VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY
UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

FINAL GROUP PROJECT

A REAL-TIME FACIAL RECOGNITION MODEL BASED ON DEEP LEARNING

MAJOR: DATA SCIENCE

Team members: Phạm Hải Nam (BI12-307)
Phùng Đức Thái (BI12-396)
Hứa Hải Minh (BI12-272)
Lê Mậu Minh Phúc (BI12-351)
Phạm Minh Hoàng (BI12-175)

Supervisor: Dr. Đoàn Nhật Quang

Hanoi, January 2024

ABSTRACT

A real time attendance check model seems to be crucial in technological advancement due to numerous benefits. Face detection, and recognition processes employing the YuNet model, ResNet34 and SVM-based classification are used during this study to propose a new model. YuNet demonstrates outstanding real-time performance at 75-85 FPS for face detection, achieving recognition of distinguishable faces with an Easy AP (Average Precision) of 0.8843. The model is also extended the accuracy to medium-difficulty and hard-to-detect faces, equivalent to MediumAP and HardAP. A real-time model combining YuNet, alignment, and ResNet34 with two SVM (Support Vector Machine) models, SVC (C-Support Vector Classification) and NuSVC (Nu-Support Vector Classification), operates at about 30 FPS. For classification algorithm selection, NuSVC outperforms the SVC in terms of processing speed while maintaining the high accuracy rate, which is important for recognising faces in the large scale. This studied model proves the fact that it can be used in real-time scenarios with high accuracy. Some future work, which can be proposed to improve the attendance check model using facial recognition, include larger identity databases testing in diverse environments, optimizing GPU deployment for latency reduction, and incorporating anti-spoofing methods to enhance security and prevent unauthorized access. These future orientations aim to improve practicality, versatility and overall performance in real-world scenarios.

Keywords: *Face detection, Face recognition, Face alignment, Yunet, ResNet34*

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLE

LIST OF ABBREVIATIONS

FPN	Feature pyramid network
TFPN	Tiny feature pyramid network
IoU	Intersection over Union
AP	Average Precision
SVM	Support Vector Machine
SVC	C-Support Vector Classification
NuSVC	Nu-Support Vector Classification
NMS	Non maximum suppression
CNN	Convolutional Neural Network

1. Introduction

In an era where advancements in artificial intelligence are reshaping daily lives, the ability to identify and individuals through facial features has emerged as a key innovation. In this introduction, the motivation of the group of researchers and a short history of face recognition are provided to inspire the reasons for this study.

1.1 Motivation

Along with the rapid development of computer vision, face recognition applications have gained significant attention. Deep Convolutional Neural Networks have achieved great success for face recognition technology, which has been widely used in various fields, including access control, attendance tracking, security systems, and mobile device authentication. Face recognition involves the identification or verification of individuals based on the facial features. In situations where hygiene and minimizing physical contact are essential, such as in the covid-19 pandemic, face recognition offers a contactless solution, unlike fingerprint or hand-based identification methods which require the users touch the machine, increasing the possibility of infection. Vision-based tasks such as Face Detection (FD), Face Recognition (FR), and Face Verification (FV) are commonly used for authentication purposes. Moreover, some advanced face recognition systems can process identification quickly, making them suitable for scenarios that require fast response such as access control at events, airports, or secure facilities. Systems designed for multi-face identification can accommodate a large number of users, which is very useful in surveillance. (Xudong Sun, 2018)

Attendance recording has become a basic requirement for almost every enterprise. However, modern attendance systems still have some drawbacks. For instance, since the current fingerprint attendance systems are sensitive, identifying a person in the case of a cut, wound or when fingerprints are smudged with dirt might be very challenging. Another example is using card attendance systems. This approach cannot handle the case of employees swiping cards for other people. Other traditional attendance methods, such as manually checking in with pen and paper or attendance check by reading names are very time-consuming and might cause delays, especially in crowded events with a large number of participants. With contactless interaction and the ability to perform in real-time, checking attendance by using facial recognition seems to be the best solution currently. (Truein, 2023)

Due to the strong development of big data and machine learning, many deep learning-based face recognition methods have been developed. Studies show that deep learning-based face recognition methods have a high level of accuracy and stability in identifying and verifying individuals based on facial features. The advantage of deep learning is the ability to automatically learn facial landmarks from large datasets. Face recognition systems applying deep learning can adapt to variations in lighting, pose, and facial expressions, making it suitable for real-time applications. (Xiao Han, 2018)

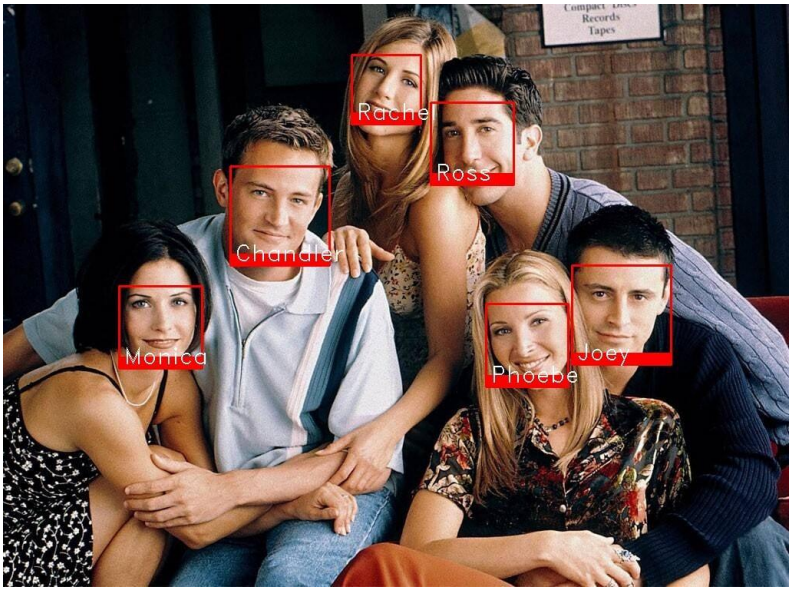


Figure 1.1: Facial Recognition

In the context of this project, a deep learning based face recognition system for attendance check is proposed and implemented. The framework of the proposed method comprises 4 main components: face detection, face alignment, feature extraction and classification, as illustrated in figure 2. In the framework, pre-trained YuNet and ResNet34 are employed. YuNet is used for the face detection stage, while ResNet34, a convolutional neural network, is utilized for the feature extraction process. For the classification stage, the recognition model deploys support vector machine (SVM).

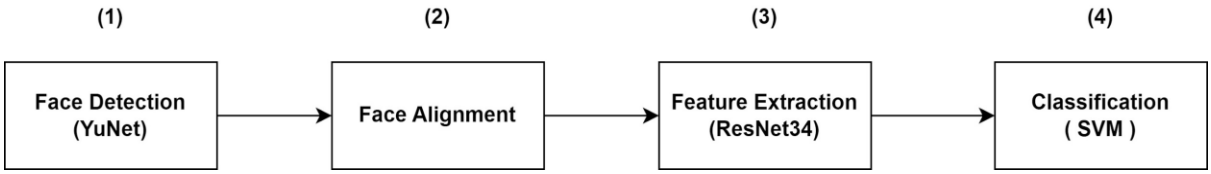


Figure 1.2: The overall framework of the proposed face recognition model

1.2 History of face recognition technology

The history of facial recognition technology started in the early 1960s, with Woody Bledsoe, Helen Chan Wolf, and Charles Bisson laying the foundation. Funded by an unnamed intelligence agency, the work focused on using computers to recognize human faces. Despite limited publication due to the sensitive nature of the project, the breakthrough efforts involved manual marking of facial landmarks and mathematical rotation for pose variation. In the 1970s, Goldstein, Harmon, and Lesk expanded this work, introducing 21 specific subjective markers including hair color and lip thickness in order to automate the recognition, but also with continued manual computations. The late 1980s witnessed a significant step when Sirovich and Kirby applied linear algebra, giving birth to Eigenface. This breakthrough showed that basic

facial features could be derived from a collection of images, requiring fewer than one hundred values for accurate coding. The work of Turk and Pentland in 1991 developed automatic facial recognition by detecting faces within images. Despite facing technological challenges, these crucial developments opened the way for the evolution of facial recognition technology nowadays. (Karl Maria Michael de Leeuw, 2007)

2. Literature review

In the era of technology, face detection and face recognition have emerged as key technologies, which greatly affect areas like security, personal devices, social media and so on. Face detection serves as the foundational step, where the system identifies and locates human faces within an image or video frame. Once faces are detected, the subsequent phase of face recognition begins. This advanced step involves analyzing the unique features of each detected face, such as the shape of the eyes, nose, and mouth, and comparing these characteristics against the face features of known faces. Through this comparison, the system can identify or verify the identity of individuals in the image. Thus, face detection paves the way for face recognition, which plays an important role in creating a comprehensive system for facial analysis.

2.1 Face detection

One of the most popular face detection algorithms is MT-CNN (Multi-task Cascaded Convolutional Networks) that has the ability to handle faces at different scales and angles, also includes three cascaded stages:

- Stage 1 – Proposal Network (P-Net): Generates bounding boxes and objectness (face/ non-face scores).
- Stage 2 – Refine Network (R-Net): Refines the candidate boxes by filtering out low-confidence detections and regresses the bounding box coordinates.
- Stage 3 – Output Network (O-Net): Further refines the results, performs facial landmark detection, and provides final bounding box predictions. (Qi, Zuo, Feng, & Naveen, 2022)

Moreover, YuNet is a lightweight face detection model designed for real-time applications to achieve fast inference speeds. The model is optimized for efficiency while maintaining good accuracy in detecting faces with a YuNet simple architecture developed by Shiqi Yu in 2018 and subsequently open-sourced in 2019. (Wei Wu, 2023)

Figure 2.1: YuNet detects 619 faces out of 1000 faces report (Wei Wu, 2023)

2.2 Face recognition

Two famous approaches include FaceNet and ResNet. FaceNet is a deep learning model designed for face recognition. It uses a deep convolutional neural network (CNN) architecture to extract high-level features from facial images. One of the key aspects of FaceNet is the ability to generate a compact numerical representation, often called an "embedding" or "embedding vector," for each face. These embeddings are used to measure the similarity between faces. Moreover, FaceNet employs a triplet loss during training, which ensures that the distance between the embeddings of images belonging to the same person is minimized, while the distance between embeddings of different people is maximized. (Qi, Zuo, Feng, & Naveen, 2022)

Otherwise, ResNet (Residual Network) is a deep neural network architecture that introduced the concept of residual learning. It allows the training of very deep networks by mitigating the vanishing gradient problem. In the context of face recognition, a pre-trained ResNet model can be used to extract features from facial images. The last layer or layers of the network can be fine-tuned for face recognition tasks (Saining Xie, 2017). ResNet34 is a state-of-the-art image classification model structured as a 34-layer convolutional neural network with the difference compared with traditional neural networks according to the residuals from each layer and the subsequent connected layers. (Huiyan Wu, 2019)

2.3 Other application of Face detection and recognition system

Florian Schroff, Dmitry Kalenichenko and James Philbin proposed a method called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. The method achieved state of the art performance with accuracy of 99.63% on the Labeled Faces in the Wild (LFW) dataset, and 95.12% on YouTube Faces DB (Florian Schroff, 2015).

On the other hand, to deal with the hard or noisy faces, in 2022, scientists introduced an Additive Angular Margin Loss function named ArcFace which can maximize class separability. Experiments showed that ArcFace can enhance the discriminative feature embedding as well as strengthen the generative face synthesis. Face recognition models using ArcFace loss functions achieved verification accuracy results of 93.3% on the LFW benchmark, outperforming other models with Softmax, SphereFace, CosFace, ArcFace (Jiankang Deng, 2015).

According to Edwin Jose, Greeshma M., Mithun Haridas T. P. and Supriya M. H. successfully built a face recognition model for a Surveillance System using MTCNN for face detection and FaceNet for face recognition. The model can perform efficiently in real-time while maintaining the impressive accuracy of 97% (Edwin Jose, 2019).

Finally, Ahmad S. Lateef and Mohammed Y. Kamil proposed various techniques in 2023 under the name of attendance registration system based on YOLOv7 with specific evaluation metrics; the dataset was collected to train about more than 2000 images from students as exemplifed in table 2.1. (Ahmad S. Lateef, 2023)

Table 2.1: Example of evaluation metrics used for technique evaluation

Techniques	Speed	Accuracy	Recall	Precision	F1-Score	No. of examined face	Configuration
YOLOv7	5 FPS	100%	100%	100%	100%	32	i7-7500U CPU

According to literature review, particular requirements are generated to evaluate a recognition model including accuracy, recall, precision, F1-Score, Number of examined faces and the speed.

2.4 Structure of an experiment

Face recognition is the task of identification and verification in a photo or video. This process has a number of main steps, including face detection, face alignment, feature extraction, and face recognition.

- Image Preprocessing: Resize and normalize images to a consistent format.
- Face Detection: Find the faces' locations in the frame and draw bounding boxes around the faces.
- Face Alignment: Align detected faces to a standardized pose to reduce variations caused by head tilts and rotations.
- Feature Extraction: Extract facial features from the aligned faces, transforming the input face images into high-dimensional feature vectors.
- Face Recognition: Compare the feature vectors to determine the face's identity.

In this study, two approaches have been used to evaluate according to evaluation metrics. Face detection is performed by YuNet, while ResNet34 is in the process of face recognition.

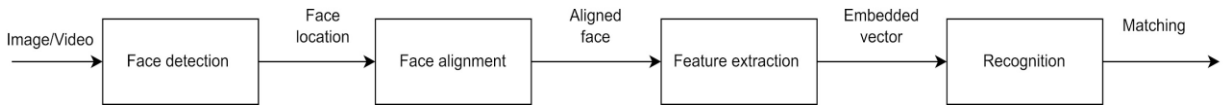


Figure 2.2: A basic pipeline of face recognition model

2.5 Objectives

The aims of this project include:

- To propose and implement a face recognition framework applied deep learning-based methods.
- Evaluate the recognition model by testing on different datasets.
- To demonstrate the model's accuracy, efficiency, and real-time recording of working in real-time.

2.6 Project's experimental design

Before starting the experiment, four datasets are collected. Three datasets, which are available on the internet, are WIDER FACE, Glint360k, AgeDB, the other dataset is collected manually by the team members. In order to achieve the objectives, the study is conducted according to the following steps:

- The YuNet face detector is chosen for the face detection task in this experiment. The model's architecture is inspired by the principles for efficient small models, consisting of three main components: a backbone, a feature pyramid, and a head. The deployed model was pre-trained with WIDER FACE, a dataset of

human images variant in posing, lighting and occlusion. YuNet face detector is tested on WIDER FACE dataset for accuracy evaluation with different confidence thresholds ranging from 0.3 to 0.8. The process's results are the average IoUs for each WIDER FACE difficulty level (Easy/Medium/Hard) and the model run-time (images/second). Additionally, the detection model is tested by webcam streaming to obtain the FPS rate with the confidence threshold of 0.9.

+ After the face detection process, the faces are aligned to a standardized pose before passing into the feature extraction network.
+ In the feature extraction process, only the face's features are extracted. Therefore, this framework utilizes the ResNet34 convolutional neural network pre-trained with the Glint360K dataset, which consists of only human faces' images.
+ For identity classification, two models of support vector machine method (SVM) are used, which are SVC and Nu-SVC. Each classification model is trained with the AgeDB images and the manual collection images, in order to serve the further testing and evaluating.
+ For real-time performance evaluation, the entire face recognition framework (YuNet, Face alignment, ResNet34, SVM) is tested on the manual collection dataset to obtain the model's speed and scalability, which are represented in the frame per second (FPS) and the minimum pixels of faces' images that can be detected. The entire model is also evaluated based on accuracy, recall, precision and F1 score, by testing on AgeDB dataset. In the end, evaluation results are given according to different multi class classification methods (SVC and Nu-SVC).

3. Methodology

Facial images are captured using a webcam featuring a video resolution of 2560 x 1440, a 4 MP camera, a field of view of 106 degrees, and a frame rate of 1080p/30 FPS. Subsequently, these images are passed on to an image preprocessing part, during which the system reads frames and preprocesses to ensure the size is appropriate (below 640x640).

Following the preprocess, the images are input into the face detection model YuNet, which identifies 5 facial landmarks and establishes bounding boxes. Next, the face alignment process uses bounding boxes, left and right eye landmarks from the face detection YuNet to rotate the faces into standardized poses. Furthermore, the aligned faces are then resized to a standardized 224 x 224 size. Subsequently, the resized images undergo analysis using a face recognition model, ResNet34, which extracts features from the cropped face images and encodes them into embedded vectors.

Simultaneously, the embedded vectors from real-time images are then compared to embedded vectors saved in SVM model. If the comparison shows a difference in the threshold, the system will mark the result as " Unrecognized". On the other hand, if the thresholds match, recognized faces display the bounding box, probability and personal name.

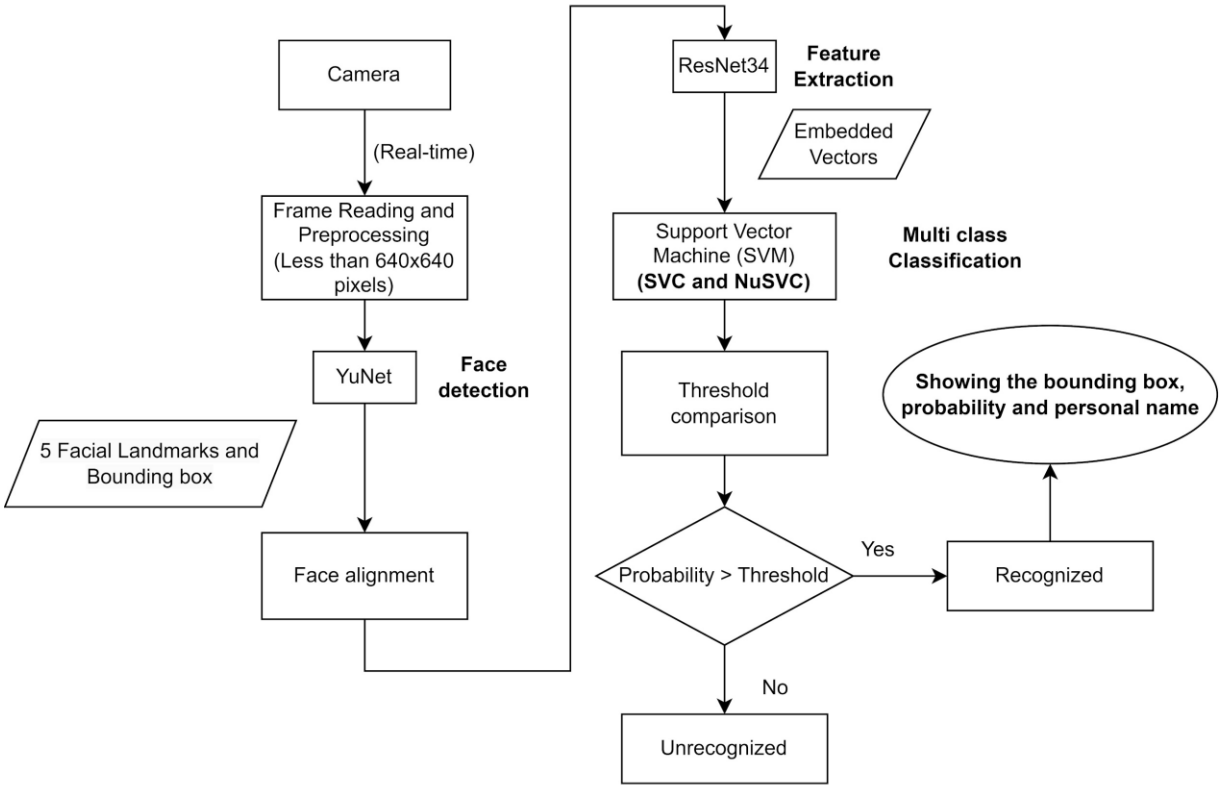


Figure 3.1: Overall Pipeline of face recognition model

3.1 Face detection

YuNet is employed for face detection, serving as a Convolutional Neural Network (CNN)-based face detector. The high accuracy and remarkable performance is of up to 1000 frames per second, YuNet is particularly notable for proficiency in recognizing side faces. The architecture of YuNet comprises three essential components: a backbone, a feature pyramid, and a head (Wei Wu, 2023).

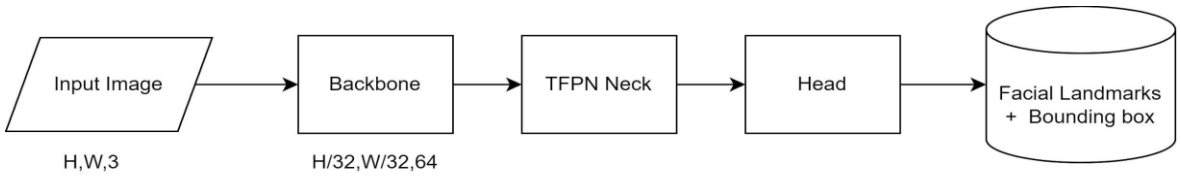


Figure 3.2: Yunet's architecture

The backbone is the central part of the network, responsible for extracting features for detection. The backbone includes 5 stages.

Stage 0 is a ConvHead module which contains a convolution layer with kernels and a stride of 2. Subsequent to Stage 0 is stage 1, which involves max pooling, a ReLU and two DWBlocks. These first two stages are used to reduce the size of feature maps to 1/4 of the input size and increase the channels from 3 to 64. From stage 2 to stage 4, the same network structure is used with DWBlocks and max pooling and subsequently passed the hierarchical output feature maps to the TFPN neck.

The neck combines multi-scale features for a higher level of features. In the model, the TFPN depthwise separable convolution, with Stage 2 as low-level features and Stage 4 as high-level features because of different depths. The number of parameters is reduced to 12% compared to FPN and the computational cost is also reduced. In the final head part, with any matching candidates, the intersection over union (IOU) is calculated by the predicted bounding box and ground truth as the soft label. The losses of bounding box regression, landmark regression and objectness are IoULoss (IOU) (Wei Wu, 2023).

Non-Maximum Suppression (NMS) is an important technique in face detection frameworks. NMS involves selecting the best bounding box when several overlapping bounding boxes are detected for the same object, which helps in reducing redundancy and improving the detection accuracy. NMS operates by comparing the overlap, or Intersection Over Union (IOU), between bounding boxes. If the IOU exceeds a specific threshold, the lesser scoring boxes are suppressed. In the context of this project, the NMS threshold is specifically set at 0.3. This threshold value is carefully chosen to balance between keeping true positives and removing duplicates, thus contributing significantly to the efficiency and accuracy of the face detection system.

3.2 Face alignment

In this experiment, face alignment is employed in order to improve the accuracy and performance of the face recognition process (Jin, 2023). The alignment process uses the output of the face detection YuNet as input, which includes the landmarks of the left and right eyes, as well as the bounding boxes. The first step of the process is taking the positions of the left and right eyes to form a third point, which is used to make a right triangle (Mishra, 2021). Subsequently, the Euclidean distance is used to calculate the edges of this triangle based on the x and y axis of every point.

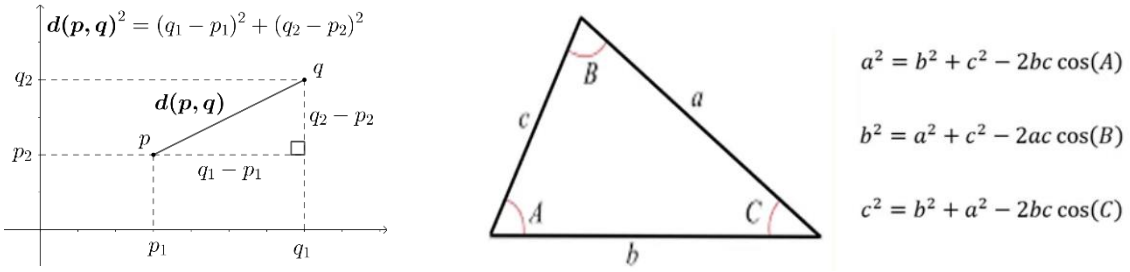


Figure 3.3: Euclidean distance

With the distances from the previous step, the cosine rule is applied to calculate the angles within the triangle, resulting in the rotation angle necessary for face alignment.

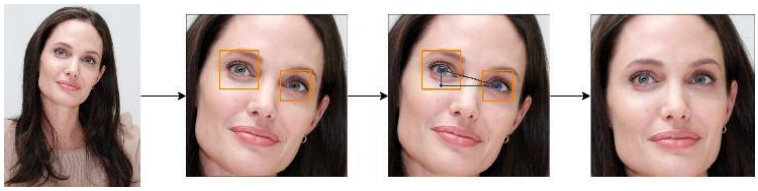


Figure 3.4: Example of face alignment

3.3 Face recognition

The network ResNet 34 is used to convert the image into embedding vectors, which are used for the matching identity process.

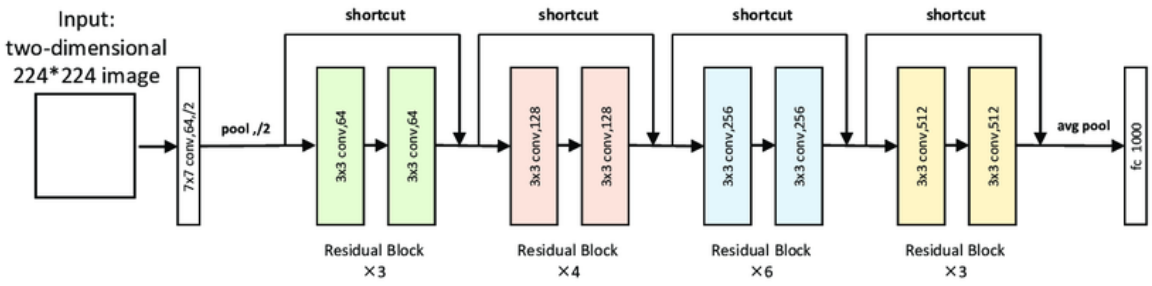


Figure 3.5: Resnet34's architecture

ResNet 34 is a Residual Network with 34 layers, including convolutional layers, pooling layers and fully connected layers. The network takes an RGB image of size 224x224 pixels as input. ResNet34 uses residual blocks, each containing 2 convolutional layers and a shortcut. The inclusion of shortcut connections enables the gradient to bypass layers to propagate the gradient back through the network during training and solve the gradient vanishing problem. Pooling layers are incorporated to reduce the spatial dimensions of the feature maps learned from the image to decrease

computational cost. In the end, fully connected layers transform high-level features into the final embedding vector, which is the final representation of an image.

The model undergoes pre-training using the glint360k dataset, which contains approximately 17 million face images distributed across 360,000 human identities. This pre-training process on a large and diverse dataset enhances the model's ability to generalize and recognize facial features effectively.

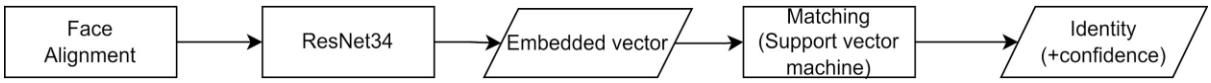


Figure 3.6: Pipeline of Recognition model to identify images

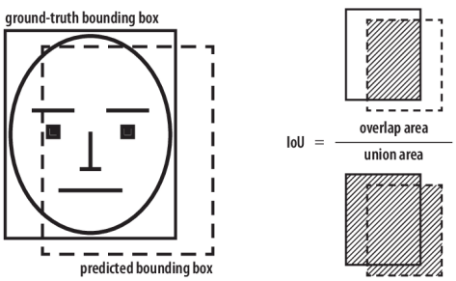
Once the faces are detected by Yunet, the cropped image of these faces as input for Resnet34. To ensure uniformity, face images with different sizes are reshaped into 224x224 pixels before going in the network for feature extraction. After the embedded vector representing the face is successfully generated by the network, the matching process using SVM assists to identify the individuals linked with the detected face based on the learned embeddings (Saining Xie, 2017).

3.4 Evaluation metrics

In this experiment, the evaluation model is divided into three parts, including YuNet evaluation, and face recognition model evaluation for real-time and datasets. In YuNet, evaluation includes he run-time, Average Intersection over Union (IOU), and Average Precision (EasyAP, MediumAP, HardAP), while in an entire model contains running time and minimum pixels for detected and recognized face for real-time, and accuary, F1-score, recall and precision for datasets.

3.4.1. Intersection over Union (IOU) score

In object detection, intersection over Union score (IoU) is a metric to estimate the accuracy of object localization. Iou is computed by the ratio of overlap area to the combined area between the ground truth bounding box and the predicted bounding box. High IoU score indicates that the model can generate boxes that closely resemble the growth truth bounding boxes (João Nuno Correia, 2019).



3.4.2 Average Precision

Average precision score (AP) measures the balance between precision and recall. The score is the area under the precision-recall curve, ranging from 0 to 1. When both precision and recall are high, the aP score is high. (Shung, 2018)

3.4.3. Recall

Recall is the fraction of instances in a class that the model correctly classified out of all instances in that class. Recall is a good measure to determine when the cost of False Negative is high. In face recognition applications, a high recall is crucial to minimize missed identifications and enhance the system's accuracy in recognizing faces (Shung, 2018).

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

3.4.4. Precision

Precision is the measure of instances correctly classified as belonging to a specific class out of all instances the model predicted to belong to that class. Precision is a good measure to determine when the cost of False Positive is high. In face recognition applications for access control, high precision is crucial to ensure that the system minimizes false positives, allowing only authorized individuals to gain access while reducing the risk of granting access to unauthorized persons (Shung, 2018).

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

3.4.5 F1 score

Accuracy is a metric for classification models that measures the number of predictions that are correct as a percentage of the total number of predictions that are made. However, accuracy is reliable only when the dataset is class balanced. In situations where there is a significant class imbalance in the data, F1 score is used to calculate the model's accuracy because the metric considers both the false positives and false negatives. F1 score is calculated by the harmonic mean of the precision and the recall, ranging from 0 to 1.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

A high F1 score indicates a good balance between precision and recall, suggesting that the model is performing well in terms of both correctly identifying positive instances and avoiding false positives and false negatives (Shung, 2018).

3.4.6 Easy, Medium and Hard Average Precision

Because of the large variations in scale, occlusion, pose and background, WIDER FACE becomes a challenging dataset with low detection rate compared to other existing datasets for face detection such as FDDB, AFW or PASCAL FACE. Based on the detection rate of EdgeBox object proposal, WIDER FACE is divided into three levels of difficulties, which are ‘Easy’, ‘Medium’ and ‘Hard’. Using three different level datasets for evaluating models gives better insight of the model’s performance in different levels of complexity. An "easy" scenario might involve well-lit frontal faces with minimal occlusions, a “medium” case might contain small faces with partial occlusions while a "hard" scenario could include faces in challenging poses or under poor lighting conditions. This evaluation method can help researchers identify specific scenarios where the model struggles. In this experiment, these types can be marked as EasyAP, MediumAP and HardAP (Shuo Yang, 2016).

3.4.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of the SVM algorithm is to find an optimal hyperplane that effectively separates data points belonging to different classes in each feature space. The hyperplane is chosen in such a way that it maximizes the margin between the classes. Hyperplane is a line which linearly divides and classifies the data. When adding the new testing data, whatever side of the hyperplane it goes will eventually decide the class that is assigned to it. The margin is defined as the distance between the hyperplane and the nearest data points (support vectors) of each class. Support Vectors are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.

3.4.8 Linear Support Vector Machine (SVM)

If a separating boundary or hyperplane can be formed to distinguish multiple class groups, the data points are linearly separable. Linear SVM are commonly used with linearly separable data points.

3.4.9 Classification

SVM includes many kinds of methods with different mathematical formulations.

- SVC: C-Support Vector Classification

SVC uses parameter C as the parameter. The default value of parameter C is 1.0.

- NuSVC: Nu-Support Vector Classification

NuSVC is similar to SVC but uses a new parameter nu which controls the number of support vectors and training errors. The parameter nu is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. The value of nu should be in the interval (0,1]. A smaller "nu" results in a stricter model, trying to avoid errors and using fewer support vectors while a larger "nu" provides a more flexible model, allowing for more errors and uses a larger number of support vectors.

The output of both Support Vector Classification (SVC) and Nu-Support Vector Classification (NuSVC) models is a predicted class label for each input sample. In multi-class classification scenarios, the model assigns each sample to one of the multiple classes. The decision is based on the learned hyperplane or decision boundary and the position of input samples with respect to it. Additionally, both models provide decision functions that return the signed distance of a sample to the hyperplane, which can be useful for obtaining confidence scores. The difference between SVC and NuSVC is the parameters used for model tuning, where SVC employs the regularization parameter `C` to control the trade-off between training error and margin width, while NuSVC utilizes the `nu` parameter, serving as an upper bound on training errors and a lower bound on support vectors (Saini, 2023).

4. Experiments and results

Experiments are designed through documentation periods leading to the decision of figure 3.1. To evaluate an entire model, the needed information includes dataset materials, evaluation results in each model and further discussion of comparison with methods in other publications.

4.1 Dataset materials

There are four different datasets for distinct purposes related to face detection, and face recognition. Almost datasets are used as pre-trained models for less time-consuming, while manual collection dataset is applied to both training and testing model in Support Vector Machine (SVM).

4.1.1 WIDER FACE

WIDER FACE dataset is used to test the face detection model as YuNet. The dataset has 32,203 images in total, containing 393,703 faces labeled. WIDER FACE data is variant in pose, occlusion and scale with a large number of scenes with diverse backgrounds, a good training source with both positive and negative samples. The whole dataset is divided into train, validation and test sets by ratio 50%, 10%, 40% respectively within each event class. In this experiment, the pretrained model with WIDER FACE dataset is used in order to save the initial training time (Shuo Yang, 2016).

Figure 4.1: WIDER FACE dataset

4.1.2 Glint360k

The deployed model ResNet34 is pre-trained using the entire Glink360k dataset, which contains 17,091,657 images of 360,232 individuals. By employing the Patial FC training strategy, baseline models trained on Glint360K can easily achieve state-of-the-art performance. The dataset has been evaluated on a large-scale test set (e.g. IFRT, IJB-C and Megaface) (Xiang An, 2021)

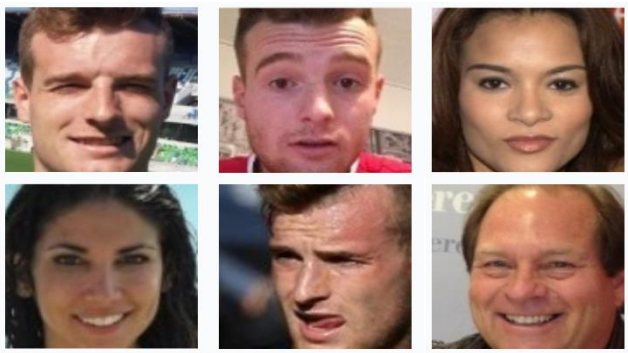


Figure 4.2: Glint360k images

4.1.3. AgeDB

AgeDB includes 16, 488 images of various famous people, with the identity, age and gender attribute in each image. 568 distinct subjects in AgeBD, together with the average number of images per subject is 29. The age ranges from 1 to 101, and the average age range for each subject is 50.3 years. Before testing, the SVM classifier is trained on the dataset with a train-test split ratio of 80%-20%, 13,191 images for testing and 3,297 images for the training. AgeDB is used for evaluating the entire model, including Face detection, Face alignment, Feature extraction and Support Vector Machine (Stylianos Moschoglou, 2017).

Figure 4.3: AgeDB images

4.1.4 Manual Collection Dataset

The dataset was created manually by group members with 416 face images (after preprocessing) of 25 people starting at 20 years old and above, as illustrated in figure 4.3. Each image contains only 1 person. The data set consists of images of 16 celebrities collected from the internet and images of the group members, friends and families captured by laptops' webcams and mobile phones' cameras. The dataset is used to train SVM and to test the entire model's real-time performance with the train-test split ratio of 80%-20%.

Figure 4.4: Manual collection Dataset

4.1.5 Device

For real-time performance evaluation, tested images are captured by the laptop webcam, a rgb camera with a video resolution of 2560 x 1440, a 4 MP camera, a field of view of 106 degrees, and a frame rate of 1080p/30 FPS.

The captured images for the manual collection dataset are taken by the laptop webcam with a resolution of 1280 x 720 pixels; and IPhone X's camera with 12Mp main camera with wide-angle f/1.8 OIS lens and 12Mp telephoto camera with f/2.4 OIS lens.

The model evaluation was carried out on a desktop computer using Ubuntu 18.04 as the operating system. The system specifications included an i5-10400 CPU running at 2.9GHz with 6 cores, 16GB of RAM operating at a speed of 2133 MHz, and a NVIDIA 3060 Ti GPU with 8GB of memory and 4864 CUDA cores.

4.2 Results

Results include three parts containing the evaluation of Face detection model of YuNet and comparison with other Face detection methods. Moreover, an entire model of Face Recognition consists of two evaluations of real-time and datasets to demonstrate the accuracy, efficiency and real-time recording.

4.2.1 Face detection model evaluation

Table 4.1 indicates the evaluation of method YuNet used in the study. YuNet achieves real-time performance, processing 17.7-19.7 images per second with the WIDER FACE validation test. This means the model can perform face detection on 17.7-19.7 images per second, suitable for applications requiring rapid processing. When evaluated with the validation set of WIDER FACE, YuNet achieves an Easy AP of 0.8843, along with an Average IOU (Intersection over Union) of 0.71524. While in the MediumAP, YuNet experiences a Medium AP of 0.8655, along with an Average IOU of 0.6943. Finally, a Hard AP is of 0.7502, and an Average IOU of 0.6237.

During testing the face detector real-time by streaming, YuNet gives an impressive result with 75-85 frames per second, indicating that the face detection model is very effective for the real-time face recognition framework.

Total results induce the fact that the model has high accuracy when dealing with easily recognizable faces. Furthermore, medium-difficulty faces seem to be effective to detect and recognize, along with relatively high accuracy with hard-to-detect faces.

Table 4.1: Evaluation results of Face detection model on WIDER FACE dataset with other methods

Method	Confidence threshold	Configurations	Dataset	Run-time	Easy AP	Average IOU	Medium AP	Average IOU	Hard AP	Average IOU
YuNet (Ours) (Resolution 640 x640)	0.3	i5-10400 CPU 2.9GHz , 6 cores (CPU) and NVIDIA 3060 Ti 8GB (GPU)	WIDERFACE	19.3 images/s	0.8843	0.7152	0.8655	0.6943	0.7502	0.6237
	0.5			19.7 images/s	0.8773	0.7314	0.8520	0.7106	0.7101	0.6360
	0.7			17.7 images/s	0.8519	0.7538	0.8155	0.7334	0.6151	0.6444
	0.8			18.4 images/s	0.8119	0.7815	0.7601	0.7637	0.4826	0.5895
YuNet (Public) (Resolution 640 x 480) *(Wei Wu, 2023)*	0.01	i7-12700K (CPU)	WIDERFACE	104 FPS	0.899	-	0.869	-	0.691	-
YuNet-s (Public) (Resolution 640 x 480) *(Wei Wu, 2023)*	0.01	i7-12700K (CPU)	WIDERFACE	128 FPS	0.876	-	0.834	-	0.591	-
SCRFD-0.5g (Resolution 640 x 480) *(Wei Wu, 2023)*	0.95	NVIDIA 2080 TI (GPU)	WIDERFACE	277 images/s	0.850	-	0.754	-	0.372	-
RetinaFace-0.25 (CPU-1) (Resolution 640 x 480) *(Wei Wu, 2023)*	0.5	i7-6700K (CPU)	WIDERFACE	60 FPS	0.765	-	0.611	-	0.271	-

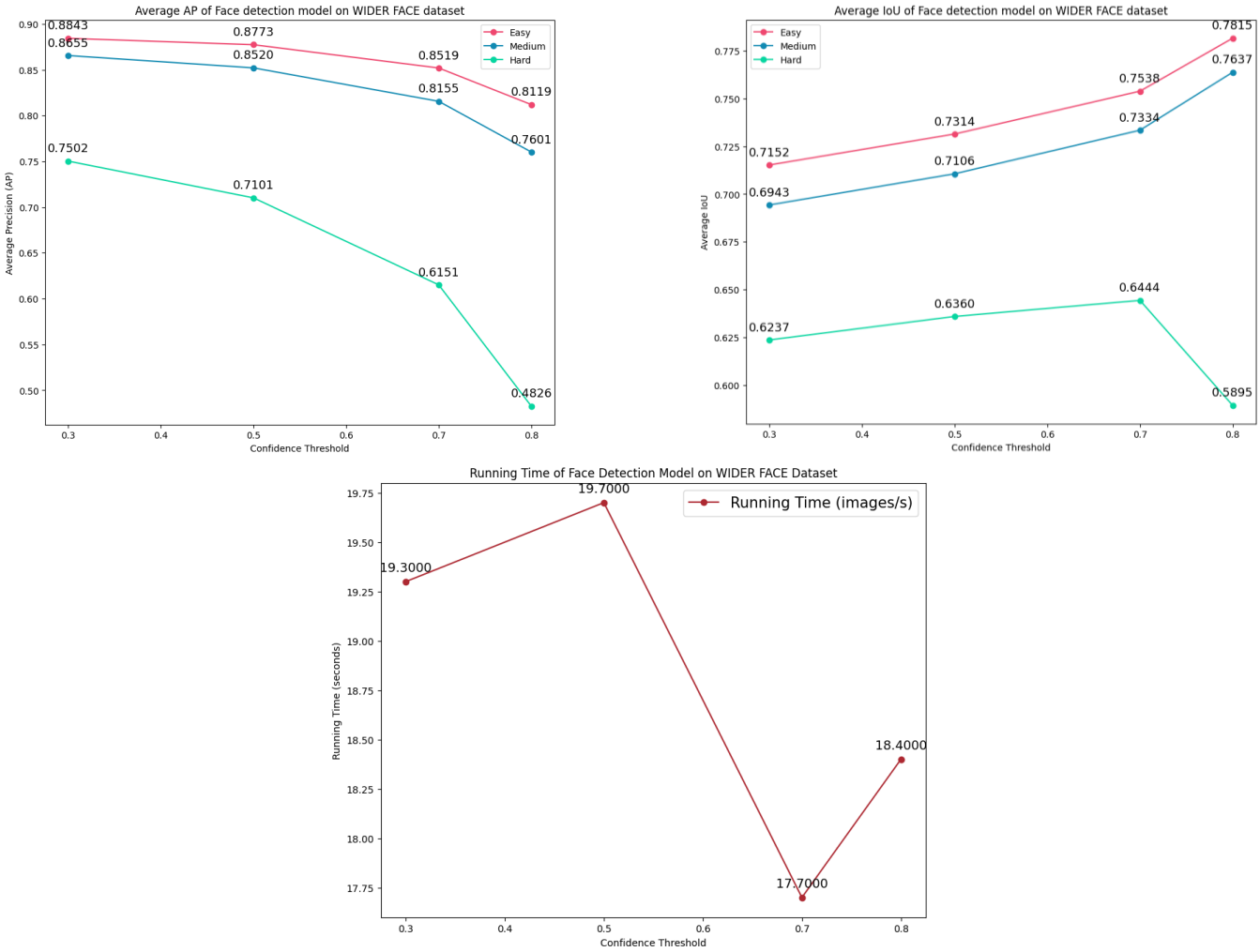


Figure 4.5: Results of Face detection model - YuNet on WIDER FACE dataset

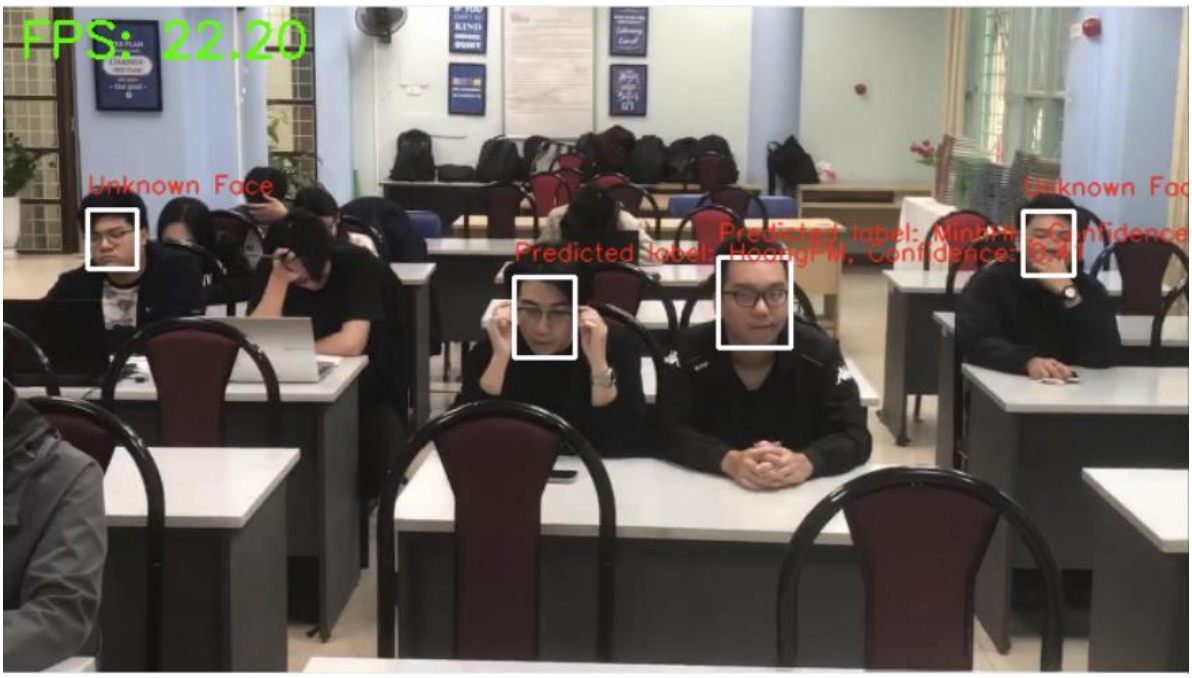


Figure 4.6: Results of Face recognition model

4.2.2 Face recognition model evaluation

Evaluation of face recognition is divided into several parts, including face detection, alignment and face recognition, and the entire model with the difference of using two other methods of C-Support Vector Classification (SVC) and Nu-Support Vector Classification (NuSVC).

Table 4.2: Evaluation of entire model in real-time video

Methods	Dataset	Speed	Output	Minimum pixels of detected face	Minimum pixels of recognized face (real-time)	Minimum pixels of recognized face (video demo)
YuNet + Alignment + ResNet34 + SVM	Manual collection dataset	30 FPS (25 users in 416 images)	Label name Probability	28 x 23 YuNet Confidence Threshold = 0.9	207 x 157 SVM Probability Threshold = 0.5	33 x 28 SVM Probability Threshold = 0.4

Moreover, table 4.2 indicates the combination of different models and techniques: YuNet for face detection, alignment for normalization, ResNet34 for feature extraction, and SVM for classification. For face detection, the face confidence threshold is 0.9 and NMS threshold is 0.3, which results in the smallest face detected in real-time testing with a size of 28x23 pixels. For classification, the threshold of whether a detected face is matched with a person in the database is set to be 0.5 for real-time testing, which results in the smallest face recognized to be 207x157 pixels and for the video demo, the threshold is set to be 0.4 and the minimum pixels of recognized face is 33 x 28 pixels. The minimum number of pixels of a detected and recognized face's image is based on the threshold for each process, as well as the color, resolution of the video and camera input. In real-time testing, the lower the threshold is, the smaller number of pixels of the face's image will be detected and recognized. The model processes, analyzes video frames in real-time and handles about 30 frames per second. In case that multiple faces are detected and classified at the same time, any number of more than three faces subsequently results in FPS drop. The speed is essential for applications that require timely responses such as live video analysis. In the output's section, labels' names are the identities or classes associated with the recognized users. Each detected face is assigned a label indicating the user it corresponds to. Furthermore, the probability is to provide the confidence or likelihood of the assigned labels, assessing the model certainty in the prediction.

When comparing table 4.1 and 4.2, it's obvious that the inference time of testing YuNet (17.7-19.7 images per second) with WIDER FACE validation set is slower than the entire model when testing real-time (30 FPS). This is due to the fact that the validation set of WIDER FACE contains 3226 images and 39708 faces in different poses and occlusions, which averaged out to be 12.3 faces per image; while in real-time testing, only three faces are detected and classified at the same time as stated above.

Table 4.3: Evaluation based on AgeDB dataset and manual collection dataset according to SVM results in different methods.

Method	Testing dataset	Configuration	SVM	Run-time	Accuracy	F1 score	Recall	Precision
YuNet + Alignment + ResNet34 + SVM (ours)	AgeDB	i5-10400 CPU 2.9GHz , 6 cores (CPU), 16GB RAM, 2133 MHz and NVIDIA 3060 Ti , 8GB, 4864 CUDA cores (GPU)	SVC (linear)	50.7 images per second	68.03%	67.25%	68.03%	72.83%
			NuSVC (nu=0.01, linear)	55.3 images per second	69.96%	64.84%	69.96%	74.41%
YuNet + Alignment + ResNet34 + SVM (ours)	Our manual collection dataset		SVC (linear)	89.6 images per second	100%	100%	100%	100%
			NuSVC (nu=0.01, linear)	94.1 images per second	98.81%	99.11%	98.81%	98.78%
YOLOv7 *(Ahmad S. Lateef, 2023)*	Author-collected dataset (32 examined faces)	i7-7500U (CPU), 2.70GHz 2.90GHz, 16 GB RAM	-	5 FPS	100%	100%	100%	100%
HOG + Dlib *(Ahmad S. Lateef, 2023*	Author-collected dataset (32 examined faces)	i7-7500U (CPU), 2.70GHz 2.90GHz, 16 GB RAM	-	4 FPS	94%	97%	94%	100%
Viola Jones + LBPH *(Ahmad S. Lateef, 2023*	Author-collected dataset (32 examined faces)	i7-7500U (CPU), 2.70GHz 2.90GHz, 16 GB RAM	-	16 FPS	91%	96%	97%	94%
FN8 (sample 3) *(Zong-Yue Deng, 2023)*	LFW and AgeDB	i7-7500U (CPU), 2.70GHz 2.90GHz, 16 GB RAM	-	4 FPS	99.2%	97.56%	-	-

Table 4.3 indicates the provided information for two different SVM models (linear kernel) within the context of face recognition using YuNet, Alignment, and ResNet34. Processing the images into face embedded vectors for identified classification, the model takes on average, 84 images per second for the AgeDB dataset and 106.6 images per seconds for the manual collection dataset. The total time for evaluating the model is based heavily on SVM, where the time complexity is O(n^3), with n being the number of data points (number of embedded vectors), which results in the run-time of evaluating the model using AgeGB dataset is much slower than using manual collection dataset. Utilizing the SVC for evaluating using the AgeDB dataset, the model achieves the speed of 50.7 images per second with an accuracy of 68.03%, indicating proficiency in correctly identifying faces. The F1 score, a balanced measure of precision and recall, is commendable at 67.25%, while the recall rate specifically demonstrates a strong performance at 68.03%. Precision, showing the accuracy of positive predictions, is notable at 72.83%. On the other hand, after training and testing the SVC classifier on the manual collection dataset, the entire model processes much faster with a run-time of 89.6 images per second, due to the smaller number of faces each frame. The model also gives better results in accuracy, f1 score, recall and precision, reaching the score of 100% in every metric. Overall, this configuration showcases a well-balanced and efficient face recognition model, providing both speed and accuracy in real-time applications.

The face recognition model employs Nu-Support Vector Classification (NuSVC) with a linear kernel and a specified nu value of 0.01. This comprehensive model operates at a speed averaging 55.3 images per second with the AgeDB dataset. In terms of performance evaluation, the model demonstrates the accuracy of 69.96%. The F1 score is particularly impressive at 64.84%, showcasing the model's effectiveness across various recognition aspects. The recall rate stands at 69.96%. Precision, denoting the accuracy of positive predictions, is commendable at 74.41%. While testing on the manual collection dataset, the NuSVC classifier has better performance in all metrics and processing time, which are 98.81%, 99.11%, 98.81%, 98.78% for accuracy, f1 score, recall and precision respectively. In summary, this configuration establishes a robust and efficient face recognition model, balancing real-time processing speed with high accuracy and precision.

Both models provide satisfactory real-time processing speeds for dynamic applications. The choice could depend on specific requirements, with the NuSVC model offering slightly improved accuracy and an F1 score. Overall, these configurations induce effective and reliable face recognition models, resulting in figure 4.6.

5. Conclusion and Discussion

In summary, the comprehensive evaluation of the face recognition model, incorporating face detection, alignment, and recognition processes along with SVM-based classification, reveals a highly capable and versatile system. The YuNet model demonstrates exceptional real-time performance, operates at 75-85 frames per second (FPS) and executes 17.7-19.7 images per second when being tested on the WIDER FACE dataset (average 12.3 faces in an image). Furthermore, YuNet particularly excels in recognizing easily distinguishable faces with an Easy AP of 0.8843. The evaluation further extends to medium-difficulty and hard-to-detect faces, which indicates consistently high accuracy.

The face feature extraction model, incorporating YuNet, alignment and ResNet34, operates at a speed of 84 images per second when using AgeDB dataset and 106.6 images per second when using the manual collected dataset, showing the ability to efficiently process images. The output of ResNet34 comprises 512-dimension embedded vectors, reflecting a representation of facial features. This model obtains a good balance between accuracy and processing speed.

Additionally, by combining YuNet, alignment, ResNet34, and SVM, the model operates real-time at 30 FPS, providing label names and probability scores for identified users. Overall, after testing on the datasets and real-time testing, the entire model deployed with SVC appears to recognize faces more accurately when tested on a small scale of people. However, when the number of users increases, resulting in more identities to recognize, NuSVC seems to be the better choice for the classification task due to its lower processing time while maintaining a high rate of accuracy.

In the future, there are many ways for improving attendance check models that use facial recognition and deep learning. Firstly, the proposed face recognition model needs to be improved to be used in real-life situations. In an instance with a larger number of identities, the model needs to calculate the similarity metrics between the vector of the detected face and the vectors of all identities in the database. The computational cost increases as the database size increases.

Secondly, the model should be tested in different places and situations, such as in schools, offices or big events. Adapting the system to fit different needs makes the model enhance the versatility and practicality. Subsequently, efficient GPU deployment should be focussed to minimize latency.

Thirdly, in future developments of this project's real-time face recognition model for classroom attendance, a key focus will be on optimal camera placement. Cameras at the classroom entrance would capture students' faces upon entry, facilitating a smooth attendance process. Additionally, overhead cameras could provide comprehensive coverage of the entire classroom. This setup would not only improve

recognition accuracy but also ensure a non-intrusive, efficient attendance system. Attention to privacy concerns and ethical standards in such deployments will be taken into account.

Figure 5.1: Camera's acquisition

Finally, an important aspect is to enhance the model's security by using anti-spoofing methods. Anti-spoofing features help prevent fake attempts to trick the facial recognition system, ensuring that it accurately identifies real faces and avoids unauthorized access.

6. Reference

Ahmad S. Lateef, M. Y. (2023). Face Recognition-Based Automatic Attendance System in a Smart Classroom. *Iraqi Journal for Electrical And Electronic Engineering*, 37-47.

Zong-Yue Deng, Hsin-Han Chiang, Li-Wei Kang, Hsiao-Chi Li(2023), A lightweight deep learning model for real-time face recognition. *The Institution of Engineering and Technology,*

Edwin Jose, G. M. (2019). Face Recognition based Surveillance System Using FaceNet and MTCNN on Jetson TX2. *5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 608-613.

Florian Schroff, D. K. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2015*.

Huiyan Wu, M. X.-M. (2019). Multi-Level Feature Network With Multi-Loss for Person Re-Identification. *IEEE Access*, pp(99).

Jiankang Deng, J. G. (2015). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *JOURNAL OF LATEX CLASS*.

Jin, M. (2023). *A Study of Face Alignment Methods in Unmasked and Masked Face Recognition.* UPPSALA UNIVERSITET.

João Nuno Correia, T. M. (2019). Evolutionary data augmentation in deep face detection. *The Genetic and Evolutionary Computation Conference Companion*.

Karl Maria Michael de Leeuw, J. B. (2007). The History of Information Security. *A Comprehensive Handbook*.

Mishra, M. (2021). Study of Nonlinear Control Techniques. *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021*.

Qi, S., Zuo, X., Feng, W., & Naveen, I. G. (2022). Face Recognition Model Based On MTCNN And Facenet. *IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022*, 1-5.

Saini, A. (2023, October 27th). *Analytics Vidhya*. Retrieved from https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Saining Xie, R. G. (2017). Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samridhi Dev, T. P. (2020). Student Attendance System using Face Recognition. *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 90-96.

Shung, K. P. (2018, March 15th). *Towards Data Science*. Retrieved from Medium: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Shuo Yang, P. L. (2016). WIDER FACE: A Face Detection Benchmark. *IEEE*.

Stylianos Moschoglou, A. P. (2017). AgeDB: the first manual collection, in-the-wild age database. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1997-2005.

Truein. (2023). *Disadvantages of the Fingerprint Attendance System*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/disadvantages-fingerprint-attendance-system-truein

Wei Wu, H. P. (2023). YuNet: A Tiny Millisecond-level Face Detector. *Machine Intelligence Research* .

Xiang An, X. Z. (2021). Partial FC: Training 10 Million Identities on a Single Machine. *Association for the Advancement of Artificial Intelligence*.

Xiao Han, Q. D. (2018). Research on face recognition based on deep learning. *Conference: 2018 Sixth International Conference on Digital Information, Networking, and Wireless Communications (DINWC)*.

Xudong Sun, P. W. (2018). Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*.

Yi Sun, D. L. (2015). DeepID3: Face Recognition with Very Deep Neural Networks. *arxiv*.

***Figure 5.1**: https://www.nea.org/nea-today/all-news-articles/cameras-classroom-big-brother-evaluating-you*

7. Appendix

In order to provide more detail about the methodologies employed and those methods architecture, this section consists of several parts. The first part is experimental protocol, which shows how the experiment is conducted, including all the preparation and parameters involved. The following sections describe the architecture of Yunet and the architecture of Resnet34 in detail.

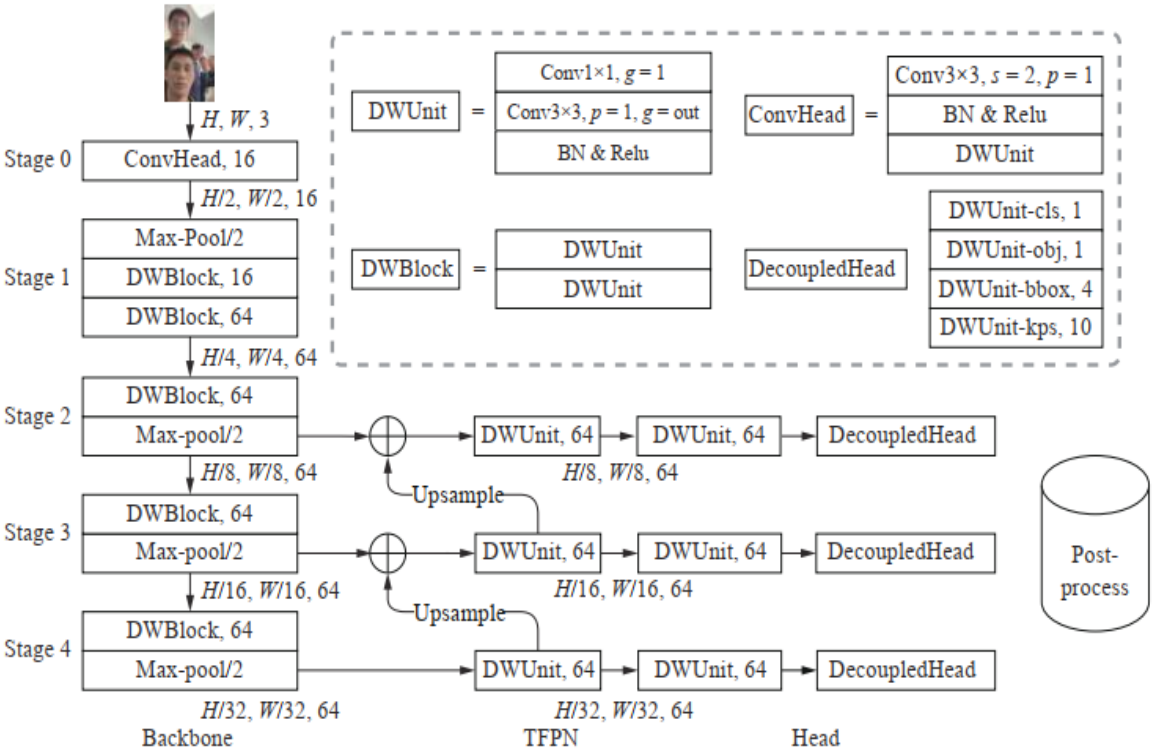
7.1 Experimental protocol

In this experiment, the YuNet face detection model was set with a confidence threshold of 0.9 and an NMS threshold of 0.3. It was tested on the WIDER FACE dataset, which categorizes detection difficulty into 'Easy', 'Medium', and 'Hard' levels. The testing set consists of 12,881 images in total.

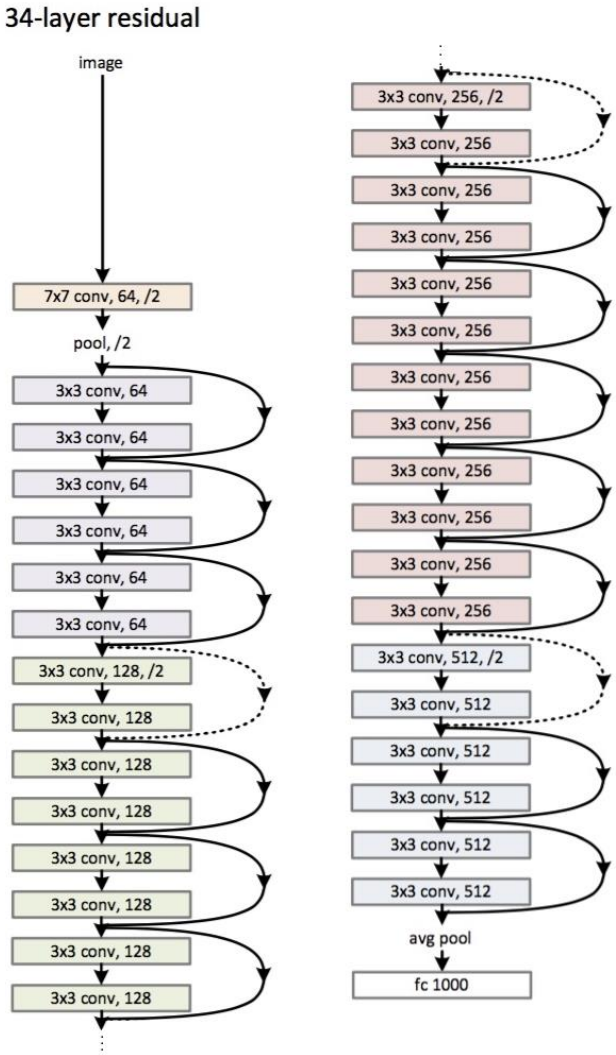
To obtain the efficiency metrics of the entire face recognition model which includes YuNet for face detection, alignment for normalization, ResNet34 for feature extraction, and SVM for classification, the framework is tested on the AgeDB dataset. Before testing, the SVM classifier is trained on the dataset with testing-training ratio of 20%-80%, 13,191 images for testing and 3,297 images for the training. In the classification process, the probability threshold for identity matching is set to 0.5.

In order to test real time performance of the entire face recognition model, SVM classifier is trained with the manual collection dataset, taken by group members. The data set consists of images of 16 celebrities collected from the internet and images of the group members captured by laptops' webcams and mobile phones' cameras. All the images are cropped around the face and resized to the standard size of 224 x 224 before utilization. After the training phase, the model is tested via laptop's webcam.

7.2 Yunet's architecture



7.3 Resnet34's architecture



ResNet-34 is a variant of the Residual Network architecture specifically designed with 34 weighted layers to solve image classification tasks. The model incorporates residual learning to facilitate the training of these deep networks by employing shortcut connections, also known as skip connections, that bypass one or more layers. The architecture starts with a single 7x7 convolutional layer with 64 filters, followed by a max pooling layer, both of which prepare the input for the deeper layers while reducing the dimensionality.

The core of the network consists of four main blocks of layers, each containing multiple sets of 3x3 convolutional layers with the same number of filters within each block. The first two blocks use 64 filters, the third block increases to 128 filters, and the fourth block uses 256 and then 512 filters. Each block has multiple residual units where the input to each unit is added to the output, thus performing identity mapping and preventing the degradation problem in training deep networks. This allows for the training of deeper networks without the performance saturation.

Between these blocks, the network employs convolutional layers with a stride of 2 to reduce the spatial size of the feature maps, effectively performing downsampling. The use of these strided convolutions is critical as it reduces the computational burden and controls the growth of the activation maps throughout the network.

The network concludes with an average pooling layer that aggregates the feature maps, reducing each map to a single number. This is followed by a fully connected (dense) layer with 1000 output units, corresponding to the number of classes in datasets like ImageNet. The fully connected layer makes the final predictions after which a softmax activation function is typically applied to obtain the probability distribution over the classes.

ResNet-34 thus is a balance trade-off between computational efficiency and the ability to learn complex features, making it a robust choice for various image recognition applications.