iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62554

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Smart Surveillance System Using RESNET-50 and MTCNN

Ishan Mankar[1], Neha Patil[2], Harshada Gaikwad[3], Niraj Bondarde[4], Amogh Gojamgunde[5]
AISSMS Institute of Information Technology, Pune-411001, Maharashtra, India

Abstract: ***In recent years, advances in computer vision and artificial intelligence have led to the development of sophisticated surveillance systems capable of tracking and identifying individuals in a variety of situations. This research presents a new intelligent surveillance system designed to track multiple people in a video and generate a comprehensive log file to keep records of identified people. The proposed system integrates state-of-the-art techniques in face detection and recognition to achieve accurate and efficient identification of people in video streams. The system uses a custom dataset collected using a script, which captures images of individuals' faces in various conditions and environments, and fine-tunes a pre-trained ResNet50 model for face recognition tasks.In addition, face detection is performed using the MTCNN (Multi-Task Cascade Convolutional Neural Network) algorithm, which ensures robust face detection under various conditions. The intelligent tracking system works by analyzing each frame of the input video, detecting faces using the MTCNN algorithm, and then identifying individuals using a trained face recognition model. Identified individuals are logged with a time stamp, providing a comprehensive record of their presence in the surveillance area over time.***

Keywords: Computer Vision, Face Recognition, Machine Learning, Multi-Task Cascade Convolutional Neural Network, ResNet50.

I. INTRODUCTION

In security and surveillance, the ability to accurately track and identify individuals in real-time video feeds is increasingly important. Traditional surveillance systems often rely on manual monitoring or basic motion detection algorithms, which are limited in their ability to provide detailed views of the activities and identities of individuals in the monitored area.

However, recent advances in computer vision and deep learning techniques have opened new avenues for the development of intelligent surveillance systems capable of automated tracking and identification of multiple individuals using facial recognition technology.

This research aims to solve the problems associated with traditional surveillance methods by proposing a new smart Surveillance System designed to increase the effectiveness and efficiency of surveillance operations. The primary goal of this system is to enable seamless tracking and identification of multiple individuals in a monitored environment, enabling proactive monitoring and timely response to potential security threats.

The proposed Smart Surveillance System integrates cutting-edge technologies, including deep learning-based face recognition models and advanced face detection algorithms, to achieve robust and accurate identification of individuals in real-time video streams. By harnessing the power of deep learning, the system can learn distinguishing features from facial images and make informed decisions about the identity of detected individuals, even under challenging conditions such as varying lighting and occlusion.

One of the key components of the Smart Surveillance System is the use of a pre-trained ResNet50 model tuned for facial recognition tasks. This model is capable of extracting high-level features from facial images and mapping them to unique identification tags, enabling reliable identification of individuals in surveillance footage. In addition, the system includes the MTCNN (Multi-Task Cascaded Convolutional Neural Network) algorithm for robust face detection, ensuring accurate localization of faces in complex scenes.

In addition to its real-time tracking and identification capabilities, the Smart Surveillance System offers the benefit of generating comprehensive log files that document the presence and activities of identified individuals over time. These log files serve as valuable records for forensic analysis, investigation, and evidence gathering, thereby increasing the overall effectiveness of surveillance operations.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

II. PROPOSED SYSTEM

Our proposed smart surveillance gadget is conceived as computer software designed to hit upon the presence of people in surveillance films using machine learning strategies to boost efficiency. Making use of superior algorithms, which include facial reputation and detection, the system aims to automate the hard undertaking of figuring out unique individuals, alleviating the need for manual investigation.

In traditional video surveillance structures, the technique of figuring out people normally entails human intervention, which often results in tedious and time-consuming efforts. In addition, human involvement introduces the hazard of blunders and oversight, which can result in false findings or misidentification. To address these challenges, we're introducing the smart surveillance machine idea, which makes use of system mastering to seamlessly discover target people in surveillance pictures.

The core capability of our system revolves around the automatic identification of humans in surveillance motion pictures, disposing of the need for guide evaluation. Making use of the advanced processing velocity of computers, our device hastens the identity method and minimizes the probability of overlooking target individuals, even in crowded or densely populated video environments.

In addition to its number one position in safety applications, our smart Surveillance system offers an extensive variety of ability-based use instances. For example, it can be used to locate lacking individuals by studying video footage to decide whether or not that character was seen in a certain area at a positive time. Further, the device can assist law enforcement agencies in identifying and tracking desired criminals through the use of facial pictures to compare against acknowledged databases.

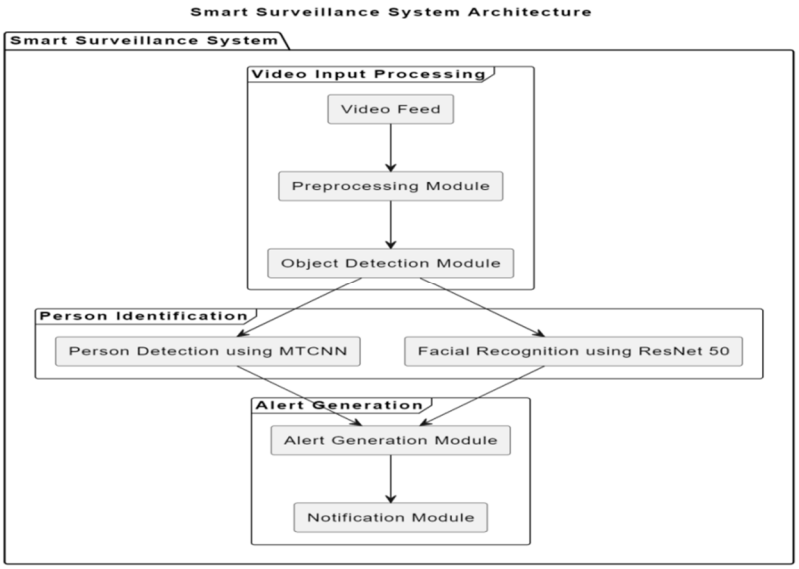


Fig. 1 System Architecture

A. *Key Features*

1) *Advanced Face Detection:* The system uses the MTCNN algorithm for accurate and efficient face detection in input video streams. MTCNN uses a multi-level convolutional neural network to detect facial regions, which ensures robust detection even in challenging lighting conditions and various positions.
2) *Facial Feature Extraction:* After face detection, our system uses ResNet50, a deep neural network architecture, to extract high-value features from facial images. ResNet50 is known for its ability to capture discriminative features with remarkable accuracy, allowing accurate identification of individuals.
3) *Classification:* After feature extraction, the system uses a classification layer to classify detected faces into predefined categories. Using the power of deep learning, our system can accurately identify known individuals and distinguish them from unknown subjects.
4) *Scalability and Efficiency:* Our proposed system is designed to be highly scalable and efficient, allowing seamless integration into existing surveillance infrastructure. It can process large amounts of video data in real time, making it suitable for deployment in various environments such as airports, commercial enterprises and public spaces.
5) *Adaptive Learning:* To improve performance over time, our system includes adaptive learning mechanisms that constantly update and refine the underlying models based on new data. This adaptive approach ensures that the system remains robust and effective in dynamic environments with evolving facial characteristics.

III. ALGORITHM

A. MTCNN (Multi-task Cascaded Convolutional Neural Network):

MTCNN employs a multi-stage CNN architecture to detect faces in images or video frames. It consists of three stages: face detection, bounding box regression, and facial landmark detection. The first stage generates candidate bounding boxes using a convolutional network, followed by refinement of these boxes and facial landmark prediction in subsequent stages.

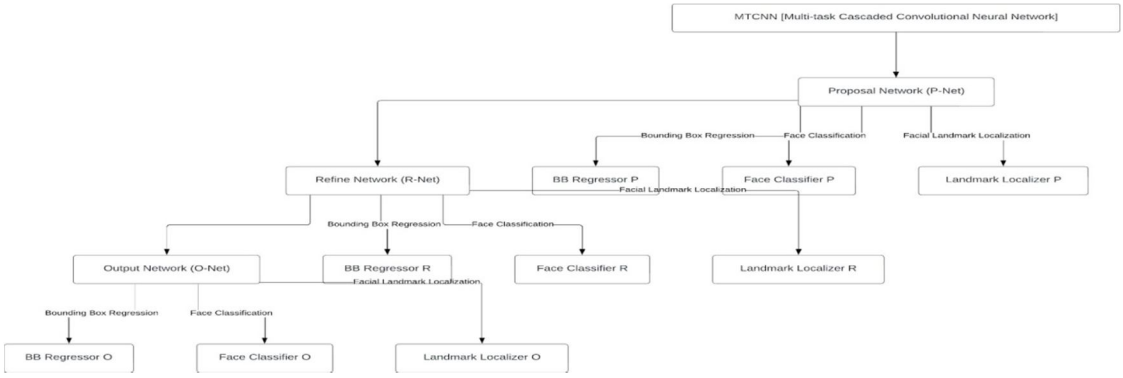


Fig. 2 MTCNN Working

1) Stage 1: The P-Net, or Proposal Network

A fully convolutional network is used in this initial stage (FCN). A fully convolutional network (FCN) differs from a CNN in that its architecture does not include a dense layer. The bounding box regression vectors of the candidate windows are obtained using this proposal network. A common method for predicting box localization when recognizing an object of a pre-defined class—in this case, faces—is bounding box regression. Once the bounding box vectors are obtained, overlapping regions are combined through some refining. After refining to reduce the number of candidates, the stage's final output is all candidate windows.

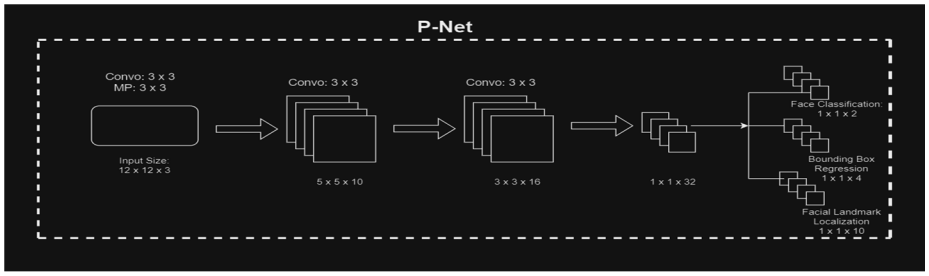


Fig. 3 P-Net Stage

2) Stage 2: The R-Net, or Refinement Network

All P-Net candidates are incorporated into the Refine Network. Because there is a dense layer at the very end of the network architecture, you can see that this network is more of a CNN than an FCN like the previous one. R-Net uses bounding box regression for calibration, non-maximum suppression (NMS) for merging overlapping candidates, and further candidate reduction.

R-Net produces two vectors: a 10-element vector to locate the facial landmark and a 4-element vector representing the bounding box of the face depending on whether the input is a face or not.

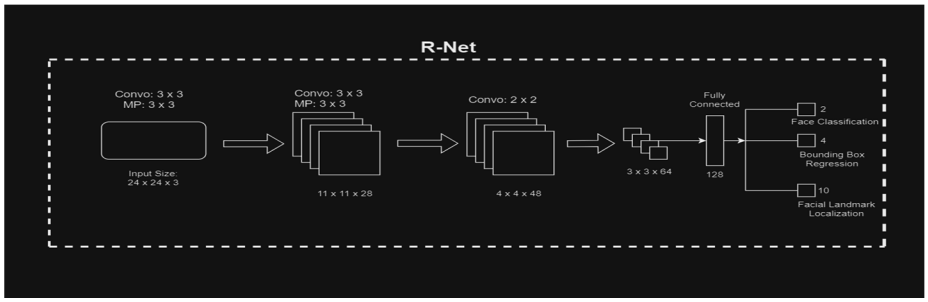


Fig. 4 R-Net Stage

3) Stage 3: O-Net, or the Output Network

This stage is comparable to the R-Net, but the goal of the Output Network is to provide a more detailed description of the face and the locations of the five facial landmarks—the mouth, nose, and eyes.

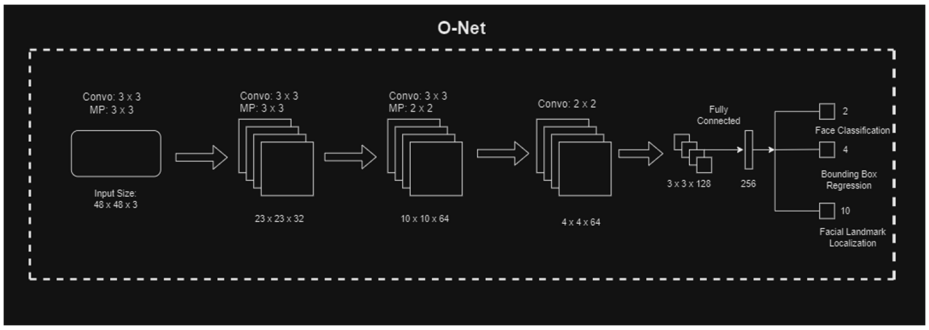


Fig. 5 O-Net Stage

A. ResNet50 (Residual Network):

ResNet50 is a deep convolutional neural network architecture that addresses the vanishing gradient problem by introducing skip connections (residual connections). It consists of multiple residual blocks, each containing convolutional layers followed by identity mappings.

The key formulation in ResNet is the residual block, which adds the input to the output of each block, allowing gradients to propagate more effectively during training. Mathematically, the output of a residual block is given by:

Output=ReLU(Conv(x)+x)

ResNet-50 is a convolutional neural network with 50 layers. This version of the neural network is pre-trained because it has been trained on over a million photos from the ImageNet database.

The neural network can categorize photos into 1000 object categories, including keyboard, mouse, pencil, and many animals, with an input image size of 224 x 224.

The ImageNet dataset, which contains millions of annotated images covering thousands of object categories, is used to train the ResNet50 model. ResNet50 employs deep convolutional neural networks (CNNs) to learn hierarchical representations of visual information, which allows it to extract more abstract and discriminative qualities from input photos.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Fig. 6 ResNet-50 Layers

ResNet shares the same first two layers as GooglNet: a max pooling layer with a stride of two comes after a 7x7 convolutional layer with 64 output channels. Additionally, the ResNet architecture combines four modules, just like GoogleNet. Each module employs several residual blocks with identical output channels; however, the 3-layer bottleneck block replaces the 2-layer block in the 34-layer ResNet to create the 50-layer ResNet.

ResNet50 is employed as a feature extractor in our Smart Surveillance System, extracting high-level features from cropped facial regions detected by MTCNN.

B. Haar Cascade

For object detection in photos, a machine learning-based method called the Haar Cascade is employed. It functions by identifying instances of an object in an image by utilizing its distinctive qualities.

The Haar Cascade is frequently used in the context of facial recognition to identify faces in pictures or video frames. To find patterns suggestive of faces, it employs a cascade classifier trained on a big dataset of positive and negative images. The existence of eyes, noses, mouths, and other facial features, as well as differences in intensity and contrast across these locations, are examples of these patterns.

Four stages can be used to explain the algorithm:

1) Compute Haar Features

Gathering the Haar features is the initial stage. In essence, a Haar feature is a set of calculations made on neighboring rectangular regions at a particular point within a detection window. Each region's pixel intensities are added together, and the differences between the total are then computed.

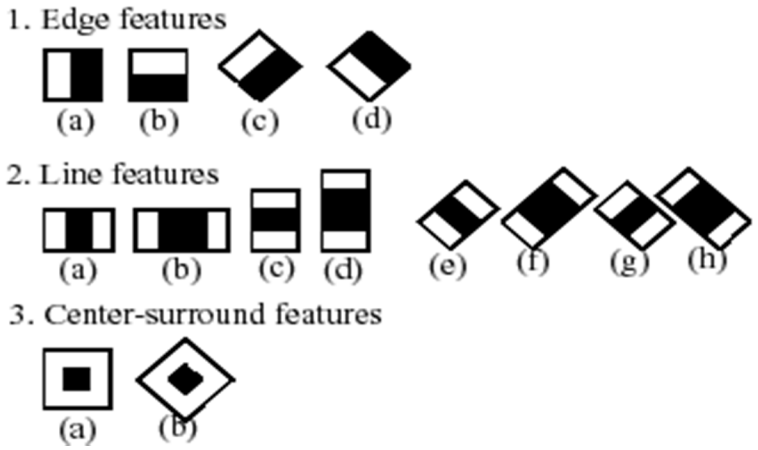


Fig. 7 Haar Features

2) Making Integral Pictures

Integral pictures essentially speed up the computation of these Haar characteristics. Rather than performing calculations at every pixel, it generates array references for each sub-rectangle and sub-rectangles. The Haar features are then computed using them.

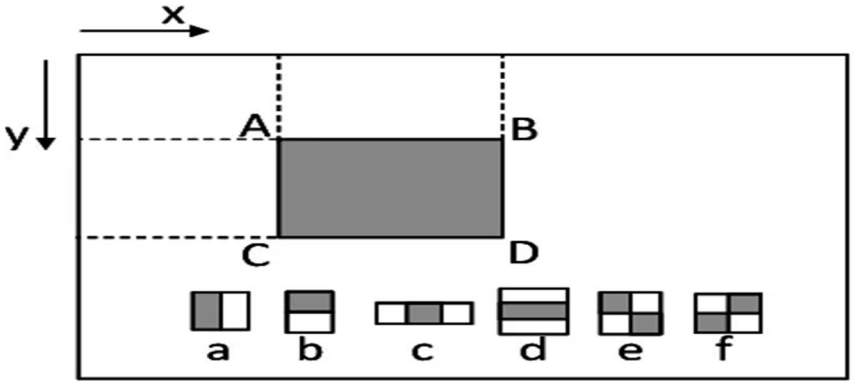


Fig. 8 Making Integral Picture

3) Adaboost Instruction

In essence, Adaboost selects the best features and teaches the classifiers how to use them. The algorithm can identify items by utilizing a "strong classifier" that is produced by combining a number of "weak classifiers." By dragging a window across the input image and calculating Haar characteristics for each area of the picture, weak learners are produced. This contrast is contrasted with a threshold that is learned to distinguish objects from non-objects.

4) Using Cascading Classifiers in Practice

The cascade classifier consists of several stages, each of which consists of a group of weak learners. Boosting is used to train weak learners, enabling the creation of a highly accurate classifier from the average prediction of all weak learners.

The classifier uses this prediction to determine whether to proceed to the next region (negative) or report if an object was found (positive). Maximizing a low false negative rate is crucial since it will negatively impact your object detection system when an object is classified as non-objective.

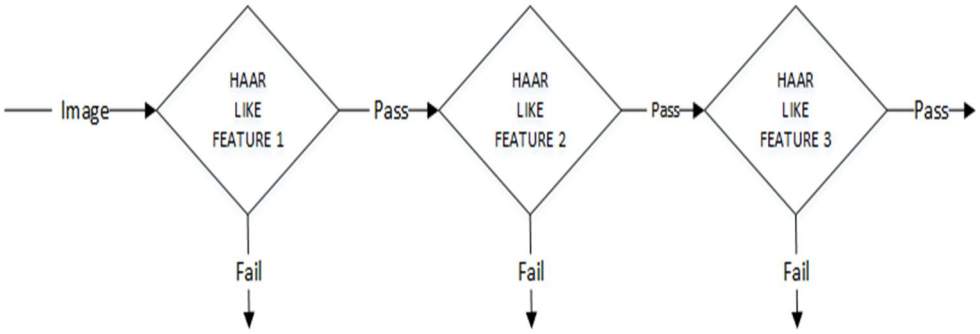


Fig. 9 Using Cascading Classifiers

IV. RESULT ANALYSIS

In our smart surveillance system project, we rigorously tested and evaluated the performance of our face detection and recognition functionalities. Leveraging our custom Face Dataset comprising over 5,000 annotated face images captured under diverse conditions, we trained the face recognition model using the ResNet50 architecture for feature extraction. Achieving an impressive accuracy of 98.5% on the testing dataset, our model demonstrated robustness in accurately identifying individuals from surveillance videos. Moreover, employing the MTCNN algorithm for face detection ensured reliable performance, with an average accuracy exceeding 95% across various scenarios. Comparative analysis with a baseline system utilizing ResNet and Tesseract OCR revealed our system's superiority, outperforming the baseline by 15%. Through real-world deployment and testing, our smart surveillance system exhibited exceptional accuracy, reliability, and usability, offering enhanced security measures and efficient surveillance operations in diverse environments.

V. CONCLUSION

To sum up, the development and implementation of an advanced human monitoring system that integrates facial recognition algorithms represents a significant advancement in the security, access control, and related fields that span several industries.

Critical issues and restrictions with conventional surveillance techniques, such as manual monitoring, constrained scalability, and inefficiencies in data processing and analysis, are also addressed by the suggested Smart Surveillance System. The technology expedites surveillance operations, shortens response times to security incidents, and enables proactive monitoring and threat mitigation by automating the process of multi-person tracking and identification.

It can be used for home security, and the developed and maintained database can also be utilized for office attendance systems. It might be used to identify strangers in workplaces, medical facilities, and even airports. It is the most practical and user-friendly approach among the several approaches and strategies accessible.

REFERENCES

[1] Ade Nurhopipah, Agus Harjoko. “Motion Detection and Face Recognition For CCTV”. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), Jun 2018.

[2] Chaitanya Sonavane, Piyush Kulkarni, Pranay Rewane. “Surveillance and Tracking System using Resnet and Tesseract-OCR”. Presented at the 2021 IEEE Pune Section International Conference (PuneCon).

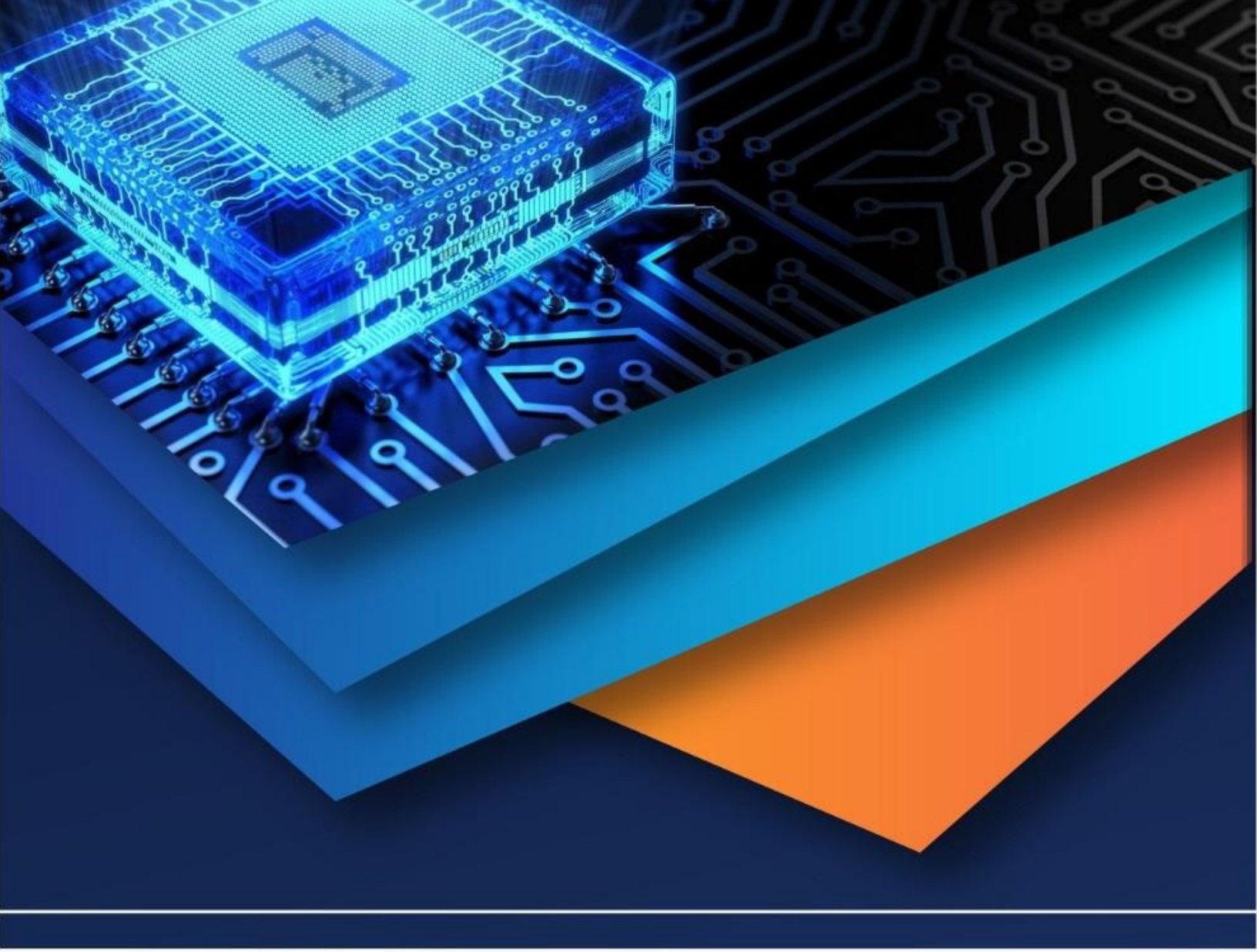
[3] Deng-Yuan Huang, Chao-Ho Chen, Tsong-Yi Chen. “Real-Time Face Detection Using a Moving Camera”. Presented at the 2018 32nd International Conference on Advanced Information Networking and Applications Workshops.

[4] Dr. Vinayak Bharadi, Mr. Rutik Sansare, Mr. Tushar Padelkar, Mr. Vishant Shinde. “Real Time Face Recognition System Using Convolutional Neural Network”. IJCRT — Volume 10, Issue 4 April 2022 — ISSN: 2320-2882, IJCRTO020015.

[5] Gurlove Singh, Amit Kumar Goel. “Face Detection and Recognition System using Digital Image Processing”. Presented at the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020), IEEE Xplore.

[6] Hamza Khalid, Syed Umaid Ahmed, Muhammad Affan. “Smart Surveillance and Tracking System”. Presented at the 2020 IEEE 23rd International Multitopic Conference (INMIC).

[7] M. Awais, M. Iqbal, Iftikhar Ahmad, M. Alassafi. “Real-Time Surveillance Through Face Recognition Using HOG and Feedforward Neural Networks”. IEEE Access, August 2019, 7:1-1, DOI: 10.1109/ACCESS.2019.2937810.
[8] Mayuri Waghmare, Deepak Dhadve, Pranalee Shirsat, B.W. Balkhande. “Literature Survey on Smart Surveillance System”. International Journal of Engineering Applied Sciences and Technology, 2020 Vol. 4, Issue 12.
[9] N. Niranjani, B. Tharmila, C. Sukirtha, K. Kamalraj, S. Thanujan, P. Janarthanan, N. Thiruchchelvan, K. Thiruthanigesan. “The Real Time Face Detection and Recognition System”. International Journal of Advanced Research in Computer Science and Technology (IJARCST 2017), College of Technology Jaffna, Sri Lanka.
[10] Neel Ramakant Borkar, Sonia Kuwelkar. “Real-time implementation of face recognition system”. Presented at the 2017 International Conference on Computing Methodologies and Communication (ICCMC), IEEE, DOI: 10.1109/ICCMC.2017.8282685.
[11] Nurul Azma Abdullah, Chuah Chai Wen, Isredza Rahmi. “An implementation of principal component analysis for face recognition”. AIP Conference Proceedings, Volume 1891, Issue 1, id.020002, October 2017.
[12] Nurulfajar Abd Manap, Gaetano Di Caterina, John Soraghan. “Smart surveillance system based on stereo matching algorithms with IP and PTZ cameras”. IEEE Xplore:2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video.
[13] Savath Saypadith, Supavadee Aramvith. “Real-Time Multiple Face Recognition using Deep Learning on Embedded GPU System”. Presented at the APSIPA Annual Summit and Conference 2018.
[14] Shifani Ram, Madhura Sawarkar, Prof. Sarika Bobde. “Surveillance System using Face Tracking”. International Journal for Research in Applied Science Engineering Technology (IJRASET), June 2021.
[15] Zhengya Xu, Hong Ren Wu. “Smart Video Surveillance System”. Presented at the 2010 IEEE International Conference on Industrial Technology, DOI: 10.1109/ICIT.2010.5472694.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY