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 are 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 is not known a priori and their names must be discovered in an unsupervised way from media content using text overlay or speech transcripts. The task is 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 multimodal 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 are provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot \( s \in S \) must 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 \( L : S \rightarrow P(\mathcal{N}) \) in the rest of the paper. The main novelty of the task is that the list of persons is not provided a priori, and person biometric models (neither voice nor face) may not be trained on external data. The only way to identify a person is 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 makes the task completely unsupervised (i.e. using algorithms not relying on pre-existing labels or biometric models).

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

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

To ensure that participants follow this strict “no biomet-
ric supervision" constraint, each hypothesized name $n \in N$ must be backed up by a carefully selected and unique shot proving that the person actually holds this name $n$: we call this an evidence and denote it by $E : N \rightarrow S$. 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 are 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 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 visible on screen).

3. DATASETS
The REPERE corpus – distributed by ELDA – serves 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 30th 2007.

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

4. BASELINE AND METADATA
This task targets researchers from several communities including multimedia, computer vision, speech and natural language processing. Though the task is multimodal by design and necessitates expertise in various domains, the technological barriers to entry is lowered by the provision of a baseline system described in Figure 2 and available as open-source software¹. 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 is segmented into speech turns, while faces are detected and tracked in the visual stream. Speech turns (resp. face tracks) are 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] is also used to compute a probabilistic mapping between co-occurring faces and speech turns. Written (resp. pronounced) person names are 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 is a two-steps algorithm: propagation of written names onto speaker clusters [26] followed by propagation of speaker names onto co-occurring speaking faces.

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

$$n_q = \arg \max_{n \in N} \rho(q, n)$$

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

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

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

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

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 platform² 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.

¹http://github.com/MediaEvalPersonDiscovery
²http://github.com/camomile-project
6. REFERENCES


