FaceRec: An Interactive Framework for Face Recognition in Video Archives

PASQUALE LISENA, EURECOM, France
JORMA LAAKSONEN, Aalto University, Finland
RAPHAËL TRONCY, EURECOM, France

Annotating the visual presence of a known person in a video is a hard and costly task, in particular when applied to large video corpora. The web is a massive source of visual information that can be exploited for detecting celebrities. In this work, we introduce FaceRec, an AI-based system for automatically detecting faces of known but also unknown people in a video. The system relies on a combination of state-of-the-art algorithms (MTCNN and FaceNet), applied on images crawled from web search engines. A tracking system links consecutive detection in order to adjust and correct the label predictions using a confidence-based voting mechanism. Furthermore, we add a clustering algorithm for the unlabelled faces, thus increasing the number of people that can be recognized. We evaluate our system that obtained high precision on datasets of both historical and recent videos. We release the complete framework as open-source at https://git.io/facerec.

CCS Concepts: • **Computing methodologies** → **Object recognition**; • **Information systems** → *Web mining*; • **Computer systems organization** → *Neural networks.*

Additional Key Words and Phrases: Face recognition, neural networks, semantic metadata, video archives

ACM Reference Format:
Pasquale Lisena, Jorma Laaksonen, and Raphaël Troncy. 2021. FaceRec: An Interactive Framework for Face Recognition in Video Archives. In *Proceedings of 2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Fig. 1. Charles de Gaulle and Dwight D. Eisenhower together in 1962 *(picture from Archives Nationales).*

1 INTRODUCTION

Identifying people appearing in videos is undoubtedly an important cue for automatically understanding video content. Knowing who appears in a video, when and where, can also lead to learning interesting patterns of relationships among characters, with interesting applications in historical research and media discovery. Such person-related annotations enable to generate more accurate segmentation and more compelling video descriptions, facilitating multimedia search and re-use of video content. Media archives contain numerous examples of celebrities appearing in the same news segment (Figure 1). However, the annotations produced manually by archivists do not always identify with precision those individuals in the videos. The presence of digital annotations is particularly crucial for large corpora, whose metadata are the only efficient way to identify relevant elements [22]. At the same time, relying on human annotations is not a scalable solution when handling large volumes of video resources.

The web offers an important amount of pictures of people and in particular of celebrities, easily findable using their full name as search terms in a general purpose search engine such as Google. While it has been considered a relevant information source in other communities – such as computational linguistics [13] and recommender system [17] – the web is still only scarcely exploited in image analysis and in face recognition in particular.

In this work, we aim to leverage pictures of celebrities crawled from the web for identifying faces of people in video archives. In doing so, we develop FaceRec, an interactive framework for face recognition in video corpora that relies on state-of-the-art algorithms. The system is based on a combination of MTCNN (face detection) and FaceNet (face embeddings), whose vector representations of faces are used to feed a classifier, which is then used to recognise faces at the frame level. A tracking system is included in order to increase the robustness of the library towards recognition errors in individual frames for getting more consistent person identifications.

The rest of this paper is organised as follows. After reporting some relevant work in Section 2, we describe our approach in Section 3. A quantitative evaluation is carried out on two different datasets in Section 4. We introduce the FaceRec API and a web application for visualizing the results in Section 5. Finally, some results and possible future work are outlined in Section 6.

2 RELATED WORK

During the last decade, there has been substantial progress in the methods for automatic recognition of individuals. The recognition process generally consists of two steps. First, faces need to be detected in a video, i.e. which region of the frame may contain a face. Second, those faces should be recognised, i.e. to whom a face belongs.

The Viola-Jones algorithm [21] for face detection and the Local Binary Pattern (LBP) features [1] for the clustering and recognition of faces were the most famous methods until the advent of deep learning and convolutional neural networks (CNN). Nowadays, two main approaches are used for detecting faces in video and both use CNNs. One implementation is available in the Dlib library [14] and provides good performance for frontal images, but it requires an additional alignment step before the face recognition step can be performed. The recent Multi-task Cascaded Convolutional Networks (MTCNN) [24] approach provides even better performance using an image pyramid approach and using face landmarks detection for re-aligning the detected faces to the frontal orientation.

After locating the position and orientation of the faces in the video frames, the face recognition process can be performed. There are several strategies available in the literature for face recognition. Currently, the most practical approach is to perform face comparison using a transformation space in which similar faces are mapped close together, and to use this representation to identify individuals. Such embeddings, computed on large collections of faces have been made available to the research community, such as the popular FaceNet [19].

In [23], MTCNN and FaceNet are used in combination and tested with eight public face datasets, reaching a recognition accuracy close to 100% and surpassing other methods. These results have been confirmed in several surveys [8, 20] and in recent works [2]. In addition, MTCNN has been recognised to be very fast while having good performance [16].

Given the almost perfect performance of the MTCNN + FaceNet face recognition setups, our work focuses on setting up a complete system built upon these technologies. In this perspective, our contribution does not consist of a new state-of-the-art performance in face recognition, but of the combination and application of available techniques in combination with images crawled on the web.

3 METHOD

This section describes the FaceRec pipeline, detailing the training and the recognition tasks, including the additional strategy for recognising unknown faces in videos.

3.1 Training the system

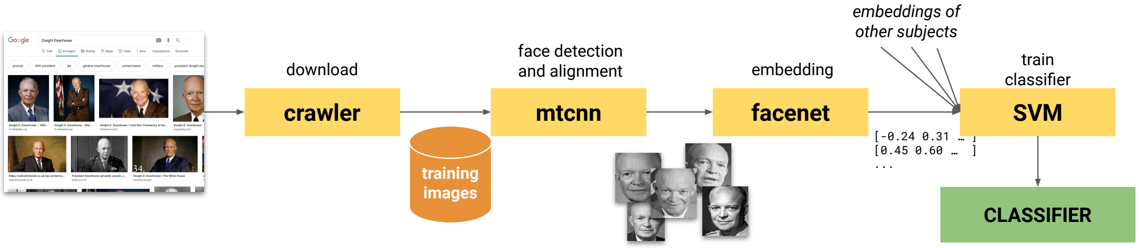


Fig. 2. FaceRec training pipeline

During training, our system retrieves images from the web for realising a face classifier (Figure 2). The first module is a **crawler**[1] which, given a person's name, automatically downloads a set of k photos using Google's image search engine. In our experiments, we have typically used $k = 50$. After converting them to greyscale, we apply to each image the **MTCNN algorithm** [24] for face detection[2]. MTCNN returns in output the bounding box of the face in the frame and the position of relevant landmarks, namely the position of eyes, nose and mouth limits. The recognised faces are cropped, resized and aligned in order to have in output a set of face images of width $w = 256$ and height $h = 256$, in which the eyes are horizontally aligned and centered. In particular, the alignment consists of a rotation of the image. Chosen the desired positions for the left (x_l, y_l) and right eye (x_r, y_r)[3] and given their original positions (a_l, b_l) and (a_r, b_r), the image is rotated by an angle α on the centre c with scale factor s, computed in the following way:

$$dX = a_r - a_l \qquad dY = b_r - b_l$$

[1] We use the *icrawler* open-source library: https://github.com/hellock/icrawler/

[2] We use the implementation at https://github.com/ipazc/mtcnn

[3] We use $x_l = 0.35w$, $x_r = (1 - x_l)$, and $y_l = y_r = 0.35h$.

$$\alpha = \arctan \frac{dY}{dX} - 180^\circ$$
$$c = (\frac{x_l + x_r}{2}, \frac{y_l + y_r}{2})$$
$$s = \frac{(x_r - x_l) \cdot w}{\sqrt{dX^2 + dY^2}}$$

Not all resulting cropped images are suitable for training a classifier. They may contain faces of other individuals, if they have been extracted from a group picture or if the original picture was not really depicting the searched person. Other cases which may have a negative impact on the system are side faces, low resolution images, drawings and sculptures. In order to exclude those images, we relied on two complementary approaches, which we used in combination:

- using face embeddings to automatically remove the outliers. This is realised by removing the face with the highest cosine distance from the average embedding vector, until the standard deviation of all differences is under an empirically chosen threshold (0.1);
- allowing the user to further improve the automatic selection by allowing the exclusion of faces via the user interface (Section 5).

On the remaining pictures, a pretrained **FaceNet** [19] model with Inception ResNet v1 architecture trained on the VGGFace2 dataset [6] is applied for extracting visual features or embeddings of the faces. The embedding vectors feed ***n*** **parallel binary SVM**[4] classifiers, where *n* is the number of distinct individuals to recognise. Each classifier is trained in a one-against-all approach [12], in which the facial images of the selected individual are used as positive samples, while all the others are considered negative samples. In this way, each classifier provides in output a confidence value, which is independent of the outputs of all other classifiers. This will allow to set – in the recognition phase – a confidence threshold for the candidate identities which does not depend on *n*, making the system scalable[5].

3.2 Recognising faces in video

The face recognition pipeline is composed of:

- operations that are performed at the frame level and are shown in Figure 3. In order to speed up the computation, it is possible to set a sampling period T. For our experiments, we set $T = 25$, in order to process one frame per second;
- operations of synthesis on the results, which take into account the tracking information across frames for providing more solid results.

In each frame, **MTCNN** detects the presence of faces, to which is applied the same cropping and alignment presented in Section 3.1. Their **FaceNet** embeddings are computed and the **classifier** selects the best match among the known faces, assigning a confidence score in the interval $[0, 1]$.

[4] SVM obtained better performance than other tested classifier, namely Random Forest, Logistic Regression and the k-Nearest Neighbours.

[5] We also performed experiments on this system using a multi-class classifier with *n* class, instead of the *n* binary classified. While the results revealed similar precision scores, the recall for the multi-class solution was considerably worse, 22 percentage points lower than the system with binary classifiers.

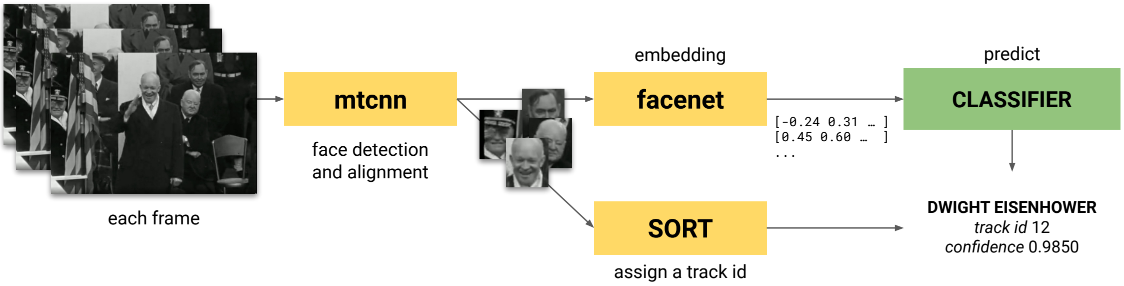


Fig. 3. FaceRec prediction pipeline

At the same time, the detected faces are processed by *Simple Online and Realtime Tracking (**SORT**)*, an object tracking algorithm which can track multiple objects (or faces) in realtime[6] [5]. The algorithm uses the MTCNN bounding box detection and tracks the bounding boxes across frames, assigning a tracking id to each face.

After having processed the entire video, we obtain a set of detected faces, each of them with a predicted label, confidence score and tracking id, as well as space and time coordinates. This information is then processed at the level of single tracking collection, integrating the data of the different recognitions having the same tracking id. For a given tracking – including a certain number of samples – we compute the mode[7] of all predictions, as well as the weighted mode with respect to the confidence scores. A unique predicted label p is chosen including among all the possible predictions if it satisfies all the following conditions:

- p is both the mode and the weighted mode;
- the ratio of samples with prediction p over the total samples is greater than the threshold h;
- the ratio of samples with prediction p over the total samples, weighting all occurrence with the confidence score, is greater than the threshold h_w.

We empirically found $h = 0.6$ and $h_w = 0.4$ as the best values for the thresholds. It is possible that a tracking does not produce a label fulfilling all the conditions. In that case, the prediction is considered uncertain and the tracking is excluded from the results. We assign to the tracking a unique confidence score from the arithmetic mean of the scores of the sample with prediction p. We intentionally exclude the minority of wrong predictions in this computation: in this way, wrong predictions – caused by e.g. temporary occlusion or turn of the head by side – do not penalise the overall scores. The final results are then filtered again by overall confidence using a threshold t, whose impact is discussed in Section 4.

3.3 Building models for unknown faces

So far, the described system is trained for recognising the faces of known people. During the processing of a video, several detected faces may not be matched with any of the individuals in the training set. However, these people may still be relevant to be tracked and inserted in the list of people to search. Therefore, in addition to the pipeline based on images crawled from the web, a face clustering algorithm is active in the background with the objective of detecting non-celebrities or more simply, any persons not present in the training set. At runtime, all FaceNet features extracted from faces in the video frames are collected. Once the video has been fully processed, these features are aggregated

[6] We used the implementation provided at https://github.com/Linzaer/Face-Track-Detect-Extract with some minor modification
[7] The mode is "the number or value that appears most often in a particular set" (*Cambridge Dictionary*)

through hierarchical clustering[8] based on a distance threshold, empirically set to 14. The clustering produces a variable number m of clusters, with all items assigned to one of them. The clusters are then filtered in order to exclude:

- those for which we can already assign a label from our training set;
- those having a distance — computed as the average distance of the elements from the centroid — larger than a second, more strict threshold, for which we have used the value 1.3;
- those having instances of side faces in the centre of the cluster. In particular, we observed that in those cases, the resulting cluster produces unreliable results and groups profile views of different people.

With MTCNN, we obtain the position of the following landmarks: left eye (a_l, b_l), right eye (a_r, b_r), left mouth corner (m_l, n_l), right mouth corner (m_r, n_r). We compute the ratio r_{dist} between the distance between mouth and eyes and the distance between the two eyes:

$$dX = a_r - a_l \qquad dG = m_l - a_l$$

$$dY = b_r - b_l \qquad dH = n_l - b_l$$

$$dist_{wide} = \sqrt{dX^2 + dY^2} \qquad dist_{high} = \sqrt{dG^2 + dH^2}$$

$$r_{dist} = \frac{dist_{high}}{dist_{wide}}$$

This value is inversely proportional to the eyes' distance on the image, increasing when the eyes are closer, e.g. in face rotation to a side. We identified as side faces the cases in which $r_{dist} > 0.6$. Finally, only the 5 faces closest to each centroid are kept, in order to exclude potential outliers.

The system returns in output the remaining clusters, which are temporary assigned to a label of type *Unknown <i>*, where *i* is an in-video incremental identifier – e.g. *Unknown 0*, *Unknown 1*, etc. The clusters can be labelled with human effort: in this case, the relevant frames are used as training images and the person is included in the training set. This strategy is particularly useful in cases when the crawler module cannot be used to obtain representative samples of the individuals appearing in the videos.

4 EVALUATION

In this section, we evaluate the FaceRec system measuring the precision and recall on two different ground-truth datasets: one of historical videos and one composed of more recent video footage.

4.1 Creation of a ground truth

In the absence of a large and rigorous ground truth dataset of faces in video, we developed two evaluation datasets of annotated video fragments from two different specialised corpora.

ANTRACT dataset. *Les Actualités françaises*[9] are a series of news programmes broadcasted in France from 1945 to 1969, currently stored and preserved by the *Institute national de l'audiovisuel (INA)*[10]. The videos are in black-and-white, with a resolution of 512×384 pixels. Metadata are collected through INA's *Okapi* platform [4, 7], which exposes a SPARQL endpoint.

[8] We used the implementation available in SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html
[9] https://www.ina.fr/emissions/les-actualites-francaises/
[10] The corpus can be downloaded from https://dataset.ina.fr/

A list of 13 historically well-known people has been provided by domain experts. From the metadata, we have obtained the reference to the segments in which these people appear and the subdivision of these segments in shots[11]. This search produced 15,628 shots belonging to 1,222 segments from 702 media. In order to reduce the number of shots and to check manually the presence of the person in the selected segments, we performed face recognition on the central frame of each shot. The final set has been realised with an iteration of automatic sampling and manual correction, adding also some shots not involving any of the specified people. At the end, it includes 198 video shots (belonging to 129 distinct media), among which 159 segments (80%) featured one or more of the 13 known people and 39 segments (20%) did not include any of the specified people.

MeMAD dataset. This dataset has been developed from a collection of news programmes broadcasted on the French TV channel *France 2* in May 2014. These videos – in colour, 455×256 pixels – are part of the MeMAD video corpus[12], with metadata available from the MeMAD's Knowledge Graph[13]. We followed the same procedure than above with the following differences. In this case, the list of people to search is composed of the six most present ones in the MeMAD Knowledge Graph's video segments. Without the information about the subdivision in shots, for each segment of duration d, we performed face recognition on the frames at positions $d/4$, $d/2$ and $3d/4$, keeping only the segments with at least one found face. We also made an automatic sampling and a manual correction as for the ANTRACT dataset. The final set includes 100 video segments, among which 57 segments (57%) featured one of the six known people and 43 segments (43%) did not include any of the specified people.

Table 1 summarises the main differences between the two datasets.

	ANTRACT	**MeMAD**
type	historical images	TV news
years	1945-1969	2014
resolution	512×384	455×256
colorspace	b/w	colour
shots division	yes	no
list of celebrities to search	13 (chosen by domain experts)	6 (most present in KG)
represented fragment and length	shot 3 seconds in avg.	segment up to 2 minutes
records	216	100
distinct fragments	198	100
distinct media (videos)	129	30
fragments without known faces	39	43

Table 1. Description of the ANTRACT and MeMAD datasets

[11] In the following, we define *media* as the entire video resource (e.g. an MPEG-4 file), *segment* a temporal fragment of variable length (possibly composed of different shots), and *shot*, a not interrupted recording of the video-camera. See also the definitions of MediaResource, Part and Shot in the EBU Core ontology (https://www.ebu.ch/metadata/ontologies/ebucore/)

[12] https://memad.eu/

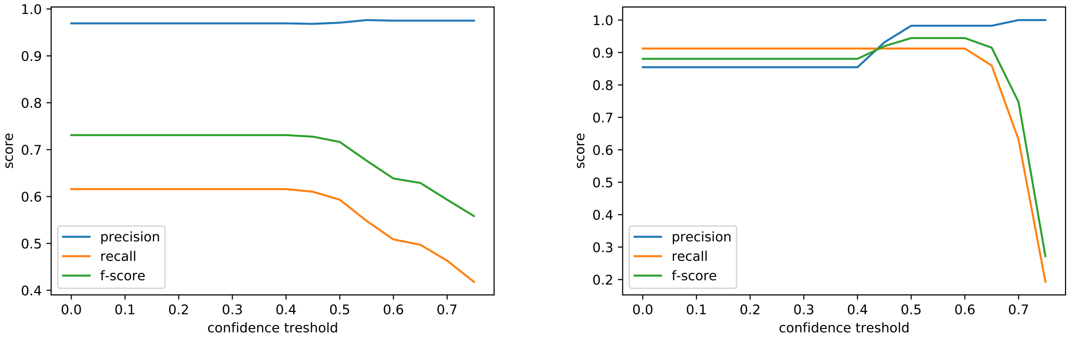
[13] https://data.memad.eu/

Person	**P**	**R**	**F**	**S**
Ahmed Ben Bella	1.00	0.46	0.63	13
François Mitterrand	1.00	0.92	0.96	13
Pierre Mendès France	1.00	0.61	0.76	13
Guy Mollet	0.92	0.92	0.92	13
Georges Bidault	0.83	0.71	0.76	14
Charles De Gaulle	1.00	0.57	0.73	19
Nikita Khrushchev	1.00	**0.38**	0.55	13
Vincent Auriol	1.00	0.46	0.63	13
Konrad Adenauer	1.00	0.53	0.70	13
Dwight Eisenhower	0.85	0.46	0.60	13
Elisabeth II	1.00	0.71	0.83	14
Vyacheslav Molotov	1.00	**0.23**	0.37	13
Georges Pompidou	1.00	0.69	0.81	13
– unknown –	0.35	0.97	0.52	39
average (unknown apart)	0.97	0.59	0.71	216

Table 2. ANTRACT dataset: precision, recall, F-score and support for each class and aggregate results. The support column corresponds to the number of shots in which the person appears.

4.2 Quantitative analysis

For each dataset, a face recognition model has been trained to recognise the individuals from the corresponding list of celebrities. The model has then been applied to the video fragments of the ANTRACT and MeMAD datasets. We varied the confidence threshold t under which we considered the face not matched as shown in Figure 4, and found the optimal values with respect to the F-score – $t = 0.5$ for ANTRACT and $t = 0.6$ for MeMAD. The overall results – with the details of each person class – are reported in Table 2 and Table 3.



(a)

(b)

Fig. 4. Precision, recall and F-score of FaceRec on different confidence thresholds for the ANTRACT (4a) and the MeMAD dataset (4b).

The system obtained high precision in both datasets, with over 97% of correct predictions. If the recall on the MeMAD dataset is likewise good (0.91), it is significantly lower for the ANTRACT dataset (0.59). This is largely due to the differences between the two datasets, which involve not only the image quality, but also different shooting approaches.

Person	P	R	F	S
Le Saint, Sophie	0.90	0.90	0.90	10
Delahousse, Laurent	1.00	1.00	1.00	7
Lucet, Elise	1.00	0.90	0.94	10
Gastrin, Sophie	1.00	0.90	0.94	10
Rincquesen, Nathanaël de	1.00	0.80	0.88	10
Drucker, Marie	1.00	1.00	1.00	10
– unknown —	0.89	0.97	0.93	43
average (unknown apart)	0.98	0.91	0.94	100

Table 3. MeMAD dataset: precision, recall, F-score and support for each class and aggregate results. The support column corresponds to the number of segments in which the person appears.

If modern news are more used to close-up shots, taken on screen for multiple seconds, in historical videos, it is easier to find group pictures (in which occlusion is more probable), quick movements of the camera, and tight editing, leaving to our approach less samples for recognition. It is also relevant to notice that the lowest recall values belong to the only two USSR politicians Khrouchtchev and Molotov: most often, they appear in group images or in very short close-up images, raising questions for historical research.

4.3 Qualitative analysis

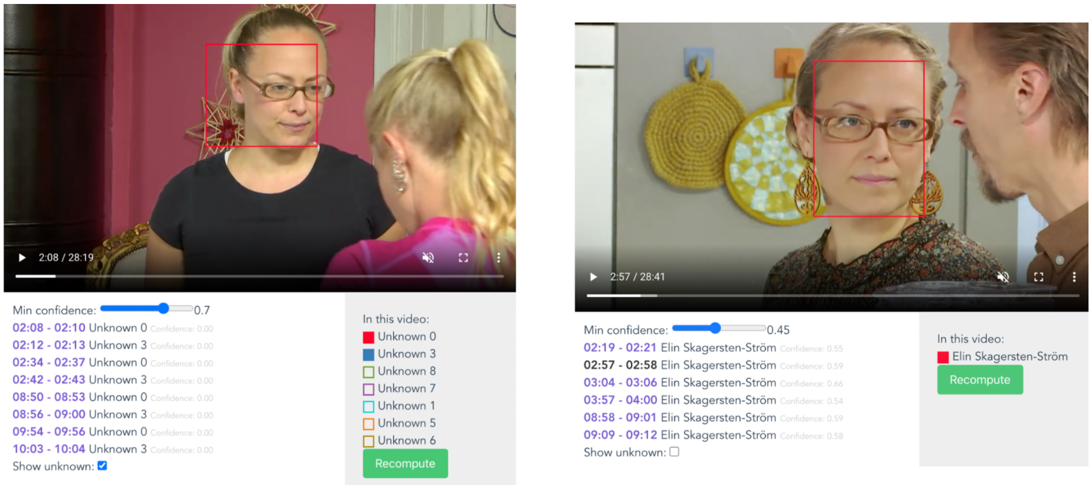
We made a qualitative analysis of the results. When inspecting the obtained recognition, we make the following observations:

- The system generally fails to detect people when they are in the background and their faces are therefore relatively small. This is particularly true for the ANTRACT dataset, in which the image quality of films is poorer.
- The cases in which one known person is confused with another known person are quite uncommon. Most errors occur when an unknown face is recognised as one of the known people.
- The recognition is negatively affected by occlusions of the face, such as unexpected glasses or other kind of objects.
- The used embeddings are not suitable to represent side faces, whose predictions are not reliable.

4.4 Unknown cluster detection evaluation

Together with the previous evaluation, we clustered the unknown faces found in the videos, as explained in Section 3.3. We then manually evaluated the resulting clusters on five randomly-selected videos for each dataset. We make the following observations:

- If more than one face is assigned to the same *Unknown <i>*, those faces actually belong to the same person. In other words, the erroneous presence of different individuals under the same label is never verified. This is due to the strict threshold chosen for intra-cluster distance.
- On the other side, not all the occurrences of that face are labelled, given that only the top five faces are kept. This may not be relevant if we are searching for new faces to add to the training set and we anyway intend to perform a further iteration afterwards.
- In one case, a single person was included in two distinct clusters, which may be reconciled by assigning the same label.



(a)

Min confidence: 0.7

02:08 - 02:10 Unknown 0 Confidence: 0.00
02:12 - 02:13 Unknown 3 Confidence: 0.00
02:34 - 02:37 Unknown 0 Confidence: 0.00
02:42 - 02:43 Unknown 3 Confidence: 0.00
08:50 - 08:53 Unknown 0 Confidence: 0.00
08:56 - 09:00 Unknown 3 Confidence: 0.00
09:54 - 09:56 Unknown 0 Confidence: 0.00
10:03 - 10:04 Unknown 3 Confidence: 0.00

Show unknown: ☑

In this video:

- 🟥 Unknown 0
- 🟦 Unknown 3
- □ Unknown 8
- □ Unknown 7
- □ Unknown 1
- □ Unknown 5
- □ Unknown 6

(b)

Min confidence: 0.45

02:19 - 02:21 Elin Skagersten-Ström Confidence: 0.55
02:57 - 02:58 Elin Skagersten-Ström Confidence: 0.59
03:04 - 03:06 Elin Skagersten-Ström Confidence: 0.66
03:57 - 04:00 Elin Skagersten-Ström Confidence: 0.54
08:58 - 09:01 Elin Skagersten-Ström Confidence: 0.59
09:09 - 09:12 Elin Skagersten-Ström Confidence: 0.58

Show unknown: □

In this video:

- 🟥 Elin Skagersten-Ström

Fig. 5. The clustering output found a set of unknown persons in the video (5a). Using the frames of *Unknown 0*, we are able to build the model for Elin Skagersten-Ström and recognise her in other videos. (5b).

- Less clusters were found in the ANTRACT videos than in the MeMAD videos – three out of five videos with no clusters. This is again explained by the lower video quality, less frequent close-up shots and faster scene changes.

For understanding the benefit that results from the face clustering, we include in Figure 5 an example use case. In Figure 5a, the clustering algorithm identified a set of unknown people, among which *Unknown 0* happens to be Elin Skagersten-Ström, who was not part of our training set. For each segment in which *Unknown 0* appeared, we extracted the four frames closer to the middle of the segment and included them as images in the training set. By re-training the classifier with this new data, it was possible to correctly detect Elin Skagersten-Ström in other videos, as seen in Figure 5b. This approach can be applied to any individuals, including those for whom one cannot find enough face images on the Web for training a classifier.

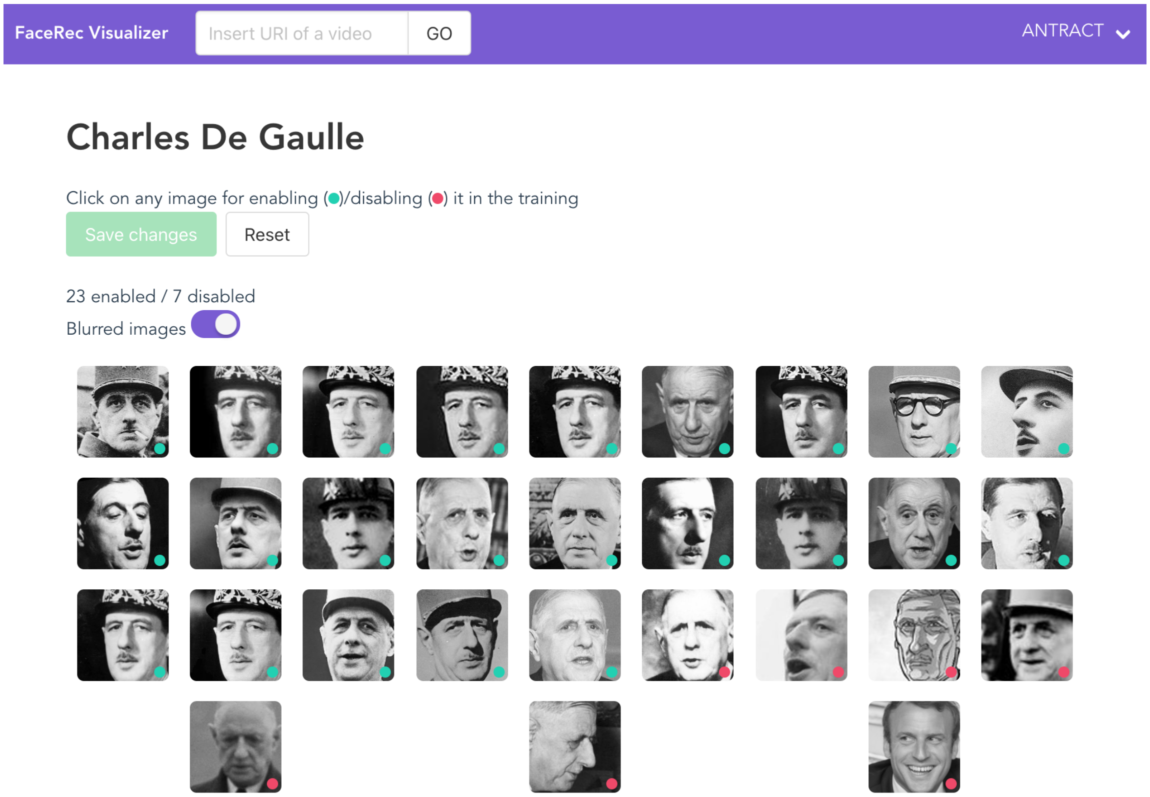
5 A WEB API AND A USER INTERFACE

In order to make FaceRec publicly usable and testable, we wrapped its Python implementation within a Flask server and made it available as a **Web API** at http://facerec.eurecom.fr/. The API has been realised in compatibility with the OpenAPI specification[14] and documented with the Swagger framework[15]. The main available methods are:

- `/crawler?q=NAME` for searching on the Web images of a specific person;
- `/train` for training the classifier;
- `/track?video=VIDEO_URI` for processing a video.

The results can be obtained in one of two output structures: a custom JSON format and a serialisation format in RDF using the Turtle syntax, relying on the EBU core[16] and Web Annotation ontologies[17]. The Media Fragment URI[18] syntax

[14] https://www.openapis.org/
[15] https://swagger.io/
[16] https://www.ebu.ch/metadata/ontologies/ebucore/
[17] https://www.w3.org/ns/oa.ttl
[18] https://www.w3.org/TR/media-frags/



Charles De Gaulle

Click on any image for enabling (●)/disabling (●) it in the training

23 enabled / 7 disabled

Fig. 6. Person page in FaceRec Visualizer: drawings, side faces, other depicted individuals and low-quality images are discarded (see the last 7 pictures marked with the red dot).

is also used for encoding the time and spatial information, with *npt* in seconds for identifying temporal fragments and *xywh* for identifying the bounding box rectangle encompassing the face in the frame. A light cache system that enables to serve pre-computed results is also provided.

In addition, a **web application** for interacting with the system has been deployed at http://facerec.eurecom.fr/visualizer. The application has a homepage in which the list of celebrities in the training set is shown. For each person, it is possible to see the crawled images and decide which of them have to be included or excluded during the training phase (Figure 6). In addition, it is possible to add a new celebrity for triggering the automatic crawling and re-train the classifier once modifications have been completed.

Finally, it is possible to run the face recognition on a video, inserting its URI in the appropriate textbox. Partial results are shown to the user as soon as they are computed, so that it is not required to wait for the analysis of the entire video for seeing the first recognised faces. The detected persons are shown on a list, whose elements can be clicked for seeking the video until the relevant moment. The faces are identified in the video using squared boxes (Figure 5). A slider enables to vary the confidence threshold, allowing to interactively see the result depending on the value chosen. Some metadata are displayed for videos coming from the MeMAD and ANTRACT corpora.

6 CONCLUSIONS AND FUTURE WORK

With FaceRec, we managed to successfully exploit images on the web for training a face recognition pipeline which combines some of the best-performing state-of-the-art algorithms. The system has shown good performance, with an almost perfect precision. A clustering system has been integrated in FaceRec with unknown person detection, the results of which can be added to the training set. A web application allows to easily interact with the system and see the results on videos. The implementation is publicly available at https://git.io/facerec under an open source licence.

This system has been successfully applied in video summarisation, in a strategy combining face recognition, automatically-generated visual captions and textual analysis [11]. The proposed approach ranked first in the *TRECVID Video Summarization Task* (VSUM) in 2020.

In future work, we plan to improve the performances of our approach and in particular its recall. While the recognition of side faces largely impacts the final results, a proper strategy for handling them is required, also relying on relevant approaches from the literature [9, 18]. With quick changes of scenes, a face can be seen in the shot for only a very short time, not giving enough frames to the system for working properly. We may propose a different local sampling period $T_{local} < T$ to be used when a face is recognised in order to collect more frames close to the detection. In addition, we believe that the system would benefit from prior shot boundary detection in videos, in order to process shots separately.

A more solid confidence score can be returned including contextual and external information, such as metadata (the dates of the video and the birth-death of the searched person), the presence of other persons in the scene [15], and textual descriptions, captions and audio in multi-modal approaches [3, 10].

The presented work has several potential applications, from annotation and cataloguing to automatic captioning, with a possible inclusion in second-screen TV systems. Moreover, it can support future research in computer vision or in other fields – e.g. history studies. An interesting application is the study of age progression in face recognition [25].

ACKNOWLEDGMENTS

This work has been partially supported by the French National Research Agency (ANR) within the ANTRACT project (grant number ANR-17-CE38-0010) and by the European Union's Horizon 2020 research and innovation program within the MeMAD project (grant agreement No. 780069).

REFERENCES

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. 2006. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 28, 12 (2006), 2037–2041.

[2] Adamu Ali-Gombe, Eyad Elyan, and Johan Zwiegelaar. 2020. Towards a Reliable Face Recognition System. In *21^{st} Engineering Applications of Neural Networks Conference (EANN)*, Lazaros Iliadis, Plamen Parvanov Angelov, Chrisina Jayne, and Elias Pimenidis (Eds.). Springer International Publishing, Cham, 304–316.

[3] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16, 6 (2010), 345–379.

[4] Abdelkrim Beloued, Peter Stockinger, and Steffen Lalande. 2017. *Studio Campus AAR: A Semantic Platform for Analyzing and Publishing Audiovisual Corpuses.* John Wiley & Sons, Ltd, Hoboken, NJ, USA, Chapter 4, 85–133.

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple Online and Realtime Tracking. In *IEEE International Conference on Image Processing (ICIP).* IEEE Computer Society, Phoenix, AZ, USA, 3464–3468.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *13^{th} IEEE International Conference on Automatic Face & Gesture Recognition (FG).* IEEE Computer Society, Xi'an, China, 67–74.

[7] Jean Carrive, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Antoine Laurent, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Benedicte Pinchemin, Geraldine Poels, and Raphael Troncy. 2021. Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quarterly, Special Issue on AudioVisual Data in DH* 15, 1 (2021).

[8] Guodong Guo and Na Zhang. 2019. A survey on deep learning based face recognition. *Computer Vision and Image Understanding* 189 (2019).

[9] Haroon Haider and Malik Khiyal. 2017. Side-View Face Detection using Automatic Landmarks. *Journal of Multidisciplinary Engineering Science Studies* 3 (2017), 1729–1736.

[10] Anand Handa, Rashi Agarwal, and Narendra Kohli. 2016. A survey of face recognition techniques and comparative study of various bi-modal and multi-modal techniques. In *11th International Conference on Industrial and Information Systems (ICIIS)*. 274–279.

[11] Ismail Harrando, Alison Reboud, Pasquale Lisena, Raphaël Troncy, Jorma Laaksonen, Anja Virkkunen, and Mikko Kurimo. 2020. Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization. In *International Workshop on Video Retrieval Evaluation (TRECVID 2020)*. NIST, Virtual Conference.

[12] Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13, 2 (2002), 415–425.

[13] Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29, 3 (2003), 333–347.

[14] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[15] Yong Jae Lee and Kristen Grauman. 2011. Face Discovery with Social Context. In *British Machine Vision Conference (BMVA)*. BMVA Press.

[16] Shan Li and Weihong Deng. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* (2020).

[17] Hao Ma, Irwin Kink, and Michael R. Lyu. 2012. Mining Web Graphs for Recommendations. *IEEE Transactions on Knowledge and Data Engineering* 24 (2012), 1051–1064.

[18] Pinar Santemiz, Luuk J. Spreeuwers, and Raymond N. J. Veldhuis. 2013. Automatic landmark detection and face recognition for side-view face images. In *International Conference of the BIOSIG Special Interest Group (BIOSIG)*. IEEE, Darmstadt, Germany.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Boston, MA, USA, 815–823.

[20] Mohammad Shafin, Rojina Hansda, Ekta Pallavi, Deo Kumar, Sumanta Bhattacharyya, and Sanjeev Kumar. 2019. Partial Face Recognition: A Survey. In *3rd International Conference on Advanced Informatics for Computing Research (ICAICR)*. Association for Computing Machinery, Shimla, India, 1–6.

[21] Paul Viola and Michael J. Jones. 2004. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 2 (2004), 137–154.

[22] Howard Wactlar and Michael Christel. 2002. Digital Video Archives: Managing through Metadata. In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Library of Congress, Washington, DC, USA, 84–99.

[23] Ivan William, De Rosal Ignatius Moses Setiadi, Eko Hari Rachmawanto, Heru Agus Santoso, and Christy Atika Sari. 2019. Face Recognition using FaceNet (Survey, Performance Test, and Comparison). In *4th International Conference on Informatics and Computing (ICIC)*. IEEE, Semarang, Indonesia.

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[25] Huiling Zhou and Kin-Man Lam. 2018. Age-invariant face recognition based on identity inference from appearance age. *Pattern Recognition* 76 (2018), 191–202.