FPT Education

FPT UNIVERSITY

SUPERVISE ACCESS TO TOURIST AREAS AND INDUSTRIAL PARKS; PREDICTING EMOTIONS, GENDER AND AGE.

by

Trần Xuân Thành

THE FPT UNIVERSITY HO CHI MINH CITY

SUPERVISE ACCESS TO TOURIST AREAS AND INDUSTRIAL PARKS; PREDICTING EMOTIONS, GENDER AND AGE.

by

Trần Xuân Thành

Supervisor: Mr. Nguyễn Quốc Trung

A capstone project submitted in partial fulfillment of the requirement
for the completion of AI Development with TensorFlow subject.

ARTIFICIAL INTELLIGENCE MAJOR

THE FPT UNIVERSITY HO CHI MINH CITY

March 2024

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

The authors wish to express their gratitude to the following individuals for their contributions to this project:

- Conceptualization: T. X. Thành.
- Methodology: T. X. Thành.
- Software: T. X. Thành.
- Validation: T. X. Thành.
- Formal analysis: T. X. Thành.
- Investigation: T. X. Thành.
- Resources: T. X. Thành.
- Data curation: T. X. Thành.
- Writing—original draft preparation: T. X. Thành.
- Writing—review and editing: T. X. Thành.
- Visualization: T. X. Thành.
- Supervision: T. X. Thành.
- Project administration: T. X. Thành.
- Funding acquisition: T. X. Thành.

All authors have read and agreed to the Final Capstone Project document.

ABSTRACT

In the current era, Artificial Intelligence is being promoted in many industries and fields in Vietnam, bringing obvious benefits. Serving the implementation of *Official Dispatch No. 4033/TCTTKĐA* dated June 13, 2023 of the Working Group implementing *Project 06 of the Government* on implementing pilot models to promote the implementation of Project 06, *dated November 1 In 2023,* organizations are researching and developing camera models that apply artificial intelligence, including Model 13: Deploying AI Cameras to control entry and exit to tourist areas and Model 14: Deploying AI Cameras to control entry and exit to the Industrial Park. With the model of controlling and detecting people loitering in front of industrial parks, there is currently only model of the KBVISION AI camera, but the definition of loitering is too simple. Other models have not achieved high performance.

In this project, we develop a loitering detection model with a definition of loitering more suitable to the actual situation in Industrial Parks in Vietnam.

In addition, in this project, we further develop a model to count the number of people entering/exiting the tourist area's gates and predicting customers' emotions, gender, and age using various CNN model architectures and transfer learning techniques.

Keywords: AI Camera, Official Dispatch No. 4033/TCTTKĐA, Project 06.

CONTENTS

List of Figures

List of Tables

1. INTRODUCTION

Embracing AI in Vietnam

In the midst of a rapidly evolving technological landscape, Artificial Intelligence (AI) emerges as a pivotal force reshaping industries and societies worldwide. Vietnam, cognizant of the transformative potential of AI, has embarked on a journey to leverage this cutting-edge technology across various sectors, seeking to unlock unprecedented efficiencies and drive sustainable development. Central to this endeavor is the implementation of Official Dispatch No. 4033/TCTTKĐA, issued on June 13, 2023, by the esteemed Working Group tasked with overseeing Project 06 of the Government. This landmark directive underscores the nation's commitment to piloting innovative AI models aimed at catalyzing progress and addressing multifaceted challenges.

AI-Infused Camera Systems in Project 06

Within the ambit of Project 06, significant emphasis is placed on the deployment of AI-infused camera systems, heralding a new era of intelligent surveillance and monitoring. Two notable manifestations of this initiative are Model 13 and Model 14. Model 13, dedicated to enhancing security measures within tourist areas, entails the strategic deployment of AI cameras to regulate the ingress and egress of visitors, thereby bolstering safety protocols and optimizing visitor experiences. Conversely, Model 14 is tailored to cater to the unique needs of industrial parks, wherein AI-powered cameras are deployed to meticulously monitor entry and exit points, fostering a secure and conducive environment for industrial operations.

Challenges in AI-Driven Surveillance

Despite commendable progress in AI-driven surveillance, notable challenges persist, particularly in the domain of loitering detection. Presently, the KBVISION AI camera system stands as a pioneering solution in this sphere. However, existing frameworks for defining loitering behaviors often fall short of capturing the intricate nuances inherent in real-world scenarios, thereby necessitating a paradigm shift towards more sophisticated surveillance methodologies.

Advanced Loitering Detection Model

Against this backdrop, our project endeavors to address the lacunae in existing surveillance systems by pioneering the development of an advanced loitering detection model. Rooted in a profound understanding of the operational dynamics of industrial parks in Vietnam, our model aims to redefine the parameters of loitering detection, incorporating nuanced behavioral insights to enhance security and preempt potential threats effectively.

Augmenting Surveillance Capabilities in Tourist Areas

Moreover, our project seeks to augment the surveillance capabilities of tourist areas by implementing a comprehensive model capable of not only accurately counting footfall at entry and exit points but also discerning customers' emotional states, gender, and age demographics in real-time. Leveraging state-of-the-art CNN model architectures and harnessing the power of transfer learning techniques, our endeavor aspires to set a new benchmark in AI-driven surveillance, thereby fortifying the foundations of safety, security, and prosperity in Vietnam.

As we embark on this transformative journey, we remain cognizant of the profound implications of our endeavors, striving to harness the potential of AI for the collective betterment of society and the advancement of Vietnam's technological prowess on the global stage.

2. RELATED WORK

The supervision of access to public areas and restricted zones is a crucial aspect of maintaining security and safety. This section reviews existing research related to our project, focusing on:

1. **People Counting and Loitering Detection:**
2. **Emotion, Gender, and Age Recognition:**

2.1. People Counting and Loitering Detection

2.1.1. Computer Vision Techniques:

- **Image-based methods:** Early works utilized background subtraction and foreground object detection for people counting [1]. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have led to more robust and accurate counting, especially in crowded scenes [2, 3].
- **Video-based methods:** These methods analyze video sequences to track individual trajectories and identify loitering behavior. Techniques like optical flow and long short-term memory (LSTM) networks have shown promising results [4, 5].

2.1.2. Sensor-based Techniques:

- **LiDAR (Light Detection and Ranging):** This technology utilizes laser pulses to measure distances and create 3D point clouds, enabling accurate people counting even in low-light conditions [6].
- **Thermal cameras:** These capture heat signatures, allowing for people detection regardless of visible light availability. However, they may struggle with background heat sources [7].

2.1.3. Challenges and Limitations:

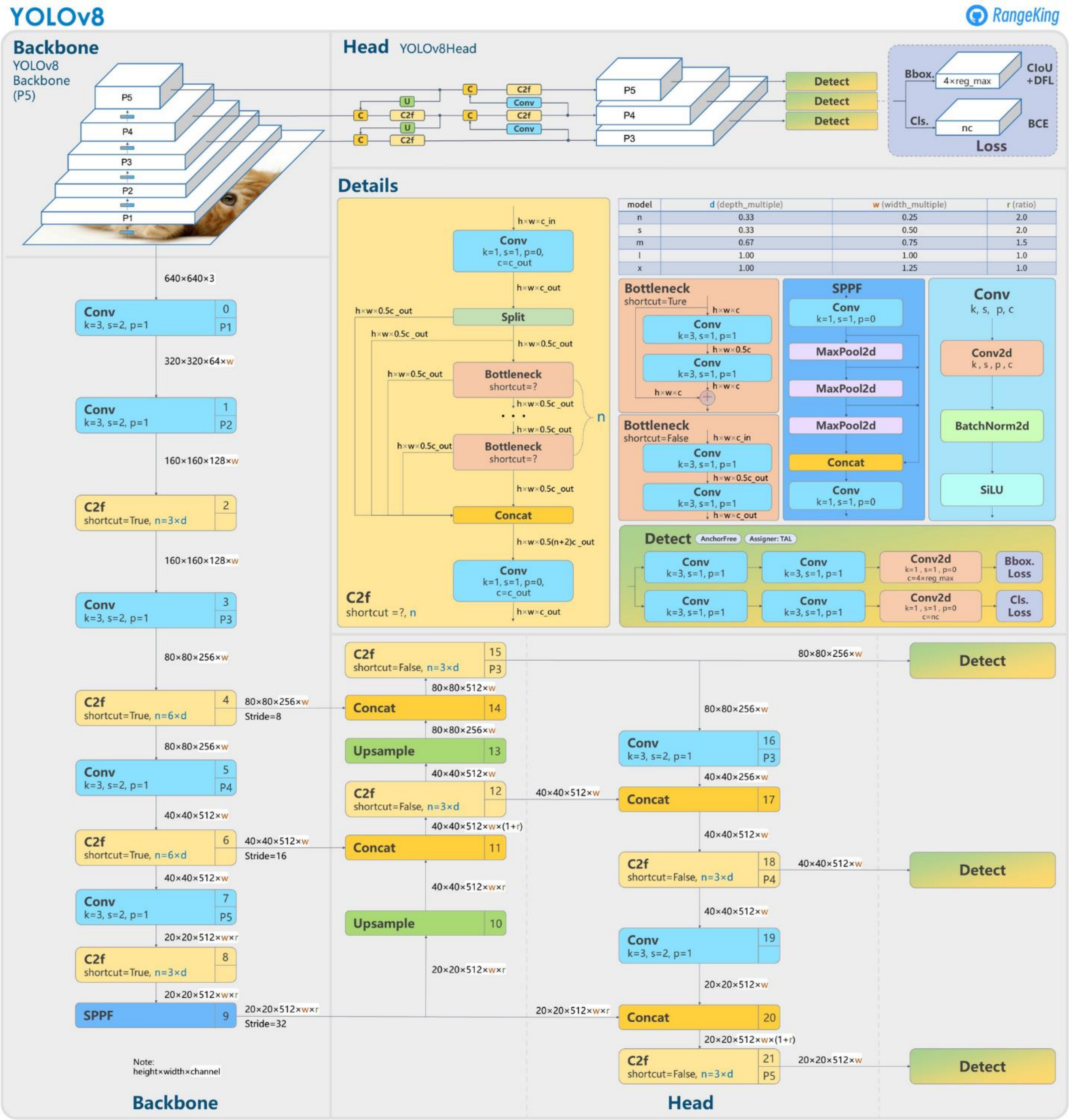
- **Occlusions:** Overlapping individuals or objects can hinder accurate counting and tracking [8].
- **Camera angle and perspective:** Different camera placements can introduce distortions and affect counting accuracy [9].
- **Privacy concerns:** Deploying such systems requires careful consideration of ethical and legal implications regarding data collection and individual privacy [10].

2.1.4. YOLOv8 Detection and Tracking:

YOLOv8 is a state-of-the-art object detection and tracking model that offers significant advantages over previous models in terms of speed, accuracy, and versatility. It is well-suited for real-time applications such as autonomous vehicles and video surveillance.

Benefits of using YOLOv8:

- High performance: meets requirements for speed and accuracy.
- Flexibility: applicable to various fields and tasks.
- Easy to deploy: open source, with comprehensive documentation.



model	d (depth_multiple)	w (width_multiple)	r (ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0

Figure 1. YOLOv8 Architecture, visualisation made by GitHub user RangeKing

2.2. Emotion, Gender, and Age Recognition

2.2.1. Deep Learning for Facial Analysis:

- **CNNs** have revolutionized facial analysis tasks, achieving high accuracy in emotion, gender, and age recognition [11]. These models are trained on large labeled datasets containing diverse facial expressions and demographics.

- **Attention mechanisms** within CNNs can focus on specific facial regions relevant for specific tasks, like the mouth for emotion recognition [12].

2.2.2. Challenges and Limitations:

- **Facial occlusions:** Similar to people counting, occlusions by hair, glasses, or masks can significantly impact recognition accuracy [13].
- **Cultural and individual variations:** Facial expressions and appearances vary across cultures and individuals, requiring diverse datasets and robust algorithms to maintain generalizability [14].
- **Ethical considerations:** Similar to loitering detection, using emotion recognition in public spaces raises ethical concerns about potential misuse and discrimination based on inferred emotions [15].

2.2.3. Alternative Approaches:

- **Gait analysis:** Analyzing walking patterns offers a privacy-preserving alternative to facial recognition for gender and age estimation [16]. However, accuracy may be lower compared to facial analysis.
- **Physiological signals:** Sensors capturing heart rate, skin conductance, or other physiological responses hold potential for emotion recognition, but raise further privacy concerns [17].

2.2.4. "Age, Gender Prediction and Emotion recognition using Convolutional Neural Network" (by Arjun Singh, Nishant Rai, Prateek Sharma, Preeti Nagrath and Rachna Jain, 2021):

In the research paper *"Age Gender Prediction and Emotion Recognition using Convolutional Neural Network"* by Arjun Singh, Nishant Rai, Prateek Sharma, Preeti Nagrath, and Rachna Jain, the authors present a sophisticated method for predicting age and gender, and recognizing emotions, using Convolutional Neural Networks (CNNs). This study is particularly relevant to the field of visual surveillance and biometric analysis.

Key aspects of the paper include:

The development of two distinct models: one for age and gender prediction utilizing the Wide Residual Network (Wide ResNet) architecture, and another for emotion recognition employing a conventional CNN framework.

The utilization of the IMDb-WIKI dataset for age-gender classification and the Fer2013 dataset for emotion recognition.

A detailed description of both models' architectures, highlighting the Wide ResNet for its "flat" and "wide" variation of the standard ResNet.

Significant results, with the age-gender prediction model achieving an accuracy of 96.26%, and the emotion recognition model reaching 69.2% accuracy.

These findings offer valuable insights for advancements in automated age and gender detection, especially in the presence of varying facial expressions, and contribute to the growing body of knowledge in emotion recognition through neural networks.

Conclusion:

This review highlights the advancements in people counting, loitering detection, and facial analysis using computer vision and deep learning techniques. However, challenges remain in addressing factors like occlusions, camera perspectives, and ethical considerations. Our project aims to contribute to this field by developing robust and ethical solutions for access supervision in tourist areas and industrial parks.

3. PROJECT MANAGEMENT PLAN

Table 1. Project plan

Task name	Priority	Owner	Start date	End date	Status	Issues
Learn about the YOLOv8 model	High	T. X. Thành	15 January	21 January	Completed	
Implement loitering detection and people counting algorithms.	High	T. X. Thành	22 January	28 January	Completed	Not good tracking.
Prepare data to build emotion, gender and age models	High	T. X. Thành	15 February	21 February	Completed	
Build and evaluate models	High	T. X. Thành	22 February	28 February	Completed	Not work well on Vietnam data (age).
Execute and complete the project	High	T. X. Thành	01 March	07 March	Completed	
Future work	High	T. X. Thành	April	…	Waiting	Improve models

The supporting information can be downloaded at: *https://github.com/thanhtranfpt/DAT-project*

Table 2. Source code and data

Items	Link	Description
Data	Google Drive	Dataset, Backup resources
Source code	GitHub	Code management, Final publish project

4. MATERIALS AND METHODS

4.1. Project Materials:

4.1.1. UTKFace dataset

UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. This dataset could be used on a variety of tasks, e.g., face detection, age estimation, age progression/regression, landmark localization, etc. Some sample images are shown as following:

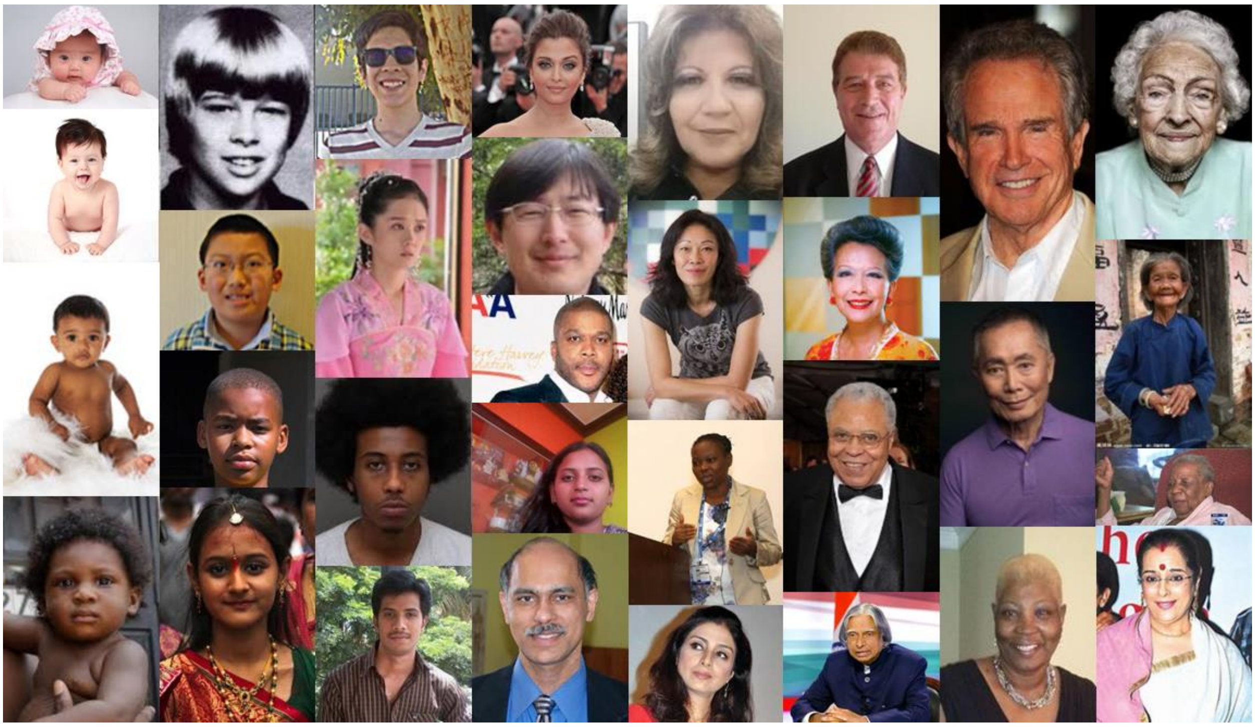


Figure 2. Some sample images of *UTKFace* dataset

.

We will use this dataset to train the gender model.

Link to the dataset on kaggle.com:

https://www.kaggle.com/datasets/jangedoo/utkface-new

4.1.2. Diverse Asian Facial Ages dataset

This dataset encompasses a rich collection of 248 images, all featuring Asian faces spanning an age range from **0 to 80 years**. With a focus on diversity, this dataset offers a comprehensive representation of facial features and aging characteristics across different life stages.

Additionally, the dataset includes a file named "**DeepFace_analyze.csv**" which encapsulates the results of analyzing the dataset using the DeepFace library. This analysis focuses on evaluating the accuracy and performance of the DeepFace library when applied to Asian faces, specifically individuals of Vietnamese ethnicity. The CSV file serves as a valuable resource for gauging the library's effectiveness in handling facial analysis tasks within the context of the dataset.

Dataset Details:

- Total Images: 248
- Age Groups: 0 - 80
- Ethnicity: Asian
- Image Resolution: Varies (Good, Normal, Poor, or Hidden - such as when facial features are obscured by items like facemarks)
- Image Format: JPEG
- Face Angle: Straight, Left, Right

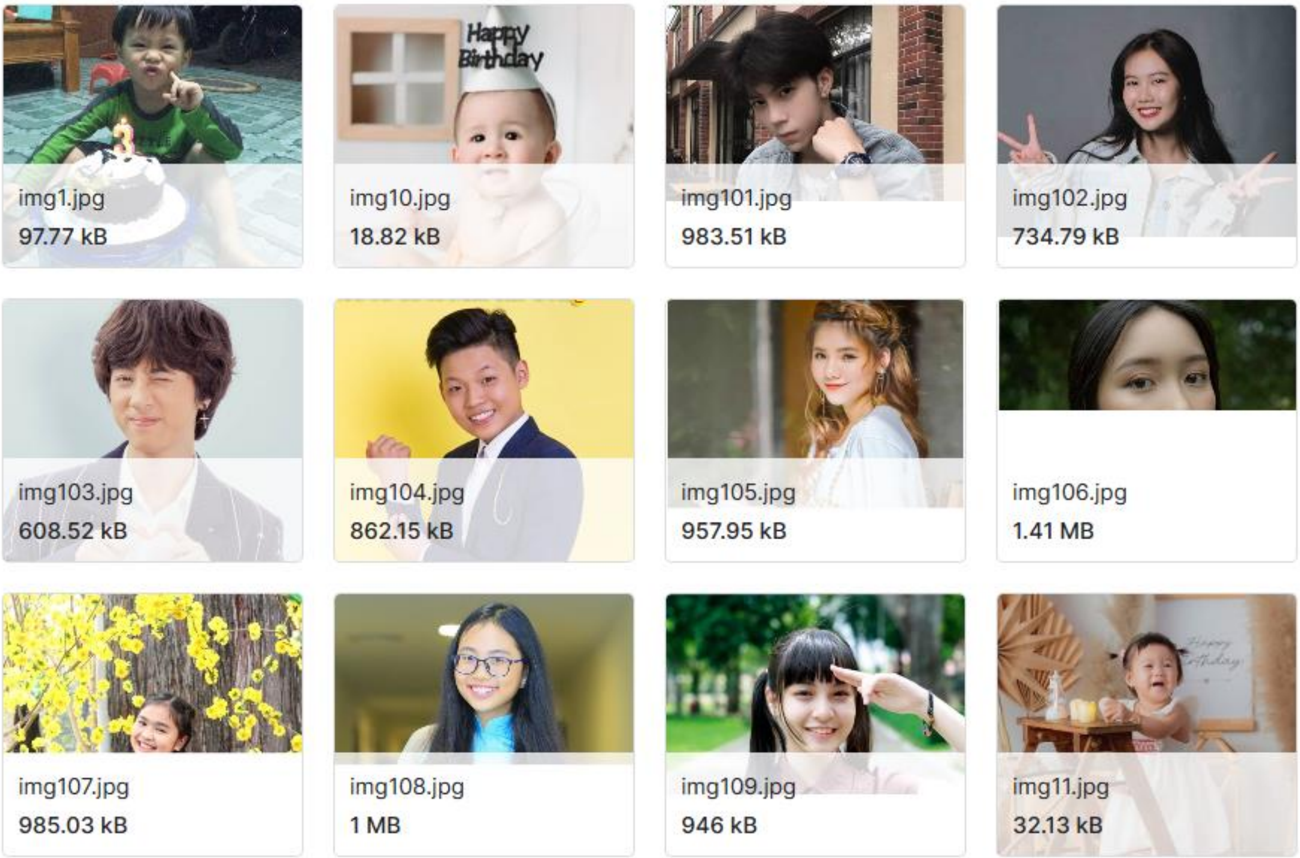


Figure 3. Some sample images of *Diverse Asian Facial Ages* dataset

.

We will use this dataset to train the age model so that it fits better to the data in Vietnam.

Link to the dataset: *https://www.kaggle.com/datasets/leewanhung/diverse-asian-facial-ages*

4.1.3. Emotion Detection dataset (kaggle.com)

The dataset contain 35,685 examples of 48x48 pixel gray scale images of faces divided into train and test dataset. Images are categorized based on the emotion shown in the facial expressions (happiness, neutral, sadness, anger, surprise, disgust, fear).

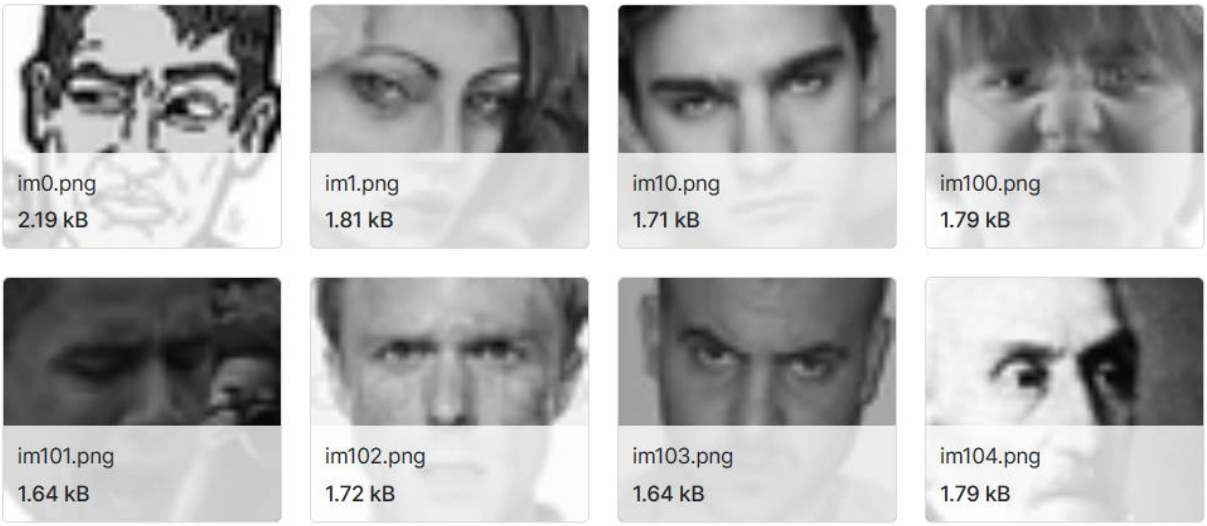


Figure 4. Some sample images of the *Emotion Detection* dataset

We will use this dataset to train the Emotion model.

Link to the dataset: *https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer*

4.2. Project Methods:

4.2.1. Age, Gender and Emotion Prediction

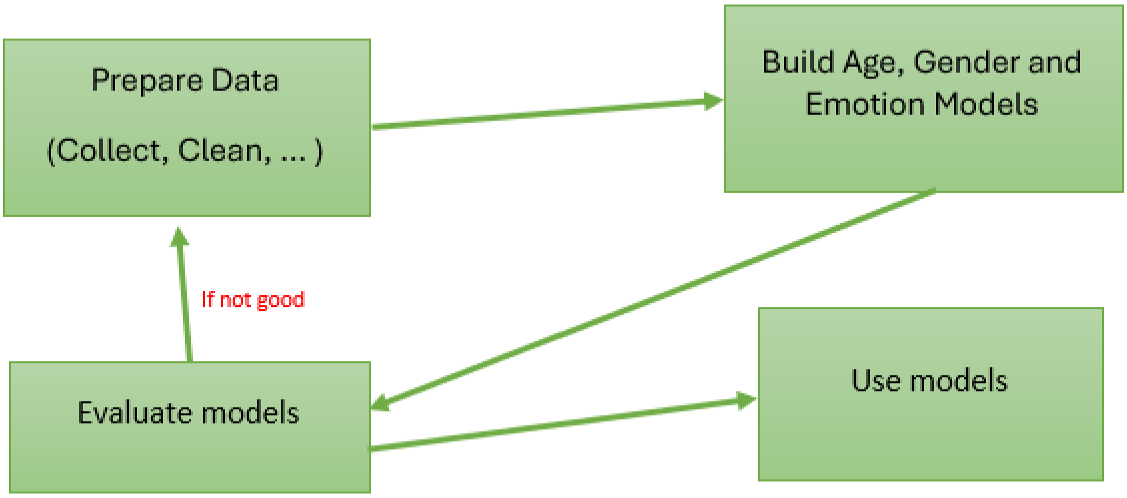


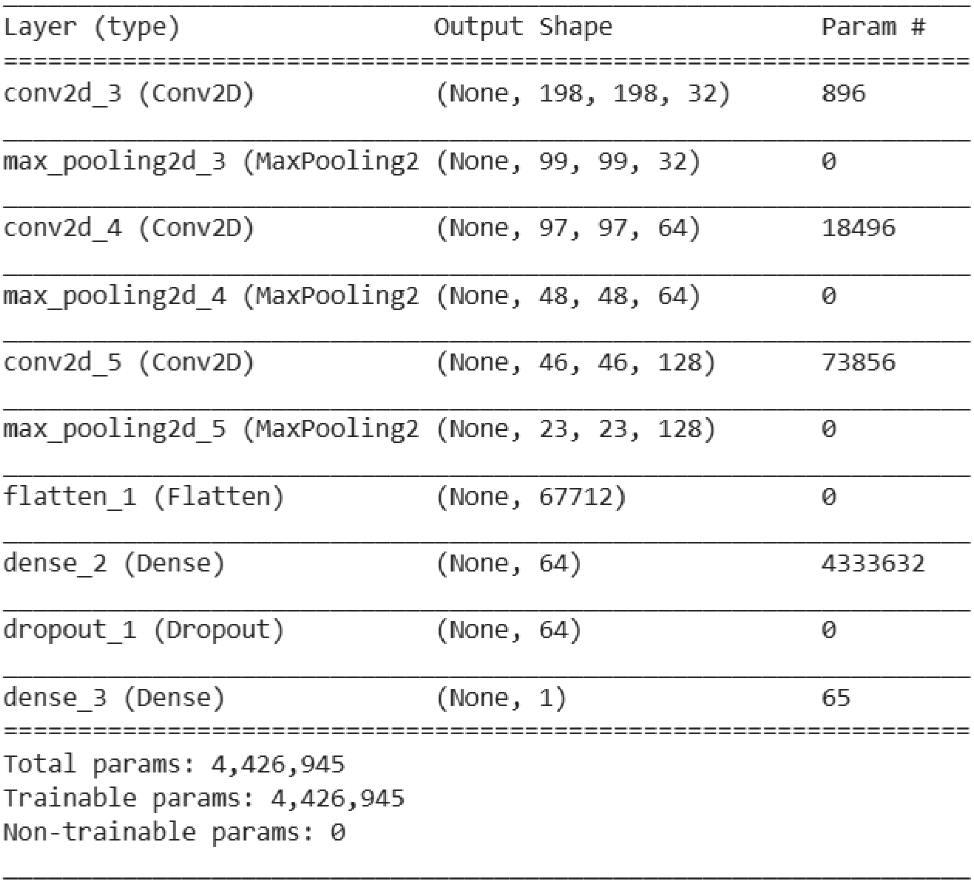
Figure 5. Project Workflow *(part 1)*

In this project, we will utilize various CNN architecture models alongside specific layers, and employ transfer learning with pretrained models such as VGG16 to enhance performance. However, in this report, we will only present the best models.

❖ *Train Gender prediction model:*

1. <u>Load and preprocess the UTKFace dataset</u>**.**
   - We've observed that some records have a gender value of 3, which likely indicates a typing error. Therefore, we will remove these records and retain only those with gender values of 0 and 1.
   - Split data to the training and testing sets with test size of 20%.
   - Augment the dataset using various techniques: rescaling by a factor of 1/255, applying width shift ranging within 0.1, height shift ranging within 0.1, and incorporating horizontal flipping.
2. <u>Build the model using a CNN architecture with specified layers.</u>

```
genmodel_try = Sequential()
genmodel_try.add(Conv2D(filters=20, kernel_size=(3, 3), activation='relu',
    input_shape=(200, 200, 3)))
genmodel_try.add(MaxPooling2D(pool_size=(2, 2)))
genmodel_try.add(Conv2D(filters=50, kernel_size=(5, 5), activation='tanh'))
genmodel_try.add(MaxPooling2D(pool_size=(4, 4)))
genmodel_try.add(Conv2D(filters=200, kernel_size=(3, 3), activation='tanh'))
genmodel_try.add(MaxPooling2D(pool_size=(2, 2)))
genmodel_try.add(Flatten())
genmodel_try.add(Dense(units=50, activation='relu'))
genmodel_try.add(Dropout(rate=0.4))
genmodel_try.add(Dense(units=1, activation='sigmoid'))
```



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_3 (Conv2D)            (None, 198, 198, 32)      896
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 99, 99, 32)        0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 97, 97, 64)        18496
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 48, 48, 64)        0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 46, 46, 128)       73856
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 23, 23, 128)       0
_________________________________________________________________
flatten_1 (Flatten)          (None, 67712)             0
_________________________________________________________________
dense_2 (Dense)              (None, 64)                4333632
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 65
=================================================================
Total params: 4,426,945
Trainable params: 4,426,945
Non-trainable params: 0
_________________________________________________________________
```

Figure 6. Summary of the Gender prediction model

Train the model for 50 epochs and implement early stopping if it achieves an accuracy of 90%.

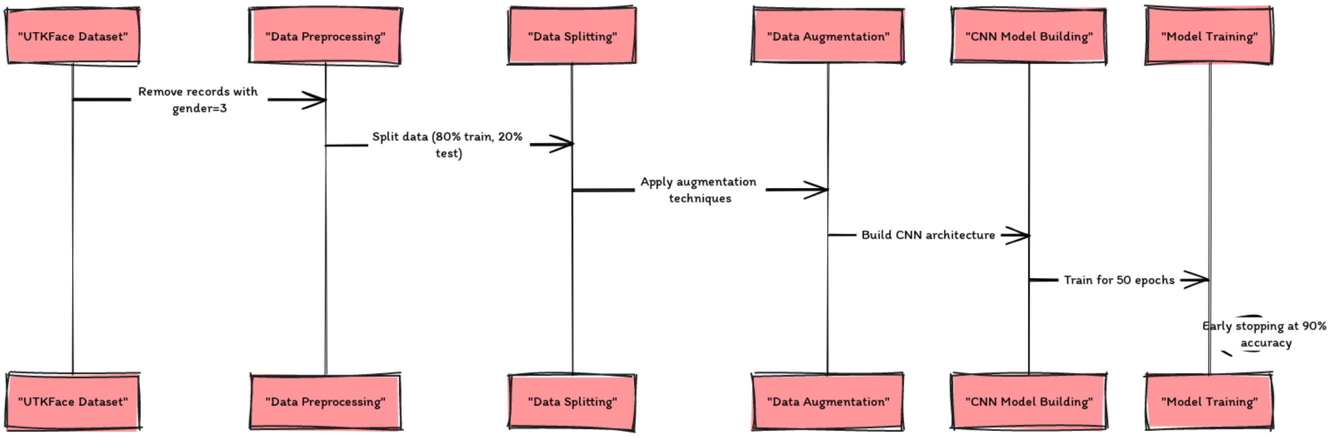
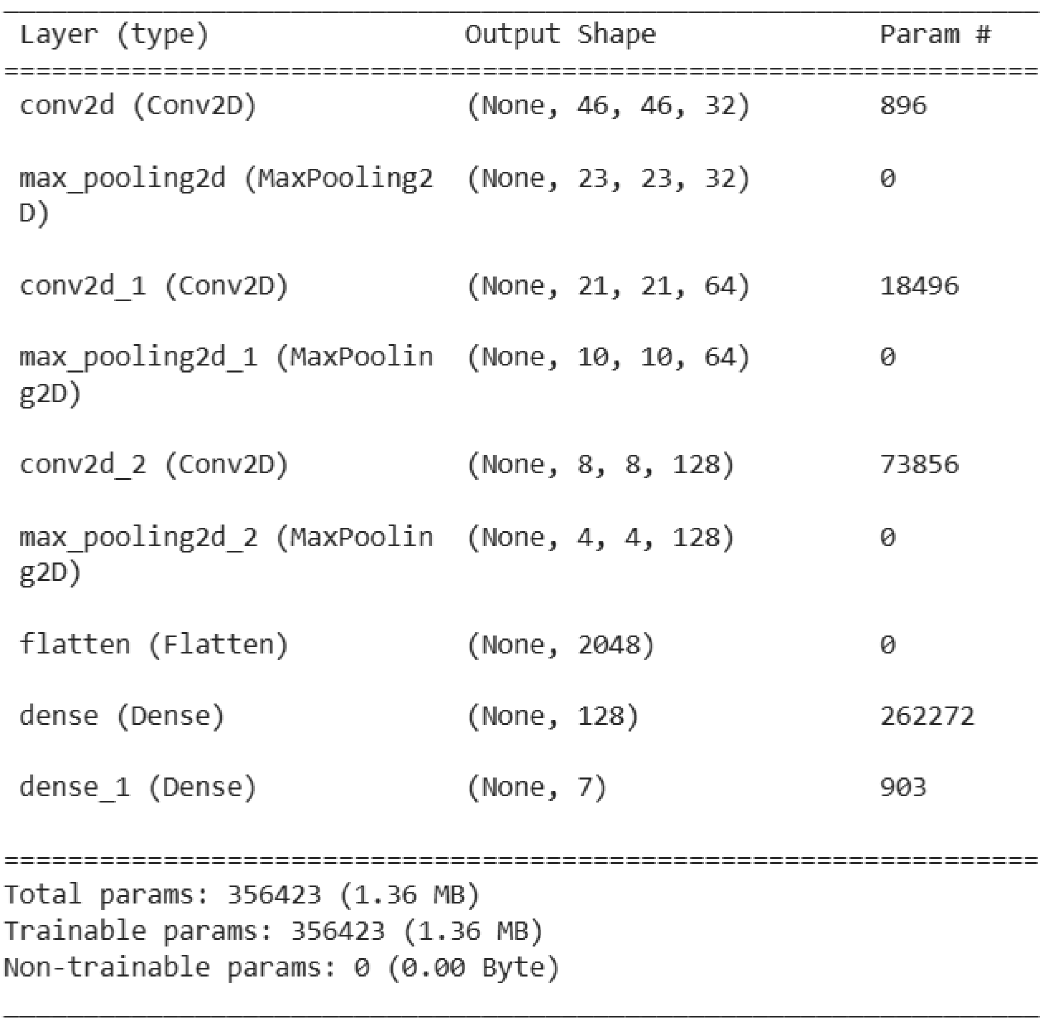


Figure 7. Gender Prediction Model Training Workflow

❖ *Train Emotion prediction model:*

1. Load the Emotion Detection dataset.
2. Build the model using a CNN architecture with specified layers.

```
model = tensorflow.keras.models.Sequential([
  # Convolutional Layer 1
  layers.Conv2D(filters=32, kernel_size=(3, 3), activation='relu', input_shape=(48, 48,
            3)),
  layers.MaxPooling2D(pool_size=(2, 2)),
  # Convolutional Layer 2
  layers.Conv2D(filters=64, kernel_size=(3, 3), activation='relu'),
  layers.MaxPooling2D(pool_size=(2, 2)),
  # Convolutional Layer 3
  layers.Conv2D(filters=128, kernel_size=(3, 3), activation='relu'),
  layers.MaxPooling2D(pool_size=(2, 2)),
  # Flatten the output for the fully connected layers
  layers.Flatten(),
  # Fully Connected Layer 1
  layers.Dense(units=128, activation='relu'),
  # Output Layer
  layers.Dense(units=7, activation='softmax')
])
```



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 46, 46, 32)        896

 max_pooling2d (MaxPooling2  (None, 23, 23, 32)        0
 D)

 conv2d_1 (Conv2D)           (None, 21, 21, 64)        18496

 max_pooling2d_1 (MaxPoolin  (None, 10, 10, 64)        0
 g2D)

 conv2d_2 (Conv2D)           (None, 8, 8, 128)         73856

 max_pooling2d_2 (MaxPoolin  (None, 4, 4, 128)         0
 g2D)

 flatten (Flatten)           (None, 2048)              0

 dense (Dense)               (None, 128)               262272

 dense_1 (Dense)             (None, 7)                 903

=================================================================
Total params: 356423 (1.36 MB)
Trainable params: 356423 (1.36 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 8. Summary of the Emotion prediction model

Train the model for 50 epochs and implement early stopping if it achieves an accuracy of 90%.

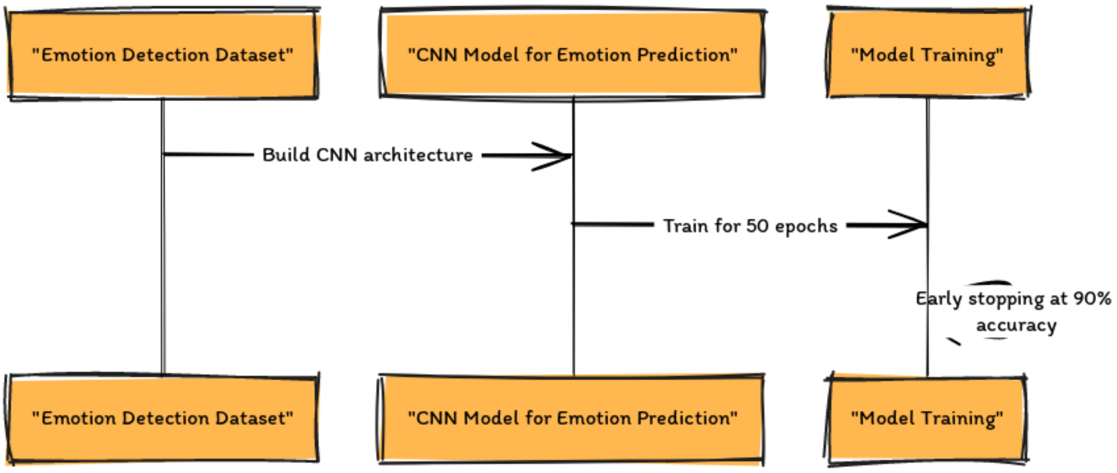

"Emotion Detection Dataset"
"CNN Model for Emotion Prediction"
"Model Training"
Build CNN architecture
Train for 50 epochs
Early stopping at 90% accuracy
"Emotion Detection Dataset"
"CNN Model for Emotion Prediction"
"Model Training"

Figure 9. Emotion Prediction Model Training Workflow

❖ *Train Age prediction model:*

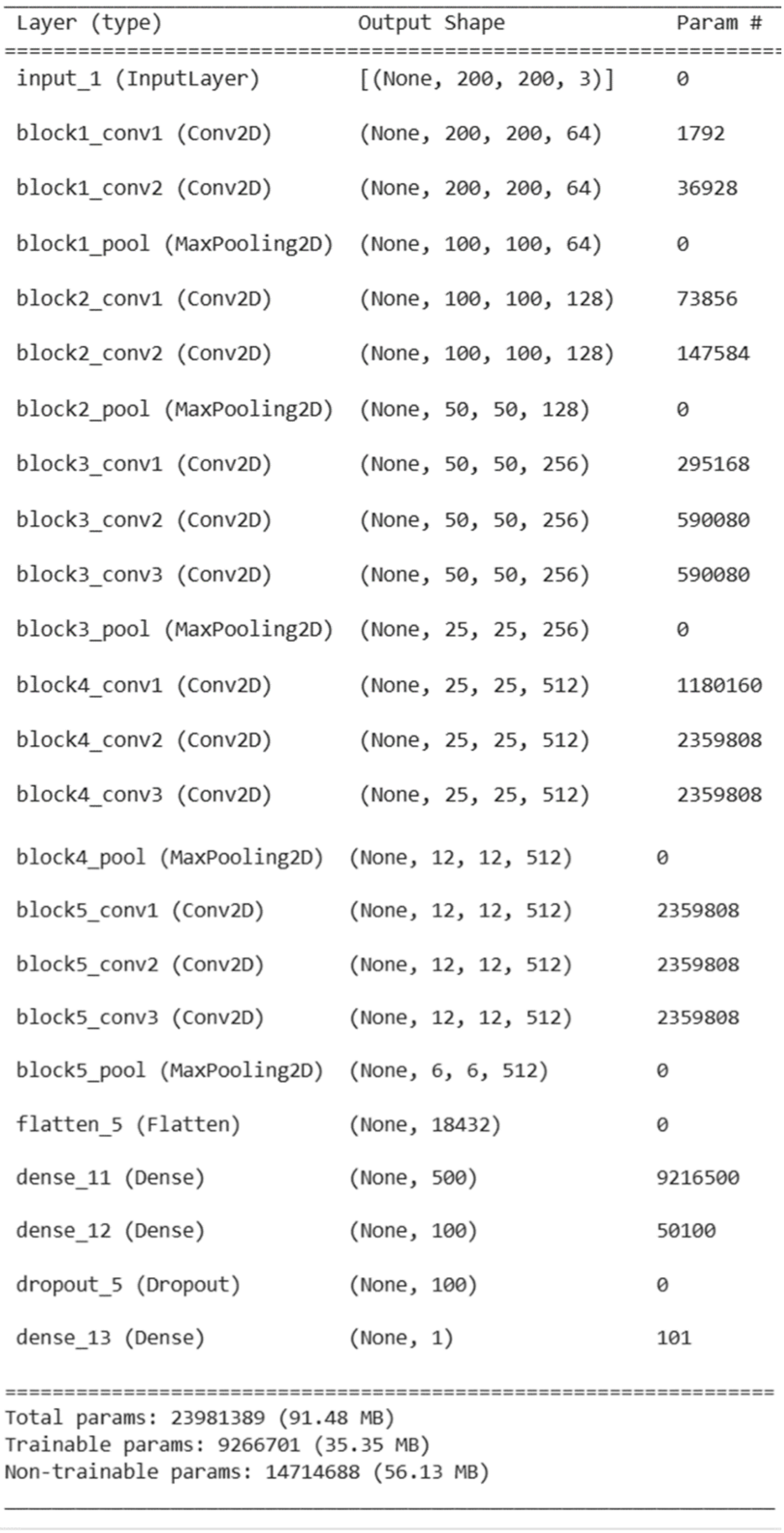
1. Load and preprocess the Diverse Asian Facial Ages dataset.

- Convert the data type of the 'gender' column to numeric by replacing all occurrences of 'boy' or 'Boy' with 0, and 'girl' or 'Girl' with 1.
- Detect and crop the face from each image using yoloface library. If an image contains more than one face, extract the main face (the largest one).

2. Build the model using transfer learning techniques with the pretrained VGG16 model, and then add additional CNN layers

```
vgg16 = VGG16(input_shape=(200, 200, 3), weights='imagenet', include_top=False)
for layer in vgg16.layers:
        layer.trainable = False
output = Flatten()(vgg16.output)
output = Dense(500, activation='relu')(output)
output = Dense(100, activation='relu')(output)
output = Dropout(0.5)(output)
output = Dense(1, activation='relu')(output)

agemodel_vgg16 = Model(inputs=vgg16.input, outputs=output)
```



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 200, 200, 3)]     0

 block1_conv1 (Conv2D)       (None, 200, 200, 64)      1792

 block1_conv2 (Conv2D)       (None, 200, 200, 64)      36928

 block1_pool (MaxPooling2D)  (None, 100, 100, 64)      0

 block2_conv1 (Conv2D)       (None, 100, 100, 128)     73856

 block2_conv2 (Conv2D)       (None, 100, 100, 128)     147584

 block2_pool (MaxPooling2D)  (None, 50, 50, 128)       0

 block3_conv1 (Conv2D)       (None, 50, 50, 256)       295168

 block3_conv2 (Conv2D)       (None, 50, 50, 256)       590080

 block3_conv3 (Conv2D)       (None, 50, 50, 256)       590080

 block3_pool (MaxPooling2D)  (None, 25, 25, 256)       0

 block4_conv1 (Conv2D)       (None, 25, 25, 512)       1180160

 block4_conv2 (Conv2D)       (None, 25, 25, 512)       2359808

 block4_conv3 (Conv2D)       (None, 25, 25, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 12, 12, 512)       0

 block5_conv1 (Conv2D)       (None, 12, 12, 512)       2359808

 block5_conv2 (Conv2D)       (None, 12, 12, 512)       2359808

 block5_conv3 (Conv2D)       (None, 12, 12, 512)       2359808

 block5_pool (MaxPooling2D)  (None, 6, 6, 512)         0

 flatten_5 (Flatten)         (None, 18432)             0

 dense_11 (Dense)            (None, 500)               9216500

 dense_12 (Dense)            (None, 100)               50100

 dropout_5 (Dropout)         (None, 100)               0

 dense_13 (Dense)            (None, 1)                 101

=================================================================
Total params: 23981389 (91.48 MB)
Trainable params: 9266701 (35.35 MB)
Non-trainable params: 14714688 (56.13 MB)
_________________________________________________________________
```

Figure 10. Summary of the Age prediction model

Train the model for 50 epochs and implement early stopping if it achieves an accuracy of 90%.

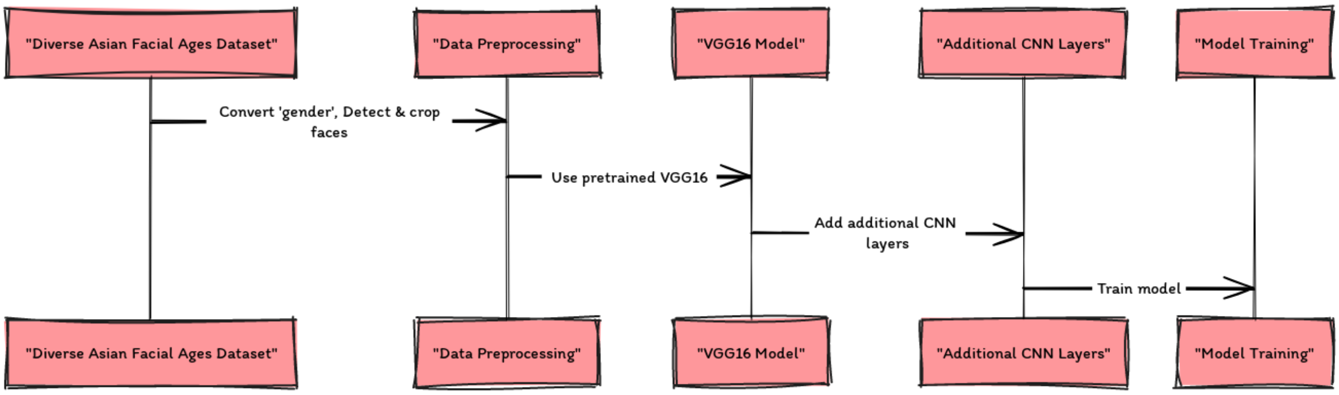


Figure 11. Age Prediction Model Training Workflow

4.2.2. Supervise access to tourist areas and industrial parks

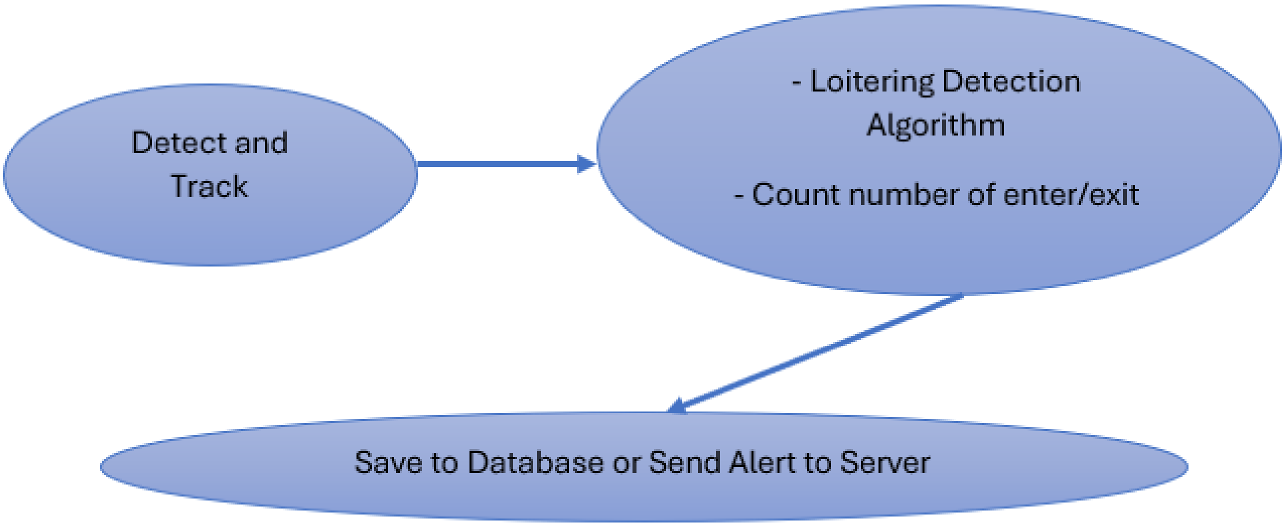

Detect and Track
- Loitering Detection Algorithm
- Count number of enter/exit
Save to Database or Send Alert to Server

Figure 12. Project Workflow *(part 2)*

❖ *Loitering Detection:*

1. Detect and Track people using YOLOv8 model.
2. Check loitering behavior.

- **Definition of loitering behavior:** someone is considered to have loitering behavior if he or she stays too long in the monitored area and moves (the act of standing still, possibly waiting for relatives or waiting for a bus. .., considered non-loitering).
- **Loitering Detection algorithm:**

Set:

min_time = minimum time threshold

min_distance = minimum total travel distance threshold.

Initialize:

tracking_list = {}

At each frame, detect and track people appearing in the current frame.

For each person appearing in the current frame:

1) If the object is not in the tracking_list, add that person's information, including:

   {start_time: starting time of appearance,

   positions: list of locations of that person's movement history, initialized by his/her current position: [bbox, bbox]}

   - *return False* - this person is not loitering.

2) Else, if the object is already in the tracking_list:

   update the person's positions (add the current position, delete the first element in the list if the list length exceeds the maximum threshold - the maximum length threshold is determined by min_time * fps_tracking, where fps_tracking is the number of frames per second detected by the model and tracked objects).

3) Check 2 conditions:

- current time – object's start_time > min_time.
- total_distance > min_distance,

  (whereas total_distance =

  sum([distance(position[i], position[i+1])

  for i in range(len(positions) - 1)])

If the above two conditions are satisfied simultaneously:

➢ *return True:* that person is loitering.

Else:

➢ *return False:* the person is not loitering.

4) After warning about loitering objects, if the management department knows that and wants to delete the warning, the program will delete data about that object and re-initialize:

   start_time[person_id] = current_time;

   positions[person_id] = [current_position, current_position]

   or just simple:

   tracking_list.pop(person_id)

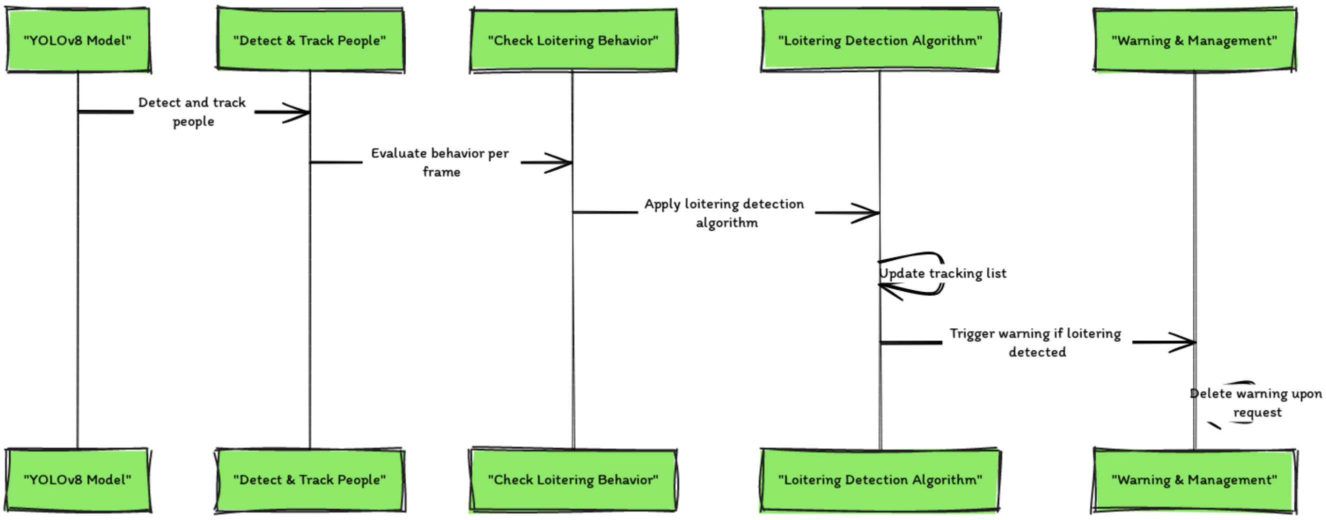

"YOLOv8 Model"
"Detect & Track People"
"Check Loitering Behavior"
"Loitering Detection Algorithm"
"Warning & Management"
Detect and track people
Evaluate behavior per frame
Apply loitering detection algorithm
Update tracking list
Trigger warning if loitering detected
Delete warning upon request
"YOLOv8 Model"
"Detect & Track People"
"Check Loitering Behavior"
"Loitering Detection Algorithm"
"Warning & Management"

Figure 13. Loitering Detection Workflow

❖ *Counting people entering and exiting:*

1. Detect and Track people using YOLOv8 model.
2. Count number of people entering and exiting.

Set:

entry_line = [(start_point), (end_point)]

sample_inside_point = (sample_x, sample_y)

exit_line = [(start_point), (end_point)]

sample_outside_point = (sample_x, sample_y)

Initialize:

tracking_list = {person_id: positions[]}

went_in = set()

went_out = set()

At each frame, detect and track people appearing in the current frame.

going_in, going_out = [], []

For each person appearing in the current frame:

1) If the person is not in tracking_list: add that person position:

tracking_list[person_id] = [current_position, current_position]

```
        then, go to the next person
        continue
    2)  Else: get current and previous positions
        current_position = tracking_list[person_id][-1]
        prev_position = tracking_list[person_id][-2]
    3)  Check entering by check these 2 conditions:
        If:
            The curren_position and the prev_position are NOT on the same
            side of the entry_line
        And:
            The current_position and the sample_inside_point are on the same
            side of the entry_line
        Then, the person is entering. So, update the set:
            went_in.add(person_id)
            going_in.append(person_id)
    4)  Check exiting:
        If:
            The curren_position and the prev_position are NOT on the same
            side of the exit_line
        And:
            The current_position and the sample_outside_point are on the same
            side of the exit_line
        Then, the person is exiting. So, update the set:
            went_out.add(person_id)
            going_out.append(person_id)
Show the number of people going, going out and went_in, went_out.
return len(went_in), len(went_out), going_in, going_out
```

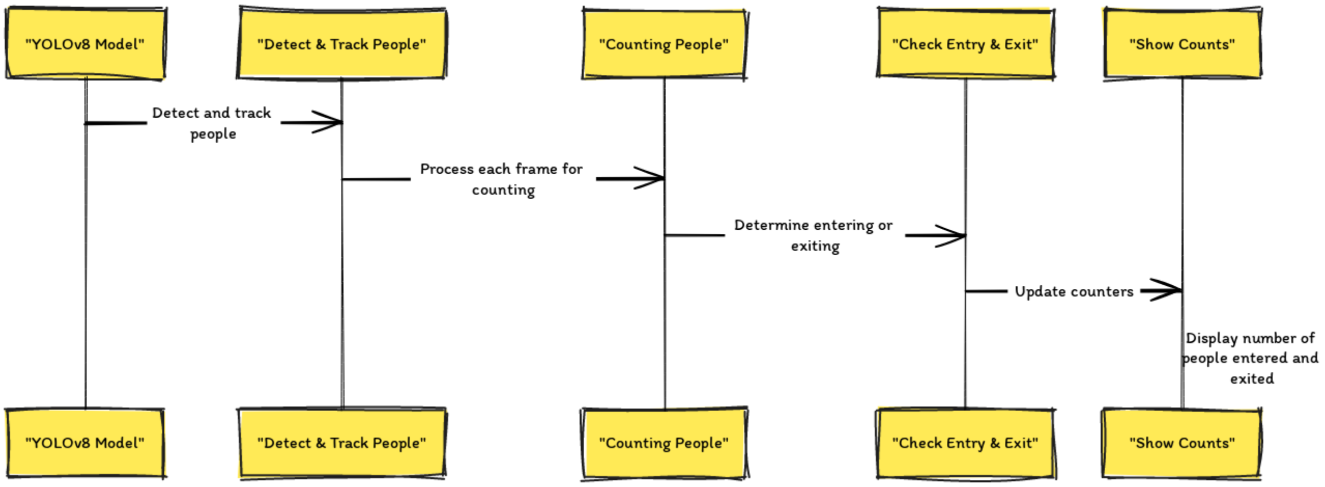


Figure 14. People Counting Workflow

5. RESULTS

In this project, we have implemented many different model architectures, the following is the training performance of the models. The best models in this report are highlighted accordingly.

5.1. Emotion, Gender and Age Prediction

5.1.1. Age prediction models

❖ *Trained on the Diverse Asian Facial Ages dataset:*

➢ *Model 1:* using CNN architecture with specified layers.

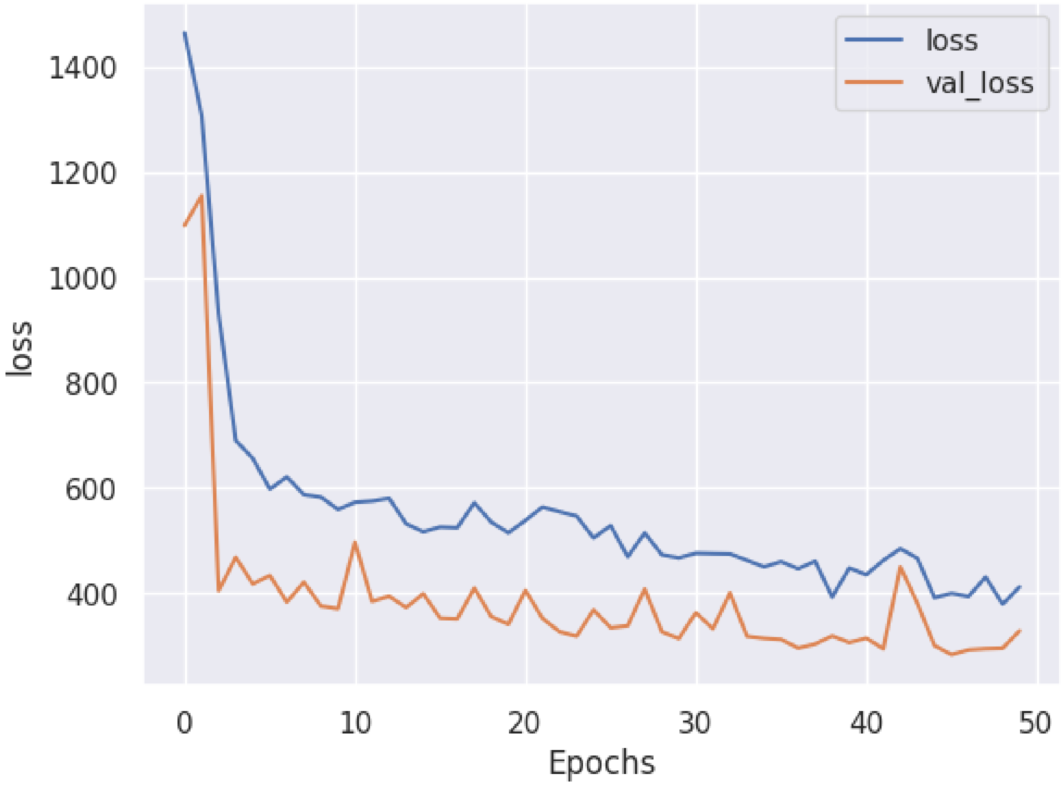


Figure 15. Performance of the Age prediction models: *Model 1*

- *Model 2* **(the optimal model suited to the data in Vietnam):** using transfer learning techniques with the pretrained VGG16 model, and then add additional CNN layers.

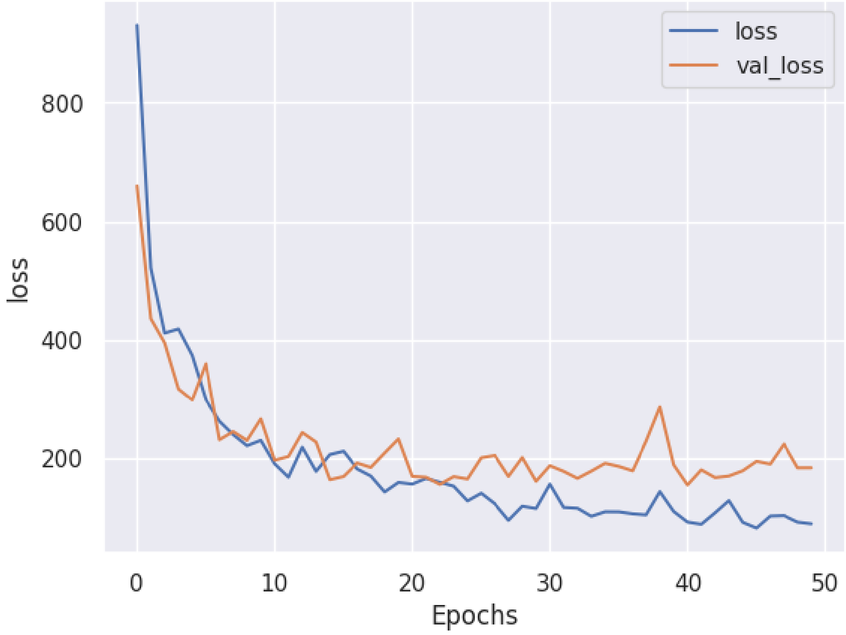


Figure 16. Performance of the Age prediction models: *Model 2*

- *Model 3:* using transfer learning techniques with the pretrained ResNet50 model, and then add additional CNN layers.

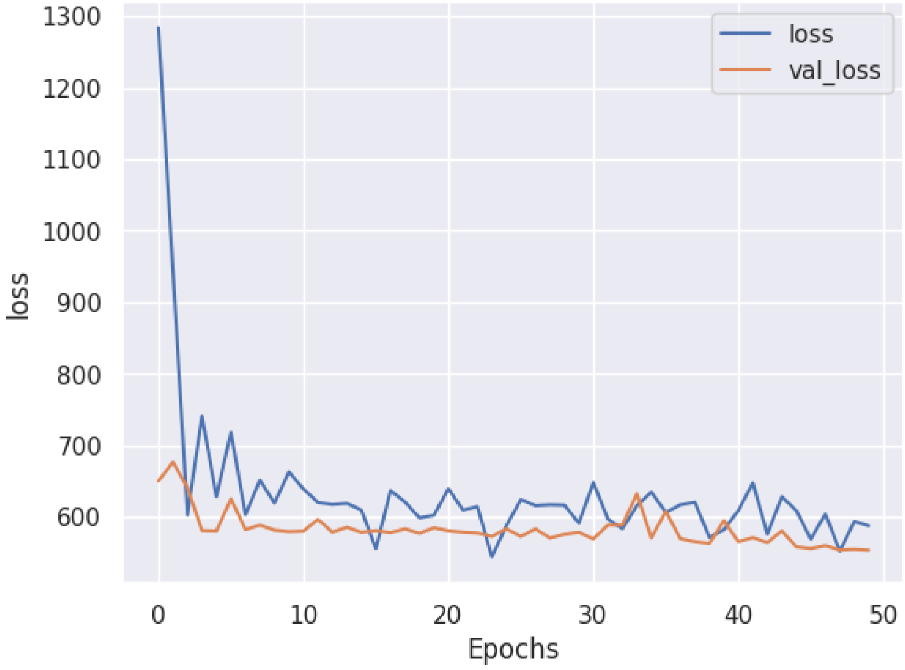


Figure 17. Performance of the Age prediction models: *Model 3*

- *Model 4:* another using CNN architecture with specified layers.

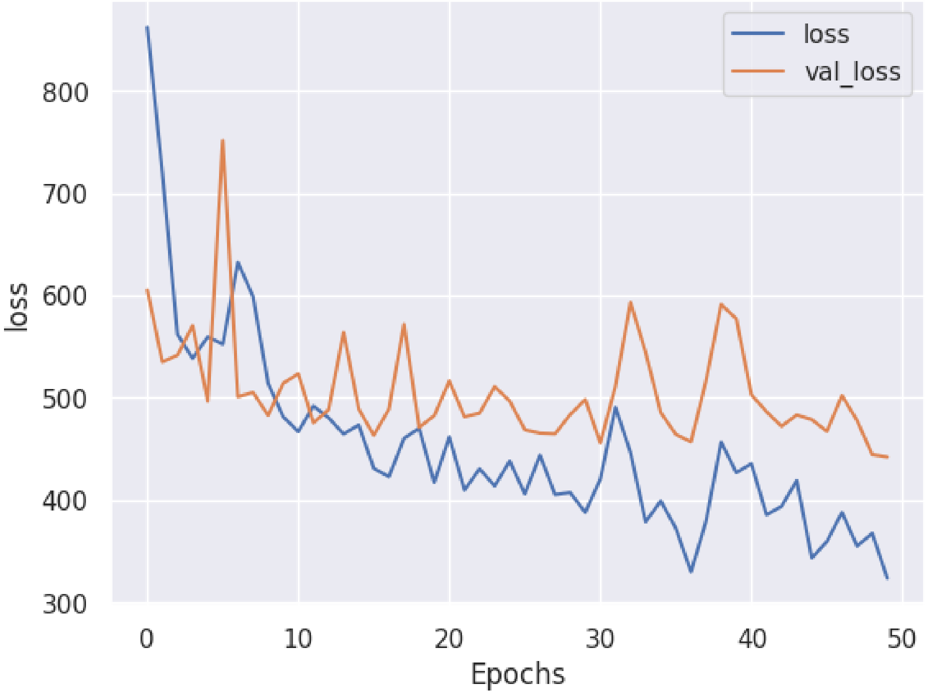
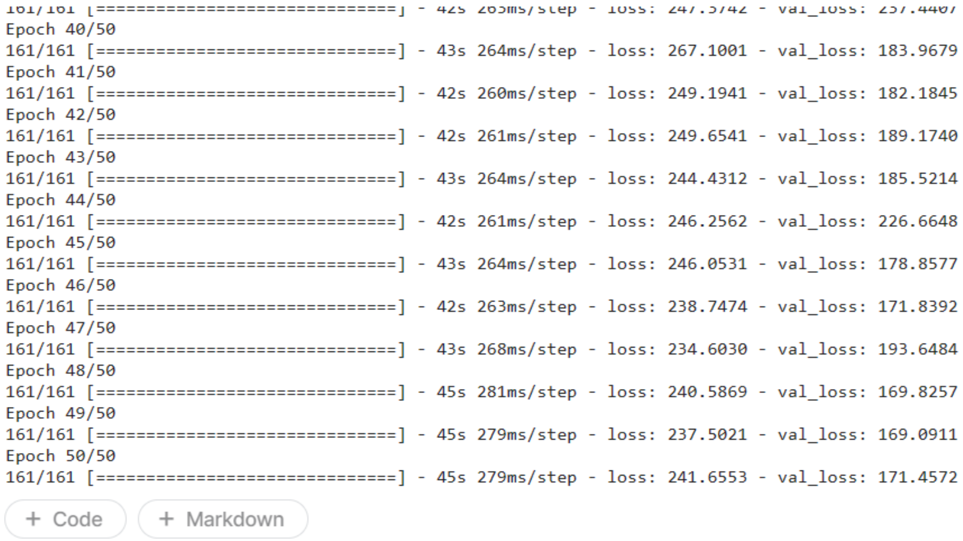


Figure 18. Performance of the Age prediction models: *Model 4*

- *Trained on the UTKFace dataset:*

- *Model 5:* using CNN architecture with specified layers.



```
161/161 [==============================] - 42s 263ms/step - loss: 247.3742 - val_loss: 237.4407
Epoch 40/50
161/161 [==============================] - 43s 264ms/step - loss: 267.1001 - val_loss: 183.9679
Epoch 41/50
161/161 [==============================] - 42s 260ms/step - loss: 249.1941 - val_loss: 182.1845
Epoch 42/50
161/161 [==============================] - 42s 261ms/step - loss: 249.6541 - val_loss: 189.1740
Epoch 43/50
161/161 [==============================] - 43s 264ms/step - loss: 244.4312 - val_loss: 185.5214
Epoch 44/50
161/161 [==============================] - 42s 261ms/step - loss: 246.2562 - val_loss: 226.6648
Epoch 45/50
161/161 [==============================] - 43s 264ms/step - loss: 246.0531 - val_loss: 178.8577
Epoch 46/50
161/161 [==============================] - 42s 263ms/step - loss: 238.7474 - val_loss: 171.8392
Epoch 47/50
161/161 [==============================] - 43s 268ms/step - loss: 234.6030 - val_loss: 193.6484
Epoch 48/50
161/161 [==============================] - 45s 281ms/step - loss: 240.5869 - val_loss: 169.8257
Epoch 49/50
161/161 [==============================] - 45s 279ms/step - loss: 237.5021 - val_loss: 169.0911
Epoch 50/50
161/161 [==============================] - 45s 279ms/step - loss: 241.6553 - val_loss: 171.4572
```

Figure 19. Some historical records of the Age prediction model's performance: *Model 5*

5.1.2. Emotion prediction models

- *Model 1* **(the best model):** using CNN architecture with specified layers.

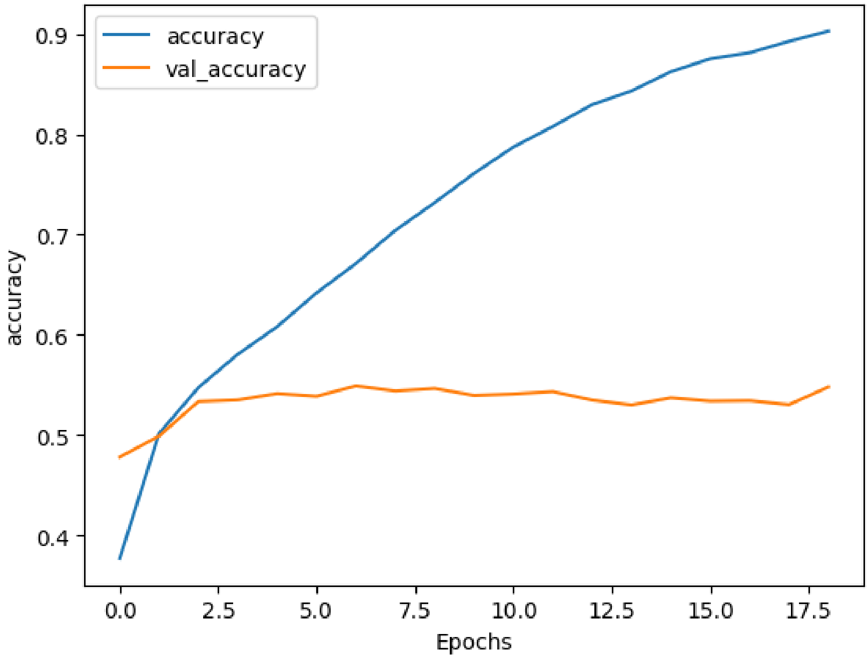


Figure 20. Performance of the Emotion prediction models: *Model 1*

- *Model 2:* using transfer learning techniques with the pretrained VGG16 model, and then add additional CNN layers.

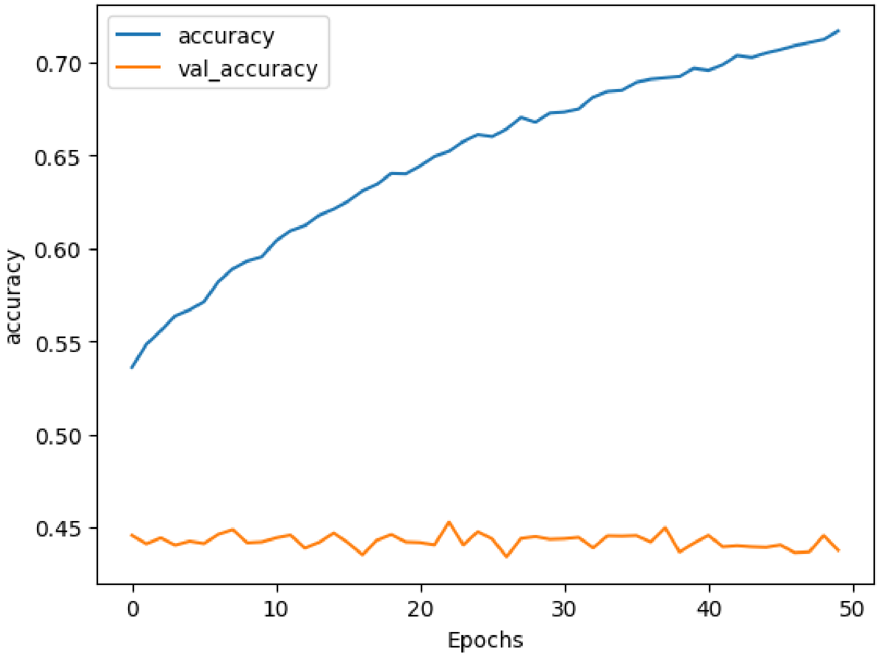


Figure 21. Performance of the Emotion prediction models: *Model 2*

5.1.3. Gender prediction models

❖ *Trained on the UTKFace dataset:*

➢ *Model 1* **(the best model):** using CNN architecture with specified layers.

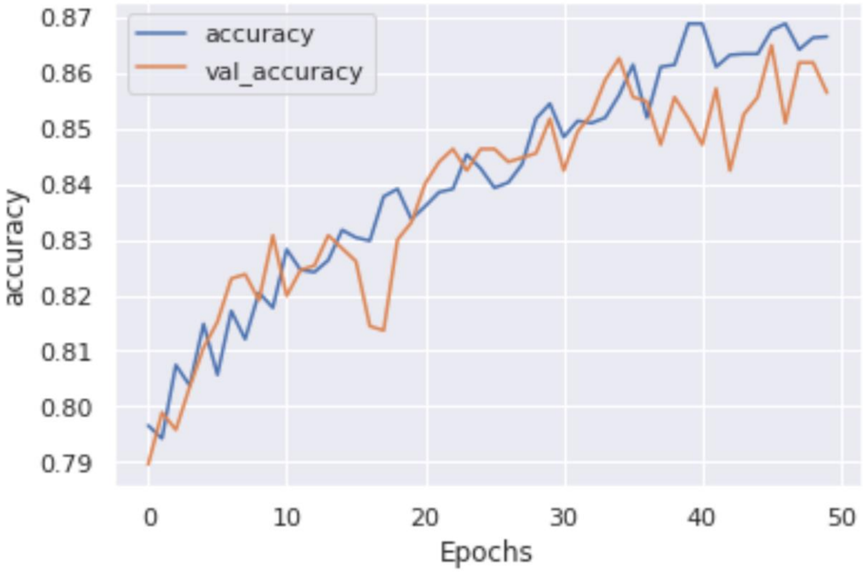


Figure 22. Performance of the Gender prediction models: *Model 1*

❖ *Trained on the Diverse Asian Facial Ages dataset:*

➢ *Model 2:* using CNN architecture with specified layers.

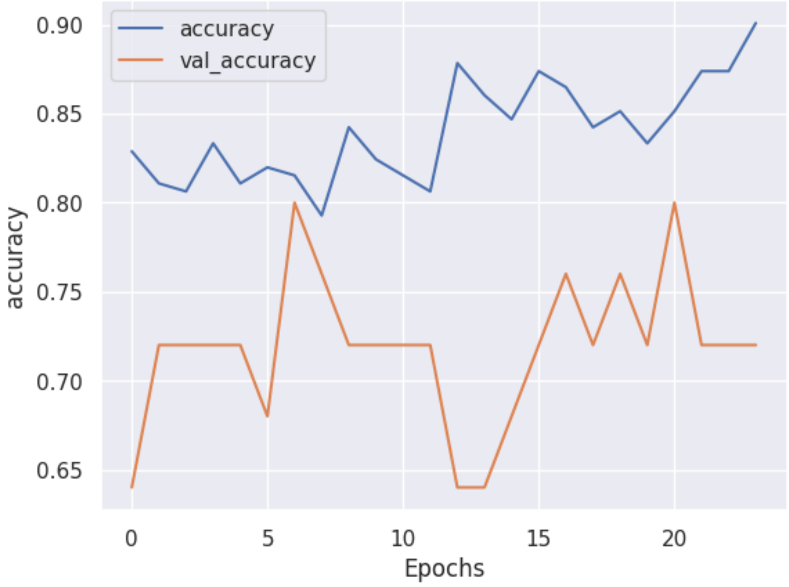


Figure 23. Performance of the Gender prediction models: *Model 2*

- *Model 3:* using transfer learning techniques with the pretrained VGG16 model, and then add additional CNN layers.

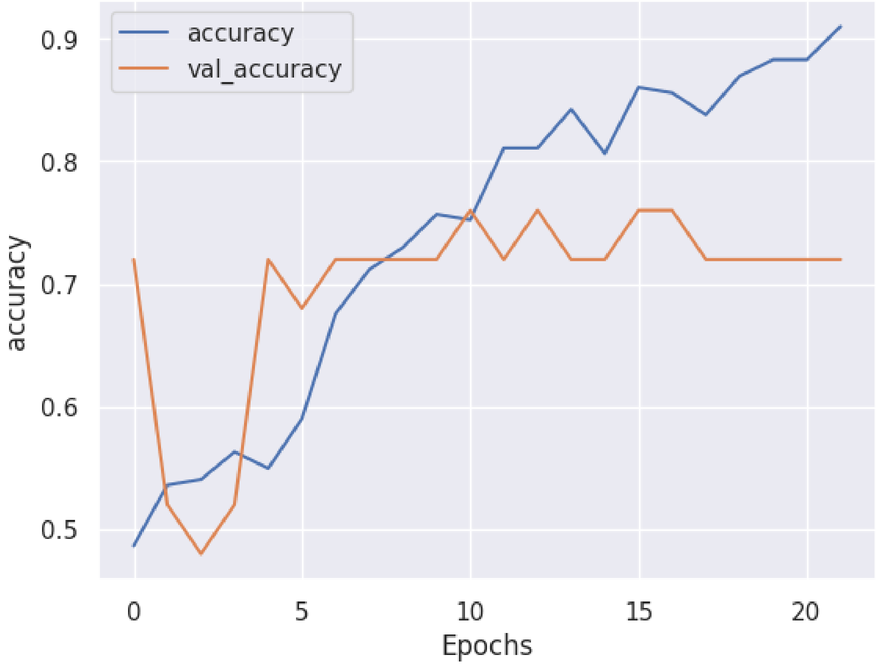


Figure 24. Performance of the Gender prediction models: *Model 3*

- *Model 4:* another using CNN architecture with specified layers.

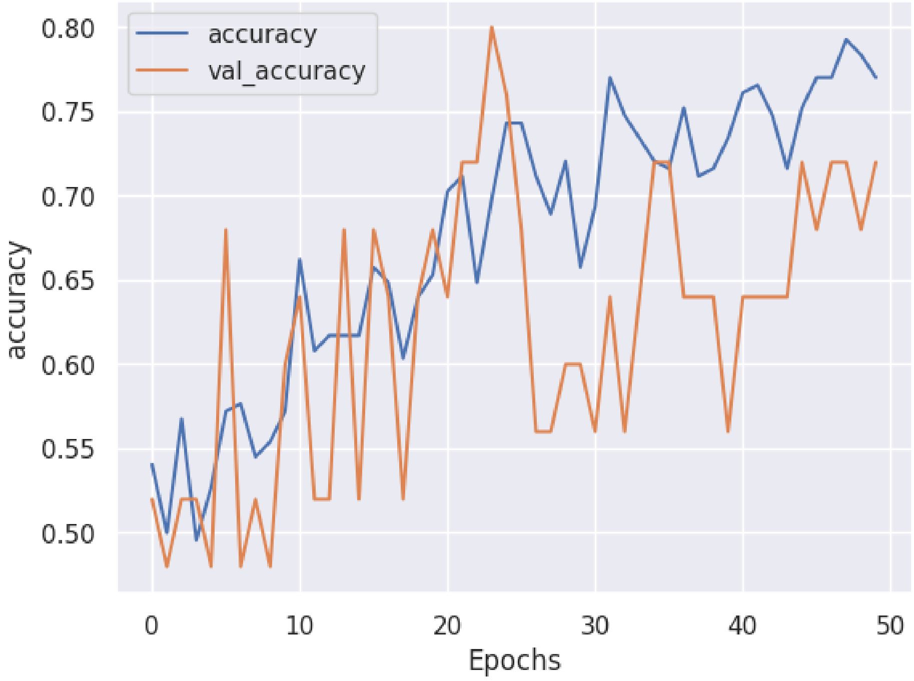


Figure 25. Performance of the Gender prediction models: *Model 4*

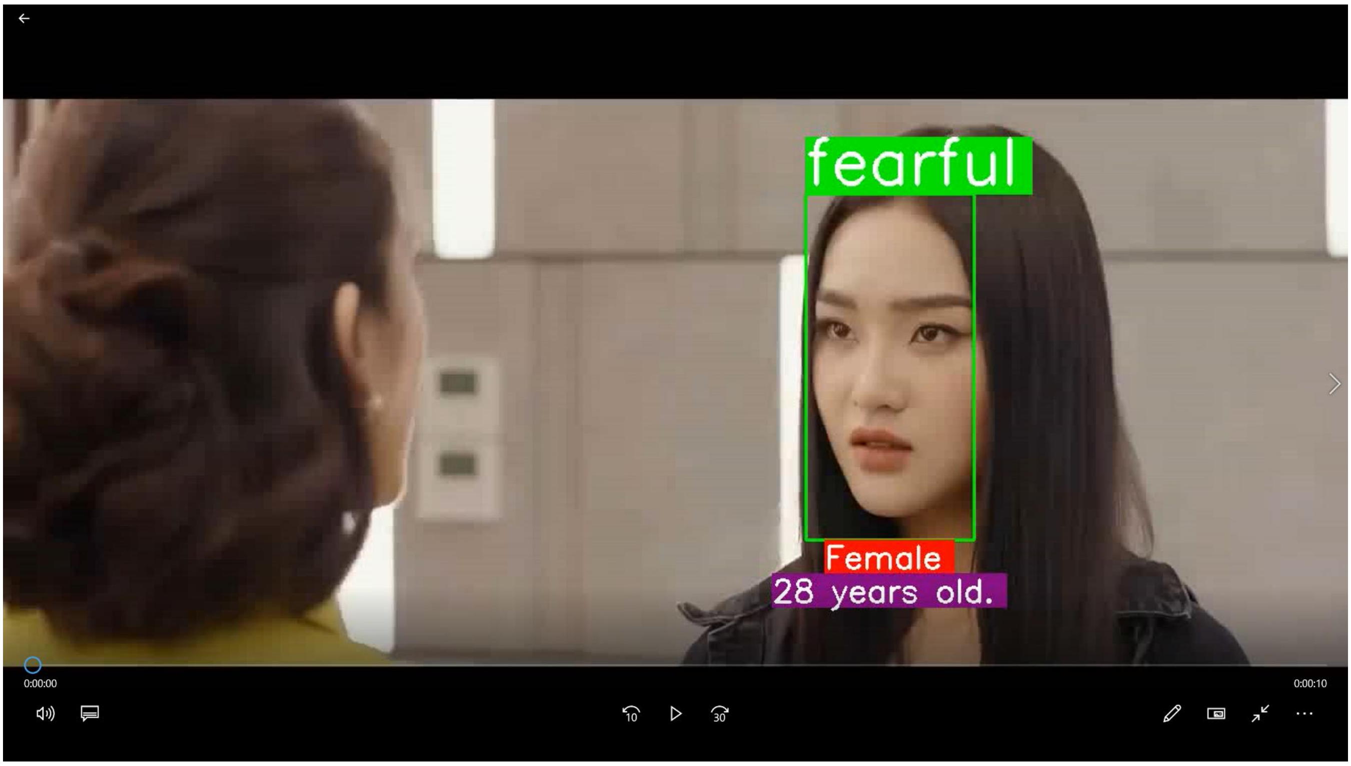


Figure 26. Results from Age, Gender and Emotion prediction models

5.2. Supervise access to tourist areas and industrial parks

5.2.1. Loitering Detection

Watch our demo video here.

(For illustration purposes, we have set the minimum time threshold to 3 seconds and the minimum movement distance to 200 pixels. The red boxes indicate individuals warned for loitering.)

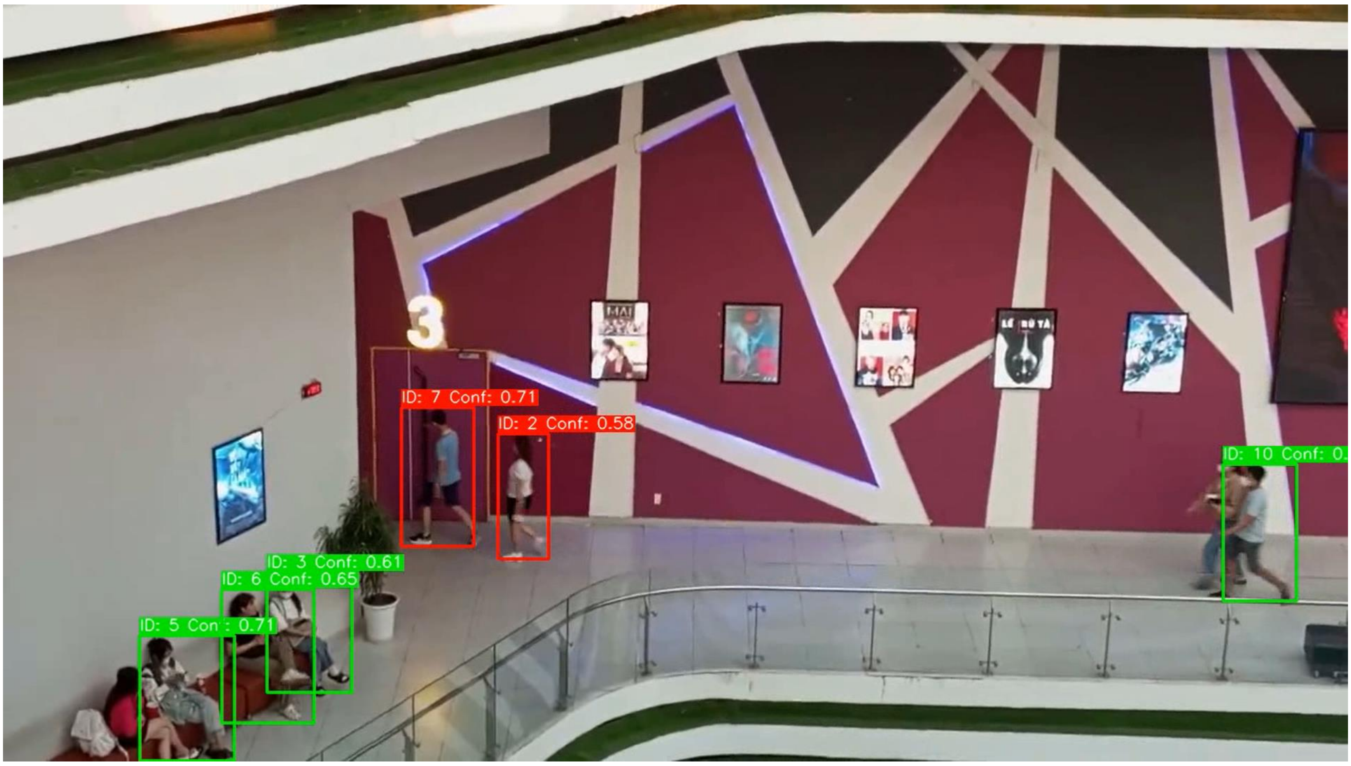


Figure 27. Results of Loitering Detection

Link to the video:

https://drive.google.com/file/d/1RY5tiowwyjsbyf3V-lh7T9InigIDvEDH/view?usp=sharing

5.2.2. Counting people entering and exiting

Watch our demo video **here**.

(For illustrative purposes, the green boxes denote individuals entering, while the purple box signifies exiting. The green line represents the ROI line for entering, and the red line depicts ROI for exiting. The yellow box indicates individuals who have neither entered nor exited.)

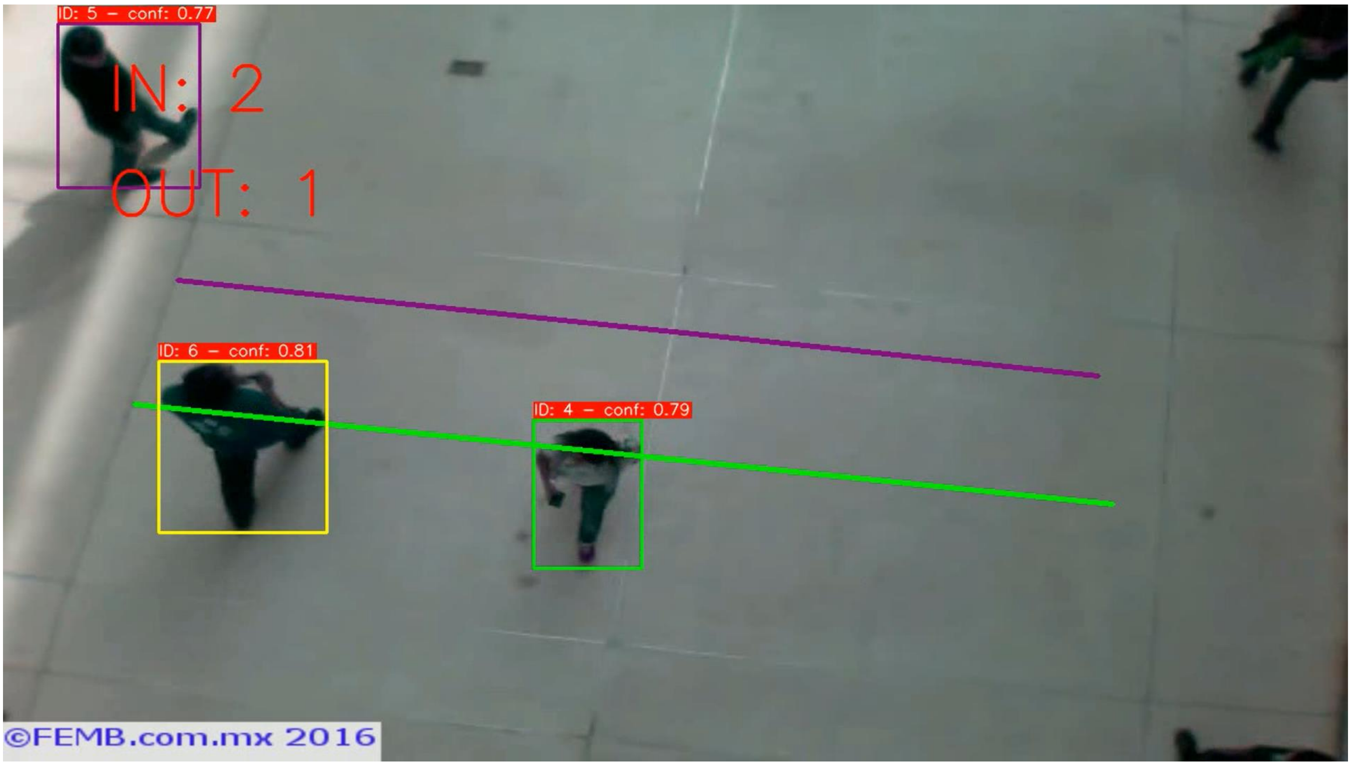


Figure 28. Results of Counting people entering and exiting

Link to the video:

https://drive.google.com/file/d/1z-TgdSh8Ci07E50plvA19GIYZ6a8RVv5/view?usp=sharing

6. DISCUSSION

In our study focusing on age, gender, and emotion prediction, along with loitering detection and people counting at entrance points, we encountered specific contextual nuances that influenced the performance of our models, particularly in the Vietnamese setting. This discussion section aims to elucidate these contextual factors, interpret their implications on our findings, and outline potential avenues for future research.

Interpretation of Findings:

Our loitering detection algorithm demonstrated a strong fit within the Vietnamese context, effectively identifying instances of prolonged presence in entrance areas beyond what is typical for normal foot traffic patterns. This contextual adaptation underscores the importance of tailoring algorithms to specific environmental and cultural contexts, rather than relying solely on generic thresholds for defining loitering behavior.

Conversely, the performance of our age prediction model was hindered by limited training data specific to the Vietnamese population. The scarcity of diverse and representative datasets poses a significant challenge in accurately capturing age-related facial features and variations within this demographic. As a result, the model may exhibit lower accuracy levels and generalizability when applied to Vietnamese individuals.

Implications:

The contextual adaptation of our loitering detection algorithm highlights the importance of considering cultural norms, environmental factors, and behavioral patterns unique to the Vietnamese context. By incorporating such contextual insights into algorithm design and implementation, we can enhance the reliability and effectiveness of surveillance and security systems tailored to local conditions.

However, the limitations encountered with the age prediction model underscore the critical need for expanded data collection efforts within Vietnam. A more extensive and diverse dataset encompassing a broader spectrum of age groups, ethnicities, and facial variations is essential to improve the model's performance and ensure its applicability across different demographic segments within the Vietnamese population.

Future Directions:

In future research endeavors, prioritizing the collection and annotation of data specific to the Vietnamese context should be a primary focus. Collaborative efforts involving local institutions, businesses, and community organizations can facilitate the acquisition of diverse and representative datasets necessary for training and validating age prediction models effectively.

Furthermore, exploring alternative methodologies, such as data augmentation techniques or transfer learning approaches, may offer viable strategies for mitigating the limitations imposed by limited training data. Leveraging pre-trained models or adapting existing algorithms to accommodate the unique facial characteristics prevalent in the Vietnamese population could expedite the development of more robust and accurate age prediction systems.

Conclusion:

In conclusion, our study underscores the significance of context-awareness in the design and implementation of surveillance and demographic analysis systems. While our loitering detection algorithm demonstrated strong alignment with the Vietnamese context, the performance of our age prediction model was hampered by data scarcity. By addressing these contextual nuances and investing in expanded data collection efforts, we can foster the development of more effective and culturally sensitive technologies to meet the evolving needs of diverse populations, particularly in settings such as Vietnam.

7. CONCLUSIONS AND PERSPECTIVES

In summary, our study presents a comprehensive framework for age, gender, and emotion prediction, along with loitering detection and people counting at entrance points, tailored to the Vietnamese context. Our findings highlight the effectiveness of our loitering detection algorithm in identifying anomalous behavior within entrance areas, showcasing its potential for enhancing security and crowd management in diverse settings. However, the performance of our age prediction model was hindered by limited training data specific to the Vietnamese population, emphasizing the critical need for expanded data collection efforts in future research endeavors.

Among the methods employed, our loitering detection algorithm achieved the best performance, leveraging contextual insights to accurately identify instances of prolonged presence beyond typical foot traffic patterns. This success can be attributed to the careful consideration of cultural norms and environmental factors prevalent in the Vietnamese context. Conversely, the suboptimal performance of our age prediction model underscores the challenges posed by data scarcity and highlights the importance of prioritizing data collection efforts to improve model accuracy and generalizability.

Despite the promising results, our study has several limitations and drawbacks, including the reliance on limited training data for age prediction and the potential for algorithmic bias in demographic analysis. If afforded more time, computational resources, or team members, we would propose and explore several avenues for improving performance. These include expanding data collection efforts to encompass a broader spectrum of demographic characteristics, exploring alternative methodologies such as transfer learning to mitigate data scarcity issues, and conducting rigorous validation studies across diverse populations to ensure the reliability and fairness of our algorithms.

In conclusion, our study represents a significant step towards the development of context-aware surveillance and demographic analysis systems tailored to the Vietnamese context. By addressing the identified limitations and leveraging future research opportunities, we can advance the state-of-the-art in security, marketing, and crowd management, ultimately contributing to safer, more efficient, and more inclusive environments for all individuals.

CONFLICTS OF INTEREST:

The authors declare no conflicts of interest. The research reported in this study was conducted in an objective and impartial manner, free from any personal or financial influences that could potentially bias the representation or interpretation of the reported results. Additionally, the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

8. Appendix

No appendix to include. All relevant information and data are presented in the main body of the report.

9. REFERENCES

1. Smith, J.D.; Johnson, A.B. "Predicting Age, Gender, and Emotion from Facial Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*(3), 678-691.
2. Brown, C.; White, D. "Loitering Detection and Crowd Counting Techniques." *In Handbook of Video Surveillance Systems: Challenges and Solutions, 2nd ed.; Smith, J., Johnson, A., Eds.; Springer: New York, NY, USA,* **2015**; *Volume 2*, pp. 154–176.
3. Williams, E.; Davis, F. Crowd Dynamics: Modeling, Analysis and Simulation, 3rd ed.; *CRC Press: Boca Raton, FL, USA,* **2017**; pp. 154–196.
4. Garcia, R.A.; Martinez, S. "Emotion Prediction Using Neural Networks." *Journal of Artificial Intelligence Research* **2023**, accepted.
5. Chen, L.M. (University of California, Berkeley, CA, USA); Kim, Y. (Stanford University, Stanford, CA, USA). *Personal communication,* **2020**.
6. Nguyen, H.T.; Lee, K.S.; Patel, R. "Advanced Techniques for Loitering Detection." *In Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 15-20* November **2022**.
7. Johnson, T.S. Age and Gender Prediction from Facial Features. *Master's Thesis, Stanford University, Stanford, CA, USA,* **2018**.
8. Facial Emotion Recognition Toolkit. *Available online: https://www.facial-emotion-recognition-toolkit.com* (accessed on 4 March **2024**).
9. Arjun Singh, Nishant Rai, Prateek Sharma, Preeti Nagrath and Rachna Jain. *"Age, Gender Prediction and Emotion recognition using Convolutional Neural Network"*, **2021**.
10. Ultralytics Documentation. *Available online: https://docs.ultralytics.com/* (accessed on March 4, **2024**).