SAPIENZA
UNIVERSITÀ DI ROMA

Biometric Systems a.y 22/23

Insight CCTV Face Recognition

Hazem Dewidar - dewidar.1883881@studenti.uniroma1.it
Riccardo Mancini - mancini.1905638@studenti.uniroma1.it
Davide Gabrielli - gabrielli.1883616@studenti.uniroma1.it

February 12, 2023

1 Introduction

1.1 Abstract

In this project[1] we tackled the problem of identification in a video captured by CCTV (closed-circuit television) in an open set context. In order to achieve our goal, we divided the problem in two phases: tracking a person giving them a temporary ID and then identifying a tracked person using their best shots. For this project we have decided to use the ChokePoint dataset [2].
The main idea was to use a model that, given two faces, it would output if they refer to the same person or not. For what concerns the tracking part, the intuition was that, given two close enough times t_1 and t_2, the face of the subject would not change much even if it was blurred or in low quality. Using this intuition, we can build a "chain" of a subject's face changing over time, tracking him and lastly trying to identify him with the five best shots of the subject.
CCTV face recognition is important because it can provide enhanced security and improve the efficiency of security systems by enabling quick and accurate identification of individuals in real-time. It can be used in a variety of settings, such as airports, banks, and public spaces, to detect and prevent criminal activity, monitor access to secure areas, and track individuals for investigations.

1.2 Identification vs Re-identification

In the context of CCTV face recognition, identification refers to the process of recognizing a face in a video or image and determining the identity of the person by matching it against a gallery of known individuals. In our project we used face identification in an open set scenario, that is, the system determines if the face that is being processed (probe) belongs to a subject in our gallery. It may happen that some subjects don't belong to the gallery so we label them as unknown.
Re-identification, on the other hand, is the task of recognizing the same person across various scenario

without linking it to an actual identity. We are not interested in who is that person but rather in knowing where a subject is across the footage that we have.

1.3 Related Work

Face Recognition Before the widespread use of neural networks in computer vision, face recognition and face detection (e.g. Viola-Jones) were mostly based on traditional computer vision and machine learning techniques. Some of these methods are:

- **Eigenfaces**: Eigenfaces involves reducing the dimensionality of face images using Principal Component Analysis (PCA) and representing each face as a set of weights for the eigenvectors.
- **Feature-based methods**: Another approach was to extract facial features such as eyes, nose, mouth, etc. and then use these features to identify and match faces.

These methods were limited by their reliance on hand-crafted features and their susceptibility to variability in lighting, pose, and expression. Neural networks and deep learning have since revolutionized the field of face recognition and detection, thus having more accuracy.
In addition, there is an ongoing work on siamese neural networks on the field of face recognition [3] [4]. The main idea is to train two identical sub-networks with shared weights, where each sub-network processes one of the inputs and the final layer of the network is used to compute a similarity score between the two features that are extracted from the inputs.

The Open-Set Problem Open-set recognition [5] consists in recognizing subject from a gallery in a context where we have to deal with probes that are both unknown (they do not belong to the gallery) or known (they are enrolled in the gallery). Open set face recognition is more suitable for real world applications where the number of possible identities is not limited and can change over time. Furthermore, we do not need to retrain the model every time a new subject is enrolled because the model is designed to handle both known and unknown individuals. The model has been trained on a large diverse set of faces and has learned to recognize and differentiate between different individuals.

1.4 Dataset

This is video dataset, designed for experiments in person identification/verification under real-world surveillance. An array of three cameras was placed above several portals (natural choke points in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way. While a person is walking through a portal, a sequence of face images (ie. a face set) can be captured. Faces in such sets will have variations in terms of illumination conditions, pose, sharpness, as well as misalignment due to automatic face localisation/detection. Due to the three camera configuration, one of the cameras is likely to capture a face set where a subset of the faces is near-frontal (see fig 1). The dataset consists of 25 subjects (19 male and 6 female) in portal 1 and 29 subjects (23 male and 6 female) in portal 2. The recording of portal 1 and portal 2 are one month apart. The dataset has

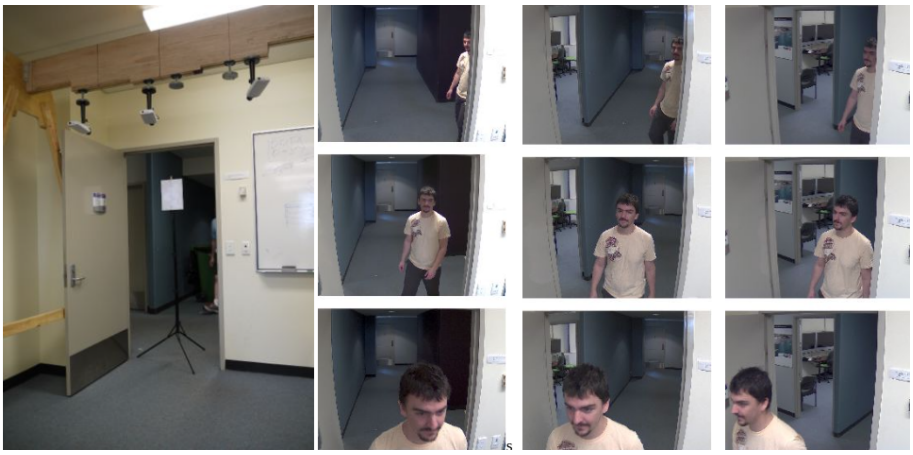

Figure 1: Setting example

frame rate of 30 fps and the image resolution is 800X600 pixels. In total, the dataset consists of 48 video sequences and 64,204 face images.

Each sequence was named according to the recording conditions (eg. P2E_S1_C3) where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate subjects either entering or leaving the portal. The numbers indicate the respective portal, sequence and camera label. For example, P2L_S1_C3 indicates that the recording was done in Portal 2, with people leaving the portal, and captured by camera 3 in the first recorded sequence.

To pose a more challenging real-world surveillance problems, two seqeunces (P2E_S5 and P2L_S5) were recorded with crowded scenario. In additional to the aforementioned variations, the sequences were presented with continuous occlusion. This phenomenon presents challenges in identity tracking and face verification according to the dataset creators. The gallery is composed by the still photos of the subject in controlled environment. Two photos were taken, one with a natural expression (see fig: 2) and one while they are smiling.

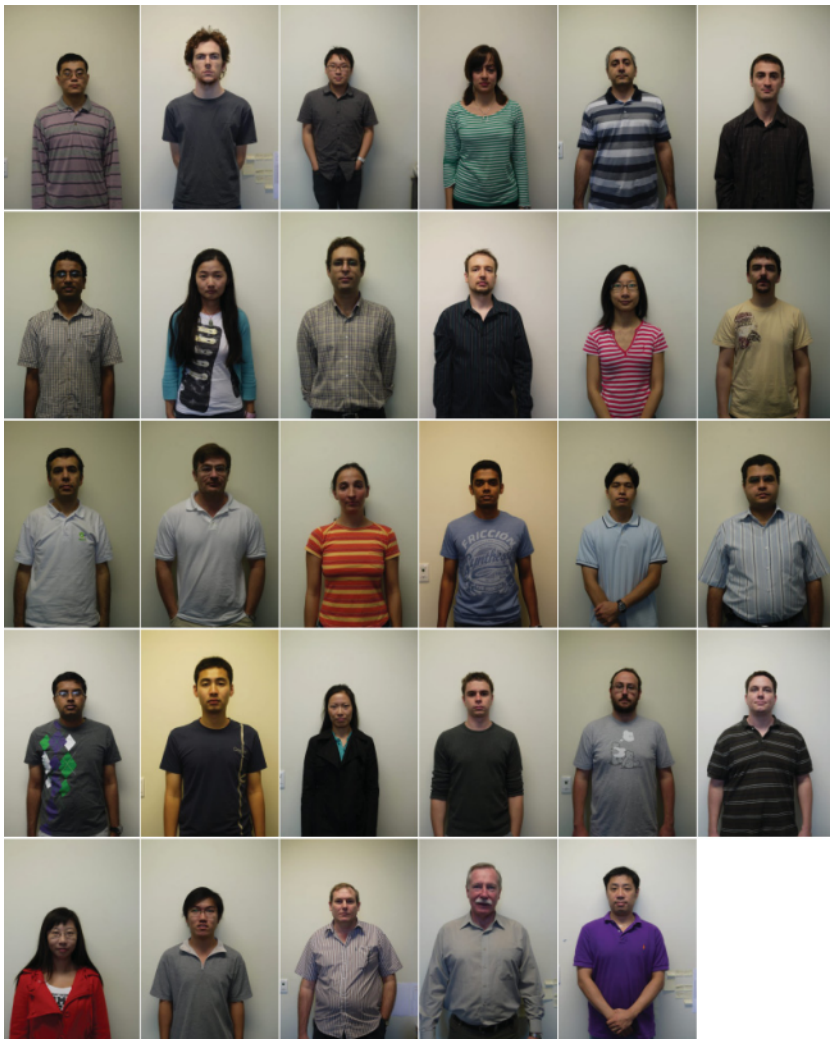

Figure 2: The gallery of the ChokePoint Dataset while the subjects have a natural expression

2 Models

2.1 InsightFace

InsightFace is an open-source project that aims to provide state-of-the-art solutions for face recognition. It is developed by the Deep Insight research group and the project is hosted on GitHub [6]. InsightFace provides pre-trained models and algorithms for various face recognition tasks, including face detection, alignment, embedding, and classification. The models in InsightFace are trained on large-scale face recognition datasets, such as the MS-Celeb-1M dataset and are based on deep neural networks, such as ResNet . We used various tools from InsightFace for this task such as:

- **SCRFD**: SCRFD is used for detecting faces in a frame and it is proposed in "Sample and

Computation Redistribution for Efficient Face Detection" [7]. It gives bounding boxes and key points.

- **ArcFaceONNX**: ArcFace is a deep learning model for face recognition, proposed in the paper "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" [8]. The model uses a loss function called "Additive Angular Margin Loss" to increase the separation between features of different classes in the feature space.

2.2 Model Proposed

Our model (see fig 3) takes as input each frame of the video to analyze. We have three perspectives (cameras), as we can see in Fig. 1. For each frame, we get the three images obtained by the three cameras and process them in the following manner:

- **Detect the faces**
- **Track the subject:**
  - If the subject hasn't been seen before, create a new temporary ID and start building a chain of faces and frames with the subject (see fig4)
  - If the subject has left the scene, try to identify him
  - If none of the conditions above are satisfied then continue to build the chain.

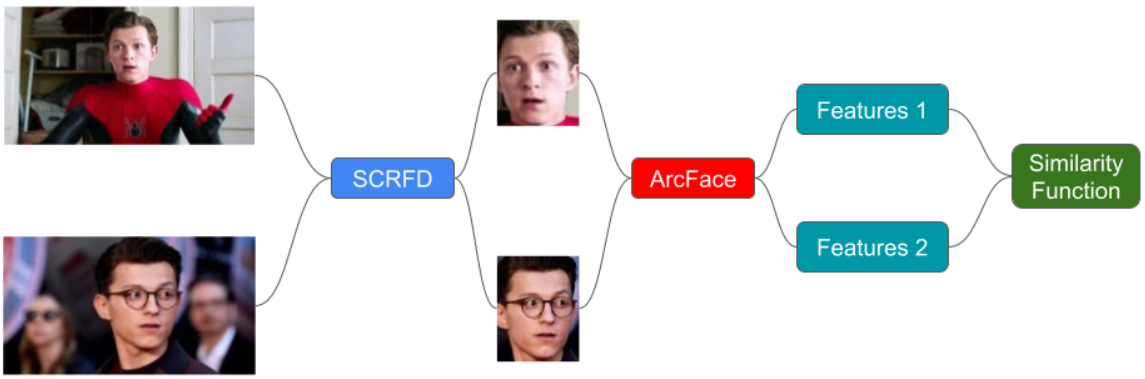

Figure 3: Model pipeline

2.3 Face Recognition

Face recognition can be applied to compare the detected faces against a database of known individuals (gallery) in order to determine the identity (or a temporary one) of the person. We use features that are extracted by a deep network model from the faces. This representation can then be compared to

representations of known individuals to make a recognition decision linking the subject to a temporarily ID or an identity in the gallery. They are learned representations of the face that are obtained by passing the face image through ArcFace.
Let's see in more detail how our system works (also see fig 3). It's divided in two macro phases:

Tracking This is the first phase. Once we have detected a face, we compare it to the unknown identities in order to build a chain of how that person's face changes over time (see fig 4). The main intuition is that given two close enough frames, the features of the faces would not change much. In this way, we can concatenate the faces in list. For each face that is not similar to any the temporary IDs (we check if he's similar to any of the last saved faces for each unknown identity), we create a new identity for that subject in order to begin to build his chain. In order to do this, we have an *unknown similarity threshold*. If the maximum similarity score is below this threshold, we create a new identity. Once a subject is not in any camera of the cameras array for a given time, the system proceeds to identify them in the second phase. See figure 6 for a summary diagram.

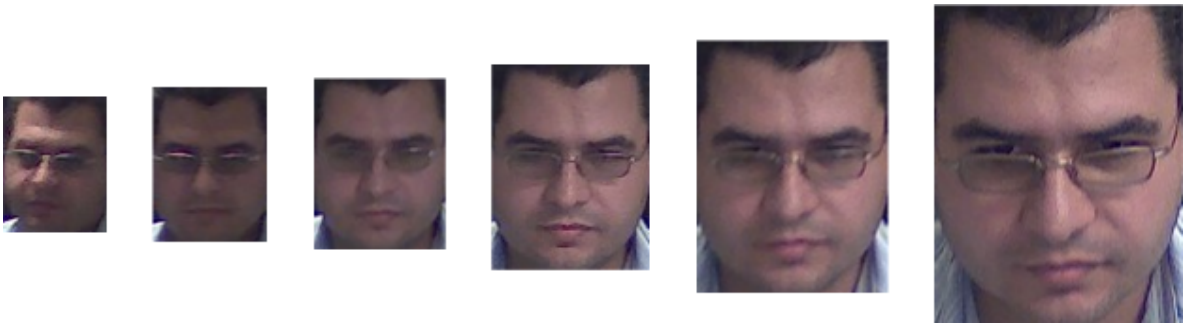

Figure 4: Example of a temporarily ID chain

Identification In this second phase, we proceed to identify the subject. We loop through each of the unknown identities, check if they are out of the array of camera of a given time and then eventually identify them. For each image in the chain of each identity, we search for the best ones, that is, those having the biggest bounding boxes. Then, for each biggest face we got in the previous step, we match with the templates in the gallery and return the identity of the subject who's template is the most similar to the probe. Once this step is completed, we give the same identity to all the frames in the chain. In this way we can assign an identity also to the faces in the frames that aren't good for identification but good enough for re-identification. In this case, we use a *gallery similarity threshold*. If the maximum similarity score is below this threshold, we will label them as unknown. See figure 7 for a summary diagram and figure

2.4 More details of the implementation

As we have already told, two modules were taken from the Deep Insight team. All the utilities that we use to interface with their models were heavily inspired by them. We tweaked them to best fit our needs. For example, we have separated the feature extraction function from the detection. We have have also separated their similarity function.

We have developed a system that given a video from a CCTV and a gallery, is able to work in real time, showing the tracking information (temporary ID) and eventually the identity of the subject once it has been recognized. You can see a snippet of how the system works (see figure 5)

Figure 5: Snapshot of our system in execution. **A)** The identity 1 and 0 are being tracked to be later identified. **B)** The previous tracked identity 1 has been identified in the gallery as subject 0003 while id 0 is an intruder in the facility

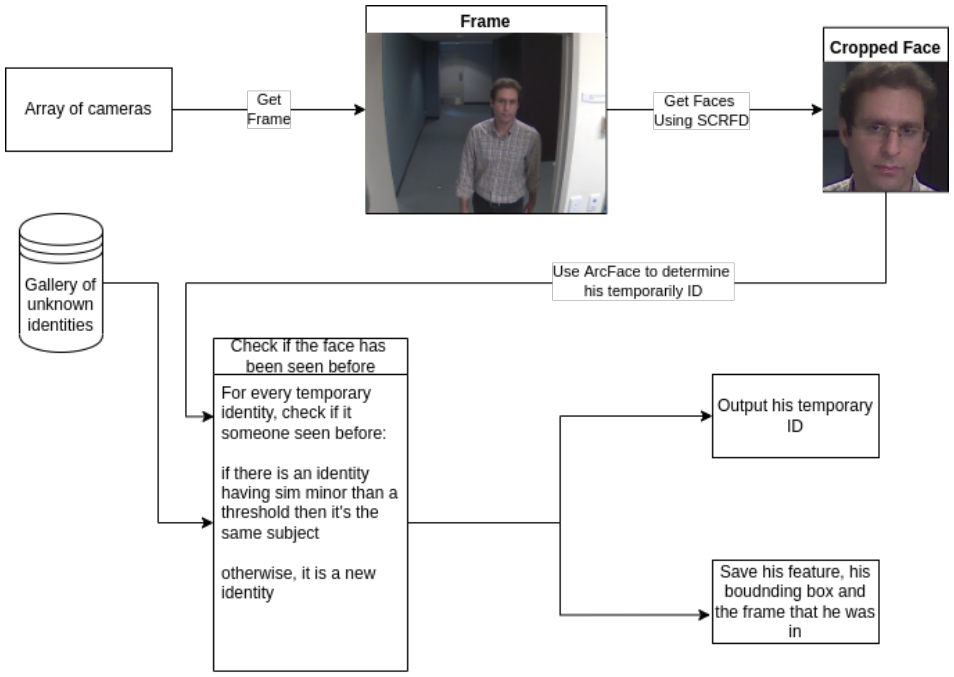

Figure 6: This diagram shows the execution flow of how we track a subject

3 Experiments

3.1 How the performance was evaluated

The ChokePoint dataset provides a ground-truth for each scenario. We have modified it by using both an automatized method and by hand. Every change made automatically was reviewed manually. The original ground-truth had a lot of missing information such as the presence of intruders (subjects that are not found in the "global" gallery). Furthermore, some of the original detected faces in the ground-truth were incomplete: entering or leaving a portal in the first/last frames for each subject were not labelled in the ground-truth but our detector would rightfully detect them. Each scenario had a subset of the global gallery (and sometimes some unknown person).
For each scenario, we have built our gallery by dividing the subjects that would appear in the scenario in two groups: intruders and authorized people. Then, for each threshold, we would compute the of false acceptance rate, false rejections rate and DIR by processing all the frames.

3.2 Metric of interest

Below, we will briefly talk about what we have measured. Measuring the accuracy isn't enough and it doesn't give a complete picture of a biometric system reliability since the consequences of incorrect results can be severe, such as security and access control.
The accuracy metric only considers the total number of correct predictions, but it doesn't take into

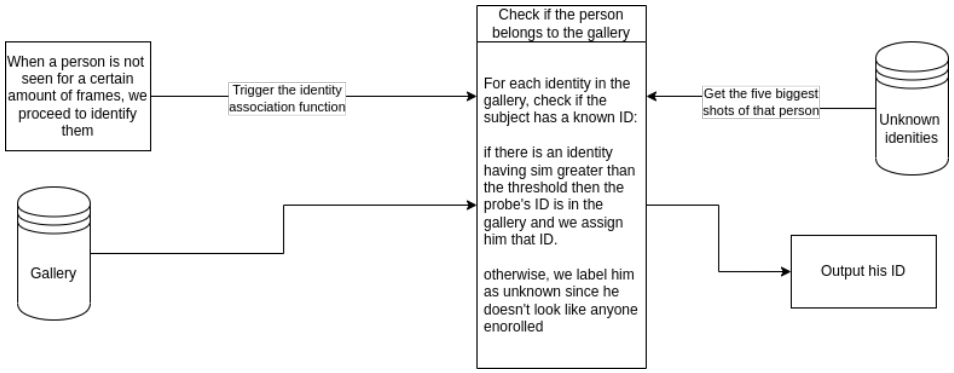

Figure 7: This diagram shows the execution flow of how we identify a subject

account the distribution of False Acceptance (incorrectly accepted instances) and False Rejections (incorrectly rejected instances). In our specific case, we are operating on a white list so we need to detect intrusions, that is, we need to know if a person is not authorized (it's not in the gallery). We would rather have a low false acceptance rate instead of a low false rejection rate. While having an alert status is of course annoying, having an intrusion is a serious security breach in a facility.

- **False Rejection Rate (FRR)**: False Rejection Rate is the rate at which genuine faces are incorrectly rejected. It is defined as the ratio of the number of False Rejections to the number of Genuine Probes:

$$FRR = \frac{False\ Rejections}{Genuine\ Probes} = 1 - DIR(1, t) \tag{1}$$

- **False Acceptance Rate (FAR)**: False Acceptance Rate is the rate at which impostor faces are incorrectly accepted. It is defined as the ratio of the number of False Acceptances to the number of Impostor Probes:

$$FAR = \frac{False\ Acceptances}{Impostor\ Probes} \tag{2}$$

- **Error Equal Rate (EER)**: Error Equal Rate is the threshold value at which both False Rejection Rate (FRR) and False Acceptance Rate (FAR) are equal. It provides a measure of the performance of the face recognition system when the threshold is set such that both error rates are equal.

3.3 Results

We have carried out different tests on different scenario, varying the threshold and by changing the impostors for each scenario. In the original interface provided by insightFace, there were two thresholds:

- **0.2**: under this threshold, the system would output that two faces are not from the same person.

- **0.28**: above this threshold, the system will output that they are from the same person.

If the similarity score is between the two thresholds, it would output that they are likely from the same person. After analyzing the results from different scenarios, the error equal rate has been found to be around 0.3 (see figure 8). This confirms the thresholds suggested by the deep insight team.
Although the definition of EER is the point where FAR and FER are the same, we were limited by the computational power at our disposal thus we used only a step of 0.1 (even with this low resolution, the tests took around 10 hours).
The values of FAR and FRR at 0.3 are:

- FAR = 0.025
- FRR = 0.019

As we can see in the table 1, the DIR at rank 1 with threshold 0.3 is quite high and it keeps having a high value up until threshold 0.5. This means that in a scenario where having a low FAR is more important, our method can keep having good identification rates (with threshold 0.5 the FAR is 0.011 and FRR is 0.032). Furthermore, starting from the threshold 0.3 we start to have just one identity above the threshold. This can be a good thing if we look at it from the prospective that, if the identity at rank 0 wasn't enrolled, no one would have got above the threshold.

Threshold	**Rank 1**	**Rank 2**	**Rank 3**
0.1	0.9962	0.9978	0.9981
0.2	0.9881	0.9995	-
0.3	0.9774	-	-
0.4	0.9737	-	-
0.5	0.9474	-	-
0.6	0.8477	-	-
0.7	0.5225	-	-
0.8	0.0491	-	-
0.9	0.0000	-	-

Table 1: DIR scores for each rank at different threshold values.

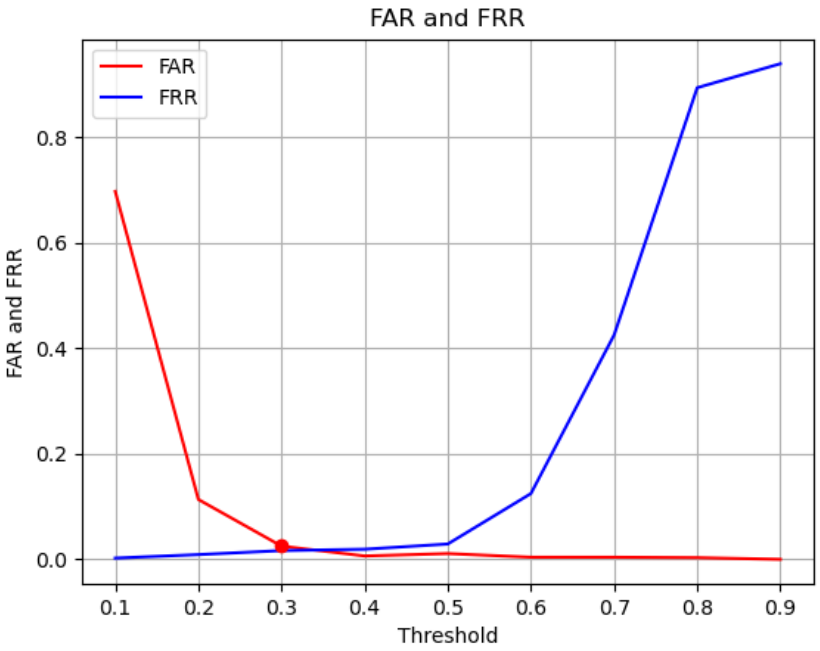

Figure 8: FAR and FER. The Error Equal Rate is highlighted by a big red dot

4 Conclusions

In conclusion, our system has shown its effectiveness in tracking an individual across frames to identifying them by using the best shots. We have used state of the art models in order to detect, extract features and to compare faces. Our model can be used in a great variety of fields ranging from law enforcement to home security. We have achieved good results in FAR and FRR, achieving a FRR of 0.029 and a FAR of 0.0109 in case we want to to prioritize the rejections of intruders.

4.1 Future development

There are a more development that can be made to improve the system such as:

- Having a better frame selection policy, maybe using a deep model as seen in the faceQnet [9]
- Improving the image quality through techniques of computer vision
- Expand the tracking module to enable not only the tracking through the face but also using other elements such as cloth and gait.

References

[1] *Project repository*. URL: https://github.com/davegabe/Suspect-Identification-CCTV.

[2] *ChokePoint dataset*. URL: https://arma.sourceforge.net/chokepoint/.

[3] Gregory R. Koch. "Siamese Neural Networks for One-Shot Image Recognition". In: 2015.

[4] Haoran Wu et al. "Face recognition based on convolution siamese networks". In: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2017, pp. 1–5. DOI: 10.1109/CISP-BMEI.2017.8302003.

[5] Manuel Günther et al. *Toward Open-Set Face Recognition*. 2017. DOI: 10.48550/ARXIV.1705.01567. URL: https://arxiv.org/abs/1705.01567.

[6] *InsightFace GitHub Project*. URL: https://github.com/deepinsight/insightface.

[7] Jia Guo et al. "Sample and Computation Redistribution for Efficient Face Detection". In: *arXiv preprint arXiv:2105.04714* (2021).

[8] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *CVPR*. 2019.

[9] Javier Hernandez-Ortega et al. "Faceqnet: Quality assessment for face recognition based on deep learning". In: *2019 International Conference on Biometrics (ICB)*. IEEE. 2019, pp. 1–8.