# EUMSSI team at the MediaEval Person Discovery Challenge 2016

Nam Le[1,2], Sylvain Meignier[3], Jean-Marc Odobez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédéral de Lausanne, Switzerland
[3] LIUM, University of Maine, Le Mans, France
{nle, odobez}@idiap.ch, sylvain.meignier@univ-lemans.fr

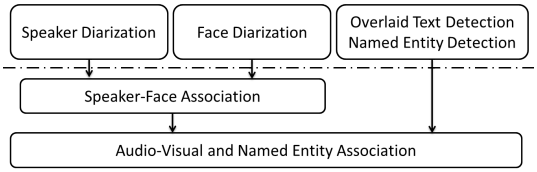


**Figure 1: Architecture of our system**

## ABSTRACT

We present the results of the EUMSSI team's participation in the Multimodal Person Discovery task. The goal is to identify all people who simultaneously appear and speak in a video corpus. In the proposed system, besides improving each modality, we emphasize on the ranking of multiple results from both audio stream and visual stream.

## 1. INTRODUCTION

As the retrieval of information on people in videos is of high interest for users, algorithms indexing identities of people and retrieving their respective quotations are indispensable for searching archives. This practical need leads to research problems on how to identify people presence in videos. Given the raw TV broadcasts, each shot must be automatically tagged with the name(s) of people who can be both seen as well as heard in the shot along with the confident score. The list of people is not known apriori and their names must be discovered from video text overlay or speech transcripts [6]. To this end, a video must be segmented in an unsupervised way into homogeneous segments according to person identity, like speaker diarization and face diarization, to be combined with the extracted names. Our goal is to benchmark our recent improvements in all components and address the fusion of multimodal results.

## 2. PROPOSED SYSTEM

The system we proposed is illustrated in Fig. 1. It consists of 4 main parts: video optical character recognition (OCR) and named entity recognition (NER), face diariation, speaker diarization, and fusion naming.

### 2.1 Video OCR and NER

To detect OCR segments in videos and exploit them for retrieval, we first relied on the approaches described in [2, 1] for text recognition in videos, and on [3, 15] for text recognition and indexing. In brief, given an input video, two main steps are applied: first the video is preprocessed with a motion filtering to reduce noise, and individual frames are processed to localize and binarize the text regions for text recognition. As compared to printed documents, OCR in TV news videos encounters several challenges: low resolution of text regions, sequence of different texts continuously displayed, or small amount of text to be recognized etc. To tackle these, multiple image segmentations of the same text region are decoded, and then all results are compared and aggregated over time to produce several hypotheses. The best hypothesis is used to extract people names for identification. To recognize names from texts, we use the MITIE open library [1], which provides state-of-the-art NER tool. To improve the raw MITIE results, a heuristics preprocessing step identifies names of editorial staff based on their roles (cameraman, editor, or writer) because they do not appear within the video, thus are not useful for identification.

### 2.2 Face diarization

Given the video shots, face diarization process consists of (i) face detection, (ii) face tracking, and (iii) face clustering.

**Detection & tracking.** Detecting and associating faces can be challenging due to the wide range of media content, where faces can appear with varied illumination and noise. To overcome these challenges, we use a fast version of deformable part-based model (DPM) [5, 11, 4] to detect faces at multiple poses and variation. Tracking is performed using the CRF-based multi-target tracking framework [7], which relies on the unsupervised learning of time sensitive association costs for different features. Because the bottle-neck of the system is detection, the detector is only applied 4 times per second. We also trained an explicit false alarm classifier at the track level to efficiently filter out false tracks. Further details can be found in [9].

**Face clustering.** We hierarchically merge face tracks across all shots using matching and biometric similarity measures similarly to [8] with two improvements: shot-constrained face clustering (SCFC) and the use of total variability modeling (TVM). SCFC is a divide-and-conquer strategy. Face clustering is first applied limiting within each group of similar shots. Then all resulting face clusters, which are now much fewer in quantity, are hierarchically merged. TVM is a state-of-the-art biometrics method that can represent faces which can appear in widely different contexts and sessions

[1] https://github.com/mit-nlp/MITIE


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

[17, 16]. To compute similarity between face clusters, we simply use the average distance between all pairs of faces using the cosine distance between i-vectors.

## 2.3 Speaker diarization

The speaker diarization system is based on the LIUM Speaker Diarization system[14], which is publicly distributed[2]. It is provided to all participants as the baseline method.

## 2.4 Identification and result ranking

After obtaining homogeneous clusters during which distinct identities speak or appear, one needs to assign each name output from NER module to the correct clusters. However, associating auditory voices with visual person clusters or names has two major difficulties. The visible person may not be the current speaker and the speaking person can be dubbed by a narrator in a different language. Although we have introduced a temporal learning method to solve the dubbing problem [10], incorporating it into an AV diarization system is still an open question. Because of these problems of AV association, we use a direct naming method [13] which finds the mapping between clusters and names to maximize the co-occurrences between them.

**Identification.** Names are propagated based on the outputs of face diarization and speaker diarization independently. The direct naming method is applied to speaker clusters to produce a mapping between names and clusters. All shots which overlap with the clusters are tagged with the corresponding names with equal confident scores. The same direct method is applied to face clusters to produce a set of named clusters. Unlike speaker naming, for one shot, a name coming from face naming is ranked based on the talking score of the cluster's segment within that shot. The talking score is predicted using lip motion and temporal modeling with LSTM [10]. Based on the two results, we propose a strategy to appropriately combine them.

**Ranking.** Let $S = \{s_k\}$ be the list of testing shots. Within each shot, $\{N_i^F, t(N_i^F)\}$ is the set of names returned by face naming and the corresponding talking scores and $\{N_i^A, 1.0\}$ is the set of names returned by speaker naming, each is ranked equally with score 1.0. The names which the two methods agree on are ranked highest. Then, names from face naming are ranked higher than speaker naming because we found that face naming is more reliable in empirical experiments. Alternative strategies that rank speaker naming equal or higher than face naming gave inferior results. Our ranking strategy is described in Algo. 1.

**Further fusion.** Finally, replacing individual component in our system with baseline NER [12] and face diarization [3] can produce complementary results. Therefore, these results are added to our final submission with lower confident scores.

# 3. EVALUATION

Participants are scored based on a set of queries. Each query is a person name in the corpus, each participant has to return all shots when that person appears and talks. The metric is Mean Average Precision (MAP) over all queries. In Tab. 1, we report our result on the test set as of 24/09/2016 [4].

[2]www-lium.univ-lemans.fr/en/content/liumspkdiarization

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

[4]The groundtruth is still updated by a collaborative annotation process.

**Algorithm 1** Ranking names within shots

**for** $s_k \in S$ **do**
  $Q_{s_k} = \emptyset$
  Face_naming($s_k$) $\Rightarrow (N_i^F, t(N_i^F))$
  Speaker_naming($s_k$) $\Rightarrow (N_j^A, 1.0)$
  **for** each $N_i^F$ **do**
    **if** $\exists N_j^A / N_j^A = N_i^F$ **then**
      $Q_{s_k} = Q_{s_k} \cup \{(N_i^F, t(N_i^F) + 2.0)\}$
    **else**
      $Q_{s_k} = Q_{s_k} \cup \{(N_i^F, t(N_i^F) + 1.0)\}$
  **for** each $N_j^A$ **do**
    **if** not $\exists N_i^F / N_i^F = N_j^A$ **then**
      $Q_{s_k} = Q_{s_k} \cup \{(N_j^A, 1.0)\}$

| | MAP@1 | MAP@10 | MAP@100 |
|---|---|---|---|
| Sub. (1) | 30.3 | 22.0 | 21.0 |
| Sub. (2) | 58.6 | 42.9 | 42.0 |
| Sub. (3) | 64.2 | 53.1 | 52.1 |
| Sub. (4) | 68.3 | 56.2 | 54.7 |
| Sub. (5) | 79.2 | 65.2 | 63.4 |

**Table 1: Benchmarking results of our submissions. Details of each submission in the text.**

Each of our 5 submissions (Sub.) is as following:

- Sub. (1) and Sub. (2) used our face naming without talking score with baseline OCR-NER (1) or with our OCR-NER (2).
- Sub. (3) used our face naming with talking score.
- Sub. (4) used the combination of talking face naming in sub. (3) with speaker naming.
- And sub. (5) used the combination of sub. (4) with other systems using baseline OCR-NER or baseline face diarization. This is also our primary submission.

When comparing sub. (1) and sub. (2), one can observe that our OCR-NER outperforms the baseline OCR-NER by a large margin. This may be contributed by the high recall of our system. Because the metric is averaged over all queries, any missing name can significantly decrease the overall MAP. On the other hand, false names are less problematic because of two reasons: they may not be associated with any clusters and they are not queried at all. In sub. (3), using talking face detection with LSTM, we can further improve by 5.6%. By combining face naming and speaker naming, we manage to increase the precision. This shows the potential for further research of better audio-visual naming. In our primary submission (5), the result are greatly boosted when other methods are added. From this we can note that these methods are complementary to each other and how to exploit their advantages is an open question in the future.

# 4. CONCLUSION

We have presented our system in MediaEval challenge 2016. This system consists of our recent advances in video processing and temporal modeling. Although each modality shows positive performance, the current system has not taken full advantage of both audio and visual streams. Therefore, the testing results serve as the basis for us to work further in this direction.

**Acknowledgement** This research was supported by the European Union project EUMSSI (FP7-611057).

## 5. REFERENCES


[1] 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.

[2] D. Chen, J.-M. Odobez, and H. Bourlard. Text detection and recognition in images and video frames. *Pattern Recognition*, 37(3):595–608, 2004.

[3] N. Daddaoua, J.-M. Odobez, and A. Vinciarelli. Ocr based slide retrieval. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 945–949. IEEE, 2005.

[4] C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[6] C. B. H. Bredin, C. Guinaudeau. Multimodal person discovery in broadcast tv at mediaeval 2016. In *Proc. of the MediaEval 2016 Workshop*, Hilversum, Netherlands, Oct. 2016.

[7] A. Heili, A. Lopez-Mendez, and J.-M. Odobez. Exploiting long-term connectivity and visual motion in crf-based multi-person tracking. *IEEE Transactions on Image Processing*, 23(7):3040–3056, 2014.

[8] E. Khoury, P. Gay, and J.-M. Odobez. Fusing Matching and Biometric Similarity Measures for Face Diarization in Video. In *ACM ICMR*, 2013.

[9] N. Le, A. Heili, D. Wu, and J.-M. Odobez. Temporally subsampled detection for accurate and efficient face tracking and diarization. In *International Conference on Pattern Recognition*. IEEE, Dec. 2016.

[10] N. Le and J.-M. Odobez. Learning multimodal temporal representation for dubbing detection in broadcast media. In *ACM Multimedia*. ACM, Oct. 2016.

[11] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, pages 720–735. Springer, 2014.

[12] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *2012 IEEE International Conference on Multimedia and Expo (ICME)*, pages 854–859. IEEE, 2012.

[13] 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*, page 4p, 2012.

[14] 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.

[15] A. Vinciarelli and J.-M. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.

[16] R. Wallace and M. McLaren. Total variability modelling for face verification. *Biometrics, IET*, 1(4):188–199, 2012.

[17] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011.