# PERCOLATTE: A Multimodal Person Discovery System in TV Broadcast for the MediaEval 2015 Evaluation Campaign

Meriem Bendris[1], Delphine Charlet[2], Gregory Senay[3], MinYoung Kim[3], Benoit Favre[1],
Mickael Rouvier[1], Frederic Bechet[1],Géraldine Damnati[2]
[1]Aix Marseille Université, [2]OrangeLabs, [3]Panasonic Silicon Valley Lab

## ABSTRACT

This paper describes the PERCOLATTE participation to MediaEval 2015 task: "Multimodal Person Discovery in Broadcast TV" which requires developing algorithms for unsupervised talking face identification in broadcast news. The proposed approach relies on two identity propagation strategies both based on document chaptering and restricted overlaid names propagation rules. The primary submission shows 10% improvement of Mean Average Precision of the baseline on the INA corpus.

## 1. INTRODUCTION

Identifying people in TV broadcasts has had a lot of attention the last decade in the literature. Current trends aim to combine traditional techniques with high level information such as prior knowledge on document structure. Indeed, TV program often have regular structure organized in homogeneous sequences. The REPERE Challenge, that ended in 2014, aimed at developing multimodal algorithms for people identification in TV broadcasts. Our PERCOLATOR system based on scene understanding features ranked first on the main task in 2014 [2]. The Mediaeval "Multimodal Person Discovery in Broadcast TV" task focuses on unsupervised talking face identification [7] for search engine applications. One novelty of this task is the metadata made available by the organizers allowing expanded participations.

This paper describes the PERCOLATTE system submitted at the MediaEval 2015. The system relies on the enrichment of broadcast news with video structure features such as shot classification (studio/report) and speaker role recognition. Two identification strategies were developed: the primary is based on chapter-restricted identity propagation to shot clusters and the secondary is based on speaker identification and rule-based speaker-face mapping. Figure 1 shows the pipeline of the PERCOLATTE system. Notice that no face-related processing (detection/identification) is used in our approach.

## 2. TOOLS

The MediaEval 2015 organizers made available different baseline mono-modal tools. In our system, we used the provided Overlaid Person Names (OPN) [6] system. In addition, we used the automatic named entities [4] and the speaking-face mapping to fix the identification scores.

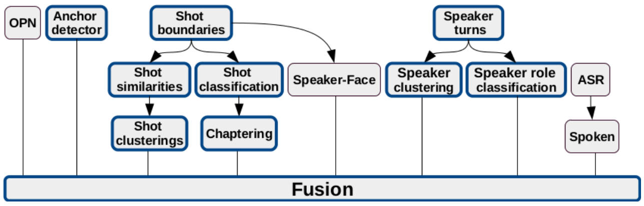

**Figure 1: The PERCOLATTE pipeline. Our modules are outlined in blue.**

### 2.1 List of names

The Audiovisual National Institute (INA) collects and enriches broadcast news with metadata such as summary, identity of journalists, etc. We collected the metadata[1] from December 2004 to December 2009 and extracted automatically the list of several journalists and anchors.

### 2.2 Overlaid anchor name detection

Anchor names were not detected by the provided OCR system. We developed an anchor name detector relying on a *Levenshtein*-based mapping of OCR results [2] (on ×2 rescaled frames) and the list of names described previously.

### 2.3 Speaker clustering

The speaker clustering follows the approach described in [1]. First, speech segments are grouped using a BIC clustering. Then, obtained clusters are modelled with GMMs in order to more accurately compare voices using a Cross-Likelihood Criterion (CLR) in a second agglomerative clustering. At each iteration, Viterbi decoding is performed to re-segment the speech data into speaker turns given the new clusters.

### 2.4 Speaker role classification

We used a simplified version of the speaker role classification approach described in [3]. First, the anchor is the speaker cluster who speaks the most and regularly. Then, a binary classification reporter/other is performed. As no speech transcript was available, in this work, the classification relies only on an acoustic GMM classifier.

### 2.5 Speaker identification

Speaker turns are identified by propagating OPNs to the speaker turns that maximise temporal overlapping and to it's cluster within the same chapter.

[1] Available on http://www.ina.fr

[2] https://github.com/meriembendris/ADNVideo

*MediaEval 2015 Workshop* Sept. 14-15, 2015, Wurzen, Germany

## 2.6 Shot boundary detection

Two systems were used based on RGB histograms peaks [10] and HSV histogram peaks on sliding window [2]. As the evaluation script needs the provided shot segmentation, a shot boundaries mapping was necessary.

## 2.7 Shot similarity and clustering

In order to measure the similarity between shots, three features where extracted: RGB histograms, HOG features on resized frames (128×64) and DNN-based frame representation (*image embeddings*). For the DNN-based features, we used the Alexnet DNN [5] to extract feature vectors at the $3^{rd}$ fully-connected layer (1000 dimension vectors). Then, shots were grouped using cosine-based distance and Integer Linear Program clustering (described in [9]).

## 2.8 Shot classification and chaptering

The shot classifier is trained on external data (8 broadcast news, 4914 shots). Four labels were annotated: studio, report, mixed and other. First, HOG features on resized frames (128x64) were extracted for each shot. Then, a Liblinear[3] classifier was trained on three quarters of data. The system reached 99.43% of accuracy on the remaining quarter. Finally, successive shots sharing the same label were grouped into chapters.

# 3. TALKING FACE IDENTIFICATION

Participants were asked to provide identified talking faces within shots with their confidence scores and evidences justifying their assertions. Two strategies were developed.

## 3.1 The primary strategy

The primary strategy relies on the fact that report chapters are independent in broadcast news. The strategy is based on a restricted OPN propagation to cluster shots within the same chapter. Precisely, we followed those rules:

- Propagate OPN to overlapping shots and their shot clusters sharing the speaker cluster within a chapter.
- Propagate anchor name to overlapping "studio" shots and their shot clusters without chapters restrictions.
- Propagate anchor name if the speaker role is an anchor.

For each identified talking face, the score was initialized by the provided OPN score and incrementally increased following those events: OPN shot overlapping, provided talking-face score $> 0.8$ and OPN pronounced around the shot(±5s).

## 3.2 The secondary strategy

The secondary strategy is based on a speaker identification followed by speaker-face rule-based mapping. This mapping relies on simple rules based on prior knowledge about broadcast news. Precisely, we considered a speaker visible when the name appears on the screen (OPN), on studio shots and on report shots when the role is not a reporter. In this strategy, no scores function was developed (score=1).

## 3.3 The evidence

To ensure that identities where detected only in unsupervised way, and to help collaborative annotations of the test set, participants were asked to select one shot per name proving his/her identity. For each name, we selected the provided OPN shot that maximizes the OCR result score.

# 4. EVALUATION

Systems were evaluated using the Mean Average Precision ($MAP$) metric and the official $C$ and $EwMAP$ metrics described in [7]. Two submission deadlines were fixed: July 1st and 8th. In our submissions, the only difference concerns shot boundary mapping. Indeed, on July the 1st, the mapping was based on overlapping shots over $0.5s$ (a rather cure strategy) while it was on overlapping coverage above 50% for the July the 8th submissions. Four runs were submitted:

- **Primary:** primary strategy with DNN- and HOG-based shot clustering.
- **Primary_DNNOnly:** primary strategy with DNN-based shot clustering.
- **Primary_RGBOnly**: primary strategy with RGB-based shot clustering.
- **Secondary:** secondary strategy based on speaker identification and speaker-face rule-based mapping.

Table 1 shows results of the PERCOLATTE runs. The secondary strategy having similar principles than the baseline [8] shows a MAP improvement of 8%. Indeed, chapter-restricted propagation in addition to simple rule-based speaker-face mapping based on shot classification and speaker roles allowed to detect less talking faces with higher precision. The primary strategy using DNN- and HOG-based shot clustering obtains the best MAP of 88.45%. This shows the consistency of the chapter-constrained propagation strategy in broadcast news. Contrastive runs with different features for shot clustering did not show significant differences. Anchor names were detected in 93% of shows. However, the primary run without anchor-specific modules performs 88.31% of MAP.

| Metrics | EwMAP | MAP | C |
|---|---|---|---|
| Baseline | 78.35 | 78.64 | 92.71 |
| Secondary on July 1st | 85.89 | 86.12 | 97.68 |
| Secondary on July 8th | 86.40 | 86.61 | 97.68 |
| Primary_DNNOnly on July 1st | 81.41 | 81.67 | 97.63 |
| Primary_DNNOnly on July 8th | 87.75 | 88.01 | 97.63 |
| Primary_RGBOnly on July 1st | 81.02 | 81.28 | 97.63 |
| Primary_RGBOnly on July 8th | 87.33 | 87.60 | 97.63 |
| Primary on July 1st deadline | 81.70 | 81.96 | 97.63 |
| Primary on July 8th | **88.19** | **88.45** | **97.63** |

**Table 1: Performances of PERCOLATTE 2015 runs.**

# 5. CONCLUSIONS

In this paper, we described the PERCOLATTE strategies for talking face identification. The system without face-related processing is based on chapter-restricted propagation of overlaid names. A significant improvement of the baseline is achieved on the INA corpus by the primary strategy (+10% of MAP). Results show that in structured programs, easy-to-establish features such as shot classification and prior knowledge about broadcast news allow to improve significantly talking faces identification.

**Acknowledgment.** This work has been carried out thanks to the support of the A*MIDEX project (n° ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR).

[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

# 6. REFERENCES

[1] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[2] F. Bechet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, G. Linares, G. Senay, P. Tirilly, and J. Martinet. Multimodal understanding for person recognition in video broadcasts. In *Interspeech, Singapore*, 2014.

[3] G. Damnati and D. Charlet. Multi-view approach for speaker turn role labeling in tv broadcast news shows. In *INTERSPEECH*, pages 1285–1288. ISCA, 2011.

[4] M. Dinarelli and S. Rosset. Models cascade for tree-structured named entity detection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[6] J. Poignant, L. Besacier, G. Quenot, and F. Thollard. From text detection in videos to person identification. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012.

[7] J. Poignant, H. Bredin, and C. Barras. Multimodal person discovery in broadcast tv at mediaeval 2015. In *MediaEval*, 2015.

[8] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. In *Interspeech 2012 - Conference of the International Speech Communication Association*, Portland, OR, United States, 2012. Poster Session: Speaker Recognition III.

[9] M. Rouvier and S. Meignier. A global optimization framework for speaker diarization. In *Speaker Odyssey*, 2012.

[10] H. Zhang, R. Hu, and L. Song. A shot boundary detection method based on color feature. In *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*, 2011.