Int J Syst Assur Eng Manag (March 2023) 14(Suppl. 1):S236–S245
https://doi.org/10.1007/s13198-022-01844-6

ORIGINAL ARTICLE

Automation of surveillance systems using deep learning and facial recognition

Arpit Singh[1] · Saumya Bhatt[1] · Vishal Nayak[1] · Manan Shah[2]

Received: 6 December 2021 / Revised: 15 August 2022 / Accepted: 25 November 2022 / Published online: 6 January 2023

Abstract While it's true that humans are skilled at recognizing faces, what about computers? Machinery is transforming conventional human labor in today's technologically sophisticated world. These cutting-edge machines may be programmed to recognize faces in inconceivable ways, a process known as facial recognition or computer vision. In this study, we present a real-time system for detecting and identifying individuals in live or recorded surveillance feeds using deep learning and face recognition algorithms such as Convolution Neural Networks (CNN). The proposed real-time database integrated system is based on the VGGFace deep learning neural architecture. Transfer learning is used to retrain the model on a smaller original tailored dataset of 7500 images of 26 distinct individuals. Details regarding the creation of original tailored datasets and brief tests with various machine learning and deep learning approaches to analyze and improve the recognition accuracy of the proposed system are also reported. The proposed approach shows the highest degree of recognition accuracy by accurately recognizing each of the 26 individuals with a confidence level ranging from 78.54 percent to 100 percent, a mean average of 96 percent on real-time inputs.

Keywords Deep learning · Transfer learning · VGGFace · Facial recognition · Original dataset

✉ Manan Shah
manan.shah@spt.pdpu.ac.in

[1] Prutha Technomarket Private Limited, Ahmedabad, Gujarat, India

[2] Department of Chemical Engineering, School of Energy Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

1 Introduction

The objective of human surveillance is to gather relevant information by studying people's movement and behavior using electronic devices, such as CCTVs, and due to this feature of observing individuals from afar, the world—in recent decades—has seen a substantial increase in the number of surveillance systems present around us. Take the United Kingdom as an example. In England and Wales, the count of security cameras grew from 100 in 1990 to over 4.2 million in 2007, implying that there are over 7 cameras for every 100 individuals (Xu 2021; Farrington et al. 2007). With such a large number of surveillance cameras installed, human monitoring of these cameras' outputs would be extremely difficult (Collings et al. 2015).

So, the question here is how can we transform this relevant information collected by these electrical instruments, into a different use case? Traditionally, through surveillance cameras, input frames are captured and then sent to a memory processing unit where these frames are saved and are later or in real-time, used by humans to monitor places that are physically far from them. But, what if these cameras instead of returning the input frames can return the details of the people detected in those frames. This is one of the core concepts of computer vision, which is an interdisciplinary scientific field that seeks to attain a high-level understanding of information extracted from images or videos.

Computer vision is a field of artificial intelligence that educates computers to perceive and grasp the visual environment, resulting in a feeling of knowledge of a machine's surroundings. It may be used for facial recognition, number plate recognition, augmented and mixed realities, position determination, and item identification. When a computer visual system identifies the face of a previously acquainted person, it is called facial recognition; which has become one

of the most critical instruments of human–computer interaction (HCI). This discipline of computer vision has been gaining attention since the 1960s owing to its inconspicuous nature and primary method of identifying a person in the frame. One of the most intriguing benefits of face recognition is that it can fluently detect and recognize people without their awareness in an uncontrolled environment.

Albeit the accuracy of face recognition systems has always been deba, the recent pivotal changes have brought a sense of complacency among its users. With the increasing computational capabilities, various pioneers in the field of facial recognition, have been able to bring revolutionary changes, and these alterations have been addressed in many pieces of research (Liu et al. 2017); (Kortli et al. 2020).

Previously, face recognition methodologies were based upon superficial learnings which were indeed subjected to various real-time obstacles such as posture & angle variation, facial concealments, and varying light conditions, etc. These superficial learning methods were able to extract the basic features of the image, but it was only after the development of Deep Learning, the term first coined in the 1980s but only after 2012, that methods based on it allowed the unsheathing of more crucial and complex facial features from an image as referenced in (Wang et al. 2020).

Unlike the traditional models and other machine learning algorithms, Deep learning-based Convolution Neural Networks (CNNs) aren't linear, instead, they are a pile-up of multiple layers having increased complexity and abstraction as we move through them. Yann LeCun, a French-American computer scientist, proposed CNNs for handwritten ZIP code recognition in 1989. Following LeCun, scientists and researchers carried through the work he started, and it was only around 2012 when a breakthrough came. A CNN called AlexNet achieved state-of-the-art performance classifying pictures in the ImageNet challenge.

Since then, deep learning has come a long way. It has risen rapidly in recent years as a promising subject of artificial intelligence, offering the advantages of great adaptability and powerful training abilities over existing machine learning (ML) methods such as support vector machines (SVM) and k-nearest neighbors (kNN). It is now widely used in pattern classification applications, computer vision, image analysis, and audio-visual identification. (Chai and Li 2019).

Research is being done on the formation of mathematical methods to allow computers to reconstruct and understand 3D images. Acquire 3D visuals of objects, object search, face recognition, active shapes of unique faces and 3D shape models, personal photo collection, instance recognition, image retrieval, an inspection of editing context, and understanding of scenes. These are just basic applications, each of the above categories can be investigated in more detail. Albeit, deep learning was created in the computer science field, its usage has proliferated in a wide range of industries, including healthcare, neurology, physics and astronomy, financial services, and business management. (Chai et al. 2013; Chai and Ngai 2020).

In 2020, due to the excruciating impact of Covid-19, biometric-based punching systems—used for attendance purposes—were losing their charm in the market, since their physicality did not support Covid-19 norms of social distance, and touchless operations. As a result, facial recognition technology became even more prevalent in the IT industry. Indeed, this was the impetus for us to create a real-time surveillance system that could also be used as an automated attendance system. In this paper, we discuss multiple applications of face recognition technology, with our primary use case centered on detecting and identifying known faces in motion using footage from surveillance cameras installed at a premises' entrance. The paper also addresses how deep learning technology has been utilized for face expression analysis, gender recognition, and calculating the age of humans in the past.

The following is how the paper is organized: Sect. 2 discusses relevant work in the field of face recognition, and Sect. 3 quickly describes the methodology, and the dataset used. Section 4: discusses results and discussion. Future scopes and challenges are mentioned in Sect. 5, whereas the conclusion is stated in Sect. 6.

2 Related work

(Ilyas et al. 2019) outlined a facial recognition system based on deep learning neural networks in their suggested study. The Viola-Jones face detection method was used to extract the face from the input image, followed by the Histogram Equalization Algorithm (AHE), which was used to change and improve the grey level of an image to a homogenous distribution. Furthermore, after preprocessing the input region of interest, they utilized VGG16, and ResNet50-based convolutional neural networks to extract facial features and classify human faces.

They tested their neural networks VGG16, and ResNet50 on two datasets: First, the Extended Yale B Face database honing to an accuracy of 96.12%, and 97.23% respectively. Second, the CMU PIE database obtained an overall accuracy of 96.55%, and 98.38% respectively. Moreover, the performance of their suggested technique was compared to other state-of-the-art methods on Yale and CMU databases, and it was evident that their method outperformed all others.

Deep learning-based neural networks are the driving force behind today's facial recognition systems; yet, training a state-of-the-art convolution neural network necessitates massive computational resources, since the model must be trained on millions of images over hundreds of hours. To overcome this restriction, the concept of transfer learning

has recently gained traction in the face recognition community. In machine learning, transfer learning is an approach in which the information obtained from one work is used to improve the learning process in another work. In their study, and (Perdana and Prahara 2019) show us how this new technique has enabled academics and professionals to disintegrate the dependency between neural networks and computational complexity.

Both of the research employs the state-of-the-art VGG16 architecture to train the input face pictures, which are pre-trained on the massive ImageNet database with over 1 million images belonging to 1000 distinct categories. Operating on different datasets where [1] chooses Yale and AT&T face databases with an input image size of 224 × 224x3, [2] chooses resized 120 × 120 images from ROSE-Youtu Face Liveness Detection Database, to re-train their respective locally altered VGG16 model constituting of the modified input layer, multiple hidden layers, and output layers. As evidenced by the above works Light convolutional neural networks re-trained for a significantly shorter timeframe on smaller datasets with transfer learning generated an accuracy of over 95% in constrained/unconstrained settings.

Furthermore, based on a discriminative model with deep feature training, (Shakeel and Lam 2019) developed an age-invariant face recognition technique. In their research, AlexNet was used as a transfer learning CNN model to learn high-level deep features. The characteristics were then encoded into a unique code word for image representation in a codebook. The encoding technique ensured that photographs of the same individual taken at various periods had the same unique code words. A linear regression-based classifier was used for face recognition, and the method was tested on three datasets, including the publicly available FGNET.

(Peng et al. 2019) in their research demonstrated how to employ a deep local descriptor learning framework for cross-modality face recognition, in which both compact local information and discriminant features are learned directly from raw facial patches. CNN is used to obtain deep local descriptors, and the method was tested on six extensively used face recognition datasets of different modes assigning to an overall accuracy of 98.68%.

Additionally, several pundits have advocated for the use of face recognition technology to automate a variety of mundane tasks, one of which is the automation of an attendance system. In this study (Harikrishnan et al. 2019), they described how Artificial Neural Networks can be taught across billions of pictures and used to detect and recognize faces with relative ease and flexibility in an instance. Their research was broken down into four main sections: detection, training, recognition, and data administration. The first stage of image preprocessing involved resizing the initial 8 bits grayscale image, enhancement of the facial features, and noise reduction in the input imagery.

OpenCV's Haar Cascade Method was used to detect the facial region from the input images, and the LBPH recognizer was trained on the dataset generated. Furthermore, live footage was sent through the LBPH recognizer during facial recognition, which generated LBPs and histograms for each recognized face. Each histogram created was compared to the ones already in the training file, and the best match was marked as present in the excel report which contained the attendance record of the students for a particular day. Achieving an accuracy of over 70% they built a prototype that was compact yet robust, and easily customizable to be utilized with a pocket-sized computer.

By now we know that deep learning is a broad domain, and neural networks based on deep learning, such as facial recognition models, have a wide range of applications. However, gender and age estimation methodologies, due to their inherent limits, have long been a source of contention. Yet, experts in deep learning, on the other hand, have devised novel techniques for improving the accuracy of these estimating models. (Smith and Chen 2019) demonstrated that their gender identification model has a state-of-the-art accuracy of over 98 percent, with an MAE of 4.1 years when predicting age.

During their research, different neural models such as VGG16, VGGFace, and VGG19 were evaluated on the Morph-II dataset, which has 55,134 pictures with subjects ranging in age from 16 to 77 years old. Transfer learning was used to retain the pre-trained weights and layers were added/dropped from the existing neural architectures. Aside from structural modifications, several other training approaches were tested to see how they affected gender identification and age estimation. Finally, VGGFace significantly outperformed other gender recognition and age estimation techniques and even surpassed human performance in real-time.

Moreover, similar to gender and age estimation models, sentiment analysis based on human facial expressions has also been a prominent area of research. In their work, (Kukla and Nowak 2015) offer a comprehensive examination of a cascade of neural networks to detect six basic human emotions and neutral expressions, including happiness, sadness, fear, anger, disgust, and surprise. They utilized single or multilayer configurations of neural networks to detect each unique expression in their multistage investigation, and classifications were generated by integrating the data from these distinct neural networks at later stages.

The Karolinska Directed Emotional Faces (KDEF) database, the John Kanade database, and a set of images obtained by the authors provided the input to the cascade of classifiers. The results obtained were different for particular emotions. Fear was the least well-detected emotion,

while happiness and surprise were the most accurate, with an overall accuracy of 76% (Table 1).

3 Dataset

One of the major setbacks that every deep learning-based project faces is 'data', especially when dealing with facial recognition models. Deep learning-based recognition models need massive amounts of training data and hundreds of hours to acquire state-of-the-art accuracy. Furthermore, diversity in training data is as important as the size of the dataset. For instance, a dataset having 1000 images of 10 different human faces (100 images of each face) can be comparatively better than a dataset having a total of 1000 images of 1000 different human faces (1 image of each face).

We created our dataset rather than using one that was readily available online. We were able to capture a face in a variety of light conditions, from various angles, and with various poses as a result of this decision. Due to the inherent variability in the facial dataset, we were pushed to achieve top-notch accuracy when identifying faces in real-world scenarios. Although capturing hundreds of images of a single face, cropping the face, and pushing this data forward for model training is a time-consuming task. On the other hand, the above process was made more efficient by simply converting videos to images (Fig. 1).

The methodology for extracting facial data from a self-made video is depicted in the diagram above. Frame by frame, the input video is fed to the facial detection model. The pre-trained face detection model detects a human face in the current input frame, and if the detection accuracy is greater than or equal to 95%, the region of interest (ROI) from that frame is extracted and saved in the output folder.

One of the main benefits of the procedure described above is that it saves time by generating a dataset with hundreds of images in a matter of seconds. In just a few hours, we were able to create 7500 live images of over 26 people, approx. 288 images per subject, in various light conditions and poses.

The above-shown face ROIs of 2 different people are generated by converting videos into images (Fig. 2).

4 Methodology & discussion

The face recognition system includes three major steps: Face Detection, Face Recognition, and Report Generation (Fig. 3).

5 Face detection

In 2017, OpenCV 3.3 was released, which included a much improved DNN module that works with a variety of deep learning frameworks, including Caffe, TensorFlow, and Torch/PyTorch. Most crucially, regardless of where it was generated, the module could be imported and utilized across any framework. The key contributor to the DNN module, Aleksandr Rybnikov, also included a state-of-art face detector based on the Single Shot Detector (SSD) framework with a ResNet base network in the official release of OpenCV which can be found in the face_detector sub-directory of the DNN samples.

With the pre-trained weights, faces could be detected in image stills, videos, video streams, and live webcams using the DNN face detector. Moreover, due to its lightweight architecture, it performs admirably in terms of operational speed, and CPU memory consumption stays comparatively low suitable for real-time detection. The input image passed to the face detector is first pre-processed using the DNN library which includes setting the image dimensions and normalization. The confidence is retrieved after each pass of the input image through the neural network, and weak detection is ruled out if the current confidence falls below the set threshold. The collected ROI is then sent through the face recognition pipeline for identification, and classification.

6 Face recognition

On the internet, there are a plethora of face recognition techniques to choose from. Choosing the right technique for the job, on the other hand, is the most difficult part of the project. We began with a deep learning-based local neural network for extracting facial features and computing embeddings. After that, the embeddings were used as a training input for the SVM classifier. The trained model was stored locally and then deployed for recognition later. Although the performance was satisfactory when tested on two individuals, the accuracy dropped as the input dataset grew larger.

In addition, we experimented with the local DNN by adding additional convolutional layers, including max pooling and dropout layers, using various optimizers, and altering the DNN's input image size. We even used a face aligner, and OpenCV's deep learning libraries for image optimization and pre-processing, and converted the input images from RGB to grayscale, and YCbCr. Unfortunately, the local DNN was not up to the task of real-time recognition, and the model would mistakenly identify two separate individuals wearing specs as the same person, and a single individual under various lighting circumstances as a different person.

To overcome the limits of our local DNN and to enhance the consistency of our face recognition system, we looked

Table 1 Relevant studies in the subject of deep learning for face detection and face recognition

S.No	Domain	Method	Results/Conclusion	References
1	Face detection and recognition	Deep Pyramid Single Shot Detector is proposed for face detection and DCNNs based pipeline for face recognition	Multiple datasets both for face detection, and face recognition are used and state-of-art accuracies are projected across all databases	(Ranjan et al. 2019)
2	Face detection and recognition	VGG16 model is used for training the purpose, PCA is used for feature dimensionality reduction, and SVM is used for face recognition	On the LFW dataset accuracy achieved by the proposed model is 97.47 percent which is higher than the DeepFace neural model	(Chen and Haoyu 2019)
3	Comparative study on Convolution Neural Networks for face detection and recognition	CNN-based models such as AlexNet, VGG16, VGG19, and MobileNet are used with the application of Transfer Learning and Fine Tuning	The original dataset is used with VGG19 giving the best validation accuracy of 87%, and MobileNet giving the best testing accuracy of 84%	(Ahmed et al. 2020)
4	Choosing the best deep learning techniques for facial recognition	Various models including multiple versions of AlexNet, VGG16, GoogLeNet, Inception, ResNet, DenseNet, and LightCNN were proposed for face recognition under different frameworks	MS-Celeb-1 M, UMDFaces, VGGFace2, and WebFace datasets were used for training, and LFW, VGGFace2-test, IJB-A, DFW, and UMDFaces-test were used for recognition purposes	(Wang and Guo 2019)
5	Popular Deep Net Architectures for facial recognition	To enhance facial biometric security, a survey was done on pre-existing popular deep neural networks	The LFW dataset was utilized, and VGG-16 & VGG-19 showed the highest levels of image recognition accuracy, as well as F1-Score	(Gwyn et al. 2021)
6	Emotion Recognition	Various sentiment analysis techniques are explored, and each method is well detailed in the article	The study examines the benefits and disadvantages of each technique	(Varghese et al. 2015)
7	Age Recognition	Deep Convolutional Neural Networks with transfer learning are used for age estimation	With state-of-the-art accuracy, their single and multimodal system can classify age into seven distinct groups	(Dong et al. 2016)
8	Gender Recognition	A Local DNN trained with small patches extracted from images is used for the gender classification	The accuracy of their model in the LFW dataset was 96.23 percent, while it was 90.23 percent in Gallagher's dataset, outperforming competing deep learning approaches	(Mansanet et al. 2016)

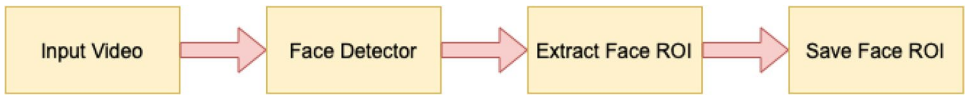

Fig. 1 Process diagram for converting videos to images

Fig. 2 ROI of two different people

at other possibilities where we could utilize the accuracy of existing state-of-art neural networks for our use case. This was made feasible by the introduction of transfer learning, which allowed state-of-the-art models to be re-trained without requiring massive computational resources.

With our original dataset, we evaluated numerous neural models, including VGG16, a 16-layer neural network that won the Imagenet competition in 2014, VGG19, which has 19 layers, VGGFace, which is based on VGG16, Inception v3 with 48 layers, and ResNet-50, which has 50 deep layers. Every model was re-trained by keeping the top layers and adding activation and dropout layers as appropriate. As a result of employing pre-trained weights and neural architecture, we were successful in improving the accuracy of our face recognition system.

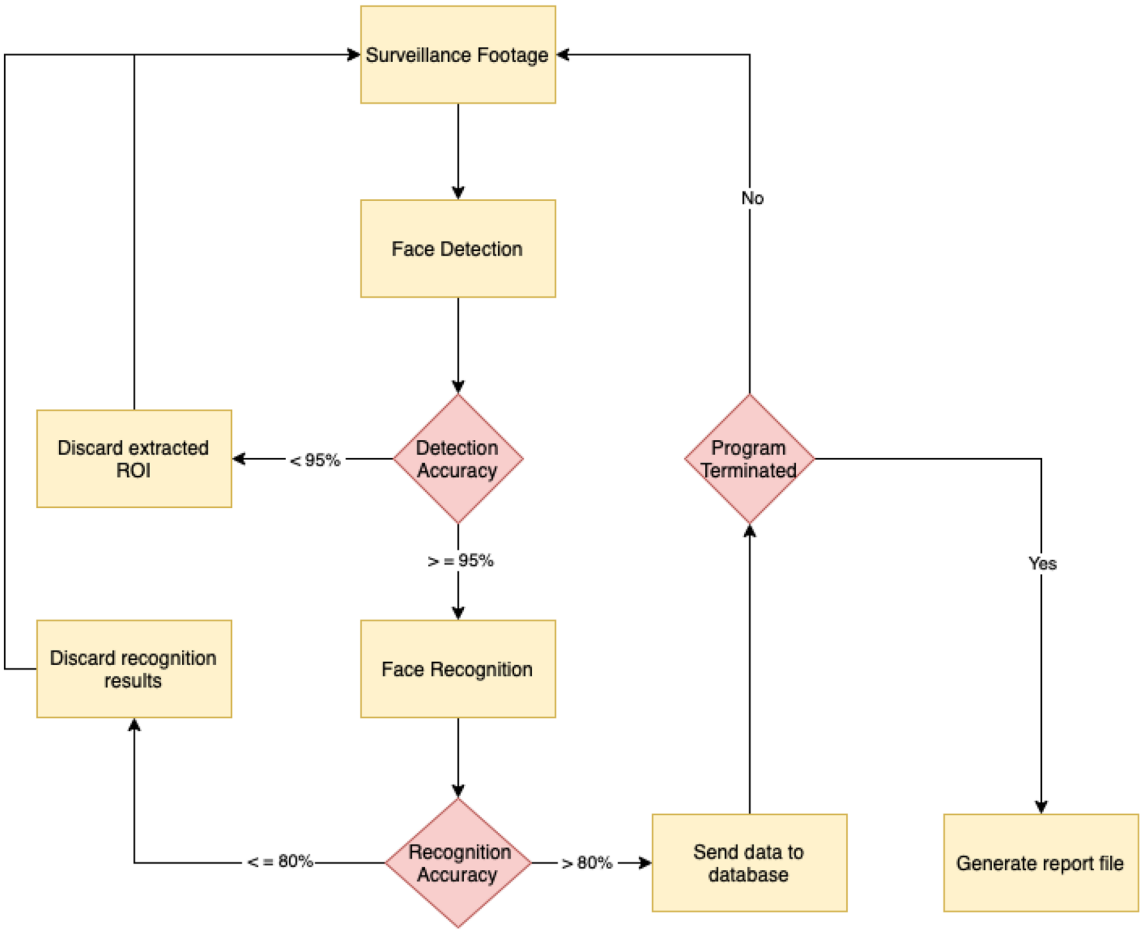

Fig. 3 Face Recognition Methodology

We re-trained the above-mentioned neural networks using Google Colab, which provided us with approximately 13 GigaBytes of RAM and a GPU including Nvidia K80s or T4s. Because of their huge size, models like Inception and ResNet, which have a relatively high number of deep layers, overshoot the RAM. As a result, we chose the VGGFace model as our leading contender, since it consumed less memory and gave better accuracy. The VGGFace represents a set of facial recognition models created by researchers of the Visual Geometry Group (VGG) at the University of Oxford and proven on benchmark computer vision datasets.

There are two principal VGG models for facial recognition: VGGFace and VGGFace2. In a nutshell, the default VGGFace model uses the same neural design as the VGG16 model (Simonyan and Zisserman 2015), which was applied in the ImageNet competition to identify 1000 classes of objects. Because the training sets were different, the primary difference between the VGG16 and VGGFace is the set of calibrated weights. The VGGFace dataset, which contains pictures of 2,622 unique identities, was used to train the VGGFace model.

VGGFace (Parkhi et al. 2015), a convolutional neural network model, is a linear sequence of convolutional, activation, pooling, and the softmax layers, trained for weeks using NVIDIA's Titan Black GPU. The convolution layer one input is a 224×224 fixed size RGB image. The first two layers have the same padding, and a filter size of 3×3 consists of 64 channels. Then, between a set of 2, stride 2×2 max pool layers, there are two 128-channel convolution layers with a 3×3 filter size.

Following that, there are two more convolution layers with a filter size of 3×3 and 256 channels. The last two sets of convolutional layers are of 512 channels, and 1-pixel padding is applied after each convolutional layer to avoid the spatial feature of the image from being lost. This results in a feature map of (7, 7, 512) after stacking 5 sets of convolution and max-pooling layers, which is further flattened into a feature vector.

Table 2 above depicts the neural architecture of our fine-tuned VGGFace model with an input image size of 150×150.

In the original model, the flattened output is passed through three Fully-Connected layers where the first two layers have 4096 channels each, and the third layer i.e. the softmax layer was used to classify 2622 unique identities. However, as seen in the above figure, we used only one dense layer having 1024 channels. Furthermore, overfitting and generalization mistakes can occur when massive neural networks are trained on limited training data. We used a dropout layer with a value of 0.5 to avoid overfitting,

Table 2 Neural architecture of fine-tuned VGGFace model

Model: VGGFace

Layer (Type)	Output Shape
input_1 (Input Layer)	(None, 150, 150, 3)
conv1_1 (Conv2D)	(None, 150, 150, 64)
conv1_2 (Conv2D)	(None, 150, 150, 64)
pool1 (MaxPooling2D)	(None, 75, 75, 64)
conv2_1 (Conv2D)	(None, 75, 75, 128)
conv2_2 (Conv2D)	(None, 75, 75, 128)
pool2 (MaxPooling2D)	(None, 37, 37, 128)
conv3_1 (Conv2D)	(None, 37, 37, 256)
conv3_2 (Conv2D)	(None, 37, 37, 256)
conv3_3 (Conv2D)	(None, 37, 37, 256)
pool3 (MaxPooling2D)	(None, 18, 18, 256)
conv4_1 (Conv2D)	(None, 18, 18, 512)
conv4_2 (Conv2D)	(None, 18, 18, 512)
conv4_3 (Conv2D)	(None, 18, 18, 512)
pool4 (MaxPooling2D)	(None, 9, 9, 512)
conv5_1 (Conv2D)	(None, 9, 9, 512)
conv5_2 (Conv2D)	(None, 9, 9, 512)
conv5_3 (Conv2D)	(None, 9, 9, 512)
pool5 (MaxPooling2D)	(None, 4, 4, 512)
flatten_1 (Flatten)	(None, 8192)
dense_1 (Dense)	(None, 1024)
dropout_1 (Dropout)	(None, 1024)
dense_2 (Dense)	(None, 26)

followed by a dense softmax layer to categorize 26 different people in our original dataset.

We only re-trained the fully connected layers, preserving the pre-trained weights of the top layers of our VGGFace model. The learning rate was set at 0.0001 with a batch size of 10 and total epochs of 30 using the Adam optimizer with categorical cross-entropy as our loss function. The training and test datasets were split into 4:1 ratios respectively with each image resized as 150×150. Apart from that, ReLU was used as the activation function for all hidden layers, as it is more computationally efficient since it speeds up learning and reduces the chance of a vanishing gradient problem.

7 Report generation

The facial recognition model received live input from the surveillance camera. The model was trained to recognize 26 distinct individuals, and whenever one was identified, an accuracy associated with the recognition was generated as well. Furthermore, we used XAMPP to set up a local Apache HTTP server with MySQL as the database management system. If an individual was identified with a precision of over 80%, the name label and a unique timestamp were sent to the database. Our algorithm was designed to run for a specific amount of time, hence, the last recorded entry of every unique individual identified by the system was pulled from the database and saved in an excel file on the local computer before the system was terminated.

8 Results

In the beginning, we proposed using our facial recognition technology to recognize 26 persons. The original dataset contained about 288 pictures per person and was used to train the VGGFace model. The model was tested in two ways: first, with random pictures of the target person, and second, using real-time video fed to the recognition model through a live camera. In each preceding example, the model correctly identified 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 detect and distinguish single to numerous persons together in a single frame.

Furthermore, it was discovered that when a new face was presented to the identification model without being trained, the recognition model's confidence level dropped below 50–55 percent. As a result, whenever a face was detected with a recognition accuracy of less than 55%, we labeled it as unknown recognition. Additionally, when the identification accuracy fell between 55 and 80 percent, it was considered a weak recognition, and such entries were excluded from the database.

9 Future scopes and challenges

The unavailability of high-end processing resources was one of our biggest challenges. Although utilizing Google Collab helped us speed up our training, we built our initial dataset with high-resolution photos, making it difficult to transfer them to the drive and employ them for training. We were able to train our model for up to 26 persons using the free resources. However, in future research, such an approach might be tweaked to minimize the size of the dataset while increasing the number of images in it. This would have a direct effect on the amount of computing power required to train the neural model.

Another challenge we faced was to incorporate a new individual into the trained model, as we trained the whole model every time we added new individuals to the training dataset. Accordingly, a technique might be investigated in future works in which additional people may be added to the trained model without having to retrain the entire model, and comprise the accuracy of the existing recognition model.

The facial features and appearance of humans change with age and time. As a consequence, for the model to remain consistent over time, it must be re-trained on new images of existing users. Therefore, to maintain the system, one must repeat the entire procedure over and over again, which is a difficult and time-consuming operation. Subsequently, methods may be investigated such that the proposed neural network can train itself using data acquired from everyday recognitions.

As part of our strategy for creating the original dataset, we gathered 5–6 short videos of the users in various light settings & postures and then used them to generate a collection of images. However, gathering data from users, such as videos including their faces, leads to the problem of privacy, as the user's information is constantly in danger of being abused or leaked. As a result, strong measures should be implemented in the future, such as enabling clients to self-train the neural model on their faces without having to share personal data with the developer team.

Finally, our facial recognition system is reliant on the live feed from the surveillance cameras. If the quality of the live input stream is low, it will directly impact the face recognition module's accuracy. In addition, too much change in the light circumstances might lead to poor detections as well. As a result, methods must be developed in the future to make the facial recognition system less reliant on light circumstances and footage quality.

10 Conclusion

In our study, we presented a system that can identify and classify individuals in surveillance footage and provide a

report based on the results of the identifications. A report such as this could be used for a variety of reasons, including monitoring attendance at educational and professional institutions. Furthermore, the proposed system could provide additional functions such as tracking individuals in the live footage, performing sentiment analysis, and making age and gender predictions on the faces detected; however, these functions are outside the scope of this study.

The most challenging element of our study was choosing the suitable face detection and recognition technology. From working with machine learning models to developing our local neural networks to eventually applying state-of-the-art neural models with transfer learning to our use case, it was an educational journey that tested our patience on every level. Creating and training local neural networks from scratch necessitates a large amount of training data and significant computing resources. On the other hand, combining transfer learning and fine-tuning to retrain neural models like VGGFace on a smaller dataset, proved to be a benefit for our research, as it yielded the highest overall accuracy.

Acknowledgements The authors are grateful to Prutha Technomarket Private Limited and Department of Chemical Engineering, School of Energy Technology, Pandit Deendayal Energy University for the permission to publish this research.

Author contribution All the authors make a substantial contribution to this manuscript. AS, SB, VN, and MS participated in drafting the manuscript. AS, SB, VN, and MS wrote the main manuscript. All the authors discussed the results and implications of the manuscript at all stages.

Funding Not Applicable.

Availability of data and material All relevant data and material are presented in the main paper.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Consent for publication Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

References

Ahmed T, Das P, Ali MF, & Mahmud MF (2020) A Comparative study on convolutional neural network based face recognition. In: 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, 1–5. doi:https://doi.org/10.1109/ICCCNT49239.2020.9225688

Chai J, Ngai EWT (2020) Decision-making techniques in supplier selection: recent accomplishments and what lies ahead. Expert Syst Appl 140:112903. https://doi.org/10.1016/j.eswa.2019.112903

Chai J, Liu JNK, Ngai EWT (2013) Application of decision-making techniques in supplier selection: a systematic review of literature. Expert Syst Appl 40(10):3872–3885. https://doi.org/10.1016/j.eswa.2012.12.040

Chai J, & Li A (2019) Deep Learning in Natural Language Processing: A State-of-the-Art Survey. In: Proceedings - International Conference on Machine Learning and Cybernetics, doi:https://doi.org/10.1109/ICMLC48188.2019.8949185

Chen H, Haoyu C (2019) Face recognition algorithm based on VGG network model and SVM. J Phys: Conf Ser. https://doi.org/10.1088/1742-6596/1229/1/012015

Collings DG, Scullion H, Vaiman V (2015) Talent management: progress and prospects. Hum Resour Manag Rev 25(3):233–235. https://doi.org/10.1016/j.hrmr.2015.04.005

Dong Y, Liu Y, Lian S (2016) Automatic age estimation based on deep learning algorithm. Neurocomputing 187:4–10. https://doi.org/10.1016/j.neucom.2015.09.115

Farrington DP, Gill M, Waples SJ, Argomaniz J (2007) The effects of closed-circuit television on crime: Meta-analysis of an English national quasi-experimental multi-site evaluation. J Exp Criminol 3(1):21–38. https://doi.org/10.1007/s11292-007-9024-2

Gwyn T, Roy K, Atay M (2021) Face recognition using popular deep net architectures: a brief comparative study. Future Internet 13(7):1–15. https://doi.org/10.3390/fi13070164

Harikrishnan J, Sudarsan A, Sadashiv A, & Remya Ajai AS (2019) Vision-face recognition attendance monitoring system for surveillance using deep learning technology and computer vision. In: Proceedings - International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN 2019. doi:https://doi.org/10.1109/ViTECoN.2019.8899418

Ilyas BR, Mohammed B, Khaled M, & Miloud K (2019) Enhanced face recognition system based on deep CNN. In: Proceedings - 2019 6th International Conference on Image and Signal Processing and Their Applications, ISPA 2019. doi:https://doi.org/10.1109/ISPA48434.2019.8966797

Kortli Y, Jridi M, Al Falou A, Atri M (2020) Face recognition systems: a survey. Sensors (switzerland). https://doi.org/10.3390/s20020342

Kukla E, Nowak P (2015) Facial emotion recognition based on cascade of neural networks. Adv Intell Syst Comput 314:67–78. https://doi.org/10.1007/978-3-319-10383-9_7

Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Mansanet J, Albiol A, Paredes R (2016) Local Deep Neural Networks for gender recognition. Pattern Recogn Lett 70:80–86. https://doi.org/10.1016/j.patrec.2015.11.015

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. Section 3. https://doi.org/10.5244/c.29.41

Peng C, Wang N, Li J, Gao X (2019) DLFace: Deep local descriptor for cross-modality face recognition. Pattern Recogn 90:161–171. https://doi.org/10.1016/j.patcog.2019.01.041

Perdana AB, & Prahara A (2019) Face recognition using light-convolutional neural networks based on modified Vgg16 model. 2019 In: International Conference of Computer Science and Information Technology, ICoSNIKOM 2019. doi:https://doi.org/10.1109/ICoSNIKOM48755.2019.9111481

Ranjan R, Bansal A, Zheng J, Xu H, Gleason J, Lu B, Nanduri A, Chen J-C, Castillo C, Chellappa R (2019) A fast and accurate system for face detection, identification, and verification. IEEE

Trans Biomet, Behav, Identity Sci 1(2):82–96. https://doi.org/10.1109/tbiom.2019.2908436
Shakeel MS, Lam KM (2019) Deep-feature encoding-based discriminative model for age-invariant face recognition. Pattern Recogn 93:442–457. https://doi.org/10.1016/j.patcog.2019.04.028
Simonyan K and Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14
Smith, P., & Chen, C. (2019). Transfer learning with deep CNNs for gender recognition and age estimation. In: Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018, 2564–2571. doi:https://doi.org/10.1109/BigData.2018.8621891
Varghese AA, Cherian JP, Kizhakkethottam JJ (2015) Overview on emotion recognition system. In 2015 international conference on soft-computing and networks security (ICSNS) pp. 1–5. IEEE
Wang Q, Guo G (2019) Benchmarking deep learning techniques for face recognition. J vis Commun Image Represent 65:102663. https://doi.org/10.1016/j.jvcir.2019.102663
Xu J (2021) A deep learning approach to building an intelligent video surveillance system. Multimed Tools Appl 80(4):5495–5515. https://doi.org/10.1007/s11042-020-09964-6