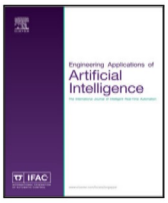
Engineering Applications of Artificial Intelligence 124 (2023) 106637

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



# Unauthorized access detection system to the equipments in a room based on the persons identification by face recognition

Yahya Zennayi [a,b,*], Soukayna Benaissa [b], Hatim Derrouz [b], Zouhair Guennoun [a]

[a] Smart Communications - ERSC Team, E3S Research Center, Mohammadia Engineering School, Mohammed V University, Rabat, Morocco
[b] Embedded System and AI Department, Moroccan foundation for advanced science innovation and research, Rabat, Morocco

ARTICLE INFO

*Keywords:*
Face recognition
System
Identification
Deep learning
Illumination
Implementation real time
Person detection
Face detection

ABSTRACT

Face recognition is a very active research topic due to the number of potential applications that use this technique and the number of challenges that still require efforts to solve them. However, new research problems arise when dealing with devices operating in real-time, and using a non-controlled image source such as surveillance cameras, due to the needs and context of end-use. A robust and efficient face recognition system is proposed in this paper to monitor unauthorized access to equipments in a room using a surveillance camera. The system is based on a person detection algorithm adapted to the context of use, a person tracking algorithm robust to occlusion, a face detection algorithm robust to orientation and scale, and a face recognition algorithm robust to illumination change. To enhance the efficiency of person identification, a novel strategy for estimating the confidence of face recognition is proposed in this paper. This is achieved by leveraging the general conditions of the image used for this purpose. This technique allows to increase the identification rate while minimizing the false alarm rate. An effort is also made to optimize the implementation of the developed algorithms to guarantee a real-time operation while keeping a very high efficiency. Also, the evaluations of the proposed algorithms on public and private data show an improvement of the accuracy on the whole context of use of the proposed system. Furthermore the evaluation of the system in general shows a high efficiency, 100% detection of unauthorized accesses and 0 false alarms.

## 1. Introduction

For safety reasons, buildings nowadays have video surveillance systems installed almost everywhere. These systems are predominantly utilized in a post-processing mode to retrieve evidence after an incident has taken place. In addition, they are employed in real-time for more intricate applications like suspect detection or identifying abnormal behavior in public spaces, as well as for access control purposes utilizing facial recognition (Elharrouss et al., 2021; Cocca et al., 2016).

The increasing value of face recognition technology in the scientific and industrial world, particularly for security purposes, is being driven by its essential applications in today's society. As a result, numerous works have been developed and enhanced to tackle the various contexts of these applications, ranging from optimal controlled conditions to extreme situations. Despite the significant achievements witnessed thus far, reliable results for uncontrolled conditions using video streams from CCTV cameras remain difficult to obtain. The challenges are exacerbated by the need for real-time applications, which require significant computational time.

For several years, the scientific community and industry have focused much attention on developing complete systems for face recognition techniques. This attention has been primarily directed towards access control applications, including smartphones or building access devices in controlled conditions (Wasnik et al., 2017; Ahuja et al., 2017; Olivares-Mercado et al., 2017). While there are other applications for more complex systems, such as students' presence verification in a classroom (Lukas et al., 2016; Arsenovic et al., 2017; Wagh et al., 2015). However, none of them combine the identification of individuals with tracking to detect unauthorized access to equipments in an open space room.

A critical challenge for video surveillance applications is posed by the optimization of the identification processing load in facial recognition systems, particularly to ensure a fast response to real-time events. To improve processing speed, face detection techniques are commonly used to identify only regions of interest in the image that contain faces, thereby reducing the time required to analyze each captured image and improving overall system performance (Lei et al., 2009; Bashbaghi et al., 2019; Ullah et al., 2022). Various methods for enhancing image quality to mitigate issues encountered by surveillance

* Corresponding author at: Smart Communications - ERSC Team, E3S Research Center, Mohammadia Engineering School, Mohammed V University, Rabat, Morocco.
*E-mail address:* zennayi.yahya@gmail.com (Y. Zennayi).

https://doi.org/10.1016/j.engappai.2023.106637
Received 18 September 2022; Received in revised form 23 March 2023; Accepted 8 June 2023
Available online 28 June 2023

cameras have been suggested in several pieces of literature (Khan et al., 2022; Muhammad et al., 2021; Khan et al., 2020). Additionally, some techniques include applying image pre-processing techniques to rectify difficulties related to unregulated settings, such as lighting and orientation variations that can affect the accuracy of facial recognition (Oh et al., 2013; Wiliem et al., 2007). The performance of facial recognition systems is aimed to be improved, and false alarms are minimized by these techniques.

An approach that optimizes the use of the detected facial images, while taking into account image quality and general conditions to optimize processing time while maintaining the same level of performance, has not been presented in any existing literature. To achieve this goal, a new approach is presented in this work to reduce the number of images subject to facial recognition operations by prioritizing images with favorable conditions such as lighting, size, and orientation. By adopting this approach, false alarms can be minimized while ensuring a high identification rate for individuals, all while optimizing processing load.

An innovative system is presented in this work, which utilizes a CCTV camera to detect and identify instances of unauthorized access to equipments within an open space. The system incorporates several optimized algorithms, including person detection, person tracking designed to overcome occlusion issues, face detection that accounts for orientation, face recognition capable of withstanding uncontrolled conditions, person identification updated over a temporal sequence, and decision support through alert notifications. This paper introduces several significant contributions, including:

- An early warning system that utilizes face identification to detect unwanted access by individuals.
- Two benchmarking studies that evaluate the performance and computation time of person detection and face detection techniques.
- An enhanced face recognition algorithm that employs a fine-tuned feature extraction pattern on a database of face images with varying lighting conditions, resulting in improved recognition rates.
- A proposed confidence estimation algorithm for person identification that considers the quality of the images used in the identification process, aiming to reduce both non-identification and false alarm rates.
- A complete optimized algorithm for the real-time alert system, which assists supervisors in anticipating instances of unauthorized access.

The rest of the paper is organized as follows: Section 2 provides an overview of systems based on face recognition techniques. Section 3 presents the concepts and terminology used in the proposed system. Section 4 covers the proposed security monitoring system with detailed description of each part's operation. The datasets used to evaluate the performance of the system modules individually are described and the system as a whole in Section 5, including experimental result. Finally, a conclusion is made in Section 6.

## 2. Face recognition systems overview

There are several facial recognition systems, and some of them are summarized below. In Afra and Alhajj (2020), an early warning system was developed and implemented to recognize individuals in a surveillance camera environment. The authors used various data sources to identify and profile these individuals and their networks. The proposed system achieves an accuracy of 93% by comparing data from social networks to identify unknown persons. Alternatively, Dong et al. proposed a new method in Bah and Ming (2020) that combines LBP with advanced IT techniques such as contrast adjustment, bilateral filtering, histogram equalization, and image blending. This system can recognize faces and be used in real-world settings as an automatic attendance management system. In D'cruz and Harirajkumar (2020), an infrared-based system was proposed for access control using face recognition with FaceNet, HoG, and SVM methods, as well as RFID and IR sensors. The authors created a web page for campus or company management to manage access authority and manually control the access system if needed. Similarly, in Pranav and Manikandan (2020), the authors proposed a deep learning-based face recognition system using the one-shot learning algorithm and Siamese CNN architecture with MySQL DB as input. They also created an interface (IHM) to interact with the user and manage the database, with the system sending an email via Gmail to employees in case of their absence using the email transfer protocol.

In Zhao et al. (2020), the authors propose the design and evaluation of a real-time face recognition system using a CNN network. The initial evaluation of the proposed design is performed using AT&T and then extended to the design of a real-time system. The proposed system has achieved an interesting results up to 95% as accuracy. In Sunaryono et al. (2021), the authors proposed a system based on self-updating template-based face recognition to address problems caused by model-based self-actualization algorithms, which preserve the expressive power of a limited set of models stored in the system database. Similarly, in Nassih et al. (2021), an efficient 3D face recognition approach based on geodesic distance (GD) using Riemannian geometry and random forest (RF) as classification method was proposed to address different problems. In Lv et al. (2021), a face recognition system was proposed to detect legal and illegal faces. In Rahouma and Mahfouz (2021), the authors presented a robust facial recognition system for Android cell phones based on pattern classification of image features, achieving an accuracy of over 95%. These previous efforts focused on specific circumstances and yielded high results when applied to specific perspectives.

In general, these works propose methods that respond to specific usage contexts, and real-time constraints are only slightly addressed. Therefore, this study addresses a well-targeted issue and presents a face recognition system capable of tracking and identifying individuals to detect abnormal activities. This system is designed, optimized, and implemented to ensure real-time operation according to a specific workflow.

## 3. Concepts and terminology

### 3.1. Persons detection

Detecting people in video streams is an important task in modern video surveillance systems. Computer vision and deep learning based methods for this task are numerous and can be divided into two main categories: two-stage detectors and one-stage detectors. The first category combines methods such as R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), and Mask R-CNN (He et al., 2017). These detection methods first find regions of interest using a selective search algorithm and then use these regions for classification. Although these detectors accuracy is good, they are time-consuming which make them less suitable for real-time applications. The second category combines methods such as the series of SSD detection methods: SSD (Liu et al., 2016), DSSD (Fu et al., 2017), ESSD (Zheng et al., 2018), MDSSD (Cui et al., 2018), the series of YOLO detection methods: YOLO (Redmon et al., 2016), YOLOv2 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Bochkovskiy et al., 2020), YOLOv5 (Jocher et al., 2020) and RetinaNet (Lin et al., 2017). These methods predict bounding boxes over the images without the regions selection step, thus, they are considerably faster and very suitable for real-time applications. YOLOv5 (Jocher et al., 2020) is a one-stage target recognition algorithm proposed in 2020. There are five network model versions of YOLOv5: YOLOv5n, YOLOv5s, YOLOv5 m, YOLOv5l and YOLOv5x. Out of these versions, the YOLOv5n network has the fastest calculation

speed and the lowest average precision while the YOLOv5x network has the opposite characteristics. The YOLOv5 network consists of three main components: Backbone, Neck and Head. Backbone aggregates the input image to form features at different scales, Neck then combines these features and pass them to the prediction layer and finally Head performs object localization and classification on these features. The YOLOv5 network uses GIOU as the network loss function.

The experimental results of the literature show that an adapted and improved method reach 89.1% and 67.8% on PASCAL VOC and MS COCO datasets. Compared to the initial YOLOv5s model, which gives 84.7% and 66.4 on the same PASCAL VOC and MS COCO datasets (Qu et al., 2022).

### 3.2. Tracking

The goal of the tracking stage is to assign a unique identifier to individuals as they move throughout the room. The challenge that must be overcome is the potential for occlusion during the movement of individuals, which can occur when two or more people cross paths and temporarily block one another from view. If the tracking algorithm is not robust enough to handle such situations, the system may fail to correctly identify individuals during favorable conditions for facial recognition.

Several approaches have been proposed in the literature to address this issue, with one of the most commonly used algorithms being DeepSORT (Bewley et al., 2016). An alternative method, Multiple Hypothesis Detection and Tracking (MHDT) (Abdelali et al., 2021), has been developed to improve upon DeepSORT and address the problem of occlusion. This method has demonstrated its effectiveness in situations similar to ours, achieving a multiple object tracking precision (MOTP) of 0.5927 and a multiple object tracking (MOTA) of 0.9415, surpassing other tracking methods.

### 3.3. Face detection

The objective of face detection is to automatically locate and extract facial regions in images or video streams. This is usually accomplished by utilizing a machine learning algorithm trained to recognize the patterns and features that characterize human faces. The algorithm analyzes the image data and generates a bounding box that encloses each detected face, along with additional details such as face position, size, and orientation.

Various face detection methods are available, ranging from simple rule-based techniques to advanced deep learning algorithms. Some commonly used face detection algorithms include the Viola–Jones algorithm, the Histogram of Oriented Gradients (HOG) method, and Convolutional Neural Networks (CNNs) based approaches.

Numerous highly accurate face detection methods are available and widely used in complete systems, including the CNN face detector in the dlib library, Retinaface (Deng et al., 2020), MTCNN (Zhang et al., 2016), Haarcascade, and the DNN Face Detector in the OpenCV library, among others.

The RefineFace model has demonstrated better results than other advanced methods, achieving an AP score of 99.45% on the PASCAL Face dataset. Additionally, it attained a score of 83.9% on the entire subset, 96.2% on the masked subset, and 95.7% on the unmasked subset of the dataset (Zhang et al., 2020).

### 3.4. Persons identification based on face recognition

The main purpose of this step is to utilize facial recognition as a biometric signature to identify the detected individuals. A variety of studies based on face recognition have been proposed in the literature to achieve this objective (Zennayi et al., 2022). These studies can be categorized into several groups, ranging from basic techniques that provide acceptable results under controlled conditions to increasingly complex methods designed to address specific challenges (Mahmood et al., 2017, Wang et al., 2017).

Techniques based on feature extraction using deep learning algorithms have proven their effectiveness on most databases in the literature (Guo and Zhang, 2019, Wang and Deng, 2021). The reason for adopting this approach resides in having a generic model that ensures an acceptable performance in systems.

Although these algorithms can achieve more than 99% accuracy on the LFW database (Deng et al., 2019; Wen et al., 2019) and 95% accuracy on Celebrities in Frontal Profile database (Ling et al., 2020); fundamental challenges such as illumination or pose changes remain unsolved.

DeepID (Sun et al., 2014) and DeepFace (Taigman et al., 2014) treat face recognition as a classification problem and use a shallow network with only 9 convolution layers. Using the same approach, the work of VGGFace (Parkhi et al., 2015) uses a larger scale of training data and a network with more shallow layers, and GoogleNet (Szegedy et al., 2015) is used to train the face recognition models. Subsequently, the adoption of a SE-ResNet-50 architecture as the backbone structure in VGGFace2 (Cao et al., 2018). Specifically, The ArcFace (Deng et al., 2019) uses ResNet-50 and ResNet-100 to perform its investigations.

The Facenet model has been evaluated on several public facial recognition databases, but the best result was obtained on the LFW (Labeled Faces in the Wild) database. This database is widely used in facial recognition research as it contains over 13,000 images of celebrity faces captured in realistic conditions, with variations in pose, expression, and lighting.

The result obtained by the Facenet model on the LFW database is 99.63%. This result is considered very performant and set a new record for facial recognition at the time of its publication in 2015. Since then, many other models have been developed that have surpassed this result on the LFW database, but Facenet remains a highly performant and widely used model in the facial recognition community (Wang and Deng, 2021).

## 4. Methodology

### 4.1. Proposed system architecture

In this section the architecture of the proposed system is detailed. As shown in Fig. 1, the face recognition system is composed of 6 core modules.

- Images acquisition: This first module is used to recover the images from a camera, then apply the preprocessing to have normalized images.
- Persons detection: This module uses the images to detect the position and bounding box of persons in the camera's field of view.
- Persons tracking: The objective of this module is to extract the positions and bounding boxes of individuals detected in single images, enabling the tracking of their movements across multiple images.
- Location analysis: This module uses a configuration of the room (a configuration is defined at the beginning, containing the position of the controlled equipments with the list of the persons authorized to access each of them) and the persons' movements information, to assign the access to the equipments to the persons during their presence in the room.
- Face detection: This module uses the cropped images of persons, to identify the position of their face.
- Face recognition: This module is used to identify people using facial recognition techniques.
- Unauthorized access identification: The latter module combines the movement information of people in the room with their identity to detect unauthorized access situations and inform supervisors.

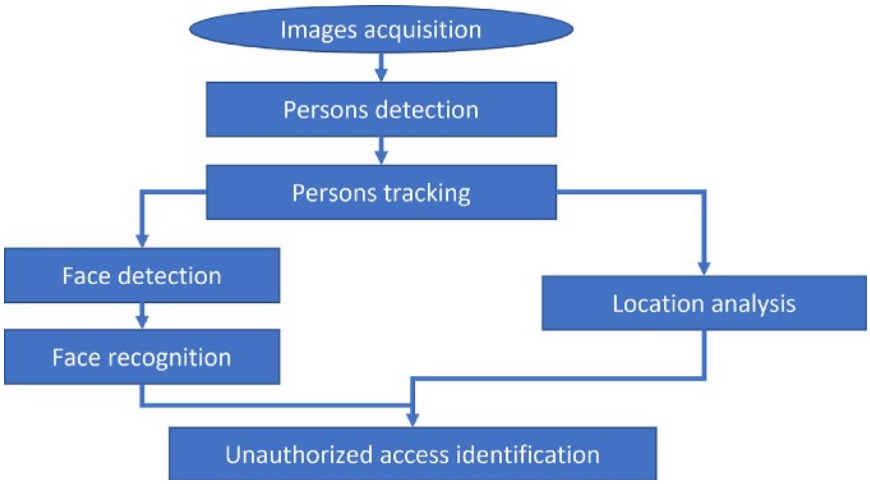



**Fig. 1.** The workflow of the proposed system for persons' face recognition.

### 4.2. Images acquisition

The proposed system contains mainly a standard CCTV camera. This camera allows to cover the whole area of interest. It has a CMOS image sensor (1280 × 960 pixels) and a focal length of 2, 8 mm. The first software model allows to recover images, to calibrate images against distortion problems, caused by the camera optics. Then, this model applies a histogram equalization to adjust the contrast of the images.

### 4.3. Persons detection

For real-time applications, processing speed is a crucial factor and model accuracy is also important to assure sound functioning of the system. Thus, YOLOv5s is the most suitable state of the art model for real-time applications as it is both fast and precise for person detection (see Section 5.2.1). YOLOv5s pretrained model was utilized in this work, which is capable of recognizing eighty different classes. However, the focus was specifically on the "person" class.

### 4.4. Tracking

The tracking stage is a crucial component of many computer vision and object recognition systems, and its main purpose is to locate and track objects of interest as they move through a scene. In the case of people tracking, the system uses a camera to detect and track individuals within the scene. Once the system has detected a person, it may assign them a unique identifier, such as a number or a name, to keep track of their movements throughout the scene. This identifier can then be used to associate the person with other information, such as their location or actions, which can be used to analyze their behavior or make decisions based on their movements.

The MHDT (Multi-Hypothesis Data Association Tracking) method was utilized in this work to track detected individuals. This method not only enables the tracking of individuals over time, but also addresses challenges such as missed detections and occlusions. The use of MHDT method improves the performance of the system and enables us to identify individuals with greater accuracy. Additionally, it is used to avoid applying time-consuming facial re-identification techniques.

### 4.5. Location analysis

The objective of this step is to analyze the trajectory of each detected individual. The system assigns a unique ID to each person upon their appearance in the scene, after which the tracking module generates a list of their positions over time. The aforementioned list will be subject of an analyzer in order to detect whether a person exists in an unauthorized area or not by applying the intersection between each bounding box and the unauthorized areas. Unauthorized persons are identified as suspected if they stayed in the unauthorized areas during a considerable moment (can be parametrized) while they are not authorized to access to the specific equipments.

### 4.6. Face detection

The aim of this step is to detect faces within the bounding boxes generated in the previous step, where individuals were detected. In this work, Retinaface pretrained model was utilized in this work due to its high accuracy and fast performance, as described in Section 5.2.2.

### 4.7. Persons identification based on face recognition

The main objective of this step is to use facial recognition as a biometric identifier to identify the detected individuals. To achieve this, a deep learning based algorithm has been adopted.

These algorithms have achieved high performance in various situations. Regarding the proposed use scenario, it is believed that improved performance can be attained while keeping the model parameters and computational complexity nearly unchanged. The performance and computational complexity of various models were evaluated to determine the most efficient model for the current use case, as discussed in Section 5.2.3.1. Then an improvement of a model is proposed for an increased performance of the system.

*Our approach*: As illustrated in Fig. 2, the proposed person identification system is designed as a multi-module system divided into two parts: a first part for processing individual images, and a second part with a memory aspect for processing the images associated with a person from its appearance in the scene to its exit.

The input of this block is a single image I0, which is a matrix having variable size (depending on the position and orientation of face regarding the camera) and represents the face of a person monitored in the scene. After storing the image in a list, a first sub-block performs a check on the image to measure the distance to the last image processed by the "Face recognition" sub-module. This operation is used to avoid processing close images which will relieve the processing load while keeping the same efficiency.

$$df = \left\| Img_l - Img_i \right\| \tag{1}$$

The distance df is calculated according to Eq. (1) based on the euclidean norm, where $Img_l$ is the last processed image and $Img_i$ is the received image. The rest of the processing of this module is conditioned by the following requirement : df > Dmin (where Dmin is a parameter designating the minimum tolerable distance).

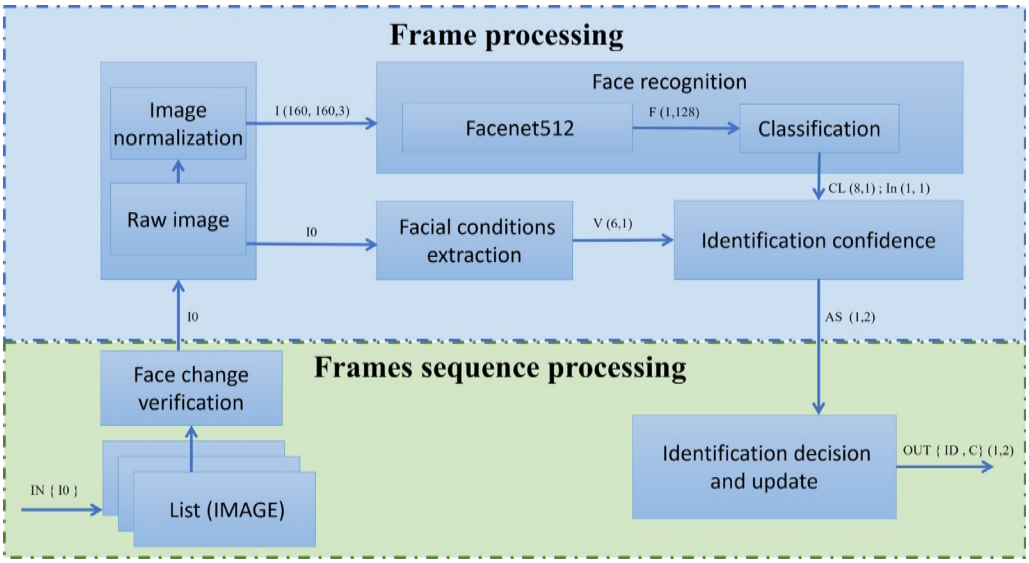



**Fig. 2.** Architecture of the face recognition identification block.

The "Facial conditions extraction" sub-module uses the I0 image to extract a vector $V = [V_1; V_2; \ldots; V_6]$ representing indicators on the quality of this image(where the optimal lighting and homogeneous illumination is explained in Han et al. (2020), and Lopes et al. (2017), Patacchiola and Cangelosi (2017) for facial expression recognition and face orientation estimation respectively), which is composed of six components with the following meanings:

- $V_1$ represents a percentage that the image I has an optimal size of $160 \times 160$.
- $V_2$ represents a percentage that the image I has an optimal lighting.
- $V_3$ represents a percentage that the image I has a homogeneous illumination.
- $V_4$ represents a percentage that the image I has a neutral expression
- $V_5$ represents a percentage of yaw orientation of I compared to the optimal orientation.
- $V_6$ represents a percentage of pitch orientation of I compared to the optimal orientation.

After normalizing the image I0 to have an image I of size (160,160, 3), it is communicated to the sub-block "FaceNet512" for feature extraction, to have a vector F of dimension (1,128) which represents a face signature. The next step is the classification of this vector to identify the person that corresponds the most to the face image. To perform this step, an MLP neural network is used, composed of four layers: a first input layer of size (1, 128); a second hidden Dense layer of size (1, 2 * N) with a CRelu activation function; a third hidden Dropout layer at 25%; and a final output layer of type Dense with a size of (1, N) and a Softmax activation function. The proposed architecture is composed mainly of an input layer with 128 nodes, two hidden layers (Dense and dropout layers), and an output layer with $N$ nodes (where $N$ is the number of persons to be identified) with the Softmax activation function. The obtained results are sorted and then filtered to keep an CL matrix of the size (8x1) representing the confidence rates of the 8 individuals who are most susceptible to being identified. The scalar In which represents the index of the person with the highest score was also issued.

The last sub-block "Identification confidence" of the "Frame processing" part will retrieve both the vector CL and the vector V to improve the confidence of the identification. The concatenation of those last mentioned vectors forms the AS vector of length (14, 1). The sub-block utilizes an MLP neural network module, which consists of four layers. The first layer is an input layer with a size of (1, 14). The second layer is a hidden Dense layer with a size of (1, 20) and a Relu activation function. The third layer is another hidden Dense layer with a size of (1, 10) and a Relu activation function. The final layer is an output layer of type Dense with a size of (1, 2) and a Softmax activation function. The model utilizes the AS vector to output two classes, indicating whether the result of the face recognition sub-block is correct or not.

The last block "Identification decision and update" aims to use the results obtained (The vector AS) from each frame to identify the person by specifying a confidence rate considering the general conditions of the face acquired images. The main advantage of this block is its ability to update the identification of the person if the conditions improve to increase the chances for a better identification.

This sub-block assigns an identity to a person if the trust exceeds a fixed threshold (in this case 90%), and updates this identification when the trust improves between frames.

### 4.8. Unauthorized access identification

This step analyzes the list of persons present in the monitoring area as a dataframe (see Fig. 3). the information about the persons will be analyzed one by one, which are an identifier "Person ID" and a list of information about the equipments accessed by the person "Equipment Access". The identifier is initialized by the value "Unknown" to mean unidentified person. This variable will be changed by an acronym in case of identification of the person. The "Equipment Access" field contains three main items: an identifier of the equipment "Equipment ID", the date and time when the person started accessing to the device "Time in" and the date and time when the access was stopped "Time out". A verification is launched on all the accessed equipment to examine if the person is part of the eligible list "Equipment List", if not an alert will be raised immediately (the alert will also be raised for an Unknown value).

### 4.9. Real-time implementation

To ensure a real-time system execution, an effort has been made to optimize the algorithm and a distribution of the execution between the CPU and the GPU (The execution platform is shown in the Section 5.3.1). The proposed implementation in this work is a parallel architecture with simultaneous threads (see Fig. 4).

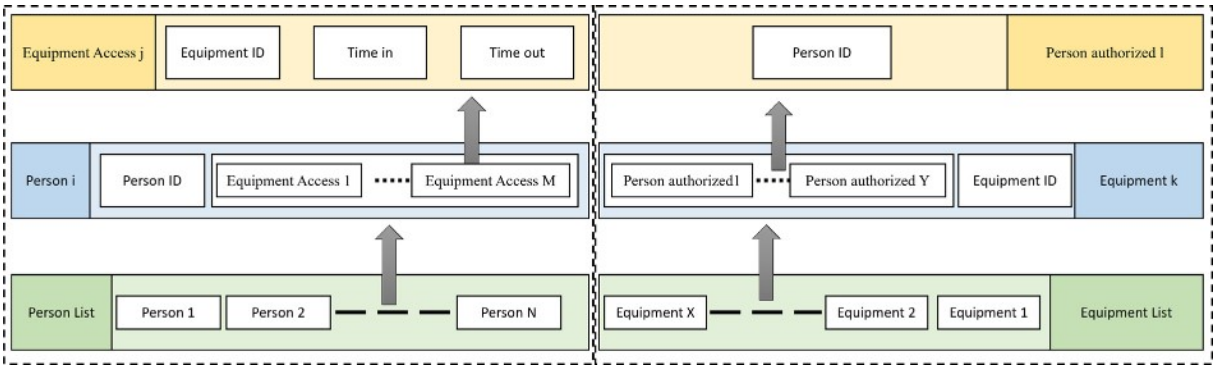



**Fig. 3.** Structure of the person and equipment data frame.

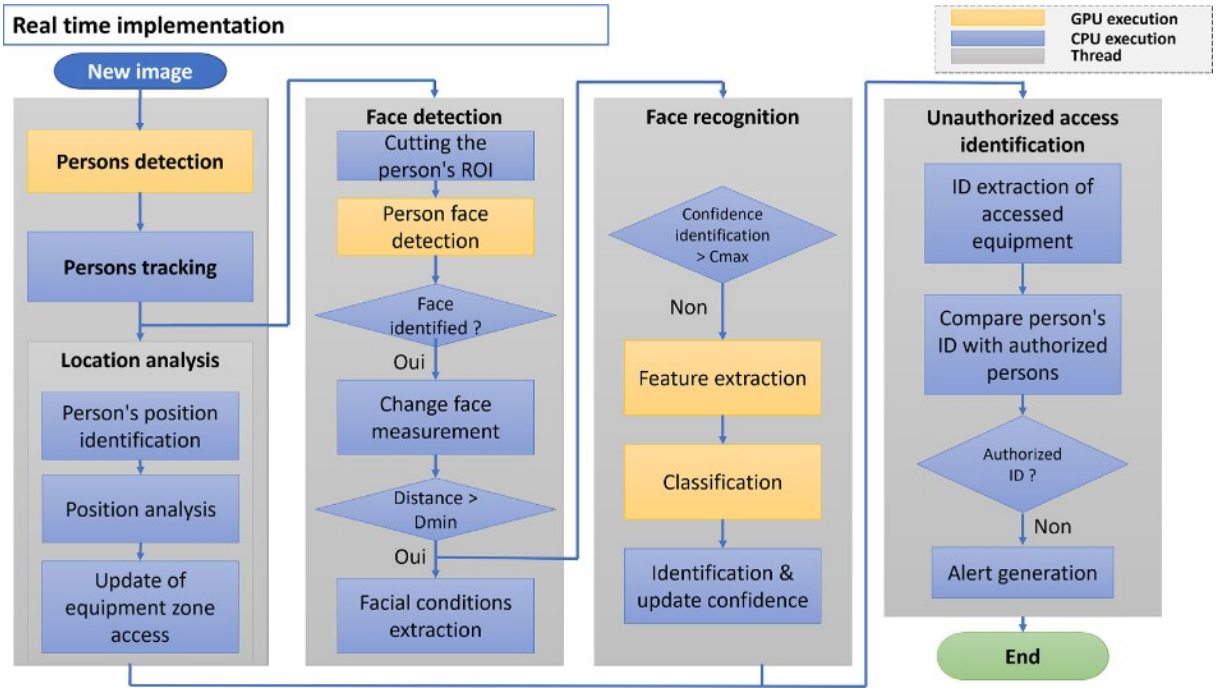



**Fig. 4.** System monitoring implementation algorithm.

- A first thread allows to acquire the camera images, to perform the pre-crossing, to execute the people detection model on the GPU, and finally to update the list of people with the help of the tracking algorithms. And secondly to analyze the position of the active persons in the session and to update the equipment access table. This thread has a uniform execution load, and does not depend significantly on the number of people in the scene.
- A second thread that retrieves the list of people present in the scene. It crops the ROI of the regions representing the people that require identification (unidentified or identification below a high confidence threshold), to launch the face detection at the GPU. This thread has a load proportional with the number of active people in the scene needing identification.
- A third thread is used to measure a distance between the detected face image and the last processed frame. This check is used to avoid additional operations (face condition extraction and face recognition), which are unnecessary in case of small changes on the face images tracked between frames. An algorithm is applied on the face to extract the vector V representing the general conditions and quality of the face image; it runs the feature extraction and classification models on the GPU for the person requiring identification (unidentified or identification below a high confidence threshold). Then it executes the final identification model to make a decision (person ID assignment with confidence). This thread has an execution load proportional to the number of persons needing an identification.
- A fourth and last thread allows the detection of unwanted accesses by comparing the equipment access lists with the eligibility lists. Also, it sends alerts of unauthorized access. This thread has a low execution load.

## 5. Experimental results

### 5.1. Datasets

The experimental procedure was setup to validate the following four points regarding the proposed real-time video processing room monitoring system: (1) the person detection module is both accurate and fast, (2) the face detection module is both accurate and fast, (3) the face recognition module is accurate for multi-subjects, fast and robust to uncontrolled environment conditions, (4) the combined blocks still assure accurate real-time video processing. To evaluate the latter points, a complex and challenging persons' video dataset has been collected using a vision-based system in different recording conditions.

**Table 1**
Face recognition databases information.

| Database | Number of persons | Train | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Number of images per person | | | Total images | Number of images per person | | | Total images |
| | | min | mean | max | | min | mean | max | |
| Aberdden | 77 | 1 | 4 | 16 | 367 | 1 | 3 | 12 | 271 |
| AR | 136 | 1 | 7 | 8 | 1015 | 1 | 5 | 6 | 742 |
| ATT | 40 | 1 | 4 | 10 | 185 | 1 | 5 | 9 | 215 |
| BioID | 23 | 2 | 34 | 99 | 785 | 1 | 28 | 70 | 657 |
| Caltech | 27 | 1 | 7 | 13 | 214 | 1 | 8 | 14 | 226 |
| CMU-PIE | 68 | 1 | 5 | 19 | 362 | 5 | 18 | 23 | 1270 |
| ESSEX_94 | 152 | 4 | 9 | 19 | 1505 | 5 | 9 | 19 | 1499 |
| FEI | 200 | 2 | 2 | 3 | 587 | 2 | 2 | 3 | 577 |
| FERET | 697 | 1 | 1 | 13 | 1221 | 1 | 1 | 15 | 1267 |
| GT | 50 | 5 | 6 | 7 | 347 | 5 | 7 | 8 | 393 |
| GTAV | 44 | 3 | 7 | 15 | 339 | 2 | 7 | 16 | 351 |
| MAScIR-DB | 28 | 1 | 5 | 15 | 149 | 1 | 5 | 13 | 147 |
| YaleB | 38 | 1 | 8 | 19 | 313 | 37 | 49 | 59 | 1891 |

This dataset was used to evaluate the overall system and each of the modules separately by creating other image datasets, a human dataset (named "MAScIR-PERSON"), to assess the person detection and the face detection modules and an face dataset (named "MAScIR-SYS-DB") to train and test the face recognition module. The experiments were conducted on the self-built and the state of the art image datasets as follows:

For person detection module evaluation, MAScIR-PERSON database was built from the videos generated by the system (see Section 5.3.2). This database is composed of 1070 images that were selected from different recording conditions. The lightning conditions go from dark to very bright. Each image contains at least one person (or part of a person) standing, sitting or walking while facing different directions. The distance between the subjects and the camera goes from 1 to 7 m. An annotation step was performed to delimit the persons' box to have a reliable source of comparison in order to measure the performance of the state of the art person detection pre trained models.

To evaluate the face detection module, the same MAScIR-PERSON database used for person detection was utilized. However, in this case, each image was associated with an annotation file indicating the location of the face in the image.

In the context of face recognition, both literature databases such as Aberdden, AR, ATT, BioID, Caltech, CMP-PI, ESSEX, FEI, FERET, GT, GTAV and YaleB, as well as a customized database called "MAScIR-DB", which was captured under two different lighting configurations (normal and low lighting), were utilized. These databases were respectively split into two sub-datasets Train and Test, Table 1 summarizes the information of these databases. Fig. 5 shows some samples of images of both sub-datasets Train and Test for each database.

To evaluate the performance of the present system, MAScIR-SYS-DB database was established using the videos generated by the camera of the system installed in the room (see Section 5.3.2). This database is composed of 4 parts: Reference, Scenario 1 (normal lighting), Scenario 2 (low lighting) and Scenario 3 (several people in the room). The first part is used to build the person identification model, while the other parts are used to test the performance of the system. Table 2 summarizes the composition of this database. An overview of the images in each of these scenarios is illustrated in Fig. 6.

**Table 2**
Information about the database generated by the system in the room "MAScIR-SYS-DB" for face recognition evaluation.

| | | Number of images per scenario | | | |
|---|---|---|---|---|---|
| | | Reference | Scenario 1 | Scenario 2 | Scenario 3 |
| Group 1 | Person 1 | 30 | 76 | 64 | 59 |
| | Person 2 | 30 | 79 | 52 | 71 |
| | Person 3 | 30 | 63 | 61 | 48 |
| Group 2 | Person 4 | 30 | 72 | 64 | 68 |
| | Person 5 | 30 | 70 | 51 | 65 |
| | Person 6 | 30 | 60 | 63 | 60 |
| Group 3 | Person 7 | 30 | 76 | 55 | 58 |
| | Person 8 | 30 | 65 | 54 | 51 |
| | Person 9 | 30 | 68 | 50 | 53 |
| Group 4 | Person 10 | 0 | 62 | 50 | 60 |
| | Person 11 | 0 | 72 | 55 | 61 |
| | Person 12 | 0 | 77 | 64 | 64 |

## 5.2. Evaluation

### 5.2.1. Person detection

To select the most suitable pre-trained model for person detection in the system, a comparison was conducted between five released YOLOv5 networks: YOLOv5n, YOLOv5s, YOLOv5 m, YOLOv5l, and YOLOv5x. The evaluation criteria were accuracy and speed. The official releases of the models provide accuracy and speed information on the COCO dataset. However, benchmarking the models on the MAScIR-PERSON dataset was necessary to select the appropriate network that best suits the system's settings and requirements.

To assess the accuracy of the person detection models, the standard metric of Average Precision (AP) was used, which evaluates the models' performance for different Intersection over Union (IoU) thresholds. The IoU thresholds ranged from 50% to 90% with a step size of 10. AP was calculated using annotated files as ground truth and detection results files that were generated on MAScIR-PERSON database after filtering "person" class from the rest of the recognition results. YOLOv5x had the highest average AP (73.16%) calculated from resulted AP for all IoU thresholds (see Table 3). Reworded: The execution time per frame was calculated for each of the models to evaluate their speed. YOLOv5n was found to be the fastest, with a time of 20.85 ms per frame, as indicated in Table 3. Considering that the proposed work is a real-time processing system, YOLOv5s was selected as the pre-trained model of choice. This model is the second fastest among the evaluated models, taking 35.25 ms per frame, while also having a high average AP of 69.54%.

### 5.2.2. Face detection

To select the appropriate face detection technique for the system, a comparison was made between the accuracy and speed of several pre-trained models including Dlib, Mtcnn, Haarcascade from OpenCV, Retinaface, and SSD (caffe model) from DNN in OpenCV. In Section 5.2.1, the evaluation of the latter models was performed using Average Precision (AP) to assess accuracy and execution time per frame to assess speed on the MAScIR-PERSON database. AP was calculated for various Intersection over Union (IoU) thresholds (from 50% to 90% with a step size of 10) using annotated files as ground truth and face detection results files that were generated on MAScIR-PERSON

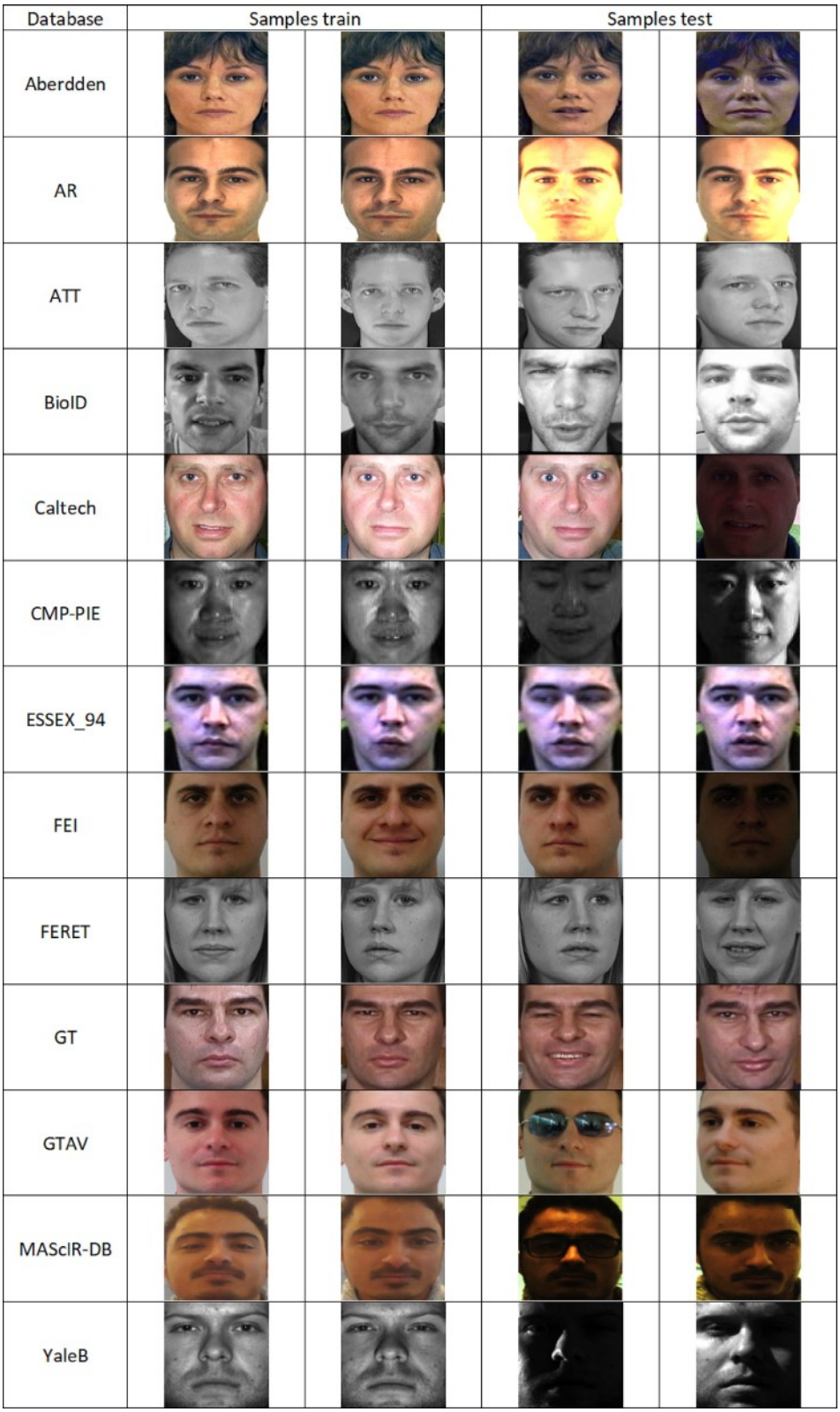



**Fig. 5.** Database samples.

database. As a result of this comparison (see Table 4), Retinaface had the highest average AP (98.9%) calculated from resulted AP for all IoU thresholds meanwhile SSD (caffe model) ranked last (3.6%). As for speed, Dlib and haarcascade face detectors work only on CPU, thus they took the most time to process a frame, whereas SSD (caffe model) consumed the least time (6.76 ms). Due to the significance of accurate face detection for the subsequent face recognition module, the Retinaface model was selected as it achieved the highest accuracy and ranked third in terms of speed, processing frames at 93.50 ms per frame.

The RefineFace model outperforms state-of-the-art methods with an AP score of 99.45% on the PASCAL Face dataset, 83.9% on the


**Fig. 6.** MAScIR-SYS-DB database samples.

whole subset, 96.2% on the masked subset, and 95.7% on the unmasked subset of the MAFA dataset.

*5.2.3. Face recognition*

*5.2.3.1. Deep learning feature extraction model benchmark.* The adopted approach in this work is based on feature extraction using Deep learning. To select the model for the proposed system, a comparison was conducted with the most successful architectures in the literature, taking into account two criteria: accuracy and computation time.

As shown in Table 5 it can be noted that the Facenet512 model presents the best performance in terms of accuracy on most databases, using KNN (K=1) as classifier, while, referring to the Table 6 the best

**Table 3**
YOLOv5 networks' accuracy and speed comparison.

| Model | AP50 | AP60 | AP70 | AP80 | AP90 | Average AP | (Average AP)/(YOLOv5x Average AP) | CT : Computation time GPU (ms) | CT/(YOLOv5n CT) |
|---|---|---|---|---|---|---|---|---|---|
| YOLOv5n | 86,36% | 83,54% | 72,79% | 47,61% | 2,50% | 58,00% | 79,28% | 20,85 | 100% |
| YOLOv5s | 95,61% | 94,98% | 85,30% | 62,48% | 9,35% | 69,54% | 95,06% | 35,25 | 169% |
| YOLOv5m | 98,20% | 97,75% | 91,88% | 64,80% | 10,72% | 72,67% | 99,33% | 71,12 | 341% |
| YOLOv5l | 99,06% | 98,34% | 93,50% | 63,48% | 10,89% | 73,05% | 99,86% | 129,11 | 619% |
| YOLOv5x | 98,88% | 98,21% | 93,08% | 64,54% | 11,09% | 73,16% | 100,00% | 218,38 | 1047% |

**Table 4**
Face detection techniques' accuracy and speed comparison.

| Model | AP50 | AP60 | AP70 | AP80 | AP90 | Average AP | (Average AP)/(retinaface Average AP) | CT : Computation time CPU (ms) | (CT/ssd CT) |
|---|---|---|---|---|---|---|---|---|---|
| dlib (shape_predictor_5_face_landmarks.dat) | 46,80% | 34,43% | 0,43% | 0,00% | 0,00% | 16,33% | 16,5% | 650,00 | 9615% |
| mtcnn | 85,45% | 83,10% | 77,23% | 56,52% | 5,61% | 61,58% | 62,3% | 68,50 | 1013% |
| opencv (haarcascade) | 33,23% | 29,92% | 0,08% | 0,00% | 0,00% | 12,65% | 12,8% | 580,00 | 8580% |
| retinaface | 98,88% | 98,88% | 98,88% | 98,88% | 98,88% | 98,88% | 100,0% | 93,50 | 1383% |
| ssd (res10_300x300_ssd_iter_140000.caffemodel) | 4,61% | 4,57% | 4,57% | 3,72% | 0,29% | 3,55% | 3,6% | 6,76 | 100% |

**Table 5**
Comparison in terms of accuracy of the face feature extraction models (VGG-Face, Facenet, OpenFace, DeepFace, ArcFace and FaceNet512)and ML classifier (KNN[K=1-5) and SVM).

| Datasets | | | | | | | | | | | | | Classifiers | Models | Statistics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aberdden | AR | ATT | BioID | Caltech | CMU-PIE | ESSEX_94 | FEI | FERET | GT | GTAV | MAScIR-DB | YaleB | | | Mean | Max | Min |
| 96% | 94% | 95% | 98% | 97% | 62% | 100% | 89% | 96% | 99% | 98% | 83% | 47% | KNN1 | | 89% | 100% | 47% |
| 94% | 93% | 71% | 98% | 97% | 56% | 100% | 86% | 76% | 99% | 96% | 83% | 44% | KNN2 | | 84% | 100% | 44% |
| 93% | 93% | 62% | 98% | 96% | 54% | 100% | 86% | 69% | 99% | 97% | 82% | 41% | KNN3 | | 82% | 100% | 41% |
| 87% | 92% | 56% | 98% | 96% | 50% | 100% | 82% | 64% | 98% | 97% | 82% | 40% | KNN4 | VGG-Face | 80% | 100% | 40% |
| 86% | 91% | 53% | 98% | 93% | 47% | 100% | 79% | 60% | 97% | 95% | 80% | 37% | KNN5 | | 78% | 100% | 37% |
| 83% | 93% | 59% | 98% | 94% | 46% | 100% | 86% | 46% | 99% | 97% | 83% | 39% | SVM | | 79% | 100% | 39% |
| 99% | 94% | 96% | 97% | 97% | 48% | 100% | 82% | 83% | 96% | 95% | 24% | 47% | KNN1 | | 81% | 100% | 24% |
| 97% | 92% | 77% | 97% | 96% | 44% | 100% | 80% | 67% | 96% | 93% | 25% | 43% | KNN2 | | 78% | 100% | 25% |
| 98% | 92% | 67% | 96% | 96% | 42% | 100% | 80% | 61% | 96% | 94% | 23% | 41% | KNN3 | | 76% | 100% | 23% |
| 92% | 89% | 60% | 96% | 95% | 37% | 100% | 75% | 57% | 95% | 92% | 21% | 39% | KNN4 | Facenet | 73% | 100% | 21% |
| 91% | 88% | 56% | 95% | 94% | 35% | 100% | 70% | 53% | 94% | 92% | 19% | 38% | KNN5 | | 71% | 100% | 19% |
| 96% | 94% | 60% | 96% | 93% | 37% | 100% | 82% | 47% | 95% | 95% | 23% | 40% | SVM | | 74% | 100% | 23% |
| 49% | 32% | 54% | 78% | 95% | 75% | 98% | 65% | 71% | 73% | 55% | 18% | 51% | KNN1 | | 63% | 98% | 18% |
| 44% | 28% | 40% | 77% | 94% | 68% | 96% | 57% | 52% | 68% | 50% | 16% | 49% | KNN2 | | 57% | 96% | 16% |
| 43% | 29% | 34% | 79% | 93% | 67% | 96% | 58% | 46% | 71% | 49% | 16% | 45% | KNN3 | | 56% | 96% | 16% |
| 45% | 28% | 31% | 77% | 92% | 65% | 96% | 58% | 42% | 68% | 50% | 20% | 43% | KNN4 | OpenFace | 55% | 96% | 20% |
| 46% | 28% | 28% | 78% | 91% | 62% | 96% | 54% | 39% | 69% | 52% | 18% | 41% | KNN5 | | 54% | 96% | 18% |
| 43% | 25% | 24% | 80% | 92% | 61% | 98% | 62% | 29% | 73% | 50% | 20% | 41% | SVM | | 54% | 98% | 20% |
| 88% | 72% | 60% | 89% | 96% | 66% | 100% | 48% | 63% | 67% | 62% | 43% | 55% | KNN1 | | 70% | 100% | 43% |
| 83% | 66% | 42% | 90% | 94% | 60% | 100% | 35% | 42% | 56% | 51% | 46% | 52% | KNN2 | | 63% | 100% | 35% |
| 83% | 64% | 40% | 87% | 94% | 58% | 100% | 34% | 34% | 56% | 49% | 39% | 49% | KNN3 | | 60% | 100% | 34% |
| 81% | 63% | 37% | 88% | 94% | 57% | 100% | 33% | 31% | 52% | 42% | 43% | 47% | KNN4 | DeepFace | 59% | 100% | 31% |
| 75% | 62% | 35% | 87% | 92% | 54% | 100% | 32% | 30% | 49% | 41% | 44% | 46% | KNN5 | | 57% | 100% | 30% |
| 65% | 51% | 18% | 76% | 91% | 51% | 100% | 50% | 18% | 69% | 44% | 40% | 41% | SVM | | 55% | 100% | 18% |
| 98% | 96% | 89% | 97% | 97% | 79% | 100% | 93% | 94% | 97% | 97% | 93% | 50% | KNN1 | | 91% | 100% | 50% |
| 94% | 93% | 71% | 97% | 97% | 74% | 100% | 93% | 79% | 97% | 92% | 91% | 46% | KNN2 | | 87% | 100% | 46% |
| 95% | 94% | 63% | 97% | 97% | 72% | 100% | 94% | 70% | 96% | 93% | 93% | 45% | KNN3 | | 85% | 100% | 45% |
| 94% | 92% | 60% | 97% | 95% | 68% | 100% | 93% | 63% | 96% | 92% | 90% | 44% | KNN4 | ArcFace | 83% | 100% | 44% |
| 91% | 91% | 58% | 97% | 92% | 65% | 100% | 92% | 58% | 95% | 92% | 91% | 42% | KNN5 | | 82% | 100% | 42% |
| 95% | 95% | 55% | 98% | 96% | 66% | 100% | 94% | 55% | 97% | 95% | 95% | 43% | SVM | | 83% | 100% | 43% |
| 99% | 98% | 99% | 98% | 97% | 78% | 100% | 92% | 97% | 99% | 98% | 89% | 66% | KNN1 | | 93% | 100% | 66% |
| 99% | 95% | 77% | 98% | 97% | 77% | 100% | 92% | 80% | 99% | 98% | 90% | 64% | KNN2 | | 90% | 100% | 64% |
| 99% | 97% | 67% | 98% | 97% | 77% | 100% | 92% | 72% | 99% | 98% | 90% | 59% | KNN3 | | 88% | 100% | 59% |
| 97% | 96% | 58% | 98% | 96% | 75% | 100% | 86% | 65% | 99% | 98% | 88% | 57% | KNN4 | Facenet512 | 86% | 100% | 57% |
| 96% | 95% | 53% | 98% | 92% | 70% | 100% | 82% | 61% | 99% | 98% | 87% | 56% | KNN5 | | 84% | 100% | 53% |
| 98% | 98% | 66% | 98% | 97% | 76% | 100% | 91% | 61% | 99% | 99% | 91% | 58% | SVM | | 87% | 100% | 58% |

**Table 6**
Comparison in terms of computation time of the face feature extraction models (VGG-Face, Facenet, OpenFace, DeepFace, ArcFace and FaceNet512).

| Model | CT : Computation Time CPU (ms) | CT/(OpenFace CT) |
|---|---|---|
| VGG-Face | 229 | 1423% |
| Facenet | 61 | 380% |
| OpenFace | 16 | 100% |
| DeepFace | 40 | 251% |
| ArcFace | 146 | 909% |
| Facenet512 | 66 | 410% |

**Table 7**
Results obtained from the face recognition module on the study database.

| Database | Train | Test |
|---|---|---|
| Aberdden | 98% | 97% |
| AR | 100% | 99% |
| ATT | 100% | 99% |
| BioID | 100% | 99% |
| Caltech | 98% | 96% |
| CMU-PIE | 100% | 80% |
| ESSEX_94 | 100% | 100% |
| FEI | 96% | 90% |
| FERET | 100% | 98% |
| GT | 100% | 99% |
| GTAV | 98% | 94% |
| MAScIR-DB | 100% | 98% |
| YaleB | 100% | 68% |

model in terms of computation time is OpenFace. The proposed system places a high emphasis on achieving high accuracy, making it the primary selection criterion, followed by computation time. Therefore, the Facenet512 model was selected as it excelled in accuracy, meeting the first criterion.

*5.2.3.2. Face recognition model experiments on literature databases.* In order to achieve superior results on literature databases, the Facenet512 model was adopted as the feature extraction model. An MLP classification model was trained and optimized separately on each database for the recognition identification function, testing several architectures and configurations: number of layers (between 1 and 3 layers), activation function (Softmax, Relu, and CRelu), the use or not of a Dropout layer. During these tests, it was observed that modifying the number of layers and nodes per layer had little effect on the obtained results, except for the convergence time of the model. However, changes to the activation function and activation of a Dropout layer did have an impact, leading to the selection of 6 significant configurations for comparison: (1) softmax without Dropout; (2) softmax with Dropout; (3) Relu without Dropout; (4) Relu with Dropout; (5) CRelu without Dropout; (6) CRelu with Dropout. The results showing the difference on the 6 configurations are shown in Fig. 7.

Considering the importance of execution time in the system, a decision was made to employ a single hidden layer with a reduced number of nodes. The CRelu activation function was chosen, and a Dropout layer was added to further improve performance. The outcomes of these decisions are presented in Table 7. The outcomes achieved on the literature databases demonstrate excellent efficacy in relatively controlled lighting scenarios, with accuracy rates exceeding 96%. However, for databases with significant or extreme lighting conditions, lower accuracy was observed, such as 80% for CMEPIE and 68% for Yale.

*5.2.3.3. Experimentation of the face recognition model on the room databases.* Since in real operating scenarios the conditions are not mastered, it is therefore necessary to test the effectiveness of this architecture on a database of this context. Fig. 7 shows the best results obtained for the 6 configurations adopted in this study. The results are displayed using the training database (in blue color), on the test database 1 which represents the normal lighting scenario (in orange color) and on the test database 2 which represents the low lighting scenario (in gray color). Overall, the outcomes obtained are favorable, with an accuracy of 90% achieved on the test 1 dataset (normal lighting) and 86% on the test 2 dataset (low lighting) by utilizing the CRelu activation function.

### *5.2.4. Proposed identification algorithm*

*5.2.4.1. Precision of the identification mono-frame.* To verify the identification of an individual using the face recognition model's output, the confidence of the model (Softmax function) was initially employed. Nevertheless, as the model is essentially a classification model, it does not directly address the scenario where the individual to be identified is not present in the reference database. Fig. 8 (a: results obtained from the basic face recognition model) shows the evolution of the identification rate accompanied by the accuracy of this identification by varying the confidence level between 1% and 100%. It is notable that the identification rate significantly decreases when aiming for high accuracy rates. For instance, to attain an accuracy rate greater than 90%, the identification rate declines to 58%. This configuration is not recommended in this context because in case of low identification rate the system may send a lot of false alarms due to non-identification of persons.

To enhance the situation, a model was created to distinguish between accurate and inaccurate facial recognition operations. The model's purpose is to decrease false alarms and simultaneously enhance the identification rate.

By utilizing the facial recognition models developed (refer to Section 5.2.3.2), results were generated on the test dataset. A CL vector was generated for each image and a target was created using the ground truth to indicate whether the result was accurate or not. On the other hand the V vector representing the conditions of the images has been generated from the same databases, and finally using the block association to concatenate the CL vector with the V vector to have a single vector named AS. Thus, at the end, an input AS vector of size (18, 1) and a binary target were obtained.

The accuracy achieved after training is 99% on the training dataset and 97% on the testing dataset. To evaluate the effectiveness of the model, it was tested on the room database, and the confidence rate was varied to obtain the facial recognition rate with the accuracy of identification. The results obtained from the additional identification model are shown in Fig. 8(b).

According to the results presented in Fig. 8, the identification rate of individuals has improved while maintaining the same precision. For instance, for a precision of 85%, the first method resulted in an identification rate of 70%, while the second method achieved an identification rate of 82%. Similarly, for a precision of 95%, the first method resulted in an identification rate of 48%, while the second method achieved an identification rate of 71%.

Moreover, regarding the computational load, this model is insignificant compared to the main face recognition model. Because, the number of parameters of this neural network is 1322, against 23.5 million for the face recognition model (i.e. 0.006%).

*5.2.4.2. Identification multi-frame.* In this section, the identification decision block is utilized to assign an identity to individuals detected and tracked in a sequence of images. The main idea of this module is to assign identity if the confidence exceeds 90%, and it will be updated if the confidence improves from one image to the next, and ultimately stop the operation of face detection and recognition if the confidence exceeds 98%. Two methods have been evaluated: (1) using the confidence of the recognition model; (2) using the confidence of the proposed identification model.

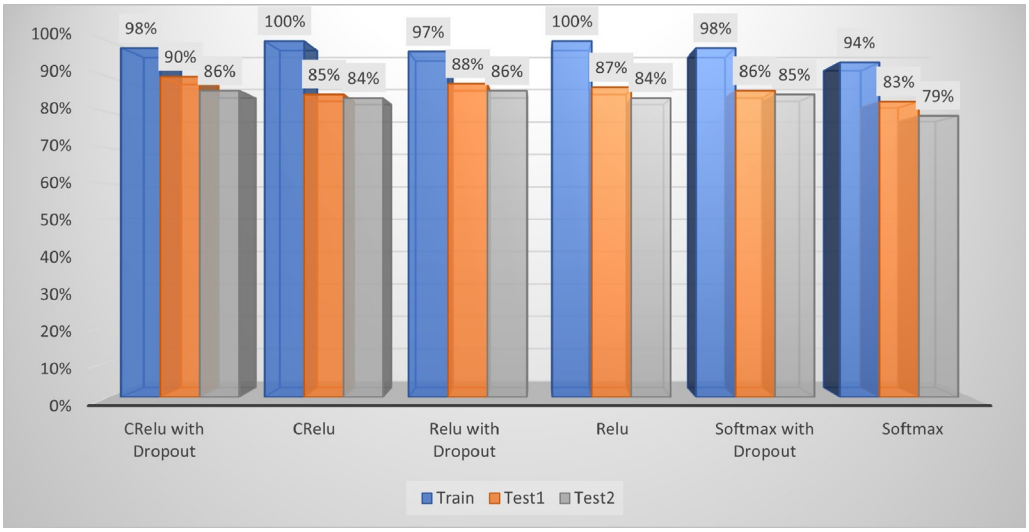



**Fig. 7.** Face recognition results on the room databases.

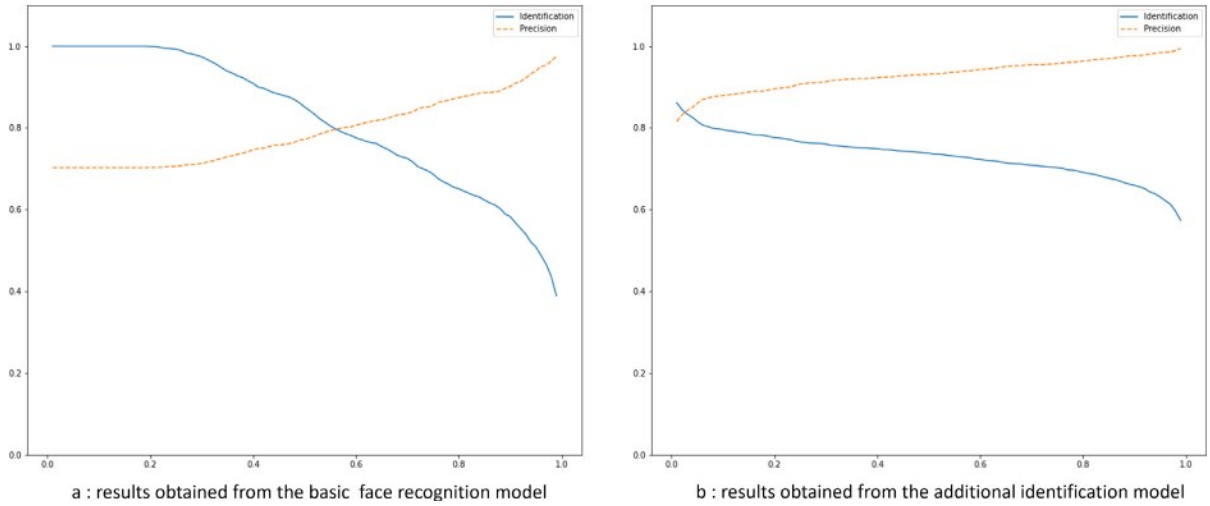



**Fig. 8.** Comparison between the use of the basic recognition model and the use of an additional identification verification module on the identification rate and the accuracy of the identification.

In Fig. 9, the identification results are presented for both approaches: the basic approach utilizing model confidence directly (right-hand side of the figure), and the proposed approach using an additional model to adjust confidence (left-hand side of the figure). Results are presented for both scenarios and for all 12 individuals. Each bar in the graph is color-coded to indicate successful identification (green) or unsuccessful identification (orange).

As shown in Fig. 9, The person identification accuracy using the proposed architecture in this work is 100%, even for individuals who are not present in the reference database (p10, p11, and p12), as the system considers them as unknown. This is essential for the system to correctly raise alerts. However, the basic architecture resulted in 6 false alarms (marked in red on the figure) out of 24 identification tests, representing an accuracy of 75%.

### 5.3. Real-world experiments

#### 5.3.1. System implementation

The proposed prototype of an unauthorized access identification system using facial recognition is implemented on a workstation with an AMD 7 3700X, 32 GB RAM, and a Nvidia GPU card (Nvidia GeForce RTX 2060 SUPER). The system runs on a Linux operating system with the Ubuntu 18.04 distribution.

#### 5.3.2. Experimental description

Twelve volunteers were selected for the experiment, with three authorized to access all equipments (group 1), three authorized to access only equipment 1 (group 2), three authorized to access only equipment 2 (group 3), and three with no authorizations (group 4). For the first nine individuals, 30 reference images were collected, resulting in a training dictionary comprising 270 images (matrices) of size (160,160,3). The monitored room has a size of 4 m $x$ 6 m. A camera is installed in the middle of the wall in front of the entrance door to ensure a total coverage of the room. The camera has a large angle (3,3 mm focal length) and a resolution of 1,8 MPixels. Two equipments are installed in this room (see Fig. 10): group 1 and 2 are allowed to access equipment 1; group 1 and 3 are allowed to access equipment 2. The experiments performed considered the lighting conditions and the level of clutter in the room. There are 4 distinct scenarios: (1) Good lighting using ceiling light; (2) Low lighting using the daylight, just

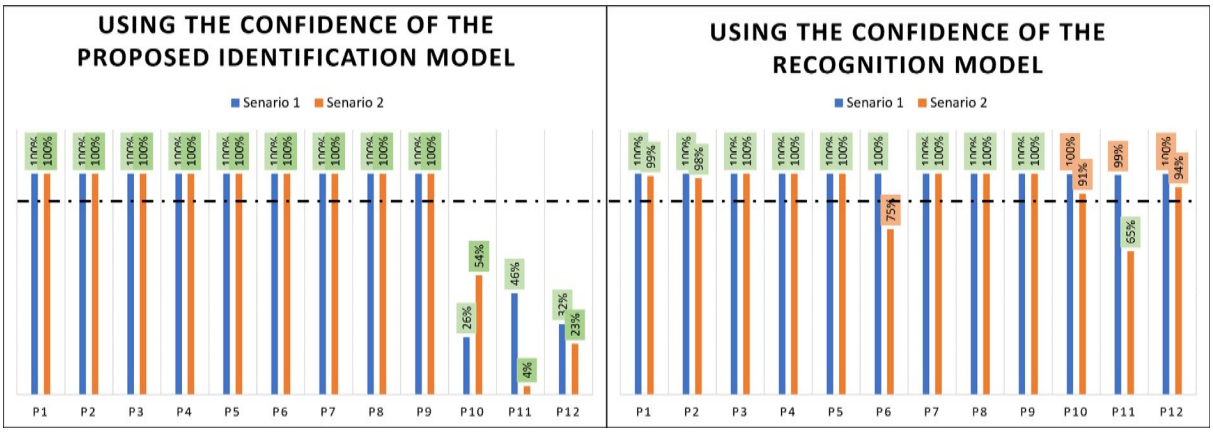



**Fig. 9.** Comparison between the use of the confidence of the basic recognition model and the confidence of the additional verification module on the final result of the identification of persons.

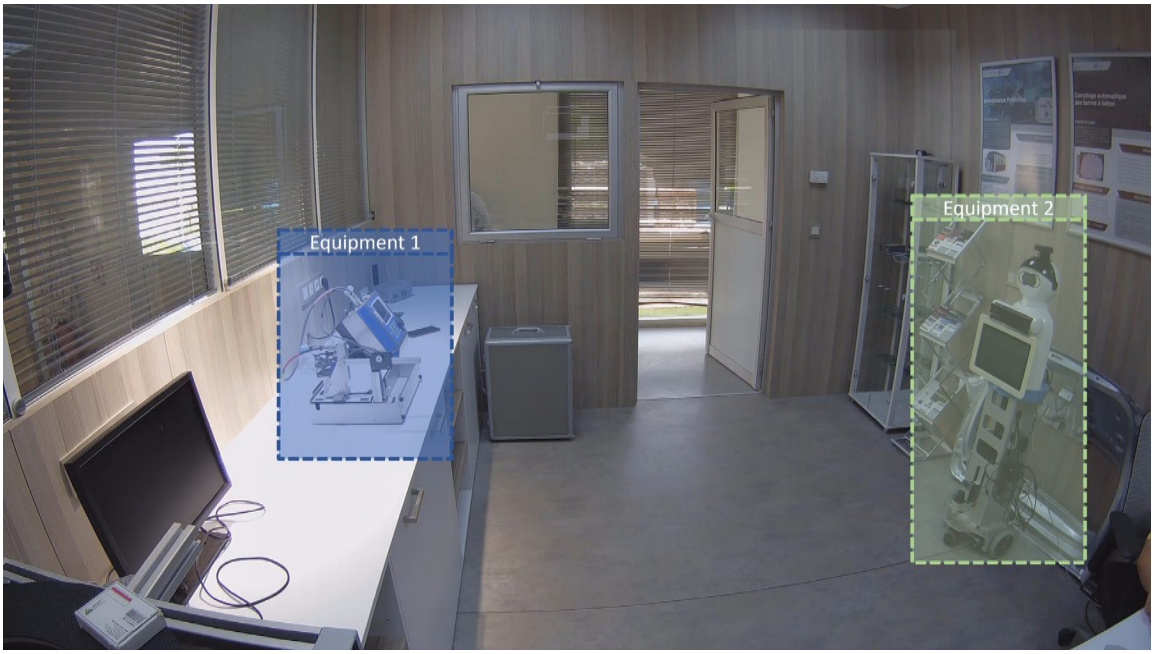



**Fig. 10.** View from the camera recording the events in the room, the two equipments to be monitored are highlighted with a blue and green rectangle respectively.

before sunset; (3) Several people moving in the room. Which makes a total of 36 videos.

Scenarios 1 and 2: a single person comes in from the door then follows a well-defined trajectory to access equipments 1 and 2, and continues his way to leave the room (see Fig. 11). scenario 3, the main person performs the same movements but with the presence of 4 additional people who move in the room in a random way (see Fig. 12)

C/C++ programming language is used for the proposed system development. OpenCV library (version: 4.5.4) is used for various modules implementation as it provides rich image processing algorithms alongside with the dnn (Deep Neural Networks) module which is used to exploit artificial intelligence models at the GPU level, using Nvidia CUDA drivers (CUDA Toolkit 11.1, and cuDNN 8.05).

### 5.3.3. Experimental results

After testing the system in real-time by executing scenarios from the MAScIR-PERSON database, it was able to detect and identify individuals and access to equipments. Table 8 summarizes the tests performed, with results presented by individuals belonging to the four groups. The "Reference" column shows the expected results. The results are then presented as the number of alerts triggered per equipment and per individual (including all three scenarios), followed by a summary of non-detections of unauthorized access and false alarms. The results are divided into two parts: the "The basic method" column for the basic method and the "The proposed method" column for the method proposed in this study.

The results demonstrate the reliability of the proposed method as it correctly identified individuals and access to equipments without any non-detections or false alarms. The proposed system is very stable under different lighting conditions and is resistant to occlusions as it works correctly when multiple individuals are present in the room. Additionally, in terms of processing time, this implementation was able to achieve real-time performance as it operates at 24 frames per second.

Overall, these results demonstrate the system's ability to accurately and efficiently recognize individuals in real-time under various conditions, making it suitable for applications such as security surveillance and access control.

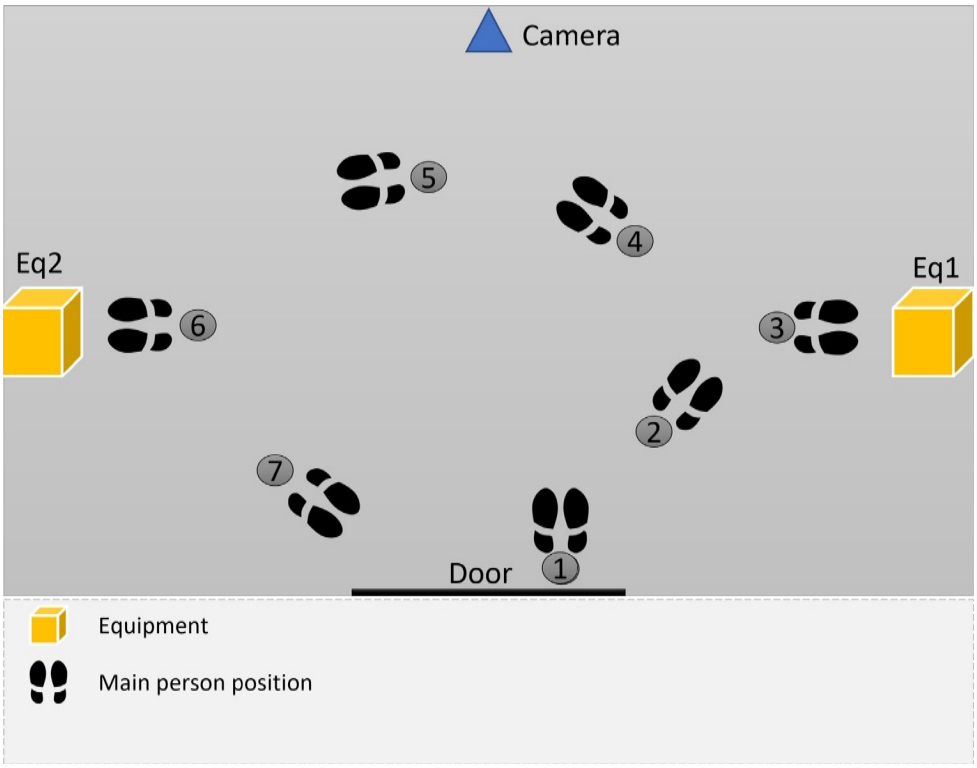



**Fig. 11.** Experimental scenario : a person moving in the room.

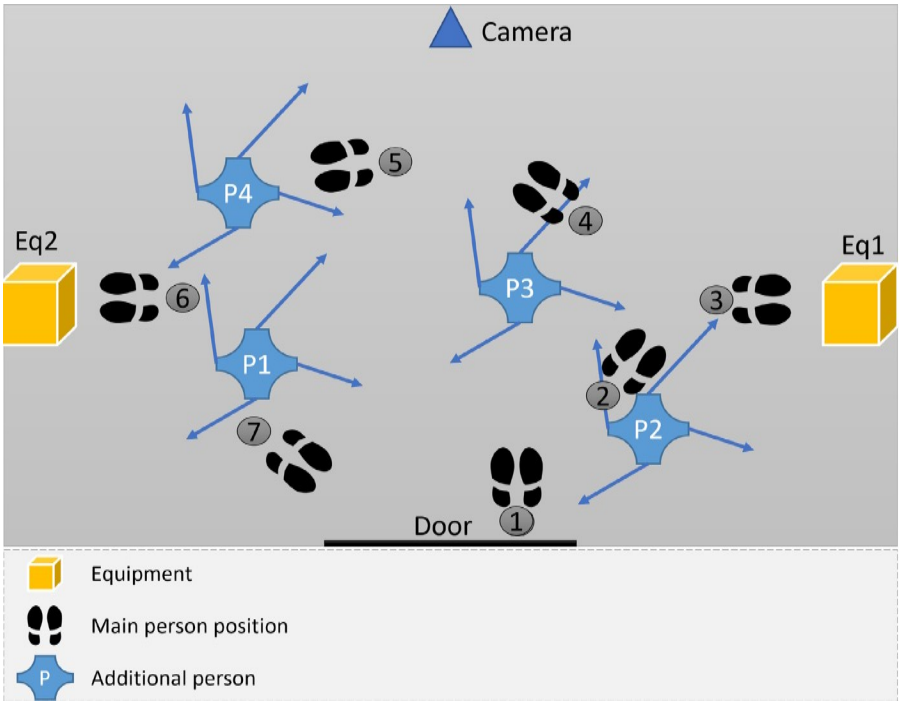



**Fig. 12.** Experimental scenario: movement of the main person in the room with other people present.

## 6. Conclusion

This article describes a new approach used to track the movement of people in a room and identify unauthorized access to equipments. The approach involves analyzing the trajectories of individuals and using facial recognition methods to identify them. The proposed system is based on a CCTV camera and a processing unit to monitor access to equipments in the room. The proposed method is a combination of algorithms optimized for real-time performance and high efficiency. This method includes a person detection algorithm, a tracking algorithm robust to occlusion, a face detection algorithm robust to orientation and scale changes, a face recognition algorithm robust to uncontrolled conditions, and to improve the robustness and stability of the identification a new algorithm is proposed based on the exploitation

**Table 8**
Results of alert generation using the complete system based on both basic methods and the proposed one.

| | | Reference | | The basic method | | | | The proposed method | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Unauthorized access alert messages | | Unauthorized access alert messages | | No-detections | False alarms | unauthorized access alert messages | | No-detections | False alarms |
| | | Equipment 1 | Equipment 2 | Equipment 1 | Equipment 2 | | | Equipment 1 | Equipment 2 | | |
| GP1 | P1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | P2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | P3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GP2 | P4 | 0 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 |
| | P5 | 0 | 3 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| | P6 | 0 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 |
| GP3 | P7 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| | P8 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| | P9 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| GP4 | P10 | 3 | 3 | 2 | 2 | 2 | 0 | 3 | 3 | 0 | 0 |
| | P11 | 3 | 3 | 0 | 0 | 6 | 0 | 3 | 3 | 0 | 0 |
| | P12 | 3 | 3 | 0 | 0 | 6 | 0 | 3 | 3 | 0 | 0 |

of information related to the general conditions of the face image. Benchmarking studies have been performed on literature databases and personalized databases adapted to the system's use context. The first study is performed to choose the algorithm for people detection, the selected algorithm is YOLOv5s. The second study allowed us to select retinaface for face detection. and finally the third study allowed to select facenet512 as feature extraction algorithm for face recognition.

Despite the promising results obtained by the proposed system, several limitations need to be highlighted. The method has been tested for a relatively small-sized room, therefore, the system's performance may decrease if the supervised area is larger and requires the use of multiple cameras. Similarly, to use this method for multi-camera mode for monitoring the same room while taking into account synchronization, additional adaptation and evaluation should be considered. On the other hand, the implementation of the proposed system requires the use of significant hardware and software resources, which may limit its deployment on a large scale and accessibility in certain contexts.

A parallel and multi-source implementation architecture is also proposed, covering the entire chain from acquisition to decision making and alerting. As a recommendation, it is advised to use and optimize lighter deep learning models in terms of execution load. This is especially applicable to the most resource-intensive modules, such as face detection and recognition, to ensure execution on lighter processing platforms.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data that has been used is confidential.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.engappai.2023.106637.

## References


Abdelali, Hamd Ait, Derrouz, Hatim, Zennayi, Yahya, Thami, Rachid Oulad Haj, Bourzeix, François, 2021. Multiple hypothesis detection and tracking using deep learning for video traffic surveillance. IEEE Access 9, 164282–164291.

Afra, Salim, Alhajj, Reda, 2020. Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people. Physica A 540, 123151. http://dx.doi.org/10.1016/j.physa.2019.123151, URL https://www.sciencedirect.com/science/article/pii/S0378437119317753.

Ahuja, Karan, Islam, Rahul, Barbhuiya, Ferdous A, Dey, Kuntal, 2017. Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recognit. Lett. 91, 17–26.

Arsenovic, Marko, Sladojevic, Srdjan, Anderla, Andras, Stefanovic, Darko, 2017. FaceTime—Deep learning based face recognition attendance system. In: 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics. SISY, IEEE, pp. 000053–000058.

Bah, Serign Modou, Ming, Fang, 2020. An improved face recognition algorithm and its application in attendance management system. Array 5, 100014.

Bashbaghi, Saman, Granger, Eric, Sabourin, Robert, Parchami, Mostafa, 2019. Deep learning architectures for face recognition in video surveillance. Deep Learn. Object Detect. Recognit. 133–154.

Bewley, Alex, Ge, Zongyuan, Ott, Lionel, Ramos, Fabio, Upcroft, Ben, 2016. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3464–3468.

Bochkovskiy, Alexey, Wang, Chien-Yao, Liao, Hong-Yuan Mark, 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Cao, Qiong, Shen, Li, Xie, Weidi, Parkhi, Omkar M, Zisserman, Andrew, 2018. Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. FG 2018, IEEE, pp. 67–74.

Cocca, Paola, Marciano, Filippo, Alberti, Marco, 2016. Video surveillance systems to enhance occupational safety: A case study. Saf. Sci. 84, 140–148.

Cui, Lisha, Ma, Rui, Lv, Pei, Jiang, Xiaoheng, Gao, Zhimin, Zhou, Bing, Xu, Mingliang, 2018. MDSSD: multi-scale deconvolutional single shot detector for small objects. arXiv preprint arXiv:1805.07009.

D'cruz, Lizzie, Harirajkumar, J., 2020. Contactless attendance system using siamese neural network based face recognition. Mater. Today: Proc.

Deng, Jiankang, Guo, Jia, Ververas, Evangelos, Kotsia, Irene, Zafeiriou, Stefanos, 2020. Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5203–5212.

Deng, Jiankang, Guo, Jia, Xue, Niannan, Zafeiriou, Stefanos, 2019. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699.

Elharrouss, Omar, Almaadeed, Noor, Al-Maadeed, Somaya, 2021. A review of video surveillance systems. J. Vis. Commun. Image Represent. 77, 103116.

Fu, Cheng-Yang, Liu, Wei, Ranga, Ananth, Tyagi, Ambrish, Berg, Alexander C, 2017. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.

Girshick, Ross, 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Girshick, Ross, Donahue, Jeff, Darrell, Trevor, Malik, Jitendra, 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Guo, Guodong, Zhang, Na, 2019. A survey on deep learning based face recognition. Comput. Vis. Image Underst. 189, 102805.

Han, Xianjun, Liu, Yanli, Yang, Hongyu, Xing, Guanyu, Zhang, Yanci, 2020. Normalization of face illumination with photorealistic texture via deep image prior synthesis. Neurocomputing 386, 305–316.

He, Kaiming, Gkioxari, Georgia, Dollár, Piotr, Girshick, Ross, 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Jocher, Glenn, Stoken, Alex, Borovec, Jirka, NanoCode012, ChristopherSTAN, Changyu, Liu, Laughing, tkianai, Hogan, Adam, lorenzomammana, yxNONG, AlexWang1900, Diaconu, Laurentiu, Marc, wanghaoyang0106, ml5ah, Doug, Ingham, Francisco, Frederik, Guilhen, Hatovix, Poznanski, Jake, Fang, Jiacong, Yuä, Lijun, changyu98, Wang, Mingyu, Gupta, Naman, Akhtar, Osama, PetrDvoracek, Rai, Prashant, 2020. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. Zenodo, http://dx.doi.org/10.5281/zenodo.4154370.

Khan, Hira, Sharif, Muhammad, Bibi, Nargis, Usman, Muhammad, Haider, Sajjad A, Zainab, Saira, Shah, Jamal H, Bashir, Yasir, Muhammad, Nazeer, 2020. Localization of radiance transformation for image dehazing in wavelet domain. Neurocomputing 381, 141–151.

Khan, Hira, Xiao, Bin, Li, Weisheng, Muhammad, Nazeer, 2022. Recent advancement in haze removal approaches. Multimedia Syst. 1–24.

Lei, Zhen, Wang, Chao, Wang, Qinghai, Huang, Yanyan, 2009. Real-time face detection and recognition for video surveillance applications. In: 2009 WRI World Congress on Computer Science and Information Engineering, Volume 5. IEEE, pp. 168–172.

Lin, Tsung-Yi, Goyal, Priya, Girshick, Ross, He, Kaiming, Dollár, Piotr, 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Ling, Hefei, Wu, Jiyang, Huang, Junrui, Chen, Jiazhong, Li, Ping, 2020. Attention-based convolutional neural network for deep face recognition. Multimedia Tools Appl. 79 (9), 5595–5616.

Liu, Wei, Anguelov, Dragomir, Erhan, Dumitru, Szegedy, Christian, Reed, Scott, Fu, Cheng-Yang, Berg, Alexander C, 2016. Ssd: Single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37.

Lopes, André Teixeira, De Aguiar, Edilson, De Souza, Alberto F, Oliveira-Santos, Thiago, 2017. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. 61, 610–628.

Lukas, Samuel, Mitra, Aditya Rama, Desanti, Ririn Ikana, Krisnadi, Dion, 2016. Student attendance system in classroom using face recognition technique. In: 2016 International Conference on Information and Communication Technology Convergence. ICTC, IEEE, pp. 1032–1035.

Lv, Xue, Su, Mingxia, Wang, Zekun, 2021. Application of face recognition method under deep learning algorithm in embedded systems. Microprocess. Microsyst. 104034.

Mahmood, Zahid, Muhammad, Nazeer, Bibi, Nargis, Ali, Tauseef, 2017. A review on state-of-the-art face recognition approaches. Fractals 25 (02), 1750025.

Muhammad, Nazeer, Khan, Hira, Bibi, Nargis, Usman, Muhammad, Ahmed, Naseer, Khan, Shahid Nawaz, Mahmood, Zahid, 2021. Frequency component vectorisation for image dehazing. J. Exp. Theor. Artif. Intell. 33 (6), 919–932.

Nassih, Bouchra, Amine, Aouatif, Ngadi, Mohammed, Azdoud, Youssef, Naji, Driss, Hmina, Nabil, 2021. An efficient three-dimensional face recognition system based random forest and geodesic curves. Comput. Geom. 97, 101758.

Oh, Sung-Kwun, Yoo, Sung-Hoon, Pedrycz, Witold, 2013. Design of face recognition algorithm using PCA-LDA combined for hybrid data pre-processing and polynomial-based RBF neural networks: Design and its application. Expert Syst. Appl. 40 (5), 1451–1466.

Olivares-Mercado, Jesus, Toscano-Medina, Karina, Sanchez-Perez, Gabriel, Perez-Meana, Hector, Nakano-Miyatake, Mariko, 2017. Face recognition system for smartphone based on lbp. In: 2017 5th International Workshop on Biometrics and Forensics. IWBF, IEEE, pp. 1–6.

Parkhi, Omkar M., Vedaldi, Andrea, Zisserman, Andrew, 2015. Deep Face Recognition. British Machine Vision Association.

Patacchiola, Massimiliano, Cangelosi, Angelo, 2017. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recognit. 71, 132–143.

Pranav, K.B., Manikandan, J., 2020. Design and evaluation of a real-time face recognition system using convolutional neural networks. Procedia Comput. Sci. 171, 1651–1659.

Qu, Zhong, Gao, Le-yuan, Wang, Sheng-ye, Yin, Hao-nan, Yi, Tu-ming, 2022. An improved YOLOv5 method for large objects detection with multi-scale feature cross-layer fusion network. Image Vis. Comput. 125, 104518.

Rahouma, Kamel Hussein, Mahfouz, Amal Zarif, 2021. Design and implementation of a face recognition system based on API mobile vision and normalized features of still images. Procedia Comput. Sci. 194, 32–44.

Redmon, Joseph, Divvala, Santosh, Girshick, Ross, Farhadi, Ali, 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, Joseph, Farhadi, Ali, 2017. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271.

Redmon, Joseph, Farhadi, Ali, 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, Shaoqing, He, Kaiming, Girshick, Ross, Sun, Jian, 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Sun, Yi, Chen, Yuheng, Wang, Xiaogang, Tang, Xiaoou, 2014. Deep learning face representation by joint identification-verification. Adv. Neural Inf. Process. Syst. 27.

Sunaryono, Dwi, Siswantoro, Joko, Anggoro, Radityo, 2021. An android based course attendance system using face recognition. J. King Saud Univ. -Comput. Inf. Sci. 33 (3), 304–312.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, Rabinovich, Andrew, 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Taigman, Yaniv, Yang, Ming, Ranzato, Marc'Aurelio, Wolf, Lior, 2014. Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1708.

Ullah, Rehmat, Hayat, Hassan, Siddiqui, Afsah Abid, Siddiqui, Uzma Abid, Khan, Jebran, Ullah, Farman, Hassan, Shoaib, Hasan, Laiq, Albattah, Waleed, Islam, Muhammad, et al., 2022. A real-time framework for human face detection and recognition in cctv images. Math. Probl. Eng. 2022.

Wagh, Priyanka, Thakare, Roshani, Chaudhari, Jagruti, Patil, Shweta, 2015. Attendance system based on face recognition using eigen face and PCA algorithms. In: 2015 International Conference on Green Computing and Internet of Things. ICGCIoT, IEEE, pp. 303–308.

Wang, Mei, Deng, Weihong, 2021. Deep face recognition: A survey. Neurocomputing 429, 215–244.

Wang, Hongjun, Hu, Jiani, Deng, Weihong, 2017. Face feature extraction: a complete review. IEEE Access 6, 6001–6039.

Wasnik, Pankaj, Raja, Kiran B, Ramachandra, Raghavendra, Busch, Christoph, 2017. Assessing face image quality for smartphone based face recognition system. In: 2017 5th International Workshop on Biometrics and Forensics. IWBF, IEEE, pp. 1–6.

Wen, Yandong, Zhang, Kaipeng, Li, Zhifeng, Qiao, Yu, 2019. A comprehensive study on center loss for deep face recognition. Int. J. Comput. Vis. 127 (6), 668–683.

Wiliem, Arnold, Madasu, Vamsi, Boles, Wageeh, Yarlagadda, Prasad, 2007. A face recognition approach using Zernike moments for video surveillance. In: Recent Advances in Security Technology: Proceedings of the 2007 RNSA Security Technology Conference. Australian Homeland Security Research Centre, pp. 341–355.

Zennayi, Yahya, Bourzeix, Francois, Guennoun, Zouhair, 2022. Analyzing the scientific evolution of face recognition research and its prominent subfields. IEEE Access 10, 68175–68201.

Zhang, Shifeng, Chi, Cheng, Lei, Zhen, Li, Stan Z., 2020. Refineface: Refinement neural network for high performance face detection. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4008–4020.

Zhang, Kaipeng, Zhang, Zhanpeng, Li, Zhifeng, Qiao, Yu, 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23 (10), 1499–1503.

Zhao, Xiaonan, Li, Jingjing, Liu, Wenqiang, Zhang, Junjie, Li, Yang, 2020. Design of the sleeping aid system based on face recognition. Ad Hoc Netw. 99, 102070.

Zheng, Liwen, Fu, Canmiao, Zhao, Yong, 2018. Extend the shallow part of single shot multibox detector via convolutional neural network. In: Tenth International Conference on Digital Image Processing, Volume 10806. ICDIP 2018, SPIE, pp. 287–293.