Integrating Deep Learning Models and Facial Recognition for Advanced Intelligent Surveillance Systems

Balamurugan M[1*], Kalaiarasi R[2], Gopinath D[3], Shanthi H J[4]

[1] Associate Professor, Department of MCA, Acharya Institute of Graduate Studies, Bengaluru, Karnataka, India.
[2] Assistant Professor, School of Computer Science, Tamil Nadu Open University, Chennai, Tamil Nadu, India.
[3] Assistant Professor, Department of Computer Science, Kristu Jayanti College (Autonomous), Bengaluru, Karnataka, India.
[4] Associate Professor, Department of MCA, Hindustan Institute of Technology and Science (Deemed to be University), Chennai, Tamil Nadu, India.
tecbalamurugan@gmail.com

Abstract. It is targeted to automate the whole surveillance systems by the deep learning face recognition. The primary objective of the first part is to develop an intelligent surveillance system that applies deep learning algorithms for timely processing data types and identify patterns. The integrated system utilizes the capabilities of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to support strong object identification and tracking, allowing for real-time monitoring in different environments. In the second component, the study deals with the integration of facial recognition techniques into the monitoring system. The system can recognize and monitor people in different scenarios as deep learning-based face recognition models are incorporated into the system. This integration enhances the ability of the system to recognize specific individuals of interest, allowing for a more proactive and targeted surveillance posture. In addition to automating surveillance procedures, the integrated strategy forms an intelligent system, capable of learning and taking steps to counter the security threats. The use of deep learning to synthesize face recognition technology is realizing a synergetic outcome, providing an innovative and convenient solution for advanced surveillance applications in a wide range of scenarios. Incorporating such innovative technologies, this research develops intelligent surveillance systems and provides a secure protective environment.

Keywords: Intelligence surveillance system, convolutional neural network, Recurrent neural network, face recognition.

1 Introduction

Human monitoring aims to be collecting pertinent information when viewing individuals wandering and talking, for instance records from equipment, for example, CCTVs. This ability to observe individuals from afar has led to the proliferation of surveillance in recent decades: the modern surveillance system in place today.

© The Author(s) 2025
S. Kannadhasan et al. (eds.), Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24), Advances in Computer Science Research 120,
https://doi.org/10.2991/978-94-6463-718-2_65

Historically, these cameras record video that humans later or in real time need to watch to oversee far-flung places. That can take a long time and be very inefficient. There's an exciting idea: Using computer vision, the field that helps computers interpret images and video, cameras could themselves analyse the footage and report back only the information that matters: Who do they see? This would conserve the available resources and ensure prompt action in a potential security incident or similar situations. It's like you have these intelligent cameras that don't just record the scene, but also know what's going on there and communicate that back to you.

Here are other things to think about:

- There are privacy concerns with this approach that need to be responsibly addressed.
- To be of any real use, the technology must have accuracy and reliability.
- And you can probably think of other areas where this could be applicable, not only surveillance but also traffic monitoring, customer service, or healthcare.

2 Related Work

Significant progress has been made in the field of intelligent surveillance systems, with numerous researchers striving to create effective systems capable of recognizing various human activities through diverse methodologies.

2.1 Human Action Recognition (HAR) Techniques:

Ionut et al., presents a review of several methods in the area of HAR with a focus on the detection of human activity via computer vision. As we can see, the challenges led researchers to develop architectures for real time action recognition system in challenging environment using the internet of things. After obtaining the motion data, the proposed method uses the Histogram of Motion Gradient (HMG) descriptor to extract motion data, and adopts the Vector of Locally Aggregated Descriptors (VLAD) for encoding. It also cites Long Short-Term Memory (LSTM) networks to build upon certain difficulties of detecting joint data due to noise, occlusion etc., demonstrating how far techniques have come in HAR.

2.2 Facial Recognition Integration:

Azher et al., discusses the integration of facial recognition technologies into surveillance systems, which enhances the ability to monitor and identify individuals in various environments. This integration is crucial for developing a more targeted surveillance approach, allowing for the identification of persons of interest.

2.3 Dataset Development:

Chen et al., explores instead of relying on publicly available datasets, the authors created their own dataset to ensure high accuracy in face detection under real-world conditions. This dataset includes images taken in different lighting, angles, and positions, which is essential for training deep learning models effectively.

2.4 Applications beyond Surveillance:

Luvizon et al., points out that the applications of intelligent surveillance systems extend beyond security. Potential areas include traffic monitoring, customer service, and healthcare, indicating the versatility of the technology.

2.5 Comprehensive System Development:

Liu et al., highlights that their contribution leads to a comprehensive system capable of individual identification and classification in surveillance footage. This system not only supports attendance monitoring but also has functionalities like live tracking and sentiment analysis, although these were not the primary focus of the study.

In summary, the literature survey within this paper illustrates the advancements in intelligent surveillance systems, particularly through the integration of deep learning and facial recognition technologies, while also addressing the challenges and applications of these systems in various fields.

3 Dataset

"Data" is the biggest obstacles that any deep learning project encounters, particularly when working with facial recognition models. To achieve state-of-the-art accuracy, deep learning-based recognition models require hundreds of hours and vast volumes of training data. Moreover, the diversity of training data has equal significance to dataset size. A dataset containing 1500 photographs of 15 distinct human faces, or 100 images of each face, may be superior to a dataset containing 1500 images of 1500 distinct human faces, or one image of each face.

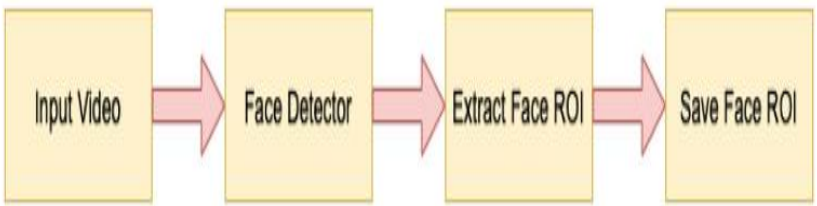

Fig. 1. Conversion of Videos to Photos.

Instead of using a widely available internet dataset, we generated the own. This choice allowed us to photograph a face in different lighting, from different perspectives, and in different positions. The task was to detect faces in real-world circumstances with the highest level of accuracy possible, given the inherent diversity in the facial dataset. While taking hundreds of pictures of a single face, it takes time to crop the image and forward this data for model training. However, by merely converting movies to photos, the aforementioned procedure was made more efficient (Fig. 1).

The delineated methodology for the extraction of facial data from a self-produced video is depicted in the diagram presented above. The input video undergoes a frame-by-frame analysis, with each frame being processed through the pre-trained facial detection model. Upon the identification of a human face within the current frame, achieving an accuracy rate of 95% or greater, the Region Of Interest (ROI) from that frame is extracted and subsequently stored in the designated output folder. A notable benefit of this technique is its efficacy in the rapid generation of a dataset comprising hundreds of images within a matter of seconds. Within a span of several hours, we successfully amassed 7500 live images representing more than 26 individuals, with an approximate distribution of 288 images per subject, thereby illustrating a variety of lighting conditions and poses.

3.1 Weizmann Action Dataset

Regarding the Weizmann Action Dataset, provided by the Weizmann Institute of Science, it encompasses two datasets: Weizmann Event-Based and Actions as Space-Time Shapes Analysis. The Weizmann Event-Based dataset includes four action classes—running, waving, jogging in place, and walking, performed by individuals in various clothing, as depicted in Fig. 2. The dataset consists of 6000 frames of video sequences, showcasing different actions.

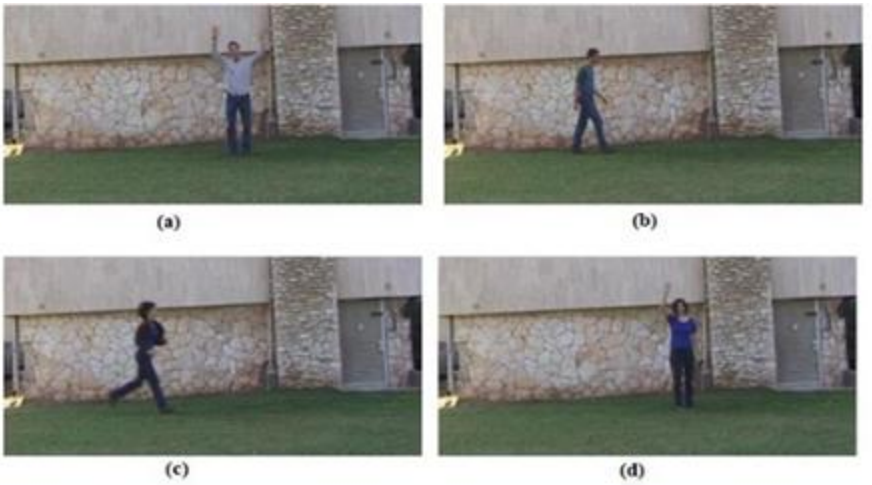

Fig. 2. Frames from Weizmann dataset.

4 Methodology

In the year 2017, OpenCV version 3.3 unveiled a markedly improved Deep Neural Network (DNN) module that demonstrated compatibility with an array of deep learning frameworks, encompassing Caffe, TensorFlow, and PyTorch. This module, notably engineered by Aleksandr Rybnikov, facilitated the seamless importation and deployment across diverse frameworks. The official iteration of OpenCV incorporated a cutting-edge face detection algorithm predicated on the Single Shot Detector (SSD) architecture, utilizing a ResNet backbone, which is available within the face finder sub-directory of the DNN sample repository. The DNN face detector, armed with pre-trained weights, exhibited efficacy in identifying faces across various mediums, including static images, video files, live video feeds, and real-time webcam streams.

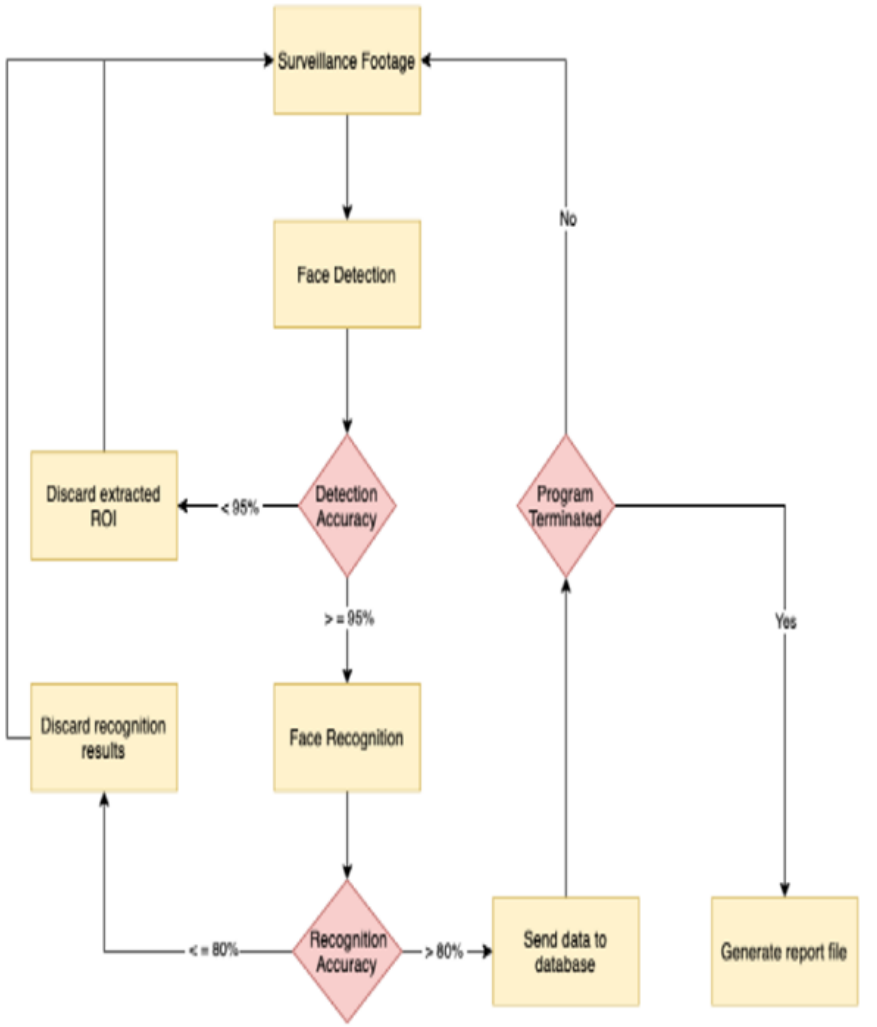

Fig. 3. Face Recognition System.

Its compact architecture exhibits a notable enhancement in operational efficiency while the CPU memory utilization is minimal and well fit for real-time detection applications. Input image pre-processing involved using the DNN library to set image sizes and perform normalization. We computed confidence scores after each pass through the neural network, eliminating poor detections where the confidence score was less than a set threshold. The corresponding Region of Interest was processed using face recognition measures for identification and classification.

It is a great challenge in the field of online facial recognition approaches to find a best solution. Our process started with the deployment of a local neural network specific to deep learning for the extraction of facial features and the calculation of embeddings. These embeddings were then used to train the Support Vector Machine (SVM) classifier. The performance was acceptable with a small dataset, but upon scaling it up outside of them became complicated. To train and optimize our local DNN we performed a set of experiments including the addition of extra convolutional layers, max pooling, dropout layers, multiple optimizers and the adjustment of the input image size of the DNN. Despite the introduction of techniques like face alignment, the use of OpenCV's deep learning facilities, and the conversion of images to RGB format grayscale and YCbCr, the local DNN struggled to meet the real-time recognition requirements. See Appendix A for failure cases such as the errors on persons with glasses, and the mistake of classifying the same person with different illumination as two distinct persons.

To address these limitations and maintain consistent application of our face recognition system, did explore alternatives by taking advantage of the accuracy of state-of-the-art neural nets using transfer learning. Top layers were held constant with activation and dropout layers for retraining of several neural models including the following: VGG16, VGG19, VGGFace, Inception v3, and ResNet-50. Because it consumes less memory and was more accurate the VGGFace model became the favorable option.

VGGFace model is developed by Visual Geometry Group (VGG) from university of Oxford with two key Models VGGFace and VGGFace2. Using transfer learning, we re-trained fully connected layers after freezing pre-trained weights. The modified VGGFace system showed better precision and adopted a single dense layer (of 1024 channels), a dropout layer to avoid overfitting, plus a dense softmax layer to classification. The approach reduced the risk of overfitting training large neural networks with limited data, resulting in an improved face recognition system. The retraining on Google Colab further validated its efficacy as a solution to overcome issues stemming from models with deeper layers. The VGGFace with improved efficiency had made it a good candidate for our studies.

5 Proposed System

A preprocessing method that combines multiple approaches is applied. First input channel is selected, then top-hat filter is applied and final threshold adjustments are made to establish intensity calibrations and contrast to ensure better image quality.Subsequent to the preprocessing phase, a weight-based segmentation strategy

is utilized for detection, wherein frame differences are computed utilizing cumulative mean, and the background is updated in accordance with the assigned weights. This methodology additionally aids in the discernment of foreground regions. Following this, a hybrid feature extraction technique is employed for the recognition of human actions. The special features that have been extracted are subjected to fusion via a serial-based methodology, and the resultant fused specific feature is utilized for the ensuing classification. Particularly, to substantiate the efficacy of the proposed system, four distinct datasets are employed, including the HOLLYWOOD dataset.

5.1 Image Processing Steps:

1. **Image Acquisition:**
   - Image acquisition involves scanning human activity photos and converting them to digital format.
   - Scanned photos are trimmed to focus on the appropriate object region.
2. **Image Binarization:**
   - Image binarization is a key area of research in image processing and serves as an essential pre-processing technique for image recognition and edge/boundary detection.
   - Selecting the appropriate threshold for each image can be challenging across different application domains.
3. **Image Improvement:**
   - Image improvement involves adjusting digital images to improve their suitability for display or further analysis.
   - This may include processes such as noise removal, sharpening, or brightening the image to facilitate the identification of key features.
4. **Feature Extraction:**
   - The method collects key characteristics from photos to generate a feature vector.
5. **Training:**
   - The resulting feature vector is used to train the system.
6. **Classification:**
   - The classification process results in a label for the human activity, with a portion of the image highlighted accordingly.

5.2 General Constraints:

- One major limitation is the necessity to train the system on both typical and suspicious activities.
- The system may classify activities incorrectly if the images are too noisy, leading to potential feature loss.

1. User:

A user manual is written and distributed with the program to help the user in operating the system.

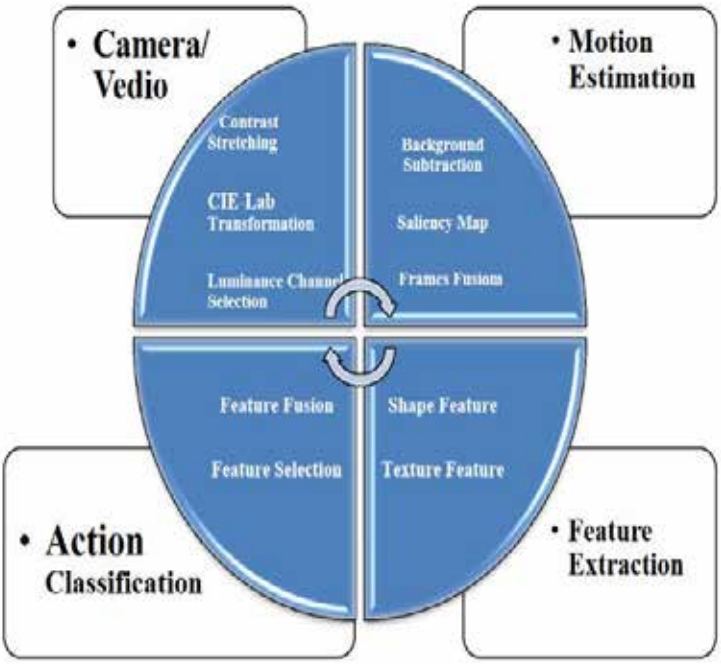

Fig. 4. Image Acquisition and Classification.

6 Results

Initially, facial recognition technology was supposed to recognize 26 people. The VGGFace model was trained using the original dataset which included around 288 photos per individual. The model was evaluated in two distinct stages: first, using random images of the target individual, and then with real-time video supplied into the recognition model via a live webcam. In each of the previous examples, the model accurately recognized each individual with a confidence level ranging from 78.54% to 100%, with a mean average confidence level of 96%. The model was able to recognize and discriminate between single and several people in a single picture.

Furthermore, it was revealed that when a new face was presented to the identification model without first being trained, the recognition model's confidence level fell below 50–55%. As a consequence, labelled every face discovered with a recognition accuracy of less than 55%, as uncertain recognition. Furthermore, if the identification accuracy fell between 55 and 80 percent, it was considered poor recognition, and such items were removed from the database.

6.1 System Validation and Performance Metrics

Image processing software such as Photoshop is used to transform non-BMP or jpg pictures to BMP or jpg images so that the system can read them. Scanned photos are additionally resized and their resolutions are adjusted.

Table 1. Comparison of different methods for metric evaluation.

Method	Feature Type	Accuracy	FNR (%)	Time Consumed (sec)
MSVM	Original FC	68.5	31.9	121.6
MSVM	Original Pooling	59.3	40.7	198.6
MSVM	Fusion	94.4	5.1	251.5
MSVM	Reduction	97.1	2.9	104.3
Ensemble	Original FC	90.6	9.4	127.9
Ensemble	Original Pooling	79.5	20.5	182.4
Ensemble	Fusion	94.6	5.4	294.9
Ensemble	Reduction	97.2	2.8	107.5
Naive Bayes	Original FC	92.5	7.5	111.5
Naive Bayes	Original Pooling	89.0	11.0	188.9
Naive Bayes	Fusion	96.3	3.7	240.5
Naive Bayes	Reduction	98.0	2.0	92.48

The system is evaluated using different methods and feature types, and the results are summarized in Table 1. The result demonstrates the accuracy, false negative rates, and time consumed for each method and feature type combination, providing insights into the system's performance under different configurations.

7 Conclusion

A method suggested for human identification and single-person activity recognition in this paper, with a focus on efficient preprocessing techniques to improve system performance. Using a weighted-based segmentation technique for edge detection, a top hat filter for intensity adjustment, and hybrid feature extraction techniques for precise human action identification were all part of the methodology. For categorization, serial-based fusion was used. Significant accuracy gains (94.1%,

96.9%, and 98.1%) were shown during validation on four datasets (HOLLYWOOD, UCF101, HMDB51, and WEIZMANN).

This contribution includes a thorough system for classifying and identifying individuals in surveillance film, producing complete reports that are helpful for tracking attendance. Although the system has other features including sentiment analysis, age-gender predictions, and live tracking, these were not included in the study's purview.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Singh, A., Bhatt, S., Nayak, V., & Shah, M. (2023). Automation of surveillance systems using deep learning and facial recognition. 14(S1), 236–245. https://doi.org/10.1007/s13198-022-01844-6.
2. Pulugu, D., Anusha, P., Srivastava, R. R., Kalaivani, R., & Pal, S. (2023). Machine learning-based facial recognition for video surveillance systems. ICTACT Journal on Image and Video Processing, 14(2), 3149–3154. https://doi.org/10.21917/ijivp.2023.0448
3. Khean, V., Kim, C., Ryu, S., Khan, A., Hong, M. K., Kim, E. Y., Kim, J., & Nam, Y. (n.d.). Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models. https://doi.org/10.32604/cmc.2024.056767
4. Tahri, M., & Arfaoui, N. (2024). E-Learning Facial Emotion Recognition Using Deep Learning Models. 215–224. https://doi.org/10.1007/978-3-031-64836-6_22
5. Mrak, M. (2023). Face recognition methods in video surveillance systems using machine learning. Ìnfokomunìkacìjnì Tehnologìï Ta Elektronna Ìnženerìâ, 3(2), 33–42. https://doi.org/10.23939/ictee2023.02.033
6. Kishor, Avhad., Vishal, Jadhav., Sahil, Kurhade., Prof., Manjiri, Raut. (2024). 1. Real-Time Surveillance with AI: A Comprehensive Approach to Security and Monitoring. International Journal of Advanced Research in Science, Communication and Technology, doi: 10.48175/ijarsct-19681
7. E, Irfan., Chinnu, Jacob., R., Resmi. (2024). 2. Facial Recognition and CCTV Integration for Enhanced Security Using Deep Learning Techniques. doi: 10.1109/raics61201.2024.10689986
8. [8] Muhammad, Mobeen, Abid., Toqeer, Mahmood., Rana, Umair, Ashraf., C., M., Nadeem, Faisal., Haseeb, Ahmad., Asim, Niaz. (2024). 4. Computationally intelligent real-time security surveillance system in the education sector using deep learning. PLOS ONE, doi: 10.1371/journal.pone.0301908
9. Amina, Catherine, Ijiga., Toyosi, Motilola, Olola., Lawrence, Anebi, Enyejo., Francis, Adejor, Akpa., Timilehin, Isaiah, Olatunde., Funminiyi, Olajide. (2024). 5. Advanced surveillance and detection systems using deep learning to combat human trafficking. Magna Scientia Advanced Research and Reviews, doi: 10.30574/msarr.2024.11.1.0091
10. S., C., Amingad., Morathoti, Rahul., Sujata, Gupta., P., S., S., Reddy. (2023). 7. Enhancing Public Safety with AI-Powered Intelligent Surveillance: An Examination of Immediate Incident Detection and Rapid Response in Urban Settings. doi: 10.1109/icaiccit60255.2023.10465690

11. Yao, Yao. (2024). 6. Research on facial recognition system based on deep learning. Applied and Computational Engineering, doi: 10.54254/2755-2721/34/20230332
12. Thiyyagura, Jagadeesh, Reddy, Marrey, Sai, Ganesh, Mandadi, Hemanth, Kumar, Reddy, Chintharapu, Bhandhavya., R., Jansi. (2024). 8. Deep Learning-Powered Face Detection and Recognition for Challenging Environments. doi: 10.1109/idciot59759.2024.10467753
13. Arjun, E, L., Sharan, Kumar, Meghana, S, D., Suresh, Maharana., Shriram, S., Tejonidhi, S., Narayana, Darapaneni., Anwesh, Reddy, Paduri. (2023). 9. Advanced Face Authentication Using Deep Learning Models. doi: 10.1109/punecon58714.2023.10450013

Appendix A: Failure Case Examples and Observations in Facial Recognition

This appendix provides an overview of failure cases encountered during the facial recognition experiments using both local DNNs and transfer learning approaches. These insights guided the transition toward more accurate models like VGGFace.

Failure Cases in Local DNN Models

1. **Eyeglasses and Facial Occlusion**
    - Description: Individuals wearing eyeglasses often resulted in partial occlusion of key facial landmarks.
    - Effect: The model frequently misclassified the same person with and without glasses as different individuals.
    - Resolution Attempted: Introduced face alignment and grayscale normalization, but issues persisted under varied lighting.
2. **Illumination Variation**
    - Description: Faces captured in different lighting conditions (e.g., shadows, bright light) produced inconsistent embeddings.
    - Effect: Same individual was recognized as two different persons under varying lighting.
    - Resolution Attempted: Applied histogram equalization and YCbCr color space transformation. Improved, but not solved.
3. **Incorrect ROI Detection**
    - Description: In videos with movement, the face detector failed to lock onto the correct region of interest (ROI) consistently.
    - Effect: Detection bounding boxes included background noise or incomplete faces.
    - Resolution Attempted: Adjusted confidence threshold to >95% to minimize false detections.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.