A CONDITIONAL RANDOM FIELD APPROACH FOR AUDIO-VISUAL PEOPLE DIARIZATION

$^{(1,2)}$*Gay Paul*, ^{1}*Khoury Elie*, ^{2}*Meignier Sylvain*, ^{1}*Odobez Jean-Marc*, ^{2}*Deleglise Paul*

^{1}Idiap Research Institute, Martigny, Switzerland, ^{2}LIUM, University of Maine, Le Mans, France

ABSTRACT

We investigate the problem of audio-visual (AV) person diarization in broadcast data. That is, automatically associate the faces and voices of people and determine when they appear or speak in the video. The contributions are twofolds. First, we formulate the problem within a novel CRF framework that simultaneously performs the AV association of voices and face clusters to build AV person models, and the joint segmentation of the audio and visual streams using a set of AV cues and their association strength. Secondly, we use for this AV association strength a score that does not only rely on lips activity, but also on contextual visual information (face size, position, number of detected faces,...) that leads to more reliable association measures. Experiments on 6 hours of broadcast data show that our framework is able to improve the AV-person diarization especially for speaker segments erroneously labeled in the mono-modal case.

Index Terms— Audiovisual, diarization, Conditional Random Field

1. INTRODUCTION

We address the problem of audio-visual (AV) person diarization in broadcast data as illustrated in Fig 1. Solving such a task would allow the design of AV person structured indexing and exploration tools useful to achieve, for instance, fast annotation of already temporally segmented documents, automatic person naming by further exploiting OCR or closed-captions [1], or in general, facilitate browsing and access to relevant video parts.

However, AV person diarization is a hard problem due to many difficulties. Speaker diarization systems can make errors due to short utterances, spontaneous speech, and background noise. In the visual modality, faces display many variations in scale and pose as shown in the top row of Fig 1. Interestingly, the joint exploitation of audio and video could help to correct those errors. However, the association between speech and face can introduce many ambiguities in case of multi-face shots, as shown in the first image of Fig 1 or shots where the talking person is not visible or difficult to detect like in the third image of Fig 1 where the talking person is the person seen from the back.

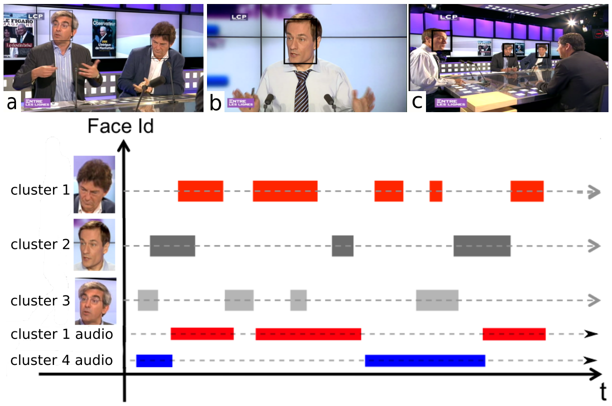

Fig. 1. Top row: Sample frames of a TV debate. Bottom row: example of an AV people diarization output. Note that cluster 1 both appears and speaks.

Related work. Earlier work on AV person diarization performs separately audio and video clustering in a first step and associate the clusters in a second step [2, 3, 4]. The most simple clue to associate faces and speakers is their temporal co-occurrence. Additionally, lips activity is an interesting cue to detect and localize a speaker in a video through the use of motion information [5, 6, 7]. However, it is hard to apply when the mouth region cannot be followed precisely motivating some authors in [5] to focus only on video segments where the AV association can be performed with high reliability (single talking face) to correct errors in the speaker diarization process. Moreover, these methods assume that the mono-modal diarizations are perfect and only perform AV association without attempting at correcting clustering errors made in the first step. More closely related to our work, the authors in [8] try to correct these errors using AV cues. They first perform the 2-steps cluster association using a greedy algorithm. Each cluster is then represented by biometric models and a rule based approach is used to refine mono-modal segments based on the AV scores given by these models. One major drawback is that this work makes a succession of local and hard decisions which might not be globally optimal. In the computer vision literature, global optimization frameworks have been successfully used to perform the segmentation task and handle multi-modality. Authors in [9] use Markov random field to combine audio and video classifiers for identifying people in TV-Series while [10] uses Condi-

The authors gratefully acknowledge the financial support from the French Research Agency (ANR) under the Project SODA and from the European Union under the EUMSSI project (grant agreement 611057).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tional Random Fields (CRF) to integrate various cues in a face clustering task.

Proposed method. This paper presents a CRF based framework which performs a global optimization over the audio segments (also called utterances) and the video segments (also called face tracks) to refine the clusters. The errors from the mono-modal diarizations are corrected by favouring couples of utterances/face tracks sharing the same label to have a high score of association. On the other hand, errors from wrong associations are prevented by representing each cluster with a biometric model and encouraging high scores between segments and their respective cluster models. This second part ensures the inner consistency of the clusters. The CRF formulation enables to learn the weights of the contribution of each type of information and to discriminantly perform inference to get the most probable clusters. In contrast to [8], the decision is taken by considering jointly all the multi-modal information in a probabilistic framework and the optimization is performed globally over all the segments.

2. PROPOSED METHOD

2.1. Problem formulation

Let $A = \{A_i, i = 1 \dots N^A\}$ denote the set of utterances and $V = \{V_i, i = 1 \dots N^V\}$ the set of face tracks. The AV person diarization problem can be formulated as the estimation of the labels field $E = \{e^a_i, i = 1..N^A, e^v_j, j = 1..N^V\}$ by maximizing the posterior distribution $P(E|A, V)$ such that the same person index is used for e^a_i and e^v_j when the utterance A_i and the face track V_j correspond to the same person. The labels e^a_i and e^v_j take value in the set of possible person indices denoted as $\mathcal{E}$. Let G be an indirect graph over the set of random observed variables A and V. We express the posterior probability for labels E as:

$$P(E|A,V) = \frac{1}{Z(A,V)} \tag{1}$$
$$\times exp\Bigg\{\lambda_{av} \sum_{(i,j)\in G_{av}} f_{av}(A_i, V_j, e^a_i, e^v_j) + \lambda_a \sum_{i=1}^{N^A} f_a(A_i, e^a_i)$$
$$+\lambda_v^{dct} \sum_{i=1}^{N^V} f_v^{dct}(V_i, e^v_i) + \lambda_v^{surf} \sum_{i=1}^{N^V} f_v^{surf}(V_i, e^v_i)\Bigg\}$$

where $Z(A, V)$ denotes the partition function, and f_{av}, f_a, and f_v^{dct} and f_v^{surf} respectively denote the AV association feature function, the audio feature function, and two video feature functions which will be defined in Sec 2.2. A graphical illustration of our model is shown in Fig. 2.

2.2. Model components

The association feature function f_{av} is defined on all the couples of overlapping utterances/face tracks: $G_{av} = \{(i, j)/t(A_i, V_j) \neq 0\}$ where $t(x, y)$ denotes the duration of the overlapping time between segments x and y. The goal is to output a high score if utterance A_i and face track V_i correspond to a talking face. It is defined as:

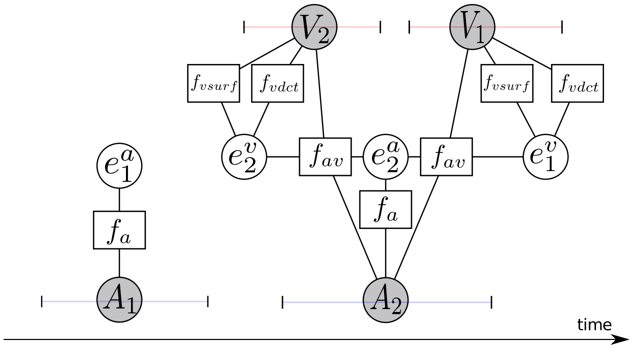

Fig. 2. Example of a factor graph representing our model with 4 segments. The face tracks V_1 is temporally overlapping with the utterance A_2, so, following the model component definition, they are dependent through the association feature function f_{av}. The utterance A_1 does not have any overlap. Thus, its likelihood depends only on its label e^a_1 through the biometric feature function f_a.

$$f_{av}(A_i, V_j, e^a_i, e^v_j) = \begin{cases} t(A_i, V_j)h(A_i, V_j) & \text{if } e^a_i = e^v_j \\ -t(A_i, V_j)h(A_i, V_j) & \text{otherwise} \end{cases}$$

where $h(A_i, V_j)$ represents the output of a SVM classifier indicating whether an utterance/track couple belongs to a talking face or not. The features used as input to the SVM not only include a lips activity measure computed using least mean square difference as done in most works, but also other contextual features that can help distinguishing a talking face track like the average face size, the average distance of the face to the center of the image. Importantly, we also take into account the presence and characteristic of other appearing faces by using the number of detected faces, the relative face size and relative lips activity as input features. Experiments showed the benefit of such contextual information for utterance/face track association.

The acoustic biometric feature function $f_a(A_i, e^a_i)$ indicates how likely the audio features of a given utterance A_i should be labeled with the person index e^a_i. This is a speaker modeling task, where we need to define an acoustic model for each label e and learn this model in an unsupervised fashion from the data currently associated to the label (and priors on model parameters). In our case, we choose a 512 GMM-UBM with diagonal covariance inspired from [11]. Feature are 12 MFCCs with first order derivatives and the features are normalized: short-term windowed mean and variance are computed to normalize the frame, and a feature warping normalization is applied.

Visual biometric feature functions. We need to proceed similarly for the visual modality. Following [12], we combine Speeded Up Robust Features (SURF) based matching and statistical models, and define two biometric feature functions f_v^{dct} and f_v^{surf}. f_v^{dct} relies on statistical models based on 45 dimensions Discrete Cosinus Transform features and a

512 GMM-UBM with diagonal covariance [13]. For f_v^{surf}, we extract SURF descriptor vectors from face images and define the score between a face track V_i and a label e_j^v as the average of pair-wise surf vector distances between the track V_i and the current face-tracks associated with this label [14].

2.3. Optimization and parameter training

The CRF inference for new data is conducted by applying the following steps: i) initialize the labels, ii) for each label, learn the biometric models from their associated data, iii) get the most probable labels. Steps ii) and iii) are then iterated in a Expectation-Maximization style by alternating model updates and inference.

Label initialization is achieved by first performing separately audio and video clustering (see Sec. 3.1) and then associating the clusters in a second step to obtain the AV person labels $\mathcal{E}$ (audio and face cluster couples). The association is done by optimizing Eq 1 while dropping the biometric terms. The optimization is conducted using the greedy Hungarian algorithm where each element C_{ij} of the cost matrix is the sum of the scores from the association function f_{av} over all couples of utterances/face tracks currently associated to labels e_i^a and e_j^v.

For each resulting person label, biometric models are learned from their associated data and used to compute the likelihood of any utterance or face track observation. Given these models, we run the loopy belief propagation inference to get the most probable labels E by solving $E = \arg\max_E P(E|A, V)$.

Note that for mono-modal labels (i.e. at a given iteration, labels associated with only faces -case of people that appear but never speak- or audio), we still need to be able to evaluate the likelihood of data in the other modality to conduct the CRF optimization. To handle this issue, a *Neutral* biometric model has been created for each modality, and is associated to the missing modality of each mono-modal person label. In practice, the score of an observation for this neutral model has been defined as the score of the corresponding biometric function obtained at the Equal Error Rate (when the number of false alarms equals the number of miss detections) of speaker and face verification experiments conducted on external data.

Parameter training. The CRF model is parameterized by the different λ values that express the reliability of each cue in the label inference. They can be learned using labeled training data. However, to avoid data mismatch (at test time, the CRF is conducted on noisy face tracks and utterances, not on cleanly segmented ones) it is important to learn parameters using the segments produced by the mono-modal automatic diarization steps but using the true person labels. Additionally, the data associated to each label and used to train the biometric models will come from the clusters produced by the mono-modal diarizations and will thus be noisy as it is will be at test time. Clusters and labels are associated by minimizing the Diarization Error Rate (DER). In other words, the CRF parameters can be trained to account for the errors made by the initial diarization steps.

Fig. 3. Examples from REPERE dataset showing the visual variability of broadcast news data.

3. EXPERIMENTAL RESULTS

3.1. Data and experimental protocol

Data and metric. Experiments are done using the dry-run of the REPERE dataset [15]. It consists of 6 hours of annotated videos recorded from 2 French TV channels (BFMTV and LCP) and 7 different shows. It includes not only TV news and debates, but also challenging talk shows with multi-head shots as illustrated in Fig 3. The data is divided into development and test sets (3 hours each). As performance metric, we used the standard Diarization Error Rate (DER). However, since false alarm and miss detection rates do not change in our comparisons, we only report the part of the DER due to the speaker and face clustering error rates, that is, the error remaining after having done the optimal cluster to ground truth association.

Initialisation. The initial speaker diarization uses the system of [16] which combines a bottom-up approach with an ILP formulation and i-vector representation. The system is state of the art and obtained the best performance at the REPERE evaluation campaign [15], with 17.14 % of DER on the dry-run thus making further improvements quite difficult. The initial face diarization uses the bottom-up system described in [12] already mentioned in Sec 2.2. It combines SURF descriptors and statistical models following a standard speaker diarization approach [17]. It achieves state of the art results on the publicly available BUFFY dataset [18].

Models. We tested 2 settings for the CRF parameters. In the first one, a single set of λ parameters (*CRF-all*) is learned from all the shows in the training set and applied to all test shows. In the second case (*CRF-spec*), specific sets of parameters are learned for each show in the training set and applied to the corresponding test shows. The rationale is that depending on the context (talkshow, report, debate), the reliabilities of the different modalities might be different. Our CRF implementation relies on [19].

3.2. Results

Association feature function. The SVM association function involved in f_{av} has been trained on the development set using an RBF kernel. We evaluated it by cross validation on 846 couples of utterances/face tracks that should be associated or not. Experiments showed that adding the contextual

Table 1. speaker, face and people error rates in percentage of scored time. The initial mono-modal face and speaker diarizations are compared with the results obtained with the CRF-based audiovisual diarization.

	speaker error rate			face error rate			AV people error rate			
show	Init	CRF-all	CRF-spec	Init	CRF-all	CRF-spec	Init	Greedy	CRF-all	CRF-spec
BFMTV_BFMStory	2.8	4.8	**2.6**	**3.5**	4.7	4.7	21.7	5.5	6.4	**6.2**
BFMTV_Planet_Showbiz	15.0	15.0	15.0	**2.9**	5.5	**2.9**	18.0	**11.4**	13.4	**11.4**
LCP_Ca_Vous_Regarde	14.8	**8.7**	**8.7**	3.0	3.1	**2.5**	26.9	10.9	8.0	**7.7**
LCP_Entre_Les_Lignes	14.7	**4.1**	**4.1**	**4.2**	9.3	10.3	40.3	9.9	**6.6**	7.0
LCP_LCPInfo	**9.6**	9.7	**9.6**	7.3	**4.9**	**4.9**	28.1	9.6	**8.4**	**8.4**
LCP_Pile_Et_Face	9.2	**2.8**	**2.8**	**2.8**	7.7	7.7	39.5	**5.8**	6.4	6.4
LCP_Top_Questions	19.3	**3.8**	**3.8**	8.2	**6.2**	**6.2**	28.9	13.1	**6.0**	**6.0**
All	10.0	6.3	**5.6**	**5.3**	5.6	5.4	27.1	9.0	7.2	**7.1**

features like the relative size and lips activity or the number of head detections drastically improved the correct association rate of utterances/face tracks, with a F-measure of 0.76 as compared to 0.69 when using the lips activity only.

Diarization tasks. We evaluate the AV people diarization produced by our CRF approach by measuring the speaker error rate, the face error rate and the AV people error rate. Results are reported on Table 1. *Init* refers to using the initial mono-modal speaker and face diarizations without any AV association or refinement (thus no label simultaneously represents a face and a speaker). *Greedy* corresponds to the direct association of the mono-modal labels using the Hungarian algorithm as explained in Sec. 2.3. Finally, *CRF-all* and *CRF-spec* corresponds to the proposed method using the two CRF learning strategies described in Sec. 3.1.

The CRF improves the AV error rate w.r.t. the mere greedy algorithm from 9.0% to 7.2% for *CRF-all* and to 7.1% for *CRF-spec*. This is mainly due to improvements in the audio modality, where the speaker error rate is reduced from 10% to 6.3% for *CRF-all* and 5.6% for *CRF-spec*. These improvements are entirely due to the integration of multi-modality, since running the CRF while dropping the association function f_{av} (thus the diarization only relies on the audio biometric model) did not alter the initial speaker diarization. A closer look at the results show that AV cues help to refine clusters in the first CRF iterations mainly when talking heads are present alone in the screen or when the lips activity has enough discriminative power. These cases are illustrated by the images a,c and e in Fig 3. In subsequent iterations, the updated biometric models benefit from these refinements and enable additional corrections. Note that the shows *Planet_Showbiz* and *BFMStory* do not exhibit improvement as the above cases are less frequent, to the contrary of multi-head shots (image d), off voices (image b and f) and missed head detections which tend to weaken the coherency of the association information.

Considering the face diarization, we note that the CRF refinement slightly deteriorates the results w.r.t. to the initial diarization, from 5.3 to 5.6 for *CRF-all* and 5.4 for *CRF-spec*. This is partly due to an undesirable side-effect of the SURF-based biometric models (however overall this term significantly contributes to the diarization). Actually, for persons whose face tracks remain splitted in several clusters after the initial diarization, some face tracks move from the dominant cluster to smaller ones during CRF optimization due to the averaging effect: their *average* SURF matching similarity can be higher with the few tracks of a small and tight cluster than with a large cluster containing more diverse track appearances. As a result, the error rate augments although the cluster purity is preserved. Although dropping the association function f_{av} results in a larger error increase, from 5.3 to 5.8 (*CRF-spec*) and 5.6 (*CRF-all*) instead of 5.6 and 5.4 with f_{av}, it is harder than in audio to rely on AV cues to improve the initial face diarization. Indeed, the situations where AV cues are useful like alone talking heads are also often easy cases for the mono-modal face diarization system and are already correctly clustered together.

Finally, we note that *a priori* information about the show generally improves the results. This is the case for *Planet_Showbiz*, which is quite different from the other shows in the dataset, and for which *CRF-spec* provides better results than *CRF-all*. As mentioned earlier, this show leads to less coherent association information. Hence, the λ parameters learned on training data specifically for this show give a lower importance to the AV association function f_{av}, thus avoiding association errors.

4. CONCLUSION

This paper presents an AV person diarization algorithm relying on a global CRF formulation of the audiovisual information association problem. Experiments show that the model is able to reduce the overall AV person diarization errors (as compared to a greedy algorithm) mainly through improvements in the speaker diarization observed when talking face presence estimation is successful.

Several improvements can be made. For instance, visual person models could be improved by adding clothes descriptors [20] or by introducing pose information [21] in the face comparison functions, while on the association side, the module could be improved using contextual information given by role recognition [22] or shot classes (studio vs field, close-ups, group, public, etc) automatically derived from scene content descriptors, movements, duration...

5. REFERENCES

[1] B. Jou, H. Li, J. G Ellis, D. Morozoff-Abegauz, and S. Chang, "Structured exploration of who, what, when, and where in heterogeneous multimedia news sources," 2013.

[2] Z. Liu and Y. Wang, "Major cast detection in video using both speaker and face information," in IEEE *Trans. on Multimedia*, 2007.

[3] E. El Khoury, G. Jaffré, J. Pinquier, and C. Sénac, "Association of audio and video segmentations for automatic person indexing," in *Proc. of CBMI*, 2007.

[4] A. Dielmann, "Unsupervised detection of multimodal clusters in edited recordings," in IEEE *Trans. on Mulitmedia Signal Processing*, 2010.

[5] S. Bozonnet, F. Vallet, N. Evans, S. Essid, G. Richard, and J. Carrive, "A multimodal approach to initialisation for top-down speaker diarization of television shows," Tech. Rep., EURECOM, Sophia Antipolis, 2010.

[6] M. Bendris, D. Charlet, and G. Chollet, "Lip activity detection for talking faces classification in tv-content," in *Proc. of ICMV*, 2010.

[7] E. El Khoury, *Unsupervised video indexing based on audiovisual characterization of persons*, Ph.D. thesis, Paul Sabatier University, Toulouse, France, 2010.

[8] M. Bendris, D. Charlet, and G. Chollet, "People indexing in tv-content using lip-activity and unsupervised audio-visual identity verification," in *Proc. of CBMI*, 2011.

[9] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "knock! knock! who is it? probabilistic person identification in tv-series," in *Proc. of CVPR*, 2012.

[10] M. Du and R. Chellappa, "Face association across unconstrained video frames using conditional random fields," in *Proc. of ECCV*, 2012.

[11] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms," in *Proc. of ICLSP*, 2004.

[12] E. Khoury, P. Gay, and J.M. Odobez, "Fusing matching and biometric similarity measures for face diarization in video," in *Proc. of ICMR*, 2013.

[13] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models," in IEEE *Trans. on Information Forensics and Security*, 2012.

[14] E. El Khoury, C. Senac, and P. Joly, "Face-and-clothing based people clustering in video content," in *Proc. of ICMIR*, 2010.

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

[16] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," 2013.

[17] C. Barras, X. Zhu, S. Meignier, and J-L Gauvain, "Multistage speaker diarization of broadcast news," IEEE *Trans. on Audio, Speech, and Language Processing*, 2006.

[18] R.G. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *Proc. of ICCV*, 2011.

[19] C. Sutton, "Grmm: Graphical models in mallet," http://mallet.cs.umass.edu/grmm/, 2006.

[20] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy–automatic naming of characters in tv video," in *Proc. of BMVC*, 2006.

[21] F. Kottelat and J.M. Odobez, "Audio-video person clustering in video databases," Tech. Rep., IDIAP, 2003.

[22] T. Schwarze, T. Riegel, S. Han, A. Hutter, S. Wirth, C. Petersohn, and P. Ndjiki-Nya, "Role-based identity recognition for telecasts," in *Proc. of AIEMP*, 2010.