CRF based context modeling for person identification in broadcast videos

Paul Gay, Sylvain Meignier, Jean-Marc Odobez, Paul Deléglise

frontiers

CRF based context modeling for person identification in broadcast videos

Paul Gay [1,2], **Sylvain Meignier** [1], **Paul Deléglise** [1] **and Jean-Marc Odobez** [2]

[1] *LIUM Laboratory, Le Mans, France*
[2] *Idiap Research Institute, Martigny, Switzerland*

Correspondence*:
Jean-Marc Odobez
Idiap Research Institute, Centre du Parc, Rue Marconi 19, Case Postale 592,
CH-1920, Martigny, Switzerland, odobez@idiap.ch

ABSTRACT

We are investigating the problem of speaker and face identification in broadcast videos. Identification is performed by associating automatically extracted names from overlaid texts with speaker and face clusters. We aimed at exploiting the structure of news videos to solve name/cluster association ambiguities and clustering errors. The proposed approach combines iteratively two Conditional Random Fields (CRF). The first CRF performs the person diarization (joint temporal segmentation, clustering and association of voices and faces) jointly over the speech segments and the face tracks. It benefits from contextual information being extracted from the image backgrounds and the overlaid texts. The second CRF associates names with person clusters thanks to co-occurrence statistics. Experiments conducted on a recent and substantial public dataset containing reports and debates demonstrate the interest and complementarity of the different modeling steps and information sources: the use of these elements enables us to obtain better performances in clustering and identification, especially in studio scenes.

Keywords: Face identification, speaker identification, broadcast videos, conditional random field, face clustering, speaker diarization

1 INTRODUCTION

For the last two decades, researchers have been trying to create indexing and fast search and browsing tools capable of handling the growing amount of available video collections. Among the associated possibilities, person identification is an important one. Indeed, video contents can often be browsed through the appearances of their different actors. Moreover, the availability of each person intervention allows easier access to video structure elements such as the scene segmentation. Both motivations are especially verified in the case of news collections. The focus of this paper is therefore to develop a program able to identify persons in broadcast videos. That is, the program must be able to provide all temporal segments corresponding to each face and speaker.

Person identification can be supervised. A face and/or a speaker model of the queried person is then learned over manually labeled training data. However, this raises the problem of annotation cost. An unsupervised and complementary approach consists in using the naming information already present in the documents. Such resources include overlaid texts, speech transcripts and metadata. Motivated by this

Figure 1. Example frames from the REPERE corpus showing the variety of the visual conditions (pose, camera viewpoint, illumination) and the name face association challenges such as: multiface images (image **c**) and name propagation (from **a** to **b**). Images **a** and **c** show an example of OPNs.

opportunity, unsupervised identification has been investigated for 15 years from the early work of Satoh et al. (1999) to the development of more complex news browsing systems exploiting this paradigm (Jou et al. (2013)), or thanks to sponsored competitions (Giraudel et al. (2012)). Whatever the source of naming information, it must tackle two main obstacles: associate the names to co-occurring speech and face segments, and propagate this naming information from the co-occurring segments to the other segments of this person.

There are several challenges related to this task. First, the named entities need to be recognized and an association step must decide if the name corresponds to people co-occurring in the document. Ambiguities arise when multiple audiovisual (AV) segments co-occur with one name. This is illustrated in Fig 1c where there is more than one face in the image. This situation is becoming more common with modern video editing. Regarding the identity propagation, it can be done with speaker and face diarization techniques (detecting and clustering person interventions). However, these two tasks have been active fields of research for more than a decade and thus are difficult problems to solve. Indeed, a person may appear in different contexts thus introducing huge intraperson variabilities. We can distinguish them in function of the modalities and the different types of videos. For the speaker diarization, the main challenge in broadcast news is background noise such as music, or a noisy environment during outside reports. If we consider debates in studio where the speech is more spontaneous, the bottleneck becomes the overlapping speech and short speech segments. Regarding face diarization, report videos usually exhibit the largest variations as location and time may change between two scenes, and so will be the illumination conditions. For the debate and studio scenes, variations come essentially from changes in the facial poses.

In this paper, we assume that closed-captions are not available as this is the case in European medias. Instead, we focus on Overlaid Person Names (OPNs) which are used to introduce the speakers as illustrated in Fig. 1a. Such names are appealing since their extraction is much more reliable than pronounced names obtained through Automatic Speech Recognition (ASR). Moreover, their association with face or speech segments is in general easier than analyzing whether pronounced names in ASR transcripts refer to people appearing in the video. The identification systems submitted at the recent REPERE campaign (Bredin et al. (2013); Bechet et al. (2014); Poignant et al. (2014)) mainly rely on such names.

Our approach offers several advantages. Faces are identified by alternating between a clustering step of faces and audio speech segments and a naming step of the resulting AV clusters. Each step is performed by a dedicated CRF. The use of CRF enables us to include heterogeneous context cues in our modeling. The use of such cues is challenging because they must use as little specific prior information as possible in order to

achieve generalization over the different types of videos. In this paper, we include different generic context cues. First, we have AV association scores which enable to associate overlapping speaker and face segments when they correspond to the same person. Then, we use uniqueness constraints between simultaneously appearing pairs of faces. Furthermore, one of the main contribution is a background recurrence descriptor which attributes a soft role to each segment. It enables to distinguish the persons which are announced by the OPNs such as guests or journalists from the anonymous persons appearing around them. Last but not least, the names contained in the OPNs are included to guide the clustering by using the probabilities obtained with the naming CRF. These different cues enable to improve the clustering by reducing errors due to monomodal intracluster variations such as facial pose or audio background noise. Eventually, the CRF formulation avoids hard local decisions by providing a joint probability distribution over all the segments.

The first CRF performs jointly the clustering of face tracks and speaker segments thanks to AV association as introduced in Gay et al. (2014c). In practice, AV association is initialized in a pre-processing step based on temporal co-occurrence and then refined inside the CRF thanks to talking head detection scores and the previously described contextual cues. The second CRF assigns a name to each cluster by using co-occurrence statistics and a uniqueness constraint preventing any two faces on the same image to receive the same name. In Gay et al. (2014b), this approach was designed for face identification. In the present case, we extend this approach for the AV case and provide results for the final evaluation of the REPERE campaign. Identification performances are discussed by investigating the algorithm behavior in different types of shows (reports, news, debates, celebrity magazines) and the relations with the clustering quality.

The rest of the article is organized as follows: Section 2 reviews related work on unsupervised identification. Then, Section 3 presents the proposed CRF-based system. Experiments and results are presented in section 4. Finally, Section 5 sums up our main findings and concludes the paper.

2 RELATED WORK

As stated in the introduction, unsupervised people identification must address the problems of local person/name association and propagation to the video parts where the names are absent. The association is conducted via the use of co-occurrence statistics between the names present in the document and the detected persons. The propagation can be seen as a clustering problem. Clustering methods can regularly benefit from new improvements in speaker and face representations. At the time of writing, the ivector approach is one of the most successful (Rouvier et al. (2013)) for the speaker diarization task. Regarding face representation, recent advances include encodings (Simonyan et al. (2013)), metric learning (Bhattarai et al. (2014)) and feature learning by deep Convolutionnal Neural Networks (Schroff et al. (2015)). However, most of the systems require explicit face alignment to obtain frontal views which is not always feasible. The work published in Zhang et al. (2015) suggests that using only face representation is a great limitation when dealing with unconstrained views of persons. For this reason, we believe that investigation into context-assisted clustering is justified, especially for broadcast news videos which exhibit a strong structure.

To identify the faces, most approaches try to solve the association and the propagation problems jointly. On one hand, co-occurrence statistics at cluster level are more discriminant and accurate than just describing a face locally with namedness features (like face position or talking activity) to assess whether the detected name should be associated. On the other hand, name/face co-occurrences are used as a contextual cue to improve the face clustering process. These principles have been used intensively since the seminal works of Berg et al. (2004) and Everingham et al. (2006) which applied to two representative use-cases: captioned images, as examplified by the *Yahoo News!* dataset, and soap series with the buffy dataset. The first case

study consists of news articles with images illustrating the subject. The initial approach described in Berg et al. (2004) is an EM clustering where the update of the model parameters takes into account the name/face co-occurrences. In this context, the work of Ozkan and Duygulu (2010) exploits the fact that a textual query enables to retrieve faces where the queried person holds the majority. The problem of finding those faces is posed as finding the densest component in a graph. This idea was later extented in Guillaumin et al. (2010) where the distance within clusters is minimized with respect to a cannot-link constraint which implies that two faces must belong to different clusters if their captions contain different names. However, those co-occurrence statistics can fail when group of people co-occur in a similar fashion, a situation commonly encountered in TV programs. In soap series, the names of the speakers can be obtained with the transcripts and the subtitles. Works in Cour et al. (2011); Wohlhart et al. (2011); Bauml et al. (2013) use those names as weak labels to improve supervised classifiers. They choose a learning setting which takes into account the label ambiguities, for example: multiple instance learning (Wohlhart et al. (2011)) and semi-supervised strategies (Bauml et al. (2013)). Talking head detection (Everingham et al. (2006); Cour et al. (2011)) and dialogue cues (Cour et al. (2010)) are also used to solve the ambiguities in the face/name association. Note that in the previous two case studies, the naming co-occurrence statistics are quite different to those in broadcast videos where the OPNs are more sporadic. Indeed, the OPN of a given person only appears one or a few times (usually for the first time utterance). This scarsity increases the dependence of the identification performance on the clustering quality.

Originally, unsupervised speaker identification in broadcast news was conducted by first performing a speaker diarization (i.e. clustering) step of the audio track and then assigning the names extracted from the transcription to the speaker clusters by using semantic classification trees (Jousse et al. (2009)) or Maximum-entropy classifiers (Ma et al. (2007)). More recently, the idea of constrained speaker clustering has been exploited in Bredin and Poignant (2013); Poignant et al. (2014). The system described in Bredin and Poignant (2013) defines a graph where the nodes are speaker segments and OPNs. OPNs are used to express must-link and cannot-link constraints between the utterances. The clustering and the naming of those segments is done using an Integer Linear Programming formulation. As first investigated by Li et al. (2001), the case study of videos allows to exploit the complementarity of audio and video modalities. AV cues such as talking head detection scores can be used to match faces and speakers and to improve the monomodal speaker and face diarizations. The scores of such cues are computed by estimating motion in the region of the lips. In addition, features such as the face size, the face position or the number of faces in the image are extracted and given to a supervised classifier (El Khoury et al. (2012); Vallet et al. (2013)) to further refine the talking assessment. In order to bring corrections to the initial monomodal diarizations, the talking head detection scores should be reliable where monomodal errors are present. Moreover, the audio and video will also be more complementary if they make errors at different moments. In other words, the improvements of the AV diarization depend on the performances of the initial monomodal ones. The work of Noulas et al. (2012) integrates faces and speech segments in a factorial hidden markov model. The assignment of a segment to a cluster label is based on biometric model and on AV links with co-occurrent segments from the other modality. The use of a graphical model enables to express dependences between variables with a global probabilistic formulation which can then be optimized jointly. In order to jointly identify faces and speakers, the authors of Poignant et al. (2015) proposed a constrained multimodal clustering. They use the simple idea that two segments which co-occur with different names implies that they should be assigned to different clusters. The authors also showed that their multimodal clustering of faces and speakers can make use of talking head detection scores to correct errors present in the monomodal systems.

The work of Bechet et al. (2014), an interesting yet not detailed contribution to the field, reports the intensive use of multimodal scene understanding cues. First, speaker diarization is performed and speakers are identified using OPNs or pre-trained models. Then, identities are propagated from the speakers to the faces. Scene segmentation, role detection, and pre-trained visual models for each TV set (and sometimes for each camera) are used to indicate how many faces are present on screen and what their roles are. Such a fine-grain modeling enables them to report the best identification on the REPERE campaign. Indeed, it permits to tell which persons are present without detecting the faces by detecting the specific shot (up to which studio camera is used). Thus, profile views and persons seen from the back can be identified. However, to learn those models, manual annotations have been made for each show. This poses the problem of human labor cost and lack of generalization. More generally, several researchers focus on exploiting the context surrounding the faces. The work in Zhang et al. (2013) uses clothes, image background, cluster co-occurrences and attribute classifiers and Tapaswi et al. (2014) build must-link and cannot-link constraints deduced from shot threads (sequence of shots obtained from the same camera angle).

Contributions: in this paper, we leverage on different contextual cues present in the state of the art, introduce new ones, and include them in our CRF model. First, instead of conducting speaker and face clustering separately [15, 5], we perform a joint clustering of face tracks and speaker segments which also benefits from the OPNs information. To be more precise, we compute Local Face visual Backgrounds (LFBs) around each face track and cluster them. This provides us with a signature for each face track characterizing the level of recurrence of its LFB in the data. Intuitively, recurrent a LFB correspond to people who are important and can be seen as a soft role assignment distinguishing faces to be named from faces of figurative people. Concretely, it enables to encourage faces tracks with recurrent LFBs to join named clusters, i.e. overlapping an OPN. Secondly, a naming CRF performs the joint identification of all person clusters, thus allowing to account for uniqueness constraints and co-occurrence statistics between clusters and OPNs. Unlike previous works which rely on extensive annotations (Bechet et al. (2014)), those elements of context have better generalization capabilities, since we can learn one single model over a large and diversified corpus, and require less annotations if we want to learn a new type of show. Thanks to the flexibility of the CRF formulation, new contextual cues could be added in the future to further improve the performances.

3 METHOD

The method will be first described globally in section 3.1. In the section 3.2, we introduce the notations. We then describe how we extract LFB and AV association features in sections 3.3 and 3.4. In section 3.5, the diarization CRF which clusters face and speech segments is presented, and in section 3.6 the naming CRF which is in charge of identifying the clusters. To conclude this part, we describe how the full system is used and optimized in section 3.7.

3.1 Method overview

The general approach is summarized in Fig 2. First, the different modalities are processed separately: monomodal speaker (Rouvier et al. (2013)) and face (Khoury et al. (2013)) diarizations are performed, LFBs are extracted around each face and clustered, Optical Character Recognition (OCR) is performed to extract the overlaid texts (Chen and Odobez (2005)) and named entities are detected (Gay et al. (2014a)).

In the second part, we perform the AV clustering and the naming of the persons. Initially, we use the Hungarian algorithm to associate face and speaker clusters based on their temporal overlap. Naming

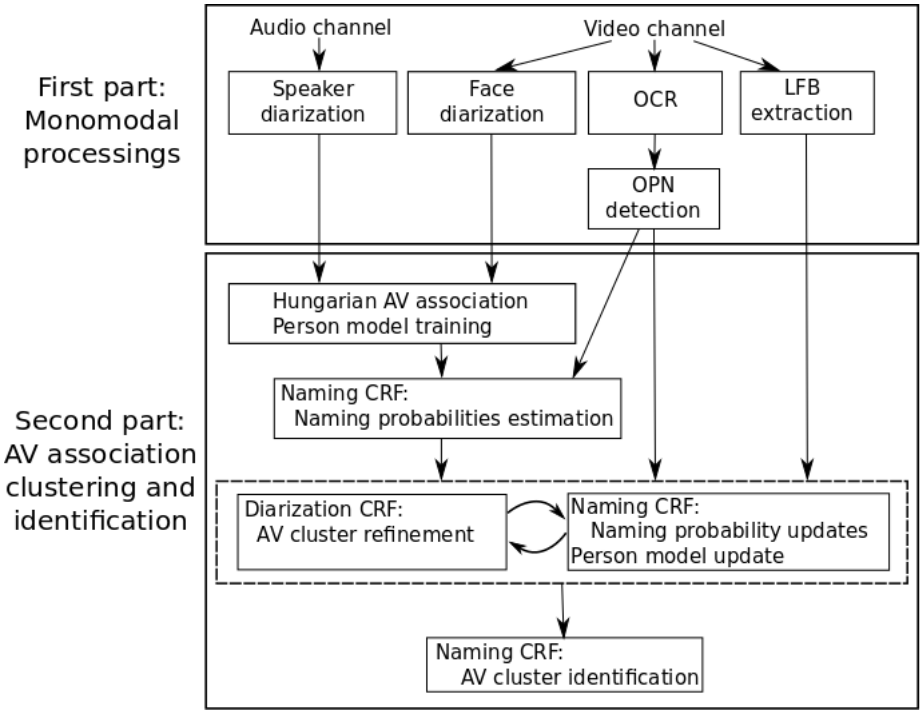

Figure 2. Overview of the system. First, face tracks and speech utterances are detected and clustered separately. They are then used with the OPNs in an iterative combination of the two CRFs. A first one to refine the diarization, and the other one to identify the clusters. The latter is eventually used for the final cluster identification.

probabilities are then computed onto those AV clusters with the naming CRF. Lastly, the system iterates over a clustering step and a naming step. In the clustering step, the diarization CRF infers a cluster label for each face track and utterance given the naming probabilities, an acoustic and visual person model for each cluster label, and various context clues including the LFBs. In the naming step, person models and naming probabilities are updated as a result of the new diarization. The motivation factor being that the diarization CRF is able to use contextual clues to correct potential clustering errors made by the monomodal diarizations and thus improve the final identification. Lastly, a name is associated to each cluster with the naming CRF.

3.2 Notations

The pre-processing includes obtaining initial monomodal face and speaker clusters, a set of OPNs and extracting the features from those elements. First, faces are detected (Viola and Jones (2004)) and tracked within each shot, resulting in a set of face tracks denoted as $V = \{V_i, i = 1 \dots N^V\}$. Each face track V_i is characterized by a set of visual features x_i^{surf} (set of Speeded-Up-Robust Features (SURF) extracted in up to 9 images of the face track (El Khoury et al. (2010))) and a set of boolean features $\{x_i^{\text{lfbv}}(k), k \in K\}$ indicating whether V_i corresponds to a recurrent LFB as explained in the next section 3.3.

Second, OCR (Chen and Odobez (2005)) and Named Entity Detection techniques based on string matching against external resources (predefined lists, freebase database, Google hits...) are applied as described in Gay et al. (2014a) to extract the set $O = \{O_i, i = 1 \dots N^O\}$ of OPNs. Each OPN O_i is characterized by its duration d_i^{opn} and its name $x_i^{\text{opn}} \in M$ where $M = \{n_j, j = 1 \dots N^M\}$ denotes the set of unique names extracted from the video.

Figure 3. Left: predefined spatial template for the background face selection area given a head detection. Center and right: two examples with head detections and the selected background areas.

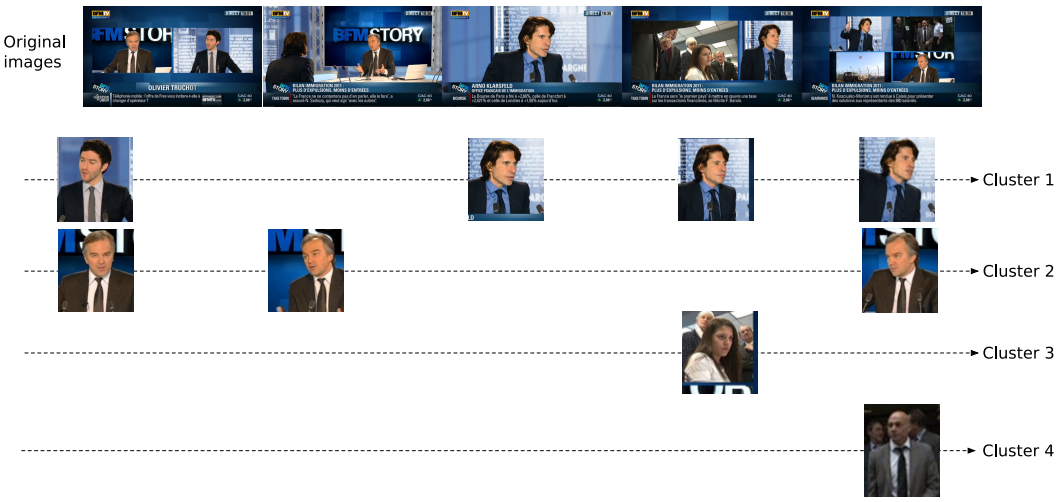

Figure 4. Top: original images. Bottom: clusters of recurrent local face backgrounds (LFB) automatically cropped from each image and clustered given the localised faces. Each row corresponds to a local background cluster. Clusters 1 and 2 are recurrent backgrounds and correspond to speakers. Clusters 3 and 4 are not recurrent. They contain non-speaking faces which appear occasionally in the video.

Finally, the audio stream is segmented into a set $A = \{A_i, i = 1 \dots N^A\}$ of continuous speech segments called utterances, each described by a set of acoustic features x_i^a. Features are 12 MFCCs with first order derivatives. Each frame is normalized with a short-term windowed mean and variance. Feature warping is also applied. In addition, a set of boolean features $\{x_i^{\text{lfba}}(k), k \in K\}$ is extracted indicating whether A_i is co-occurring with a recurrent LFB, as described in the next section. Finally, talking head detection features $x_{\text{ij}}^{\text{av}}$ are extracted between each couple (A_i, V_j) with a non-zero overlap as described in section 3.4.

3.3 Local Face Background recurrence

We want to capture whether a face appears with a recurrent visual background. This feature will be included in the diarization CRF. To this end, we focus on an area around each face track V_i to capture the background context of this face. We do not consider full images as the same image might include different face visual contexts (see the first, fourth and fifth images from the left in the top row of Fig 4). Instead, we select a rectangle area around each face as local face background (LFB) representative by following a predefined spatial template between the face and this rectangle as can be seen in Fig 3. In practice, the fixed proportions were chosen manually so as to avoid a potential overlap with other parts of the images in typical edited videos like in the 4th and 5th images from the left on the top of Fig 4. We then characterize

each obtained rectangle area with SURF features and, in order to cluster them, we use a hierarchical clustering approach (El Khoury et al. (2010)). Then, we set $x_i^{\text{lfbv}}(k)$ to true if face track V_i belongs to a local background cluster whose number of elements is higher than k. In practice, multiple values of k can be used to characterize different levels of recurrence and reduce the importance of the stopping criterion of the hierarchical clustering. Fig 4 shows examples of obtained recurrent and non-recurrent patterns.

3.4 Talking head detection features

In order to integrate AV association information in the CRF, we detect talking heads. To characterize talking heads, we use the following measures. These features are extracted for each overlapping utterance/face track couple and include:

- **lip activity:** the lip activity of a given face at frame k is computed as described in El Khoury et al. (2012) and consists in the mean intensity difference between frame k and $k+1$ after local image registration in predefined regions corresponding to the lips. In addition, we focus on the relative lip activity by dividing by the sum of all the lip activities measured from all people in the image.
- **Head size:** the interest of this feature relies on the hypothesis that the face of the speaker is usually larger than the faces of other people in the image. Put simply, we take the diagonal size of the detection bounding boxes. We also use the relative head size.

The previous features are computed from each frame of the face track. Eventually, the final feature $x_{\text{ij}}^{\text{av}}$ is an average over all values from the frames included in the overlap between the utterance A_i and the face track V_j. This corresponds to the method used in Gay et al. (2014c). To assess whether a couple of utterance/face track corresponds to a talking head given the features, we use an SVM with gaussian kernel denoted as h.

3.5 Audio-visual (AV) person diarization CRF

The clustering of face tracks and utterances defines itself by estimating the label field $E^d = \{e_i^a, i = 1 \dots N^A, e_j^v, j = 1 \dots N^V\}$ as such, the same person index is used for e_i^a and e_j^v when the utterance A_i and the face track V_j correspond to the same person. The labels e_i^a and e_j^v take value in the set of possible person indices denoted as P. To achieve this, let G be an undirected graph over the set of random variables A, V, O, and E^d. We then seek to maximize the CRF posterior probability formulated as:

$$P(E^d|A,V,O) = \frac{1}{Z(A,V,O)} \times exp\left\{ \sum_{i \in \mathcal{F}} \sum_{c \in G_i} \lambda_i f_i(A_c, V_c, O_c, E_c^d) \right\} \quad (1)$$

where each triplet (f_i, G_i, λ_i) is composed of a feature function f_i, a weight λ_i learned at training time and the set G_i of cliques where this function is defined. (A_c, V_c, O_c, E_c) denotes the set of nodes contained in the clique c. $\mathcal{F}$ is a set of abstract functions indices. We use 6 types of feature functions which will be described in the next sections. A graphical representation of this model is illustrated on Fig 5.

The association function f_{av} favors the association of talking heads to utterances. The function is defined on all overlapping utterance/face track couples $\{(i,j)/t(A_i,V_j) \neq 0\}$ where $t(A_i,V_j)$ is the overlapping time duration between segments A_i and V_j:

$$f_{\text{av}}(A_i, V_j, e_i^a, e_j^v) = \begin{cases} t(A_i,V_j)h(x_{\text{ij}}^{\text{av}}) & \text{if } e_i^a = e_j^v \\ -t(A_i,V_j)h(x_{\text{ij}}^{\text{av}}) & \text{otherwise} \end{cases} \quad (2)$$

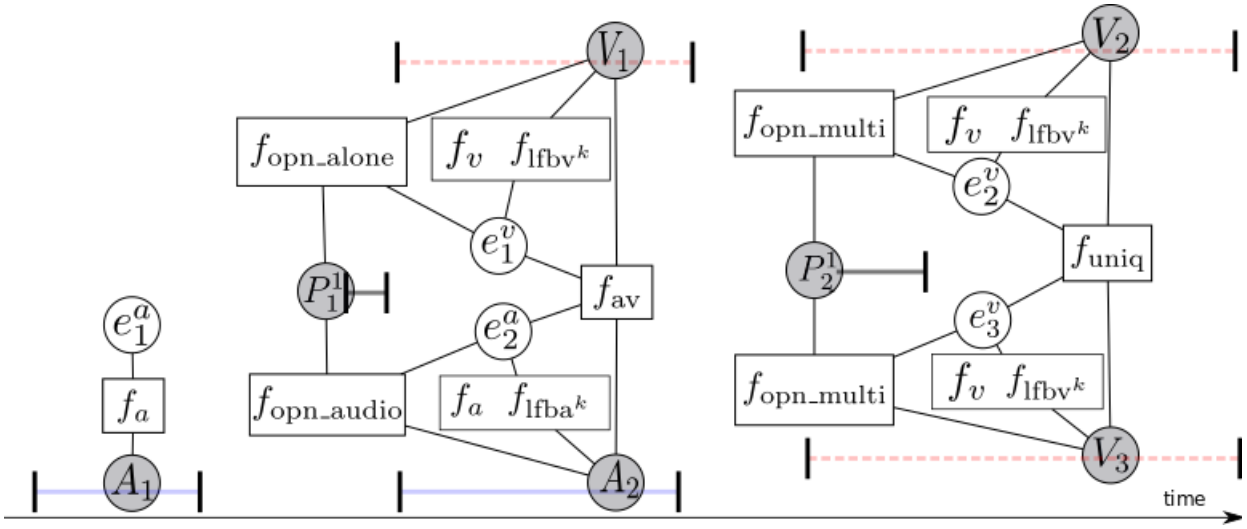

Figure 5. Factor graph illustrating the diarization CRF using talking head information (f_{av}) and the context from the OPNs (f_{opn}, f_{lfb}). The blank circle nodes correspond to hidden variables, the shaded circle nodes correspond to observations while the squares represent the feature functions. The x-axis represents time and the drawing shows also segments corresponding to the time intervals during which a specific observation (track, utterance, OPN) occurs. Red dot segments illustrate the face track temporal segments while blue plain segments the utterances.

where $h(x_{\text{ij}}^{\text{av}})$ represents the binary output of the SVM classifier introduced in section 3.4. It corresponds to 1 when the face and the speaker correspond to the same person and -1 otherwise. We chose a SVM classifier since it shows good results in El Khoury et al. (2012); Vallet et al. (2013). Other techniques could be employed but we leave this problem for future research.

The visual feature function $f_v(V_i, e_i^v)$, defined for all face tracks $V_i \in V$, indicates how likely the visual features x_i^{surf} of V_i should be labeled with the person index e_i^v. This is a face modeling task in which for each label e_i, we need to define a visual model that is learned from the data currently associated to the label. Practically, f_v computes as score between V_i and a label e_i^v the 10th percentile SURF vector distances between x_i^{surf} and all the SURF features of the current face tracks associated with this label. The distance between two face tracks is computed following (El Khoury et al. (2010)). Although the use of SURF features could be discussed regarding other more modern representations, we observe that their matching power is useful for similar faces of the same person viewed from a similar view point. The previous work in Gay et al. (2014b) uses an average of the distances. By using the percentile, we found a slight improvement for the diarization task (0.2 points on the development REPERE corpus). We believe that the use of a percentile instead of averaging enables to merge 2 clusters of the same identity but containing samples whose poses are dominantly from different poses.

The acoustic function $f_a(A_i, e_i^a)$, defined over all utterances $A_i \in A$, is the audio equivalent of f_v. We chose a 512 GMM-UBM with diagonal covariance following Ben et al. (2004). We did not use ivectors since we might need to learn a model on small clusters containing only a few seconds of speech. $f_a(A_i, e_i^a)$ computes the likelihood score of the features x_i^a given the GMM model learned over the data currently associated to the cluster label e_i^a.

The LFB feature function is driven by the assumption that faces inside a recurrent LFB are likely to correspond to a speaker announced by an OPN. To favor face tracks identified as recurrent LFB to join a

person cluster which could be named, we define the following feature function. For each face track V_i,

$$f_{\text{lfbv}^k}(V_i, e_i^v) = \begin{cases} 1 & \text{if } x_i^{\text{lfbv}}(k) \text{ and } e_i^v \in \mathcal{E}^{\text{opn}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{E}^{\text{opn}}$ is the set of person clusters indices co-occurring with an OPN, i.e. the set of clusters which are currently associated with a name.

This principle is extended to each utterance A_i with the function f_{lfba^k} which employs the feature $x_i^{\text{lfba}}(k)$. To this end, we assume that the utterances co-occurring with a recurrent LFB should be assigned a cluster label from the set $\mathcal{E}^{\text{opn}}$. Thus, as discussed in section 3.3, $x_i^{\text{lfba}}(k)$ is set to true if utterance A_i is overlapping with a face track V_j such that $x_j^{\text{lfbv}}(k)$ is true. We then introduce the same function as in the video case:

$$f_{\text{lfba}^k}(A_i, e_i^a) = \begin{cases} 1 & \text{if } x_i^{\text{lfba}}(k) \text{ and } e_i^a \in \mathcal{E}^{opn} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Interestingly, these functions act as a namedness feature (Pham et al. (2008)) in the sense that they favor the naming of the corresponding face tracks and utterances. They also softly constrain the number of clusters. In other words, the clusters whose labels belong to $\mathcal{E}^{\text{opn}}$ will attract the segments identified as recurrent LFB. Note that if the constraint was strictly enforced, each concerned audio or visual segment would only be assigned to a member of $\mathcal{E}^{\text{opn}}$.

The OPN feature functions bring a special treatment to the segments co-occurring with OPNs. The idea is to favor segments (face tracks or utterances) co-occurring with an OPN O_j to be assigned to a person cluster likely to be labeled with the name x_j^{opn}. Thus, we define:

$$f_{\text{opn\_alone}}(V_i, O_j, e_i^v) = \begin{cases} p(e_i^c = x_j^{\text{opn}} | C, P) & \text{if } V_i \text{ is alone in the image and co-occurs with OPN } O_j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $p(e_i^c = x_j^{\text{opn}} | C, P)$ is the probability that the name contained in the OPN O_j corresponds to the cluster label e_i^v given the clustering C and the set of OPNs P. Here, we denote as e_i^c the naming label of cluster label e_i^v. This probability is computed with the naming CRF as defined in section 3.6.

Similarly, we use $f_{\text{opn\_multi}}$ if V_i co-occurs with other faces:

$$f_{\text{opn\_multi}}(V_i, O_j, e_i^v) = \begin{cases} p(e_i^c = x_j^{\text{opn}} | C, P) & \text{if } V_i \text{ co-occurs with OPN } O_j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

We also define $f_{\text{opn\_audio}}$ for each co-occurring couple (A_i, O_j):

$$f_{\text{opn\_audio}}(A_i, O_j, e_i^a) = \begin{cases} p(e_i^c = x_j^{\text{opn}} | C, P) & \text{if } A_i \text{ co-occurs with OPN } O_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Differentiating these 3 cases enables to learn specific λ weights so that the model behavior is adapted to each situation.

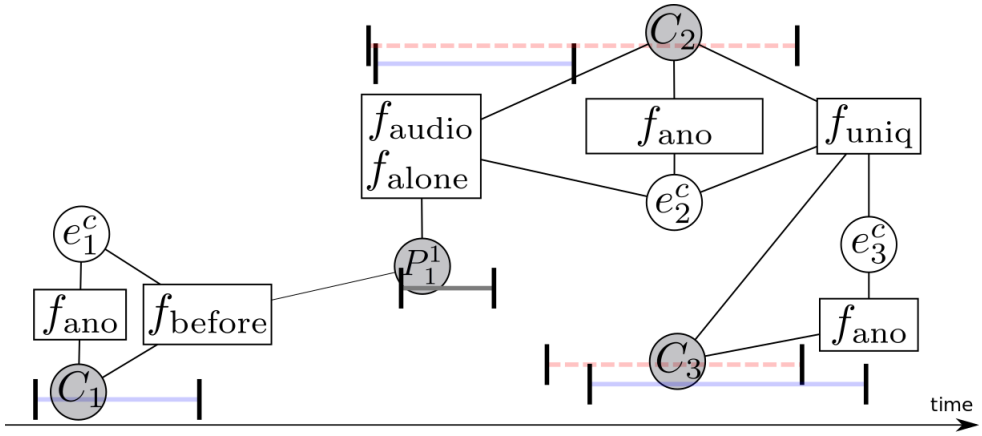

Figure 6. Factor graph illustrating the naming CRF using the co-occurrence functions with the OPNs and the uniqueness constraint. The conventions are the same as in Fig 5.

The uniqueness feature function ensures two faces that co-occur in the same shot to have different labels (Berg et al. (2004); Pham et al. (2013)). For such a pair V_i, V_j:

$$f_{\text{uniq}}(V_i, V_j, e_i^v, e_j^v) = \begin{cases} -\text{Inf} & \text{if } e_i^v = e_j^v \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

It is crucial to use this function because due to the OPN feature functions, multiple faces co-occurring with the same OPN will tend to be assigned to the same person cluster.

3.6 Cluster identification

The previous diarization CRF provides us a set of AV person clusters $C = \{C_i, i = 1 \dots N^C\}$. Thus, in the naming step, the goal incorporates estimating the label field $E^N = \{e_i^c, i = 1 \dots N^C\}$ such that the label e_i^c corresponds to the name of the cluster C_i. The label e_i^c takes value in the set of names M augmented by an anonymous label which should be assigned to anonymous persons. For this naming CRF, the posterior probability uses 6 feature functions:

$$P(E^N|C, O) = \frac{1}{Z(C, O)} \times exp\left\{ \sum_{i=1}^{6} \sum_{c \in G_i} \lambda_i f_i(E_c^N, C_c, O_c) \right\} \tag{9}$$

Fig 6 represents an illustration of this. This naming model exploits four different co-occurrence statistics between clusters and OPNs. The first function f_{alone} is defined over each triplet (e_i^c, C_i, O_j), where the OPN O_j must co-occur with a face track which belongs to C_i and which is alone in the image. Let us denote as $\delta(C_i, O_j)$ the co-occurring time between the face tracks which occurs alone in the cluster C_i and O_j. Then, we have:

$$f_{\text{alone}}(e_i^c, C_i, O_j) = \begin{cases} \frac{\delta(C_i, O_j)}{d_j^{\text{opn}}} & \text{if } x_j^{\text{opn}} = e_i^c \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

As for the OPN diarization model components, we define similarly two other functions f_{multi} and f_{audio} which measure the overlapping time between O_j and the face tracks of C_i which occur with other faces

on one hand, and with the audio segments of C_i on the other hand. Moreover, we exploit the assumption that a person does not usually appear or speak before the first apparition of his name in an OPN to define $f_{\text{before}}(e_i^c, C_i, O_j)$, which returns the number of audio segments from cluster C_i that occur before the first apparition of the name x_j^{opn} associated to the OPN O_j.

$$f_{\text{before}}(e_i^c, C_i, O_j) = \begin{cases} \sharp\{A_i \in C_i, \text{end}(A_i) < \text{start}(O_j)\} & \text{if } x_j^{\text{opn}} = e_i^c \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

We also introduce prior knowledge over the anonymous label by defining a fifth feature function $f_{\text{ano}}(e_i^c, C_i)$ which returns 1 if e_i^c is the anonymous label. When applied, it allows the model to penalize the fact of not identifying a person and improves the recall.

Lastly, we define a uniqueness function $f_{\text{uniq}}(e_i^c, C_i, e_j^c, C_j)$ over visually overlapping clusters just as in the diarization step. For each cluster pair (C_i, C_j) with overlapping face tracks:

$$f_{\text{uniq}}(e_i^c, C_i, e_j^c, C_j) = \begin{cases} -\infty & \text{if } e_i^c = e_j^c \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

3.7 Optimization

The joint use of the two CRFs is conducted by applying the following steps: i) the diarization labels are firstly initialized by separately performing audio and video clustering and then associating the clusters to obtain the potential AV person labels P (audio and face cluster couples). The association is conducted using the Hungarian algorithm (Kuhn (1955)) where the cost for a cluster couple is defined as the sum of the scores from the function f_{av} over all its utterance/face track pairs. ii) For each resulting person label p_i, biometric models are learned from their associated data and naming probabilities for each label are estimated by using the naming CRF. iii) Given these models, we run the loopy belief propagation inference to get the most probable diarization labels E^d by solving $E^d = \arg\max_{E^d} P(E^d|A, V, O)$.

Eventually, Steps ii) and iii) are iterated in a Expectation-Maximization style by alternating model updates and inference. Ideally, one would iterate until convergence, i.e. when the label for each segment becomes stable. In practice, as there is no guarantee that the algorithm converges, a fixed number of iterations is tuned over the development set since we observe only small modifications after a few iterations.

The computational bottleneck with the Loopy Belief Propagation algorithm occurs in the presence of big cliques. This is the case, in our graphs, when the uniqueness constraint is applied to images where there are more than 20 faces. In such cases, uniqueness constraints can be dropped from the graph during inference and enforced in a post-processing step.

4 RESULTS AND DISCUSSION

This section will firstly present our experimental set-up: the corpus (section 4.1), implementation details (section 4.2) and the metrics used for the evaluation (section 4.3). Then in section 4.4, we present our results showing identification and clustering performances in function of the different parts of the model.

4.1 Corpus description

We used the REPERE corpus (Giraudel et al. (2012)) for our experiments. It involves broadcast data videos containing 4 main types of shows: i) debates in indoor studio (Fig 1a,b); ii) modern format

information shows which contain reports and interviews with dynamic picture compositions (Fig 1c,d); iii) extracts from parliamentary sessions "Questions to the government" (Fig 1e) iv); celebrity news (Fig 1f).

We evaluate our approach on the final test set which contains 37 hours during which 10 are annotated. A development set is used to optimize the number of LFB functions and the number of iterations between the two CRFs. It consists of 28 hours among which 6 are annotated. The SVM h used in the f_{av} feature function and the CRF parameters are learned on the test set of the first REPERE evaluation composed of 3 hours of annotated data.

4.2 Parameter settings and algorithm details

We set the K value to $\{3, 4, 5\}$ for the LFB feature functions. We set the number of iterations between the two CRFs to 3 as we noticed that no major changes usually occur after that point. It is important to note that these CRF parameters are learned on automatic detections and automatic clusters and not on cleanly segmented ones. Therefore, it enables us to take into account the noise present at test time. We use the GRMM toolbox (McCallum" (2002)) for the CRF implementation.

The initial speaker diarization system is the LiumSpkDiarization toolbox[1] which combines ivector representation and ILP clustering (Rouvier et al. (2013)). It has achieved state-of-the-art results in several speaker diarization benchmarks (Rouvier and Meignier (2012)). The initial face diarization uses the system described in (Khoury et al. (2013)) which combines SURF based distances and DCT features whose distribution is modeled with GMMs. This system has been evaluated on the public Buffy dataset (Cinbis et al. (2011)) and compares favorably to other metric learning methods. The use of state-of-the-art systems enables us to verify that our CRF is able to correct errors which are proven difficult to solve in the monomodal case.

4.3 Performance measures

The overall identification performance is measured with the Estimated Global Error Rate (EGER) which is the REPERE evaluation metric. It is defined as follows:

$$\text{EGER} = \frac{\sharp \text{ conf} + \sharp \text{ miss} + \sharp \text{ false}}{\sharp \text{ total}} \tag{13}$$

where $\sharp$ conf is the number of wrongly identified persons, $\sharp$ miss, the number of missed persons, $\sharp$ false, the number of false alarms and $\sharp$ total, the total number of persons to be detected. It should be noted that the metric ignores the spatial position of the faces and simply uses a person list for each annotated image. The behavior of this metric is illustrated on figure 7. Wrong predictions are counted as false alarms only if the number of predictions exceeds the number of persons in the annotation. Otherwise, they are counted as confusions. Similarly, missing persons are reported only if the number of predictions is smaller than the number of persons.

We also use the clustering error rate (CER) to study the correlation between clustering and identification performances as our work is motivated by an interdependence between those two tasks. Initially, the CER has been introduced for the speaker clustering task (NIST (2003)) and is defined as:

$$\text{CER} = \frac{\sum_{\text{seg}\in\text{Segs}} \text{dur(seg)}(\min\,(N_{\text{Ref}}(\text{seg}),N_{\text{Sys}}(\text{seg}))\text{-}N_{\text{Correct}}(\text{seg}))}{\sum_{\text{seg}\in\text{Segs}} \text{dur(seg)}N_{\text{ref}}(\text{seg})} \tag{14}$$

[1] http://www-lium.univ-lemans.fr/diarization

```
Show: BFMStory_12 frame: 4312
Head Ref: Barack_OBAMA Augusta_ADA_KING
Head Hyp: Augusta_ADA_KING David_HAMILTON Alan_TURING
```

Figure 7. Extract of an evaluation file for face identification. The second row is the reference name list and the third row is the predicted list. Augusta_ADA_KING will be counted as correct. One of the two remaining names will be counted as confusion with Barack_OBAMA, and the third one will be a false alarm. Since there are 2 persons in the reference and the system made 2 errors, the corresponding EGER of this example is 1.

Speaker diarization results

	Initial monomodal	CRF Dia	CRF Dia without OPNs
News	6.9%	7.0%	**6.8%**
Debates	6.6%	**4.0%**	6.5%
Parliament	6.9%	**5.0%**	9.5%
Celebrity	**14.6%**	15.1%	**14.6%**
All	7.4	**6.8%**	7.4%

Face diarization results

	Initial monomodal	CRF Dia	CRF Dia without OPNs
News	**4.8%**	5.4%	5.9%
Debates	4.6%	**1.9%**	4.4%
Parliament	11.2%	**10.4%**	13.7%
Celebrity	**3.5%**	7.9%	6.4%
All	5.2%	**5.0%**	6.1%

Table 1. Speaker and face diarization performances in terms of CER. The first column presents the initial monomodal systems Khoury et al. (2013); Rouvier et al. (2013). The second one is the diarization CRF presented in this paper. The third one is the same as the second one, however, we remove the OPN related functions f_{lfb} and f_{opn}.

where the audio file is divided in continuous segments at each speaker change and:

- dur(seg) is the duration of the segment seg.
- N_{Ref}(seg) is the number of active speakers during segment seg.
- N_{Sys}(seg) is the number of speakers detected by the system.
- N_{Correct}(seg) is the number of speakers correctly detected by the system. A match needs to be made between the clusters and the speaker references in order to compute this term.

We applied this measure to the face clustering task. With the audio CER, a detected speech segment is matched to a reference during their temporal overlap. The only modification to tackle visual modality is that a face detection must have a temporal AND spatial overlap to be matched with a reference. In addition, note that we do not consider false alarms and missed detections that are usually considered in NIST to compare the effects of the different systems since the only error that changes with methods given the setup (fixed face tracks and utterances) is due to the final clustering of the face and speech segments. Thus, miss detections and false alarms are identical.

4.4 Identification and clustering results with the CRF combination

Diarization results: we first describe the diarization results presented in Table 1. We can see that the full CRF model has a slightly lower error rate over the whole corpus than the initial monomodal systems (6.8% vs 7.4% for the speakers and 5.0% vs 5.2% for the faces). On the other hand, the performances depend

function	f_a	f_v	f_{av}	f_{lfbv^k}	f_{lfba^k}
λ	λ_a	λ_v	λ_{av}	λ_{lfbv^k}	λ_{lfba^k}
λ value	0.4	1.8	0.2	1.9	1.7

Table 2. The λ parameter values for some of the feature functions used by the diarization CRF. For λ_{lfbv^k} and λ_{lfba^k}, the value of k is 5, which is the highest parameter value. It corresponds to the most common case as 90% of the segments are inside background clusters which contain more than 5 elements.

strongly on the type of shows. For instance, an important part of the global improvement comes from the debate videos (4.0% vs 6.6% for the speakers and 1.9% vs 4.9% for the faces). In debates, most of the scenes are in the same studio thereby reducing the visual variability of the background image and most of the persons present are speakers announced by an OPN. Thus, most faces and utterances are featured as recurrent (i.e. x_i^{lfbv} is set to true) and the f_{lfb} functions have a positive impact on the diarization. They enable to solve clustering confusion errors by constraining the number of clusters toward the number of detected OPNs. Indeed, if we remove the OPN related functions f_{lfb} and f_{opn} (cf third column of the Table 1), most of the improvements are lost. It appears that the use of multimodality does not help to correct clustering errors. This is somewhat surprising as past works (Gay et al. (2014c)) reports improvements in the audio modality with this very system on the same type of data. The difference with this previous work is that our initial monomodal speaker diarization system has become much more efficient, essentially thanks to a careful selection of the data used to train the generic speaker model UBM. This way, there are much fewer errors to correct.

In the case of celebrity magazines, the diarization CRF increases the error rate (15.1% vs 14.6% for the speakers and 7.9% vs 3.5% for the faces). Those videos contain very few OPNs and essentially short outdoor scenes. Thus, the f_{lfb} functions cannot help the CRF to take appropriate decisions. Moreover, previous experiments reported in Gay et al. (2014c) showed that the use of the biometric person models inside the CRF framework appears to be less efficient than when it is used in the hierarchical monomodal systems.

The importance of the OPN related functions is also visible if we consider the λ parameters learned by the CRF in Table 2. During training, the weight λ_{av} are indeed set to a relatively low value as compared to the other terms (although those values are ponderated by the amplitude of the feature functions). We have found that for a majority of segments, the f_{lfb} function is dominant. This is further illustrated in table 4.

Identification results: We now turn to the identification results reported in Table 3. We compare 3 systems: we denote by N the naming CRF applied on top of the initial monomodal diarizations described in Khoury et al. (2013); Rouvier et al. (2013), $N + D$ is the joint use of the naming and the diarization CRF, and the last one is an oracle. Note that the oracle still produces errors, since, as we deal with automatic face detection and tracking, there are errors that a perfect clustering and naming cannot correct: false alarms, missed faces and face tracks for which the identity is not introduced by an OPN (see more about this in Fig 8). Adding the diarization CRF permits to globally reduce the error rates in both modalities (31.4% vs 33.4% for the speakers and 52.2% vs 54.5% for the faces), especially for debate and parliament videos. This is not surprising as we previously showed that the diarization CRF have less confusion errors for studio scenes than the initial monomodal systems.

Regarding news videos, although we saw that clustering confusion errors were not reduced globally, the use of the diarization CRF also improves the identification. This is probably due to the correction of confusion errors in studio scenes which have a greater impact on the identification than errors concerning anonymous persons in reports.

	Audio			Visual		
	N	$N+D$	Oracle	N	$N+D$	Oracle
News	31.6%	30.8%	25.7%	58.2%	56.4%	37.7%
Debates	18.0%	14.0%	11.3%	42.0%	38.0%	35.6%
Parliament	11.3%	8.7%	5.2%	62.2%	59.6%	47.4%
Celebrity	85.6%	85.8%	82.1%	83.9%	86.6%	75.3%
All	33.4%	31.4%	27.2%	54.5%	52.2%	40.2%

Table 3. Identification performances measured in EGER. The system N is the naming CRF on top of the monomodal diarizations and the system $N+D$ is the naming and diarization CRF combination.

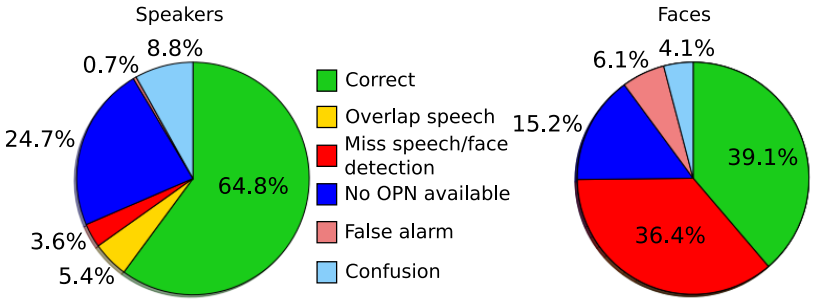

Figure 8. Different errors for the speaker (left) and face (right) identification tasks. Percentages are expressed relatively to the number of annotations.

The structure of celebrity magazines differs from the other shows as it contains very few OPNs and recurrent LFB. In those cases, the diarization CRF degrades both diarization and identification performances. We design an oracle on the diarization and the identification to measure the potential improvements. It uses automatic face/speech segment detections and automatic OPN extraction. Then, the association between these segments and the OPNs is done with the manual reference. Thus, the errors made by the oracle correspond to missing OPNs or missing segment detections. In the case of celebrity shows, with an error rate of 75.3%, the OPN-based approach is clearly not suitable.

Error analysis: the proportion of the different error types can be visualized globally on the pie charts in Fig 8. Regarding the speaker identification task, the lack of OPNs explains most of the errors as 24.7% of the annotated persons are not announced, most of them being journalists. As for the faces, the detection step is more crucial as 36.4% of the persons faces are not detected. This corresponds usually to profile faces or persons seen from the back. Most of the false alarms are anonymous persons incorrectly identified.

We also illustrate the correlation between diarization and identification performances in figure 9. We plot the performance differences for each video between the full system ($N+D$) and the CRF naming alone (N). We observe that they are unique to their type of show. The debate videos appear in the top-right part of the plane, which means that the diarization CRF improves the diarization and the identification. Concerning news and parliament videos, the correlation between CER and DER is not as strong. The presence of anonymous persons and off voices imply that a change in the diarization does not necessarily correspond to a change in identification performances.

Finally, the table 4 shows the performance of the model when adding the different components of the diarization CRF one by one. If we focus on the first and second lines, we see that the CRF with only 3 feature functions degrades the performances compared to the monomodal diarizations. We find that, used alone, the monomodal representations present in the CRF (see the f_a and f_v functions) do not compare favourably with the monomodal diarization frameworks. This could be improved in a future work by using better person representations. However, each other component enables to reduce the error rate and the

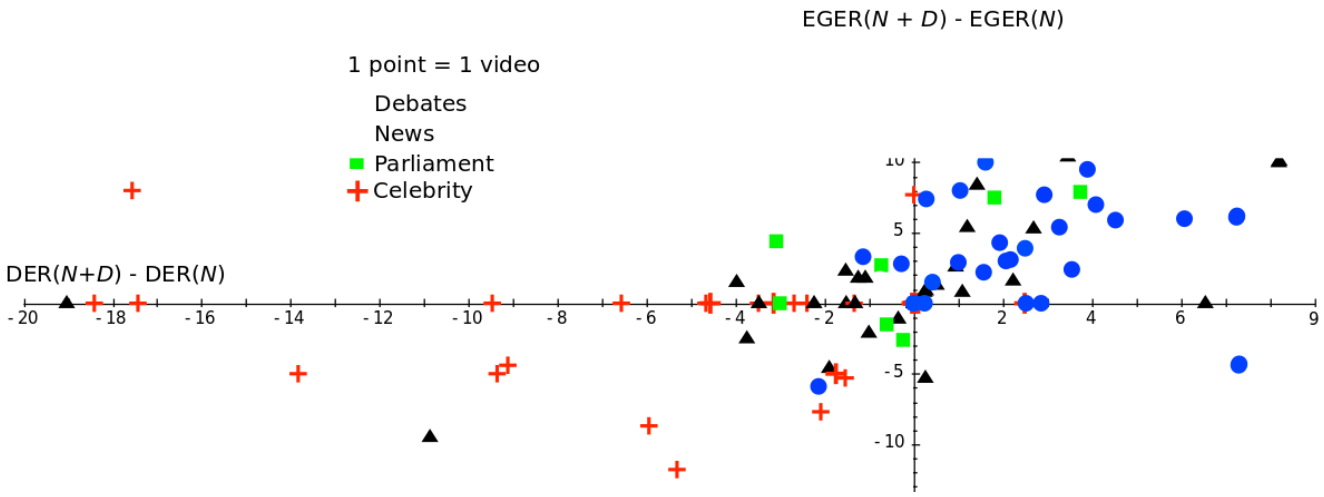

Figure 9. The Y-axis is the EGER difference between the CRF combination and the naming CRF alone measured for the faces. The X-axis is the DER difference between the diarization CRF and the initial monomodal face diarization Khoury et al. (2013).

	Audio	Visual
N	33.4%	54.5%
$N + D\ (f_a + f_v + f_{\text{av}})$	34.1%	56.2%
$N + D\ (f_a + f_v + f_{\text{av}} + f_{\text{opn}})$	33.9%	56.4%
$N + D\ (f_a + f_v + f_{\text{av}} + f_{\text{opn}} + f_{\text{uniq}})$	33.9%	55.6%
$N + D\ (f_a + f_v + f_{\text{av}} + f_{\text{opn}} + f_{\text{uniq}} + f_{\text{lfb}})$	31.4%	52.7%

Table 4. Contribution of the different diarization model components on the naming task (results in EGER). As in table 3, the system N is the naming CRF with the monomodal diarizations. The other lines correspond to the combination of the naming and the diarization CRF, using as feature functions in the diarization CRF those given in parenthesis.

full model provides the best performances. It should also be noticed that, although it might generates big cliques in some cases, the uniqueness function is essential to benefit from the f_{opn} feature functions. If not applied, an OPN will be propagated to all the faces overlapping with him.

Comparison with state of the art and discussion: on the same dataset, the system described in Bechet et al. (2014) obtains an EGER of 30.9% for the speakers and 39.4% for the faces. Thus, it proves to have a better performance especially regarding the faces. This is possible with the help of pre-trained models for each show which enable to indicate how many faces should be present on screen and what their roles are. For instance, when it detects the configuration shown in Fig 1c, it deduces that the announced guest is present on the right even if no faces have been detected. In fact, this approach does not even use a face diarization module. However, it requires a large amount of learning and *a priori* information. By comparison, our method is much simpler to implement, especially since it has better generalization capabilities, we learn one single model over a large and diverse corpus, and what is more, it requires less annotations if we need to process a new type of show.

The constrained hierarchical clustering detailed in Poignant et al. (2015) obtains an EGER of 35.9% for the speakers and 44.3% for the faces. Compared with our system, it has better performances on the faces, but worst for the speakers. As we do, they only rely on OPNs without other specific supervised information on the show. According to their paper, it seems that their constrained multimodal clustering, which avoids clustering together faces which co-occur with different OPN names, is one of the contributions which improves results and that we do not use, and could explain the difference. Nevertheless, the influence of

each pre-processing (speaker and face detections, monomodal clusterings and OPN detection) makes it hard to analyse the performance difference.

5 CONCLUSION

In this paper, we presented our contribution for AV person diarization and identification from OPNs. Our system uses an iterative combination of 2 CRFs. One performing the AV diarization at a person level, and a second one associating the names and the clusters. Several context modeling cues are used to solve the person/name association problem and the diarization issues. While it is clear that more supervised learning and *a priori* information on the context can improve the performances, our approach provides an interesting trade-off between performance on one hand and generalization/low annotation cost on the other hand. The principal contextual cue consists in the face image background. It allows us to distinguish the faces and the speakers which are announced by OPNs and guide the clustering accordingly.

In this work, we did not address the issue of non-frontal face detection. As a short term perspective, it would be interesting to increase the recall of the face detector, for instance, by adding a profile view detector. This would render the face clustering task more challenging and the potential benefit from context modeling would be greater. Secondly, our context modeling assumes that speakers are announced by an OPN the first time they talk. For the REPERE dataset, this is the case. However, this assumption could be sensible to broadcaster's editing policies. Actually, the optimal choice of the context for unsupervised person identification is a difficult problem if we want to avoid the need for specific annotations for each show. One solution to consider is to learn the setting of each show or a part of the setting from a corpus in an unsupervised way.

REFERENCES

Bauml, M., Tapaswi, M., and Stiefelhagen, R. (2013). Semi-supervised learning with constraints for person identification in multimedia data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609, Portland, USA.

Bechet, F., Bendris, M., Charlet, D., Damnati, G., Favre, B., Rouvier, M., Auguste, R., Bigot, B., Dufour, R., Fredouille, C., Linares, G., Martinet, J., Senay, G., and Tirilly, P. (2014). Multimodal understanding for person recognition in video broadcasts. In *Proceedings of InterSpeech*, pages 146–151, Singapore.

Ben, M., Betser, M., Bimbot, F., and Gravier, G. (2004). Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *Proceedings of International Conference on Spoken Language Processing*, pages 523–538, Jeju, Island.

Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y. W., Learned-Miller, E. G., and Forsyth, D. A. (2004). Names and faces in the news. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 836–848, Washington, USA.

Bhattarai, B., Sharma, G., Jurie, F., and Pérez, P. (2014). Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 160–172, Zurich, Switzerland.

Bredin, H. and Poignant, J. (2013). Integer linear programming for speaker diarization and cross-modal identification in TV broadcast. In *Proceedings of InterSpeech*, pages 49–54, Lyon, France.

Bredin, H., Poignant, J., Tapaswi, M., Fortier, G., Le, V., Napoleon, T., Gao, H., Barras, C., Rosset, S., Besacier, L., et al. (2013). QCOMPERE@REPERE 2013. In *Workshop on Speech, Language and Audio in Multimedia*, pages 49–54, Marseille, France.

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

Cinbis, R. G., Verbeek, J., and Schmid, C. (2011). Unsupervised metric learning for face identification in TV video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1559–1566, Barcelona, Spain.

Cour, T., Sapp, B., Nagle, A., and Taskar, B. (2010). Talking pictures: Temporal grouping and dialog-supervised person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1021, San Francisco, USA. IEEE.

Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536.

El Khoury, E., Senac, C., and Joly, P. (2010). Face-and-clothing based people clustering in video content. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 295–304, Philadelphia, USA.

El Khoury, E., Sénac, C., and Joly, P. (2012). Audiovisual diarization of people in video content. *Multimedia Tools and Applications*, 68(3):747–775.

Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello! my name is... buffy–automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference*, volume 2, pages 2365–2371, Edinburgh, Scotland.

Gay, P., Dupuy, G., Lailler, C., Odobez, J.-M., Meignier, S., and Deléglise, P. (2014a). Comparison of two methods for unsupervised person identification in TV shows. In *Proceedings of The Content Based Multimedia Indexing Workshop*, pages 1–6, Klagenfurt, Austria.

Gay, P., Elie, K., Sylvain, M., Jean-Marc, O., and Paul, D. (2014b). A conditional random field approach for face identification in broadcast news using overlaid text. In *Proceedings of the IEEE International Conference on Image Processing*, pages 318–322.

Gay, P., Khoury, E., Meignier, S., Odobez, J.-M., and Deleglise, P. (2014c). A conditional random field approach for audio-visual people diarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 116–120, Florence, Italy.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The REPERE corpus: a multimodal corpus for person recognition. In *Proceedings of The International Conference on Language Resources and Evaluation*, pages 1102–1107, Istanbul, Turkey.

Guillaumin, M., Verbeek, J., and Schmid, C. (2010). Multiple instance metric learning from automatically labeled bags of faces. In *Proceedings of the European Conference on Computer Vision*, pages 634–647. Crete, Greece.

Jou, B., Li, H., Ellis, J. G., Morozoff-Abegauz, D., and Chang, S. (2013). Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *Proceedings of ACM International conference on Multimedia*, pages 357–360, Barcelona, Spain.

Jousse, V., Petit-Renaud, S., Meignier, S., Esteve, Y., and Jacquin, C. (2009). Automatic named identification of speakers using diarization and ASR systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4557–4560, Taipei, Taiwan.

Khoury, E., Gay, P., and Odobez, J. (2013). Fusing matching and biometric similarity measures for face diarization in video. In *Proceedings of the IEEE International Conference on Multimedia Retrieval*, pages 97–104, Dallas, USA.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97.

Li, D., Wei, G., Sethi, I. K., and Dimitrova, N. (2001). Person identification in tv programs. *Journal of Electronic Imaging*, 10(4):930–938.

Ma, C., Nguyen, P., and Mahajan, M. (2007). Finding speaker identities with a conditional maximum entropy model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 253–261, Honolulu, USA.

McCallum", A. K. (2002). "MALLET: A machine learning for language toolkit". "http://mallet.cs.umass.edu".

NIST (2003). The rich transcription spring 2003 (rt-03s) evaluation plan.

Noulas, A., Englebienne, G., and Krose, B. J. (2012). Multimodal speaker diarization. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 34(1):79–93.

Ozkan, D. and Duygulu, P. (2010). Interesting faces: A graph-based approach for finding people in news. *Pattern Recognition*, 43(5):1717–1735.

Pham, P., Moens, M.-F., and Tuytelaars, T. (2008). Linking names and faces: Seeing the problem in different ways. In *Proceedings of the European Conference on Computer Vision*, pages 68–81, Marseille, France.

Pham, P. T., Deschacht, K., Tuytelaars, T., and Moens, M.-F. (2013). Naming persons in video: Using the weak supervision of textual stories. *Journal of Visual Communication and Image Representation*, 24(7):944–955.

Poignant, J., Besacier, L., and Quénot, G. (2014). Unsupervised speaker identification in TV broadcast based on written names. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):57–68.

Poignant, J., Fortier, G., Besacier, L., and Quénot, G. (2015). Naming multi-modal clusters to identify persons in tv broadcast. *Multimedia Tools and Applications*, pages 1–25.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *Proceedings of InterSpeech*, pages 547–552, Lyon, France.

Rouvier, M. and Meignier, S. (2012). A global optimization framework for speaker diarization. In *Proceedings of the Odyssey workshop*, pages 546–552, Singapore.

Satoh, S., Nakamura, Y., and Kanade, T. (1999). Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, Boston, USA.

Simonyan, K., Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2013). Fisher vector faces in the wild. In *Proceedings of the British Machine Vision Conference*, pages 867–879, Bristol, United Kingdom.

Tapaswi, M., Parkhi, O. M., Rahtu, E., Sommerlade, E., Stiefelhagen, R., and Zisserman, A. (2014). Total cluster: A person agnostic clustering method for broadcast videos. In *Proceedings of the Indian Conference on Computer Vision Graphics and Image Processing*, pages 7–15, Bengaluru, India.

Vallet, F., Essid, S., and Carrive, J. (2013). A multimodal approach to speaker diarization on TV talk-shows. *IEEE Transactions on Multimedia*, 15(3):509–520.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wohlhart, P., Köstinger, M., Roth, P. M., and Bischof, H. (2011). Multiple instance boosting for face recognition in videos. *Pattern Recognition*, 6835:132–141.

Zhang, L., Kalashnikov, D. V., and Mehrotra, S. (2013). A unified framework for context assisted face clustering. In *Proceedings of the IEEE International Conference on Multimedia Retrieval*, pages 9–16, Dallas, USA.

Zhang, N., Paluri, M., Tagiman, Y., Fergus, R., and Bourdev, L. (2015). Beyond frontal faces: Improving person recognition using multiple cues. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813.