iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.56500

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 11 Issue XI Nov 2023- Available at www.ijraset.com

Tracking People Using Face Recognition

Varanasi Amit Maheshwar
Department of Computer Science and Engineering, Vellore Institute of Technology, AP

Abstract: ***In recent years, face recognition has emerged as an active field within computer vision, driven by potential applications and the parallel development of algorithmic techniques alongside the increasing availability of cost-effective computers with sufficient computational power to support these algorithms. As a biometric information process, a face recognition system offers wider applicability and operational range compared to other biometric methods such as fingerprinting, iris scanning, or signature recognition. The system employs a blend of methodologies across two primary domains: face detection and recognition. Face detection is executed on real-time images without a specific application context in mind. Processes involved in the system encompass white balance correction, segmentation of skin-like regions, extraction of facial features, and the extraction of face images from a face candidate. The integration of a face classification method utilizing a Feedforward Neural Network is a key component of the system. The system's performance is evaluated using a database containing 10 individuals with 20 to 30 photos of each person. The system demonstrates acceptable performance within the specified parameters, successfully recognizing faces and capable of detecting and recognizing multiple faces within live image captures.***
The primary focus of this study revolves around the implementation of the FaceNet Neural Network using a database of 10 individuals. The objective is to test the network's capacity within a specific application where multiple individuals are simultaneously recognized across different input streams. Additionally, the study explores additional features to track and identify the person of interest within the recognized group.

I. INTRODUCTION

A facial recognition system is a technology designed to identify or verify a person using a digital image or a video frame from a video source. There exist various methods employed by facial recognition systems, but generally, they function by comparing specific facial features from a provided image with faces stored within a database. It is often defined as a Biometric Artificial Intelligence application capable of uniquely identifying an individual through the analysis of patterns derived from the person's facial textures and shape.

II. BASIC CONCEPTS USED AND LITERATURE REVIEW

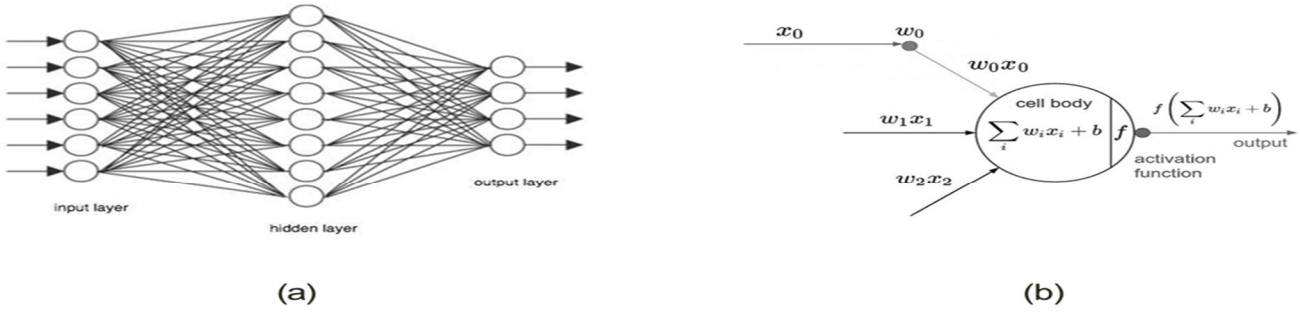


Fig 1. (a) General structure of a Neural Network, (b) an example of a neuron

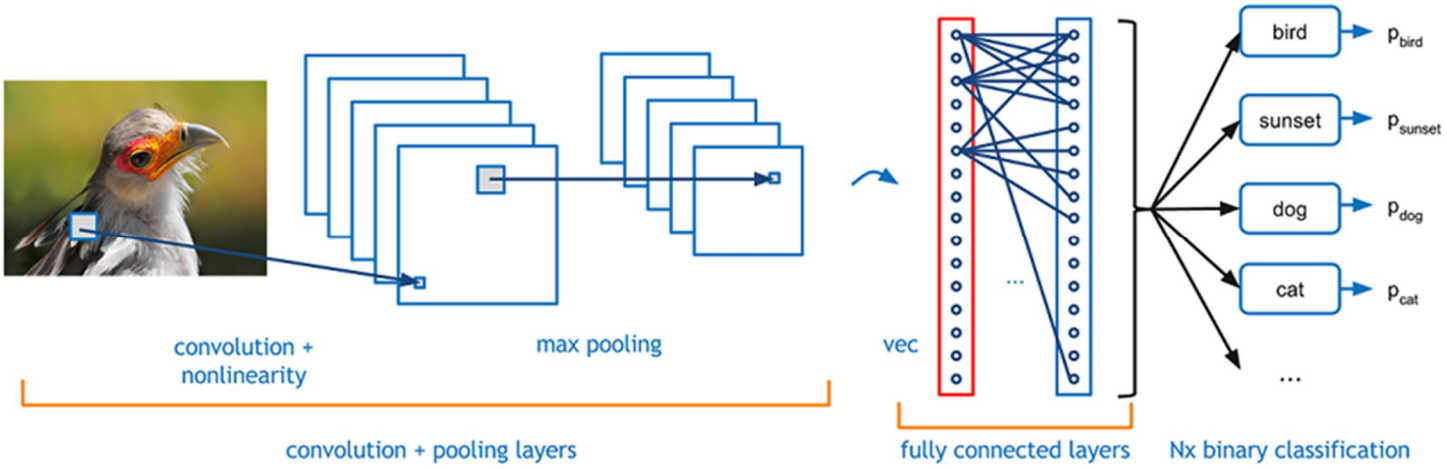
In this network, the initial column of perceptrons, termed the first layer, is responsible for making three straightforward decisions by evaluating the input evidence. However, the subsequent layer of perceptrons, referred to as the second layer, makes decisions based on the outcomes derived from the first layer. Consequently, these second-layer perceptrons are capable of making decisions at a more intricate and abstract level compared to the perceptrons in the first layer. Similarly, more complex decisions can be tackled by the perceptrons in the third layer. Consequently, a multi-layer network of perceptrons can engage in sophisticated decision-making processes.

Although I initially defined perceptrons as having a single output, it may appear in the network diagram that the perceptrons possess multiple outputs. However, each perceptron still outputs a single result. The illustration of multiple output arrows serves as a convenient way to signify that the output from a perceptron is being utilized as the input for several other perceptrons. This representation is more practical than drawing a single output line that subsequently branches off.

A. Convolutional Neural Networks

Convolutional neural networks, often appearing as a fusion of biology and mathematics, enriched with a sprinkle of computer science, have emerged as a significant innovation in the realm of computer vision. The year 2012 marked a pivotal moment when neural networks gained significant recognition, notably through Alex Krizhevsky's triumph in that year's ImageNet competition. This competition is essentially the pinnacle event in the domain of computer vision. Krizhevsky's application of neural nets substantially reduced the classification error rate from 26% to 15%, a remarkable and groundbreaking improvement at that time.

Since this milestone, numerous companies have integrated deep learning as a fundamental component of their services. Major tech entities such as Facebook utilize neural networks for their automatic tagging algorithms, Google employs them for photo searches, Amazon incorporates them into their product recommendation systems, Pinterest utilizes them for personalized home feed curation, and Instagram relies on these networks for their search infrastructure. These applications showcase the widespread adoption of convolutional neural networks in various aspects of modern technology and services.



B. General CNN Based Face Recognition Schema

Image classification involves the process of taking an input image and providing an output class (such as identifying a cat, dog, etc.) or a probability indicating the classes that best describe the image. For humans, recognition is a fundamental skill acquired from the earliest stages of life, becoming almost instinctive and effortless in adulthood. Without conscious effort, we swiftly and seamlessly identify the environment we're in and the objects surrounding us. Observing an image or the world around us, we often instinctively characterize the scene and assign labels to each object, almost subconsciously.

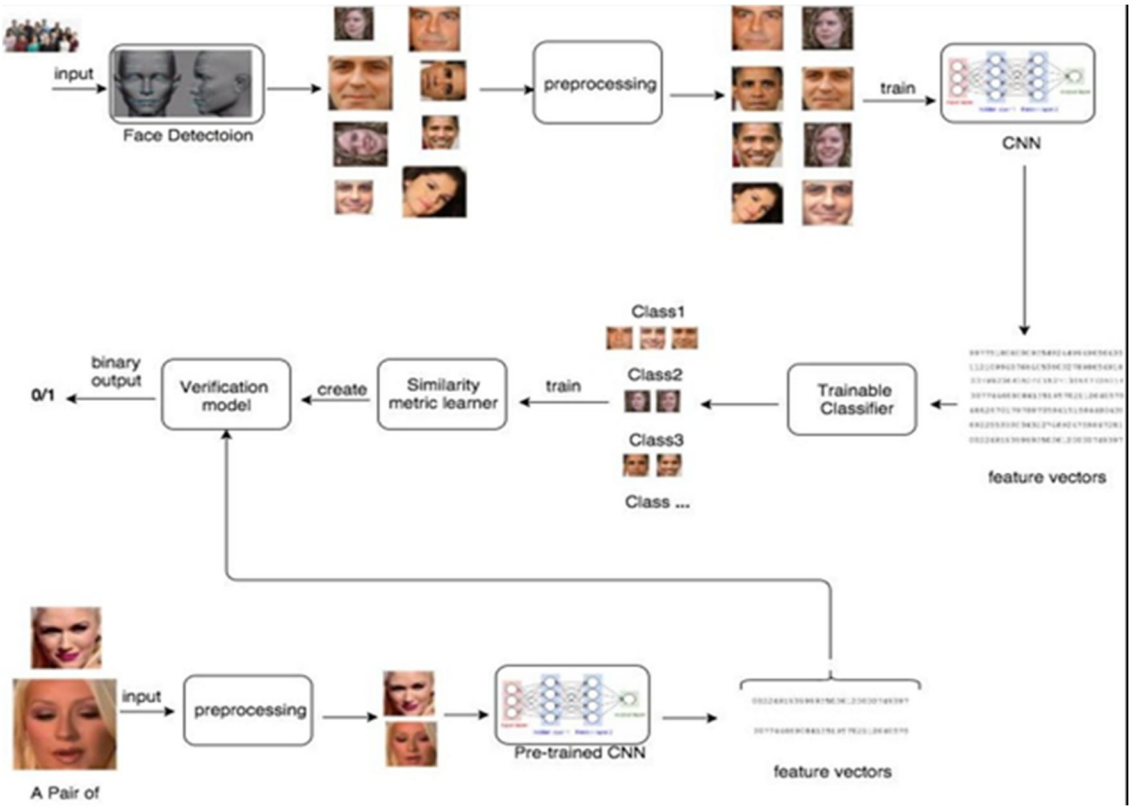
This innate ability to rapidly recognize patterns, draw from previous knowledge, and adapt to various visual environments is a skill that sets us apart from machines. It's a capacity that machines do not naturally possess and is inherent to human cognition and perception.


What We See


What Computers See

The Common Steps involved in Face recognition using convolutional neural networks are Face Detection, preprocessing, CNN training, Face Identification, Metric Learning and Face Verification.

The general face recognition pipeline



III. PROBLEM STATEMENT AND REQUIREMENT SPECIFICATIONS

A. Problem Statement

The challenge of face recognition can be framed as follows:

Face recognition involves identifying and extracting human facial characteristics, such as the mouth, nose, and eyes, from a frontal face image. To accomplish this, we will employ a multi-step approach. Initially, we will utilize a skin-color segmentation technique to detect the face region. Subsequently, morphological methods will be employed to address any gaps or irregularities that may arise during the segmentation process. The next step involves skeletonization, which yields a skeletal representation of the face, enabling the extraction of facial landmark points. These facial features are typically located within the interior of the face contour. To evaluate the effectiveness of our method, we will employ various differential images.

B. System Design

1) Data Abstraction and Preprocessing

- Face Detection: Use face detection algorithms to abstract faces from the training data.
- Data Labeling: Preprocess the detected faces and label the data for training purposes.

2) Model Training and Feature Extraction

- Model Training: Utilize the preprocessed data to train a model for face recognition.
- Feature Extraction: Extract features from the trained model to identify individuals in images.

3) Real-time Video Processing

- Webcam Image Acquisition: Capture images from the webcam at 30 frames per second.
- Image Classification: Classify and recognize individuals in the captured images.

4) Multi-stream Face Recognition

- Multithreading: Implement multithreading to enable recognition from multiple video streams simultaneously.

5) Real-life Scenario Implementation

- Search Feature: Develop a feature to search for a specific person across multiple video streams.
- Functionality: When a person's name is provided, the system will locate them in different streams.
- Output: Provide timestamp, camera ID, and the person's location within the different streams.

C. Components

1) Face Detection Module: Utilizes algorithms for face detection and extraction from images or videos.
2) Data Processing and Training: Preprocesses data, labels faces, and trains the model for feature extraction.
3) Real-time Video Capture and Processing: Captures webcam images and performs real-time face recognition.
4) Multi-stream Recognition Engine: Uses multithreading to process and recognize faces from multiple video streams simultaneously.
5) Search and Retrieval System: Accepts a person's name as input and returns their location across various streams with timestamps and camera IDs.

D. Technologies

1) Face Detection: OpenCV, Dlib, MTCNN
2) Model Training: Deep learning frameworks like TensorFlow or PyTorch
3) Real-time Video Processing: OpenCV, FFmpeg
4) Multithreading: Python's threading or multiprocessing libraries
5) Data Storage: Database to store labeled data and recognition results

E. Workflow:

1) Data Collection and Preprocessing

- Collect and label face data for training.
- Preprocess and train the model.

2) Real-time Face Recognition

- Capture and classify images from the webcam.
- Implement multithreading for processing multiple video streams simultaneously.

3) Search and Retrieval

- Enable the system to locate and display a specific person's location across streams with timestamps and camera IDs.

This design will serve as the foundation for a system capable of recognizing and locating individuals across multiple video streams in real-time.

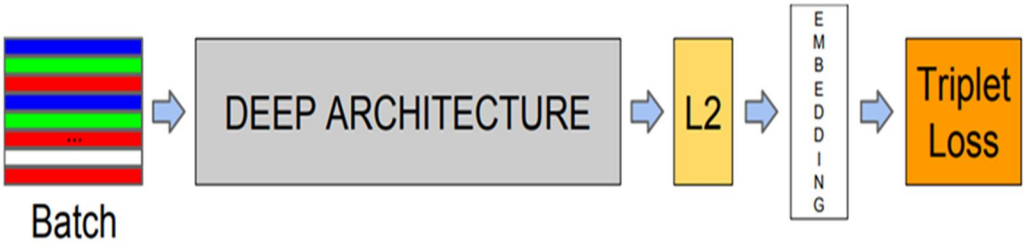
IV. IMPLEMENTATION

A. FaceNet Neural Network

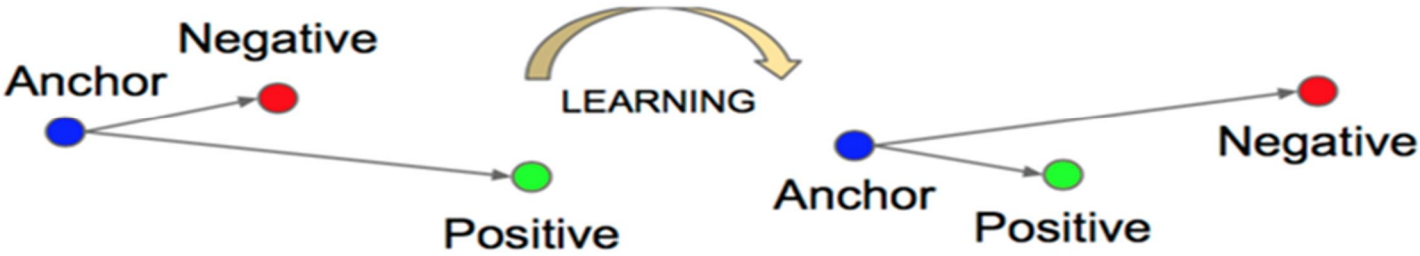
FaceNet is a system designed to create a condensed representation of facial images in a way that distances within this representation signify the similarity between faces. It accomplishes this by training a deep convolutional neural network (CNN) to generate embeddings, which directly relate to facial characteristics. Rather than relying on intermediate bottleneck layers, FaceNet optimizes the embedding space directly.

The training process involves using triplets of images: one as the reference, another of the same individual , and a third of a different person . The advantage lies in the efficient representation of faces, enabling tasks like recognition, verification, and clustering using these embeddings as features. Notably, FaceNet achieves exceptional performance, reaching a record accuracy of 99.63% on LFW and 95.12% on Youtube Faces DB while requiring only 128 bytes per face.

B. Triplet Loss



The triplet-based loss function integrated into FaceNet's learning mechanism is a modified version inspired by the LMNN classifier. This adaptation aims to bring similar face images closer while simultaneously pushing apart images of different individuals. Such an approach significantly aids the training process of deep neural networks. Sun et al. implemented an ensemble of networks trained using a blend of classification and verification loss. Their verification loss aligns with FaceNet's triplet loss by minimizing squared L2 distances between images of the same person and imposing a boundary between images of distinct individuals. However, their method contrasts as it evaluates pairs of images, diverging from FaceNet's encouragement of triplet-based comparisons. Similarly, Wang et al. adopted a loss function similar to FaceNet's triplet loss to rank images based on both semantic and visual similarity. Their emphasis on handling relative distance constraints in the learning process closely mirrors FaceNet's approach.



Our objective is to guarantee that an image x_i_a of a particular individual is closer to all other images x_i_p of the same person than it is to any image x_i_n of any other person by a margin ε_i. In other words, we want to enforce the following condition:
$d(x\_i\_a, x\_i\_p) + \varepsilon_i < d(x\_i\_a, x\_i\_n)$ for all i
Therefore, the loss function (L) can be defined as:
$L = \max(0, d(x\_i\_a, x\_i\_p) + \varepsilon_i - d(x\_i\_a, x\_i\_n))$
This loss function encourages the embeddings of images from the same person (x_i_a and x_i_p) to be closer to each other by at least a margin ε_i compared to any image from a different person (x_i_n). If this condition is met, the loss is zero, indicating that the embeddings are properly separated. If the condition is violated, the loss becomes a positive value, indicating that the embeddings need to be adjusted to meet the desired margin ε_i.

$$L = \sum_{i}^{N} \left[\left\| f(x_{q,i}) - f(x_{p,i}) \right\|_2^2 - \left\| f(x_{q,i}) - f(x_{n,i}) \right\|_2^2 + \alpha \right]_+$$

Of all possible triplets (N of them), many would easily satisfy the above constraint. So it'd be a waste to look at these during training (wouldn't contribute to adjusting parameters, would only slow down convergence); it's therefore important to select "hard" triplets (which would contribute to improving the model) to use in training. How do we do that?
we could select (xp,i) and (xn,i) :
(1) hard positive- argmax (xp,i) $\|f(xq,i) - f(xp,i)\|$ 2 2
(2) hard negative- argmin (xp,i) $\|f(xq,i) - f(xn,i)\|$ 2 2

C. Triple Selection

The original concept suggests selecting the "hardest" positive and negative images concerning an anchor image within a dataset to form triplets. The hardest positive and the hardest negative (closest in the dataset) are chosen. However, due to computational infeasibility in identifying these images across the entire dataset and potential training issues due to mislabeled or poorly imaged faces dominating these selections, an alternative method is proposed.

To tackle this challenge, the generation of triplets dynamically takes place within the training mini-batches, not across the complete dataset. In this methodology, rather than considering the entire dataset for each anchor image to identify the "hardest" positive, the approach involves using the pool of positive images present within the batch. Unlike selecting a single hardest positive for an anchor, this method utilizes all anchor-positive pairs within the batch. Meanwhile, it still selects hard negatives, ensuring one hard negative is picked corresponding to each anchor image.

The batch structure is designed to include approximately 40 images per identity for each mini-batch, ensuring a meaningful representation of the anchor-positive distances. For negative faces, they are randomly sampled for each mini-batch.This strategy of dynamically generating triplets within mini-batches, using all anchor-positive pairs in a batch while selecting hard negatives, has been found to yield a more stable and faster-converging solution compared to exclusively choosing the hardest positive for each anchor. This adaptation mitigates the challenges posed by mislabeled or poorly imaged faces dominating the selection of hard positives and negatives, contributing to more effective training.

D. Overview

FaceNet, treating the CNN architecture as a Blackbox, emphasizes the crucial aspect of end-to-end learning in its system. It focuses on generating an embedding, f(x), which maps face images to a feature space $\mathbb{R}d$. Within this space, the objective is to minimize the squared L2 distance between face images of the same individual, even when captured under different imaging conditions, resulting in their embeddings being closely positioned. Simultaneously, FaceNet aims to maximize the distance between pairs of face images from different individuals, ensuring distinct separation.

Traditional loss functions typically aggregate all images of the same identity into a single point in $\mathbb{R}d$. However, FaceNet goes beyond this approach. The triplet loss utilized by FaceNet not only brings together faces of the same person (anchor and positive) but also enforces a margin between each pair of faces from one individual and those of all other identities. This margin significantly separates the embeddings, enhancing the model's ability to discern between different individuals.

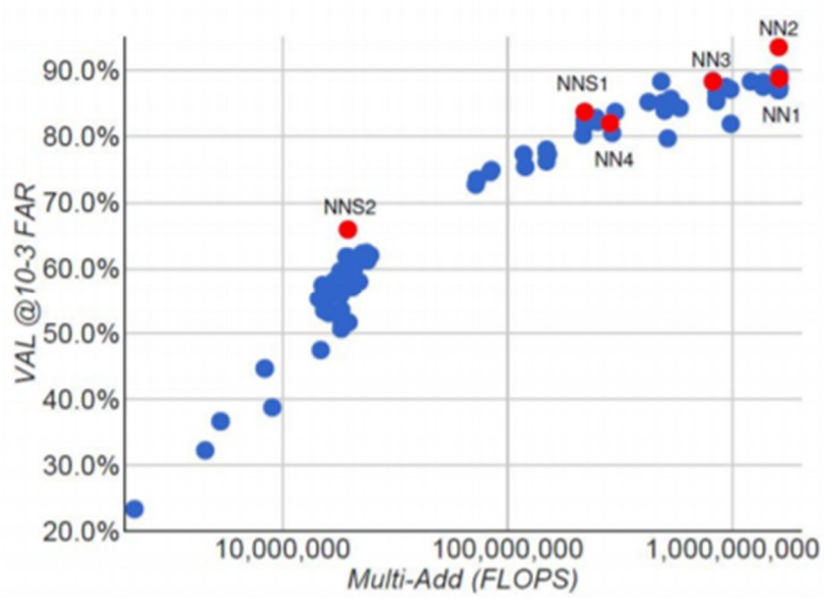
E. Experiments- Computation vs Accuracy Tradeoffs

1) *Dataset:* Comprised 100-200 million face thumbnails from 8 million unique identities.
2) *Pre-processing:* Detected faces, generated bounding boxes, and resized images to varying dimensions (96x96 to 224x224) for network inputs.

F. Trade-offs Observed

1) *Accuracy vs. FLOPS:* Strong correlation noted; higher computational demands (FLOPS) tend to correspond with improved accuracy in facial recognition tasks.
2) *Accuracy vs. Number of Parameters:* No significant correlation found; the number of model parameters did not consistently determine performance.

architecture	VAL
NN1 (Zeiler&Fergus 220×220)	87.9% ± 1.9
NN2 (Inception 224×224)	89.4% ± 1.6
NN3 (Inception 160×160)	88.3% ± 1.7
NN4 (Inception 96×96)	82.0% ± 2.3
NNS1 (mini Inception 165×165)	82.4% ± 2.4
NNS2 (tiny Inception 140×116)	51.9% ± 2.9



G. Impact of CNN Models

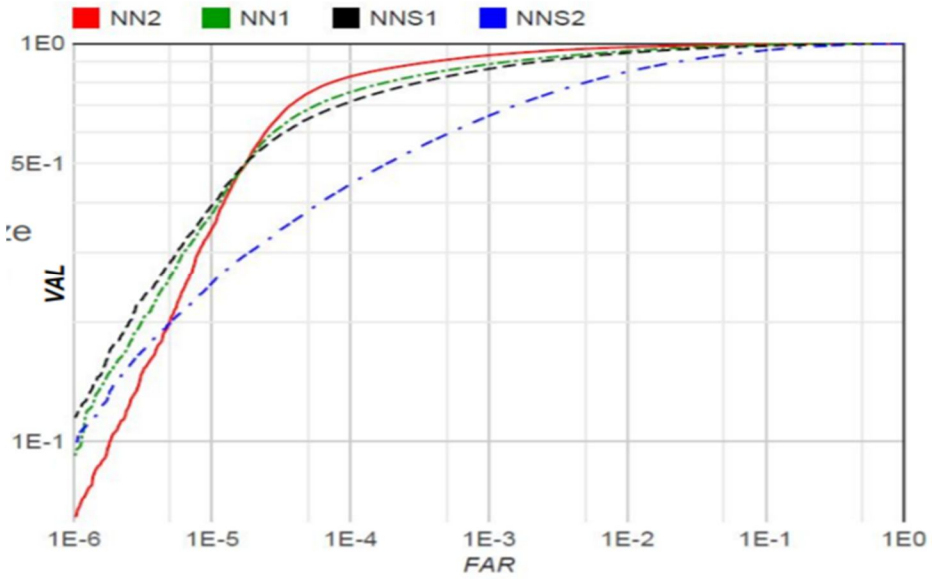
1) Zeiler&Fergus (NN1) vs. Inception (NN2): Despite a 20-fold difference in the number of parameters between the Zeiler&Fergus-based (NN1) and the GoogLeNet-based (NN2), both achieve similar performance levels in facial recognition tasks.

2) NNS2: A scaled-down version of NN2, denoted as NNS2, designed with an input size of 140x116, offers the advantage of running on mobile devices with a processing speed of 30 milliseconds per image. Despite its smaller size and faster processing, it maintains a good enough accuracy for face recognition tasks, achieving a validation accuracy (VAL) of 51.9%.

H. Summary

1) NN1 and NN2 demonstrate comparable performance despite a substantial difference in parameter count.

2) NNS2, a reduced-size variant of NN2, optimized for mobile devices, delivers acceptable face recognition accuracy with a processing speed of 30ms per image and a validation accuracy of 51.9%.



V. RESULTS AND ANALYSIS

A. *Performance on LFW*

The FaceNet system utilizes an optimal threshold of 1.242 for calculating the L2 distance.

To prepare the input data, two distinct pre-processing methods are applied:

1) The first method involves a fixed center crop of the provided LFW thumbnails.

2) The second method utilizes a proprietary face detector. In cases where this method doesn't provide alignment, LFW alignment is applied.

When using method "a," the system achieves an accuracy rate of 98.87%. On the other hand, method "b" achieves an exceptional accuracy of 99.63%, demonstrating a state-of-the-art performance.

False Accept

False Reject

B. *Performance on YouTube Faces DB*

1) Utilizing their proprietary face detector, the method involves assessing the average similarity among all face pairs detected within the initial 100 frames.

2) The achieved classification accuracy is 95.12%, representing cutting-edge performance.

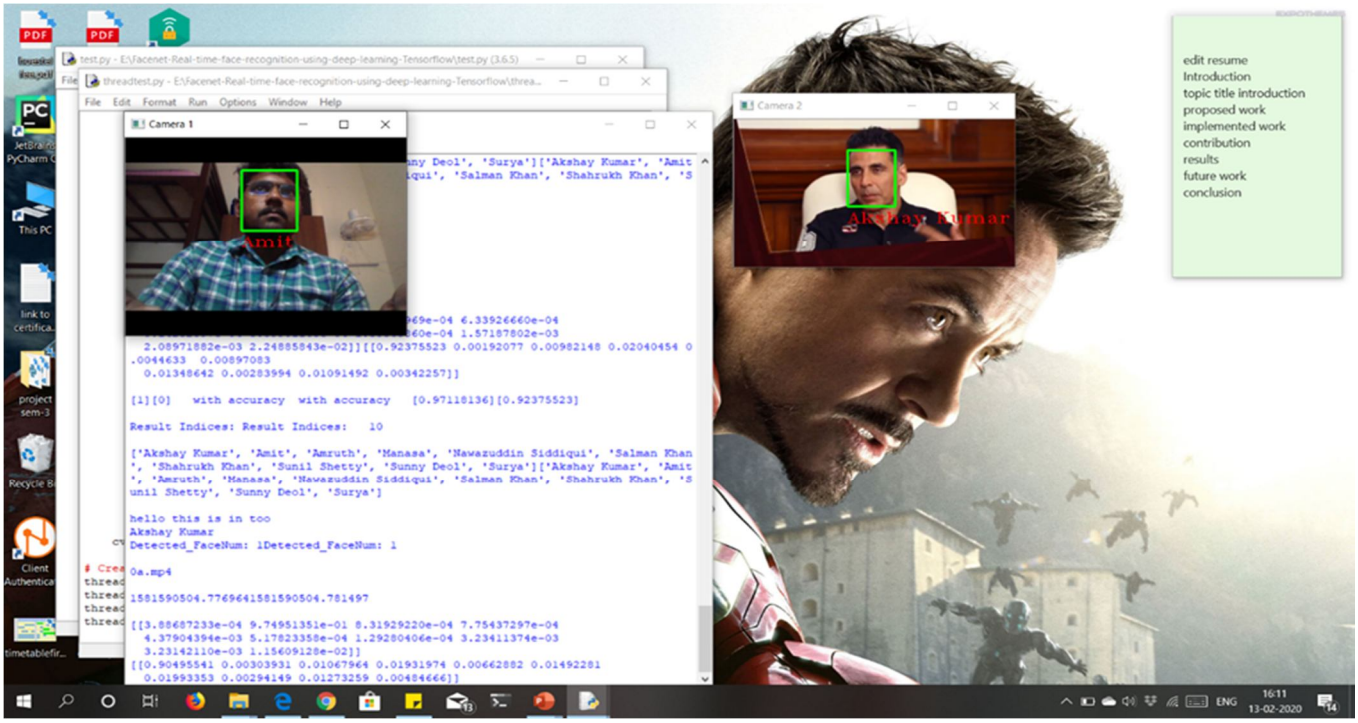
3) Extending the analysis to the first 1000 frames yields a consistent accuracy of 95.18%, showing no marked improvement compared to the smaller subset.

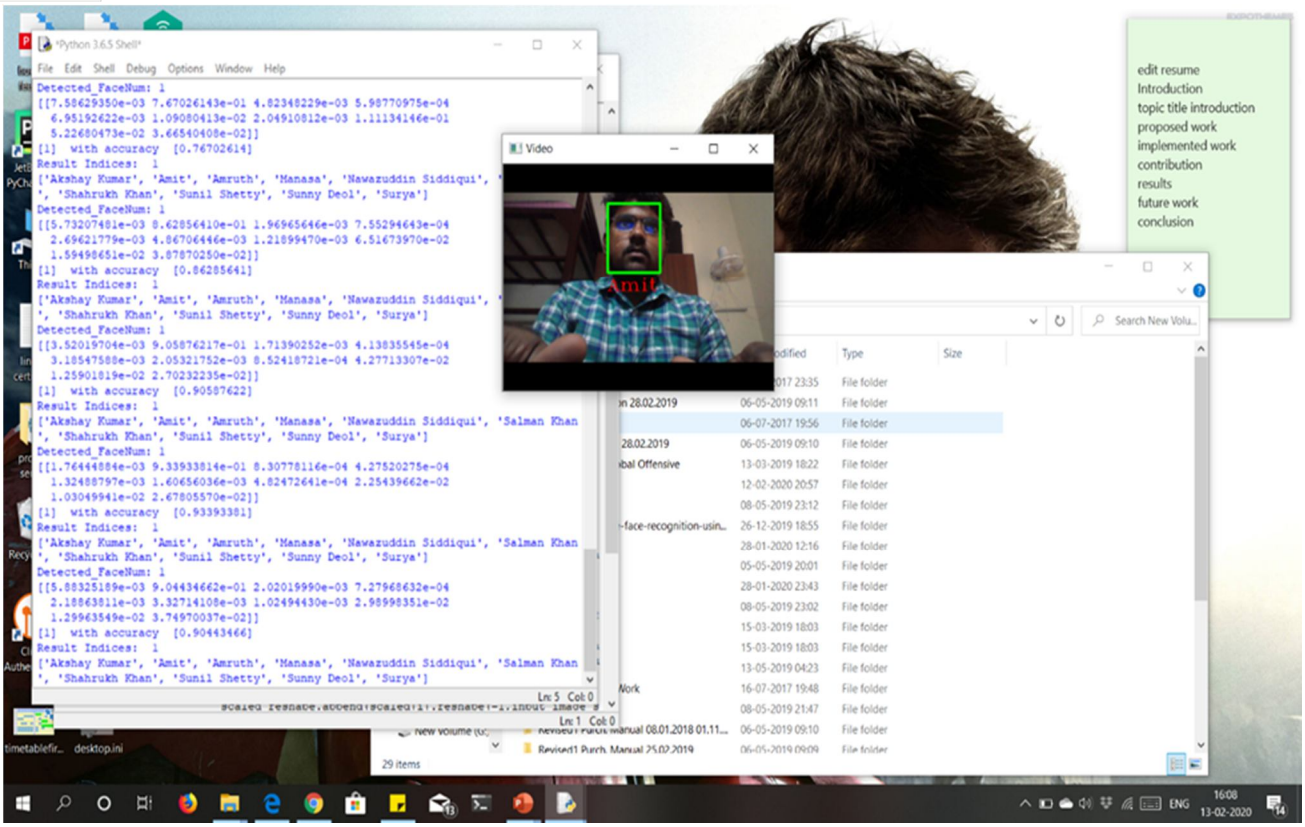
4) In contrast, previous work by DeepId2+ (Sun et al.) attained an accuracy of 93.2%.

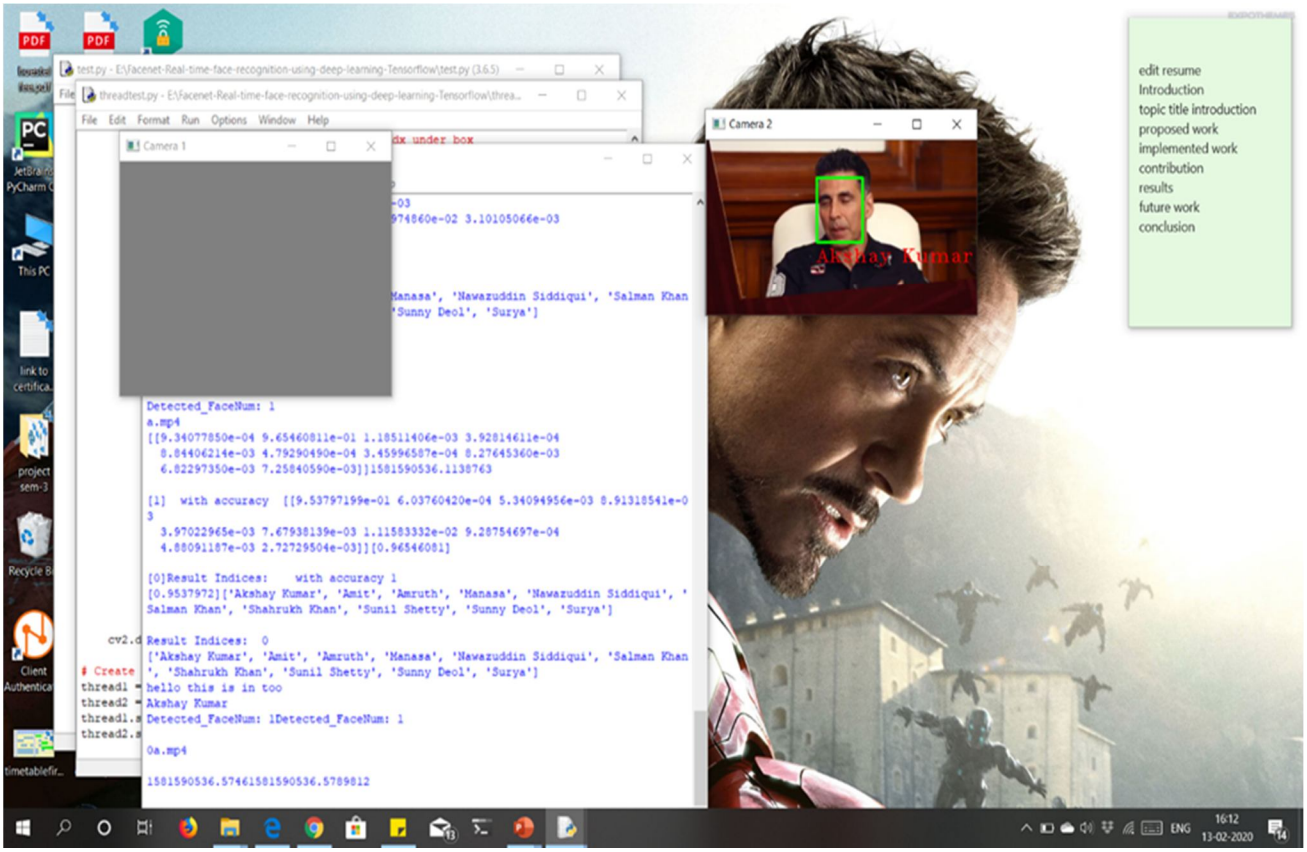
C. *Performance on the Personal DB*

1) This is recognizing the people in the videos/live stream with an accuracy of almost 95%

2) It can recognize people in different videos/live streams at the same time




edit resume
Introduction
topic title introduction
proposed work
implemented work
contribution
results
future work
conclusion
Video
Amit


edit resume
Introduction
topic title introduction
proposed work
implemented work
contribution
results
future work
conclusion
Camera 2
Akshay Kumar
Camera 1

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

VI. STANDARDS ADOPTED

The system design for face recognition and multi-stream video processing adheres to various industry standards and best practices to ensure robustness, security, and compliance. The design's approach towards face detection and recognition aligns with ISO/IEC 19794-5 and ISO/IEC 30107-3, providing standardization for biometric data interchange formats and performance testing in face recognition systems. Data labeling and preprocessing follow ISO/IEC 25000 series guidelines for software quality, complemented by GDPR compliance to ensure data privacy and protection. Model training and feature extraction methodologies are based on best practices from TensorFlow or PyTorch, emphasizing model evaluation through metrics such as precision, recall, and F1-score. Real-time video processing adheres to standards like MPEG-4 or H.264 for video compression and ITU-R BT.500 for video quality assessment. Concurrency standards and system integration practices support multi-stream recognition, while security measures follow ISO/IEC 27001 guidelines.

The system's workflow complies with ISO/IEC 12207 for software lifecycle processes and ISO 9001 standards for quality management, ensuring a comprehensive and standardized approach throughout development. Legal and ethical compliance, such as adherence to GDPR and other regional laws, is incorporated, ensuring the system's alignment with privacy and data handling regulations. Overall, these standards collectively form a robust foundation for the system's functionality, reliability, and ethical operation in recognizing and locating individuals across multiple video streams in real-time.

VII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

Over the past two decades, face recognition technology has made significant advancements. Presently, machines possess the capability to automatically authenticate identity details for secure transactions, surveillance, security measures, and building access control, among other applications. These applications typically function in controlled settings, leveraging environmental constraints to achieve high accuracy in recognition algorithms.

However, the next phase of face recognition systems is expected to expand into smart environments, wherein computers and machines operate as more supportive aides. For this to become a reality, computers must proficiently identify individuals nearby in a manner that seamlessly integrates into typical human interactions without necessitating special prompts. They should align with human expectations about when recognition is probable, indicating the need for future smart environments to utilize the same modalities as humans and have similar limitations.

Although these objectives seem attainable, considerable research is still necessary to ensure the reliability of person recognition technology across vastly different conditions using information from single or multiple modalities. The focus lies on making face recognition technology work consistently in a variety of situations to fulfill the potential of future smart environments.

B. Future Scope

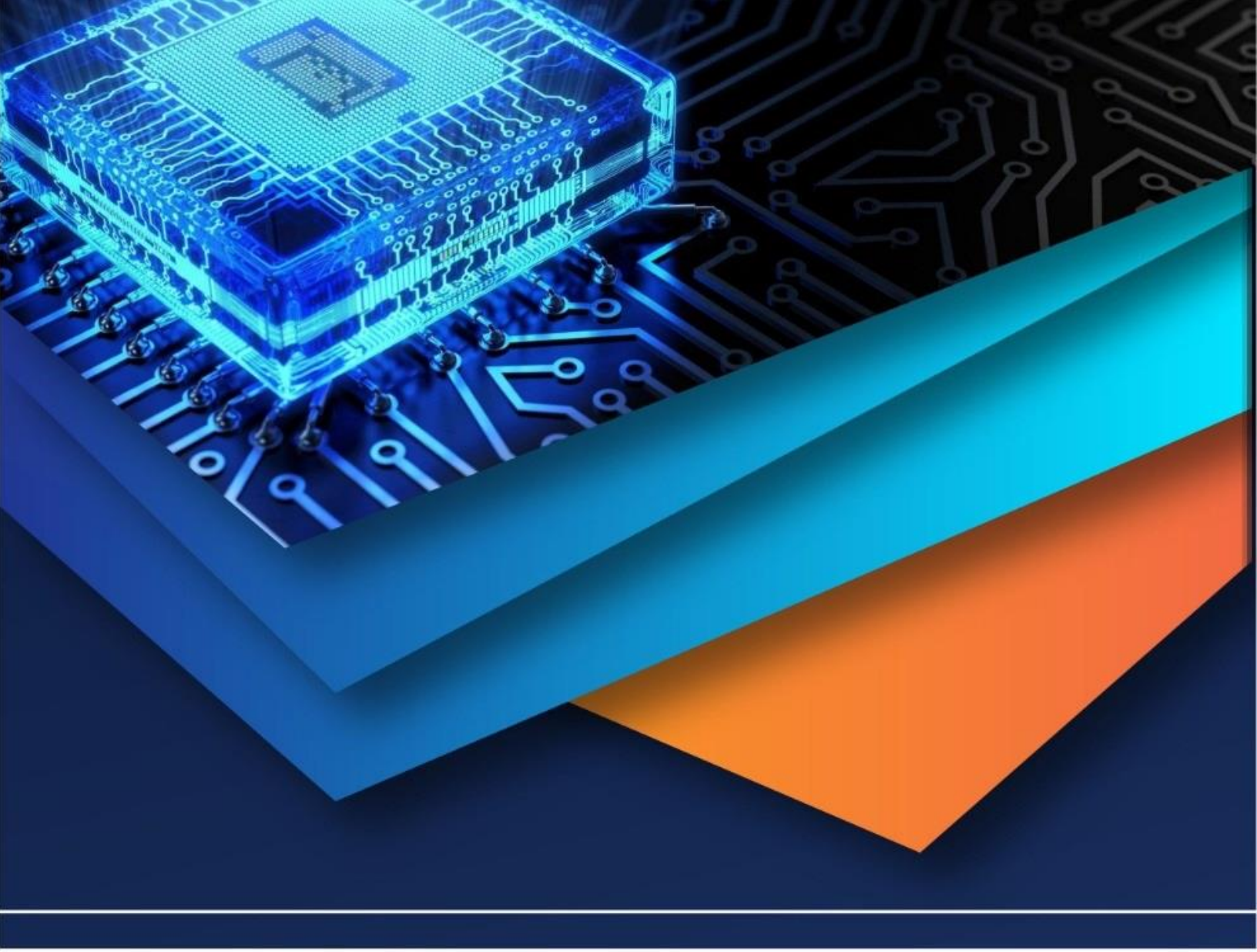
Today, facial recognition technology is significantly utilized in the realm of security. It serves as an effective tool for law enforcement agencies to identify criminals, while software companies leverage this technology for user access. There's extensive potential for further development, enabling its utilization in various other areas such as ATMs, accessing confidential files, or other sensitive materials. This could potentially render conventional security measures like passwords and keys obsolete.

Innovators are exploring the integration of facial recognition in public transportation, particularly within subways and similar outlets. The objective is to use faces as a form of credit card for paying transportation fees. Rather than purchasing tickets at a booth, facial recognition technology would capture the individual's face, process it through a system, and debit the predetermined account. This innovation has the potential to significantly streamline processes and optimize traffic flow, offering a glimpse into a futuristic approach.

The amalgamation of personal data has granted marketers and advertisers an unprecedented opportunity to refine their approaches to target markets. Facial recognition could similarly allow companies to discern specific demographics. For instance, upon recognition of a male aged between 12 and 21, a display screen might present an advertisement for the latest FIFA game. Major retailer Tesco plans to implement Optimizes screens at 450 petrol stations in the UK, tailoring targeted ads to customers. According to company CEO Simon Sugar, this technology could revolutionize British retail. However, the accuracy of these systems in identifying customers is paramount, as being misclassified in terms of age or gender could lead to significant misunderstandings, comparable to the frustration of having one's name misspelled on a Starbucks cup.

REFERENCES

[1] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical- view faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2
[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 8
[3] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. 2, 4, 6
[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014.2,4,5,6,9
[5] K.Q. Weinberger, J.Blitzer,and L.K.Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press, 2006. 2, 3
[6] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8 [7] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. CoRR, abs/1404.4661, 2014. 2



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY