EBU

OPERATING EUROVISION AND EURORADIO

TECHNOLOGY & INNOVATION

Get an edge

AI-BENCHMARKING AT THE EBU

ALBERTO MESSINA (RAI),
ALEXANDRE ROUXEL (EBU),
SEBASTIEN DUCRET (RTS)

EBU

OPERATING EUROVISION AND EURORADIO

TECHNOLOGY & INNOVATION

Get an edge

AI-BENCHMARKING AT THE EBU

INTRODUCTION

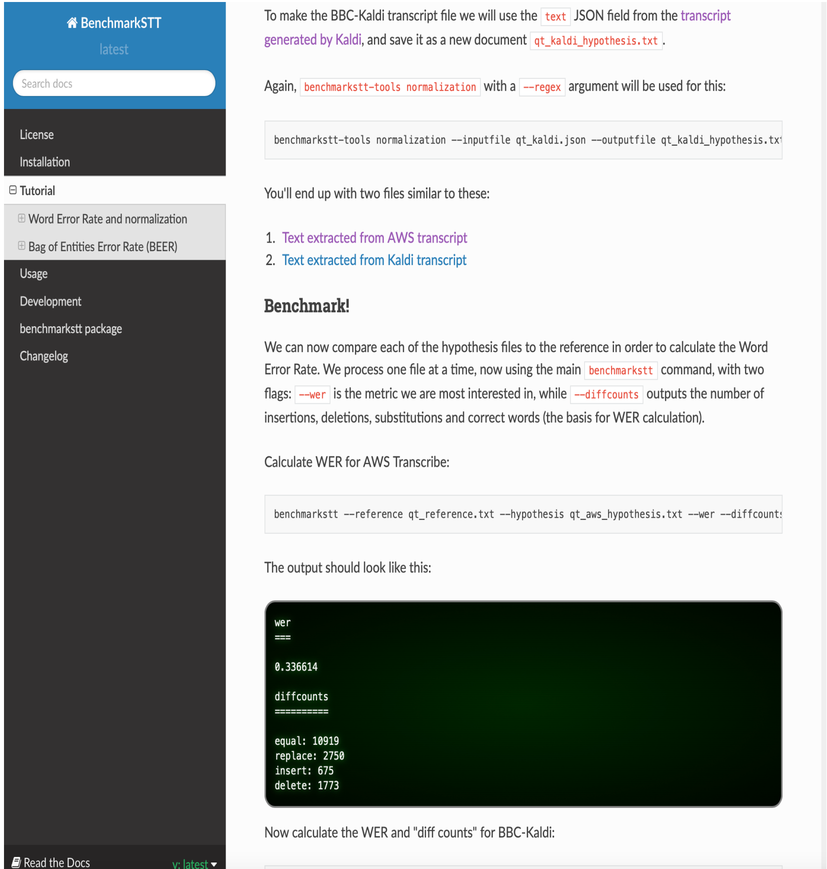
After successfully delivering an open source tool to evaluate speech-to-text solutions, the EBU AI-Benchmarking Group is now working on facial recognition.

Our goal is to go further and develop an open-source platform to evaluate facial recognition systems for video, provide metrics and share the evaluations of open source and market solutions.

1. **BENCHMARKSTT**
2. **FACIAL RECOGNITION**
3. **DATASET**
4. **MODULAR PIPELINE**
5. **OPEN QUESTIONS**

BENCHMARKSTT IN A NUTSHELL

- GitHub:
  - Docker image
  - JSON-RPC API
- Read the Doc
  - Rich documentation
- PyPi
  - Easy to integrate in your workflows



To make the BBC-Kaldi transcript file we will use the `text` JSON field from the transcript generated by Kaldi, and save it as a new document `qt_kaldi_hypothesis.txt`.

Again, `benchmarkstt-tools normalization` with a `--regex` argument will be used for this:

```
benchmarkstt-tools normalization --inputfile qt_kaldi.json --outputfile qt_kaldi_hypothesis.tx
```

You'll end up with two files similar to these:

1. Text extracted from AWS transcript
2. Text extracted from Kaldi transcript

Benchmark!

We can now compare each of the hypothesis files to the reference in order to calculate the Word Error Rate. We process one file at a time, now using the main `benchmarkstt` command, with two flags: `--wer` is the metric we are most interested in, while `--diffcounts` outputs the number of insertions, deletions, substitutions and correct words (the basis for WER calculation).

Calculate WER for AWS Transcribe:

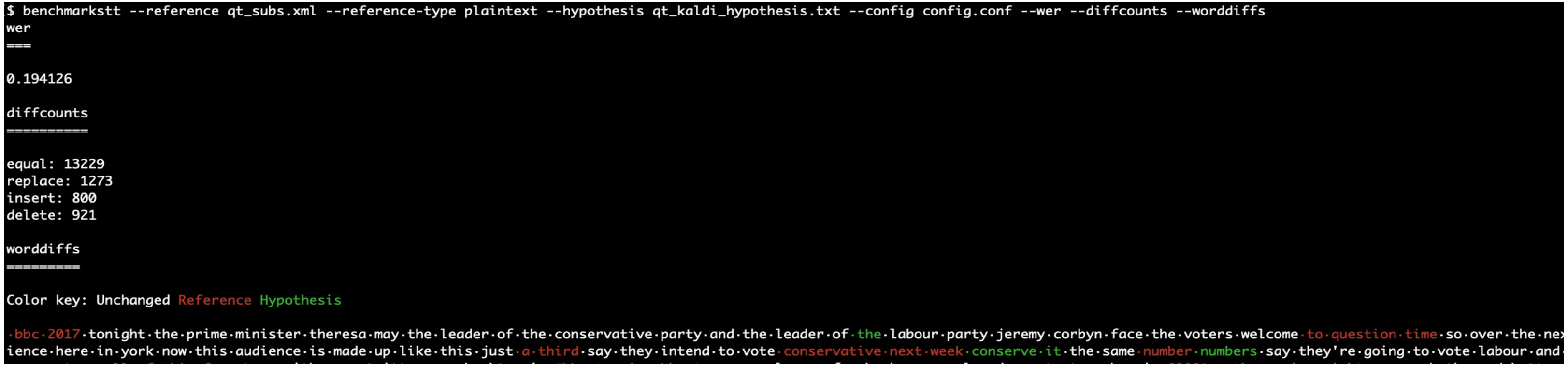
```
benchmarkstt --reference qt_reference.txt --hypothesis qt_aws_hypothesis.txt --wer --diffcount
```

The output should look like this:

```
wer
===

0.336614

diffcounts
==========

equal: 10919
replace: 2750
insert: 675
delete: 1773
```

Now calculate the WER and "diff counts" for BBC-Kaldi:

WORD ERROR RATE



```
$ benchmarkstt --reference qt_subs.xml --reference-type plaintext --hypothesis qt_kaldi_hypothesis.txt --config config.conf --wer --diffcounts --worddiffs
wer
===

0.194126

diffcounts
==========

equal: 13229
replace: 1273
insert: 800
delete: 921

worddiffs
=========

Color key: Unchanged Reference Hypothesis

·bbc·2017·tonight·the·prime·minister·theresa·may·the·leader·of·the·conservative·party·and·the·leader·of·the·labour·party·jeremy·corbyn·face·the·voters·welcome·to·question·time·so·over·the·nex
ience·here·in·york·now·this·audience·is·made·up·like·this·just·a·third·say·they·intend·to·vote·conservative·next·week·conserve·it·the·same·number·numbers·say·they're·going·to·vote·labour·and·
```

- WER is the most common metric to evaluate the quality of the transcripts
- WER is sensitive to the normalisation of the reference and hypothesis texts. But the normalisation rules depend on the languages and the use cases.
- BenchmarkSTT provides simple commande to normalize and compute the metric in one line !

$$WER = \frac{R + I + \ D}{N}$$

- R = number of replacement, substituion
- I = number of insertion
- D = number of deletion
- N = number of words in the reference document

BAG OF ENTITIES ERROR RATE

$$BEER(entity) = \frac{\left|n_{hyp} - n_{ref}\right|}{n_{ref}}$$

n_{ref}= number of occurences of entity in the reference document

n_{hyp} = number of occurences of entity in the hypothesis document

- The WER treats all words as equally important but in reality some words, like proper nouns, key words, phrases are more significant than common words.
- The BEER measures the quality of the transcriptions of a set of entities. An entity is a word or an ordered list of words including capital letters and punctuation.

FACIAL RECOGNITION BENCHMARKING

- Why do we need to develop an AI benchmarking framework for broadcasters?
  - There is no open source framework with state-of-the-art models available for video
  - No benchmarking of open source and market solutions for video
- We want to provide both!

OPEN-SOURCE FOR FACIAL RECOGNITION

- It is an active topic in research and many published solutions are available on GitHub with open source licences. The state of the art is moving fast:
    - Detection : MTCNN, RetinaFace
    - Recognition : Inception-ResNet, Arcface
    - Alignment : SDUNet
- There are many open source frameworks that integrate state of the art models
    - Open source Frameworks:
        - InsightFace : RetinaFace; SCRFD ; SDUNet ; ArcFace …
        - FaceNet : MTCNN and Inception-ResNet
        - DeepFace : FaceNet, OpenFace, ArcFace …
- But not for videos !

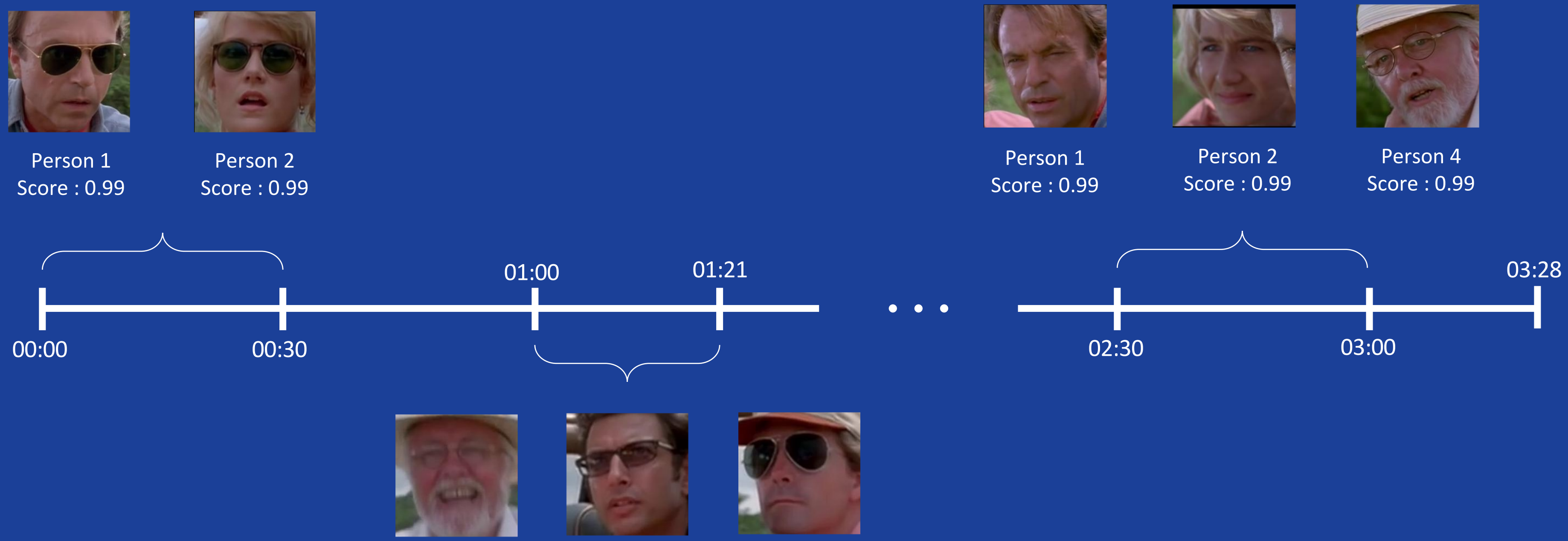
MARKET PLACE SOLUTIONS

- A plethora of paid solutions
  - Amazon Rekognition
  - Azure Face API
  - Google Cloud Patform (restricted access to media and entertainment companies)
  - Deep Vision AI, Kairos ...
- We have already integrated AWS recognition and developed serverless workflows with EBU MCMA, we will integrate more.
- Is it better than open source ?

ANNOTATED VIDEO DATASET

- How did we proceed ?
- Manual annotation of videos is time consuming and expensive:
  - extract each frame with celebrities
  - label all the faces
- Semi-automatic annotation process : a strategy to reduce the cost of the annotation
  - we defined intervals
    - 30 second maximum
    - maximum 3 different persons in it
  - we segment automatically the videos in intervals
    - with a face detection and face clustering strategy

› Max length: 30s / Max persons : 3



Person 1
Score : 0.99

Person 2
Score : 0.99

Person 1
Score : 0.99

Person 2
Score : 0.99

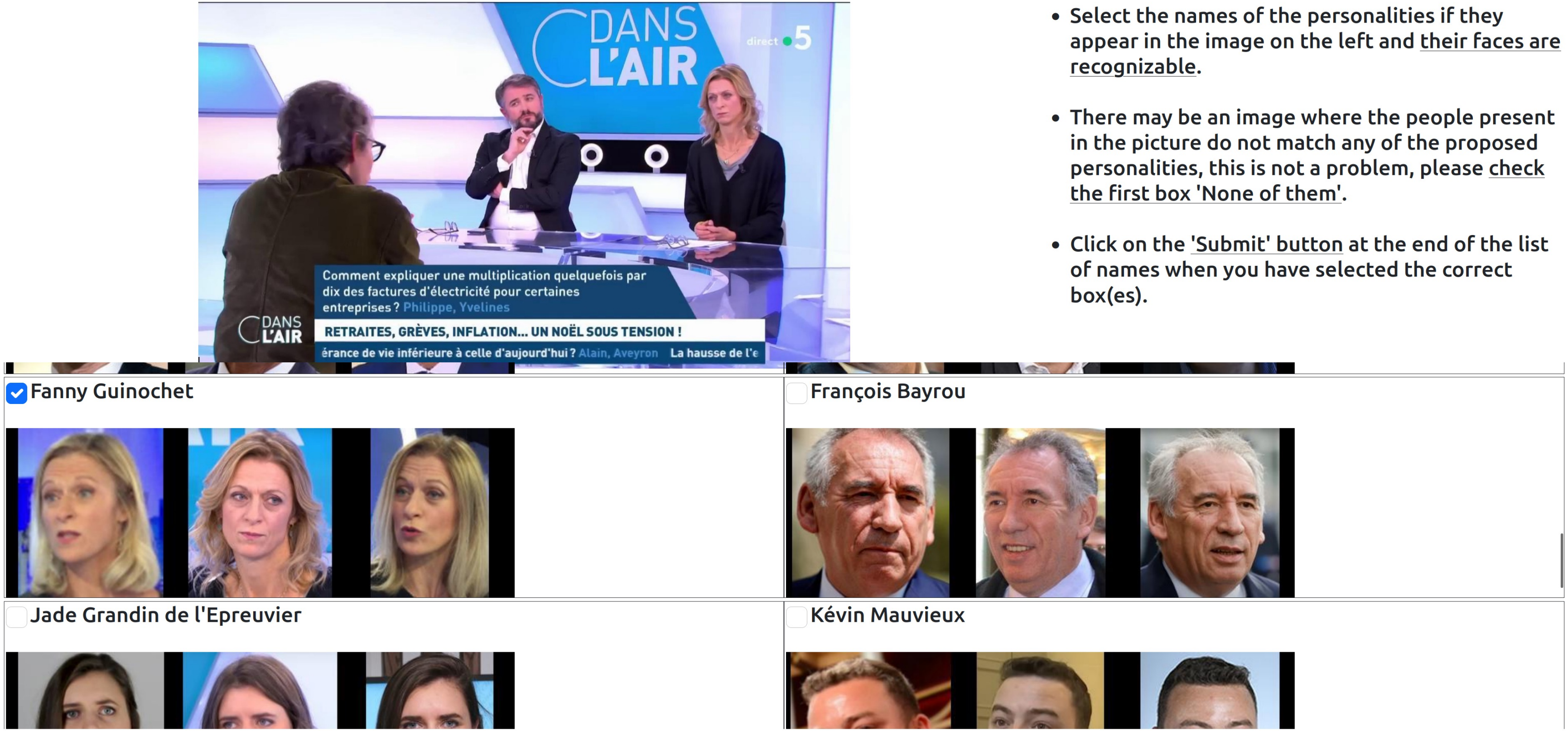
Person 4
Score : 0.99

Person 4
Score : 0.98

Person 3
Score : 0.98

Person 5
Score : 0.96

AWS MKTURC WORKERS INTERFACE


DANS L'AIR
direct 5
Comment expliquer une multiplication quelquefois par dix des factures d'électricité pour certaines entreprises ? Philippe, Yvelines
RETRAITES, GRÈVES, INFLATION... UN NOËL SOUS TENSION !
érance de vie inférieure à celle d'aujourd'hui ? Alain, Aveyron La hausse de l'e

- Select the names of the personalities if they appear in the image on the left and their faces are recognizable.

- There may be an image where the people present in the picture do not match any of the proposed personalities, this is not a problem, please check the first box 'None of them'.

- Click on the 'Submit' button at the end of the list of names when you have selected the correct box(es).

☑ Fanny Guinochet

□ François Bayrou

□ Jade Grandin de l'Epreuvier

□ Kévin Mauvieux

EBU DATASET

- 335 videos – 78h13m27s
- 700 celebrities
- 1 key frame every 5s
- 3 or 5 annotators per keyframe
- TV content from :
  - RAI, RTS,FTV,BBC
- We select videos with metadata
  - hosts and guests names
- Post-processing :
  - verification of annotated frames that seem suspicious

- RTS : 5 videos (4:58:49)
- RAI : 20 videos (2:45:01)
- BBC : Graham Norton Show : 24 videos (19:06:51)
- Taratata : 50 videos (8:32:41)
- C dans l'air : 10 videos (10:37:36)
- C politique : 8 videos (1:20:01)
- C l'Hebdo : 40 videos (10:11:54)
- Vivement Dimanche : 40 videos (3:02:58)
- C ce soir : 11 videos (3:01:28)
- Télé Matin: 39 videos (4:46:46)
- C à vous : 88 videos (9:49:22)

RESULT

- JSON format

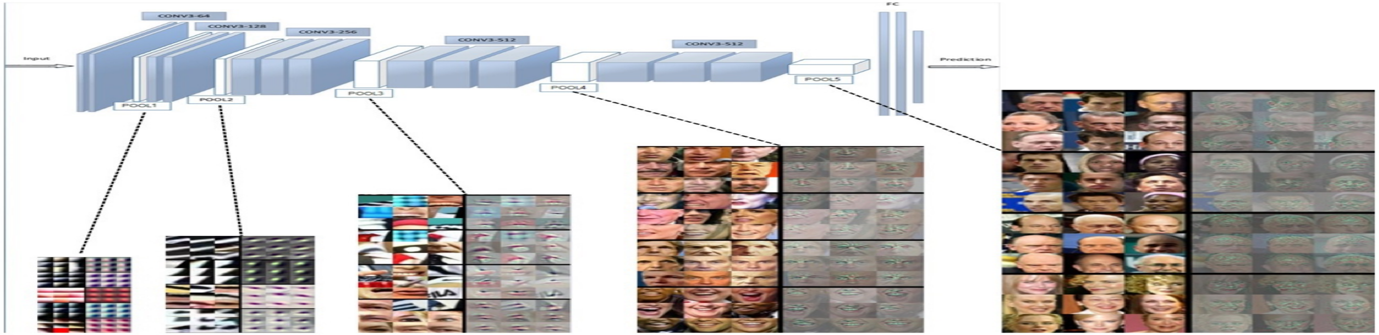
```
{
video_title: Title of the video,
video_url: YouTube or Dailymotion url (if available),
program_name: Name of the program or broadcaster,
all_personalities: [Personality 1, Personality 2, …]
annotation: {
   Interval_id: {
      time_interval:[start_time, end_time, step_second],
      frame_interval:[start_frame, end_frame, step_frame],
      personalities: [Personality 1, Personality 2, …]
   }
}
```

- Post-processed dataset available on an AWS s3 bucket

FACIAL RECOGNITION PRINCIPLE

- Let’s have a quick look at the concepts
- The main blocks can be isolated and explained simply.
- We start by explaining facial recognition on images, then move on to the video pipeline.

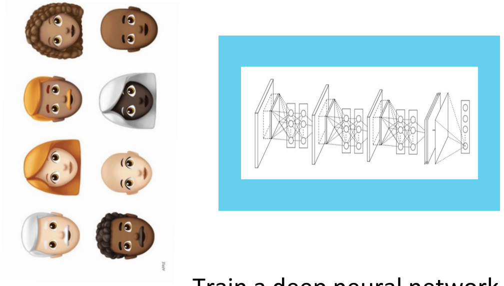
DEEP NEURAL NETWORKS



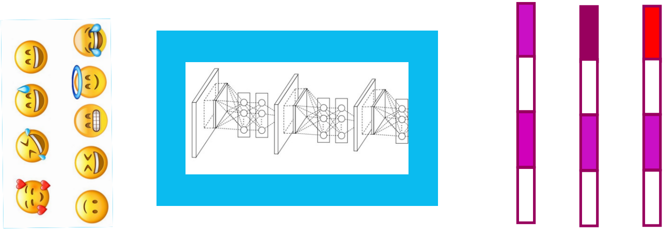
I. Masi, Y. Wu, T. Hassner and P. Natarajan, "Deep Face Recognition: A Survey," 2018, SIBGRAPI

- Deep Neural Networks generate hierarchical features
  - The first layer is similar to filters created by humans for images processing, 30 years ago
  - The higher layers learn more complex features that are humanly understandable

EMBEDDINGS



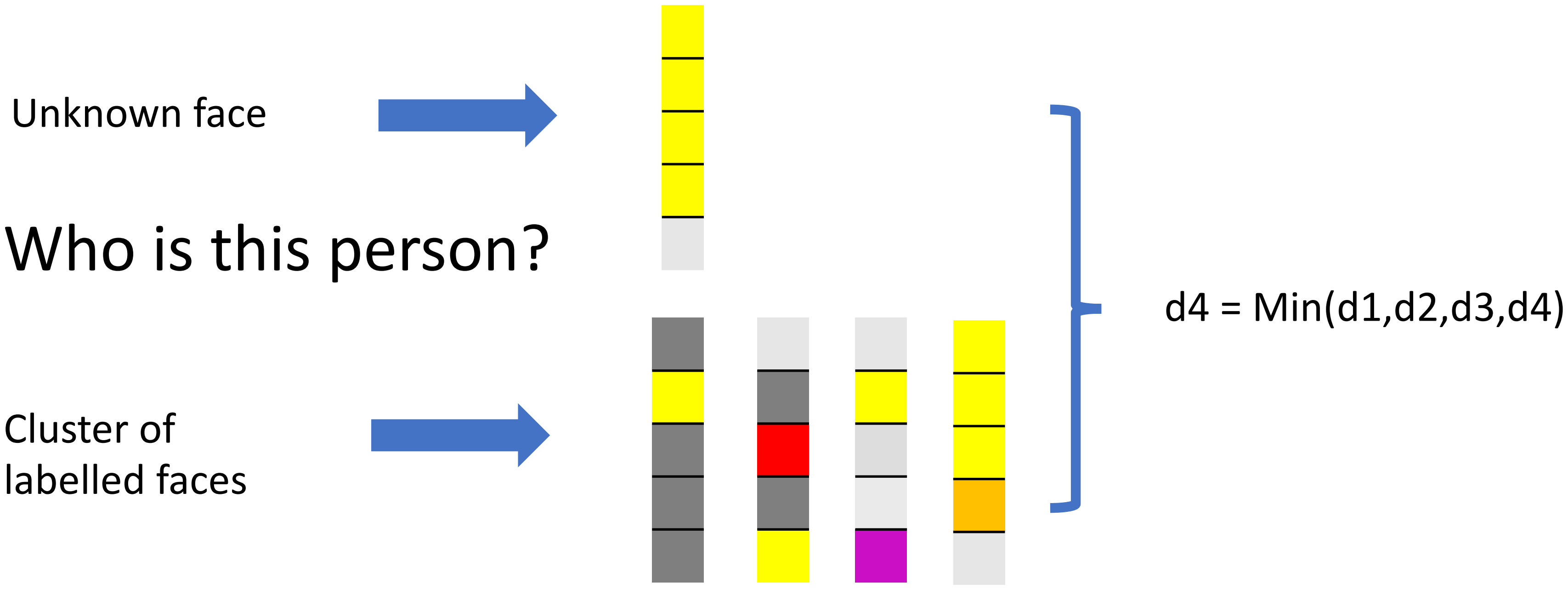
Train a deep neural network



Generate vectors to represent each person
of your gallery

- Generate a labelled cluster of embeddings for each celebrity
  - Build a dictionary of celebrities and select a set of images per celebrity
  - Generate your embeddings and label it
- Zero-shot learning approach
  - A new person can be added to your database without retraining the deep neural network

FACIAL RECOGNITION WITH EMBEDDINGS



Unknown face

Who is this person?

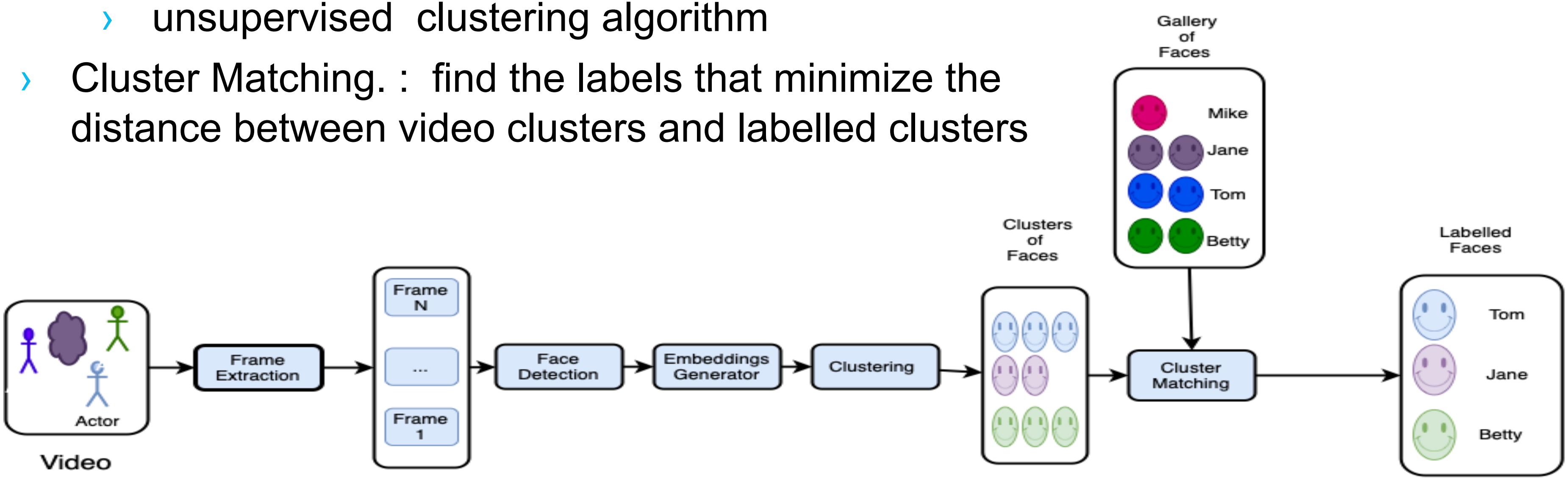
Cluster of labelled faces

d4 = Min(d1,d2,d3,d4)

- Recognise a face by comparing vectors with a mathematical distance and take the closest one

FACIAL RECOGNITION FOR VIDEO

- Generate unlabelled clusters from videos :
  - face detection - embeddings generation
  - unsupervised clustering algorithm
- Cluster Matching. : find the labels that minimize the distance between video clusters and labelled clusters

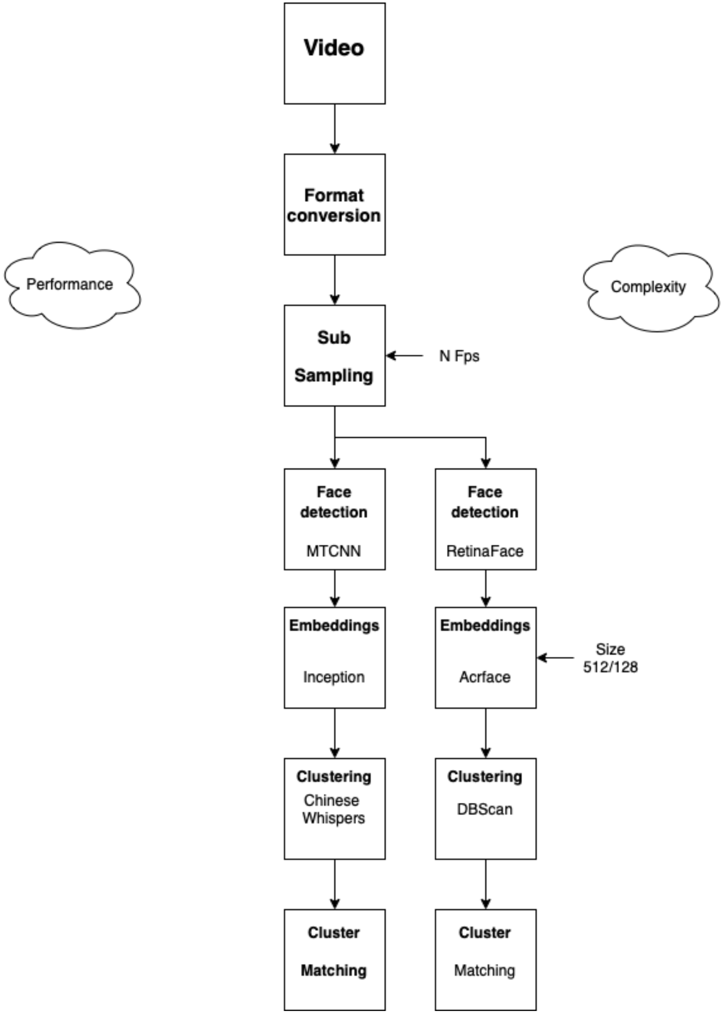


MODULAR PIPELINE

- The pipeline is generic and the processing blocks can be adapted and integrate open source models from research

- This is an efficient way to take advantage of the state of the art by adapting it to specific needs

- It is a way to evaluate different options depending on your use cases and constraints on
  - performances
  - complexity



Video

Format conversion

Sub Sampling

N Fps

Performance

Complexity

Face detection MTCNN

Face detection RetinaFace

Embeddings Inception

Embeddings Acrface

Size 512/128

Clustering Chinese Whispers

Clustering DBScan

Cluster Matching

Cluster Matching

SOME ADVANTAGES OF OPEN SOURCE

- You know exactly what the processing is and can extend it for your own needs
- Once you have the embeddings, you can use them for
  - Similarity search
  - Gender classification
  - Estimate age, emotion
  - Identify the most present persons in your archive(unlabelled)
  - Identify unlabelled clusters in your archive

WORKING GROUP DELIVERY

- Development
  - Open source modular pipeline
  - Cloud hosted application
    - easy to call like the other market place for evaluation only
- Dataset sharing
  - Data sharing strategy is ready (avoid copy right issue)
- Benchmarking of both Market Place and Open source solutions
- Report sharing
  - Users of the benchmarking framework will share the results
    - Scores
    - Use case description

JOIN THE TEAM !

Rai

RTS Radio Télévision Suisse

EBU

- The group is open to developers and users
  - One bi-weekly meeting

WORK IN PROGRESS

- Optimise the gallery considering the pipelines
  - cluster matching algorithms can be adapted to gallery properties and vice versa
- Specific metrics for video
  - define user-centric metrics
  - define bloc level metric
- Open sourcing of the dataset

QUESTIONS ?

CONTACT ME

Alexandre Rouxel

rouxel@ebu.ch