Journal of Information Systems Engineering and Management
2025, 10(16s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Video Indexing through Human Faces by Combined Deep Learning Neural Networks

Sanjoy Ghatak and Debotosh Bhattacharjee
[1] *Sikkim Manipal University, Majhitar, 737136, India*
sanjoy1cs@yahoo.co.in
[2] *Jadavpur University, Kolkata-700052, India*
debotoshb@hotmail.com

ARTICLE INFO

Received: 02 Dec 2024

Revised: 22 Jan 2025

Accepted: 06 Feb 2025

ABSTRACT

This research aims to suggest an algorithm that uses the human face as a cue for detecting faces and recognition from input video. Face recognition has become popular because it has various applications, such as information security, smart cards, video surveillance, and law enforcement. The suggested approach, divided into two parts, combines the Multitask Convolution Neural Network (MTCNN) and Shuffle Net. Face detection from a video series using MTCNN and Shuffle Net is the first phase, and face recognition using eigenvalue recognition—a method for identifying faces using principal component analysis (PCA)—is the second. A Multitask convolution neural network with Shuffle Net and Eigen face recognition, a method for facial identification utilizing principal component analysis (PCA), is used to carry out these two processes, a deep learning-based tending type of neural network. Numerous experiments are run on various test datasets to assess the suggested strategy. The outcomes of the simulations are very intriguing and demonstrate how practical the suggested approach is. The human face plays a crucial role in applications like security systems, credit and debit card verification, and surveillance on identifying illegal public venues. So, face recognition is one of the most crucial techniques for video indexing. Facial recognition is becoming increasingly crucial in many aspects of our lives, such as security (discovering missing children speeds up searches for missing persons), attendance, healthcare, the retail sector, and banking. Detecting and recognizing faces are utilized for indexing after the human face has been identified in the input video. Using this video indexing method, we can quickly and effectively search for human activity in the input video. The suggested face detection approach is contrasted with the MTCNN and Shuffle Net algorithms for video indexing. After comparing, it is found that using the combined MTCNN and Shuffle Net algorithm for face detection is more effective and time-saving than just MTCNN and Shuffle net. Using the combined MTCNN and Shuffle Net method, more faces are discovered. Among other advantages, Eigen face recognition using PCA is straightforward, efficient, and precise. It can work with low-resolution photographs and adjust to variations in lighting, facial expressions, and head tilt. However, it has some limitations, such as its sensitivity to changes in face size and shape and its inability to handle partial faces, occlusions, or disguises.CNN also uses facial recognition, but the volume of images is where they confront their biggest obstacle. CNN needs many training images to attain a high level of recognition accuracy. Eigen facial recognition using PCA produces good results on fewer training images. For the indicated strategy, it is 99.35%. Human faces in the input video are the results of face detection and identification, and videos are indexed using these faces as cues.

Keywords: Video Indexing, MTCNN, Shuffle Net, Keyframe, Human face, Shuffle channel, P-Net, R-Net, O-Net, PCA.

1 INTRODUCTION

The study of biometrics is both highly intriguing and intricate. It is feasible to distinguish between individuals employing advanced mathematical methodologies, necessitating the operation within a highly diverse environment through the utilisation of biometrics. The wide range of facial recognition algorithms that have been developed also reflects this diversity. The human visage constitutes a complex, multivariate structure that can convey significant information regarding an individual, encompassing their expressions, emotions, and distinctive facial characteristics [1]. Over the past few decades, face recognition has been thoroughly investigated, and facial data analysis has grown to be a challenging and time-consuming task [2]. In addition to face recognition and identification, these technologies are highly valued in other domains, including robotic manufacturing [4], clinical psychology [5], multimedia [6], intelligent security [3], and automotive security [7]. Recent advances in memory and processor speed have made it possible to process video in real time for computer vision tasks. One of the most well-known face detection and identification algorithms developed as a result of this advancement is convolutional neural networks (CNN), which have greatly improved the performance of computer vision tasks [8].

The change in facial posture is still an issue for face recognition systems, even with CNN's benefits. Additional investigations are offered in order to lessen this problem. A method for developing two CNN models that compute the posture distribution of the training data and correspond to the frontal and profile faces was presented by Masi et al. [9]. Similarly, to represent align-free faces, Liao et al. developed a partial face recognition localisation method using multi-key-point descriptors [10], where the size of the descriptors was determined by the face image and picture content. This research [11] aims to investigate in detail the usage of the local binary pattern (LBP) in conjunction with the convolutional neural network (CNN) for real-time face detection and person recognition. This is because deep learning techniques do exceptionally well on a variety of identification and recognition tasks. The CNN technique actually performs better as the number of epochs grows, stabilising until it reaches the desired learning rate. Viola and Jones [12] developed a cascade face detector that uses AdaBoost and Haar-Like features to train cascaded classifiers that perform well in real time. Even with more sophisticated features and classifiers, this detector may experience multiple degradations in real-world applications with more visual diversity of human faces, according to a number of articles [13, 14, 15]. When incorporated into the cascade structure for face identification in [16, 17, 18], deformable part models (DPM) yield exceptional results.

But they are computationally costly, and they usually need costly annotation at the training stage. In a number of computer vision applications, including image classification and face recognition, convolutional neural networks (CNNs) have made substantial strides in recent years [19–20]. The effectiveness of CNNs in computer vision tasks has led to the release of certain CNN-based face detection systems in recent years. Using deep learning neural networks with high reactivity in face regions, Yang et al. [21] produce candidate windows of faces for facial attribute recognition. In reality, though, this approach takes a lot of time because of the intricate CNN structure. In order to identify faces, Li et al. [22] use cascaded CNNs. But their method ignores the inherent relationship between bounding box regression and facial landmark localisation, instead requiring bounding box calibration from face detection at an additional computational cost. The face's alignment is also very important. There is considerable use of template-fitting techniques [26, 27, 28] and regression-based approaches [23–25]. To improve the performance of face alignment, Zhang et al. [28] recently suggested using facial attribute identification as an auxiliary job using deep convolutional neural networks. The majority of face alignment and identification techniques, however, do not take into consideration the obvious relationship between these two tasks. They have limits even if many works attempt to solve them together. For example, Chen et al. [29] use the properties of pixel value difference to simultaneously conduct alignment and detection with random forests. The use of handcrafted parts, however, limits the performance. Zhang et al.'s [30] multitask CNN improves multi-view face identification accuracy. The early detection windows created by a poor face detector, still continue to restrict the accuracy.

Yet, in contrast, enhancing the detector's effectiveness during training requires mining complex samples. On the other hand, traditional hard sample mining usually takes place offline, which greatly increases the number of manual processes. It would be ideal to create an online method for hard sample mining that automatically adjusts to the current training procedure for face alignment and detection. It is proposed in Paper [31] to use multitask learning and unified cascaded CNNs to integrate these two tasks. There are three stages in the proposed CNNs. It uses a shallow CNN to rapidly generate candidate windows in the first stage. A more sophisticated CNN is then used to refine the windows in order to reject a significant portion of the non-face windows. It then outputs the positions of the face landmarks after refining the results with a stronger CNN. A paper titled "Video Indexing through human face" is discussed [32]. In Paper [32,34], faces are detected by the Viola-Jones algorithm, which makes use of the AdaBoost and Haar-Like features.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun created the highly computationally efficient CNN architecture known as Shuffle Net [35] especially for mobile devices with extremely low processing power (such as

10-150 MFLOPs). By employing two novel operations—pointwise group convolution and channel shuffle—the new architecture dramatically reduces computation costs while maintaining object detection (face) accuracy. In this paper, it is mentioned that out of all the leading and well-known designs, Shuffle Net is one of the finest for compact devices. However, some essential frames (faces) are skipped while recognizing faces in real-time video. So, face detection accuracy is reduced when only a shuffle net is employed. In order to overcome these problems, the paper "Video indexing through human face as a cue using Combined Shuffle net and MTCNN algorithm and recognition of face image" suggests an innovative approach. In this work, essential problems such posture change, illumination invariant aspects, facial angular changes, increased face detection, storage capacity limitation, inability to store crucial frames from movies, and time and spatial complexity are addressed. The research's main contribution is the recognition of face pictures and video indexing that uses the human face as a cue by combining the MTCNN and shuffle net algorithms.

- To extract the frames from the frame and get around the key frame storage problems, cropped facial portraits were created using the MTCNN, Shuffle Net, and Combined Shuffle Net algorithms.
- Using the combined MTCNN and Shuffle Net method, shorten the face detection time and increase the number of faces detected from the video for indexing the input video using the face as a cue.
- Then, using Eigen face recognition, a method for facial identification that uses principal component analysis (PCA) for security (such as detecting criminals and preventing crimes, finding missing children, and accelerating investigations), monitoring attendance, healthcare, the retail industry, and finance from the input video, the face is recognized.

Therefore, problems with posture change, storage capacity limitations, storing video key frames, and time and space complexity are all addressed by the suggested method of indexing videos.

This is how the remainder of the paper is structured: In Section 2, the relevant works from the previous Methodology are described, the suggested methodology and the roles of the different algorithms are explained in Section 3, results and discussion of the experiment are presented in Section 4, and the offered research is concluded in Section 5.

2. LITERATURE SURVEY

The development of digital technology, web streaming, and social networking has made it possible for more people to alter video objects and try to utilise them for a wider range of purposes. Experts are particularly interested in camera expressions since they are so common daily. Because of this, a person's face is now regarded as an essential attribute for video indexing. Face detection techniques have advanced significantly recently thanks to the technology's quick development and algorithmic advancements. Many CNN-based methods have been proposed recently. These techniques perform better, thanks to their strong discriminative abilities. Three data sources or modalities are taken into account while creating a video report for video indexing, following the publication [36]. Semantic analysis-based video search was the primary focus of this study. These three forms of communication consist of written, auditory, and visual modes. This study discusses the three components of a video—visual, audio, and textual—as well as their unique characteristics. The hybrid's face recognition The work [37] addressed the clustering of video system indexing using the hidden Markov model and supporting vector machine architecture. The five components of the human face—the forehead, eye, nose, mouth, and chin—are categorised in this study using the SVM in order to find any independent characteristics of each component. A.I. algorithms are then used in paper [38] to break the video's narrative structure into smaller, easier-to-understand segments so that viewers may swiftly read the content and find a certain video part. Using discourse interactions and visual substance-extricated labels, a deep learning architecture was developed to divide the film into narratives and comment on related major outlines.This structure included verbal discussion, literary content, audio, and images. provided a technique for video indexing and retrieval using a sparse representation of the Bag-of-Faces [39].This technique encodes a face track as a sparse representation of a single bag of faces, which efficiently processes a lot of face data. With an emphasis on deep learning of the binary hash format, Mingtao Pei [40] created a face video retrieval system.The researcher created a deep convolution neural network (deep CNN) that can learn from compact and discriminative binary representations using face-to-face video retrieval.

Publications [41] and [42] focused on the significant intraclass face variability and the pressing need for time and space conservation, which this work addressed and resolved.A article [43] discusses a solution to the issue of low-level video indexing-related programs.An integrated video indexing system that combines face detection and face recognition was the subject of this investigation.In order to identify and recognise faces, the researcher used a neural network, a pseudo-two-dimensional HMM, and a K-Means clustering algorithm.The faces are untraceable using this procedure.Speech is one of the essential indexing elements in lip-based audio-based video indexing.In the audio modality, speech fluctuation presents a barrier for video indexing.The technique to control this variability was explained in the research paper [44].It entailed using the temporal analysis of lip movements to choose key frames from movies.A method for automatically identifying a human face in a generic video series was covered in a publication [45].To provide a confidence estimate for the presence or absence of faces in the video images, the author employed an iterative method.A unique method for processing, categorising, and retrieving independent data from video using the human face is proposed in Paper [32].Keyframe extraction from information videos, face recognition from keyframes, face differentiating proof using standardised identification, and barcode-based video sorting are the primary areas of attention for this work.The study by Gayathri et al. [46] found a number of drawbacks to preprocessing video frames prior to accessing private video libraries.In order to avoid preprocessing errors, feature extraction and classification techniques are considered.Video indexing, which provides a variety of extraction options and standard frame construction for the incoming video frame, has been expected to be used in this scenario.The frame structures are categorised into dominating structures using a fuzzy-based SVM classifier.The multidimensional histogram of directional gradients (HOGs) and colour attribute extraction are used to remove texture information from a video clip.Classifiers using this approach are unable to focus on signal descriptions for video processing applications due to storage capacity constraints.

Lin et al. [47] claim that artefacts in deep neural networks, especially for facial recognition, are commonly detected by profound learning models.This study recommends a deep learning cloud-based video recovery technique as a consequence.The dataset is then extracted and preprocessed to produce a dataset suitable for CNN templates after the remaining images have been matched.The final dataset is then generated to pre-train the CNN models called Face Net, Arc Face, and VGG Face for face recognition.Neither the system's efficiency nor the ability to acquire additional datasets are improved by this.To solve the problems in the link management program's smart town security video retrieval, Li et al. [48] initially suggested a traffic location quantisation index based on backbone traffic characteristics to quantitatively assess the traffic region characteristics in the back-end communication.Important frame abstraction and deep learning-based retrieval are suggested in order to improve the effectiveness and accuracy of video recovery.Key frame features are extracted using the convolutional neural network architecture, an adaptive key frame selection technique is created, and supervised, semi-supervised, and unsupervised retraining models are built.The technique does not save keyframes, nor does it maintain space and time complexity.The authors of Paper [49] are Bastanfard et al. This research introduces an approach called E-appearance to predict the main effects of facial pictures with different looks.Given the same person's facial image data in two different appearances, the quotient image captures the appearance characteristic of the image.A new facial appearance is then created by applying the trait to the face of a different individual utilising a warping process.The enormous variations in lighting, facial expressions, and other elements may make this technique inaccurately depict people's traits.In the study by Bastanfard et al. [50], face regeneration is modelled using two distinct approaches.The wrinkle-painting technique is used to remove wrinkles, while the face anthropometric theory is used to explain facial deformation.For example, we must be able to evaluate the differences in facial features between young and elderly when presented with a range of various faces in order to create a collection of outlines that will form the basis of the Face Rejuvenationsimulation.Due to significant variations in lighting, facial expressions, and other elements, this metho d may not precisely capture people's traits.

According to Dutta et al. [51], the objective is to use deep learning to detect attributes in order to create a statement for a picture.The same method used for image captioning will also be used to generate a sentence for video frames.While the video is being made, keyframes can be recovered from it using the program's integrated critical frame extraction architecture.The same image captioning model that is used to generate captions for images is applied to the keyframes that were extracted from the video. The video frame images are entered into a preprogrammed, pertained model to generate captions from the video frames.Nonetheless, there were other challenges, such as translating video pictures into a coherent frame sequence and generating suitable speech patterns when confronted with a large vocabulary.[52] Jacob et al. By integrating video storytelling with indexing techniques, this study presents a novel method for analysing video content and extracting the needed video clip from a long video. To make a video explanation, the video storytelling technique is paired with video content analysis.The wormhole approach is then used to construct an index from the video description, ensuring that a keyword of a specific length may be located as quickly as feasible. Due to the term's frequent occurrence in the video index's keyword search, video search engines may use this index to locate the needed episode. The user has the option to transfer only the pertinent video clip rather than downloading and sending the full video.This process may therefore take some time.Despite the fact that cloud services offer effective picture indexing, Krishna Raj et al. [28] found that this issue still exists since the user query's semantics do not match the diverse semantics of the large database. A visual semantic indexing-based RTI paradigm for cloud platform photography will be illustrated in this research.The standard semantic and visual descriptor space is first established by an interactive optimisation model. An RTI architecture is then combined with the semantic visual space-sharing model to identify the best way to search for larger data sets.Lastly, the distributed model Spark is expanded to include an online picture retrieval service.

The performance of the suggested system is evaluated using two popular datasets, Holidays 1 M and Oxford 5 K, in terms of average precision (mAP) and processing time across a range of data set sizes.However, neither machine learning nor calculation performance are improved by this approach. Using the channel shuffle function, Paper [35] proposes a novel Shuffle Net unit that is particularly well suited for small networks.On mobile devices and other small devices, it reduces the amount of time needed for real-time object detection (face), but it skips more objects (face) in the process.[46] Storage space is limited, and classifiers made expressly for video processing cannot be used to establish signals.[47] An attempt has yet to be made to create classifiers that can collect more information or perform better.[48] It fails to save important frames and is unable to deal with the challenges of time and space.Due to the wide variations in lighting, facial expressions, and other factors, these techniques may not be able to capture faces [49, 50].Video picture conversion into a meaningful sequence of frames is challenging [51], time-consuming [52], and [53] neither the machine learning method nor the computation time is improved.

An approach for video indexing utilizing face detection and identification is discussed in the paper [54]. In this study, human faces are first recognized using a 2-D hidden Markov model and DCT coefficients after being recognized from the input video using a neural network-based method. However, these are time-consuming, outdated methods. If there are more image data sets, it will not function effectively. The face identification hybrid Hidden Markov model and its supporting vector machine-based clustering of video system indexing were discussed in paper [55].In this work, the SVM is used to classify and search for features in the five parts of the human face: the chin, nose, mouth, eye, and forehead. The human face is separated into these five sections.According to the study [56], AI algorithms are then used to divide the movie's narrative structure into more manageable, shorter segments, allowing users to quickly view the information and retrieve a particular video clip.Using visual substance-extricated labels and discourse interactions, a deep learning architecture was created to separate the film into narratives and provide commentary on pertinent, significant topics. Literary substance, sound, pictures, and discourse were all incorporated into this architecture.A sparse version of the Bag-of-Faces was used by the author to suggest a method for indexing and retrieving video [57].This method encoded a face track as a sparse representation of a single bag of faces, which allowed it to analyse large amounts of face information efficiently.Mingtao Pei [58] developed a face video retrieval system that focusses heavily on deep learning of the binary hash representation.In this paper, the researcher developed a deep convolution neural network (deep CNN) to learn from face-to-face video retrieval discriminative and compact binary representations.The problems mentioned in studies [59] and [60] (significant intraclass face variability and the urgent need to conserve time and space) were addressed and handled in the research.A solution to the problem of low-level programs connected to video indexing is examined in a paper [61].In this study, the author described a video indexing system that uses face detection and identification techniques.The researcher employed a neural network-based face identification and recognition system, a pseudo-two-dimensional HMM, and a K-Means clustering technique for face detection.In this manner, it is impossible to trace the faces. In audio-based video indexing, which depends on the lips, speech is one of the most important indexing variables. A challenge for audio modality video indexing is speech variance.The author of the paper [62] described a method for handling this unpredictability by using the temporal analysis of the lip movements to choose key frames from video.The research explored a technique for automatically recognising human faces in a generic video series [63].The presence or absence of faces in the video pictures was given a confidence measure by the author using an iterative method.

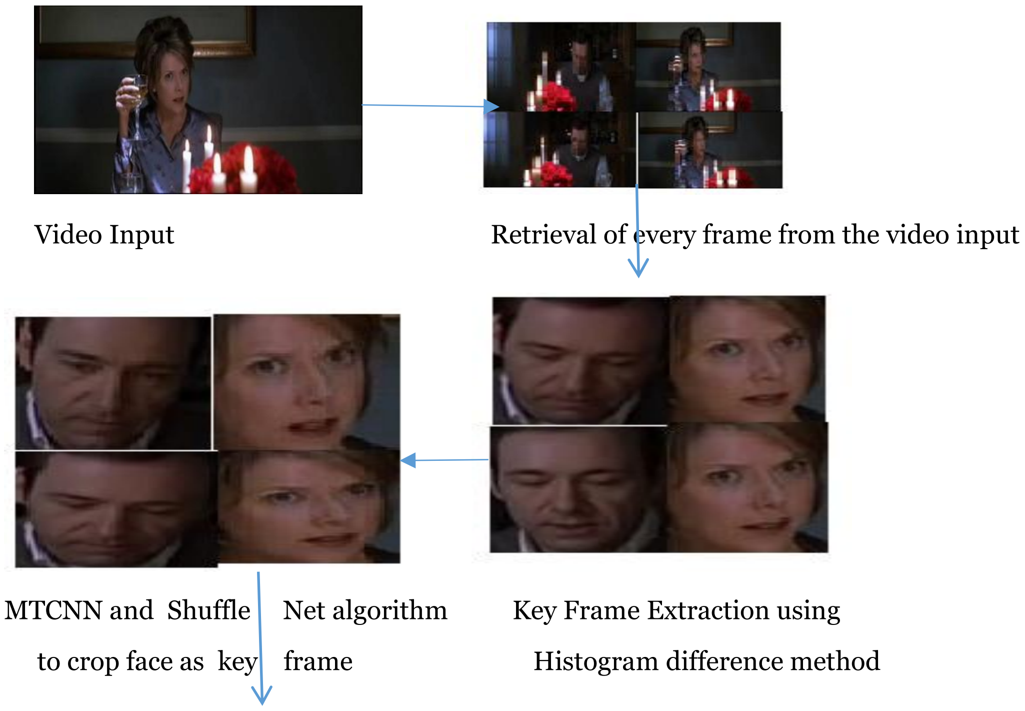
The author of the research [64] investigated software areas, possible issues with video ordering and recovery, possible headings for video ordering, and executive system video recovery.The majority of video-based identification systems on the market today [65] try to do the following: recognise the face first, then follow it throughout time.Recognition using the still-to-even-now technique is only possible if a casing that meets certain criteria (size and posture) has been obtained.For this purpose, the face portion is removed from the edge and modified or enrolled as necessary.

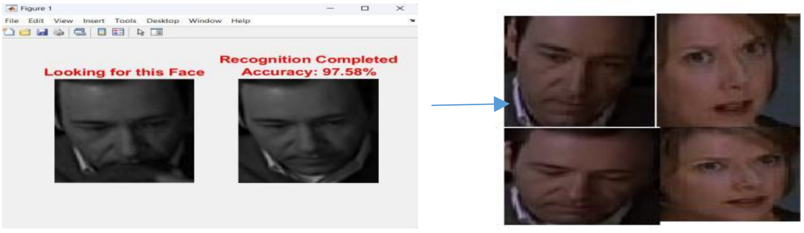
The method of acknowledging after something happens has several unanswered problems.The author of the study [66] addressed this problem in "Probabilistic Recognition of Human Face from the Video."A novel approach to processing, classifying, and recovering independent data from video through the human face is proposed in Paper [32].The main goals of this research effort are to extract key frames from the information video, recognise faces from keyframes, use standardised identification to differentiate faces, and use the linear EAN-8 barcode to organise videos.To tackle important problems such as shifting posture, storage capacity constraints, the inability to save important video frames, and the complexity of time and space,In order to address the indexing and retrieval problems with the aforementioned videos, as well as the problems with LGFA and the Sliding Window Technique, Paper [34] proposes a novel technique called Video Indexing utilising Human Face Images.The Keyframe, or human face, is identified in this study using the Viola-Jones technique, and the bar code is generated using the EAN-8 standard.Despite being quicker than MTCNN, the Viola-Jones algorithm has the drawback of being unable to identify an angular face in the video.The MTCNN approach to face detection is still a laborious process, despite being an improvement over the Viola-Jones (discussed in Paper [34]) and Shuffle net (discussed in Paper [35]) methods in Paper [31].

In order to overcome all of these problems, this study suggests a novel method for video indexing (video indexing using the human face as a cue using Combined Shuffle net and MTCNN algorithm and recognition of face picture).This method indexes the face in different video formats and recognises the human face in the input video by combining the MTCNN and Shuffle Net algorithms.The primary tasks of the MTCNN and shuffle net algorithms in this approach are face detection and alignment in angular keyframes.This method also reduces the time and increases the number of keyframes (faces) detected from the input video. Following that, a face is identified using Eigen face recognition, a PCA-based technique for facial identification. The use of detected and recognized faces for indexing follows face detection and recognition.

3. SUGGESTED METHODOLOGY AND THE ROLE OF SEVERAL ALGORITHMS

Fig. 1 depicts a block diagram for each of the steps in the method described in this text. (a) Extracting frames from the input video is the first stage.(b) Keyframes are distinguished from the frames taken from the input video by the difference in the colour histogram.(c) Using MTCNN and the Shuffle Net technique to identify faces in the key frame. (d) Next, Eigen face recognition, a PCA-based method for facial identification, is used to identify a face. (e) Detecting and recognizing faces are used for video indexing purposes.


Video Input
Retrieval of every frame from the video input
MTCNN and Shuffle Net algorithm
to crop face as key frame
Key Frame Extraction using
Histogram difference method



Eigen face recognition, a PCA-based technique for facial identification.

Detected and Recognized faces used for indexing purposes.

Fig. 1: The suggested system's block diagram

3.1 Frame extraction

The combination of the scene, shot, and frame creates dynamic video.Therefore, the first step is to extract the still images, which are represented as a scene, shot, and picture, from the input movies.A shot is a sequence of frames, and a scene is a group of shots.The classic video is particularly information-dense and has a frame rate of 20 to 30 frames per second.The frame, a still image that is a part of a video, contains extraneous information.Figure 2 displays the frame from the Holly Wood movie American Beauty-00222.

Fig.2. The frame of American Beauty -00222 Video

3.2 Extraction of the Key Frame

3.2.1 The colour histogram approach

The Keyframe contains the most important components of every shot.In this piece, human features in various stances, lighting conditions, and illuminations are essential frames.The likelihood scale, the results of the curve saliency motion capture, and a few others are examples of such tactics.However, the key frame is extracted from each frame of a specific film at this point using the Colour Histogram approach.If the detected change is larger than the threshold, keyframes can be retrieved from the frames using the Colour Histogram difference. When the threshold for colour histogram disagreement is reached, the frame is chosen as the following Keyframe.A few key frames from the Hollywood film American Beauty -00222, starring Holly Wood, are shown in Figure 3.

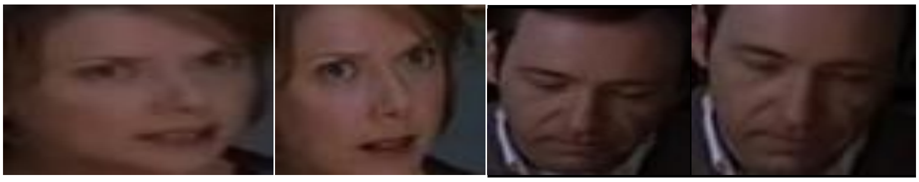


Fig. 3 Using the face as the primary component, Fig. 3's Key Frame for "American Beauty"-00222

In a paper, the formula and algorithm for splitting the colour histogram's two successive frames are mentioned [34].

3.3 Localization of face using a combined MTCNN and Shuffle Net method.

3.3.1 MTCNN Algorithm

From the retrieved keyframes, faces are extracted using the MTCNN object detection technique. The MTCNN algorithm has three steps.In the first stage, a shallow CNN is used to quickly produce candidate windows.The windows are then refined using a more sophisticated CNN in order to reject a significant portion of the non-face windows.Lastly, it enhances the output and placement of facial landmarks by utilising a more potent CNN.

The Proposal Network (P-Net), a fully convolutional network, is used in Stage 1 to get the candidate windows and their bounding box regression vectors.The candidates are calibrated using the bounding box regression vectors that have been calculated.Non-maximum suppression (NMS) is then used to mix highly overlapped candidates.

The Refine Network (R-Net), a distinct CNN that combines NMS candidates, rejects a large number of inefficient candidates, and calibrates using bounding box regression, receives all of the candidates.

Stage 3 (O-Net): This stage aims to provide a more detailed description of the face, despite its similarities to Stage 2.Specifically, the network will output the location of the five face landmarks.

The MTCNN method, which is covered in the publication [31], is broken down into its several steps in the accompanying graphic.

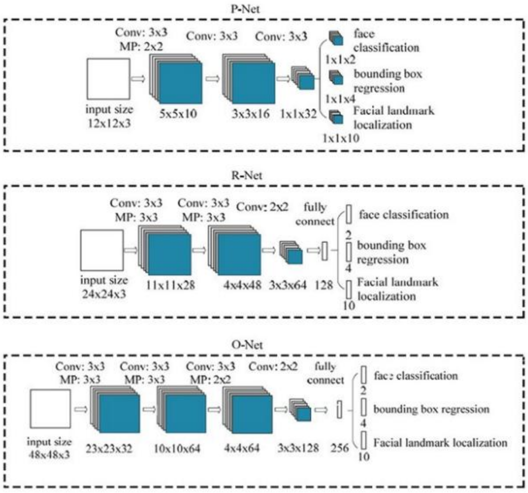


Fig 4 Architectural diagram of MTCNN [31]

3.3.2 Shuffle Net Algorithm

Megvii Inc. launched Shuffle Net [35], a CNN architecture designed for mobile devices with processing capability of 10–15 MFLOPs (also known as Face++).Using pointwise group convolution and channel shuffle, the Shuffle Net reduces computation costs without sacrificing accuracy.In real speed, it beats Alex Net by over 13 times and can categorise photos from ImageNet with fewer top-1 error than the Mobile Net system.It requires few parameters and low computing costs to attain high precision. By incorporating depth-wise separable or group convolutions into the building blocks, Xception and Res Net achieve an outstanding balance between representational ability and computing cost. However, they must partially account for 1x1 convolutions, also known as pointwise convolutions.The number of channels available to meet the complexity criterion is limited by costly pointwise convolutions, which can significantly reduce the accuracy of tiny networks.

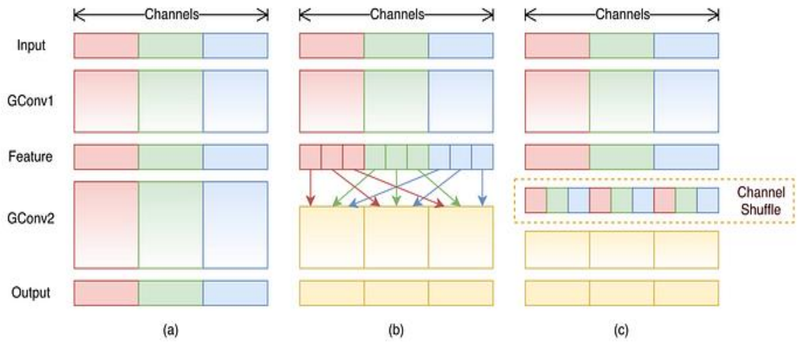


Figure 5: Channel shuffling and two stacked group convolutions. GConv is the acronym for group convolution. a) A pair of convolution layers layered with an identical number of groups. Each output channel only connects to

the input channels in the group. no cross-talk; b) input and output channels are fully connected after GConv1; c) a channel shuffle implementation equal to b) when GConv2 receives data from different groups.[35]

Group convolutions are used as examples in this paper [35] to show how they can lower computing costs. Figure 5(a) shows two stacked group convolution layers, which impair representations and impede information transmission between channel groups.As shown in Figure 5(b), the group convolution can receive input data from several groups.Observe how closely the input and output channels are related. In Figure 5(c), the feature map from the previous group layer is configured and put into practice using a channel shuffle operation. Thanks to channel shuffle operation, multiple group convolutional layers can create more muscular structures.Although the convolutions in Figure 5(b) have different groups, the design in Figure 5(c) is preferred since the channel shuffle operation still applies in the stacked layers. It is also differentiable, which allows it to be included in network topologies.

3.3.3 Shuffle Net and MTCNN Combined Algorithm

To attain great accuracy while keeping a low computational cost, the Shuffle Net and MTCNN combined algorithm combines the two algorithms. Preprocessing the input image with Shuffle Net helps MTCNN run more quickly. The preprocessed image is then subjected to precise face detection using MTCNN. Pseudocode for face detection from input video using a mixed Shuffle Net and MTCNN is discussed in the following pseudocode.

Face detection using a mixed Shuffle Net and MTCNN algorithm.

Input: Input as Video

Output: Face as output from input video

Steps

1. Start.
2. Read the user input video.
3. From the input video, extract the frame.
4. Store the total number of frames in the video.
5. Generate the KeyframeKeyframe (face) from the extracted frame using color histogram difference.
6. It detected the face from the KeyframeKeyframe using MTCNN and the Shuffle Net algorithm.
7. The procedure is terminated when all of the input video's keyframes have been processed, and all faces are found using MTCNN and Shuffle Net.
8. If not, repeat the procedure.
9. Store all faces present in the video.
10. Stop

Fig. 6 shows screenshots of the face detection function in Keyframes using a combined Shuffle Net and MTCNN algorithm.

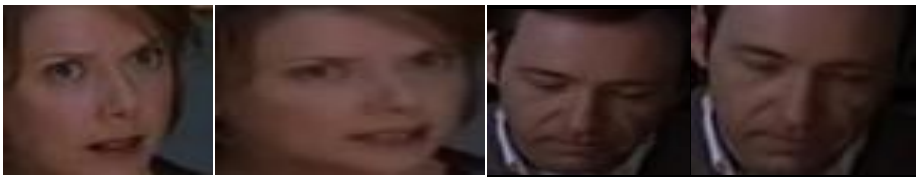


Fig. 6 Key scenes from the movie "Holly Wood's American Beauty-00222 with faces taken out

3.4 Eigen face recognition, a PCA-based method for facial identification, is used to identify a face.

There is debate concerning the advantages and disadvantages of facial recognition technology. While many participants emphasize the advantages, detractors frequently consider the disadvantages. Concerns regarding facial recognition technology include invasion of privacy, abuse of authority, and what rogue government employees might do with it. Facial recognition is getting more media attention than ever before.Due to recent historical events, expenditures in face recognition have rapidly expanded.The global COVID-19 pandemic may lead to increased investment in biometric technologies, such as facial recognition.Due to the high contagiousness of COVID-19, contactless encounters are highly valued.Security precautions remain the primary use of facial recognition technology.In a number of sectors, such as improved public safety, aviation and transportation, retail, access and authentication, faster processing, seamless integration, and more, facial recognition is recognised as one of the most straightforward and dependable techniques for identifying people.Principal Component Analysis (PCA) is a facial recognition technique called Eigen face recognition that uses human faces to identify objects. Using PCA's Eigen face recognition approach, global features are retrieved. Facial recognition and face views are two elements that influence the recognition system's accuracy. The difficulty is that all photos must have the exact dimensions and color depth to ensure accurate feature extraction. The number of images is CNN's major problem because many training images are necessary for the system to acquire high recognition accuracy. Due to this, this work employs the Eigen face recognition method.This approach is based on the idea that faces can be represented as a linear combination of "Eigenfaces" created from a set of training images. The training set's covariance matrix contains these Eigenfaces sorted in ascending order by eigenvalues. The Eigenfaces capture the main aspects of facial images, such as lighting, facial expressions, and head tilt. The system must first be trained using facial photo examples when using Eigen face recognition. The system computes the Eigenfaces and projects each training image into the eigenface space to produce feature vectors representing each face. During the recognition phase, the system compares the feature vectors of an input face with those of the training faces to determine which match is the closest. The recognition performance can be improved by selecting the most crucial eigenvectors and including more training images. Among other advantages, Eigen face recognition is straightforward, efficient, and accurate. It can work with low-resolution photographs and adjust to variations in lighting, facial expressions, and head tilt. However, it has some limitations, such as its sensitivity to changes in face size and shape and its inability to handle partial faces, occlusions, or disguises.

Algorithm for Eigen face identification, which uses principal component analysis (PCA) to identify faces:

Input: Detected face dataset of a human face from the input video.

Output: Identified faces from input video

Steps:

1. Start

2. Open the detected database of human faces from the input video.

3. Pick a random face picture from the face database.

4. Eliminate the selected picture from the database after identifying the selected face.

5. PCA is applied to the remaining images to generate a set of Eigenfaces.

6. Determine each remaining face image's signature.

7. Calculate the image's signature.

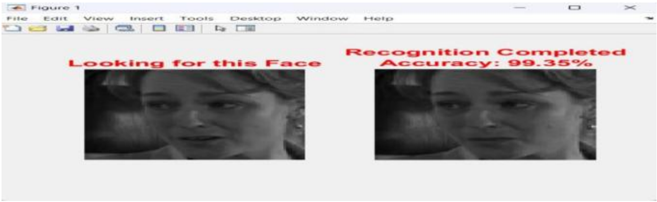
8. Determine the distance between signatures.

9. Look up the nearest image.

10. Determine the accuracy of recognition. The recognition accuracy is determined as (1 - z(i) / (norm (s, 2)) *100 to get the percentage.

11. Stop.

Fig. 7 shows the screenshot of face recognition accuracy from the movie "Holly Wood’s American Beauty-00222 with faces taken out.



Looking for this Face

Recognition Completed
Accuracy: 99.35%

Fig. 7 Screenshot of face recognition accuracy from the movie "Holly Wood’s American Beauty-00222 with faces taken out using Eigen Face Recognition, which uses principal component analysis.

3.5 Detecting and recognizing faces are used for video indexing purposes.

Once the faces in the input video have been identified and detected, the faces are used for video indexing.

4 Results and discussion of the experiment

The effectiveness of the system, comparison strategies, and implementation outcomes are all thoroughly explained in this section.

4.1 Setup for the experiment

The system specification required to do this work in MATLAB 2023a or later is provided below.

The platform should be MATLAB 2023a or later.

Windows O.S.

Processor: a minimum of an i3.

Four gigabytes of RAM.

The KeyframeKeyframe for this study was first acquired from this cropped face in the human face-based video collection. The human face visible in the video is then used for indexing. The method was validated using three separate video data sets. The Hollywood video dataset is used as the project's starting point.There are also video clips from 32 human action flicks in here.At least one of the eight categories must be applied to the sample.A 20-film data set is created by splitting the test set into two 12-film practice sets.The automated learning set consisted of 233 video recordings that were compiled using automatic script-based action labelling, with around 60% of the labels being accurate.There are 211 video examples with manually tested labels and 219 video samples with manually checked labels in a clean training collection of Hollywood results.

Then, using 113 movie trailers from YouTube with a 2010 release date, we created the Movie Trailer Face Dataset, which can be used to train celebrities in our enhanced PublicFig+10 dataset. The Methodology mentioned above was used in these videos to produce face tracks. The final dataset consists of 3,585 face tracks with 514 known identities and 63% unknown identities (not present in PubFig+10). Finally, a realistic T.V. series video dataset of 27 episodes of 6 well-known T.V. shows is available. Breaking Bad (3), How I Met Your Mother (8), Mad Man (3), Modern Family (6), Sons of Anarchy (3), and 24 (4) were among the 27 episodes. These videos are 16 hours long overall. This film contains 6231 and 30 acts and actions in total. It can be seen through a comparison of these algorithms that MTCNN is the most precise and specialized face detection technique. However, due to its high computational cost, it is inappropriate for embedded and mobile systems. Although lightweight and practical, Shuffle Net might not be as precise as MTCNN. The combined method provides an excellent compromise between computational efficiency and precision, making it appropriate for embedded and mobile systems.

It is also observed that a thin neural network architecture called Shuffle Net was created for mobile and embedded devices. Thanks to a group convolutional operation, it maintains good accuracy while requiring less processing and memory. Although Shuffle Net can be trained to recognize faces, its accuracy could not match that of other, more specialized algorithms. MTCNN is a specialized detection technique that finds and locates faces in an image using a multi-stage neural network. It is renowned for its exceptional accuracy and versatility in handling faces with different sizes, orientations, and lighting situations. To attain great accuracy while keeping a low computational cost, the Shuffle Net and MTCNN combined algorithm combines the two algorithms. Preprocessing the input image with Shuffle Net helps MTCNN run more quickly. The preprocessed image is next accurately detected faces using MTCNN.

The results of face detection using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods on various video data sets are shown in the following tables, along with the time needed to find faces. Table 1 shows the number of faces found on various video clips from the Hollywood Data set using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods. Table 2 displays the times for face detection using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods on several video clips from the Hollywood Data set. In the movie trailers, there are videos and T.V. shows. Tables 3 and 4 discuss the number of faces detected in the video data set and the time it takes to find faces using the Shuffle Net, MTCNN, and combined Shuffle Net and MTCNN methods. Face detection is followed by Eigen face recognition, which recognizes faces using principal component analysis

(PCA). The most outstanding accuracy for face identification in this experiment is 99.35%. However, it varies depending on the faces' size, shape, and quality.

Table 1: Table for Number of faces detected in Shuffle net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm

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 2: Table for Execution time (in seconds) in Shuffle net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm

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

Table 3: Table for Number of faces detected in Shuffle net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm of Movie

Trailer Face Dataset and T.V. series video dataset

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 4: Table for Execution time (in seconds) in Shuffle net vs. MTCNN vs. Shuffle Net and MTCNN Combined Algorithm

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

4.2 Comparison strategies

The performance of the proposed approach—the combined Shuffle Net and MTCNN algorithm—is compared to that of the MTCNN and Shuffle Net algorithm in this section. Once face detection is complete, it is evident that using both Shuffle Net and MTCNN results in a more accurate method than using only MTCNN or Shuffle Net alone. With the suggested method, more faces are discovered, and less processing time is needed than with MTCNN and Shuffle Net. The main differences between the combined Shuffle Net and MTCNN, MTCNN, and Shuffle Net algorithms are shown in a graph in Figure 8. This graph contrasts the algorithmic performance using the same movie clip from "Holly Wood Movie." After the experiments, the graph displays how many faces were found by combining the Shuffle Net, MTCNN, and Shuffle Net algorithms. Figure 9's graph illustrates the critical distinctions between the combined Shuffle Net and MTCNN, MTCNN and Shuffle Net algorithms. This graph compares algorithmic performance using the same "Holly Wood Movie video clip." A graph showing the execution time required for face detection using a combination of the Shuffle Net and MTCNN, MTCNN, and Shuffle Net algorithms is presented after the testing. Fig. 10 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, Shuffle Net, and MTCNN combined Algorithm for Face detection in data set of Movie Trailer and T.V. series video data set on number of faces detected.Fig. 11 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, Shuffle Net, and MTCNN combined Algorithm for Execution time required for Face detection in the data set of Movie Trailer and T.V. series video data set.

MTCNN is the most precise and specialized face detection algorithm, according to a comparison of various methods. However, due to its high computational cost, it is inappropriate for embedded and mobile systems. Although lightweight and practical, Shuffle Net could not be as precise as MTCNN. The combined method provides an excellent compromise between computational efficiency and precision, making it appropriate for embedded and mobile systems. In conclusion, the application's requirements will determine the method used. MTCNN is the best option if accuracy is the top requirement. Shuffle Net or the combination method might be preferable if computing efficiency is more crucial for face detection.

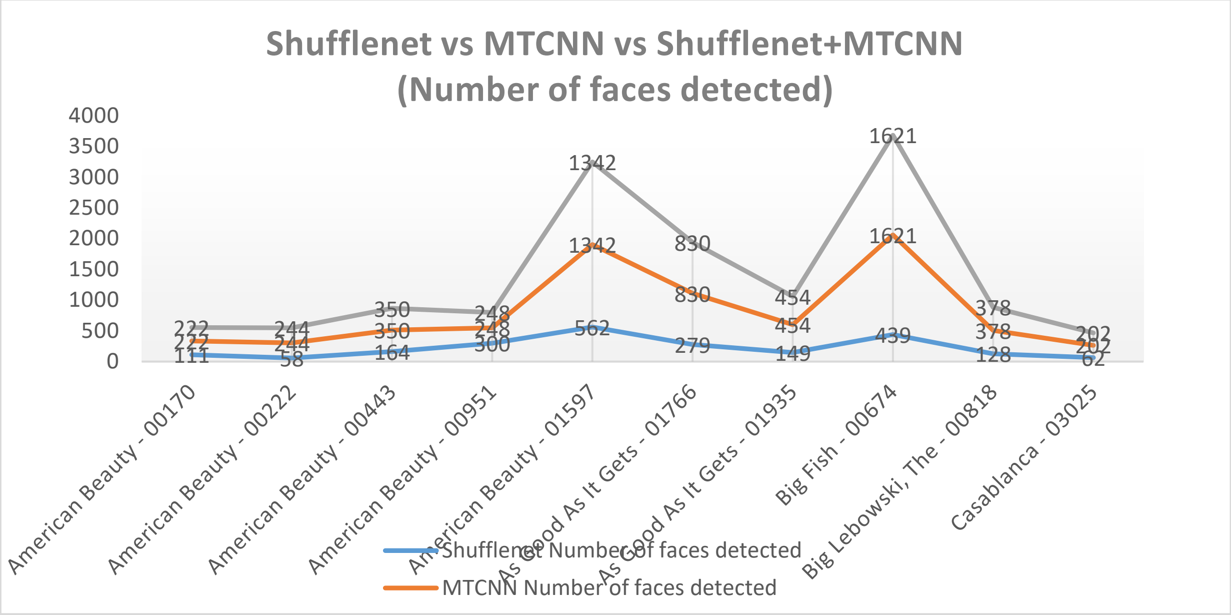


Fig. 8 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, ShuffleNet, and MTCNN combined Algorithm for Face detection in 10 different video clips of Holly Wood video data set on several faces detected.

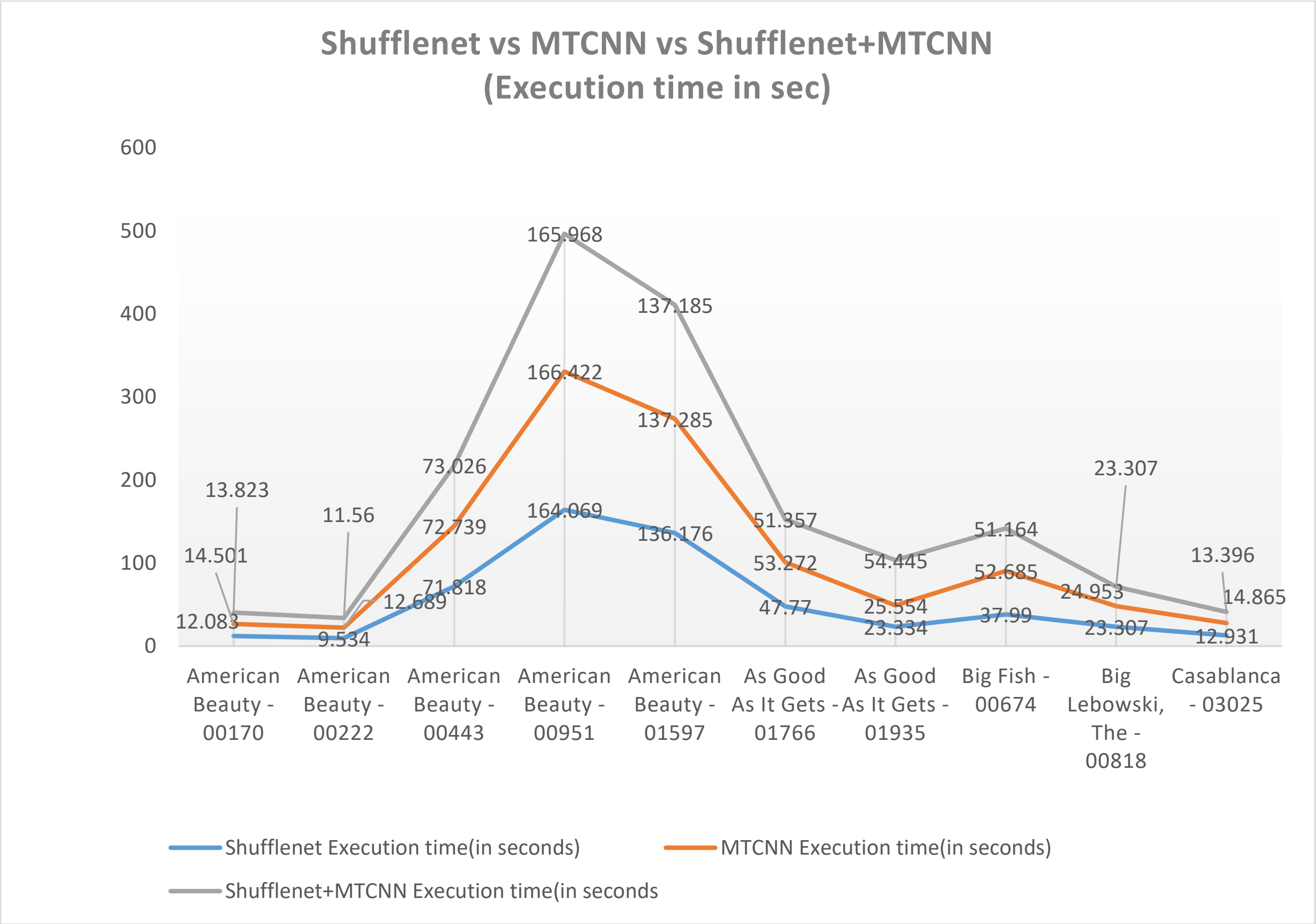


Fig. 9 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, ShuffleNet, and MTCNN combined Algorithm for Execution time required for Face detection in 10 different video clips of Holly Wood video data set.

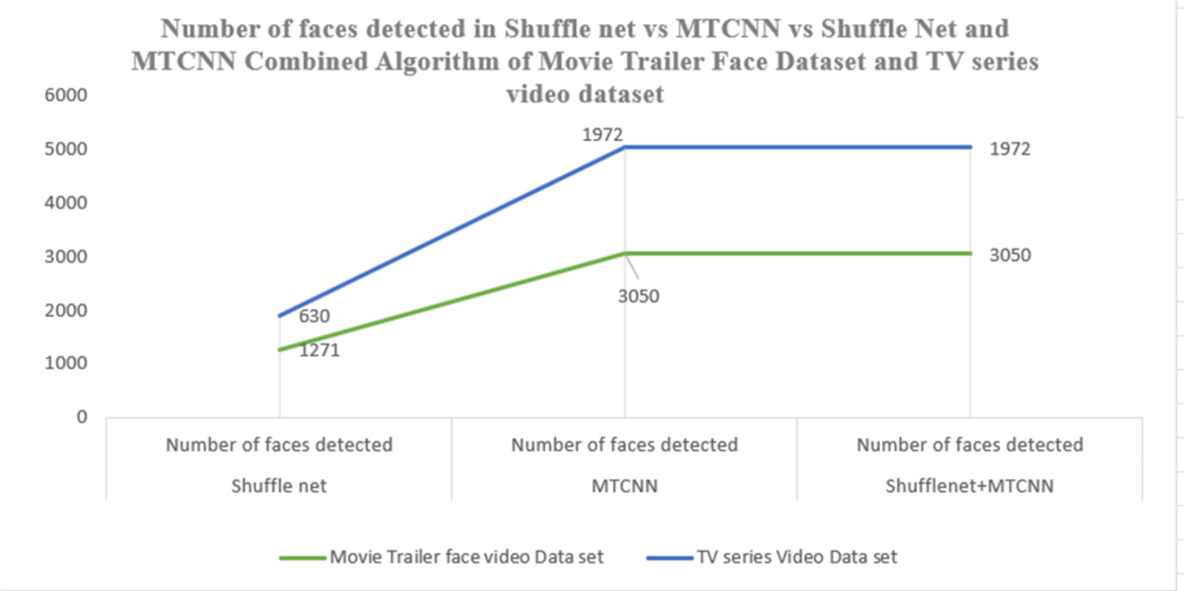


Fig. 10 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, ShuffleNet, and MTCNN combined Algorithm for Face detection in data set of Movie Trailer and T.V. series video data set on number of faces detected.

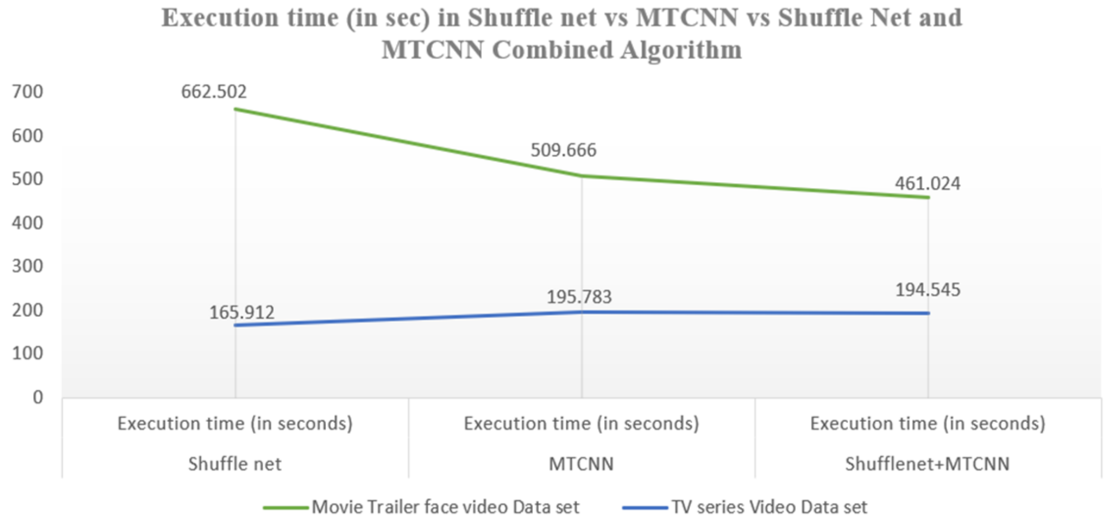


Fig. 11 Graph for Comparative Study between Shuffle Net Algorithm, MTCNN, Shuffle Net, and MTCNN combined Algorithm for Execution time required for Face detection in the data set of Movie Trailer and T.V. series video data set.

It is evident from this comparison that the combined shuffle Net and MTCNN algorithm is the best for face detection from the input video, and the suggested method is employed for this purpose. Face recognition is done using the Eigen face recognition technique after face detection. Once recognition is complete, the detected or recognized faces are used for video indexing from the provided input footage.

4 CONCLUSION

In this study, face recognition accuracy can be raised using Eigenfaces, while face detection accuracy can be increased by combining Shuffle Net and MTCNN. First, faces are detected using a combination of the Shuffle Net and MTCNN methods, and then faces are identified using the Eigen face approach and principal component analysis (PCA). Therefore, the work demonstrates the integration of the two components to produce a complete, reliable, real-time face detection and identification application. The acquired findings demonstrate the suggested method's ability to identify individuals in various situations and in real-time, with a face recognition learning rate of 99.35%. The findings are encouraging because the proposed technique can be implemented on a device due to the short processing time. The Hollywood video dataset, the Movie Trailer face video dataset, and a video collection for T.V. shows have been employed in the test. For an account, the video indexing technique can define authentication, affirmation, and personal search.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] I. Masi, S. Rawls, G. Medioni," Pose-aware face recognition in the wild," Conference on Computer Vision and Pattern Recognition (2016), pp 4838–4846.

[10] 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.

[11] 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.

[12] P. Viola and M.J. Jones, "Robust real-time face detection, International Journal of computer vision," vol.57, no.2, pp.137-154,2004.

[13] 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.

[14] 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," in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 942-949

[15] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.

[16] 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.

[17] 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.

[18] 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.

[19] 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

[20] 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

[21] 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.

[22] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multitask learning," in European Conference on Computer Vision, 2014, pp. 94-108.

[29] 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.

[30] C. Zhang, and Z. Zhang, "Improving multi-view face detection with multitask deep convolutional neural networks," IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036-1041.

[31] 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.

[32] 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

[33] Y. Matveev, G. Kukharev, N. Shchegoleva, 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 (Academic et al., 2014), pp. 213–220.

[34] 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

[35] 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, USA, 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[36] Cees G.M. Snoek, Marcel Worring, "Multimodal video indexing: A Review of state of the art," Multimedia Tools and Applications, 25, 5–35, 2005

[37] Yuehua Wan1, Shiming Ji1, Yi Xie2, Xian Zhang1, and PeijunXie "Video program clustering indexing based on faced recognition hybrid model of Hidden Markov model and support vector machine," IWCIA 2004, LNCS 3322, pp. 739–749, 2004.

[38] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara, "Neural Story: An interactive Multimedia system for Video indexing and re-use," In proceedings of CBIM, Florence, Italy, June 19-21, 2017.

[39] Bor-Chun Chen, Yan_Ying Chen, Yin-His Kuo, Thanh Duc Ngo, Duy-Dinh Le, Shin Ichi Satoh, Winston H Hsu, "Scalable face Track Retrieval in Video archives using Bag-of-faces sparse Representation," IEEE Transactions on Circuits and Systems for video technology,2015

[40] 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 Intelligence (AAAI-16).

[41] Li Y, Wang, R, Huang Z, Shan S, and Chen X, "Face video retrieval with image query via hashing across Euclidean space and Riemannian manifold," In CVPR, 4758-4767, IEEE, 2015b.

[42] Chen Y.C, Patel V.M, Shekhar S., Chellappa R. and Phillips P.J "Video-based face recognition via sparse joint representation," In F.G.,1–8, IEEE, 2013.

[43] Stefan Eickeler, Frank Wallhoff, Uri Iurgel, Gerhard Rigoll, "Content-based indexing of images and video using face detection and recognition methods, "published in ICASSP 2001, IEEE Xplore.

[44] 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.

[45] CsabaCzirjek, 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

[46] 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

[47] Lin FC, Ngo HH, Dow CR (2020) A cloud-based face video retrieval system with deep learning. J Supercomputing 76(11):8473–8493

[48] Li C, Zhou B (2020) Fast keyframe image retrieval of intelligent city security video based on deep feature coding in the high concurrent network environment. Journal of ambient intelligence and humanized computing 1-9.

[49] 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

[50] 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

[51] Dutta G (2021) Create captions by extracting features from images and videos using a deep learning model.

[52] 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

[53] Krishnaraj N, Elhoseny M, Lydia EL, Shankar K, and Aldabbas O (2020) An efficient radix tire-based semantic visual indexing model for large-scale image retrieval in a cloud environment. Software: Practice and Experience

[54] Eickeler, Stefan & Muller, Stefan & Rigoll, Gerhard. (1999). Video Indexing Using Face Detection and Face Recognition Methods.

[55] Yuehua Wan1, Shiming Ji1, Yi Xie2, Xian Zhang1, and PeijunXie "Video program clustering indexing based on faced recognition hybrid model of Hidden Markov model and support vector machine,” IWCIA 2004, LNCS 3322, pp. 739–749, 2004.

[56] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara, "Neural Story: An interactive Multimedia system for Video indexing and re-use," In proceedings of CBIM, Florence, Italy, June 19-21, 2017.

[57]Bor-Chun Chen, Yan_Ying Chen, Yin-His Kuo, Thanh Duc Ngo, Duy-Dinh Le, Shin Ichi Satoh, Winston H Hsu, "Scalable face Track Retrieval in Video archives using Bag-of-faces sparse Representation," IEEE Transactions on Circuits and Systems for video technology,2015

[58] 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 Intelligence (AAAI-16).

[59] Li Y, Wang, R, Huang Z, Shan S, and Chen X, "Face video retrieval with image query via hashing across Euclidean space and Riemannian manifold," In CVPR, 4758-4767, IEEE, 2015b.

[60] Chen Y.C, Patel V.M, Shekhar S., Chellappa R. and Phillips P.J "Video-based face recognition via sparse joint representation," In F.G.,1-8, IEEE, 2013.

[61]Stefan Eickeler, Frank Wallhoff, Uri Iurgel, Gerhard Rigoll, “Content-based indexing of images and video using face detection and recognition methods, “published in ICASSP 2001, IEEE Xplore.

[62] 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.

[63] CsabaCzirjek, 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.

[64] M. Ravinder, T. Venu Gopal, and T. Venkat Narayana Rao, “Video Indexing and Retrieval Applications and Challenges, “Oriental Journal of Computer Science & Technology, Vol. **3**(1), 125-137 (2010).

[65] 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.

[66] Shaohua Zhou, * Volker Krueger, and Rama Chellappa, “Probabilistic recognition of human face from the video, “Computer Vision and Image Understanding 91 (2003) 214–245.