Multimodal understanding for person recognition in video broadcasts

Frederic Bechet[1], *Meriem Bendris*[1], *Delphine Charlet*[4], *Geraldine Damnati*[4], *Benoit Favre*[1], *Mickael Rouvier*[1], *Remi Auguste*[2], *Benjamin Bigot*[3], *Richard Dufour*[3], *Corinne Fredouille*[3], *Georges Linares*[3], *Jean Martinet*[2] *Gregory Senay*[3], *Pierre Tirilly*[2]

[1]Aix Marseille Universite, CNRS-LIF ; [2]Universite de Lille, LIFL
[3]Universite d'Avignon, LIA; [4]Orange Labs, France

Abstract

This paper describes a multi-modal person recognition system for video broadcast developed for participating in the Defi-Repere challenge. The main track of this challenge targets the identification of all persons occurring in a video either in the audio modality (speakers) or the image modality (faces). This system is developed by the PERCOL team involving 4 research labs in France and was ranked first at the 2014 Defi-Repere challenge. The main scientific issue addressed by this challenge is the combination of audio and video information extraction processes for improving the extraction performance in both modalities. In this paper, we present the strategy followed by the PERCOL team for speaker identification based on enriching the speaker diarization with features related to the "understanding" of the video scenes: text overlay transcription and analysis, automatic situation identification (TV set, report), the amount of people visible, TV set disposition and even the camera when available. Experiments on the REPERE corpus show interesting results on the speaker identification system enriched by the scene understanding features and the usefulness of the speaker to identify faces.

Index Terms: multi-modal speaker recognition, multi-modal fusion.

10.21437/Interspeech.2014-146

1. Introduction

Identifying people in TV broadcast is naturally a multi-modal task. Indeed, people can be identified thanks to biometric information (face or voice), a reference to their identity in spoken words, or their name in overlaid text. The *Defi-Repere*[1] challenge consists in identifying people in TV-broadcast using cues from both audio and video information [1]. The challenge provides a set of videos manually annotated with speaker segmentation, speech transcription, overlaid texts and face annotation. All image-related annotations are sampled every 10 seconds on so-called key-frames. Systems taking part in the challenge must generate a list of sequences with person names according to their presence (talking and/or visible), using both biometric models and context analysis. The different systems that were submitted to previous editions of the challenge were based on fusion of mono-modal systems results such as speaker/face diarization, ASR, OCR and speaker/face identification [2, 3]. The main source of information for these systems is biometric models of speakers and faces trained on collected identities.

Results obtained during the 2013 edition of the *Defi-Repere* challenge have shown the strength of this approach for non-ambiguous situations where a speaker is the only person visible on screen, in which case the identity can be propagated between the audio and visual modalities. However, this setting is too simplistic to cover all cases. Indeed, in TV-broadcast, the speaker is completely or partially visible only in a subset of sequences. Moreover, most of the visible people are not talking and overlay text to identify them is usually available once per show. In addition, for non-talking people, relying only on visual biometric models is a strong limitation due to the variety of pause and condition for face identification. The cost of maintaining dictionaries of biometric models is prohibitive for both speaker and face models. Our system revolves around making predictions in the speaker modality, which is less ambiguous as usually only one person is talking at a given time, and propagating them to the visual modality thanks to scene modelling and multi-modal understanding.

In this paper, we describe the PERCOL system ranked the first at the 2014 edition of the *Defi-Repere* challenge. As speaker ambiguities are limited, the backbone of the system is to enrich the speaker identification module with features related to the *understanding* of complex TV broadcasts scenes: text overlay transcription and analysis, automatic situation identification (TV set, report), the number of people visible, TV set disposition and even the camera when available. The fused system improves the performance of the speaker identification system, and predicts the presence of people in the image modality. Section 2 presents some related work to the problem of multi-modal person identification; Section 3 describes the multi-modal talking people identification module of our system; Section 4 specifies the visual features used to build a semantic representation of a scene and how they are integrated into the speaker module. Finally Section 5 shows some contrastive results on the *Defi-Repere* corpus.

2. Multimodal Person Recognition: the *Defi-Repere* challenge

Multimodal video processing has been addressed differently by the two scientific communities of image processing and speech processing. For example, in the speech processing community, previous studies focused on topic segmentation [4] or speaker role recognition [5], using only the audio signal and speech transcriptions. Similarly, in the image processing community, previous works have explored scene analysis [6] and action recognition, based only on images, for detecting actions such as "person walking" or "closing door".

In addition to these monomodal studies, video segmentation has been carried out with audio and video features [7, 8]. However these approaches limit the cooperation between the modalities to a fusion process, either an early fusion of multimodal features, or a late fusion of monomodal decisions.

[1]http://www.defi-repere.fr

Nevertheless, for more challenging tasks than video segmentation, there is a need for more complex multimodal models. The Defi-Repere challenge illustrates this need [9]. The main task of the challenge is the identification of persons in video broadcasts (talk-shows, news, debates, and reports). The persons to be identified can be well known celebrities (politicians, movie/music stars) and TV presenters, for whom biometric models can be trained in advance, as well as guests potentially unknown from any available database and for whom only a multimodal analysis of the video can help find an identity through text overlay or speech transcription analysis.

In the 2014 edition of the evaluation, the test corpus consists of about 10 hours of French TV from seven shows (*BFMStory, CaVousRegarde, LCPInfo, EntreLesLignes, PileEtFace, TopQuestion, RuthElkief*). While "*visible*" persons are only annotated on keyframes approximately every 10 seconds, speaker identities are given on the whole audio signal. For the image modality, the task is a superset of face identification as all persons occurring in a keyframe have to be identified, whether or not the face of this person is visible. For example, participants must be able to identify back-facing persons as well as partially occluded faces and small faces, often preventing the application of biometric identification [10]. In our corpus, about 10% of persons are considered as impossible to identify with biometric models because of these factors.

The spoken modality does not show such ambiguity: most of the time only one speaker is audible and there is no "occlusion" phenomenon. However, even with perfect speaker diarization and identification, attributing an identity to a person in an image from the speaker modality is not straightforward. For example [11] have shown, studying a TV talk-show corpus, that speakers are only visible 60% of the time, and that on average a face is talking only in 30% of its visible time.

Because of this asynchrony between the audio and video channels, multimodal speaker and visible person recognition is a challenging task. Besides the fusion of monomodal features, we claim that a system can benefit from a richer level of *understanding* of the video scene to process. We present such a system in the next sections, firstly by introducing the speaker identification module based on audio and OCR features, then describing how video scene analysis features can be added to turn this speaker-oriented system into a multimodal person recognition system.

3. From Speaker diarization to person recognition

In our system, the speaker identification strategy consists in the combination of two processes: a speaker diarization process in charge of producing clusters of speaker turns and a naming process in charge of associating identity hypotheses to each speaker turn. For the speaker diarization process, first, speech and non-speech are separated, segmenting the speech into turns. Then, speaker turns are grouped using the diarization system presented in [12].

The naming process relies on three sources of possible identity described in the next subsections: a speaker identification module based on biometric models (*Speaker ID*); an overlaid text processing module (*Overlaid Person Name - OPN*) and a name spotting module from Automatic Speech Recognition transcriptions (*ASR Name Spotting*).

3.1. Speaker ID

The biometric speaker ID system is built on the Alize platform [13], and takes advantage of i-vectors for modeling speakers. For Repere-2014, we trained 533 models of the most frequent speakers of the Repere training corpus and political figures prominent on the test period. While the minimum duration of training data is fixed to 30s to estimate an i-vector-based speaker model, the large amount of training data possibly available for some speakers (notably anchor speakers) was handled by training an i-vector every 2'30s of speech segment available. For scoring, the maximum score representing the best training i-vector for a given speaker regarding an identity test was selected.

The simple cosine distance, joined to the Within Class Covariance Normalization (WCCN) session compensation technique, was used for the scoring between training and testing i-vectors. Finally, in the perspective of the fusion, both i-vector-based speaker models and speaker id scores were computed at the segment and cluster level according to the speaker segmentation and clustering produced by the preliminary speaker diarization system.

3.2. Overlaid Person Name

In TV shows such as debates and news, the name of speakers is often overlaid the first time they contribute to the show, to make it easier to catch up when changing channels. Our overlaid person name (OPN) recognition module takes advantage of Optical Character Recognition to detect displayed text, transcribe it and locate names specifically identifying the person on screen.

The module is based on Video OCR technology developed at Orange Labs using a Character Recognition Convolutional Neural Network [14]. A set of rules on text box size, relative position, location and word content are tasked with identifying OPNs from other text (titles, topics, locations, time...) Since a single error in OCR would lead to a misidentification, captured names are normalized according to a large static list of people names from Wikipedia and knowledge bases, and dynamic lists gathered from newswires and websites dating from the same day as the processed show. The matching is performed in the FST framework by composing the OCR output with an automaton performing distortions (insertion, deletion, substitution) and a transducer converting character strings to full names from the lists. The output string of the shortest path in this composition leads to the normalized name. Names suffering high distortion costs are discarded.

3.3. ASR Name Spotting

The ASR Name Spotting system takes as input the automatic transcriptions of the speech segments provided by the speaker diarization process and several lists of person names selected as possible candidates for being speakers in the shows to process (the same lists as used for OPN recognition).

The spotting strategy is a three step process:

1. Search for full match hypotheses (firstname + lastname) in ASR transcription.
2. Search for partial match hypotheses, guessing from all possible compatible full names from the person lists.
3. Score each name hypothesis thanks to a phonetic alignment in the ASR confusion network.

At the end of this process we have a list of person name detections with time span and confidence scores.

4. Multimodal understanding and decision process

The main novelty for the 2014 Defi-Repere challenge is the development of a *multimodal understanding* component in charge of dealing with identification ambiguities. In our system, this understanding is performed at four levels: metadata, audio, overlaid image, video scene analysis. Several decision strategies, using all or part of these features have been developed and integrated into our primary submission to the challenge.

4.1. Multimodal understanding features

4.1.1. Metadata features

The *metadata* features includes all *a priori* knowledge that can be collected: the TV channel, the program and the broadcast date. The name lists mentioned in the previous section are considered as *metadata*. Each name in the person lists is associated, when available, with biographic features such as: date of birth and death, nationality, main related topic, profession, list of presence in TV shows from the training data. Finally, when available, we used the structure of the show in terms of "chapters" (anchor, studio, debate, talk show, report, ...).

4.1.2. Additional audio features

In addition to speaker ID hypotheses as presented in Section 3.1, speaker gender and speaker role labelling [5] are performed for every cluster of speakers.

4.1.3. Overlaid image features

Overlaid texts are extracted following the method described in Section 3.2. Moreover, since logos are very good markers for chaptering a given show, we performed logo detection and marked chapter transitions.

4.1.4. Additional visual features

Two types of visual features are used: the face and clothes color histograms and a characterization of the scene.

For features extracted form face and clothes, we assume that within a single show there is a bijection between people and their clothes. Therefore, after performing face detection using OpenCV's frontal and profile cascade detectors on every frame of the video, we cluster facetracks using the signature of clothes colors. The clothing area is estimated by taking a rectangle under the detected face proportional to its size, then the HSV color histogram determines the features vector and a cosine-based distance is used to measure the similarity between "clothtracks."

Scene characterization is an active research topic in the image processing community, as for example in sports and CCTV videos [15, 16, 17]. Following previous work on scene analysis in news video, we extract three types of descriptors at the visual level: type of shot (studio, report, composite, other), role of people on screen (anchor, journalist, invited, other) and show-specific camera identifiers when in studio. The system uses image-level HOG and RGB features for each shot and is trained with liblinear[2]. This system does not rely on face or person detection for determining roles but rather uses frame-level regularities to jointly identity all roles in a shot at once.

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

4.2. Decision process

The decision process is in charge of aggregating all previously described features in order to answer these two simple questions: who is talking and who can be seen in each frame of a video? The main source of information is the segment-level multimodal speaker identity presented in section 3. From these segments, the decision process performs the following actions:

1. Assign an identity to each speaker segment, either from speaker ID, OPN or name spotting.
2. Decide if the identified speaker for a given segment is visible on screen.
3. Decide if non-speaking persons are visible on screen, either during a speaker segment, or in a non-speech segment, and assign an identity to each of them.

All these actions are performed thanks to the following features:

- Metadata: show name, segment duration, number of name hypotheses, name gender, state (alive or dead), spoken language, topics, etc.
- Optical Character Recognition: confidence of OCR, OCR extended to speaker segment or speaker clusters, OCR extended to shot clusters.
- Speaker ID: appears in speaker-id 10-best, 1-best, speaker-id score, speaker gender and agreement with name gender, whether the name is chosen by the ASR-based speaker naming system, speaker role.
- ASR: time since last detection of name in transcript.
- Scene descriptors: type of shot, visual roles, camera id

Four different decision systems have setup to produce the primary submission of the PERCOL team which was ranked first at the *Defi-Repere* challenge 2014:

- S1: a rule-based decision system similar to the one used by the PERCOL team in the 2013 Defi-Repere challenge [3], with additional use of scene-level features.
- S2: in this decision system, the rules are replaced by a supervised decision process trained on the REPERE corpus [2]. Three classifiers are used to rerank names hypotheses: icsiboost [18], bonzaiboost with 3-level decision trees [19], and C4.5 8-level decision trees. The idea is that these classifiers can better model local interactions between features. At test time, the classifier is chosen at the show level according to its performance on a development set.
- S3: this decision system, described in [20], takes advantage of constraint propagation from the speaker modality to the head modality in order to seed an ILP-based head clustering system with multimodal cues.
- S4: a camera identification system which models the stage in order to retrieve which persons are being filmed from the camera angle and uses this scene information to propagate names.

For building the primary submission to the challenge, we chose on a development corpus the best system for each test condition (channel, kind of show, modality). The results obtained during this evaluation are presented in the next section.

5. Experiments

5.1. Official results

Systems participating in the *REPERE* challenge are evaluated in term of Estimated Global Error Rate (EGER), computed as the keyframe-level number of inserted or deleted names divided by the number of names in the reference [21]. Lower EGER means a better system. Evaluation results are given on the 2014-edition test set (phase2_test) for the speaker modality, the head modality and the official metric which is the mix of both modalities. In addition, systems are evaluated in two conditions: with biometric models and without biometric models.

Condition	Speaker	Head	Both
Biometric models	18.7	37.4	28.9
No biometric models	30.9	39.4	35.5

Table 1: EGER results for PERCOL's primary submission by modality and according to the availability of biometric models.

The results for PERCOL's primary system are listed in Table 1. It is remarkable that while biometric models have a large impact on the speaker identification system even though the number of covered speakers is limited, such benefit is less pronounced on the visual modality due to the effect of scene understanding descriptors which are not biometric models and therefore apply to both conditions.

Since the PERCOL's primary system is a selection of subsystems according to the show, Table 2 lists which subsystem was selected for each show according to its performance on a development set (marked by a stars), and the corresponding performance on the test set. This table shows that the selection of the best system for the speaker identification task is not always robust, probably due to differences between the development set and the test data.

		Speaker		Head		
Show	**Type**	**S1**	**S2**	**S1**	**S3**	**S4**
Story	news/deb.	18.6	**20.5***	**49.4***	50.0	-
Culture	tabloid	35.4	**34.4***	**77.5***	83.9	-
R&K	news/deb.	**25.3***	26.7	**44.2***	54.4	-
CVR	debate	15.8	**17.8***	42.5	54.0	**27.4***
EEL	debate	**16.1***	13.7	36.1	33.3	**20.8***
Actu	news	**12.0***	11.1	34.1	**31.8***	51.2
Info	news	**17.8***	16.8	50.6	**43.5***	43.4
P&F	debate	6.2	**7.1***	16.2	15.7	**8.9***
TopQ	political	**9.6***	9.6	**62.2***	69.2	-

Table 2: Show-level results for each subsystem by modality, with biometric models. Subsystems used in the primary system are denoted with a star (they were selected according to their performances on the development set). S1 is the rule-based system, S2 is the classifier-based system, S3 is the constraint-propagation system and S4 is the camera-id system.

5.2. Contrastive results

In this section, we highlight contrastive results that demonstrate the relevance of scene descriptors for multimodal person identification. Four subsets of features are studied: mono-modal only using speaker diarization and speaker ID, multi-modal without speaker ID (only using OCR and scene descriptors), multi-modal with scene features, and all features. The mono-modal result is obtained by performing speaker clustering and using speaker ID on each cluster. The multi-modal without speaker ID subset is the output of the S1 system (rule-based) without biometric models. The last two results are obtained with the S2 system (classifier) by reducing the features.

Results, detailed in Table 3, show that even though speaker ID features and OCR are the backbone of the system, it benefits from adding scene descriptors. In addition, this difference is intensified by only computing the performance for the speakers who are visible at the same time as they speak.

Condition	All speakers	Only talking heads
SpkID	35.2	36.0
OPN+scene	30.9	20.2
SpkID+OPN	21.2	14.7
SpkID+OPN+Scene	19.2	12.2

Table 3: EGER on the speaker modality according to the feature subset, on all instances or restricted to visible speakers.

Contrastively, Table 4 shows a similar break out according to feature sets for the head modality (predicting the name of all visible people). In this table, we look at using speaker ID to identify faces (essentially assuming that all people are talking every time we can see them), at naming faces using OCR-detected names only, and adding scene information to these features. This results in an improvement due to scene features, and shows the usefulness of speaker features for identifying faces. In particular, when results are restricted to people who talk at the same time as they are visible, speaker ID performance is much better, emphasizing the usefulness of speaker features in a multi-modal person identification system.

Condition	All heads	Only talking heads
SpkID	72.6	23.0
OPN	80.5	49.7
OPN+Scene	39.4	15.6
SpkID+OPN+Scene	37.4	13.1

Table 4: EGER on the head modality according to the feature subset, on all instances, or restricted to visible speaking faces.

6. Conclusion

This paper describes a multi-modal person recognition system for video broadcast developed for participating to the Defi-Repere challenge. We have presented in this paper the strategy followed by the PERCOL team enriching the speaker identification system by multi-modal features. Those features are based on analyzing and transcribing text overlay, recognizing the situation (TV set, report), the amount of people visible, the disposition of the TV set and even the camera when available. Encouraging results are obtained on the REPERE corpus, showing that video scene analysis provides good features for the speaker identification task in TV-broadcast.

Concerning the head identification task, using the speaker features in addition to the scene understanding allowed us to improve the head identification results (-2% in EGER), showing the usefulness of speaker information for identifying faces.

7. Acknowledgements

This work is funded by ANR under project PERCOL 2010-CORD-102-01.

8. References

[1] 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.

[2] H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V.-B. Le, A. Roy, C. Barras, S. Rosset, A. Sarkar, Q. Yang, H. Gao, A. Mignon, J. Verbeek, L. Besacier, G. Quénot, H. K. Ekenel, and R. Stiefelhagen, "QCompere @ REPERE 2013," in *SLAM*, 2013, pp. 49–54. [Online]. Available: http://ceur-ws.org/Vol-1012/paper-09.pdf

[3] B. Favre, G. Damnati, F. Bechet, M. Bendris, D. Charlet, R. Auguste, S. Ayache, B. Bigot, A. Delteil, R. Dufour, C. Fredouille, G. Linarès, J. Martinet, G. Senay, and P. Tirilly, "PERCOLI: a person identification system for the 2013 REPERE challenge," in *SLAM*, 2013, pp. 55–60. [Online]. Available: http://ceur-ws.org/Vol-1012/paper-10.pdf

[4] C. Guinaudeau, G. Gravier, and P. Sébillot, "Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation," *Computer Speech & Language*, vol. 26, no. 2, pp. 90–104, 2012.

[5] G. Damnati and D. Charlet, "Robust speaker turn role labeling of tv broadcast news shows," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5684–5687.

[6] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 2782–2795, 2013.

[7] E. Dumont and G. Quénot, "Automatic story segmentation for tv news video using multiple modalities," *International journal of digital multimedia broadcasting*, vol. 2012, 2012.

[8] X. Wang, L. Xie, B. Ma, E. S. Chng, and H. Li, "Modeling broadcast news prosody using conditional random fields for story segmentation," *Proc. APSIPA ASC*, pp. 253–256, 2010.

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

[10] 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*. IEEE, 2011, pp. 1–8.

[11] M. Bendris, D. Charlet, and G. Chollet, "Introduction of quality measures in audio-visual identity verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1913–1916.

[12] D. Charlet, C. Barras, and J. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," *ICASSP*, 2013.

[13] A. Larcher, J.-F. Bonastre, and B. e. a. Fauve, "Alize3.0 - open source toolkit for state-of-the-art speaker recognition," in *Interspeech 2013*, 2013.

[14] K. Elagouni, C. Garcia, F. Mamalet, and P. Sébillot, "Text recognition in multimedia documents: a study of two neural-based ocrs using and avoiding character segmentation," *International Journal on Document Analysis and Recognition (IJDAR)*, 2013.

[15] M. Bertini, A. Del Bimbo, and P. Pala, "Content-based indexing and retrieval of TV news," *Pattern Recogn. Lett.*, 2001.

[16] C.-C. Ko and W.-M. Xie, "News video segmentation and categorization techniques for content-demand browsing," *Image and Signal Processing, Congress on*, 2008.

[17] L. Chaisorn and T. seng Chua, "The segmentation and classification of story boundaries in news video," in *in IEEE International Conference on Multimedia and Expo*, 2002.

[18] B. Favre, D. Hakkani-Tür, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[19] C. Raymond, "Bonzaiboost," http://bonzaiboost.gforge.inria.fr, 2007.

[20] M. Bendris, B. Favre, D. Charlet, G. Damnati, and R. Auguste, "Multiple-view constrained clustering for unsupervised face identification in TV-broadcast," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[21] O. Galibert and J. Kahn, "The first official repere evaluation," in *First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013)*, 2013.