EUMSSI team at the MediaEval Person Discovery Challenge

Nam Le[1,2], Di Wu[1], 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, dwu, odobez}@idiap.ch, sylvain.meignier@univ-lemans.fr

ABSTRACT

We present the results of the EUMSSI team's participation in the Multimodal Person Discovery task at the MediaEval challenge 2015. The goal is to identify all people who simultaneously appear and speak in a video corpus, which implicitly involves both audio stream and visual stream. We emphasize on improving each modality separately and benchmarking them to analyze their pros and cons.

1. INTRODUCTION

Nowadays, viewers, journalists, or archivists have access to a vast amount multimedia data. The need for browsing and retrieval tools of these archives has led researchers to devote effort to developing technologies that create searchable indices [14]. In this view, as humans are very interested in other people while consuming multimedia contents, 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 and answer 'who appears when?' or 'who speaks when?'.

In particular, in the MediaEval Person Discovery task, the goal is the following. 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. The list of people is not known a priori and their names must be discovered in an unsupervised way from video text overlay or speech transcripts. This situation corresponds to cases where at the moment a content is created or broadcast, some of the appearing people are relatively unknown but may later on become a trending topic on social networks or search engines. In addition, to ensure high quality indexes, algorithms should also help human annotators double-check these indexes by providing an evidence of the claimed identity (especially for people who are not yet famous).

2. PROPOSED SYSTEM

The participation of the EUMSSI team was to enable the assessment of the different modules developed by the authors in the past [11, 7, 8, 17, 4]. In this view, starting from the baseline provided by the organizer, the goal was to replace baseline components by the team's components, whenever they have been made compatible and their processing speed was enough to address the data provided in the challenge, and test their performance to understand their advantages.

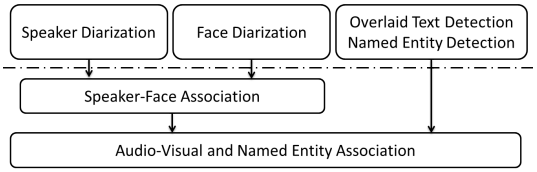

Figure 1: Architecture of proposed system

The used system, as illustrated in Fig. 1, consists of 2 main stages. The first stage detects and clusters speakers, faces and overlaid person names, including extracting Named Entities (NE). The second one associates speakers to faces using co- occurrence statistics and the overlaid person names are propagated to the speakers, or faces, in order to give the identities of the persons in the show.

2.1 Speaker diarization

The speaker diarization system ("who speak when?") is based on the LIUM Speaker Diarization system[16], freely distributed[1]. This system has achieved the best or second best results in the speaker diarization task on REPERE French broadcast evaluation campaigns 2012 and 2013 [6].

The diarization system is first composed of an acoustic Bayesian Information Criterion (BIC)-based segmentation followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. A Viterbi decoding re-segments the signal using GMMs with 8 diagonal components learned by EM-ML, for each cluster. Segmentation, clustering and decoding are performed with 12 MFCC+E, computed with a 10ms frame rate. Music and jingle regions are removed using a Viterbi decoding with 8 GMMs (trained on french broadcast news data) for music, jingle, silence, and speech (with wide/narrow band variants for the last two, and clean or noised or musical background variants for wideband speech).

In the above steps, features were used unnormalized in order to preserve information on the background environment, which may help differentiating between speakers. At this point however, each cluster contains the voice of only one speaker, but several clusters can be related to a same speaker. The background environment contribution must be removed from each GMM cluster, through feature gaussianization. Finally, the system is completed with clustering method based on the i-vectors paradigm and Integer Linear Programming (ILP). This new clustering method is fully described in [17] and [4]. The ILP clustering along with i-

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

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

vectors speaker models gives better results than the usual hierarchical agglomerative clustering based on GMMs and cross-likelihood distances [1].

2.2 Face diarization

Given the video shots, face diarization process consists of (i) face detection, detecting faces appearing within each shot, (ii) face tracking, extending detections into continuous tracks within each shot, and (iii) face clustering, grouping all tracks with the same identity into clusters.

Face detection. Detecting faces in broadcasting media can be challenging due to the wide range of media content. Faces can appear in widely different situations with varied illumination and noise such as in studio, during live coverage, or during political debate. To overcome these challenges, we employ deformable part-based model (DPM) [5, 12], which can detect faces at multiple poses and variation. Because, the main disadvantage of DPM is its long running time, face detector is only applied 2 times per second.

Face tracking. The goal of this step is to create continuous face tracks in one video shot, which raises the need for association individual detections. Because of long gaps between detected faces, we exploit long term connectivity using CRF-based multi-target tracking [10]. This framework relies on the unsupervised learning of time sensitive association costs for different features. First, similarities between detections are computed based on low level features (color histogram, position, motion, SURF keypoint descriptors) which can be computed fast. Then, for each feature type, the corresponding pairwise factor of the CRF is defined as the probability of similarity measurements between pairs of detections under two distinct hypotheses that they correspond to the same label or not. By optimizing a graph labeling posterior, we assign the same label to detections belonging to the same face, and different labels to different faces.

Face clustering. Given the face tracks across all video shots, we hierarchically merge face tracks tracks using matching and biometric similarity measures [11]. Matching cluster similarity is calculated based on average of distances between sparse keypoints of two clusters. Meanwhile, biometric model-based similarity measures how densely extracted features from one cluster are likely to belong to the model of the other cluster, as compared to the likelihood given by the statistical model, and vice-versa. Face tracks are first clustered using only feature-based matching, yielding clusters with sufficient data to adapt the biometric models. Then, model-based similarity is combined with matching similarity to merge clusters until stopping criteria are met. Similarly to speaker diarization, face diarization produces face segments during which distinct identities appear.

2.3 Person Naming

Identity candidate retrieval. OPNs can be more reliably extracted using Optical Character Recognition (OCR) techniques [2, 13] than from automatic speech transcripts. Therefore, we only exploit name entities detected from OCR by [3] as potential identity candidates.

Direct one-to-one tagging. As mentioned earlier, our goal is to benchmark improvements of each modality in the system. Hence, there is one assumption that the temporal clusters of the diarization processes are trustable. In this work, we use a simple one-to-one naming method provided by [15] which finds the mapping between clusters and named entities to maximize the co-occurrences between them.

Method	EwMAP	MAP	C	#(2485)
Baseline	49.98	50.32	58.75	617
SpkDia	**65.31**	66.70	72.50	2817
FaceDia	**66.38**	67.98	71.67	1691

Table 1: Results on REPERE test 2 (dev set)

Method	EwMAP	MAP	C	#(21963)
Baseline	78.35	78.64	92.71	12066
FaceDia	83.04	83.33	90.77	7237
SpkDia*	**89.75**	**90.14**	**97.05**	30583
SpkFace	89.53	89.90	96.52	20601
* Primary submission				

Table 2: Results on INA (test set)

3. EXPERIMENTS

We evaluated 3 methods: SpkDia, FaceDia, and SpkFace. In *SpkDia* (primary submission), we apply naming based on audio information only (this is equivalent to assumption that all speakers which are associated with a name are visible and speaking). This is our primary submission for the challenge. Second, in *FaceDia*, we apply naming based on visual information only, and assume that all visible faces (which are associated with a name) are talking. Third, in *SpkFace*, we apply naming based on audio information only, but validate if there exists visible faces during the speech segments (if not, the segment is discarded). Because our approaches are monomodal and fully unsupervised, we did not use the information provided by leaderboard to improve performance.

The results using the challenge performance measures are reported in Tab. 1 for the REPERE test 2 data [9] as the initial development data and in Tab. 2 for the challenge testing part of the INA dataset. SpkDia is the most robust and performs the best even without any face information, which might be explained by two points. First, there is usually only one speaker at a time, and not much noise in the challenge data. Meanwhile, face diarization can be difficult due to multiple faces, facial variation, missed detections, etc. Hence, speech clusters tend to be more reliable than face clusters. Second, when a speaker is not visible, it is often the anchor of the show, who is counted as one query equally to those appearing for short duration. Therefore, SpkDia is not penalized much by the visibility of speakers. We can observe this effect more in the last column of Tab. 2 which shows the number of person presence with names predicted by each scheme. Using faces to filter 1/3 of speech segments does not help to increase precision because these segments correspond to a small number of repetitive speakers. Also, though face diarization gives only 1/3 of possible names, these names are precise person-wise. This interesting fact may provide outlook on combining 2 modalities.

4. FUTURE WORKS

We have presented our system in MediaEval challenge. The testing result serves as our basis for improving each component. We are working on speeding up the tracking process as well as investigating alternative face representations such as total variability modeling. On another hand, current system has not taken full advantage of both audio and visual streams, which we plan to proceed in the future.

5. REFERENCES

[1] C. Barras, X. Zhu, S. Meignier, and J. Gauvain. Multi-stage speaker diarization of broadcast news. 14(5):1505–1512, Feb. 2006.

[2] 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.

[3] M. Dinarelli and S. Rosset. Models cascade for tree-structured named entity detection. In *IJCNLP*, pages 1269–1278, 2011.

[4] G. Dupuy, S. Meignier, P. Deléglise, and Y. Estève. Recent improvements towards ILP-based clustering for broadcast news speaker diarization. 2014.

[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] O. Galibert and J. Kahn. The first official REPERE evaluation. In *Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, 2013.

[7] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise. A Conditional Random Field approach for Audio-Visual people diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014.

[8] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise. Face identification from overlaid texts using Local Face Recurrent Patterns and CRF models. In *IEEE International Conference on Image Processing (ICIP)*, 2014.

[9] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The repere corpus : a multimodal corpus for person recognition. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[10] 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.

[11] E. Khoury, P. Gay, and J.-M. Odobez. Fusing matching and biometric similarity measures for face diarization in video. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 97–104. ACM, 2013.

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

[13] 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.

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

[15] 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.

[16] 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.

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