Fusion of speech, faces and text for person identification in TV broadcast

Hervé Bredin[1], Johann Poignant[2], Makarand Tapaswi[3], Guillaume Fortier[4], Viet Bac Le[5], Thibault Napoleon[6], Hua Gao[3], Claude Barras[1], Sophie Rosset[1], Laurent Besacier[2], Jakob Verbeek[4], Georges Quénot[2], Frédéric Jurie[6], and Hazim Kemal Ekenel[3]

[1] Univ Paris-Sud / CNRS-LIMSI UPR 3251, BP 133, F-91403 Orsay, France
[2] UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS-LIG UMR 5217, F-38041 Grenoble, France
[3] Karlsruher Institut fur Technologie, Karlsruhe, Germany
[4] INRIA Rhone-Alpes, 655 Avenue de lEurope, F-38330 Montbonnot, France
[5] Vocapia Research, 3 rue Jean Rostand, Parc Orsay Université, F-91400 Orsay, France
[6] Université de Caen / GREYC UMR 6072, F-14050 Caen Cedex, France

Abstract. The REPERE challenge is a project aiming at the evaluation of systems for supervised and unsupervised multimodal recognition of people in TV broadcast. In this paper, we describe, evaluate and discuss QCOMPERE consortium submissions to the 2012 REPERE evaluation campaign dry-run. Speaker identification (and face recognition) can be greatly improved when combined with name detection through video optical character recognition. Moreover, we show that unsupervised multimodal person recognition systems can achieve performance nearly as good as supervised monomodal ones (with several hundreds of identity models).

1 Introduction

Over the years, a growing amount of multimedia data has been produced and made available, fostering the need for automatic processing systems allowing efficient search into multimedia archives.

Person recognition is one of the main keys for structuring a video document. Face recognition in images or videos [1] and speaker identification in audio [2] are already very active research fields in this domain.

As illustrated in Figure 1, the REPERE challenge[1] aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. It takes the form of an annual evaluation campaign and debriefing workshop.

[1] `http://www.defi-repere.fr`

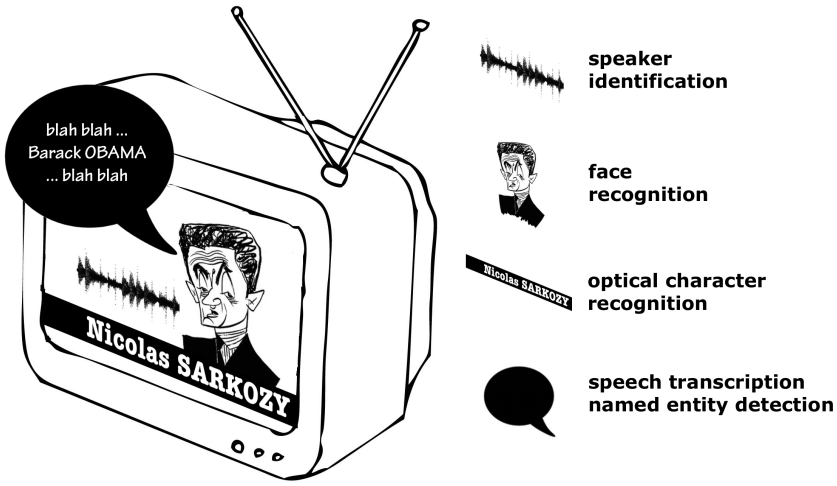


Fig. 1. One identity, four modalities

In this paper we describe QCompere consortium submissions to the 2012 Repere evaluation campaign dry-run. The Repere corpus and evaluation protocol is described in Section 2. Mono-modal person recognition components are introduced in Section 3, while Section 4 is dedicated to their supervised and unsupervised combination. Finally, results are reported and discussed in Section 5.

2 The REPERE challenge

The Repere evaluation campaign dry-run was organized in January 2012. We first describe the corresponding Repere corpus which is meant to be extended throughout the duration of the project, with a final total of 60 hours of annotated videos. Then, the main tasks and the corresponding evaluation metric are quickly summarized.

2.1 Corpus

The 2012 Repere corpus contains a total of 6 hours of annotated videos recorded from 2 French TV channels (BFMTV and LCP) and 7 different TV shows (TV news and talk shows). It is divided into development and test sets (3 hours each). Annotations are provided for four main modalities:

Speaker Each speech turn is described with its start and end timestamps and the normalized speaker identity (e.g. `Nicolas_SARKOZY`).

Head Each head track is described with its appearance and disappearance timestamps and the associated normalized identity.

Written Every overlaid text box is transcribed with its appearance and disappearance timestamps and written person names are tagged with the normalized identity.

Spoken Each speech turn is transcribed and spoken person names are tagged with the normalized identity (e.g. `Barack_OBAMA`).

People whose identity cannot be infered from the rest of the video (and who are not famous people) are tagged as such in a consistent way (e.g. `Unknown_1` $\neq$ `Unknown_2`). Moreover, a set $\mathcal{F}$ of video frames was sampled (one every 10 seconds on average) and annotated more precisely with the position of each face and overlaid text bounding boxes.

2.2 Main tasks

The main objective of the REPERE challenge is to answer the two following questions at any instant of the video:

"who is speaking?" *"who is seen?"*

While the former question can be seen as the usual speaker diarization and tracking problem, the latter cannot be reduced to basic face recognition. As a matter of fact, a person who is seen from the back must also be recognized if a human could infer his/her identity from the context.

In the context of the REPERE challenge, we distinguish mono- and multi-modal conditions as well as supervised and unsupervised person identification.

In the **mono-modal** case, only the raw acoustic signal can be used to detect and identify speakers (using its automatic transcription is not allowed). Similarly, visual person recognition cannot rely on name detection in overlaid text, for instance. On the other hand, in the **multi-modal** case, any of the four modalities (speaker, head, written or spoken) can be used to answer both questions.

In the **supervised** case, any previously trained identity model can be used to recognize a person. However, these models are strictly forbidden in the **unsupervised** conditions: person names can only be inferred from the **written** and **spoken** modalities. Therefore, any unsupervised method is – by design – multi-modal.

2.3 Estimated Global Error Rate

Though the whole test set is processed, evaluation is only performed on the annotated frames $\mathcal{F}$. For each frame f, let us denote $\#\text{total}(f)$ the number of persons in the reference. The hypothesis proposed by an automatic system can make three types of errors:

False Alarms (#fa) when it contains more persons than there actually are in the reference.

Missed Detections (#miss) when it contains less persons than there actually are in the reference.

Confusion (#conf) when the detected identity is wrong. For evaluation purposes, and because unknown people cannot – by definition – be recognized in any way, they are excluded from the scoring.

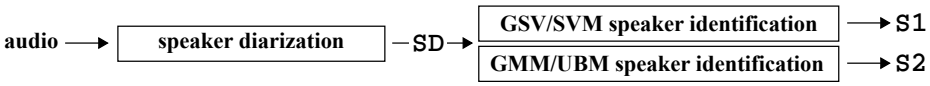
The Estimated Global Error Rate (EGER) is defined by:

$$\text{EGER} = \frac{\sum_{f \in \mathcal{F}} \#\text{conf}(f) + \#\text{fa}(f) + \#\text{miss}(f)}{\sum_{f \in \mathcal{F}} \#\text{total}(f)}$$

3 Monomodal components

3.1 Who is speaking?

Speaker diarization is the process of partitioning the audio stream into homogeneous clusters without prior knowledge of the speaker voices. Our system `SD` relies on two steps: agglomerative clustering based on the BIC criterion to provide pure clusters followed by a second clustering stage using more complex models and cross-likelihood ratio (CLR) as distance between clusters [3].



Unsupervised speaker diarization is followed by a cluster-wise **speaker identification**. We implemented two systems [4]. The GSV-SVM system `S1` uses the supervector made of the concatenation of the UBM-adapted GMM means to train one Support Vector Machine classifier per speaker. Our baseline system `S2` follows the standard GMM-UBM paradigm. For both systems, each cluster is scored against all gender-matching speaker models, and the best scoring model is chosen if its score is higher than the decision threshold.

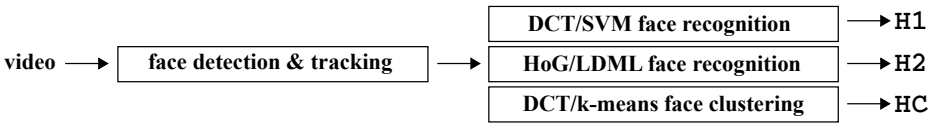
Three data sources were used for training 535 different speaker models in our experiments: the REPERE development set, the ETAPE[2] evaluation data and French radio data annotated into politicians speaking times.

3.2 Who is seen?

Figure below summarizes how our two submissions to the monomodal face recognition REPERE task are built and differ from each other.

Face detection and tracking is achieved using a detector-based face tracker in a particle-filter framework [5]. Face tracks are first initialized by scanning the first frame of every shot, and the subsequent fifth frame, using frontal, half-profile

[2] `http://www.afcp-parole.org/etape.html`



and profile face detectors – making face detection independent of the initial pose. Tracking is performed in an online manner, using the state of the previous frame to infer the location and head pose of the faces in the current frame. Head pose is explicitly incorporated in the continuous tracked state (alongside face position and size) as the head yaw-angle. A total of 11 yaw-angle-dependent face detectors are combined to score each particle of a track.

Features used in `H1` are based on a local appearance-based approach [6]. Each face is normalized to a canonical pose and size and then split into 8×8 blocks. The top five Discrete Cosine Transform (DCT) coefficients are stored for each block. For recognition, one-vs-all second order polynomial kernel SVMs are trained for each person in the development set. Normalized classification scores are then accumulated over each track to obtain face identity scores in the range from 0 to 1.

In `H2` approach, nine facial points located around the eyes, nose and mouth are automatically detected [7]. Each of them is described by a 490-dimensional HOG descriptor [8], yielding a 4410-dimensional feature vector per face. Logistic discriminant metric learning [9] is then used to project this vector into a 200-dimensional feature vector space where the ℓ_2 distance is combined with a nearest neighbor classifier for face recognition.

Alongside these supervised face recognition approaches, a **face clustering** system `HC` is also implemented for later use in multimodal unsupervised face recognition. It uses DCT-based descriptors from `H1`. Seven representative face samples are extracted from each face track using k-means algorithm. Then, hierarchical agglomerative clustering is performed until the elbow point of the distortion curve is reached – in order to get pure clusters.

3.3 Whose name is written?

As illustrated in Figure 1, voice and appearance are not the only sources of information available to identify a person on TV. Hence, guests or reporters are sometimes introduced to the viewer using overlaid text containing their name.

A video `OCR` system was designed to automatically extract this information, which is especially useful in an unsupervised framework [10]. Overlaid text boxes are first detected using a coarse-to-fine approach with temporal tracking. Then, `Google Tesseract` open-source OCR system provides one transcription for every corresponding frames. They are finally combined to produce one single better transcription for each text box.

Using the shows from the development set and a list of famous people names extracted from `Wikipedia`, we were also able to extract the positions most likely

used by each type of show to introduce a person. Only the detected names at these positions are used in later fusion.

3.4 Whose name is pronounced?

Person names are also often pronounced by the anchor or other guests – providing a fourth source of information to identify them. Though we could not integrate this information in the final system in time for the first campaign, we did develop a system aiming at extracting these names.

First, a state-of-the-art speech-to-text system (STT) based on statistical modeling techniques [11] is used to automatically obtain the speech transcription. Then, a named entity recognition system `NE` [12] automatically detects several kind of named entities in the STT output, including the `<pers>` entity that is of interest in this work. It has a tree structure that is summarized in Figure 2.

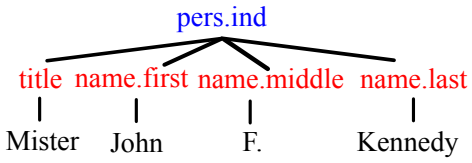


Fig. 2. Structured person entity

For precision concerns, we only detect `<pers>` entities for which both a first name and a last name are available (regardless of their order) – thus leaving room for great future improvement.

4 Multimodal fusion

Once all monomodal components have been run on a video, their outputs can be combined to improve the overall person recognition performance. Figure 3 draws up their list, along with two slightly modified versions of `OCR`: extended to the whole speech turns ($\texttt{OCR}^+$) or speaker diarization clusters ($\texttt{OCR}^*$).

4.1 Supervised person recognition

Since each modality relies on its own temporal segmentation, the first step consists in aligning the various timelines onto the finest common segmentation. The final decision is taken at this segmentation granularity. For each resulting segment $\mathcal{S}$, a list of possible identities is built based on the output of all modalities. For each hypothesis identity $\mathcal{P}$, a set of features is extracted:

– Does the name of $\mathcal{P}$ appear in `OCR`? in $\texttt{OCR}^+$? in $\texttt{OCR}^*$?

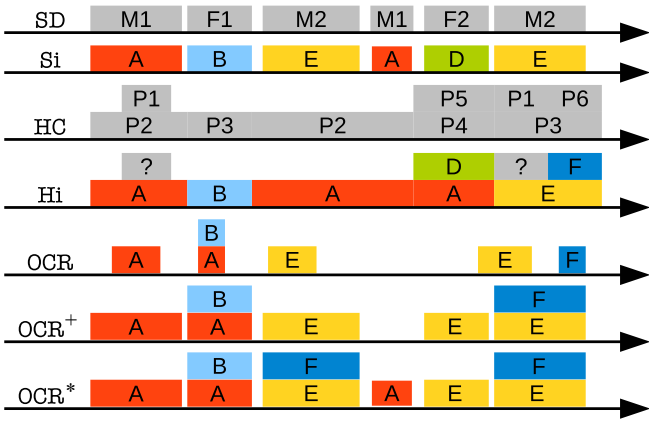


Fig. 3. Several annotation timelines

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

Two additional features were added for face recognition:

- Face recognition scores for identity $\mathcal{P}$ provided by `H1` and `H2`.
- Is $\mathcal{P}$ the most likely identity according to `H1` or `H2`?

Based on these features, we trained several classifiers using Weka[3] to answer to the following question:

"is $\mathcal{P}$ speaking (or seen) for the duration of $\mathcal{S}$?"

Since these features can be either boolean or (unbounded) float, several classifiers insensitive to numerical types were used. As shown in Table 1, the best classifier for each task was selected using 2-fold cross-validation on the development set.

The best performance was obtained using multi-layer perceptron for `speaker` identification and random forest for its `face` counterpart. The identity with the highest score is selected for the `speaker` task and the N-best hypotheses for the `head` task – where N is the number of detected heads.

4.2 Unsupervised person recognition

As stated in Section 2, the Repere challenge also includes an unsupervised track, for which no previously trained identity model can be used to perform

[3] `http://www.cs.waikato.ac.nz/ml/weka`

Classifier	Speaker	Head
NaiveBayes	32.49	66.42
RBFNetwork	32.12	65.61
RandomTree	31.09	66.55
RandomForest	29.41	**61.63**

Classifier	Speaker	Head
J48	28.20	63.12
ADTree	27.82	62.31
NBTree	26.98	64.73
MultilayerPerceptron	**26.24**	63.86

Table 1. Estimated Global Error Rate on development set

person recognition. Hence, none of `S1`, `S2`, `H1` and `H2` systems can be used for people identification in these conditions, as they all rely on trained identity models. Both our unsupervised person identification systems `Su` (for speaker) and `Hu` (for head) rely on a similar 3-steps approach that can be schematized as follows:

$$\texttt{Su} = \texttt{SD} \otimes \texttt{OCR} \qquad\qquad \texttt{Hu} = \texttt{HC} \otimes \texttt{OCR}$$

First, speaker diarization (`SD`, introduced in Section 3.1) or face clustering (`HC`, from Section 3.2) labels every occurrence of the same person with a unique anonymous tag (e.g. `head#1` or `speaker#2`). Let us denote $\mathcal{K} = \{k_1, \ldots, k_L\}$ the set of L resulting (speaker or face) clusters. Then, `OCR` (from Section 3.3) provides a short list of M possible names $\mathcal{N} = \{n_1, \ldots, n_M\}$. Finally, each person cluster (speaker or face) k is renamed after the name $\widehat{n}$ with the largest co-occurrence duration $\texttt{C}_{kn}$. In case a cluster has no co-occurring name, its tag is set to `Unknown`:

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

Note that this approach can lead to the propagation of one name n to multiple clusters. It does not blindly trust the speaker diarization or face clustering systems. In particular, it assumes that they may produce over-segmented clusters (for instance, split speech turns from one speaker into two or more clusters) that can be merged afterwards.

5 Results

Table 2 summarizes the performance of both mono- and multi-modal approaches, as well as of the unsupervised ones.

As expected, `S1` (based on GSV-SVM) brings significant improvement (-3.3% EGER) over the simpler system `S2` (based on GMM/UBM) for mono-modal speaker recognition. Why mono-modal **speaker** approaches (EGER $\approx$ 50%) work much better than their **head** counterpart (EGER $\approx$ 80%) can be explained by looking at Table 3. Indeed, only one third of known persons in test set actually had a previously trained head model (vs. 49% for speaker recognition). Even an *oracle* capable of correctly identifying any previously modeled person

Conditions	Speaker	Head
Supervised & monomodal	`S1` — 48.1%	`H1` — 77.4%
	`S2` — 51.4%	`H2` — 82.5%
Supervised & multimodal	`Ss` — 25.8%	`Hs` — 61.5%
Unsupervised	`Su` — 52.2%	`Hu` — 68.0%

Table 2. Estimated Global Error Rate

	# persons	# modeled	Oracle EGER
Speaker	116	57 (49%)	33.8%
Head	145	50 (34%)	50.8%

Table 3. Number of persons with trained identity model & best possible performance for a monomodal supervised person recognition oracle.

(from the development set) could not reach better performance than 50% for head-based people recognition.

One of the most interesting contribution of this paper is the improvement brought by multi-modal fusion of the **written** modality with **speaker** and **head** ones: around 20% absolute EGER decrease for both of them (`Ss` vs. `S1`, and `Hs` vs. `H1`).

Finally, the other major result highlighted in this paper is that multi-modal unsupervised person recognition can achieve performance as good as monomodal supervised approaches (`Su` vs. `S1` and `Hu` vs. `H1`). Yet, Table 4 shows that one can expect much better performance from `Su` and `Hu`. An *oracle* capable of giving the correct name to a person – as long as his/her name appears at least once during the show – can indeed reach around 42% (respectively 32%) EGER, when relying on perfect **speaker** diarization (resp. **head** clustering) and perfect **written** name detection.

	# persons	# written	Oracle EGER
Speaker	116	74 (64%)	41.7%
Head	145	82 (56%)	32.5%

Table 4. Is unsupervised recognition even possible? Number of persons whose name is written at least once & oracle performance.

6 Conclusion

In this paper, we described, evaluated and discussed QCompere consortium submissions to the 2012 Repere evaluation campaign dry-run. We showed that

speaker identification (and face recognition) can be greatly improved when combined with name detection through video optical character recognition; and that unsupervised multimodal person recognition systems can achieve performance nearly as good as supervised monomodal ones.

Yet, there is plenty of room for improvement – in particular for our face recognition algorithms that showed their limits on this particular type of videos. Moreover, the **spoken** modality has not yet been added to the game. It might indeed be very useful, especially in the unsupervised conditions: talk-show anchors, for instance, tend to introduce their guest by pronouncing their name. These are issues we will address for next year Repere evaluation campaign.

7 Acknowledgment

This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

References

1. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: a Literature Survey. ACM Comput. Surv. **35**(4) (December 2003) 399–458
2. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A.: A Tutorial on Text-Independent Speaker Verification. EURASIP J. Appl. Signal Process. **2004** (January 2004) 430–451
3. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Multi-Stage Speaker Diarization of Broadcast News. IEEE Transactions on Audio, Speech and Language Processing **14**(5) (September 2006) 1505–1512
4. Le, V.B., Barras, C., Ferràs, M.: On the use of GSV-SVM for Speaker Diarization and Tracking. In: Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, Brno, Czech Republic (June 2010) 146–150
5. Baeuml, M., Bernardin, K., Fischer, M., Ekenel, H., Stiefelhagen, R.: Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In: Advanced Video and Signal-based Surveillance. (2010)
6. Ekenel, H., Stiefelhagen, R.: Analysis of Local Appearance Based Face Recognition: Effects of Feature Selection and Feature Normalization. In: CVPR Biometrics Workshop. (2006)
7. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – Automatic Naming of Characters in TV video. In: British Machine Vision Conference. (2006)
8. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: International Conference on Computer Vision & Pattern Recognition. (2005) 886–893
9. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Face Recognition from Caption-based Supervision. International Journal of Computer Vision **96**(1) (2012) 64–82
10. Poignant, J., Besacier, L., Quénot, G., Thollard, F.: From Text Detection in Videos to Person Identification. In: IEEE ICME, Melbourne, Australia (2012)

11. Gauvain, J., Lamel, L., Adda, G.: The LIMSI Broadcast News Transcription System. Speech Communication **37**(1-2) (2002) 89–109
12. Dinarelli, M., Rosset, S.: Models Cascade for Tree-Structured Named Entity Detection. In: Proceedings of International Joint Conference of Natural Language Processing (IJCNLP), Chiang Mai, Thailand (November 2011)