Fusion of Speech, Faces and Text for Person Identification in TV Broadcast

REPERE challenge

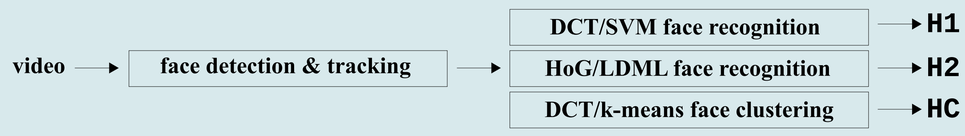
The REPERE challenge is an evaluation campaign in the field of people recognition in multimedia TV documents.

The main objective is to answer:

Who is speaking?
Who is seen?

Any modality can be used.

Who is seen?



FACE RECOGNITION	EGER
DCT/SVM approach (H1)	77.4 %
HoG/LDML approach (H2)	82.5 %
Oracle (50 identity models covering 34 % of test set)	50.8 %

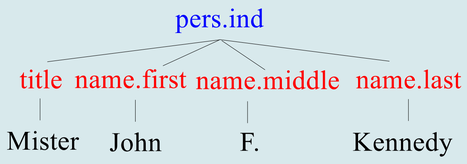
Whose name is written or spoken?

Named Entity Detection in Speech

1. Automatic speech transcription
2. Named entity detection
3. Person tags filtering

Video Optical Character Recognition

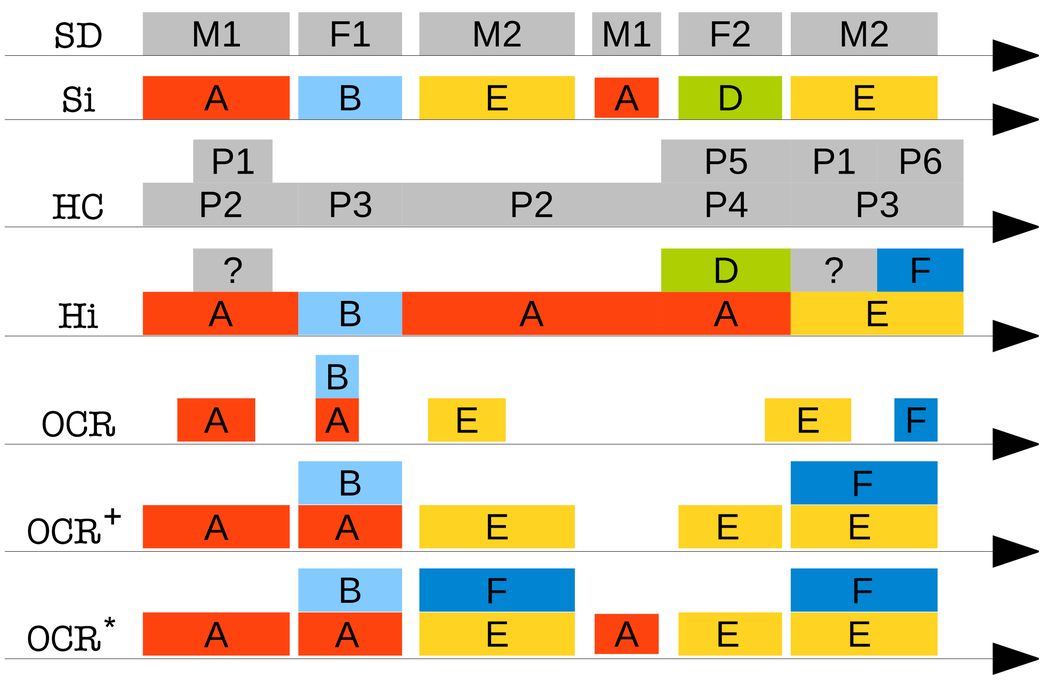
1. Overlaid text boxes detection
2. Automatic transcription (Tesseract)
3. Temporal filtering/smoothing
4. Named entity detection



Who is speaking?

SPEAKER DIARIZATION	DER
2-stages BIC/CLR approach (SD)	9.9 %

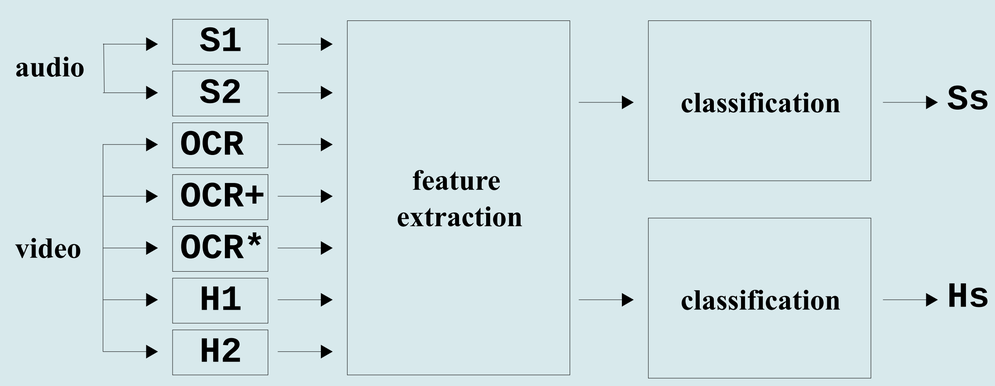
SPEAKER IDENTIFICATION	EGER
GSV/SVM approach (S1)	48.1 %
GMM/UBM approach (S2)	51.4 %
Oracle (57 identity models covering 49 % of test set)	33.8 %



Supervised person identification

Multiple classifiers are trained to answer the following question for each possible identity P during segment S

« is P speaking (or seen) for the duration of S? »



Feature extraction

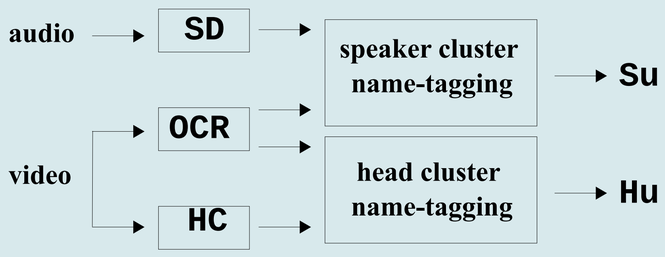
- Does the name of P appear in OCR? in OCR+? in OCR*?
- Duration of appearance of the name of P in OCR+, in OCR*.
- Duration of appearance of any name in OCR+, in OCR*.
- Their ratio.
- Speaker recognition scores for identity P provided by S1 and S2.
- Their difference to the best scores of any other identity.
- Is P the most likely identity according to S1 or S2?
- Do the gender of P and the detected gender of the speaker cluster match?

- Face recognition scores for identity P provided by H1 and H2.
- Is P the most likely identity according to H1 or H2?

Experimental results

CLASSIFIER	SPEAKER EGER	HEAD EGER
Naive Bayes	32.5 %	66.4 %
RBF Network	32.1 %	65.6 %
Random Tree	31.1 %	66.5 %
Random Forest	29.4 %	61.6 %
J48	28.2 %	63.1 %
AD Tree	27.8 %	62.3 %
NB Tree	27.0 %	64.7 %
Multilayer Perceptron	26.2 %	63.9 %
(Mono-modal) Oracle	33.8 %	50.8 %

Unsupervised person identification



Speaker diarization & face clustering group similar tracks into one cluster.
Each cluster is then tagged with a name obtained with the video OCR module.

Each person cluster (speaker or face) k is renamed after the name n with the largest co-occurrence duration Ckn. In case a cluster has no co-occurring name, its tag is set to Unknown:

$$\forall k \in \mathcal{K}, \quad \widehat{n_k} = \begin{cases} \underset{n \in \mathcal{N}}{\operatorname{argmax}}\, \mathtt{C}_{kn} & \text{if } \exists\, n \in \mathcal{N} \text{ such that } \mathtt{C}_{kn} > 0, \\ \mathtt{Unknown} & \text{otherwise.} \end{cases}$$

Experimental results

APPROACH	SPEAKER EGER	HEAD EGER
Automatic name-tagging	52.5 %	68.0 %
Oracle (OCR covers 60 % of test set)	41.7 %	32.5 %

Conclusion & Future work

For face recognition, **unsupervised multimodal** systems can be **as good as supervised monomodal** approaches.

For speaker identification, **visual modalities** (face recogntion and optical character recognition) bring **significant improvements**: every supervised multimodal approach beats the monomodal oracle.

The **whose name is spoken?** modality will be added to the game in the future.

Existing systems rely on **late fusion** approaches.
Earlier fusion techniques will be investigated.

Authors

Hervé Bredin
Johann Poignant
Makarand Tapaswi
Guillaume Fortier
Viet Bac Le
Thibault Napoleon
Hua Gao
Claude Barras
Sophie Rosset
Laurent Besacier
Jakob Verbeek
Georges Quénot
Frédéric Jurie
Hazim Kemal Ekenel

Supported by

OSEO / Quaero program
ANR / Qcompere project

Glossary

BIC > bayesian information criterion
CLR > cross-likelihood ratio
DCT > discrete cosine transform
DER > diarization error rate
EGER > estimated global error rate
GMM > gaussian mixture model
GSV > gaussian super vector
HC > head clustering
HoG > histogram of gradient
LDLM > logistic discriminant metric learning
OCR > optical character recognition
SD > speaker diarization
SVM > support vector machine
UBM > universal background model

Evaluation

REPERE corpus

7 different shows

27 hours raw videos (full video condition)

6 hours annotated videos (standard condition)

2195 annotated frames

Metric

$$\text{EGER} = \frac{\#\text{fa} + \#\text{miss} + \#\text{conf}}{\#\text{total}}$$

#total : number of person utterances to be detected

#conf : number of utterances wrongly identified

#miss : number of missed utterances

#fa : number of false alarms

DER is the fraction of time that is not attributed correctly to a speaker or to non-speech.

ECCV IFCVCR Workshop – October 2012 – Firenze, Italy