Multimodal Person Discovery in Broadcast TV at MediaEval 2015

Johann Poignant, Hervé Bredin, Claude Barras
LIMSI - CNRS - Rue John Von Neumann, Orsay, France.
firstname.lastname@limsi.fr

ABSTRACT

We describe the “Multimodal Person Discovery in Broadcast TV” task of MediaEval 2015 benchmarking initiative. Participants were asked to return the names of people who can be both seen as well as heard in every shot of a collection of videos. The list of people was not known *a priori* and their names had to be discovered in an unsupervised way from media content using text overlay or speech transcripts. The task was evaluated using information retrieval metrics, based on *a posteriori* collaborative annotation of the test corpus.

1. MOTIVATION

TV archives maintained by national institutions such as the French INA, the Netherlands Institute for Sound & Vision, or the British Broadcasting Corporation are rapidly growing in size. The need for applications that make these archives searchable has led researchers to devote concerted effort to developing technologies that create indexes.

Indexes that represent the location and identity of people in the archive are indispensable for searching archives. Human nature leads people to be very interested in other people. However, when the content is created or broadcast, it is not always possible to predict which people will be the most important to find in the future. For this reason, it is not possible to assume that biometric models will always be available at indexing time. For some people, such a model may not be available in advance, simply because they are not (yet) famous. In such cases, it is also possible that archivists annotating content by hand do not even know the name of the person. The goal of this task is to address the challenge of indexing people in the archive, under real-world conditions (*i.e.* when there is no pre-set list of people to index).

Canseco et al. [8, 9] pioneered approaches relying on pronounced names instead of biometric models for speaker identification [13, 19, 22, 30]. However, due to relatively high speech transcription and named entity detection errors, all these audio-only approaches did not achieve good enough identification performance. Similarly, for face recognition, initial visual-only approaches based on overlaid title box transcriptions were very dependent on the quality of overlaid name transcription [18, 29, 32, 33].

Started in 2011, the REPERE challenge aimed at supporting research on multimodal person recognition [3, 20] to overcome the limitations of monomodal approaches. Its main goal was to answer the two questions *“who speaks when?”* and *“who appears when?”* using any available source of information (including pre-existing biometric models and person names extracted from text overlay and speech transcripts). To assess the technology progress, annual evaluations were organized in 2012, 2013 and 2014. Thanks to this challenge and the associated multimodal corpus [16], significant progress was achieved in either supervised or unsupervised mulitmodal person recognition [1, 2, 4, 5, 6, 7, 14, 15, 23, 25, 26, 27, 28]. The REPERE challenge came to an end in 2014 and this task can be seen as a follow-up campaign with a strong focus on unsupervised person recognition.

2. DEFINITION OF THE TASK

Participants were provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot $s \in \mathbb{S}$ had to be automatically tagged with the names of people both speaking and appearing at the same time during the shot: this tagging algorithm is denoted by $\mathcal{L} : \mathbb{S} \mapsto \mathcal{P}(\mathcal{N})$ in the rest of the paper. The main novelty of the task is that the list of persons was not provided *a priori*, and person biometric models (neither voice nor face) could not be trained on external data. The only way to identify a person was by finding their name $n \in \mathcal{N}$ in the audio (*e.g.* using speech transcription – ASR) or visual (*e.g.* using optical character recognition – OCR) streams and associating them to the correct person. This made the task completely unsupervised (*i.e.* using algorithms not relying on pre-existing labels or biometric models).

Because person names were detected and transcribed automatically, they could contain transcription errors to a certain extent (more on that later in Section 5). In the following, we denote by $\mathbb{N}$ the set of all possible person names in the universe, correctly formatted as `firstname_lastname` – while $\mathcal{N}$ is the set of hypothesized names.

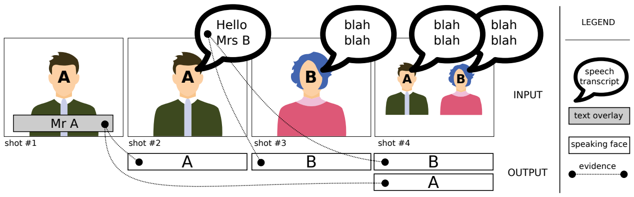

Figure 1: For each shot, participants had to return the names of every speaking face. Each name had to be backed up by an evidence.

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop Sept. 14-15, 2015, Wurzen, Germany

To ensure that participants followed this strict *"no biometric supervision"* constraint, each hypothesized name $n \in \mathcal{N}$ had to be backed up by a carefully selected and unique shot prooving that the person actually holds this name n: we call this an evidence and denote it by $\mathcal{E} : \mathcal{N} \mapsto \mathbb{S}$. In real-world conditions, this evidence would help a human annotator double-check the automatically-generated index, even for people they did not know beforehand.

Two types of evidence were allowed: an *image* evidence is a shot during which a person is visible, and their name is written on screen; an *audio* evidence is a shot during which a person is visible, and their name is pronounced at least once during a [shot start time $- 5s$, shot end time $+ 5s$] neighborhood. For instance, in Figure 1, shot #1 is an *image* evidence for Mr A (because his name and his face are visible simultaneously on screen) while shot #3 is an *audio* evidence for Mrs B (because her name is pronounced less than 5 seconds before or after her face is visble on screen).

3. DATASETS

The REPERE corpus – distributed by ELDA – served as development set. It is composed of various TV shows (around news, politics and people) from two French TV channels, for a total of 137 hours. A subset of 50 hours is manually annotated. Audio annotations are dense and provide speech transcripts and identity-labeled speech turns. Video annotations are sparse (one image every 10 seconds) and provide overlaid text transcripts and identity-labeled face segmentation. Both speech and overlaid text transcripts are tagged with named entities. The test set – distributed by INA – contains 106 hours of video, corresponding to 172 editions of evening broadcast news *"Le 20 heures"* of French public channel *"France 2"*, from January 1st 2007 to June 30st 2007.

As the test set came completely free of any annotation, it was annotated *a posteriori* based on participants' submissions. In the following, task groundtruths are denoted by function $\mathbb{L} : \mathbb{S} \mapsto \mathcal{P}(\mathbb{N})$ that maps each shot s to the set of names of every speaking face it contains, and function $\mathbb{E} : \mathbb{S} \mapsto \mathcal{P}(\mathbb{N})$ that maps each shot s to the set of person names for which it actually is an evidence.

4. BASELINE AND METADATA

This task targeted researchers from several communities including multimedia, computer vision, speech and natural language processing. Though the task was multimodal by design and necessitated expertise in various domains, the technological barriers to entry was lowered by the provision of a baseline system described in Figure 2 and available as open-source software[1]. For instance, a researcher from the speech processing community could focus its research efforts on improving speaker diarization and automatic speech transcription, while still being able to rely on provided face detection and tracking results to participate to the task.

The audio stream was segmented into speech turns, while faces were detected and tracked in the visual stream. Speech turns (resp. face tracks) were then compared and clustered based on MFCC and the Bayesian Information Criterion [10] (resp. HOG [11] and Logistic Discriminant Metric Learning [17] on facial landmarks [31]). The approach proposed in [27] was also used to compute a probabilistic mapping between co-occuring faces and speech turns. Written (resp. pronounced) person names were automatically extracted from the visual stream (resp. the audio stream) using open source LOOV Optical Character Recognition [24] (resp. Automatic Speech Recognition [21, 12]) followed by Named Entity detection (NE). The fusion module was a two-steps algorithm: propagation of written names onto speaker clusters [26] followed by propagation of speaker names onto co-occurring speaking faces.

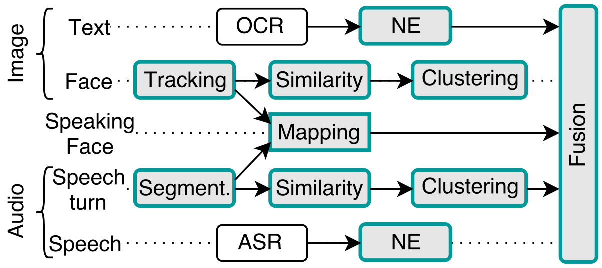

Figure 2: Multimodal baseline pipeline. Output of greyed out modules is provided to the participants.

5. EVALUATION METRIC

This information retrieval task was evaluated using a variant of Mean Average Precision (MAP), that took the quality of evidences into account. For each query $q \in \mathbb{Q} \subset \mathbb{N}$ (`firstname_lastname`), the hypothesized person name n_q with the highest Levenshtein ratio ρ to the query q is selected ($\rho : \mathbb{N} \times \mathcal{N} \mapsto [0, 1]$) – allowing approximate name transcription:

$$n_q = \underset{n \in \mathcal{N}}{\arg\max}\, \rho\,(q, n) \;\text{ and }\; \rho_q = \rho\,(q, n_q)$$

Average precision AP(q) is then computed classically based on relevant and returned shots:

$$\begin{aligned} \text{relevant}(q) &= \{s \in \mathbb{S} \mid q \in \mathbb{L}(s)\} \\ \text{returned}(q) &= \{s \in \mathbb{S} \mid n_q \in \mathcal{L}(s)\}_{\substack{\text{sorted by} \\ \text{confidence}}} \end{aligned}$$

Proposed evidence is Correct if name n_q is close enough to the query q and if shot $\mathcal{E}(n_q)$ actually is an evidence for q:

$$C(q) = \begin{cases} 1 & \text{if } \rho_q > 0.95 \text{ and } q \in \mathbb{E}(\mathcal{E}(n_q)) \\ 0 & \text{otherwise} \end{cases}$$

To ensure participants do provide correct evidences for every hypothesized name $n \in \mathcal{N}$, standard MAP is altered into EwMAP (Evidence-weighted Mean Average Precision), the official metric for the task:

$$\text{EwMAP} = \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} C(q) \cdot \text{AP}(q)$$

Acknowledgment. This work was supported by the French National Agency for Research under grant ANR-12-CHRI-0006-01. The open source CAMOMILE collaborative annotation platform[2] was used extensively throughout the progress of the task: from the run submission script to the automated leaderboard, including *a posteriori* collaborative annotation of the test corpus. We thank ELDA and INA for supporting the task by distributing development and test datasets.

[1] `http://github.com/MediaEvalPersonDiscoveryTask`

[2] `http://github.com/camomile-project`

6. REFERENCES

[1] F. Bechet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, G. Linarès, J. Martinet, G. Senay, and P. Tirilly. Multimodal Understanding for Person Recognition in Video Broadcasts. In *INTERSPEECH*, 2014.

[2] M. Bendris, B. Favre, D. Charlet, G. Damnati, R. Auguste, J. Martinet, and G. Senay. Unsupervised Face Identification in TV Content using Audio-Visual Sources. In *CBMI*, 2013.

[3] G. Bernard, O. Galibert, and J. Kahn. The First Official REPERE Evaluation. In *SLAM-INTERSPEECH*, 2013.

[4] H. Bredin, A. Laurent, A. Sarkar, V.-B. Le, S. Rosset, and C. Barras. Person Instance Graphs for Named Speaker Identification in TV Broadcast. In *Odyssey*, 2014.

[5] H. Bredin and J. Poignant. Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. In *INTERSPEECH*, 2013.

[6] H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V.-B. Le, A. Sarkar, C. Barras, S. Rosset, A. Roy, Q. Yang, H. Gao, A. Mignon, J. Verbeek, L. Besacier, G. Quénot, H. K. Ekenel, and R. Stiefelhagen. QCompere at REPERE 2013. In *SLAM-INTERSPEECH*, 2013.

[7] H. Bredin, A. Roy, V.-B. Le, and C. Barras. Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identification in TV broadcast. In *IJMIR*, 2014.

[8] L. Canseco, L. Lamel, and J.-L. Gauvain. A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In *ASRU*, 2005.

[9] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain. Speaker diarization from speech transcripts. In *INTERSPEECH*, 2004.

[10] S. Chen and P. Gopalakrishnan. Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion. In *DARPA Broadcast News Trans. and Under. Workshop*, 1998.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[12] M. Dinarelli and S. Rosset. Models Cascade for Tree-Structured Named Entity Detection. In *IJCNLP*, 2011.

[13] Y. Estève, S. Meignier, P. Deléglise, and J. Mauclair. Extracting true speaker identities from transcriptions. In *INTERSPEECH*, 2007.

[14] B. Favre, G. Damnati, F. Béchet, M. Bendris, D. Charlet, R. Auguste, S. Ayache, B. Bigot, A. Delteil, R. Dufour, C. Fredouille, G. Linares, J. Martinet, G. Senay, and P. Tirilly. PERCOLI: a person identification system for the 2013 REPERE challenge. In *SLAM-INTERSPEECH*, 2013.

[15] P. Gay, G. Dupuy, C. Lailler, J.-M. Odobez, S. Meignier, and P. Deléglise. Comparison of Two Methods for Unsupervised Person Identification in TV Shows. In *CBMI*, 2014.

[16] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The REPERE Corpus : a Multimodal Corpus for Person Recognition. In *LREC*, 2012.

[17] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *IJCV*, 96(1), 2012.

[18] R. Houghton. Named Faces: Putting Names to Faces. *IEEE Intelligent Systems*, 14, 1999.

[19] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Estève, and C. Jacquin. Automatic named identification of speakers using diarization and ASR systems. In *ICASSP*, 2009.

[20] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P.Joly. A presentation of the REPERE challenge. In *CBMI*, 2012.

[21] L. Lamel, S. Courcinous, J. Despres, J. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V.-B. Le, H. Ney, M. Nussbaum-Thom, I. Oparin, T. Schlippe, R. Schlëter, T. Schultz, T. F. da Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. Vu, A. Waibel, and C. Woehrling. Speech Recognition for Machine Translation in Quaero. In *IWSLT*, 2011.

[22] J. Mauclair, S. Meignier, and Y. Estève. Speaker diarization: about whom the speaker is talking? In *Odyssey*, 2006.

[23] J. Poignant, L. Besacier, and G. Quénot. Unsupervised Speaker Identification in TV Broadcast Based on Written Names. *IEEE/ACM ASLP*, 23(1), 2015.

[24] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *ICME*, 2012.

[25] J. Poignant, H. Bredin, L. Besacier, G. Quénot, and C. Barras. Towards a better integration of written names for unsupervised speakers identification in videos. In *SLAM-INTERSPEECH*, 2013.

[26] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *INTERSPEECH*, 2012.

[27] J. Poignant, G. Fortier, L. Besacier, and G. Quénot. Naming multi-modal clusters to identify persons in TV broadcast. *MTAP*, 2015.

[28] M. Rouvier, B. Favre, M. Bendris, D. Charlet, and G. Damnati. Scene understanding for identifying persons in TV shows: beyond face authentication. In *CBMI*, 2014.

[29] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and Detecting Faces in News Videos. *IEEE Multimedia*, 6, 1999.

[30] S. E. Tranter. WHO REALLY SPOKE WHEN? FINDING SPEAKER TURNS AND IDENTITIES IN BROADCAST NEWS AUDIO. In *ICASSP*, 2006.

[31] M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP*, volume 1, 2012.

[32] J. Yang and A. G. Hauptmann. Naming every individual in news video monologues. In *ACM Multimedia*, 2004.

[33] J. Yang, R. Yan, and A. G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *ACM Multimedia*, 2005.