Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Real-Time Multiple Face Detection using MTCNN

Dr. P Venkateswara Rao[1], Dr. G S Ramesh[2], Sahith[3], Nikhil[4], Sai Rashmitha[5], Sandhya[6]

[1,2]Assisant Professor, Department of CSE, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana – 500090 [3,4,5,6]Students, Department of CSE, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana – 500090

ABSTRACT

For the security and investigation process, face recognition and identification are essential. It is an important part of computer vision. Human faces are important and need to be recognized for a variety of reasons. For the purpose of identifying people, a variety of biometric applications are available, such as face, iris, and fingerprint identification. A facial recognition system can identify or confirm a person from a digital image or video frame. Attendance control systems, mobile phone unlocking, social media tagging, payments, advertising, disease diagnosis, etc. are only a few examples of real-time facial recognition applications. The more sophisticated facial recognition systems can identify several faces in a single image, video, or livestream, aswell as faces from different angles. High identification rates and short training times are key prerequisites for real-time facial recognition applications. The FaceNet and MTCNN are combined in the suggested face recognition model to extract and classify features from faces that are embedded in images, respectively. The idea of transfer learning is utilized to shorten training time and boost recognition rates. The Multi-Task Cascaded Convolution Neural Network (MTCNN) model is used to get the 5-point landmarks from face frames, and the extracted face frame is then sent to FaceNet to extract embedding with a similarity of 99.8.

Keywords: Face detection, FaceNet, Face recognition, MTCNN, ResNet.

I. INTRODUCTION

Facial recognition systems have spread everywhere across the world. It is one of the most significant applications of face detection. It is used to identify an individual's identity using their face. It analyzes facial features from video or live stream to compare the incoming face toa dataset of trained images.

In today's world, face recognition has become one of the key components of the verification process for users during the online exam process. The system takes a photo of the use attempting the exam and it is compared with a photo stored in the system at the start of the online exam. Candidates transferring from the exam tab to other browser windows, straying from the webcam,

using a mobile device, and having numerous people present throughout the exam are a few examples of common malpractices during online exams. Such actions compromise the integrity of an online exam. Face recognition and multiple face detection hence contribute to a decrease in malpractice.

In this paper, a system is designed for recognizing multiple persons faces in real time or live video. This focusses on multiple person detection during proctoring in frames. It finds weather there is a single person or multiple persons in the frame and displays the name of known person and marks other person as unknown.

A deep learning model called MTCNN (Multi-Task Cascaded Convolutional Neural Network) is used to recognize faces and facial landmarks. It is a three-stage neural network architecture made up of three separate neural networks, each of which is designed to perform a particular task. The first network generates candidate regions that may contain faces using convolutional neural networks (CNNs). The second network filters the candidate regions to reject false positives and refine the boundingboxes around the faces. The third network performs facial landmark detection on the refined regions to locate the keyfacial features. Google researchers created FaceNet, a deep learning network for facial recognition. It extracts information from faces using a deep convolutional neural network and maps those features into ahigh-dimensional space where the distances between vectors reflect how similar the faces are to one another.

The three photos used to train the triplet loss function utilized by the FaceNet algorithm are an anchor image, a positive image (one of the same person as the anchor), and a negative image (one of a different person than the anchor). The algorithm is now able to minimize the distance between the anchor and positive photos while increasing the distance between the anchor and negative images in the feature space. During inference, FaceNet maps a new face image to the feature space and compares it to the feature vectors of the known faces in the database using a similarity metric (typically cosine similarity). The algorithm can then recognize the person in the image if the similarity score is above a certainthreshold. It can detect and recognize the faces in the live video and give out accuracy.

Section 2 of this paper introduces Face recognition using Multi-Task Cascaded Convolution Neural Network (MTCNN). Section 3 of this paper contains literature survey of three papers, read and understood which were further used for research. The steps of the suggested algorithm and its operation are described in Section 4. The experimental data are presented with tables and graphs in Section 5 of this publication. displays the model's performance as well. The model's conclusion and the paper's future scope are provided in Section 6.

II. RELATED WORK

In this section, various existing research papers on Face Recognition methodologies are collected and analyzed.

Real-time multiple face detection using active illumination [1] – 2020

This study's multiple face detector is founded on accurate pupil detection. The active lighting used by the pupil detector accelerates detection by utilizing the ability of the eyes to retroreflect light.The detection range of this method is appropriate for interactive desktop and kiosk applications. Heuristic techniques are utilized to filter and group the pupil candidates into pairs that match faces after computing the positions of the student candidates. A dual mode face tracker that is initialized using the most conspicuous detected face was created to show theresilience of the face detection technique. The tracked face is always kept in the center of the image by moving the camera in real-time using a pan-tilt servo mechanism that is controlled bythe estimated position of the face.

Accuracy: Active illumination can provide high-quality and accurate depth information, resulting in more accurate face detection and recognition compared to other techniques such aspassive cameras.

Speed: The active illumination technique can operate in real-time, which is critical in applications such as surveillance, security, and robotics.

Low Light: Active illumination can perform well in low-light conditions, as the structured light source can help to overcome challenges such as shadows and reflections.

Non-Intrusive: Active illumination is a non-intrusive technique that does not require any contact with the subjects being detected, making it ideal for use in public places.

Cost: The active illumination technique can be more expensive than other face detection techniques, as it requires specialized hardware such as a structured light source and a camera that can capture the reflected light.

Complexity: The active illumination technique is more complex than other face detection techniques, requiring more advanced image processing algorithms and calibration procedures to ensure accurate and reliable results.

Eye safety: Active illumination can emit bright light, which can be harmful to the human eye if not properly controlled, making safety concerns a priority.

Limited Range: Active illumination is limited by the range and field of view of the structured light source, which may not be suitable for larger spaces or scenarios that require more extensive coverage.

Multiple face detection algorithm using color skin modelling [11]- 2021

Due to the wide range of skin tones and other constraints on photo taking within a certain environment, face detection is a challenging issue in image processing for surveillance and security applications. This paper proposes a color-based skin model for a wide range of skin tones. The proposed skin model outperforms a number of existing skin models, it is found. Using the suggested skin model, a framework for soft decision-based multiple face detection is developed. The robustness of the proposed framework has been evaluated for a wide range of poses, expressions, complex backgrounds, and blur noises. The recommended skin model and soft decision-based strategy have restricted the range of Cb and Cr values to achieve less misleading and more accurate results than the earlier methods.

Simplicity: The color skin modeling technique is relatively simple and can be implemented efficiently in real-time applications.

Robustness: Skin color is a robust feature for detecting faces, as it remains relatively stable across variations in illumination, facial expressions, and pose.

Non-Intrusive: Color skin modeling is a non-intrusive technique that does not require any contact with the subjects being detected, making it ideal for use in public places.

Computational efficiency: Color skin modeling is computationally efficient compared to other techniques such as template matching or feature-based detection.

Accuracy: Skin color may not always be a reliable feature for face detection, as it can be affected by factors such as makeup, skin tone, and lighting conditions, resulting in false positives or false negatives.

Limited use: The color skin modeling technique can only be used for detecting faces of light-skinned people, as it is not suitable for detecting faces of dark-skinned people.

Vulnerability: The color skin modeling technique can be vulnerable to spoofing attacks, where attackers can manipulate the color information of their skin to evade detection.

High dependence on calibration: The accuracy of the calibration of the skin color model, which can be difficult to achieve, is crucial for the performance of the color skin modelling technique.

A new face detection method based on shape information [12] - 2020

One of the trickiest problems in pattern recognition is automatic facial identification. Face photographs frequently have a plain background in many practical applications (such aspersonal identification utilizing photo-IDs). In this study, we provide a novel approach based on shape data. For photographs with a plain background, our system can be highly effective. Histogram equalization is used to improve the input image before edge detection using a multiple-scale filter.

Then, a technique based on an energy function is used to link the extractededges. Finally, the face contour is retrieved using the associated edges' direction data. The efficiency of this approach is confirmed by the experimental results.

Accuracy: Shape-based face detection methods can be highly accurate as they can analyze the geometrical features of facial components to detect faces, which can be highly distinctive.

Robustness: Shape-based face detection methods are relatively robust to variations in lighting conditions, facial expressions, and pose, as they can focus on the geometric shapes of facial components rather than color or texture.

Reliability: Shape-based face detection methods can be highly reliable as they can analyze multiple facial components simultaneously, improving the overall detection accuracy.

Security: Shape-based face detection methods can be more secure than other techniques such as color skin modeling, as they are less vulnerable to spoofing attacks.

Computational complexity: Shape-based face detection methods can be computationally complex, as they involve analyzing multiple facial components and performing complexmathematical operations to detect faces.

Limited use: Shape-based face detection methods may not be suitable for detecting faces of individuals with unusual facial features or facial deformities, as these can cause errors in the shape-based analysis.

Difficulty with occlusions: Shape-based face detection methods may struggle with detecting faces that are partially occluded, as it can be challenging to determine the exact shape of facial components that are not fully visible.

Sensitivity to training data: Shape-based face detection methods are highly dependent on the quality and diversity of the training data, which can limit their generalizability to new datasets.

These papers helped to understand various existing research methodologies and techniques and problems faced by them, which were analyzed and understood. The proposed paper solves the problem faced in the existing system.

III. RESEARCH METHODOLOGY

Face detection and face recognition would be the two primary parts of the system that is suggested for a face detection project employing MTCNN and FaceNet. Face recognition would be accomplished using MTCNN (Multi-Task Cascaded Convolutional Networks), a cutting-edge deep learning method that identifies faces in a picture and outputs the bounding box coordinates of each face.The output of the face detection component would then be passed to the face recognition component, which would use FaceNet, a deep neural network that extracts high-

dimensional feature vectors from faces, to recognize and identify the detected faces.

We would first need to train the FaceNet model on a sizable dataset of face photos in order to learn the high-dimensional feature vectors for each face before we could apply the suggested method. Once trained, we would use the model to encode each face in our dataset into a feature vector, which would act as a distinctive representation of that face. These feature vectors and their accompanying names would then be kept in a database.

The MTCNN algorithm would identify the faces in the image during runtime and output the bounding box coordinates of each face. With the help of these bounding box coordinates, the face would be removed from the original image and processed by the FaceNet model to produce a feature vector for that face. Then, using a distance metric like cosine similarity, this feature vector would be compared to the feature vectors in the database. If the distance between the vectors is below a predetermined threshold, the face would be recognized and given the matching name from the database. Then, the names associated with the recognized faces would be labelled and a bounding box would be used to highlight them in the original image. The predicted bounding box coordinates (x, y, width, height) are adjusted based on the learnedtransformation parameters (Δx, Δy, Δw, Δh).

$New\_x = x + \Delta x$, $New\_y = y + \Delta y$, $New\_width = width + \Delta w$,
$New\_height = height + \Delta h$

Face Recognition Through Multi-Task Cascaded Convolution Neural Network (MTCNN)

When given inputs like images, voice, or audio, for example, convolutional neural networks perform better than other neural networks. A CNN normally has three layers: convolutional, pooling, and fully connected.

In the kernel convolution procedure, we take a small number matrix, pass it over our image, and then adjust it in accordance with the filter's findings. The following formula yields the values of the subsequent feature map, where the input picture is denoted by letter f and our kernel by letter h. The numbers m and n, respectively, are used to represent the rows (m)and columns(n) of the result matrix's indexes.

Output feature map $O(i, j) = \sum(I(i + m, j + n) * F(m, n))$

Several Tasks Cascaded A neural network called a convolutional neural network can recognise faces and other facial landmarks in photos. MTCNN is one of the most popular and trustworthy face detection systems on the market right now. This system is a cascade coupling of three neural networks. It can still be used to undertake numerous real-time face detection tests. The three stages of the MTCNN are as follows:

i. The Proposal Network (P-Net)

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Fully convolutional network (FCN) is the first stage. Dense layer is not a component of the architecture in FCN. Candidate window and their bounding box regression are achieved using this P-Net.

ii. The Refine Network (R-Net)

R-Net receives the whole P-Net candidate pool. This network is currently CNN but is not an FCN because the ultimate step of building involves many layers.

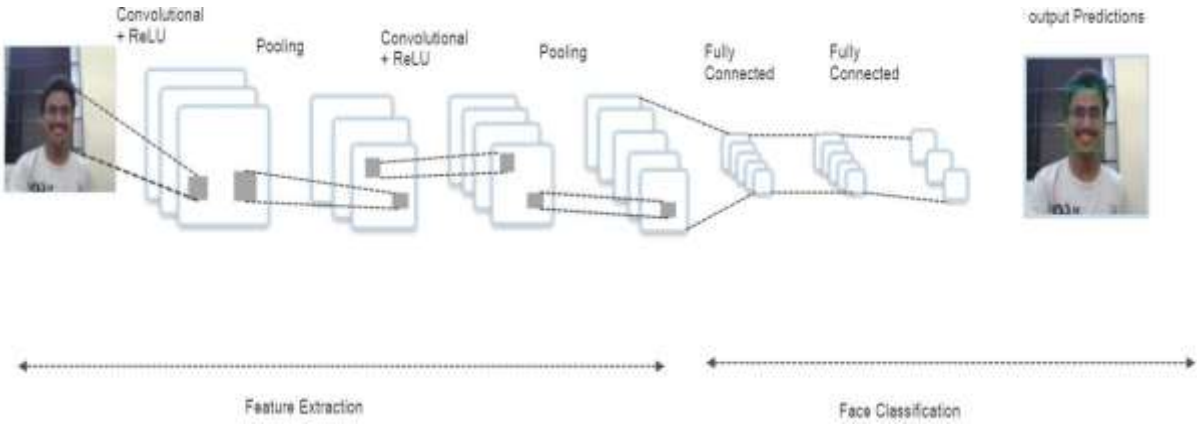
Depending on whether the input is a face or not, R-Net produces a 4-element vector and a 10-element vector for localizing the facial landmarks.

Max Pooling:

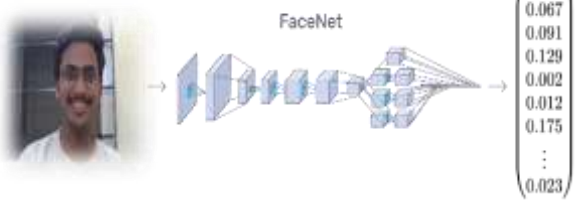
$$O(i, j) = \max(I(i + m, j + n))$$

iii. The Output Network (O-Net)

The goal of this network is to provide a more detailed description of the face as well as the locations of the five facial markers for the eyes.



A deep neural network called **FaceNet** is used to recognize features in a facial image. Google researchers published it in 2015. FaceNet creates a vector of 128 integers from a face image as input, representing the most significant facial features.



In machine learning, this vector is referred to as an embedding. Images of the same person are mapped to (about) the same position in the coordinate system when the hash code is embedded. FaceNet uses a face image as its input to produce the embedding vector.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Fig 1 FaceNet

Output vector $Y = W * X + b$

The development of ResNets, which also introduced the concept of residual blocks, addressed the vanishing/exploding gradient problem. In this network, we use a technique known as skip connection. To connect layer activations to next layers, the skip connection skips over some intermediate levels. Consequently, a block is left over. To build resnets, these leftover blocks are piled.

Instead of having layers learn the underlying mapping, the approach used by this network is to let the network fit the residual mapping. So, instead of using, example, the initial mapping of H(x), let the network fit.

ResNet-L's overall output, Y, can be calculated as follows:

Y = H_n(H_{n-1} (...(H_2(H_1(X)))...). In metric learning tasks, particularly those involving face recognition and picture retrieval, the triplet loss function is employed. TripletAloss aims to learn an embedding space with a maximum distance between dissimilar samples and a minimum distance between comparable examples. The distance between an anchor sample, a positive sample (which is comparable to the anchor), and a negative sample (which is dissimilar to the anchor) is compared to accomplish this.

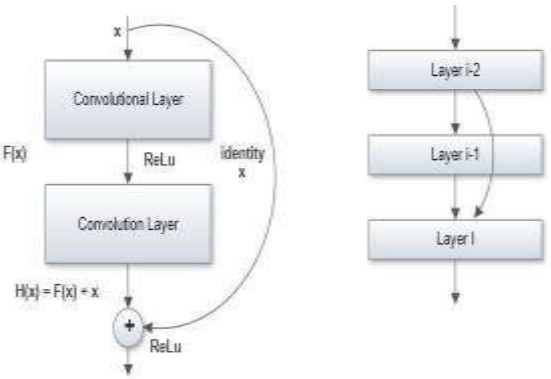
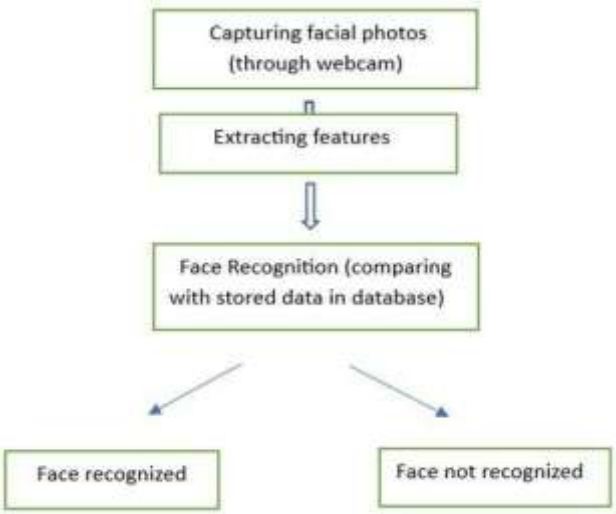


Fig 2 Residual Mapping

The ResNet architecture is used as the backbone network to extract deep features from the input images, and the triplet loss is applied to learn an embedding space where similar samples are closer together and dissimilar samples are farther apart. Combining ResNet with triplet loss is a common approach in deep metric learning.



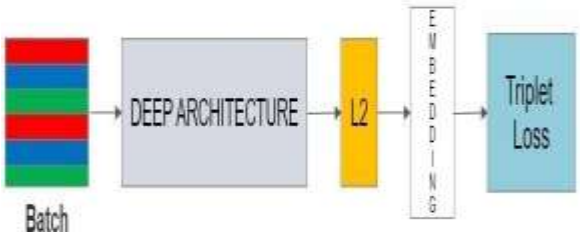


Fig 4 ResNet Architecture

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

User Face recognised unknown user detected

User Face recognized

Fig 3 Multi face detection flowchart

IV. DATASET

The "Labelled Faces in the Wild" (LFW) dataset is the one most frequently linked to FaceNet. LFW is a well-liked benchmark dataset used to assess FaceNet and other facial recognition software. Around 5,000 people are included among the more than 13,000 face photos in the LFW collection, which was compiled from web-based sources.

Due to changes in stance, lighting, facial emotions, and image quality, the LFW dataset presents a number of hurdles for face identification. It is separated into a training set and a test set. The test set contains pairs of images that are either labelled as "same person" or "different person" in order to assess the model's performance on face verification tasks.

V. RESULTS

In the context of multiple face detection in video, similarity refers to the process of assessing whether the identified faces in consecutive frames of a video belong to the same person. A higher similarity suggests that the faces caught in the video are more likely to match those in the database. It is a numerical value which measures the degree of resemblance between the facial features of the individuals.

The MTCNN algorithm would identify the faces in the image during runtime and output the bounding box coordinates of each face. These bounding box coordinates would be used to crop out the face from the original image, which would then be passed through the FaceNet model to generate a feature vector for that face.

This feature vector would then be compared to the feature vectors in the database using a distance metric such as 25 cosine similarity, and if the distance between the vectors is below a certain threshold, the face would be recognized and labeled with the corresponding name from the

database.

Finally, the recognized faces would be highlighted in the original image with a bounding box and labeled with their corresponding names. If the model finds multiple people in the frame, then it displays the name of a known person and marks the other person as unknown.

The below are the results of various models and CNN architectures available

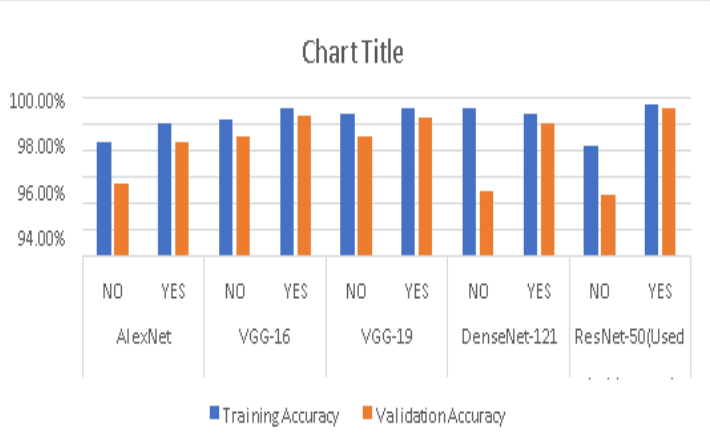


Fig 5 training and validation accuracy

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Input	Actual Result
Testing with person's image	Detected person's image
	
	Similarity = [99.79]
	
	Unknown face detected
	
	Similarity = [99.85]
	
	Similarity = [99.82] and unknown
	
	Similarity = [99.82] , [99.74]

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Multiple Unkonwn faces detected

Training and Validation Accuracy: A model's accuracy when applied to new data is validation accuracy, but its accuracy when applied to the data it was trained on is training accuracy.

Since the model has never seen the validation data before, validation accuracy is typically lower than training accuracy. The goal is to generalize the model effectively on new data, which means it should perform well on examples which are unknown.

CNN architectures	Pre-trained	Training Accuracy	Validation Accuracy
AlexNet	NO	96.64%	93.59%
	YES	98.01%	96.63%
VGG-16	NO	98.32%	97.13%
	YES	99.22%	98.58%
VGG-19	NO	98.71%	97.06%
	YES	99.15%	98.53%
DenseNet-121	NO	99.20%	92.90%
	YES	98.73%	98.01%
ResNet-50(Used Architecture)	NO	96.33%	92.72%
	YES	99.43%	99.17%

Fig 6 Results

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, a strong solution for real-time face detection and recognition is provided by the union of FaceNet and MTCNN. MTCNN is a highly precise face detection technique that can find faces in real-time video streams, whereas FaceNet is an efficient deep learning model for creating face embeddings. These two technologies work well together because they can accurately identify and recognize faces in real-time, which makes them perfect for uses like surveillance, access control, and facial recognition.

To boost the system's performance, more work can be done. For instance, by fine-tuning the FaceNet model on a larger dataset of faces or by training it to recognize faces in various lighting conditions and angles, one can increase the accuracy of face identification and recognition. The system may also be expanded with new functions, such the capacity to follow faces over time or recognize facial

emotions.

All things considered, real-time face detection and recognition utilizing FaceNet and MTCNN is a promising technology with many potential uses. It might potentially get even more precise, effective, and secure with more study and improvement. Multi-face detection is useful in video surveillance systems, where it helps track and monitor multiple individuals simultaneously. It can be used for security purposes in public places, crowd monitoring, or identifying suspicious activities by detecting multiple faces in real-time.

By capturing and analyzing multiple faces during the exam, it becomes possible to detect if someone else attempts to take the exam on behalf of the registered candidate. (Exams like TOEFLexam, GRE). Accurate Attendance Recording an be employed with the help of multi face detection algorithm, attendance systems can accurately capture the presence of everyone in real time.The system can detect faces, match them against a pre-registered database, and mark attendance for each recognized individual.

Future Scope would be integration of speech modules with multi-face detection which can open exciting possibilities for multimodal interaction and analysis. Multi-face detection can be combined with speech analysis which assists in health monitoring and diagnosis. By analyzing facial cues and speech patterns, systems can detect and monitor various health conditions, such as stress levels, cognitive impairments, or speech disorders, contributing to early detection and intervention.

VII. REFERENCES

[1] Morimoto, Carlos Hitoshi, and Myron Flickner. "Real-time multiple face detection using active illumination." In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 8-13. IEEE, 2000.

[2] Wu, Chunming, and Ying Zhang. "MTCNN and FACENET based access control system for face detection and recognition." Automatic Control and Computer Sciences 55 (2021): 102112.

[3] Zhang, Ning, Junmin Luo, and Wuqi Gao. "Research on face detection technology based on MTCNN." In 2020 international conference on computer network, electronic and automation (ICCNEA), pp. 154-158. IEEE, 2020.

[4] Jose, Edwin, M. Greeshma, Mithun TP Haridas, and M. H. Supriya. "Face recognition-based surveillance system using facenet and mtcnn on jetson tx2." In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 608-613. IEEE, 2019.

[5] Xiang, Jia, and Gengming Zhu. "Joint face detection and facial expression recognition with MTCNN." In 2017 4th international conference on information science and control engineering (ICISCE), pp. 424-427. IEEE, 2017.

[6] Ku, Hongchang, and Wei Dong. "Face recognition based on mtcnn and convolutional neural network." Frontiers in Signal Processing 4, no. 1 (2020): 37-42.

[7] Li, Xiaochao, Zhenjie Yang, and Hongwei Wu. "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks." IEEE Access 8 (2020): 174922174930.

[8] Arora, Mehul, Sarthak Naithani, and Anu Shaju Areeckal. "A web-based application forface detection in real-time images and videos." In Journal of Physics: Conference Series, vol. 2161, no. 1, p. 012071. IOP Publishing, 2022.

[9] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[10] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[11] Galvez, Reagan L., Argel A. Bandala, Elmer P. Dadios, Ryan Rhay P. Vicerra, and Jose Martin Z. Maningo. "Object detection using convolutional neural networks." In *TENCON 2018-2018 IEEE Region 10 Conference*, pp. 2023-2027. IEEE, 2018.

[12] Zhiqiang, Wang, and Liu Jun. "A review of object detection based on convolutional neural network." In *2017 36th Chinese control conference (CCC)*, pp. 11104-11109. IEEE, 2017.

[13] Yanagisawa, Hideaki, Takuro Yamashita, and Hiroshi Watanabe. "A study on object detection method from manga images using CNN." In *2018 International Workshop on Advanced Image Technology (IWAIT)*, pp. 1-4. IEEE, 2018.

[14] Dhillon, Anamika, and Gyanendra K. Verma. "Convolutional neural network: a review of models, methodologies and applications to object detection." *Progress in Artificial Intelligence* 9, no. 2 (2020): 85-112.

[15] Sultana, Farhana, Abu Sufian, and Paramartha Dutta. "A review of object detection models based on convolutional neural network." *Intelligent computing: image processing based applications* (2020): 1-16.

[16] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time

object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).
[17] Kang, Kai, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang et al. "T-cnn: Tubelets with convolutional neural networks for object detection from videos." *IEEE Transactions on Circuits and Systems for Video Technology* 28, no. 10 (2017): 2896-2907.
[18] "Domain adaptive faster r-cnn for object detection in the wild," by Christos Sakaridis, Dengxin Dai, Chen, Yuhua, Wen Li, and Luc Van Gool. 2018; 3339–3348 in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.