurate result for input videos with more angular and lighting-invariant faces. The MTCNN algorithm is often slower because the video's average frame rate was 8 frames per second. More time and processing power are needed for MTCNN. However, it takes longer to detect faces utilising lightweight real-time devices that use the MTCNN and Viola-Jones techniques. Furthermore, it has been discovered that MTCNN and Viola

Jones have trouble identifying little faces in videos. Scanning QR codes on damaged surfaces or in poorly light environments can be challenging.

- Chapter 5 discussed an automated method for video indexing that uses deep learning models and facial photos. In this chapter, we discovered that when the input video had more angular and lighting-invariant faces, the suggested "Video indexing through the Face Images using Deep Learning Models" could not generate a result that was 100% accurate. This model offers low accuracy if the face size is very small and the facial picture resolution is very low.

- The removal of small items in the original Yolov5s object detection approach leads to low object detection accuracy. YOLO v8n might have trouble identifying little items in face images. The problem with minimal pixel-sized objects is that the receptive field of the model may not be able to collect enough information, which could affect their accuracy. The two approaches (YOLOv5s and YOLOv8n) are unable to extract the human face from the input video if the face changes direction more than that.

6.3. Future Scope

In this dissertation, a thorough examination of the objectives and accomplishments is presented. The future directions of the work presented in this dissertation have been elaborately discussed in this part.

- Chapter 3 describes an automated method that utilizes barcodes to index videos based on face images. The development of a system that can produce a stable bar code from more angular, overlapping, tiny, and illumination-invariant face images of input videos is the future scope of this work. Human age affects the production of stable bar codes. This was observed in the FG-NET aging dataset. Therefore, this issue will also be a focus of our future efforts. This chapter covers the various techniques for using a person's face as a cue for video indexing.
- Nevertheless, none of the algorithms are successful in recognizing faces in videos due to various aspects, such as shifts in face orientation, brightness, and illumination. Moreover, if there is a detection system that works, it does not have a clear picture of faces. When

scanning images of the face horizontally, a linear bar code provides a concise representation, but it loses a significant amount of information. We are resolving some of the issues, but we are not achieving 100% success. In the future, we will work towards achieving more accurate results to fix all these issues.

- A QR code-based automated system for video indexing using a human face was covered in Chapter 4. Three face video datasets—FDDB, LFW, and WIDER—as well as one video dataset—the Hollywood movie video dataset—were used to evaluate this automated method. However, other complex video datasets, such as the 3D face dataset, should be used to test the system. For input videos with more angular and lighting-invariant faces, the suggested video indexing through the face photos utilizing a QR code could not yield a 100% accurate result. Since the average frame rate of the video was 8 frames per second, the MTCNN algorithm is frequently slower. MTCNN requires more processing power and time. However, face detection using lightweight real-time devices that employ the Viola-Jones and MTCNN methods takes longer. Moreover, Viola Jones and MTCNN have been found to struggle with recognizing little faces in videos. It can be difficult to scan QR codes on damaged surfaces or in dimly lit areas. Even if some of the problems in Chapter 5 have been fixed, any researcher can still concentrate on all of these problems in subsequent studies.

- An automated approach for video indexing using face images and deep learning models was covered in Chapter 5. In this chapter, we found that the proposed "Video indexing through the Face Images using Deep Learning Models" was unable to produce a 100% accurate result when the input video contained more angular and lighting-invariant faces. If the face size is very small and the resolution of the facial image is very poor, this model provides low accuracy. We can concentrate on all of these issues in subsequent work. Low object detection accuracy is the result of the original Yolov5s object detection method's exclusion of small items. It may be difficult for YOLO v8n to recognize little objects in face photos. The issue with minimal pixel-sized objects is that their accuracy may be impacted by the model's receptive field's inability to gather sufficient information. If there is a greater directional change in the face, the two methods (YOLOv5s and YOLOv8n) cannot extract the human face from the input video. This is the study's drawback. This issue will be taken into account when determining the path of future research.

In conclusion, it is worth noting that over the last few years, numerous advances have been made in the field of video indexing and retrieval, utilizing human faces as cues. In order to further investigate the topic of video indexing and retrieval, emerging researchers must yet overcome some obstacles.

References

[1] D.W. Oard. The state of the art in text filtering. User Modelling and User Adapted Interaction, 7(3):141–178, 1997.

[2] A. Hampapur, R. Jain, and T. Weymouth. Feature based digital video indexing. In IFIP 2.6 Third Working Conference on Visual Database Systems, Lausanne, Switzerland, 1995.

[3] Sanjoy Ghatak, Abhishek Pradhan, Dhiraj Khandelwal and Pema Lhamu Tamang, "News Video Indexing and Retrieval Using Combination of S.A.D and E.C.R Scoring Techniques" in International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 11, Page 370-373, November 2012.

[4] Tjondronegoro, D. W. (2005, May). Content-based Video Indexing for Sports Applications. 320.

[5] Wernicke, A. L. (2000). On the segmentation of text in videos. Multimedia and Expo, IEEE International.

[6] Cees G.M. Snoek, Marcel Worring, "Multimodal video indexing: A Review of state of the art," Multimedia Tools and Applications, 25, 5-35, 2005.

[7] G. Davenport, T. Aguierre Smith, and N. Pincever. Cinematic principles for multimedia. IEEE Computer Graphics & Applications, 11(4):67–74, 1991.

[8] Ghatak, S., Bhattacharjee, D. "A review of state-of-the-art in video indexing and retrieval using human faces as cues from video surveillance systems." Book chapter published in Book "Advances in Computer Science" Volume - 25", page 123-152, 2025

[9] J.M. Boggs and D.W. Petrie. The Art of Watching Films. Mayfield Publishing Company, Mountain View, U.S.A., 5th edition, 2000.

[10] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):4–37, 2000.

[11] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1):34– 58, 2002.

[12] J. Li *et al.*, "D.S.F.D.: Dual Shot Face Detector," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (C.V.P.R.)*, Long Beach, CA, U.S.A., 2019, pp. 5055-5064, doi: 10.1109/CVPR.2019.00520.

[13] Blaze face: Sub-millisecond neural face detection on mobile gpus V Bazarevsky, Y Kartynnik, A Vakunov, K Raveendran, M GrundmannarXiv preprint arXiv:1907.05047, 2019•arxiv.org.

[14] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. Computer Science, arXiv: 1804. 02767.http://arxiv.org/abs/1804.02767.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[16] X. Zhang, X. Zhou, M. Lin and J. Sun, "Shuffle Net: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, U.S.A., 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[17] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular Eigen spaces for face recognition. In IEEE International Conference on Computer Vision and Pattern Recognition, Seattle, U.S.A., 1994.

[18] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigen faces vs. fisher faces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711–720, 1997.

[19] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. IEEE Multimedia, 6(1):22–35, 1999.

[20] Caifeng Shan: Face Recognition and Retrieval in video. In Book Chapter on Video Search and Mining, Publisher Springer Berlin Heidelberg, Print ISBN: 978-3-642-12899-8 Electronic ISBN: 978-3-642-12900-1 SCI 287, pp. 235-260(2010).

[21] X. Jiwei, X. Jiyuan, F. Yi and C. Dongfang: "Research on video face retrieval method based on deep learning and key frame", *Proc. 4th Int. Conf. Digit. Signal Process.*, pp. 75-80, Jun. 2020.

[22] Cotsaces, N. Nikolaidis and I. Pitas, "Face-Based Digital Signatures for Video Retrieval," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 549-553, April 2008, doi: 10.1109/TCSVT.2008.918458.

[23] S. Eickeler, F. Wallhoff, U. Lurgel and G. Rigoll, "Content based indexing of images and video using face detection and recognition methods," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Salt Lake City, UT, U.S.A., 2001, pp. 1505-1508 vol.3, doi: 10.1109/ICASSP.2001.941217.

[24] Csaba Czirjek, Noel O'Connor, Sean Marlow, and Noel Murphy, "Face detection and clustering for video indexing applications" In ACIVS 2003 - Advanced Concepts for Intelligent Vision Systems, 2-5 September 2003.

[25] D. Cazzato, M. Leo, P. Carcagnì, C. Distante, J. Lorenzo-Navarro and H. Voos, "Video Indexing Using Face Appearance and Shot Transition Detection," *2019 IEEE/CVF International Conference on Computer Vision Workshop (I.C.C.V.W.)*, Seoul, Korea (South), 2019, pp. 2611-2618, doi: 10.1109/ICCVW.2019.00319.

[26] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognizing faces across pose and age. In International Conference on Automatic Face and Gesture Recognition, 2018.

[27] Kalirajan, K., Sudha, M.: Moving object detection for video surveillance. Sci. World J. 10, 8 (2015).

[28] Prasad, D.K., Rajan, D., Rachmawati, L., et al.: Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey. IEEE Trans. Intell. Transp. Syst. 18(8), 1993–2016 (2017).

[29] Yuan, Y., Xiong, Z., Wang, Q.: An incremental framework for video-based traffic sign detection, tracking, and recognition. IEEE Trans. Intell. Transp. Syst. 18(7), 1918–1929 (2017).

[30] YouTube During COVID-19. Accessed: Apr. 2, 2021. [Online]. Available:https://www.youtube.com/trends/articles/what-it-means-to-stayhome-on YouTube/

[31] Cisco Annual Internet Report (20182023) White Paper. Accessed: Apr. 2, 2021. [Online]. Available: https://www.cisco.com /c/en/us/solutions/collateral/executive-perspectives/annual-internet report/white-paper-c11-741490.html.

[32] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, A survey on visual content-based video indexing and retrieval, IEEE Trans. Syst., Man, Cybern., C (Appl. Rev.), vol. 41, no. 6, pp. 797819, Nov. 2011, doi: 10.1109/TSMCC.2011.2109710.

[33] N. Spolaor, H. D. Lee, W. S. R. Takaki, L. A. Ensina, C. S. R. Coy, and F. C. Wu, A systematic review on content-based video retrieval, Eng. Appl. Artif. Intell., vol. 90, Apr. 2020, Art. no. 103557, doi: 10.1016/j.engappai.2020.103557.

[34] D. Jain, S. Agrawal, S. Sengupta, P. De, B. Mitra, and S. Chakraborty, Prediction of quality degradation for mobile video streaming apps: A case study using YouTube, in Proc. 8th Int. Conf. Commun. Syst. Netw. (COMSNETS), Jan. 2016, pp. 12.

[35] G. Aceto, G. Bovenzi, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescape, Characterization and prediction of mobile-app traffic using Markov mod eling, IEEE Trans. Netw.ServiceManage.Vol.18, no.1, pp.907925, Mar. 2021, doi: 10.1109/TNSM.2021.3051381.

[36] H. Yoon and J. -H. Han, “Content-Based Video Retrieval with Prototypes of Deep Features," in *IEEE Access*, vol. 10, pp. 30730-30742, 2022, doi: 10.1109/ACCESS.2022.3160214.

[37] V. S. Subrahmanian, Principles of multimedia database systems. San Franciso, Calif.: Morgan Kaufmann Publishers, 1998.

[38] A. K. Elgmagarmid, H. Jiang, A. A. Helal, A. Joshi, and M. Admed, Video database systems: issues, products, and applications. Boston: Kluwer Academic Publishers, 1997.

[39] Chandran, R., Raman, N.: A review on video-based techniques for vehicle detection, tracking, and behaviour understanding. Int. J. Adv. Comp. Electr. Eng. 2(5), 7–13 (2017).

[40] G. J. Lu, Multimedia database management systems. Boston; London: Artech House, 1999.

[41] Proceedings. IEEE Workshop on, 1999. R. Tusch, H. Kosch, and L. Böszörményi, "VIDEX: an integrated generic video indexing approach," presented at The eighth ACM international conference on Multimedia, Marina del Rey, California, United States, 2000.

[42] C. Djeraba, "Content-based multimedia indexing and retrieval," Multimedia, IEEE, vol. 9, pp. 18-22, 2002.

[43] H. J. Zhang, "Content-based video browsing and retrieval," in Handbook of Internet and multimedia systems and applications, B. Furht, Ed.: CRC press LLC, 1999.

[44] R. M. Leonardi, P., "Semantic indexing of multimedia documents," Multimedia, IEEE, vol. 9, pp. 44-51, 2002.

[45] D. Ponceleon, S. Srinivasan, A. Amir, D. Petkovic, and D. Diklic, "Key to effective video retieval: Effective cataloging and browsing," presented at IEEE Internation Workshop on Content-based image and video databases, Bombay, India, 1998.

[46] L. Zeinik-Manor and M. Irani, "Event-based analysis of video," presented at Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, The Weizmann Institute of Science, 2001.

[47] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," Multimedia, IEEE Transactions on, vol. 4, pp. 68-75, 2002.

[48] A.Sasithradevi and S.Mohamed Mansoor Roomi," Video Classification and Retrieval through Spatio-Temporal Radon Features", Pattern Recognition, vol.99, pp.107099-107134, 2020.

[49] Nitin Janwe and Kishor Bhoyar," Semantic Concept based Video Retrieval using Convolutional Neural Network", SN Applied Sciences, vol.22, pp.80-88, 2019.

[50] D.Asha, Madhavee Lata and V.S.K. Reddy, "Content based Video Retrieval System using Multiple Features", International Journal of Pure and Applied Mathematics, vol.118, pp.287-294, 2018.

[51] Yang H, Meinel C. Content based lecture video retrieval using speech and video text information. IEEE Trans Learn Technol. 2014; 7(2):142– 154.

[52] Li K, Wang J, Wang H, et al. Structuring lecture videos by automatic projection screen localization and analysis. IEEE Trans Pattern Anal Mach Intell. 2015; 37(5):1233–1246.

[53] Nguyen NV, Coustaty M, Ogier JM. 2014 Multi-modal and cross-modal for lecture videos retrieval. Proceedings of IEEE International Conference on Pattern Recognition (ICPR), Stockholm, Sweden; 2014. p. 2667–2672.

[54] Gayathri N, Mahesh K (2020) Improved fuzzy-based SVM classification system using feature extraction for video indexing and retrieval. International Journal of Fuzzy Systems 22:1716–1729.

[55] Lin FC, Ngo HH, Dow CR (2020) A cloud-based face video retrieval system with deep learning. J Supercomput 76(11):8473–8493.

[56] Li C, Zhou B (2020) Fast key-frame image retrieval of intelligent city security video based on deep feature coding in high concurrent network environment. Journal of ambient intelligence and humanized computing 1-9.

[57] Jacob J, Sudheep Elayidom M, Devassia VP (2020) Video content analysis and retrieval system using video storytelling and indexing techniques. International Journal of Electrical & Computer Engineering 10(6):6019.

[58] G.G. Lakshmi Priya, S Domnic (2014) Shot based key frame extraction for ecological video indexing and retrieval. International Journal of Ecological Informatics ,Volume 23, September 2014, Pages 107-117.

[59] Markos, Z. (2014): Integrating motion and colour for content-based video classification. Int. J. Innov. Res. Comput. Commun. Eng. 2, 4.

[60] Ali, W.: Multimodal approach for video surveillance indexing and retrieval. Int. J. Comput. Appl. 64, 1 (2013).

[61] Dutta G (2021) Create caption by extracting features from image and video using deep learning model.

[62] Krishnaraj N, Elhoseny M, Lydia EL, Shankar K, and Aldabbas O (2020) An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in cloud environment. Software: Practice and Experience.

[63] L. Baraldi, C. Grana, R. Cucchiara, Neural story: an interactive multimedia system for Video indexing and re-use, in *Proceedings of CBIM*, Florence, Italy, June 19–21 (2017).

[64] B. C. Chen, Y.Y. Chen, Y.-H. Kuo, T.D. Ngo, D.-D. Le, S.I. Satoh, W.H. Hsu, Scalable face track retrieval in video archives using bag-of-faces sparse Representation. IEEE Trans. Circ. Syst. Video Technol. (2015).

[65] Zhen Dong, Su Jia, Tianfu Wu, and Mingtao Pei, "Face video Retrieval via Deep learning of binary hash Representations "Proceeding of the Thirtieth AAAI conference on Artificial Intel ligence (AAAI-16).

[66] Z. Mbarki, B. Miladi, C. J. Seddik, M. Fadhly, and H. Seddik, "Real-time face detection and identification from video sequences combining LBP algorithm and convolutional neural network," *2022 IEEE Information Technologies & Smart Industrial Systems (ITSIS)*, Paris, France, 2022, pp. 1-8, doi: 10.1109/ITSIS56166.2022.10118424.

[67] Li, X., Zhao, B., & Lu, X. (2018). Key frame extraction in the summary space. IEEE transactions on cybernetics, 48(6), 1923-1934.

[68] Bahroun S, Abed R, Zagrouba E (2020) KS-FQA: Key frame selection based on face quality assessment for efficient face recognition in video. IET Image Process 15:77–90.

[69] Gawande U, Hajari K, Golhar Y (2020) Deep learning approach to key frame detection in human action videos. Recent Trends in Computational Intelligence. IntechOpen

[70] Li C, Zhou B (2020) Fast key-frame image retrieval of intelligent city security video based on deep feature coding in high concurrent network environment. Journal of ambient intelligence and humanized computing 1-9.

[71] Chao G, Tsai Y, Jeng S. Augmented 3-D key frame extraction for surveillance videos. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). 2010;20(11):1395-1408.

[72] Kumar, K., Shrimankar, D. D., & Singh, N. (2017). Eratosthenes sieve based key-frame extraction technique for event summarization in videos. Multimedia Tools and Applications, 1-22.

[73] Kuanar, S. K., Panda, R., & Chowdhury, A. S. (2013). Video key frame extraction through dynamic Delaunay clustering with a structural constraint. Journal of Visual Communication and Image Representation, 24(7), 1212-1227.
[74] Usman Saeed, Jean-Luc Dugely, “Temporally consistent key frame selection from video for face recognition,” 18th European signal processing conference, 23-27th Aug.2010, IEEE Xplore, 30th April 2015.
[75] Dang C, Radha H. RPCA-KFE: Key frame extraction for video using robust principal component analysis. IEEE Transactions on Image Processing (TIP). 2015;24(11):3742-3753.
[76] Mentzelopoulos M, Psarrou A. Key frame extraction algorithm using entropy difference. In: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR; 15–16 October, 2004; New York, NY, USA. pp. 39-45
[77] Rasheed Z, Shah M. Detection and representation of scenes videos. IEEE Transactions on Multimedia. 2005;7(6):1097-1105.
[78] J Wu, S H Zhong, J Jiang, Y Yang A novel clustering method for static video summarization Multimedia Tools and Applications, volume 76, issue 7, p.9625-9641 Posted:2017.
[79] Mademlis, I., Tefas, A., & Pitas, I. (2018). A salient dictionary learning framework for activity video summarization via key-frame extraction. Information Sciences, 432, 319-331.
[80] Nasreen A, Roy K, Roy K, Shobha G. Key frame extraction and foreground modelling using K-means clustering. In: International Conference on Computational Intelligence, Communication Systems and Networks (CICSYN); Latvia; 2015. pp. 141-145.
[81] K. Wu, "Simple Implementations of Video Segmentation, Key Frame Extraction and Browsing," 2011.
[82] Zhang Q, Yu S-P, Zhou D-S, Wei X-P. An efficient method of key-frame extraction based on a cluster algorithm. Journal of Human Kinetics. 2013;39(1):5-13
[83] M. Asim, N. Almaadeed, S. Al-Máadeed, A. Bouridane, and A. Beghdadi, "A key frame based video summarization using colour features," in 2018 Colour and Visual Computing Symposium (CVCS), 2018, pp. 1-6: IEEE.
[84] C. Huang and H. Wang, "A Novel Key-frames Selection Framework for Comprehensive Video Summarization," IEEE Transactions on Circuits and Systems for Video Technology, 2020.

[85] V. Benni, R. Dinesh, P. Punitha, and V. Rao, "Key frame extraction and shot boundary detection using Eigen values," International Journal of Information Electronics Engineering, vol. 5, no. 1, p. 40, 2015.

[86] Zheng R, Yao C, Jin H, Zhu L, Zhang Q, et al. (2015) Parallel Key Frame Extraction for Surveillance Video Service in a Smart City. PLOSONE10(8): e0135694. https://doi.org/10.1371/journal.pone.0135694

[87] Viola, P., Jones, M.J. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 137–154 (2004). https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[88] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng and S. Avidan, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, 2006, pp. 1491-1498, doi: 10.1109/CVPR.2006.119.

[89] M. -T. Pham, Y. Gao, V. -D. D. Hoang and T. -J. Cham, "Fast polygonal integration and its application in extending haar-like features to improve object detection," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 942-949, doi: 10.1109/CVPR.2010.5540117.

[90] X. Zhu, and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879-2886.

[91] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for multi-view face detection,” in IEEE International Joint Conference on Biometrics, 2014, pp. 1-8.

[92] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in European Conference on Computer Vision, 2014, pp. 720-735.

[93] J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2497-2504.

[94] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.

[95] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in European Conference on Computer Vision, 2014, pp. 109-122.

[96] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.

[97] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 5325-5334, doi: 10.1109/CVPR.2015.7299170.

[98] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision, 2014, pp. 94-108.

[99] C. Zhang, and Z. Zhang, “Improving multiview face detection with multi-task deep convolutional neural networks,” IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036-1041.

[100] Sign Modou Bah, Fang Ming IEEE Conference on Computer Vision and Pattern Recognition, « An improved face recognition algorithm and its application in attendance management system» Array, vol. 5, (2020), p 100014.

[101] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramid box: A context-assisted single-shot face detector. In Proceedings of European Conference on Computer Vision (ECCV), 2018. 2, 3, 5

[102] Jialiang Zhang, Xiongwei Wu, Jianke Zhu, and Steven CH Hoi. Feature agglomeration networks for single-stage face detection. arXiv preprint arXiv:1712.00721, 2017. 2, 3

[103] Wei J, Liu G, Liu S, Xiao Z. A novel algorithm for small object detection based on YOLOv4. Peer J Comput Sci. 2023 Mar 22;9: e1314. doi: 10.7717/peerj-cs.1314. PMID: 37346537; PMCID: PMC10280595.

[104] Qi, D., Tan, W., Yao, Q., Liu, J. (2023). YOLO5Face: Why Reinventing a Face Detector. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds) Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol. 13805. Springer, Cham. https://doi.org/10.1007/978-3-031-25072-9_15.

[105] Sirisha, U., Praveen, S.P., Srinivasu, P.N. et al. Statistical Analysis of Design Aspects of Various YOLO-Based Deep Learning Models for Object Detection. Int J Comput Intell Syst 16, 126 (2023). https://doi.org/10.1007/s44196-023-00302-w.

[106] Lenc L, Král P. Local binary pattern based face recognition with automatically detected fiducial points. *Integrated Computer-Aided Engineering*. 2016;23(2):129-139. doi:10.3233/ICA-150506

[107] Fenggao Tang et al., « An end-to-end face recognition method with alignment learning, » Optik - International Journal for Light and Electron Optics, vol 205, (2020), p 164238.
[108] I. Masi, S. Rawls, G. Medioni," Pose-aware face recognition in the wild," Conference on Computer Vision and Pattern Recognition (2016), pp 4838–4846.
[109] S. Liao, A.K. Jain, S.Z. Li, "Partial face recognition: alignment-free approach," IEEE Trans. Pattern Anal. Mach. Intel. Vol 35 (5), (2013), pp 1193–1205.
[110] Chen Y.C, Patel V.M, Shekhar S, Chellappa R and Phillips P.J "Video-based face recognition via sparse joint representation," In FG,1-8, IEEE, 2013.
[111] Shaohua Zhou, * Volker Krueger, and Rama Chellappa, "Probabilistic recognition of human face from the video, "Computer Vision and Image Understanding 91 (2003) 214–245.
[112] Forczmanski, P., Kukharev, G., Shchegoleva, N.: An algorithm of face recognition under difficulty lighting conditions. Electr. Rev. (10 b), 201–204 (2012).
[113] Roy, H., Bhattacharjee, D.: Local-Gravity-Face (LG-face) for illumination–invariant and heterogeneous face recognition. IEEE Trans. Inf. Forensics Secur. **11**(7), 1412–1424 (2016).
[114] Kukharev, G., Matveev, Y., Shchegoleva, N.: Barcode generation for face images. Data Anal. Intellect. Syst. Bus. Inf. 3(29), 201.
[115] Carcagnì P, Del Coco M, Leo M, Distante C. Facial expression recognition and histograms of oriented gradients: a comprehensive study. Springer plus. 2015 Oct 26; 4:645. doi: 10.1186/s40064-015-1427-3. PMID: 26543779; PMCID: PMC4628009.
[116] Dakin, S. C., & Watt, R. J. (2009). Biological "bar codes" in human faces. Journal of Vision, 9(4):2, 1–10, http://journalofvision.org/9/4/2/, doi:10.1167/9.4.2.
[117] Matveev, Y., Kukharev, G., Shchegoleva, N.: A simple method for generating facial Barcodes. In: WSCG2014 Conference on Computer Graphics, Visualization and Computer Vision in Co-operation with EUROGRAPHICS Association Exchange Anisotropy, pp. 213–220. Academic, Czech Republic (2014).
[118] S. Ghatak, Facial representation using linear barcode, in *Advanced Computational and Communication Paradigms*, vol. 2, pp.791–801 (2018).
[119] Ghatak, S., Bhattacharjee, D. (2020). Barcode Representation of Face Image Combining LGFA and Windowing Technique. In: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds)

Human Interaction and Emerging Technologies. IHIET 2019. Advances in Intelligent Systems and Computing, vol 1018. Springer, Cham. https://doi.org/10.1007/978-3-030-25629-6_73.

[120] S. Tiwari, "An Introduction to QR Code Technology," *2016 International Conference on Information Technology (ICIT)*, Bhubaneswar, India, 2016, pp. 39-44, doi: 10.1109/ICIT.2016.021.

[121] Linglong Tan et al. 2021 *J. Phys.: Conf. Ser.* 1944 012020**DOI** 10.1088/1742-6596/1944/1/012020.

[122] Kratochvíl, Miroslav, et al. (2020) Som-hunter: video browsing with relevance-to-som feedback loop. International conference on multimedia modeling. Springer, Cham, SOM-Hunter: Video Browsing with Relevance-to-SOM Feedback Loop.

[123] Rossetto L, Gasser R, Lokoc J, Bailer W, Schoeffmann K, Muenzer B, Soucek T, Nguyen PA, Bolettieri P, Leibetseder A, Vrochidis S (2020) Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Transactions on Multimedia 23:243–256.

[124] Sauter L et al. (2020) Combining boolean and multimedia retrieval in vitrivr for large-scale video search. International conference on multimedia modeling. Springer, Cham, Combining Boolean and Multimedia Retrieval in vitrivr for Large-Scale Video Search.

[125] Ji LY, Yang Z (2017) Design and implementation of medication recommending system for chronic hepatitis B. Chinese Medical Equipment Journal 38(7):48–51

[126] Li D, Liao X, Xiang T, Wu J, Le J (2020) Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. Computers & Security 90:101701

[127] Yürekli A, Bilge A, Kaleli C (2021) Exploring playlist titles for cold-start music recommendation: an effectiveness analysis. Journal of Ambient Intelligence and Humanized Computing 1–20

[128] Adeyemo J, Oyebode O, Stretch D (2018) River flow forecasting using an improved artificial neural network. EVOLVE-A Bridge Between Probability, Set Oriented Numerics, and Evolutionary Computation VI. Springer, Cham 179–193

[129] Saritha RR, Paul V, Kumar PG (2019) Content based image retrieval using deep learning process. Clust Comput 22(2):4187–4200

[130] Ghatak S, Bhattacharjee D (2020) “Video indexing through human face”, present the paper in third international conference on communications, circuits and systems held at school of electronics engineering. Kalinga Institute of Industrial Technology, Bhubaneswar, Video Indexing Through Human Face.

[131] Hoy MB (2018) Deep learning and online video: advances in transcription, automated indexing, and manipulation. Medical reference services quarterly 37(3):300–305
[132] Zhang C, Lin Y, Zhu L, Liu A, Zhang Z, Huang F (2019) CNN-VWII: an efficient approach for large-scale video retrieval by image queries. Pattern Recogn Lett 123:82–88
[133] Tian H, Tao Y, Pouyanfar S, Chen SC, Shyu ML (2019) Multimodal deep representation learning for video classification. World Wide Web 22(3):1325–1341
[134] Dong Z, Wei J, Chen X, Zheng P (2020) Face detection in security monitoring based on artificial intelligence video retrieval technology. IEEE Access 8:63421–63433
[135] Ullah A, Muhammad K, Hussain T, Baik SW, De Albuquerque VHC (2020) Event-oriented 3d convolutional features selection and hash codes generation using PCA for video retrieval. IEEE Access 8:196529–196540
[136] Deng Y, Yu Y (2019) Self-feedback image retrieval algorithm based on annular color moments. EURASIP Journal on Image and Video Processing 2019(1):1–13
[137] Yan C, Gong B, Wei Y, Gao Y (2020) Deep multi-view enhancement hashing for image retrieval. IEEE Trans Pattern Anal Mach Intell 43:1445–1451
[138] Kumar GN, Reddy VSK (2019) Key frame extraction using rough set theory for video retrieval. In Soft Computing and Signal Processing:751–757
[139] Matveev, Y.N.: Technologies of biometric identification of a person by voice and other modalities. Vestnik MGTU.Priborostroenie, Special Issue “Biometric Technologies”, pp. 46–61 (2012). (in Russian)
[140] Wang, XY., Wu, JF. & Yang, HY. Robust image retrieval based on colour histogram of local feature regions. Multimedia Tools Appl 49, 323–345 (2010). https://doi.org/10.1007/s11042-009-0362-0
[141] Paul Viola and Michael Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” Accepted conference on Computer Vision and Pattern Recognition 2001.
[142] Barcode generation algorithm for notes.spb.ru/barcode_ean8.htm (15.03.2014).
[143] S. Kundu, “Gravitational clustering: A new approach based on the spatial distribution of the points,” Pattern Recognition, vol.32, no. 7, pp. 1149-1160, 1999.
[144] Ram Zheng, Chuanwei Yao, Hui Jin, Lei Zhu, Qin Zhang, and Wei Deng, “Parallel key frame extraction surveillance video service in a Smart City,” PloSOne,10(8), e0135694, doi:10.137/Journal pone. 0135694.

[145] T. Choudhury, B. Clarkson, T. Jebara, A. Pentland, Multimodal person recognition using unconstrained audio and video, in Proceedings of International Conference on Audio- and Video-Based Person Authentication, 1999, pp. 176–181.

[146] Eyuphan Bulut &Tolga Capin, “Key Frame Extraction from Motion Capture Data by Curve Saliency.”

[147] Haiyan Xie,” Key Frame Segmentation in Video Sequences,” 2008.

[148] R.P. Olenick, T.M. Apostol, and D.L.Goodstein, The Mechanical Universe: Mechanics and Heat. New York, NY, USA: Cambridge Univ. Press,1985.

[149] Ghatak, S., Bhattacharjee, D. Video indexing through human face images using LGFA and window technique. Multimedia Tools https://doi.org/10.1007/s11042-022-12965-2.

[150] Ghatak, S., Kollman, C., Bhattacharjee, D. (2024). Video Indexing Through QR Code of Human Faces Using MTCNN Algorithm. In: Das, N., Khan, A.K., Mandal, S., Krejcar, O., Bhattacharjee, D. (eds) Proceedings of International Conference on Data, Electronics and Computing. ICDEC 2023. Lecture Notes in Networks and Systems, vol. 1103. Springer, Singapore. https://doi.org/10.1007/978-981-97-6489-1_1

[151] Roeland De Geest, EfstratiosGavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, Tinne Tuytelaars, “Online Action Detection,” arXiv: 1604.06506v2[cs.CV] 30 Aug.2016.

[152] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587756.

[153] L. Wolf, T. Hassner and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," *CVPR 2011*, Colorado Springs, CO, USA, 2011, pp. 529-534, doi: 10.1109/CVPR.2011.5995566.

[154] Face94 database:http://cswww.essex.ac.uk/mvallfaces/face94.html.

[155] http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

[156] CHUK face Sketch FERET database:http://mmlab.ie.cuhk.edu.hk/cufsf.

[157] FG-NET Aging database”, http://www.fgnet.rsunit.com,2010.

[158] S. Z. Li, D. Yi, Z. Lei, and S. Liao, “The Casia NIR-VIS 2.0 face database,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 348–353.

[159] In Advances in Face Detection and Facial Image Analysis, edited by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, Springer, pages 189-248, 2016.

[160] D. Yi, Z. Lei, and S. Z. Li. “A robust eye localization method for low quality face images”. In International Joint Conference on Biometrics (IJCB), pages 15–21, Washington, DC, USA, Oct. 11-13 2011.
[161] S. Yang, P. Luo, C.-C. Loy, and X. Tang. W wider face: A face detection benchmark. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
[162] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
[163] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097-1105
[164] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, 2001
[165] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, “Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model,” in IEEE International Conference on Computer Vision, 2013, pp. 1944-1951.
[166] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in IEEE International Conference on Computer Vision, 2013, pp. 1513-1520.
[167] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” International Journal of Computer Vision, vol 107, no. 2, pp. 177-190, 2012.
[168] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment,” in European Conference on Computer Vision, 2014, pp. 1- 16.N
[169] Information Technology-Automatic Identification and Data Capture Techniques-QR code Bar code symbology specification (Adopted ISO/IEC 18004: 2015, Third Edition, 201502- 01)
[170] Heming Zhang et al. "Fast face detection on mobile devices by leveraging global and local facial characteristics" Signal Processing: Image Communication, vol78, (2019), pp1–8.
[171] R. Wang, B. Fang, Affective computing and biometrics-based HCI surveillance system, in Proceedings of the International Symposium on Information Science and Engineering, 2008, pp. 192–195.
[172] W. Weiguo, M. Qingmei, W. Yu, Development of the humanoid head portrait robot system with flexible face and expression, in Proceedings of the 2004 IEEE International

Conference on Robotics and Biomimetic, 2004, pp. 757–762, doi: 10.1109/ROBIO.2004.1521877.

[173] M.H. Su, C.H. Wu, K.Y. Huang, Q.B. Hong, H.M. Wang, Exploring microscopic fluctuation of facial expression for mood disorder classification, in Proceedings of the International Conference on Orange Technologies, 2017, pp. 65–69.

[174] M.B. Mariappan, M. Suk, B. Prabhakaran, Face fetch: a user emotion driven multimedia content recommendation system based on facial expression recognition, Proceedings of the 2012 IEEE International Symposium on Multimedia (2012) 84–87.

[175] S.A. Patil, P.J. Deore, Local binary pattern based face recognition system for automotive security, in Proceedings of the International Conference on Signal Processing, Computing, and Control, 2016, pp. 13–17.

[176] Peng Wang, Lingqiao Liu, Chunhua Shen, Heng Tao Shen, "Order-aware convolutional pooling for video based action recognition", Pattern Recognition, Volume91,2019, Pages357-365, ISSN0031-3203, https://doi.org/10.1016/j.patcog.2019.03.002.

[177] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, Yisheng Zhong, "Person re-identification by unsupervised video matching", Pattern Recognition, Volume 65,2017, Pages 197-210, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2016.11.018.

[178] Jingke Meng, Ancong Wu, Wei-Shi Zheng, "Deep asymmetric video-based person re-identification", Pattern Recognition, Volume 93,2019, Pages 430-441, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.04.008.

[179] Zhen Dong, Chenchen Jing, Mingtao Pei, Yunde Jia, Deep CNN based binary hash video representations for face retrieval, Pattern Recognition, Volume 81,2018, Pages 357-369, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2018.04.014.

[180] Yafeng Li, Wavelet-based fuzzy multiphase image segmentation method, Pattern Recognition Letters, Volume 53,2015, Pages 1-8, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2014.10.013.

[181] Song, Jingkuan, Yang, Yi, Huang, Zi, Shen, Heng Tao, and Luo, Jiebo (2013). Effective multiple feature hashing for large-scale near-duplicate video retrieval. IEEE Transactions on Multimedia 15 (8) 6553136 1997,2008, https://doi.org/10.1109/TMM.2013.2271746.

[182] Bochkovskiy, Alexey, Chien-YaoWang, and HongYuanMarkLiao. "YOLOv4: Optimal Speed and Accuracy of Object Detection (link is external)." arXiv:2004.10934 [cs.CV] (2020).

[183] Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature Toward Digital Manufacturing and Industrial Defect Detection. Machines 2023, 11, 677. https://doi.org/10.3390/machines11070677.

[184] E. G. Ortiz, A. Wright and M. Shah, "Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-Based Classification," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 3531-3538, doi: 10.1109/CVPR.2013.453.

PhD Thesis v2

ORIGINALITY REPORT

8%

SIMILARITY INDEX

PRIMARY SOURCES

1	ebin.pub Internet	810 words — 1%
2	www.researchgate.net Internet	226 words — < 1%
3	link.springer.com Internet	219 words — < 1%
4	staff.science.uva.nl Internet	194 words — < 1%
5	smu.edu.in Internet	136 words — < 1%
6	technodocbox.com Internet	125 words — < 1%
7	www.jaduniv.edu.in Internet	120 words — < 1%
8	arxiv.org Internet	110 words — < 1%
9	actorsfit.com Internet	91 words — < 1%
10	hdl.handle.net Internet	

Sanjoy Ghosh

Bhatta

85 words — < 1%

11	springerplus.springeropen.com Internet	75 words — < 1%
12	www.academiafa.edu.pt Internet	68 words — < 1%
13	"Cognitive Computing – ICCC 2019", Springer Science and Business Media LLC, 2019 Crossref	65 words — < 1%
14	pubmed.ncbi.nlm.nih.gov Internet	64 words — < 1%
15	zhzhanp.github.io Internet	63 words — < 1%
16	"Computer Vision – ECCV 2018", Springer Nature America, Inc, 2018 Crossref	60 words — < 1%
17	Jiangshu Wei, Gang Liu, Siqi Liu, Zeyan Xiao. "A novel algorithm for small object detection based on YOLOv4", PeerJ Computer Science, 2023 Crossref	59 words — < 1%
18	Chuhong Li, Bo Zhou. "Fast key-frame image retrieval of intelligent city security video based on deep feature coding in high concurrent network environment", Journal of Ambient Intelligence and Humanized Computing, 2020 Crossref	58 words — < 1%
19	juew.org Internet	58 words — < 1%

20	Qiang Zhang, Shao-Pei Yu, Dong-Sheng Zhou, Xiao-Peng Wei. "An Efficient Method of Key-Frame Extraction Based on a Cluster Algorithm", Journal of Human Kinetics, 2013 Crossref	55 words — < 1%
21	Shishi Qiao, Ruiping Wang, Shiguang Shan, Xilin Chen. "Deep video code for efficient face video retrieval", Pattern Recognition, 2021 Crossref	53 words — < 1%
22	www.semanticscholar.org Internet	52 words — < 1%
23	livros01.livrosgratis.com.br Internet	49 words — < 1%
24	personales.upv.es Internet	45 words — < 1%
25	www.akinik.com Internet	44 words — < 1%
26	Ladislav Lenc, Pavel Král. "Local binary pattern based face recognition with automatically detected fiducial points", Integrated Computer-Aided Engineering, 2016 Crossref	42 words — < 1%
27	Studies in Computational Intelligence, 2010. Crossref	42 words — < 1%
28	assets.researchsquare.com Internet	41 words — < 1%
29	Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun. "ShuffleNet: An Extremely Efficient	38 words — < 1%

Convolutional Neural Network for Mobile Devices", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018
Crossref

30 Geetha Rani E, Mounika E, Gopala Krisnan C, Tanuep Bellam, Bhuvaneswari P, Kanagavalli Rengaraju. "Comparative Analysis of Deepfake Video Detection Using Inception Net and Efficient Net", 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2022 — 37 words — < 1%
Crossref

31 coek.info — 37 words — < 1%
Internet

32 Bandyopadhyay, Ambar. "Neural Learning: Can we Make it a Little More Bio-Inspired!", Indian Statistical Institute - Kolkata, 2021 — 36 words — < 1%
ProQuest

33 www.bose.res.in — 34 words — < 1%
Internet

34 Fenggao Tang, Xuedong Wu, Zhiyu Zhu, Zhengang Wan, Yanchao Chang, Zhaoping Du, Lili Gu. "An end-to-end face recognition method with alignment learning", Optik, 2020 — 33 words — < 1%
Crossref

35 Siriki Atchuta Bhavani, C. Karthikeyan. "An attention based deep learning with effective SVM-ConvFaceNeXt model for face recognition in unconstrained environment", Signal, Image and Video Processing, 2025 — 33 words — < 1%
Crossref

36	Biswas, Utpal. "Some Studies on Wavelength Establishment Algorithms for All Optical Networks.", Jadavpur University (India), 2020 ProQuest	32 words — < 1%
37	http://195.170.12.01/DAEI/PRODUCTS/Informtc/Ovire/Ovire.htm Internet	32 words — < 1%
38	Liu, Yuguang. "Scale-Aware Multi-Path Deep Neural Networks for Unconstrained Face Detection.", McGill University (Canada), 2021 ProQuest	31 words — < 1%
39	R. J. Poovaraghan, P. Prabhavathy. "Chapter 47 Video Indexing and Retrieval Techniques: A Review", Springer Science and Business Media LLC, 2023 Crossref	31 words — < 1%
40	www.ijert.org Internet	30 words — < 1%
41	Wan En Ng, Muhammad Syafiq Mohd Pozi, Mohd Hasbullah Omar, Norliza Katuk, Abdul Rafiez Abdul Raziff. "Chapter 16 A Video Summarization Method for Movie Trailer-Genre Classification Based on Emotion Analysis", Springer Science and Business Media LLC, 2024 Crossref	29 words — < 1%
42	doras.dcu.ie Internet	29 words — < 1%
43	www.ijraset.com Internet	29 words — < 1%

44	Bin Jiang, Hongbin Jiang, Huanlong Zhang, Qiuwen Zhang, Zuhe Li, Lixun Huang. "4AC-YOLOv5: an improved algorithm for small target face detection", EURASIP Journal on Image and Video Processing, 2024 Crossref	28 words — < 1%
45	Jun Yu, Changwei Luo, Chang Wen Chen. "Multi-Modal Human Modeling, Analysis and Synthesis", CRC Press, 2025 Publications	28 words — < 1%
46	theses.hal.science Internet	28 words — < 1%
47	tudr.thapar.edu:8080 Internet	28 words — < 1%
48	Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, Alberto Del Bimbo. "Investigating Nuisances in DCNN-Based Face Recognition", IEEE Transactions on Image Processing, 2018 Crossref	27 words — < 1%
49	www.eurasip.org Internet	27 words — < 1%
50	N. Babaguchi, Y. Kawai, T. Kitahashi. "Event based indexing of broadcasted sports video by intermodal collaboration", IEEE Transactions on Multimedia, 2002 Crossref	26 words — < 1%
51	easychair.org Internet	26 words — < 1%
52	deepai.org	

Internet

25 words — < 1%

53 Sahbi Bahroun, Rahma Abed, Ezzeddine Zagrouba. "KS-FQA: Keyframe selection based on face quality assessment for efficient face recognition in video", IET Image Processing, 2020

Crossref

24 words — < 1%

54 Chao, Gwo-Cheng, Yu-Pao Tsai, and Shyh-Kang Jeng. "Augmented 3-D Keyframe Extraction for Surveillance Videos", IEEE Transactions on Circuits and Systems for Video Technology, 2010.

Crossref

23 words — < 1%

55 Juan Du. "High-Precision Portrait Classification Based on MTCNN and Its Application on Similarity Judgement", Journal of Physics: Conference Series, 2020

Crossref

23 words — < 1%

56 Krishan Kumar, Deepti D. Shrimankar, Navjot Singh. "Eratosthenes sieve based key-frame extraction technique for event summarization in videos", Multimedia Tools and Applications, 2017

Crossref

23 words — < 1%

57 www.arxiv-vanity.com

Internet

22 words — < 1%

58 Enrique G. Ortiz, Alan Wright, Mubarak Shah. "Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-Based Classification", 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013

Crossref

21 words — < 1%

59	Jialiang Zhang, Xiongwei Wu, Steven C.H. Hoi, Jianke Zhu. "Feature agglomeration networks for single stage face detection", Neurocomputing, 2020 Crossref	20 words — < 1%
60	escholarship.org Internet	20 words — < 1%
61	Communications in Computer and Information Science, 2014. Crossref	19 words — < 1%
62	Roy, Kaushik. "On the Development of an Optical Character Recognition System for Indian Postal Automation.", Jadavpur University (India), 2020 ProQuest	19 words — < 1%
63	S. C. Dakin, R. J. Watt. "Biological "bar codes" in human faces", Journal of Vision, 2009 Crossref	18 words — < 1%
64	acadpubl.eu Internet	18 words — < 1%
65	iopscience.iop.org Internet	18 words — < 1%
66	unsworks.unsw.edu.au Internet	18 words — < 1%
67	www.jisem-journal.com Internet	18 words — < 1%
68	Costas Cotsaces. "<![CDATA[Face-Based Digital Signatures for Video Retrieval]]>", IEEE Transactions on Circuits and Systems for Video Technology, 4/2008 Crossref	17 words — < 1%

69	G.G. Lakshmi Priya, S. Domnic. "Shot based keyframe extraction for ecological video indexing and retrieval", Ecological Informatics, 2014 Crossref	17 words — < 1%
70	Z. Rasheed. "Detection and Representation of Scenes in Videos", IEEE Transactions on Multimedia, 12/2005 Crossref	17 words — < 1%
71	ijetae.com Internet	17 words — < 1%
72	m.moam.info Internet	17 words — < 1%
73	"6th International Technical Conference on Advances in Computing, Control and Industrial Engineering (CCIE 2021)", Springer Science and Business Media LLC, 2022 Crossref	16 words — < 1%
74	Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, Gang Hua. "A convolutional neural network cascade for face detection", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 Crossref	16 words — < 1%
75	Lecture Notes in Computer Science, 2015. Crossref	16 words — < 1%
76	Madhura Phatak, Manasi Patwardhan, Prashant Borkar. "Mood Detection in Aesthetically Appealing Video Based on Color Association", Wireless Personal Communications, 2024 Crossref	16 words — < 1%

77	Minh-Tri Pham. "Fast polygonal integration and its application in extending haar-like features to improve object detection", 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 06/2010 Crossref	16 words — < 1%
78	www.coursehero.com Internet	16 words — < 1%
79	Amin Ullah, Khan Muhammad, Tanveer Hussain, Sung Wook Baik, Victor Hugo C. de Albuquerque. "Event-Oriented 3D Convolutional Features Selection and Hash Codes Generation using PCA for Video Retrieval", IEEE Access, 2020 Crossref	15 words — < 1%
80	Xue Jiwei, Xin Jiyuan, Fang Yi, Chen Dongfang. "Research on Video Face Retrieval Method Based on Deep Learning and Key Frame", Proceedings of the 2020 4th International Conference on Digital Signal Processing, 2020 Crossref	15 words — < 1%
81	tel.archives-ouvertes.fr Internet	15 words — < 1%
82	vdoc.pub Internet	15 words — < 1%
83	www-nlpir.nist.gov Internet	15 words — < 1%
84	"Proceedings of 4th International Conference on Frontiers in Computing and Systems", Springer Science and Business Media LLC, 2024 Crossref	14 words — < 1%

85	"Proceedings of Data Analytics and Management", Springer Science and Business Media LLC, 2025 Crossref	14 words — < 1%
86	Nasreen, Azra, Kaushik Roy, Kunal Roy, and G. Shobha. "Key Frame Extraction and Foreground Modelling Using K-Means Clustering", 2015 7th International Conference on Computational Intelligence Communication Systems and Networks, 2015. Crossref	14 words — < 1%
87	dspace5.zcu.cz Internet	14 words — < 1%
88	ink.library.smu.edu.sg Internet	14 words — < 1%
89	moam.info Internet	14 words — < 1%
90	www.springerprofessional.de Internet	14 words — < 1%
91	A Debnath, K Sreenivasa Rao, Partha P Das. "Similarity-based Multi-Modal Lecture Video Indexing and Retrieval with Deep Learning", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023 Crossref	13 words — < 1%
92	B. Reddy Mounika, Om Prakash, Ashish Khare. "Key Frame Extraction using Uniform Local Binary Pattern", 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), 2018 Crossref	13 words — < 1%

93	Dias, Pedro Henrique Sampaio. "Design and Evaluation of a Cognitive Vehicle System: Emphasizing User Routine Learning and Interaction", Universidade do Minho (Portugal), 2025 ProQuest	13 words — < 1%
94	Muhammad Abulaish, Ashraf Kamal, Mohammed J. Zaki. "A Survey of Figurative Language and Its Computational Detection in Online Social Networks", ACM Transactions on the Web, 2020 Crossref	13 words — < 1%
95	att.aptisi.or.id Internet	13 words — < 1%
96	research.ijcaonline.org Internet	13 words — < 1%
97	"Advances in Multimedia Information Processing - PCM 2016", Springer Science and Business Media LLC, 2016 Crossref	12 words — < 1%
98	"Neural Information Processing", Springer Science and Business Media LLC, 2021 Crossref	12 words — < 1%
99	"Third International Conference on Image Processing and Capsule Networks", Springer Science and Business Media LLC, 2022 Crossref	12 words — < 1%
100	Adak, Subhas. "Land Resource Evaluation of Mondouri Farm by Using GIS Technology.", Bidhan Chandra Krishi Viswavidyalaya University (India), 2020 ProQuest	12 words — < 1%

101 Chenxi Bai, Kexin Zhang, Haozhe Jin, Peng Qian, Rui Zhai, Ke Lu. "SFFEF-YOLO: Small object detection network based on fine-grained feature extraction and fusion for unmanned aerial images", Image and Vision Computing, 2025
Crossref
12 words — < 1%

102 Enting Guo, Peng Li, Shui Yu, Hao Wang. "Efficient Video Privacy Protection Against Malicious Face Recognition Models", IEEE Open Journal of the Computer Society, 2022
Crossref
12 words — < 1%

103 Hanna F. Menezes, Arthur S. C. Ferreira, Eanes T. Pereira, Herman M. Gomes. "Bias and Fairness in Face Detection", 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2021
Crossref
12 words — < 1%

104 Iacopo Masi, Stephen Rawls, Gerard Medioni, Prem Natarajan. "Pose-Aware Face Recognition in the Wild", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
Crossref
12 words — < 1%

105 MD. SHAFAEAT HOSSAIN, KHANDAKER ABIR RAHMAN, MD. HASANUZZAMAN, M. A. BHUYIAN, H. UENO. "VIDEO IMAGE CLUSTERING BASED ON HUMAN FACE AND SHIRT COLOR", International Journal of Image and Graphics, 2012
Crossref
12 words — < 1%

106 Muhammad Asim, Noor Almaadeed, Somaya Al-maadeed, Ahmed Bouridane, Azeddine Beghdadi. "A Key Frame Based Video Summarization using Color Features", 2018 Colour and Visual Computing Symposium (CVCS), 2018
12 words — < 1%

Crossref

107 N. Krishnaraj, Mohamed Elhoseny, E. Laxmi Lydia, K. Shankar, Omar ALDabbas. "An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in cloud environment", Software: Practice and Experience, 2020 — 12 words — < 1%
Crossref

108 Shaohua Zhou, Volker Krueger, Rama Chellappa. "Probabilistic recognition of human faces from video", Computer Vision and Image Understanding, 2003 — 12 words — < 1%
Crossref

109 Umurerwa Marie Adeline, Harerimana Gaspard, Kabandana Innocent. "Chapter 8 A Real-Time Face Recognition Attendance Using Machine Learning", Springer Science and Business Media LLC, 2023 — 12 words — < 1%
Crossref

110 Vasu Namala, S. Anbu Karuppusamy. "Efficient feature based video retrieval and indexing using pattern change with invariance algorithm", Journal of Intelligent & Fuzzy Systems, 2023 — 12 words — < 1%
Crossref

111 digitalcommons.unl.edu — 12 words — < 1%
Internet

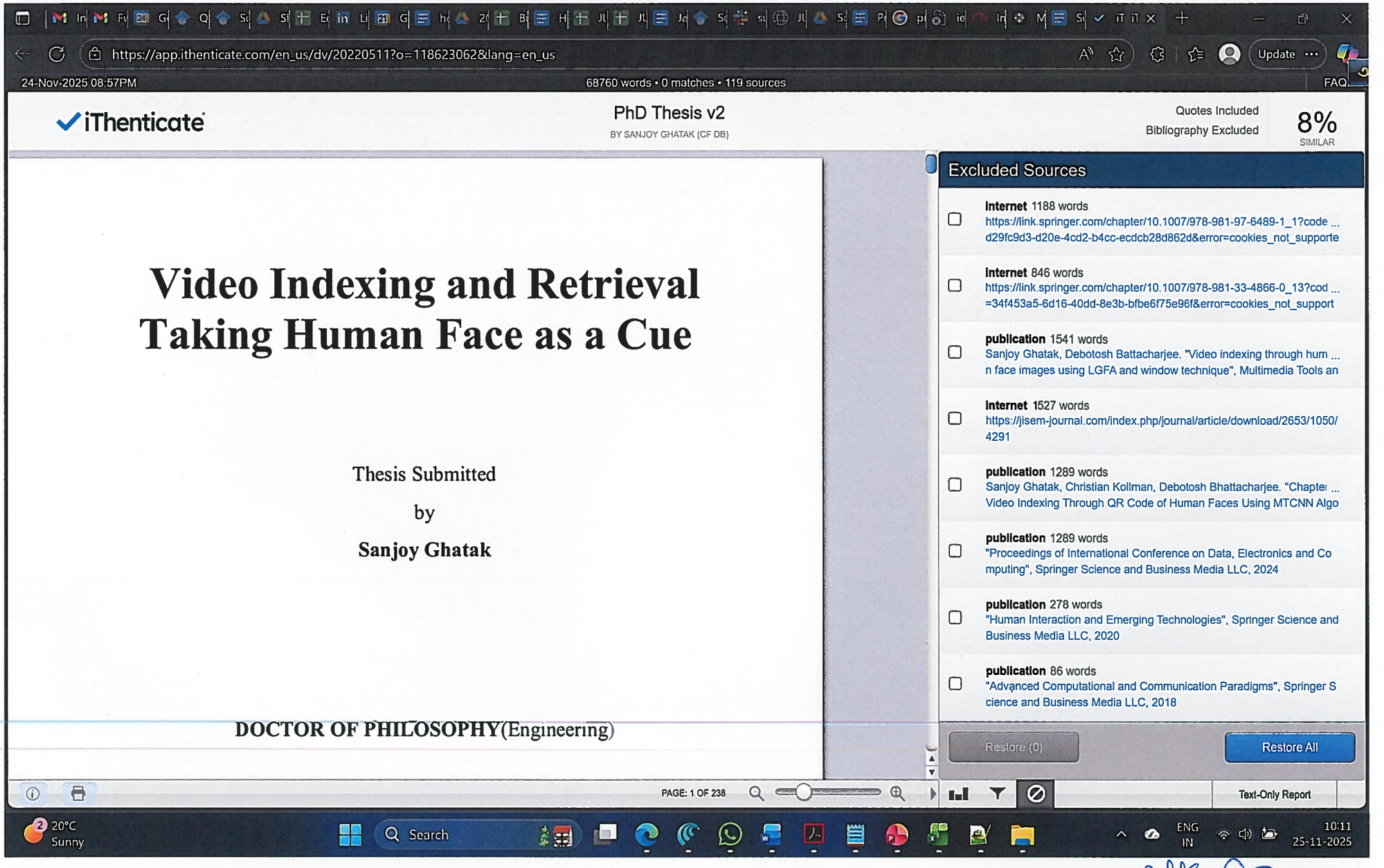
112 dl.lib.uom.lk — 12 words — < 1%
Internet

113 emicsoft-iphone-converter.com-download.net — 12 words — < 1%
Internet

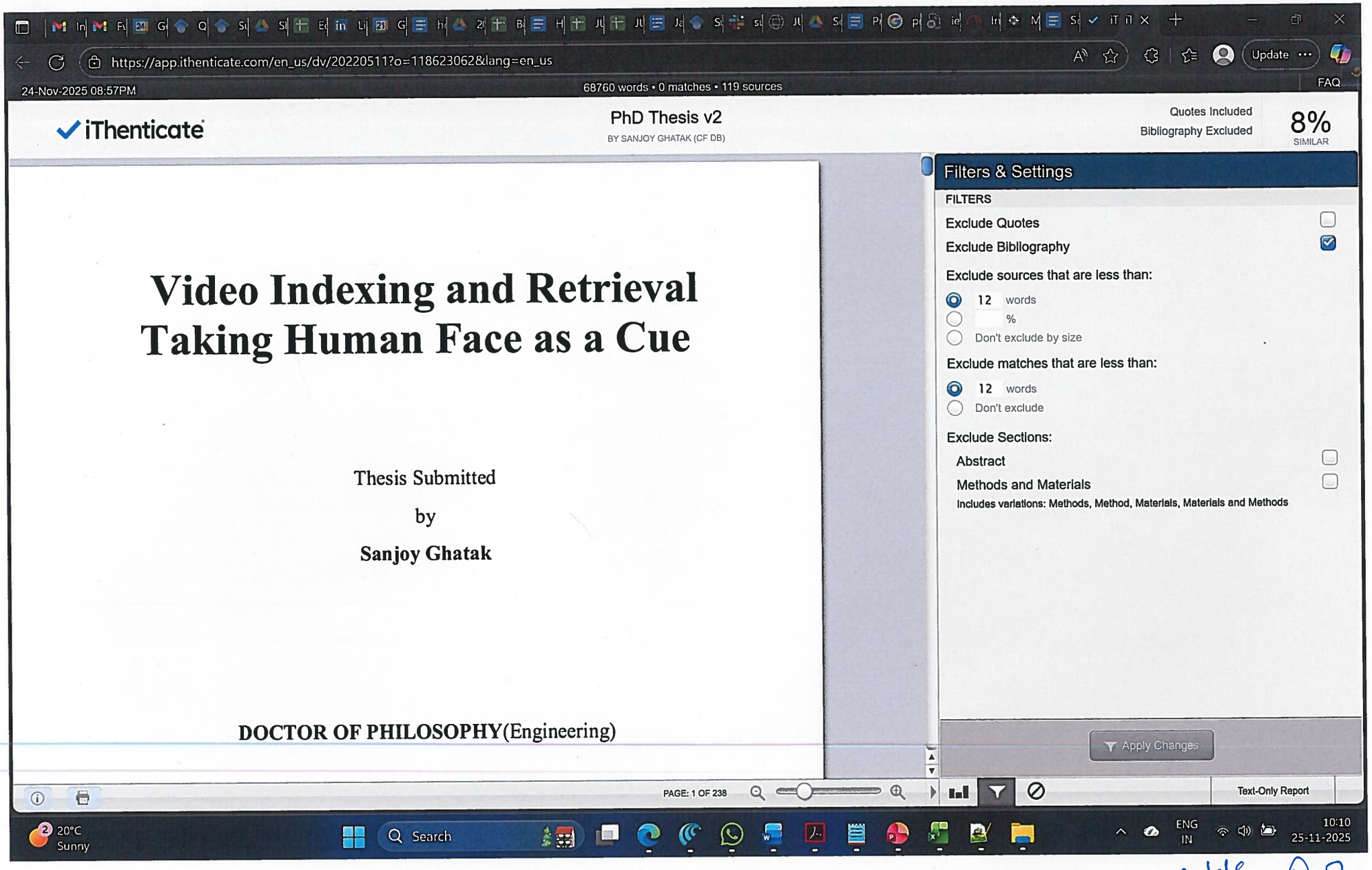
114 export.arxiv.org — 12 words — < 1%
Internet

115	ia800501.us.archive.org Internet	12 words — < 1%
116	iovsoft.hypermart.net Internet	12 words — < 1%
117	publications.eai.eu Internet	12 words — < 1%
118	www.diplomarbeiten24.de Internet	12 words — < 1%
119	www.ijitr.com Internet	12 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS



Sanjoy Ghatak



Sanjoy Ghatak

Multimedia Tools and Applications (2022) 81:31509–31527
https://doi.org/10.1007/s11042-022-12965-2

Video indexing through human face images using LGFA and window technique

Sanjoy Ghatak[1] · Debotosh Battacharjee[2]

Received: 22 May 2021 / Revised: 8 February 2022 / Accepted: 13 March 2022 /
Published online: 9 April 2022

Abstract
Adaptive video monitoring settings have been extensively deployed in recent years. Smart video monitoring technology enables the acquisition and analysis of movies from various devices, as well as automatic analysis based on knowledge gathering. However, the storage capacity is restricted, and the important frames from the movie cannot be saved. Though, if the movie employs face as keyframes, it creates space and time complexity. To address this issue, the Viola-Jones Algorithm was used to detect faces from extracted keyframes in Video Indexing through Human Face Images using LGFA and the sliding window technique. The image gradient for brightness is created by integrating the sliding windowing method with LGFA, and scanning the input image horizontally takes up 70% of the facial image. As a result, Barcode as an index using the sequence table of the EAN 8 approach converts a video's human face into an EAN-8 linear video indexing barcode and thereby reducing bandwidth, storage space, and time complexity. Regular TV series video datasets, datasets of YouTube faces, and data sets of Hollywood clips were used to evaluate the proposed technique, and shown to be effective for indexing videos based on human faces.

Keywords Indexing · Facial image · EAN-8 barcode · Viola-Jones algorithm · Image gradient · Illumination invariant face image · LGFA

1 Introduction

The use of video data, ranging from conventional entertainment and radio broadcasting to camera systems for the development of, medicine, intelligent urban environments, transport, or even wearable devices, has become all-around in our everyday lives. It allows systems to

✉ Sanjoy Ghatak
scholar.sanjoyghatak@gmail.com

[1] Sikkim Manipal Institute of Technology, Staff Quater L-304, Majitar, Rangpo, East Sikkim, India

[2] Jadavpur University, 188, Raja S.C. Mallick Rd, Kolkata, West Bengal 700032, India

30/06/25 — 30/06/2025

effectively and efficiently archive, arrange and handle video data while allowing users to easily fulfill a basic demand for information on top of the large video collections [16, 22, 25]. The image probably is notified of the visual formation and its subsequent sections shall be annihilated. Owing to its approach, the description of photographs or videos on the current structures operates with file name searching rather than the inside of it [15, 20, 29]. Deep learning is one of the classifications of soft computing in which data can be obtained using the phenomenon from millions of segregated images. The efficiency of an image retrieval system based on the material depends crucially on the feature representation and similarity calculation, which multimedia scientists have researched extensively for decades [1, 24]. Video content has unique accessibility challenges: indexing and searching video has always been very labor-intensive. New software that uses deep learning techniques can automate video indexing and search, making video content more discoverable and useful [12]. These deep learning resources include modern and fascinating ways of transcribing, decoding and managing Video. This column explores the concept of deep learning and investigates several new stuff that can be achieved by video deep learning techniques [13].

The video of a movie can be found on the Internet by users via a snap with no tag. Another application is for a person to scan for a recording of a lecture online with a slide. The scientific group has applied a lot of work to the issue of video recovery. Any video frame is called a single image as a most basic solution to this problem. Thus, it becomes a challenge to figure out whether a picture is contained in a video if the background is identical to a picture [30]. Deep neural networks have been used in various real-world systems such as autonomous cars, sports, research, and even art as an important advance in machine learning. Deep learning has contributed to pioneering advancement in numerous areas, such as computer vision, natural language processing (NLP), and voice processing. Most of the research deal with the issue of single modal deep learning instead of multimodal learning problems. In a multimedia system, however, the final identification and recovery output can be dramatically improved for different data types, particularly when errors or missing values in one or more modes are present [26].

Videos retain rich content than images, which concurrently contain a lot of redundant content. In addition, video processing and analytics involve substantial computational complexity, including browsing and recovery, for their efficient use [7, 10]. Consequently, it is difficult to retrieve related videos from a wide archive in many ways relative to image retrieval. Video recovery is a boring and time-consuming process for humans manually. In addition, people are vulnerable to mistake, so there is a risk of inappropriate findings [27]. Although video feature retrieval burden is reduced and system efficiency in a highly competitive network environment is increased, the system also needs a lot of resources and bandwidth [9, 23]. A concerning issue to address immediately is how to process and interpret videos rapidly and efficiently. Video keyframe removal is an easy way to easily search and capture huge video data. It is also the base for video applications [6].

In video keyframe applications, the contributions are in poor representation and redundancy of the keyframes retrieved and in low continuity of the user-extracted keyframes. The purpose is to eliminate keyframes that are more in line with the human vision system and are capable of generalizing the original video content [19]. The redundant and identical keyframes in a video without impacting visual content in semantic information are removed by traditional frame extraction processes [10]. Previous researchers use multiple/cross-mode hashing to solve the complexities of the video image keyframe fusion problem with different modalities [28]. The video evidence, based on frames such as the original frame, middle frame, and final frame, is

drawn from the standard techniques. The previous approach is easy to execute, but certain videos in these mainframes could be skipped which reduces the number of computations and time [18]. Do not concentrate on the ordering of facial images for keyframe extraction. In this way, the consistency of the face in such frames is considered, not involved in ordering the face images. It shows that extracting frames of good face content increases the precision of face recognition [3].

This paper seeks to provide a solution to major difficulties such as posture change, limited storage capacity, and inability to preserve important frames from video, as well as time and space complexity. As a result, Video Indexing with Human Face Images Using LGFA and Sliding Window Technique is presented to address these difficulties, which is the paper's major contribution.

- Color histogram is used to find the distinction between two images and hence remove the issue in posture change.
- Cropped face portraits using the Viola-Jones algorithm from the keyframes have been used to extract the keyframes from the face and thereby solve the issues to store keyframes.
- The face image gradient is computed from grayscale using LGFA and Sliding Window technique determines the image gradient from greyscale image and the stable bar code can be created from this gradient of the image and thereby removes the time and space complexity.
- Barcode as an index using the sequence table of EAN-8 is employed to index human faces recognized from any form of video since EAN-8 barcodes are a linear representation of face images and thereby remove the issue in limited storage capacity.

Thus the proposed video indexing approach solves the issues such as posture change, limited storage capacity, and issues to store keyframes from video and time and space complexity problems.

The rest of the paper is organized as follows: Section 2 describes the related works of the prior method, Section 3 describes the proposed methodology and working characteristics of the various algorithms, Section 4 describes the experimental results and discussion and Section 5 concludes this paper.

2 Literature survey

As seen in the introduction, this paper discusses adaptive video monitoring, video indexing [15, 16, 20, 22, 25, 29] through deep learning techniques [13, 26, 30], also video keyframe removal [6, 9, 10, 19, 23, 27], and video image keyframe fusion [3, 18, 28] problems. In this section, some recent related researches are reviewed and the gaps are identified in order to motivate for proposing a novel video indexing technique.

Gayathri et.al [11] analyzed that in pre-processing video frames before accessing secret video collections, there are several shortcomings. Feature extraction and classification methods are considered to resolve the pitfalls in pre-processing. Video indexing with multiple extracting capabilities with prevailing frame formation for the input video frame has been awaited here. A fuzzy-based SVM classifier is used to categorize frame structures into dominant structures. To remove texture features from a video clip, the multidimensional histogram of directed gradients (HOGs) and the color attribute extraction are used. However, in this method, the

classifiers cannot concentrate on video processing applications to specify signals, and storage capacity is limited.

Lin et.al [21] discussed that deep neural networks, especially for facial recognition, have been intensively investigated and profound learning models used widely to detect artifacts. This research, therefore, suggested a deep learning cloud-based video recovery system. Next, it extracts and preprocesses a dataset, to create a useful dataset for templates of CNN, distorted images are omitted and the remaining images are matched. The final dataset is then developed and used for pre-training of the CNN models for face recognition (VGGFace, ArcFace, and FaceNet). However, in this method, the system is not improved to get more datasets and the efficiency is not enhanced.

Li et.al [19] focused on tackling the problems in the smart town protection video retrieval of the link management program in a single packet processing, suggested firstly a traffic location quantization index based on backbone traffic characteristics to evaluate the traffic region characteristics in the backend communication in a quantitative way. To increase the performance and accuracy of video recovery, the keyframe abstraction and retrieval of videos based on deep learning is proposed where an adaptive keyframe selection algorithm is developed and the current convolutionary neural network architecture is used to extract features of keyframes, and unsupervised, semi-supervised, and supervised retraining models are built. However, this method does not maintain space and time complexity and keyframes are not stored.

Bastanfard et al. [4] To anticipate the fundamental impacts of facial pictures with varied appearances, an E-appearance method is presented. The quotient image captures the appearance characteristic of an image given facial image data of the same individual in two distinct appearances. Then, using a warping approach, we transfer the feature to any other specific person's face to create a new facial look. This technique may fail to accurately depict the faces because of huge differences in lighting conditions, facial expression, and other variables.

Bastanfard et al. [5] With two approaches, this study presents a unique face rejuvenation modeling algorithm. These approaches explain facial deformation using the face anthropometrics theory and erase wrinkles using a process known as wrinkle inpainting. For example, if we are given a few different faces, we must be able to assess the differences between the facial features of the young and the elderly and then establish a set of outlines that will serve as the foundation for the Face Rejuvenation simulation. This technique may fail to accurately depict the faces because of huge differences in lighting conditions, facial expression, and other variables.

Dutta et al. [8] The goal is to produce a sentence for an image by detecting characteristics using deep learning techniques, as well as to generate a phrase for video frames using the same model used for image captioning. Key-frames are retrieved from the video by sending it through the keyframe extraction framework integrated within the program during the production of cation for the video. The retrieved keyframes from the video are put into the same image captioning model that was used to create the captions for the photos. By putting the pictures into a preset pre-trained model, captions are produced from the frames retrieved from the movie. However, one of the difficulties was generating correct meaningful phrases from a large vocabulary, while another was turning video pictures into a meaningful sequence of frames.

Jacob et al. [14] This study presents a unique way of analyzing video content and retrieving the desired video clip from a long video using video storytelling and indexing techniques. The video storytelling approach is used to analyze video footage and create a video explanation. The video description is then utilized to construct an index using the wormhole method,

ensuring that a keyword of fixed length L may be found in the shortest possible time. Because of the frequency of the term in the video index's keyword search, this video index may be utilized by video searching algorithms to obtain the relevant portion of the movie. Rather than downloading and uploading the entire video, the user can download and transfer only the video clip that is required. Hence in this method time complexity may occur.

Anayat et.al [2] analyzed that a recent specialism has been the search for video in large libraries. Due to technological developments, well-established search methods for the video recovery of pictures must be adopted. This research primarily aimed to classify the best video retaining technologies based on the image. This research revealed the importance of image base videos in the search area and solved the issue of choosing the most precise technique for I2V retrieval. Different procedures vary in precision and recovery time. This research showed that there are many visual search strategies, with different precision and velocity, all these techniques work differently. However, in this method, the retrieval of video clips is not expanded as well as time complexity increased.

Krishnaraj et.al [17] analyzed that although cloud services provide efficient image indexing, it remains an important problem because of the semantical distance between the user query and various semantics of the broad database. An RTI model based on visual semantic indexing for photography from cloud platforms will be displayed in this Article. Initially, the typical semantic and visual descriptor space is being defined through an interactive optimization model. Next, an RTI architecture is used for combining the semantic visual space sharing model to find an optimal solution for looking for larger data sets. Finally, an online image retrieval service is added to the distributed model Spark. Two standard datasets, Holidays 1 M and Oxford 5 K in terms of average precision (mAP) and processing time in various sizes of data sets, validate the efficiency of the proposed system. However, this method is not enhanced in machine learning also computation time is not minimized.

[11] Cannot focus classifiers on the application for video processing to set signals and the ability for storage is small [21] has not been developed to collect more knowledge and performance has not been improved. [19] It cannot handle the complexity of space and time and does not save keyframes. These techniques may fail to accurately depict the faces because of huge differences in lighting conditions, facial expression, and other variables [4, 5]. Turning video pictures into a meaningful sequence of frames is difficult [8] and in [14] time complexity may occur. [2] Video clip retrieval is not improved and time complexity is raised, in addition, [17] the machine learning technique is not improved as well as computation time is not diminished. Thus motived by the aforementioned issues presented in the video indexing and retrieval, we propose a novel technique by solving the above issues named Video Indexing with Human Face Images Using LGFA and Sliding Window Technique.

3 Video indexing and retrieval based on LGFA and sliding windowing technique through human FACES

With the advent of cheap video cameras and high process capacity, video-based facial recognition has easily surpassed image-based approaches. That is why video-based facial recognition has drawn many researchers' interest in recent years. Face detection provides a high degree of precision in a controlled environment. This function remains the biggest difficulty in contrast to the crowded atmosphere because of head changes, lighting situations, facial gestures, the occlusion of other items or clothes, resolution and blurring triggered by

people's movements in front of the cameras, e.g. sunglasses, scarves, etc. Face recognition is more interesting in images on the one hand, rather than just one shot single picture. Secondly, because of the time it takes to work with all frames, the process of this vast volume of data for each video is difficult. Furthermore, faces will either be obsolete or irrelevant due to a lack of accuracy in these pictures. For identifying the issue of a large number of data frames within a video, some frames have to be chosen. Do not focus on ordering facial images for keyframe extraction. Thus, only the accuracy of the face in these frames is regarded and not included in the ordering of face pictures which provokes the storage is limited yet time and space complexity is increased. Following a thorough examination and analysis of this research, it has been shown that video-indexing using low-level characteristics is challenging for individual recognition. All of these approaches include the use of time and space in terms of storage space and indexing. The video's face picture's facial expression, posture, mood, illumination shifts, and occlusions are all key factors in video indexing via the human face. Video-based identification is also shown to be hampered by intrinsic ambiguities such as poor resolution sensitivity, posture changes, and partial blockage of facial cavities. To address these issues with the implementation of face detection, a technique that can do the following must be developed: With small differences in lighting conditions, facial emotions, and face position, the system will allow for sufficiently exact performance, Facial image lighting that is invariant, The technique should produce fairly detailed results without regard to the individual's age or skin color.

The existing techniques have significant issues such as storage capacity being limited and are not able to store keyframes from video in addition, time and space complexity occurs, posture changes, and partial blockage of facial cavities. Hence to tackle these issues, a Color histogram is used to find the distinction between two images and hence remove the issue in posture change. Cropped face portraits using the Viola-Jones algorithm from the keyframes have been used to extract the keyframes from the face. A novel hybrid sliding windowing technique and LGFA has been introduced to determine image gradients and hence reduce the time and space complexity and increase storage capacity. Barcode as an index using the sequence table of EAN-8 is proposed in which the keyframes can store from video through EAN-8 barcode and thereby solves the issue in storing keyframes.

The different phases of the Proposed strategy are described in this work and seen in Fig. 1 with a block diagram (a) The first step is to recover all video input pictures (b) Keyframes are retrieved using the color histogram discrepancy, from the frames obtained from the video input. (c) Use the Viola-Jones algorithm to detect faces from the main image. (d) All facial pictures are converted into face grayscale. (e) The gradient of the facial photograph is calculated with LGFA and Window sliding technology. (f) The barcode EAN-8 is generated by the image gradients using the EAN-8 sequence table.

3.1 Extracting frame

A combination of the scene shot and the frame is dynamic video. The first step, therefore, is to extract the still images that are depicted as a scene, shot, and picture from input videos. The scene is a series of frames that are filmed and shot. Traditional video contains 20 to 30 frames per second and has an abundance of content. The frame is a still image and includes redundant data that is present in a film. Figure 2 reveals the frame present in the as Good as It Gets-01766 Holly wood movie scene.

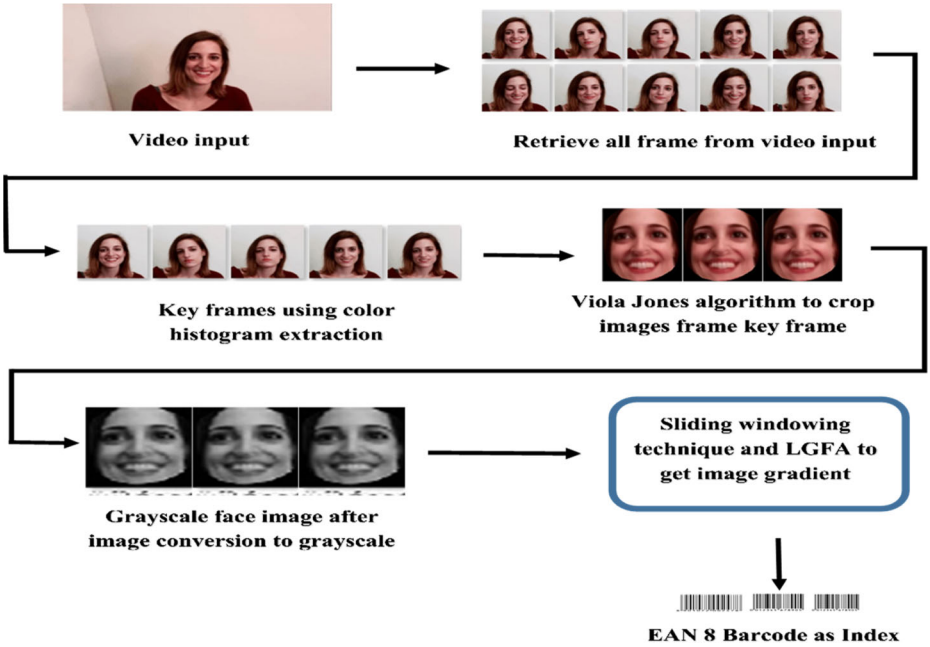


Fig. 1 Block Diagram for Proposed System

3.2 Extracting Keyframe

3.2.1 Color histogram technique

The frame that reflects each shot's vital information is known as the keyframe. Human face with distinct expression, posture, lighting condition, and illumination are considered as keyframes in this paper. A few of the similarities, the likelihood scale, the effects of the curve saliency motion capture, and several more methods. But in this step, the Color histogram method is used to extract the keyframe from all frames of a given film. Keyframes are retrieved from the frames using the Color Histogram difference. If the observed difference is greater than the magnitude of the threshold. The color histogram discrepancy is set to the threshold then the frame will be selected as the next keyframe. Figure 3 displays some of the keyframes based on a Hollywood movie's human face (As Good as it Gets-01766).

Fig. 2 The frame of as Good as It Gets-01766 Hollywood movie scene

Fig. 3 KeyFrame (based on the face as key content) of as Good as It Gets-01766

A color histogram is used for the keyframe extraction as a color-dependent frame difference technique. The explanation behind it is that color is one of the most important visual features to describe a picture. In the case of camera motion and a simple method of computation, a color histogram is robust. The concept behind this technique is that their corresponding histograms in two frames with uniform context and uniform (though moving) objects would have little difference. An important part of this approach is threshold selection. The formula for determining the distinction between the two consecutive frames in the color histogram is as follows:

$$D(F_I, F_{I+1}) = \sum_{J=1}^{n} \frac{h_I(J) - h_{I+1}(J)^2}{h_{I+1}(J)}$$

Where h_I and h_{I+1} indicate the histogram of the two consecutive frames F_I and F_{I+1}respectively. When D (F_I, F_{I+1}) is higher than the given threshold, then a shot transition occurs. Where 'n' is the number of the frame. The algorithm for the color histogram is described below:

Algorithm 1: Algorithm for finding the color histogram difference between two images.

Input: The two different images. (Image1, image2)
Output: The Histogram difference of two images (Image1 and Image2)
Step1: Start
Step2: Calculate the image histograms along each channel RGB (1, 2, and 3) for each image1 and Image2.
Step3:Normalize each histogram for all 3 channels for each image as:
histogram nor= histogram/max (histogram).
Step4: Calculate histogram error (The difference in terms of Euclidean distance) between two images for each channel as:
heR=($histgramR1 - histogramR2$)2 // for red color, heR means histogram error for red color.
heG=($histgramG1 - histogramG2$)2 // for Green color, heG means histogram error for green color.
heB =($histgramB1 - histogramB2$)2 // for the blue color, heB means histogram error for blue color.
Step5: Calculate color histogram difference by considering each channel contribution towards it as:
Histogram Difference = 0.2989*heR+0.5870*heG +0.1140*heB
Step6: End

Algorithm 2 Algorithm for generating keyframes from an input video.

Input: Video file
Output: Keyframes
Step1: Start
Step2: Read the user input video
Step3: Store the number of frames present in the video.
Step4: Take the two corresponding frames for comparing and identifying the keyframe from all the frames present in the video.
Step5:Calculate the color histogram difference of the two corresponding frames.
Step6:Then compare with the threshold value. // threshold is the amount of variation in the color histogram.
Step7:If the difference in the color of the selected frame is greater than or equal to the threshold.
Step8: Then this frame is considered as the keyframe.
Step9: Else continue steps 4 to 7.
Step10: When all the keyframe is identified after comparing all frame present in the video then the process is stopped.
Step11: Stop

3.3 Cropped face portraits using the Viola-Jones algorithm from the keyframes

Faces are detected from the extracted keyframes using the Viola-Jones object detection technique. The benefit of using the Viola-Jones algorithm is that detection is fast, but training is slow. Considering an image(this grayscale algorithm works), this algorithm sees several smaller subregions and attempts to locate a face by searching in each subregion for particular characteristics. It involves verifying several different sizes since an image will contain several sides of various sizes. Instead of facing upwards, sideways and backward faces Viola-Jones has been designed for frontal faces. Until a face is found, the images are transformed to grayscale, so the encoding is faster and fewer details are available. First, the Viola-Jones algorithm senses the face of the gray picture and then the position of the colored picture. Some screenshots of Keyframe face detection are shown in Fig.4.

3.4 Grayscale face image after image conversion to grayscale

To calculate the gradient of the face image, face image to grayscale conversion is needed after getting the faces from keyframes.

Fig. 4 Cropped faces from Keyframes of as Good as It Gets-01766 (Hollywood movie)

3.5 The face image gradient is computed from grayscale using LGFA and sliding window technique

Image gradients are determined from this grayscale image (faces) using LGFA and the Sliding Window technique. LGFA is used for calculating image gradient from illumination invariant face image. In paper [18], it is seen that the sliding window technique is not good for generating stable bar code from illumination invariant face images. For this reason, a combination of LGFA and sliding window techniques are used to generate a stable bar code. The sliding window technique is used to scan the input image from the forehead to the upper portion of the lips. The image gradient values are calculated using this method, considering the upper 70% of the face image. Up to 70% of the face image is taken for secure barcode creation. If the sliding window moves the upper 70% of the face image; the stable bar code can be created from this gradient of the image.

3.6 Barcode as an index using the sequence table of EAN 8

The method of EAN-8 bar code generation from the gradient value obtained will be discussed in this section. In this methodology, a linear representation of the face image identified from the video is the barcode. With the support of these barcodes, it is possible to index human faces recognized from any form of video since these barcodes are a linear representation of face images. The advantage of human face barcode indexing is that these barcodes require less storage space and minimum indexing time. The examples of certain LG- faces and their barcodes are shown in Fig.5, and those are detected from various face video datasets in Hollywood (As Good as It Gets-01766).

Fig. 5 The as Good as It Gets-01766 video scene and its corresponding EAN 8 bar-code have cropped faces

The following are barcode generation algorithms from the obtained gradient values of the face images:

Algorithm 3:

Input: Gradient of face image obtained from the taken video.
Output: EAN 8 Face picture linear barcode.
Step1: Start
Step2: The maximum value of the gradients
Step3: Initialize max_gradient with the maximum value of gradient.
Step4: for i= 1 to T-(windowsize+1) do
gradient (i) = gradient (i)/max_gradient// NORMALIZATION
end
Step5: for i=1 to T do
gradient(i)=floor(gradient(i)*10) // QUANTIZATION
end
Step6: Create a zeros matrix named barcode
Step7:Initialize scale=9.5
Step8: for i=1 to 7 do
initialize num with value 0
for j=1 to m do
num=num+gradient((i-1) *m+j)
end
barcode(i)=round((scale*num)/m)
end
Step9: S_odd=barcode (1) +barcode (3) +barcode (5) +barcode (7) // S_ odd means summation of odd position number present in EAN 8 barcode
Step10:S_even= barcode (2) +barcode (4) +barcode (6) // S_even means summation even position number present in EAN 8 barcode
Step11: barcode (8) = mod(10-mod((3*S_odd+S_even),10)10 // CALCULATE CHECKSUM VALUE
Step12: for i= 1 to odd do
Store values of the barcode(i) into barcodeStr
End
Step13:Use EAN 8 sequence table and gradient values to generate graphical storage and store in barcode Image
Step14: Stop

3.7 Bar code determination accuracy

The accuracy of the human face bar code is calculated based on an 8 digit EAN bar code number with two similar or different photographs of the face. The checksum digit is not considered here since it is an error correction digit. The remaining 7 digits are taken into account. Accuracy is determined when the digit present in the two EAN bar codes of the same position is compared between the two human faces of the same or different image. At the same time, two different barcode numbers are paired. If two different bar codes have the same position number, the same face picture assumes to be the same face. For measurement, all the bar code is produced from the face image of the input video is compared to the precision of the particular bar code. If the accuracy is greater than 80%, then two bar codes are assumed to be the same faces. After comparing the two bar codes and if it is less than 60%, then the barcode

of the image is considered to be different faces. However, if the accuracy is less than 60% and the face images bar codes are the same faces, then these faces' barcode forms are not treated as indexing. But it can be indexed for record-keeping purposes and may generate uncertainty in this situation. The formula stated to calculate the barcode's accuracy is as follows:

$$Accuracy = \left(\frac{(F-I)}{F}\right) \times 100(2)$$

Where F is the size of the barcode, and I is the initial value, which is considered as zero.

4 Experimental results and discussion

This section provides a comprehensive description of the implementation results, performance, and comparison strategies of this proposed system.

4.1 Experimental setup

This work has been implemented in the MATLAB working platform with the following system specification is given below.

Platform: MATLAB 18 a.

OS: Windows 7.

Processor: Intel Core.

RAM: 4GB RAM.

The keyframe is first extracted from the human face-based video dataset in this research and then created the barcode from this cropped face. After that indexing is performed based on the human face barcode present in the video. Some video datasets, however, explicitly contain the face image keyframe, so this generates the bar-code directly from this dataset without extracting the keyframe (Face video dataset for YouTube). To validate the method, utilizing three distinct video data sets.

Initially, present a realistic video dataset of TV series consisting of 27 episodes from six popular TV series. Breaking Bad (3), How I Met Your Mother (8), Crazy Man (3), Modern Family (6), Sons of Anarchy (3), and 24(4) were 27 episodes. The total size of these videos is 16 h.6231 are the total number of acts and 30 are the actions in this video. After this, the video dataset of Hollywood is taken for the experiment. This includes video clips are taken from 32 human action films. For the sample to be labeled, one or more of eight groups are needed. The 20-film data set is divided into a test set and two 12-film practice sets, isolated from the test set. The automated learning set includes 233 video recordings, collected through automatic script-based action marking with approximately 60% right labels. In a clean training collection of Hollywood results, there are 219 video samples with manually-checked labels and 211 video samples with manually tested labels.

Finally, for the experiments, YouTube facial data sets are taken. This data collection features 3425 videos of 1595 outstanding individuals as well as have taken the entire YouTube video. For each subject, a standard 2.15 recording is available. 48 outlines are the most restricted clasp word, the longest clasp is 6070 edges, and 181.3 edges are the usual length of the video clasps. Table 1 shows the barcode comparative study created using the window technique from some of the Hollywood Movies and combining LGFA and window technique.

After analyzing the result, it was observed that get better results after using combining LGFA and window technique than only window technique. Quality of barcode also increases and no of barcode also increase in this method than window technique. All videos of this data set are in AVI format. After observing the result of Table 1 see that the number of the face detected from Butterfly Effect −00696, Forrest Gump-01277, and Gandhi-02262 datasets are 15, 35, and 5 respectively, but from these datasets are 3, 16, and 4 respectively, the number of valid bar code of the individual face is created using window technique but no of bar code is increased (8, 20 and 5) after using combining LGFA and window technique. The explanation behind this is that when calculating the barcode accuracy, the accuracy is less than 60% and the barcode of the face images is the same. So all of these barcodes for face images are not deemed to be indexed.

The result of the barcode created from the video dataset for the TV series is shown in Table 2. All videos from this dataset are available in MP4 format. There is no barcode here for every face recognized from keyframes. If this takes into account the bar code number of the same picture faces, will be diminished.

Table 3 is the product of our YouTube Face Video Face Data Set job.

4.2 Performances metrics

Accuracy The accuracy defines the total number of correct predictions returned within the face recognition

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Precision Precision is defined as the ratio of the number of relevant returned images divided by the sum of the returned ones

$$Precision = \frac{TP}{TP + FP}$$

Table 1 Statistics of the number of generated mainframes, the number of detected faces, and the number of barcodes generated from some of the Hollywood movies

Video Name	Size(KB)	No. of Keyframes	No. of faces detected	No. of valid barcodes using window technique	No. of valid barcodes using combining LGFA and window technique
American Beauty - 00170	1420	2	1	1	1
As Good As It Gets – 01766	6150	18	5	5	5
Big Fish - 00664	1640	6	1	1	1
Butterfly Effect, The - 00696	1940	15	15	3	8
Casablanca – 00250	628	7	0	0	0
Crying Game, The - 01482	3210	18	1	1	1
Forrest Gump – 01277	22,300	424	35	16	20
Gandhi - 02262	18,000	18	5	4	5

Table 2 Number of generated keyframes, number of recognized faces, and number of valid barcodes generated from some episodes of the TV series video dataset

Video name	Size (KB)	No. of Key Frames	No. of faces detected	No. of the valid barcode.	No. of valid barcodes using combining LGFA and window technique	Video name
24_ep2	848,000	4471	1970	1970	1970	24_ep2
Breaking_bad_ep1	1,130,000	2473	567	567	567	Breaking_bad_ep1
How_I_Met_Your_Mother_ep1	435,000	553	532	532	532	How_I_Met_Your_Mother_ep1
Mad_men_ep1	980,000	3092	1983	1983	1983	Mad_men_ep1
Modern_Family_ep1	465,000	2356	1415	1415	1415	Modern_Family_ep1
Sons _of _Anarchy_ep1	1,110,000	3964	1648	1648	1648	Sons _of _Anarchy_ep1

Table 3 Keyframe no statistics generated from some TV series video database video files, faces no. detected, and barcode no. generated

Video Name	Size (KB)	No. of Key Frame detected	No. of faces detected	No. of valid barcode generated	No. of valid barcodes using combining LGFA and window technique
Aligned_video_0	556	84	84	84	84
Aligned_video_5	819	166	166	166	166
Aligned_video_3	1171	173	173	173	173
Aligned_video_1	1942	307	307	307	307
Aligned_video_2	3158	468	468	468	468
Aligned_video_4	664	119	119	119	119

Where TP represents the true positive pairs, which means that the system assigns similar images pairs to the same cluster. FP denotes the false positive pairs. In other words, assigning a dissimilar pair of images to the same class.

Recall The recall is defined as the ratio of the correctly clustered images pairs by the total number of.

images of the same cluster. It means that the recall measurement is the percentage of the retrieved relevant images from all returned ones.

$$Recall = \frac{TP}{TP + FN}$$

FN refers to assigning two similar pairs of images to different clusters. Named false negatives pairs.

The F1-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$F1{-}Score = 2 \times \frac{Precision \times Recall}{Precisiom + Recall}$$

4.3 Comparison strategies

In this section, the performance of the proposed method is compared with prior technologies such as Fuzzy based Support Vector Machine classifier (FBSVM), Video Retrieval Model on Convolution Neural Network (VRMCNN), Cloud-based Face Video Retrieval System (CBFVRS). The Comparison Process is represented graphically in the below figure.

Figure 6 shows the resultant plot for comparison for accuracy with prior methodologies such as FBSVM, VRMCNN, CBFVRS. The figure reveals that the proposed framework obtained high accuracy when compared with previous techniques.

Figure 7 shows the resultant plot for comparison for Precision with prior methodologies such as FBSVM, VRMCNN, CBFVRS. The figure reveals that the proposed framework obtained high Precision when compared with previous techniques.

Figure 8 shows the resultant plot for comparison for recall with prior methodologies such as FBSVM, VRMCNN, CBFVRS. The figure reveals that the proposed framework attained high recall when compared with previous techniques.

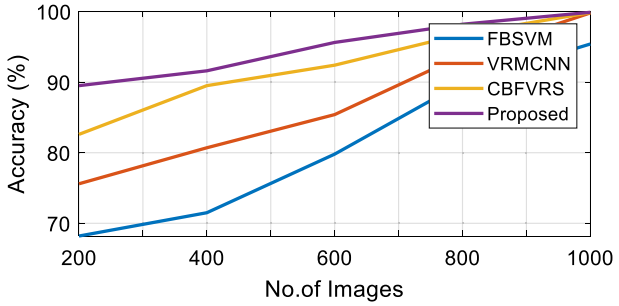


Fig. 6 Comparison for Accuracy

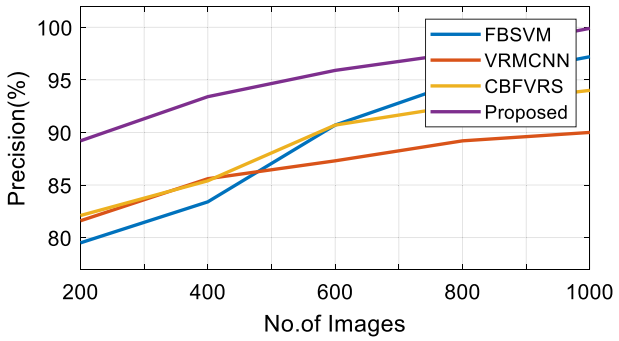


Fig. 7 Comparison for Precision

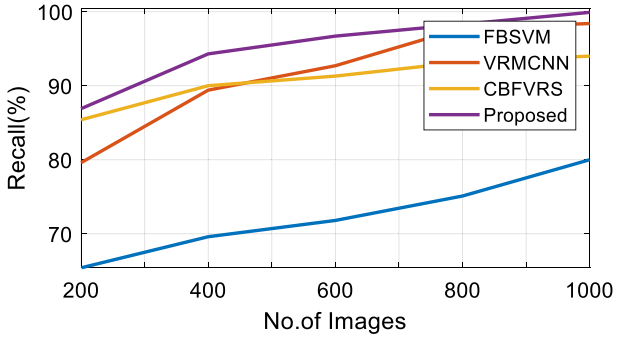


Fig. 8 Comparison for recall

Figure 9 shows the resultant plot for comparison for F1-score with prior methodologies such as FBSVM, VRMCNN, CBFVRS. The figure reveals that the proposed framework obtained a high F1-score when compared with previous techniques.

Above Table 4 shows the comparison for various methods with various parameters. The number of images is varying from 200, 400, 600, 800, and 1000. The proposed method achieves 99.89% accuracy when compared with the prior method as FBSVM achieves 95.4% accuracy, VRMCNN achieves 99.8% accuracy and CBFVRS achieves 99.84% accuracy. The

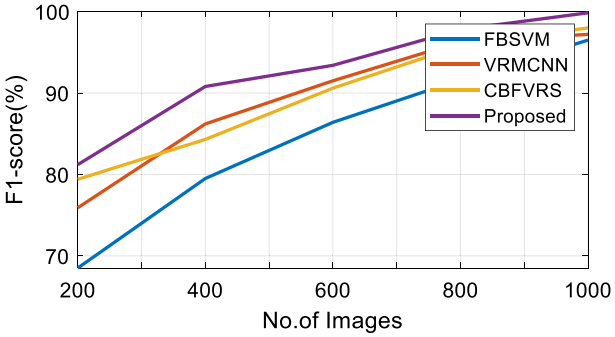


Fig. 9 Comparison for F1-score

Table 4 Shows a comparison for various parameters with various methodologies

Methodologies	Number of images	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FBSVM	200	68.2	79.5	65.4	68.5
	400	71.5	83.4	696	79.5
	600	79.	90.7	71.8	86.4
	800	89.9	94.9	75.1	91.65
	1000	95.4	97.2	80	96.5
VRMCNN	200	75.6	81.6	79.6	75.9
	400	80.7	85.6	89.4	86.2
	600	85.4	87.3	92.7	91.5
	800	93.8	89.2	97.8	96.3
	1000	99.8	90	98.4	97.2
CBFVRS	200	82.6	82.1	85.4	79.4
	400	89.5	85.4	90	84.3
	600	92.4	90.7	91.3	90.6
	800	96.8	92.6	93.2	95.8
	1000	99.84	94	94	98
Proposed	200	89.5	89.2	86.9	81.2
	400	91.6	93.4	94.3	90.8
	600	95.62	95.9	96.7	93.4
	800	98.2	97.6	98.3	97.8
	1000	99.89	99.9	99.91	99.85

precision of the proposed method obtained 99.9% when compared with the previous technique FBSVM has 97.2%, VRMCNN has 90%, and CBFVRS has 94%. Recall attained with the prior technique as FBSVM has 80%, VRMCNN has 98.4%, and CBFVRS has 94% and proposed method obtained 99.91%. F1-score obtained in existing methods as FBSVM has 96.5%, VRMCNN has 97.2%, and CBFVRS has 98% and the proposed method obtained 99.85%. Thus the proposed method achieves higher efficiency in video indexing and retrieval technique when compared with the previous technique.

5 Conclusion

In video indexing and retrieval in Image Processing and deep learning is used to enhance the storage capacity for storing the keyframes from the video along with diminishing the space and

time complexity in capturing angular faces from videos. Hence, to mitigate these issues, the Viola-Jones Algorithm has been proposed for use as a face detector, a novel hybrid sliding windowing and LGFA technique has been introduced, as well as the EAN 8 bar code is utilized as a bar code to face index. In addition, the case of illumination of the invariant face image is improvised after using this form and it is computationally straightforward. Furthermore, the proposed method enhanced the storage capacity and also minimized the time and space complexity. Thus, the proposed method is efficiently enhanced in video indexing and retrieval techniques. The test was conducted on the Hollywood video dataset, YouTube video data array, and a video set for TV shows. The video indexing scheme can be used for an account to define authentication, affirmation, individual search. In the future, we will create a linear barcode from angular face images.

Declarations

Conflict of interest None.

References

1. Adeyemo J, Oyebode O, Stretch D (2018) River flow forecasting using an improved artificial neural network. EVOLVE-A Bridge Between Probability, Set Oriented Numerics, and Evolutionary Computation VI. Springer, Cham 179–193
2. Anayat S, Sikandar A, Rasheed SA, Butt S (2020) A deep analysis of image based video searching techniques. Int J Wirel Microw Technol (IJWMT) 10(4):39–48
3. Bahroun S, Abed R, Zagrouba E (2020) KS-FQA: Keyframe selection based on face quality assessment for efficient face recognition in video. IET Image Process 15:77–90
4. Bastanfard A, Takahashi H, Nakajima M (2004) Toward E-appearance of human face and hair by age, expression and rejuvenation. International conference on Cyberworlds. IEEE
5. Bastanfard A, Bastanfard O, Takahashi H, Nakajima M (2004) Toward anthropometrics simulation of face rejuvenation and skin cosmetic. Computer Animation and Virtual Worlds 15(3–4):347–352
6. Deng Y, Yu Y (2019) Self-feedback image retrieval algorithm based on annular color moments. EURASIP Journal on Image and Video Processing 2019(1):1–13
7. Dong Z, Wei J, Chen X, Zheng P (2020) Face detection in security monitoring based on artificial intelligence video retrieval technology. IEEE Access 8:63421–63433
8. Dutta G (2021) Create caption by extracting features from image and video using deep learning model
9. Feng Y, Zhou P, Xu J, Ji S, Wu D (2018) Video big data retrieval over media cloud: a context-aware online learning approach. IEEE Transactions on Multimedia 21(7):1762–1777
10. Gawande U, Hajari K, Golhar Y (2020) Deep learning approach to key frame detection in human action videos. Recent Trends in Computational Intelligence. IntechOpen
11. Gayathri N, Mahesh K (2020) Improved fuzzy-based SVM classification system using feature extraction for video indexing and retrieval. International Journal of Fuzzy Systems 22:1716–1729
12. Ghatak S, Bhattacharjee D (2020) Video indexing through human face, present the paper in third international conference on communications, circuits and systems held at school of electronics engineering. Kalinga Institute of Industrial Technology, Bhubaneswar, Video Indexing Through Human Face.
13. Hoy MB (2018) Deep learning and online video: advances in transcription, automated indexing, and manipulation. Medical reference services quarterly 37(3):300–305
14. Jacob J, Sudheep Elayidom M, Devassia VP (2020) Video content analysis and retrieval system using video storytelling and indexing techniques. International Journal of Electrical & Computer Engineering 10(6): 6019
15. Ji LY, Yang Z (2017) Design and implementation of medication recommending system for chronic hepatitis B. Chinese Medical Equipment Journal 38(7):48–51
16. Kratochvíl, Miroslav, et al. (2020) Som-hunter: video browsing with relevance-to-som feedback loop. International conference on multimedia modeling. Springer, Cham, SOM-Hunter: Video Browsing with Relevance-to-SOM Feedback Loop

17. Krishnaraj N, Elhoseny M, Lydia EL, Shankar K, and Aldabbas O (2020) An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in cloud environment. Software: Practice and Experience
18. Kumar GN, Reddy VSK (2019) Key frame extraction using rough set theory for video retrieval. In Soft Computing and Signal Processing:751–757
19. Li C, Zhou B (2020) Fast key-frame image retrieval of intelligent city security video based on deep feature coding in high concurrent network environment. Journal of ambient intelligence and humanized computing 1-9.
20. Li D, Liao X, Xiang T, Wu J, Le J (2020) Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. Computers & Security 90:101701
21. Lin FC, Ngo HH, Dow CR (2020) A cloud-based face video retrieval system with deep learning. J Supercomput 76(11):8473–8493
22. Rossetto L, Gasser R, Lokoc J, Bailer W, Schoeffmann K, Muenzer B, Soucek T, Nguyen PA, Bolettieri P, Leibetseder A, Vrochidis S (2020) Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Transactions on Multimedia 23:243–256
23. Rossetto L et al. (2021) Interactive video retrieval in the age of deep learning – detailed evaluation of VBS 2019. In IEEE transactions on multimedia 23: 243-256
24. Saritha RR, Paul V, Kumar PG (2019) Content based image retrieval using deep learning process. Clust Comput 22(2):4187–4200
25. Sauter L et al. (2020) Combining boolean and multimedia retrieval in vitrivr for large-scale video search. International conference on multimedia modeling. Springer, Cham, Combining Boolean and Multimedia Retrieval in vitrivr for Large-Scale Video Search
26. Tian H, Tao Y, Pouyanfar S, Chen SC, Shyu ML (2019) Multimodal deep representation learning for video classification. World Wide Web 22(3):1325–1341
27. Ullah A, Muhammad K, Hussain T, Baik SW, De Albuquerque VHC (2020) Event-oriented 3d convolutional features selection and hash codes generation using pca for video retrieval. IEEE Access 8: 196529–196540
28. Yan C, Gong B, Wei Y, Gao Y (2020) Deep multi-view enhancement hashing for image retrieval. IEEE Trans Pattern Anal Mach Intell 43:1445–1451
29. Yürekli A, Bilge A, Kaleli C (2021) Exploring playlist titles for cold-start music recommendation: an effectiveness analysis. Journal of Ambient Intelligence and Humanized Computing 1–20
30. Zhang C, Lin Y, Zhu L, Liu A, Zhang Z, Huang F (2019) CNN-VWII: an efficient approach for large-scale video retrieval by image queries. Pattern Recogn Lett 123:82–88