# Understanding videos with face recognition: a complete pipeline and applications

Pasquale Lisena, Jorma Laaksonen, Raphaël Troncy

# Understanding Videos with Face Recognition: A Complete Pipeline and Applications

Pasquale Lisena[1*], Jorma Laaksonen[2] and Raphaël Troncy[1]

[1]EURECOM, Sophia Antipolis, France.
[2]Aalto University, Espoo, Finland.

*Corresponding author(s). E-mail(s): pasquale.lisena@eurecom.fr;
Contributing authors: jorma.laaksonen@aalto.fi;
raphael.troncy@eurecom.fr;

**Abstract**

When browsing or studying a video corpus, particularly relevant information consists in knowing who are the people appearing in the scenes. In this paper, we show how a combination of state of the art techniques can be organised in a pipeline for face recognition of celebrities. In particular, we propose a system which combines MTCNN for detecting faces and FaceNet for extracting face embeddings, which are used to train a set of classifiers. The face recognition results obtained at a frame level are then combined with those in consecutive frames, relying on automatic object tracking. Differently from previous work, we use images automatically retrieved by web search engines. We evaluate the systems one three datasets including historical videos from 1945-1969 and contemporary videos, obtaining a good precision score. In addition, we show how the obtained results can be applied to foster historical studies.

**Keywords:** Computer Vision, Face recognition, Web image retrieval, Knowledge Graphs

## 1 Introduction

TV archives host large amounts of video resources containing loads of hours of content. Often, these archives contain metadata that are being added by archivists during recurrent annotation tasks. These metadata may include

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

knowledge about the people appearing in the video, crucial information for searching, browsing and discovery, as well as for using the extracted data for obtaining statistics and training intelligent systems. For instance, person-related annotations may lead to learning interesting patterns of relationships among characters based on their appearance in the same news segment (Figure 1). This would enable interesting applications in historical research and media discovery.

For large corpora, relying solely on human annotations is not a scalable solution. Using artificial intelligence for computing digital annotations becomes necessary for identifying relevant people in videos [1].

The web offers an important amount of pictures of people and in particular of celebrities, easily findable using their full names as search terms in general-purpose search engines such as Google. While it has been considered a relevant information source in other communities – such as computational linguistics [2] and recommender system [3] – the web is still only scarcely exploited in image analysis and face recognition in particular.

In this paper, we aim to show that face recognition algorithms can be successfully trained on images crawled from the web and be applied for extracting relevant knowledge about the studied video corpus. We describe FaceRec, a pipeline combining state-of-the-art techniques for face recognition with an image crawling system from the web. In particular, FaceRec relies on Multi-task Cascaded Convolutional Networks (MTCNN) [4] for face detection and FaceNet [5] for computing face embeddings, in order to train a classifier for recognising faces at the frame level. A tracking system is included to increase the robustness of the library towards recognition errors in individual frames for getting more consistent person identifications. To this aim, the identification at frame level is then compared to those made in consecutive frames for the same face, which has been automatically tracked. We test our method on two datasets: ANTRACT composed of b/w videos from 1940s-60s, and MeMAD that includes TV news broadcasted from 2014.

While this work makes use of state-of-the-art technologies, without claiming to improve methods that already have a very low margin of error, this paper has two main contributions:

- for the first time, images automatically crawled from the web are used for training a face recognition system;
- we show how these technologies are performing in a complete pipeline and on two different video archives.

This paper is a follow up to a previous publication [6]. We have evaluated the system against an additional bigger dataset (ANTRACT Full), and we have included an in-depth analysis of the results obtained. Furthermore, we introduce some use cases for such a system that have proved to be very useful in the field of historical research.

The remaining of this paper is organised as follows. After highlighting some relevant work in Section 2, we describe our approach in Section 3. A quantitative evaluation is carried out on both a historical and a modern TV corpus in Section 4. An application case study is reported in Section 5, while access methods to the algorithm (an API and a web application) are described in Section 6. Finally, some conclusions and possible future work are outlined in Section 7.

## 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.

A survey of methods for face detection and tracking has been carried out in [7]. The Viola-Jones algorithm [8] for face detection and the Local Binary Pattern (LBP) features [9] for the clustering and recognition of faces were among 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 [10] 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) [4] 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. A boosted version of Multitask Joint Sparse Representation (MTJSR) has been used in [11], using multiple video frames for face identification. In [12], a common metric learning scheme is proposed for both the Euclidean space and a Riemannian manifold, in order to fuse appearance mean and pattern variation. In the video surveillance domain, we can mention the Trunk-Branch Ensemble CNN model (TBE-CNN) [13],

that combines two CNNs which deal with the entire face and patches around smaller details respectively.

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 [5].

In [14], 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 [15, 16] and in recent works [17]. In addition, MTCNN has been recognised to be very fast while having good performance [18].

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 algorithm for face recognition, but in the combination and application of available techniques 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

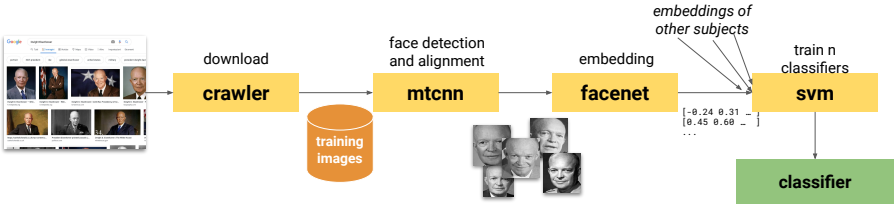



**Fig. 2**: FaceRec training pipeline

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[2], we apply to each image the **MTCNN algorithm** [4] for

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

[2] We decided to convert to greyscale because some preliminary experiments revealed not enough improvement, considering the increment of computation complexity of using 3 colour channels.

face detection[3]. 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 $(x_r, y_r)$[4] and given their original positions $(a_l, b_l)$ and $(a_r, b_r)$, the image is rotated by an angle $\alpha$ on the centre $c$ with scale factor $s$, computed in the following way:

$$dX = a_r - a_l \qquad dY = b_r - b_l \tag{1}$$

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

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[5]. This is realised by removing the faces with the highest cosine distance from the average vector among all FaceNet embeddings (using the same pre-trained model mentioned in the following) of the same person, until the standard deviation of all differences is under an empirically chosen threshold $\theta_{outlier} = 0.1$;
- allowing the user to further improve the automatic selection by allowing the exclusion of faces via a dedicated user interface (Section 6).

On the remaining pictures, a pretrained **FaceNet** [5] model with Inception ResNet v1 architecture trained on the VGGFace2 dataset [19] is applied

[3]We use the implementation provided at https://github.com/ipazc/mtcnn
[4]We use $x_l = 0.35w$, $x_r = (1 - x_l)$, and $y_l = y_r = 0.35h$.
[5]In this context, we are not taking care of high visual diversity in the images of one person, which can be due for example to ageing. In cases like Elizabeth II, with pictures publicly available for several decades, we decided to modify the search keyword for images to "Elizabeth II 1960. In other cases with high visual variation in less time – e.g. for Charles De Gaulle in 1940 and 1960 –, the Facenet embeddings were similar enough to not require splitting into two different classifiers.

for extracting visual features or embeddings of the faces. The embedding vectors feed ***n* parallel binary SVM**[6] classifiers, where $n$ is the number of distinct individuals to recognise. Each classifier is trained in a one-against-all approach [20], 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 setting – in the recognition phase – a confidence threshold for the candidate identities which does not depend on $n$, making the system scalable[7].

## 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. 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.

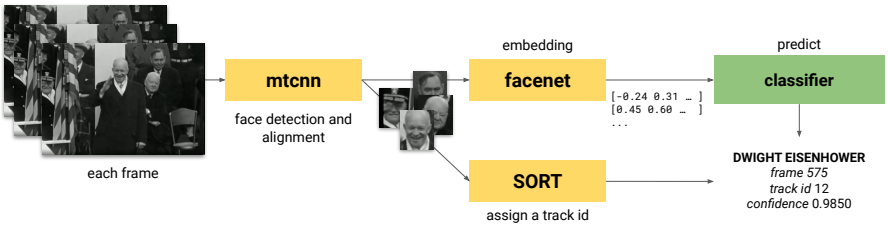



**Fig. 3**: FaceRec prediction pipeline

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 $[0, 1]$.

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 real-time[8] [21]. The algorithm uses the MTCNN bounding box detection and tracks the bounding boxes across frames, assigning a tracking id to each face.

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

[7] We also performed experiments on this system using a multi-class classifier with $n$ class, instead of the $n$ binary classifiers. 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.

[8] We used the implementation provided at https://github.com/Linzaer/Face-Track-Detect-Extract with some minor modification

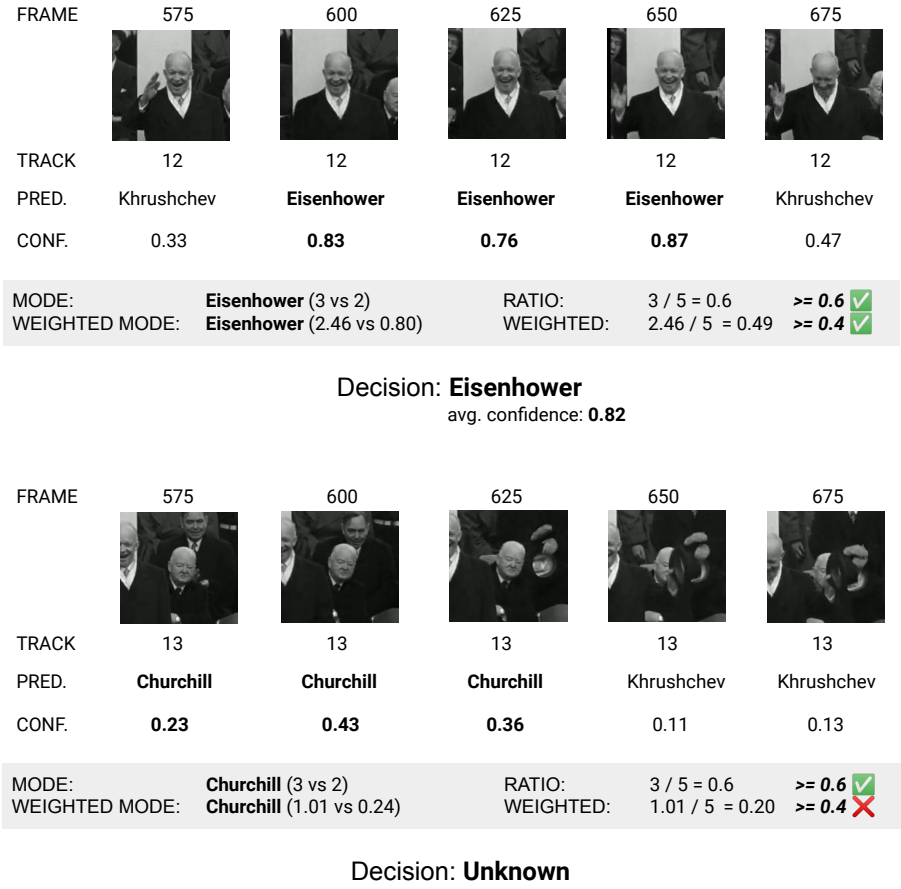



**Fig. 4**: The application of the decision method for assigning a unique label to tracking-ids for a positive (top) and a negative example (bottom)

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 track $t$ – including a certain number of samples $n_t$ – we compute the mode[9] of all predictions, as well as the weighted mode with respect to the confidence scores. The weighted mode $m_w$ is computed using the following formula, where $P$ are all the predicted labels

[9]The mode is "the number or value that appears most often in a particular set" (*Cambridge Dictionary*)

$p$ for a single tracking id, $c_p$ is the confidence score for the prediction $p$, $x$ is any distinct predicted value[10]:

$$m_w = \arg\max_x \left(\sum_{p=x}^{p \in P} c_p\right)$$

A unique predicted label $p$ is chosen including among all the possible predictions if it satisfies all the following conditions:

- $p$ is equal to both the mode and the weighted mode;
- the ratio of samples with prediction $p$ over the total number of samples $n_t$ is greater than the threshold $h$;
- the ratio of samples with prediction $p$ over the total number of samples $n_t$, weighting all occurrences with the confidence score, is greater than the threshold $h_w$.

Two examples of applications of these conditions are shown in Figure 4.

We empirically found that $\theta_m = 0.6$ and $\theta_w = 0.4$ are the best values for the thresholds. It is possible that the tracking process does not produce a label fulfilling all the conditions. In that case, the prediction is considered uncertain and the tracking id is excluded from the results. We assign to the track 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 who may re-occur. The applied method is represented in Figure 5.

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 through hierarchical clustering[11] based on a distance threshold, empirically set to $\theta_d = 14$. The clustering produces a variable number $m$ of

[10] The mode can be seen a generalisation of the weighted mode, putting all weights (in our formula, $c_p$) to 1.

[11] We used the implementation available in SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

clusters, with all items assigned to one of them. The clusters are then filtered to exclude:

- the clusters for which we can already assign a label from our training set;
- the clusters 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 $\theta_{clustering} = 1.3$;
- the clusters 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 ($dX$ and $dY$ have been defined in (1)):

$$ddG = m_l - a_l \qquad dH = n_l - b_l$$

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

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

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 $n_{clustering} = 5$ 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 counter – 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.


**Fig. 5**: The clustering strategy in FaceRec

# 4 Evaluation

In this section, we evaluate the FaceRec system measuring the precision and recall on three different datasets: two composed of historical videos and one composed of more recent TV footage. We report in Table 1 all the required parameters, with the values used in this paper.

| Parameter | Value | Description |
|---|---|---|
| $k$ | 50 | Number of photos to crawl from the web |
| $w$, $h$ | 256, 256 | Dimensions of the images in the training set |
| $xl$, $xr$, $yl$, $yr$ | $0.35w$, $1xl$, $0.35h$, $0.35h$ | Desired position for the eyes on face images |
| $\theta_{outlier}$ | 0.1 | Threshold for the outlier detection |
| $T$ | 25 | Sampling period for frame processing |
| $\theta_m$ | 0.6 | Threshold for the mode |
| $\theta_w$ | 0.4 | Threshold for the weighted mode |
| $t$ | 0.5 – 0.6 | Confidence threshold |
| $\theta_d$ | 14 | Distance threshold for clustering |
| $\theta_{clustering}$ | 1.3 | Final threshold for clustering |
| $r_{dist}$ (max) | 0.6 | Max value of face rotation |
| $n_{clustering}$ | 5 | Number of faces kept for each cluster |

**Table 1**: Overview of parameters for FaceRec, with the final values used in evaluation

## 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 datasets**. *Les Actualités françaises*[12] 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)*[13]. The videos are in black-and-white, with a resolution of 512×384 pixels. Metadata are collected through INA's *Okapi* platform [22, 23], which exposes a SPARQL endpoint.

Two lists of historically well-known people have been provided by domain experts, and we derived from these list two subsets.

The first list includes 13 celebrities. From the metadata, we have obtained the reference to the segments in which these people appear and the subdivision of these segments in shots[14]. This search produced 15,628 shots belonging to 1,222 segments from 702 distinct media resources. 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

[12] https://www.ina.fr/emissions/les-actualites-francaises/
[13] The corpus can be downloaded from https://dataset.ina.fr/
[14] 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/)

manual correction, adding also some shots not involving any of the specified people. In the end, it includes 198 video shots (belonging to 129 distinct media resources), 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. This **ANTRACT Gold** dataset can be considered a gold standard.

A second list includes 121 celebrities. In comparison to the first list, the amount of videos in which these celebrities appear does not allow a manual inspection of each video. As a result, the studied temporal fragments are less granular. In addition, the actual presence of the face of the person in the video has not been confirmed by human observation, so it cannot be considered a gold standard. This **ANTRACT Full** dataset contains over 5,000 records, belonging to over 1,000 media resources.

**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[15], with metadata available from the MeMAD's Knowledge Graph[16]. We followed the same procedure as 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 we did 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. This dataset can be considered a gold standard.

Table 2 summarises the main differences between the 3 datasets.

| | **ANTRACT Gold** | **ANTRACT Full** | **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) | 121 (chosen by domain experts) | 6 (most present in KG) |
| represented fragment and length | shot 3 seconds in avg. | segment 77 seconds in avg. | segment up to 2 minutes |
| records | 216 | 5413 | 100 |
| distinct fragments | 198 | 2698 | 100 |
| distinct media (videos) | 129 | 1117 | 30 |
| fragments without known faces | 39 | 0 | 43 |

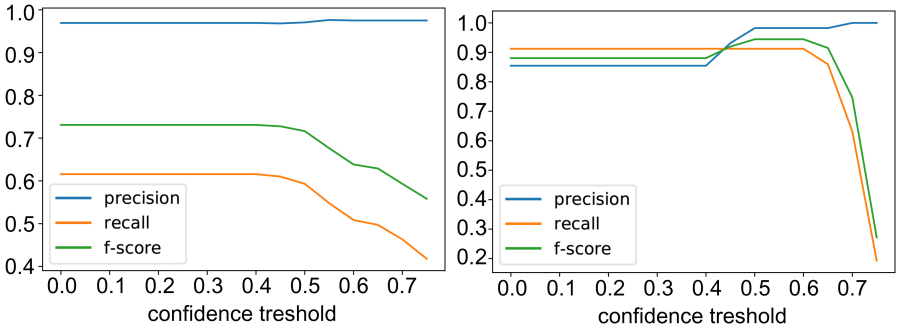
**Table 2**: Description of the ANTRACT and MeMAD datasets

[15] https://memad.eu/
[16] https://data.memad.eu/

## 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 training set consists of images crawled on the web, using the method described in Section 3.1. The model has then been applied to the video fragments of the ANTRACT and MeMAD datasets (shot or segment), of which we processed 1 frame per second. For each fragment, we check if we have found the expected person.

We varied the confidence threshold $t$ under which we considered the face not recognised as shown in Figure 6, 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 3 and Table 4.




(a) (b)

**Fig. 6**: Precision, recall and F-score of FaceRec on different confidence thresholds for the ANTRACT Gold (6a) and the MeMAD datasets (6b).

The system obtains 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 Gold 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. While modern news are more used to close-up shots, taken on screen for multiple seconds, we observe that 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 fewer samples for recognition. It is also relevant to notice that the lowest recall values belong to the only two USSR politicians Khrushchev and Molotov: most often, they appear in group images or in very short close-up images, raising questions for historical research.

The gap between precision and recall obtained on the ANTRACT Gold dataset is confirmed by the results obtained on ANTRACT Full, reported in Table 5. The percentiles show that more than half of the people in the class set

| 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 3**: ANTRACT Gold 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.

| 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 4**: 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.

have been always correctly predicted. On the other side, the recall is dropping fast, being less than 1% of average. This is due to the possible actual absence of these people in the image, as already mentioned in Section 4.1. In addition, we should mention that the training set of faces is more prone to include noise (wrong or low-quality images), because of the high number of celebrities to search for, and the shortage of available images for some of them. Further work is required to filter out the noisy images in a pre-processing step.

## 4.3 Qualitative analysis

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

| Person | P | R | F | S |
|---|---|---|---|---|
| Min (6 different celebrities) | 0.00 | 0.00 | 0.00 | - |
| 25% | 0.50 | 0.00 | 0.00 | - |
| 50% | 1.00 | 0.06 | 0.11 | - |
| 75% | 1.00 | 0.14 | 0.23 | - |
| Max (Fernandel) | 1.00 | 0.55 | 0.71 | 4 |
| average (unknown apart) | 0.69 | 0.08 | 0.12 | 5413 |

**Table 5**: ANTRACT Full dataset: aggregate statistics about precision, recall, F-score and support, with percentiles. The support column corresponds to the number of segments in which the person appears.

- The system generally fails to detect people when they are in the background and their faces are therefore relatively small. This is particularly true in 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 kinds of objects.
- The used embeddings are not suitable to represent side faces, whose predictions are not reliable.

Figure 7 shows some examples of faces which were not predicted.

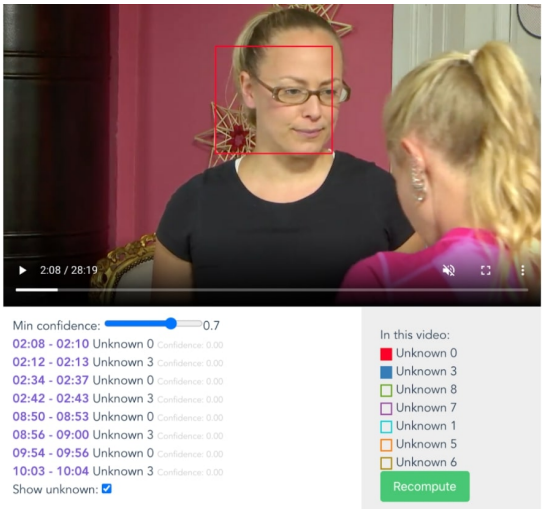
(a) (b) (c)

**Fig. 7**: Examples of not predicted faces for occlusion (a), side face (b) and low resolution (c).

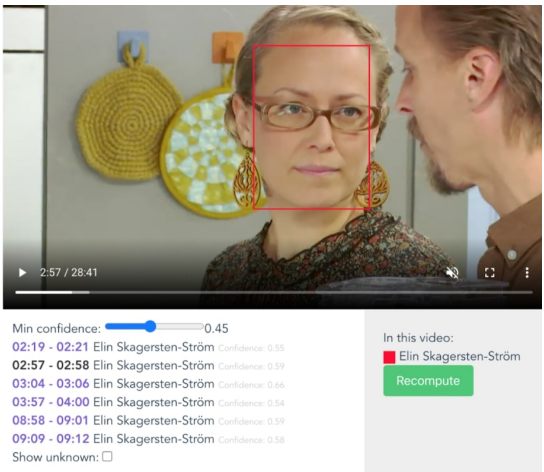
## 4.4 Unknown cluster detection evaluation

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




(a)




(b)

**Fig. 8**: The clustering output found a set of unknown persons in the video (8a). Using the frames of *Unknown 0*, we are able to build the model for Elin Skagersten-Strm and recognise her in other videos. (8b).

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.
- Fewer clusters were found in the ANTRACT dataset than in the MeMAD dataset – 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 8 an example use case. In Figure 8a, the clustering algorithm identified a set of unknown people, among which *Unknown 0* happens to be Elin Skagersten-Strm, 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 8b. 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 Face Recognition for Understanding Video Corpus

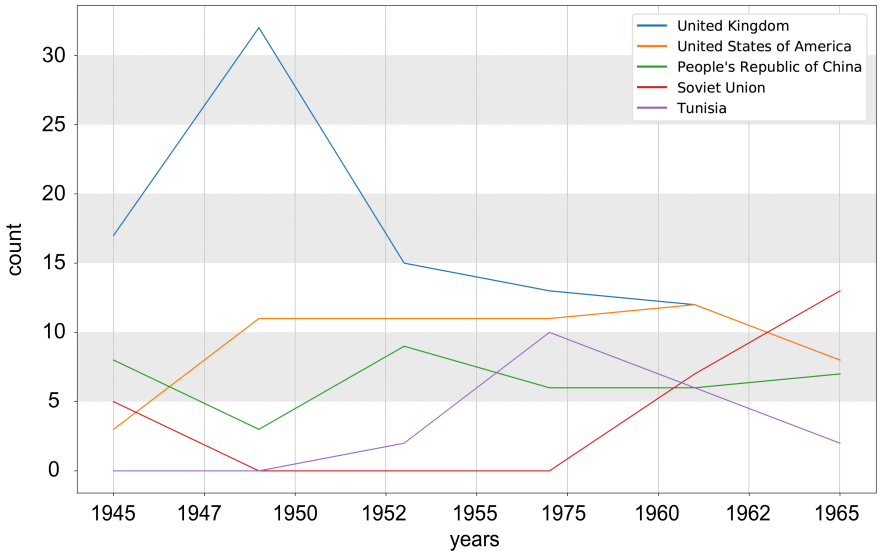



**Fig. 9**: Presence in video in 3-years windows of people coming by the first 5 represented countries (France excluded).

How can face recognition results give relevant insights about a particular video corpus? We used the results obtained on the ANTRACT Full corpus

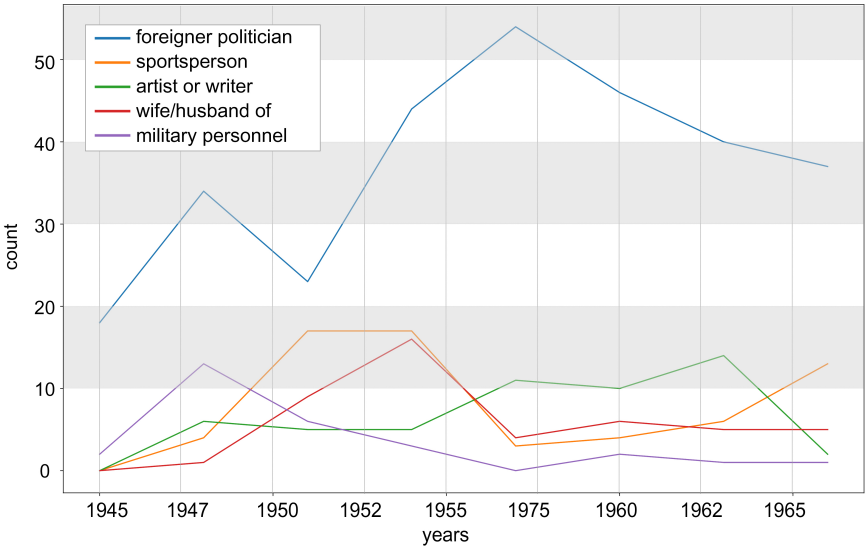



**Fig. 10**: Presence in video in 3-years windows of people belonging to 5 different groups (French politicians excluded).

in order to extract some statistics about the people involved in French news between 1945 and 1965.

Matching the results with the available metadata, we can study the evolution of media presence throughout the years. In Figure 9, we plotted the presence in video of people aggregated by nationality, excluding France – largely over-represented. We can make the following observations:

- the Allies' members had a greater exposition in media immediately after World War II in comparison to others;
- for a time window of more than 10 years (1947-1959), the relationships with the Soviet Union were absent from media or sovietic personalities are shown only in groups, making harder the recognition by the system;
- the pick of attention for Tunisia matches with the independence of the country in 1956.

In Figure 10, we aggregate the presence of people according to their role: French politicians (excluded in the figure because they are over-represented), foreign politicians, sportspeople, artists and writers, partners of other celebrities, military people. We can observe that the military people group was quite exposed in the news between 1945 and 1947, and slowly disappear to make way for artists and sportspeople, demonstrating the beginning of a peace period. From 1950 to 1960 the presence of international politicians doubled, a probable sign of a more interconnected world.

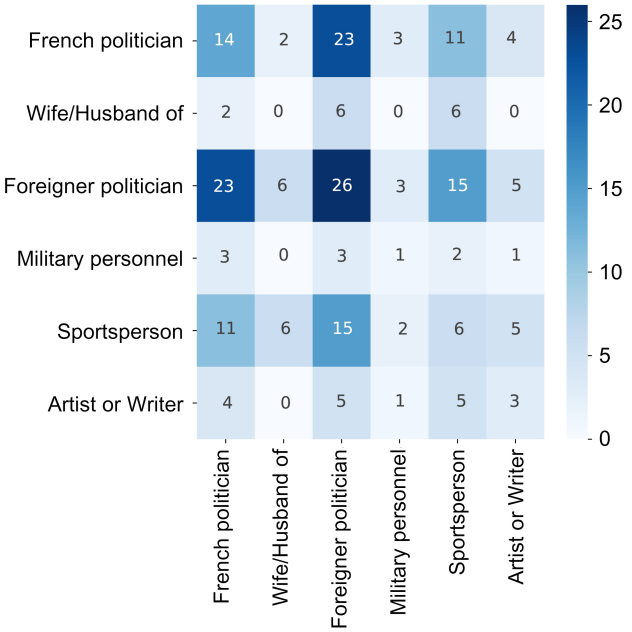



**Fig. 11**: Co-occurrences in video in a 2 minutes window of people, grouped by categories.

We also wanted to study how celebrities co-occur in videos. For doing so, we extracted all couples of people who have been recognised in the same video at a maximum time distance of 2 minutes. The aggregated results, grouped using the same classification as before, are reported in Figure 11. The high number of co-occurrences of foreign politicians suggests the presence of segments of the news dedicated to foreign politics. It is interesting to observe that politicians have "encountered" sportspeople more often than artists, possibly because they were involved in ceremonies following sports competitions.

In Figure 12, the co-occurrence is shown aggregating by gender. The data shows a quite underrepresented female presence. The data shows that encounters among women were almost absent in the news (only 2%), but this is also due to a strongly unbalanced training set, in which only 10% of individuals are women.

## 6 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 to be compatible with

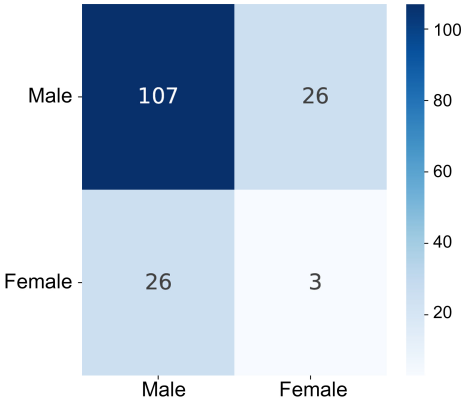



**Fig. 12**: Co-occurrences in video in a 2 minutes window of people, grouped by gender.

the OpenAPI specification[17] and documented with the Swagger framework[18]. 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 semantic-rich format in RDF using the Turtle syntax, relying on the EBU Core[19] and Web Annotation ontologies[20]. The Media Fragment URI[21] syntax 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 13). 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

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

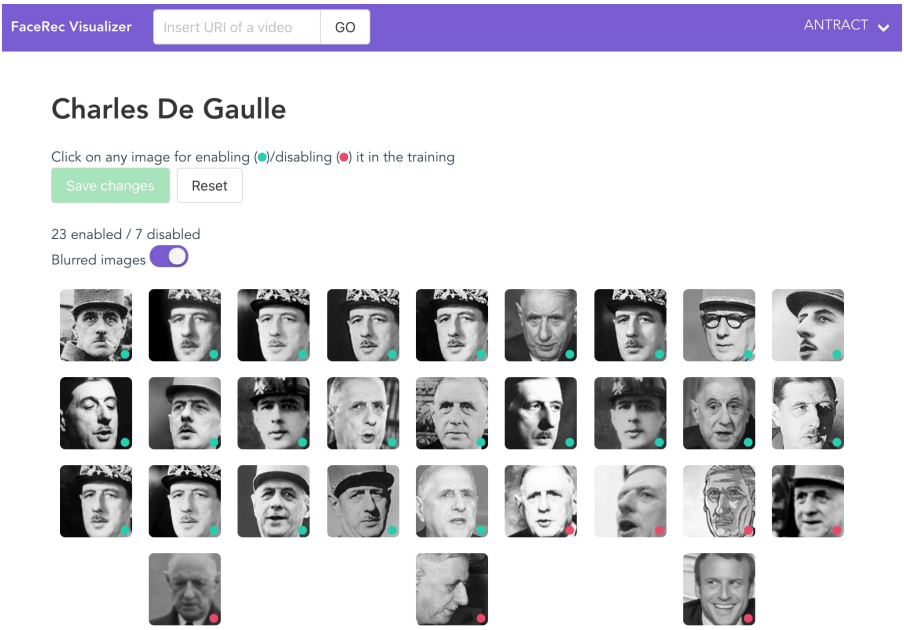



**Fig. 13**: Person page in FaceRec Visualizer: drawings, side faces, other depicted individuals and low-quality images are discarded (see the last 7 pictures marked with the red dot).

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 as 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 8). 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.

# 7 Conclusions and future work

In this paper, we have shown how face recognition can be trained on celebrities' pictures from the web and be applied for studying a video corpus. To do this, we relied on FaceRec, a pipeline combining some of the best-performing state-of-the-art algorithms. The followed approach revealed good performance, with almost perfect precision, with some margin for improvement on the recall, in particular when the original video quality is challenging – i.e. in the case of historical videos. The system is publicly available at https://git.io/facerec under an open source licence.

The results of FaceRec can be applied to different tasks. In this paper, we have shown how aggregate results can tell us more about a video corpus.

Another application is video summarisation, and we have proven that combining face recognition, automatically-generated visual captions and textual analysis is an effective strategy [24].

In future work, we plan to improve the performance 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 relying on relevant approaches from the literature [25, 26]. 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, 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 [27], and textual descriptions, captions and audio in multimodal approaches [28, 29].

The presented work has several potential applications, from annotation and cataloguing to automatic captioning, with possible inclusion in second-screen TV systems. Moreover, it can support future research in computer vision or in other fields – e.g. history studies. Currently, FaceRec is being used by history researchers who are studying past meetings between political figures using the ANTRACT corpus. They are able to more easily jump to the parts of the video where such an encounter between two or more political figures took place thanks to the automatic detection we are providing. They can also filter by the roles that a given person had at a time, thanks to external knowledge about each personality available in Wikipedia or in the Wikidata knowledge graph.

An interesting application is the study of age progression in face recognition [30]. Finally, we intend to use the results obtained on historical corpora in order to extract patterns about the on-screen presence of relevant people, in particular regarding the field size (close-up, full size, etc.), the duration of the shots, the presence or not of other people, and the correlation of these elements with the role and importance of the studied celebrities.

## Acknowledgement

The authors would like to thank Bénédicte Pincemin for the valuable feedback, which helped to improve the paper. 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 Academy of Finland (project grants n. 329268 and 345791).

## References

[1] Wactlar, H., Christel, M.: Digital Video Archives: Managing through Metadata. In: Building a National Strategy for Digital Preservation: Issues

in Digital Media Archiving, pp. 84–99. Library of Congress, Washington, DC, USA (2002)

[2] Kilgarriff, A., Grefenstette, G.: Introduction to the Special Issue on the Web as Corpus. Computational Linguistics **29**(3), 333–347 (2003)

[3] Ma, H., Kink, I., Lyu, M.R.: Mining Web Graphs for Recommendations. IEEE Transactions on Knowledge and Data Engineering **24**, 1051–1064 (2012)

[4] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)

[5] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. IEEE Computer Society, Boston, MA, USA (2015)

[6] Lisena, P., Laaksonen, J., Troncy, R.: FaceRec: An Interactive Framework for Face Recognition in Video Archives. In: 2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021), New York, USA (2021). https://doi.org/10.5281/zenodo.4764632

[7] Vij, R., Kaushik, B.: A survey on various face detecting and tracking techniques in video sequences. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 69–73 (2019). https://doi.org/10.1109/ICCS45141.2019.9065483

[8] Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision **57**(2), 137–154 (2004)

[9] Ahonen, T., Hadid, A., Pietikinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence **28**(12), 2037–2041 (2006)

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

[11] Liu, L., Zhang, L., Liu, H., Yan, S.: Toward Large-Population Face Identification in Unconstrained Videos. IEEE Transactions on Circuits and Systems for Video Technology **24**(11), 1874–1884 (2014). https://doi.org/10.1109/TCSVT.2014.2319671

[12] Huang, Z., Wang, R., Shan, S., Van Gool, L., Chen, X.: Cross euclidean-to-riemannian metric learning with application to face recognition from

video. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2827–2840 (2018). https://doi.org/10.1109/TPAMI.2017.2776154

[13] Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 1002–1014 (2018). https://doi.org/10.1109/TPAMI.2017.2700390

[14] William, I., Ignatius Moses Setiadi, D.R., Rachmawanto, E.H., Santoso, H.A., Sari, C.A.: Face Recognition using FaceNet (Survey, Performance Test, and Comparison). In: $4^{th}$ International Conference on Informatics and Computing (ICIC). IEEE, Semarang, Indonesia (2019)

[15] Guo, G., Zhang, N.: A survey on deep learning based face recognition. Computer Vision and Image Understanding **189** (2019)

[16] Shafin, M., Hansda, R., Pallavi, E., Kumar, D., Bhattacharyya, S., Kumar, S.: Partial Face Recognition: A Survey. In: $3^{rd}$ International Conference on Advanced Informatics for Computing Research (ICAICR), pp. 1–6. Association for Computing Machinery, Shimla, India (2019)

[17] Ali-Gombe, A., Elyan, E., Zwiegelaar, J.: Towards a Reliable Face Recognition System. In: Iliadis, L., Angelov, P.P., Jayne, C., Pimenidis, E. (eds.) $21^{st}$ Engineering Applications of Neural Networks Conference (EANN), pp. 304–316. Springer, Cham (2020)

[18] Li, S., Deng, W.: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing (2020)

[19] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: $13^{th}$ IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 67–74. IEEE Computer Society, Xi'an, China (2018)

[20] Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks **13**(2), 415–425 (2002)

[21] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple Online and Realtime Tracking. In: IEEE International Conference on Image Processing (ICIP), pp. 3464–3468. IEEE Computer Society, Phoenix, AZ, USA (2016)

[22] Beloued, A., Stockinger, P., Lalande, S.: 4. Studio Campus AAR: A Semantic Platform for Analyzing and Publishing Audiovisual Corpuses, pp. 85–133. John Wiley & Sons, Ltd, Hoboken, NJ, USA (2017)

[23] Carrive, J., Beloued, A., Goetschel, P., Heiden, S., Laurent, A., Lisena, P., Mazuet, F., Meignier, S., Pinchemin, B., Poels, G., Troncy, R.: Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. Digital Humanities Quarterly, Special Issue on AudioVisual Data in DH **15**(1) (2021)

[24] Harrando, I., Reboud, A., Lisena, P., Troncy, R., Laaksonen, J., Virkkunen, A., Kurimo, M.: 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 (2020)

[25] Santemiz, P., Spreeuwers, L.J., Veldhuis, R.N.J.: Automatic landmark detection and face recognition for side-view face images. In: International Conference of the BIOSIG Special Interest Group (BIOSIG). IEEE, Darmstadt, Germany (2013)

[26] Haider, H., Khiyal, M.: Side-View Face Detection using Automatic Landmarks. Journal of Multidisciplinary Engineering Science Studies **3**, 1729–1736 (2017)

[27] Lee, Y.J., Grauman, K.: Face Discovery with Social Context. In: British Machine Vision Conference (BMVA). BMVA Press, Dundee, UK (2011)

[28] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems **16**(6), 345–379 (2010)

[29] Handa, A., Agarwal, R., Kohli, N.: A survey of face recognition techniques and comparative study of various bi-modal and multi-modal techniques. In: $11^{th}$ International Conference on Industrial and Information Systems (ICIIS), pp. 274–279. IEEE, Roorkee, India (2016)

[30] Zhou, H., Lam, K.-M.: Age-invariant face recognition based on identity inference from appearance age. Pattern Recognition **76**, 191–202 (2018)