Multimodal Person Discovery in Broadcast TV at MediaEval 2016

Hervé Bredin, Claude Barras, Camille Guinaudeau
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France.
firstname.lastname@limsi.fr

ABSTRACT

We describe the "Multimodal Person Discovery in Broadcast TV" task of MediaEval 2016 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 has to be discovered in an unsupervised way from media content using text overlay or speech transcripts for the primary runs. 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 broadcasted, it is not always possible to predict which people will be the most important to find in the future and biometric models may not yet be available at indexing time The goal of this task is thus 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.

Started in 2011, the REPERE challenge aimed at supporting research on multimodal person recognition [3, 16]. 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). Thanks to this challenge and the associated multimodal corpus [13], significant progress was achieved in either supervised or unsupervised multimodal person recognition [1, 2, 4, 5, 6, 7, 11, 12, 17, 20, 21, 22, 24]. After the end of the REPERE challenge in 2014, the first edition of the "Multimodal Person Discovery in Broadcast TV" task was organized in 2015 [19]. This year's task is a follow-up of last year edition.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

2. DEFINITION OF THE TASK

Participants are provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot $s \in \mathbb{S}$ has to be automatically tagged with the names of people both speaking and appearing at the same time during the shot.

As last year, the list of persons is not provided *a priori*, and person biometric models (neither voice nor face) can not be trained on external data in the primary runs. 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). The main novelty of this year task is that participants may use their contrastive run to try brave new ideas that may rely on any external data, including textual metadata provided with the test set.

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 $\mathbb{N}$ the set of all possible person names in the universe, correctly formatted as `firstname_lastname` – while $\mathcal{N}$ is the set of hypothesized names.

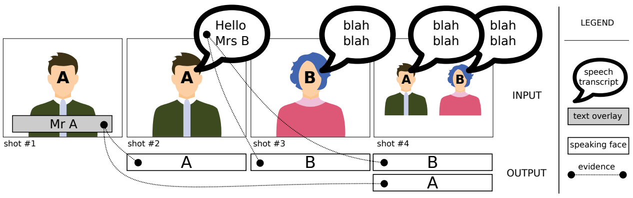

Figure 1: For each shot, participants have to return the names of every speaking face. Each name has to be backed up by an evidence.

3. DATASETS

The 2015 test corpus serves as development set for this year's task. It contains 106 hours of video, corresponding to 172 editions of evening broadcast news *"Le 20 heures"* of the French public channel *"France 2"*, from January 1st 2007 to June 30st 2007. This development set is associated with *a posteriori* annotations based on last year participants' submissions.

The test set is divided into three datasets: INA, DW and

3-24. The INA dataset contains a full week of broadcast for 3 TV channels and 3 radio channels in French. Only a subset (made of 2 TV video channels for a total duration of 90 hours) needs to be processed. However, participants can process the rest of it if they think it might lead to improved results. Moreover, this dataset is associated with manual metadata provided by INA in the shape of CSV files. The DW dataset [14] is composed of video downloaded from Deutsche Welle website, in English and German for a total duration of 50 hours. This dataset is also associated with metadata that can be used in contrastive runs. The last dataset contains 13 hours of broadcast from 3/24 Catalan TV news channel.

As the test set comes completely free of any annotation, it will be annotated *a posteriori* based on participants' submissions. In order to ease this annotation process, participants are asked to justify their assertion. To this end, each hypothesized name $n \in \mathcal{N}$ has 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. 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 time in a video when a person is visible, while his/her name is written on screen; an *audio* evidence is the time when the name of a person is pronounced, provided that this person is visible in a $[\text{time}-5s, \text{time}+5s]$ neighborhood. For instance, in Figure 1, shot #1 contains an *image* evidence for Mr A (because his name and his face are visible simultaneously on screen) while shot #3 contains an *audio* evidence for Mrs B (because her name is pronounced less than 5 seconds before or after her face is visible on screen).

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 available partially as open-source software.

For instance, a researcher from the speech processing community can 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. Figure 2 summarizes the available modules.

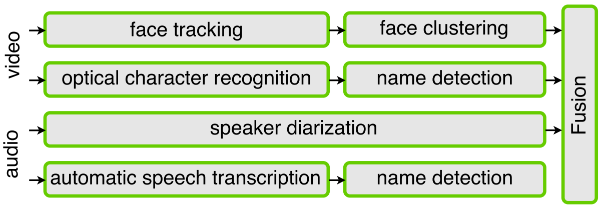

Figure 2: Multimodal baseline pipeline.

4.1 Video processing

Face tracking-by-detection is applied within each shot using a detector based on histogram of oriented gradients [9] and the correlation tracker proposed by *Danelljan et al.* [10]. Each face track is then described by its average *FaceNet* embedding and compared with all the others using Euclidean distance [25]. Finally, average-link hierarchical agglomerative clustering is applied. Source code for this module is available in *pyannote-video*[1].

Optical character recognition followed by name detection is contributed by IDIAP [8] and UPC. UPC detection was performed using LOOV [18]. Then, text results were filtered using first and last names gathered from internet and an hand-crafted list of negative words. Due to the large diversity of the test corpus, optical character recognition results are much more noisy than the ones provided in 2015.

4.2 Audio processing

Speaker diarization and speech transcription for French, German and English are contributed by LIUM [23, 15]. Pronounced person names are automatically extracted from the audio stream using a large list of names gathered from the Wikipedia website.

4.3 Multimodal fusion baseline

Three variants of the name propagation technique proposed in [21] are proposed. Baseline 1 tags each speaker cluster by the most co-occurring written name. Baseline 2 tags each face cluster by the most co-occurring written name. Baseline 3 is the temporal intersection of both. These fusion techniques are available as open-source software[2].

5. EVALUATION METRIC

Because of limited resources dedicated to collaborative annotation, the test set cannot be fully annotated. Therefore, the task is evaluated indirectly as an information retrieval task, using the folllowing principle.

For each query $q \in \mathbb{Q} \subset \mathbb{N}$ (`firstname_lastname`), returned shots are first sorted by the edit distance between the hypothesized person name and the query q and then by confidence scores. Average precision $\text{AP}(q)$ is then computed classically based on the list of relevant shots (according to the groundtruth) and the sorted list of shots. Finally, Mean Average Precision is computed as follows:

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

Acknowledgment

This work was supported by the French National Agency for Research under grants ANR-12-CHRI-0006-01 and ANR-14-CE24-0024. The open source CAMOMILE collaborative annotation platform[3] 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. The task builds on Johann Poignant involvement in 2015 task organization. Xavier Trimolet helped design and develop the 2016 annotation interface. We also thank INA, LIUM, UPC and IDIAP for providing datasets and baseline modules.

[1] http://pyannote.github.io

[2] http://github.com/MediaEvalPersonDiscoveryTask

[3] 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] D. Chen and J.-M. Odobez. Video text recognition using sequential monte carlo and error voting methods. *Pattern Recognition Letters*, 26(9):1386 – 1403, 2005.

[9] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893 vol. 1, June 2005.

[10] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate Scale Estimation for Robust Visual Tracking. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[11] 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.

[12] 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.

[13] 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.

[14] J. Grivolla, M. Melero, T. Badia, C. Cabulea, Y. Esteve, E. Herder, J.-M. Odobez, S. Preuss, and R. Marin. EUMSSI: a Platform for Multimodal Analysis and Recommendation using UIMA. In *International Conference on Computational Linguistics (Coling)*, 2014.

[15] V. Gupta, P. Deléglise, G. Boulianne, Y. Estève, S. Meignier, and A. Rousseau. CRIM and LIUM approaches for multi-genre broadcast media transcription. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 681–686. IEEE, 2015.

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

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

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

[19] J. Poignant, H. Bredin, and C. Barras. Multimodal Person Discovery in Broadcast TV at MediaEval 2015. In *MediaEval 2015*, 2015.

[20] 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.

[21] 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.

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

[23] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech*, Lyon (France), 25-29 Aug. 2013.

[24] 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.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.