Digital Forensics
Project 3 - Face Recognition

Francesco Agostini, Nicola Gastaldello
Department of Information Engineering
University of Padua

A.Y. 2019-20

Abstract

Face Detection and Face Recognition have become interesting areas in research application of pattern recognition and computer vision during the last years. They play a vital role in several applications, from surveillance systems to criminology. However, the Face Recognition task is still complex in computer vision, especially due to the high degree of varation of human faces. The main aim of this work is to present an Artificial Intelligence (AI) - based face recognition system adapted to the temporal dimension for videos face recognition. After a brief introduction, some theorethical insights regarding the face recognition task and CNNs will be proposed. Then, results and the performance of our face recognition strategy will be presented and discussed. We conclude this work with some discussions on how the strategy could be further imporved.

1 Introduction

In the last years, the improvement of technology have made possible the development of real-time vision modules interacting with individuals. Object detection is one of the main task which has recently become interesting in research applications, and it is connected to computer vision and image processing. It basically consists in detecting instances of objects, which can be taken from video frames or digital images, from specified class: human faces, cars, buildings etc. Specifically, face detection algorithms aim to determine automatically whether there is a human face in an image or not[**2**].
Thus, face detection is one of the main task in which the research have focussed on in the last years: several pattern recognition methods have been proposed so far, aiming to detect faces with different accuracy, both in given images or real time surveillance systems. Moreover, it is important to highlight that machine learning represents the main tool to achieve this task in static and video mode. The ever increasing attention on the face detection problem is also due to the wide range of its usage aplications in law enforcement and commerce. In fact, face detection is the first phase in many face processing systems, such as automatic focusing on cameras, driver drowsiness detection in cars, criminal identification, access control, face recognition etc[**2**]. Together with face detection, face recognition has received a great deal of attention over the last years, due to its many applications in various domains. Just think of the several biometric-based techniques emerged recently; these techniques, based on the examination of the individual's physiological characteristics, are considered as the most promising for the recognition of individuals, instead of authentication based on passwords, PINs, keys and so forth, which can be stolen, forgotten or guessed and are hard to remember. The biological traits, instead, cannot be stolen, forged or misplaced.
Another advantage of face recognition over biometric methods is the following: voluntary actions by the user (i.e., place the hand for fingerprint, keeping a fixed position in front of the camera for retina identification etc.) are not required. Since face images can be acquired from a distance, face recognition is a task that can be performed passively, and this is beneficial, especially for surveillance and security purposes. Moreover, for other biometrics, data acquisition is affected by several problems, such as

damaged epidemic tissue for fingerprints etc. In addition, retina-based identification systems are very sensitive to any body motion and require expensive equipment. However, face recognition can be performed with inexpensive fixed cameras, while ad-hoc algorithms can compensate for noise and variation in illumination and scale[**1**].
Convolutional Neural Networks (CNNs) have significantly improved the state of the art in several applications[**5**]. However, since neural networks require large quantities of training data, and since large scale public datasets have been lacking, most of the advances in the world of face recognition remain restricted to Internet giant, such as Google and Facebook, which has the capability of building such large datasets[**1**]. In addition to this problem, the challenging issue in face detection and recognition is linked to diversity in faces, such as shape, color, texture etc. Furthermore, the photographic occurrence could cause additional differences in terms of lightning conditions, facial expressions and head pose.
The rest of this work is organized as follow: section 2 offers some theorethical insights about CNNs; section 3 presents a general overview on the face recognition task, showing the general strategy to adopt; in section 4 we describe our processing pipeline, and describes the creation of the dataset, the training of the CNN and the algorithms we used. In section 5 we report and discuss the obtained results, and do some performance analysis. Finally, we conclude this work with section 6.

2 Convolutional Neural Networks: overview

As a discipline, Digital Forencsis systems have been evolving, like artificial intelligence (AI) and its subcategories. To make digital forencsis applications to be predictive and proactive, the use of AI is needed. At the state-of-art, there are a lot of attempts to use AI to improve this field of study, because digital forensics is becoming increasingly important, and often requires the intelligent analysis of large amounts of complex data. The main application are developed in problems like face detection and recognition, packet classification, law enforcement and more other.

AI is a branch of computer science whose goal is the simulation of the human intelligence over machines. The most developed application of AI developed in recent times is machine learning (ML) and it aims to create algorithms which can automatically learn and improve from the past experience, without explicit programming. There are three main subclasses of ML:

- **Supervised learning**: is the task of learning a function, based on input-output pairs, which maps an input to an output. Therefore, given training data, which are composed by the samples and the labels (correct answer), the algorithm will try to create the map function between the samples and the label.
- **Unsupervised learning**: is the task of leaning as more as possible from the data, since we have the input data but not the corresponding output; thus, the model doesn't need to be supervised.
- **Reinforcement learning**: is the task of finding a tradeoff between exploration and exploitation, namely between the unknown knowledge and the current knowledge. Therefore, there will not be input/output pairs, but the algorithm will take decision based on a target to reach, in order to maximize reward in a special situation.

An important class of supervised learning is Deep Learning (DL), whose main feature is to use artificial neural networks (ANNs) in order to extract high level features from the input. ANNs are computing systems inspired by human brain, namely an interconnected group of nodes. The nodes are artificial neurons and they are arranged in several layers; the artificial neurons positioned in differents layers are linked between them.

In this work we use Convolutional Neural Networks (CNNs), a category of Neural Networks designed for processing structured data such as images. The main difference between a feed-forward neural network and a CNN comes from a special kind of layer, called convolutional layer, which is capable of recognizing sophisticated shapes[**6**]. A typcal CNN architecture is shown in Figure 1.

CNNs have been successful in natural language processing for text classification and they are widely used in computer vision, for example in faces classification and recognition, which is the aim of our work.

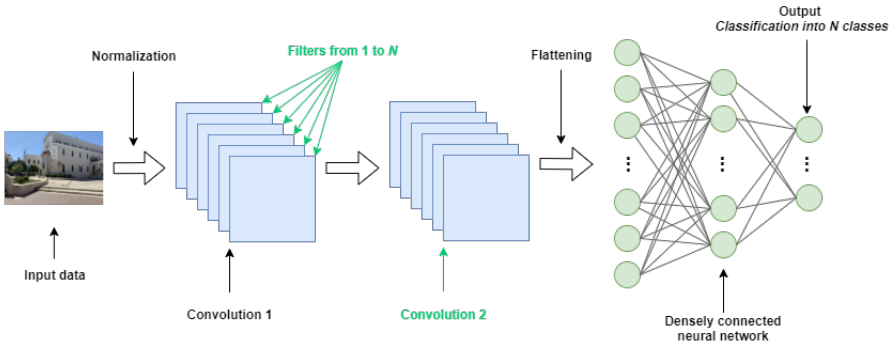

Figure 1: Traditional CNN design.

3 Face Detection and Face Recognition: overview

3.1 Face Recognition Applications

As mentioned in the section above, the Face Recognition technology has many applications in various domains. However, we can classify all these application usages into two primary tasks:

- **Identification (one-to-many matching)**: in this case, the image of an unknown individual is given, and the task is to determine the person's identity, performing a comparison between the given image and a database of images of known individuals
- **Verification (one-to-one matching)**: it basically consists on identity verification. Given a face image of an unknown person along with a claim of identity, the objective is to ascertain whether the individual is who she/he claims to be.

Some of the various application areas which exploit face recognition systems are listed below:

- Surveillance: with the installation of CCTVs it is possible to keep monitored and to look for drug offenders, sex offenders, abductors and known criminals, also notificating the authorities when one of them is located.
- Security: several access control systems to biuldings, seaports, airports, as well as border checkpoints and ATM machines exploit the face recognition technology. In addition, it is also used in network security and email authentication on multimedia workstations.
- Criminal justice systems: forensics, post-event analysis etc.
- General identity verification: banking, e-commerce, national IDs, electoral registration, passports, driving licenses, employee IDs.
- Image database investigations: for immigrants control, police bookings, benefit recipients etc.
- Smart Card applications: the face-print can be stored in smart cards, bar codes or magnetic stripes, and their authentication is performed by matching the live image with the stored one.
- Expression recognition, gender classification, facial feature recognition: for intensive care monitoring in the field of medicine, or tracking driver's eyes and monitoring his fatigue and stress detection.

3.2 Face Recognition - Typical Issues

FD algorithms can tolerate several performance affecting factors, including facial expression, occlusion, posture, lack or existance of structural elements, illumination, camera distorsion and noise, resolution, image orientation and the time and speed of computation[**4**]. Some of these issues are briefly discussed below:

- Facial expression: one of the most temperament, influential and instant means for human to converse their meanings and emotions. The appearence of faces, like happiness or angriness, is strictly related to the facial expression and directly impact on an individual's face.
- Head pose: on images, the location of the face can vary, due to different profile or frontal plane rotation.
- Image rotation: the orientation of an image depends on the nature of the image itself. It may appear rotated, upside-down or inversed from left to right.
- Occlusion: we have occlusion when faces on images result partially or fully covered, due to the presence of other elements.
- Computation time and speed: the execution time is a critical factor, especially for real time application systems, which require very fast processing. Numerous researches have been carrieid out n face detection to enhance the performance of recognition in terms of execution time, as well as accuracy.
- Illumination: it is another crucial factor in determining the quality of images and thus it can have much effect on the evaluation, the detection and the recognition tasks. Moreover, when facial images are taken under natural environment, the illumination condition could be drastic and the image background could be complex.

Although many of the current FR systems work well under constrained conditions (i.e., at least a few of these factors are controlled), their performance degrades rapidly when none of these factors are regulated.

3.3 Face Recognition - Generic Framework

The methodology for acquiring face images depends on the specific usage application. For instance, image database investigations require static intensity images taken from a standard camera; while for surveillance applications, images should be captured by a videocamera. Other applications, such as access to security domains, may require the forgoing of the non-intrusive quality of FR, capturing facial images by means of an infra-red sensor or a 3D scanner.
Therefore, based on the data acquisition method, FR techniques can be classified into three categories: methods requiring sensory data such as infra-red imagery or 3D information, those dealing with video sequences, and techniques operating on intensity images.
However, video-based FR systems work by choosing some good frames and by applying the recognition techniques for intensity images to them, in order ti identify the individual.
Concerning FR methods for intensity images, they can be classified into two categories:

- **Feature-based**: fetaure-based methods aim to first process the input image in order to indentify, extract and measure distinctive facia features, such as mouth, eyes, skin color, nose etc., and other marks, for computing the geometric relationships among these facial points, transforming

the original input facial image to a vector of geometric features. The main advatage of featured-based techniques is the robustness to position variations in the input image, since the exytaction of feature points is performed before the matching phase. Other benefits are high speed matching and the compactness of representation of the face images. On the contrary, the main disadvantage of these kind of approach is that the implementation of these techniques requires to make arbitrary decisions about which features are important or not; this translates in difficulty of automatic feature detection.

– **Holistic**: holistic approaches aim to identify faces using global representations; thus, the descriptions are not based on local features of the face, but on th entire image. The main advantage of these methods is that image information is not destroyed, since they do not focus on only limited regions and point of interest. However, this factor also represents the main disadvantage of these approaches; in fact, the assumption that all the image pixels are equally important translates in higher computational complexity. Furthermore, a high degree of correlation between training and test images is required. Finally, another disadvantage of these kind of approaches is that they do not effectively perform under large variations in illumination, pose, and scale. Neural networks, which will be discussed in the next session, fit into this class of methods.

As mentioned before, the most recent FR solutions are based on Deep Learning, and the standard face recognition pipeline is the following[**3**]: face preprocessing (detection, alignement, etc.), choice of architecture and loss functions for transfer learning, face recognition for identification and verification.

3.3.1 Face Detection, Preprocessing and Alignement

The main purpose of Face Detection is to determine whether human faces appear in a given image, and where these faces are located with respect to the background. Once.the face is localized, it is required make the FR system more robust with some preprocessing. Face preprocessing exploit strong domain knowledge, such as facial landmark detection, pose estimation, rendering and data augmentation, in order to modify input faces to ease the learning of representations. Generally, the preprocessing phase corresponds to the facie alignement; after the detection, a face is always localized with a bounding box, but it may be not aligned. So, it is important to compensate for spatial changes, scale, in-plane rotations and traslations; a common approach to achieve this is 2D or 3D data augmentation when training the network[**1**].

3.3.2 Network Architectures, Disentanglement and Recognition

After the preprocessing of the training data, the standard FR pipeline proceeds defining a CNN architecture and a loss function for classification: the CNN is trained on a closed pool of elements and then it is used as a descriptor extractor on unknown faces.
There are several different network architectures that are optimized for FR, such as DeepFace, which exploits 3D frontalization: Furthermore, many attempts to modify CNNs architectures have been made so far, in order to better handling pose variations and to efficiently map back a profile face to a canonical pose, for simplifying the classification task.
It is also important to consider the fact that there is an alterntive way for data augmentation: disentanglement of signals in the learned feature space. In other words, while data augmentation injects confounding factors on the pixel space (such as pose, expression, illumination, etc.), disentanglement aims to decouple the factors in feature space; the factors present in the learning representations are linked only to the individual's identity.
Finally, after the training, the CNN is used to extract face descriptors, which will be classified (using Softmax Classifier, see Section 4.2), in order to complete the verification or identification task[**1**].

4 Processing pipeline

In this section our solution is presented. The workflow structure of our work is composed of dataset preparation, learning framework and face detection and recognition on a video.

4.1 Dataset preparation

The dataset was create starting by multiple images taken from 13 actors: Adam Sandler, Alyssa Milano, Bruce Willis, Denise Richards, George Clooney, Gwyneth Paltrow, Hugh Jackman, Jason Statham, Jennifer Love Hewitt, Lindsay Lohan, Mark Ruffalo, Robert Downery Jr and Will Smith. To build an enough complex model and to avoid overfitting we used all the photos from all the 13 actors. The size of the starting pictures were 256x256x3 and all images contains the complete face of the actors. The figure below shown an example of dataset images.

Figure 2: Example of images for each actor in alphabetic order taken from the original dataset

To prepare the dataset we follow a specific process, done by 5 steps:

- **Grayscale conversion**: we pass our images from 3 channels to one, from 256 x 256 x 3 to 256 x 256 x 1 , so we transform RGB space to grayscale using :

$$\text{RGBtoGray} : Y \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

- **Face detection**: to have a more linear dataset and to minimaze the discrepancy between dataset and face detected in videos we use the Viola e Jones face detector in all the images. We are going to explain Viola e Jones in next section.
- **Resize**: resizing images means change the width and the height. So, we pass from size 256 x 256 x 1 to 64 x 64 x 1 to have a more linear and light dataset, although quality could be lost.
- **Blur**: we applied a gaussian filter to blur the images, with the aim of obtain a more uniform dataset. Gaussian filtering is done by convolving each point in the picuture with a Gaussian kernel and then summing them all to produce the smoothed picture.
- **Label**: finally, we labeled all the resulting images, so we associate to each detected face the name of the correspondent actor.

In the figure below we shown the results of our dataset preparation use to train our convolutional neural network.

Figure 3: Example of preprocessed actor image

4.2 Learning framework

For the experiment we used a machine learning approach for actor face recognition. The classification is carried out with a Convolutional-Neural-Network (CNN)

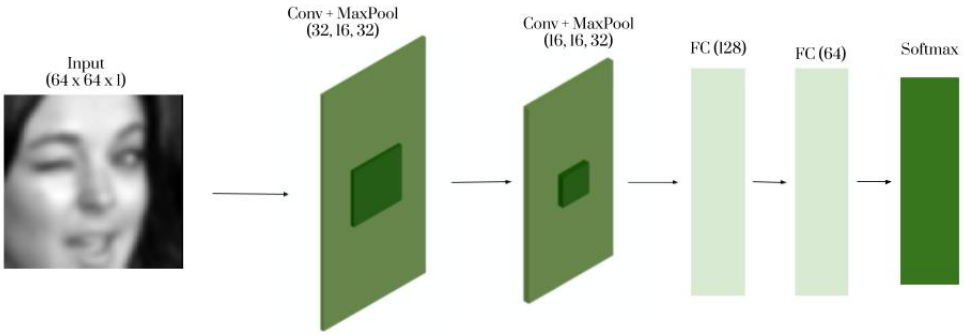

Figure 4: Architecture of the CNN for classification.

As shown in Figure 6, the classification is done using two blocks composed by a convolutional layer to detect the features, a max pooling layer to reduce the dimension and a dropout layer to regularize the data during the train to reduce overfitting. Then, it is achieved by feeding the output with two dense layers followed by two SoftMax. This network's structure is used to classify subjects in the input image. After the dataset creation, we obtain 3771 samples, which are splitted, after a data shuffle, in train set, validation set and test set. The first two contains respectively 3550 and 200 samples, and they are used for the CNN training. In the train phase of our Feed-Forward-Neural-Network we used an adam optimizer to update network weights iteratively, and we set 100 epochs and a batch size of 16, so, in this way the number of samples that will be propagated through the network are more, and the model can learn better. In the last part of the training, we used a categorical cross entropy, which is a loss function that is used in multi-class classification tasks. For this actor recognition we identified 13 different subjects listed above. The validation set was used to held back from training our model that is used to give an estimate of model skill while tuning models hyperparameters. This is different from the test set, because with test data we give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models. This latter is composed by 21 samples. The metrics used are the accuracy and the loss.

Formally, we can consider a training set composed of (I, γ) where I is the input image with the face and γ the actor which correspond to I. To learn the functions, we used the multiclass cross-entropy loss of the subject:

$$L_{actor} = -\sum \gamma_{ij} \log P(\gamma_j | I_i)$$

4.3 Face detection and recognition

The main goal of this work is to done face recognition adapted to the temporal dimension for videos face recognition. Our approach is based on the video frames processing, namely, for each video we had as input, we analyze each frame which compose the entire video. The main two steps of the analysis of the single frame are the face locations and then the face recognition. During the first phase we transform the frame from RGB space to grayscale and we used the Viola and Jones face detector. They proposed an object detector use Haar feature-based cascade classifiers, which is an algorithm based on an image preprocessing, *AdaBoost* and *Cascade Classifier*. *AdaBoost* technique consists in a classifier which reduce the number of relevent features and the *Cascade Classifier* consists in a set of classifier with increasing complexity.

After the face location, to develop the face recognition we crop the region of interest, we preprocessed it as in the section 4.1 and then we use the trained CNN to make the predictions of which actor is present in the image.

5 Results and Performance Evaluation

In this section the results are presented for both the CNN and the face detection and recognition performance.

5.1 CNN

The Convolutional Neural Networks used for the recognition reaches an accuracy of 99%, whic is a very good result. Obviously, we do not expect to achieve these results in practice, basically because in real-time processing of the video, the acquired images may be much more unprecise with respect to the images of the training dataset. In Figure 5 the traininig process and the accuracy increasement over the epochs of the models is shown.

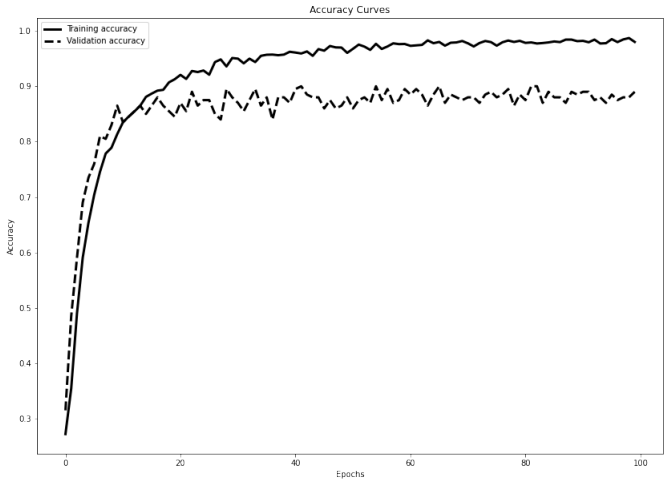

Figure 5: Results of the accuracy over the epochs.

As reported above, the accuracy of the model reaches good values. For more accurate evaluation of the model, in Figure 6 below we can observe the confusion matrix. As we can notice, some classes reach

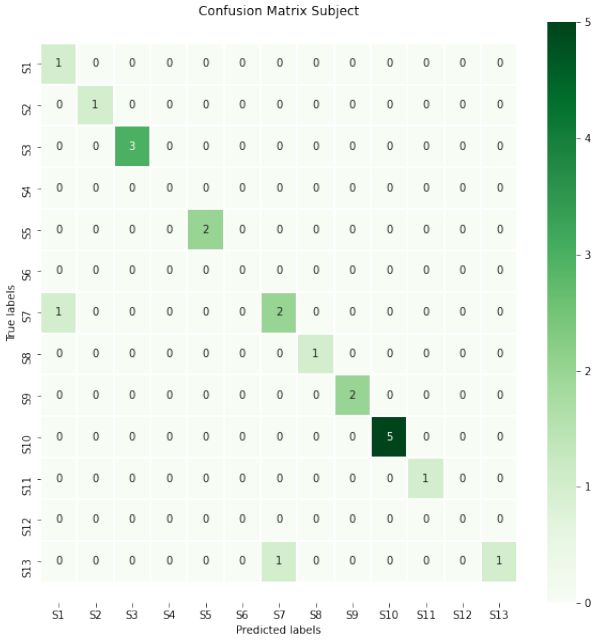

Figure 6: Confusion matrix of the CNN model.

lower accuracy than others: this could be due to the fact that some human faces present a few similar features, making the recognition task even more challenging.

5.2 Face Recognition

This section shows the results of our FR system.
Regarding the testing video, the duration is of 26s, and the number of frames is 643. The approach that we adopted is the following: we set the *accuracy threshold* to 90%, namely, the prediction is considered valid if and only if the accuracy is at least of 90%. Furthermore, we decided not to focus the recognition task on each of the 643 frames of the video, but to focus on small video clips lasting 1s (namely, 26 video clips), thus by grouping 25 frames per clip. We consider this evaluation to be reasonable, since a generic movie film, for example, last at least a few seconds, and the objective is to recognize the individual on that specific scene, rather on the specific frame. In other words, considering the 25 frames that form a clip, we compute the the most common name found.

An example of the output predictions of the Convolutional Neural Network are below reported (Figure 7). As we can see, we reported three different cases: correct recognition, false positive, and unknown face.
Regarding the unknown face case, we found an accuracy of 74%, which is rightly lower than the fixed accuracy threshold (90%): in fact, the individual in that frame is not present in the training dataset, and thus the behaviour of the CNN is correct. An individual which is not included in the training dataset can not be recognized. However, the face detection task is correctly computed, since the face is well localzed on the frame.
The false positive case presents an accuracy of 91%, but the recongition is not correct, since the CNN recognize the individual as *Will Smith* instead of *Bruce Willis*. The face localization, however, is corrected also in this case.

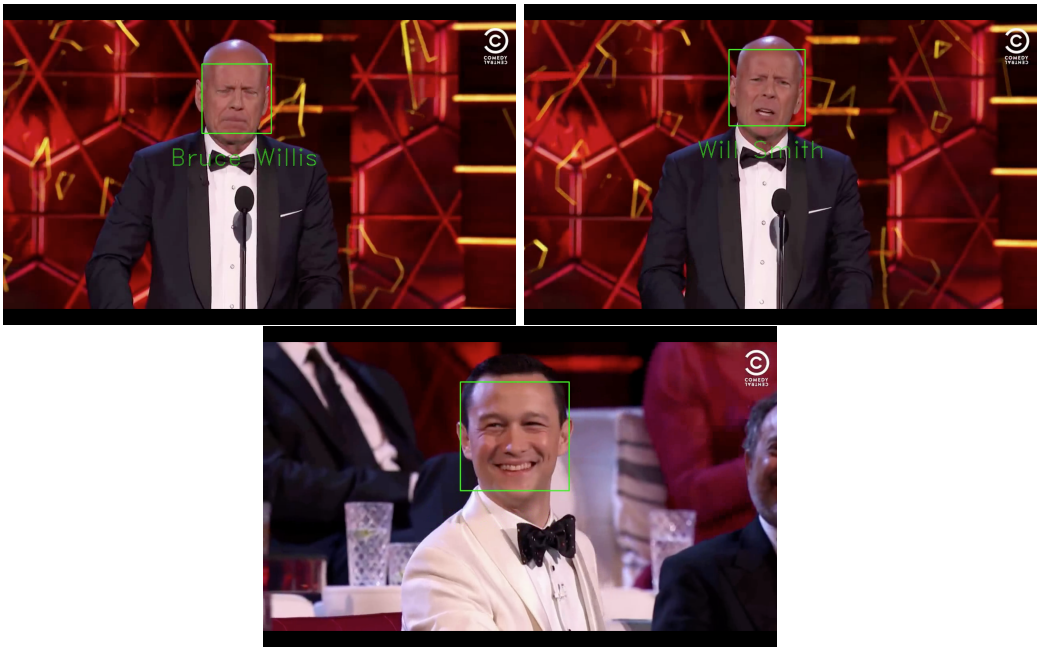

Figure 7: Top left: corect recognition (accuracy: 97%); Top right: false positive (accuracy: 91%); Bottom: unknown face (accuracy: 74%)

Finally, the correct recognition case reaches an accuracy of 97%, far above the fixed threshold of 90%, and both the face detection and face recognition task are corrected (*Bruce Willis*).
As a summary, with respect to the approach descried above, we obtained a correct face recogntion on 24 of the 26 clips that make up the video, which results in a success rate of 92%.

6 Conclusion

We have seen how Face Detection and Face Recognition task are important nowadays, since they have many applications an various domanis, such as surveillance and security systems, criminal justice systems, image database verifications, identity verification, etc.
However, FR is still considered a difficult problem, as faces can vary a great deal in their lighting conditions, orientation and facial expression. We also reported that Convolutional Neural Networks are optimized and widely used in computer vision, specially for FD and FR tasks. This is the reason why most recent FR solutions are based on AI.
The goal of this work was to provide a CNN-based FR systems for real time recognition on a video. The CNN we used exploits Viola and Joness face detector and Haar feature-based cascade classifier. As reported above, we obtained a success rate of 92%, which can be considere a good result. The performance in terms of accuracy could be further improved, but, since our goal was to implement a system able to perform FR in real time on a video, it was necessary to find a trade-off between accuracy and processing speed, to ensure that the network was not too slow. Furthermore, in order to improve the accuracy, it would be advisable to deal with images without a background, also making less computational demanding the preprocessing and detection phases, facilitating real-time processing, as well as having a larger dataset for training the network.
Beyond that, considering the obtined results, we can conclude that the performance of our FR system are satisfactory.

References

[1] Deep face recognition: A survey. Masi, Iacopo and Wu, Yue and Hassner, Tal and Natarajan, Prem. 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), 2018, IEEE.

[2] A survey of face recognition techniques. Jafri, Rabia and Arabnia, Hamid R. Journal of information processing systems, 2009, Korea Information Processing Society.

[3] Face recognition based on convolutional neural network. Musab and Uçar, Ayşegül and Yildirim. 2017 International Conference on Modern Electrical and Energy Systems (MEES), 2017, IEEE.

[4] Deep face recognition. Parkhi, Omkar M and Vedaldi, Andrea and Zisserman, Andrew, 2015, British Machine Vision Association.

[5] Face recognition: A convolutional neural-network approach. Lawrence, Steve and Giles, C Lee and Tsoi, Ah Chung and Back, Andrew D. IEEE transactions on neural networks, 1997, IEEE.

[6] How convolutional neural network see the world-A survey of convolutional neural network visualization methods. Qin, Zhuwei and Yu, Fuxun and Liu, Chenchen and Chen, Xiang. ArXiv preprint arXiv:1804.11191, 2018