A CONDITIONAL RANDOM FIELD APPROACH FOR FACE IDENTIFICATION IN BROADCAST NEWS USING OVERLAID TEXT

[(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 face identification in broadcast programs where people names are obtained from text overlays automatically processed with Optical Character Recognition (OCR) and further linked to the faces throughout the video. To solve the face-name association and propagation, we propose a novel approach that combines the positive effects of two Conditional Random Field (CRF) models: a CRF for person diarization (joint temporal segmentation and association of voices and faces) that benefit from the combination of multiple cues including as main contributions the use of identification sources (OCR appearances) and recurrent local face visual background (LFB) playing the role of a namedness feature; a second CRF for the joint identification of the person clusters that improves identification performance thanks to the use of further diarization statistics. Experiments conducted on a recent and substantial public dataset of 7 different shows demonstrate the interest and complementarity of the different modeling steps and information sources, leading to state of the art results.

Index Terms— Face recognition,Conditionnal Random Fields, Broadcast news,Audio-visual

1. INTRODUCTION

Due to the growing amount of multimedia documents, there is a crucial need for search and fast browsing tools. A practical way to index multimedia documents is to identify the faces that appear in them. This approach has started to be investigated 15 years ago [1] and has raised since then a large amount of work, especially for face clustering tasks [2], face naming in captioned images [3], and more recently for actual automatic naming within broadcast videos [4, 5, 6, 7] that allows to monitor who said what and when in news programs.

In this paper, we propose an approach to identify faces using overlaid person names (OPNs) extracted by OCR techniques. These OPNs are used to introduce the most important people in the videos, like journalists and guests as illustrated in Fig. 1, making them very appealing [8, 7]: their extraction is much more reliable than pronounced names obtained through Automatic Speech Recognition (ASR), and their association with faces or people in the video is easier than analysing whether pronounced names in ASR transcripts refer to people appearing in the video.

Fig. 1. *Example frames from the REPERE corpus showing the variety of the visual conditions: pose, camera viewpoint, illumination, and the name face association challenges: multi face images (image **b**) and name propagation (from **a** to **c**).*

Identification using OPNs requires to solve two main tasks. The first one is the association of OPNs to faces, which is ambiguous when a name co-occurs with several faces as shown in Fig 1b, a situation that occurs more often with modern video editing. Secondly, in parallel, as only few appearances of a person are announced by an OPN, it is crucial to propagate the identity information to all the other face occurrences of that person (see Fig 1a-c). This later task is closely related to face clustering (also called face diarization), which is a difficult problem due to illumination, pose and camera view point changes.

Previous works have addressed automatic face clustering and identity recognition, for captioned images [9, 3], or in weakly labelled soap series [10]. There, the co-occurrence

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).

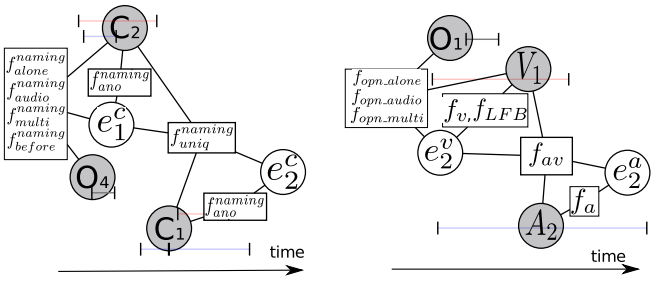

Fig. 2. *Factor graphs showing the naming CRF (left) which operates over clusters and OPNs, and the diarization CRF (right) which operates over segments and OPNs.*

statistics between face clusters and names is the main cue for association, but can fail when groups of people co-occur in a similar fashion [9]. Note that in these use-cases, the naming co-occurrence statistics to rely on are quite different than in broadcast videos where the OPNs are much sparser. To remove clustering ambiguities contextual and multimodal cues along with constraints has been investigated, including clothing [11, 12, 13, 14], cluster/name uniqueness constraints in images containing multiple faces [15, 12, 10], or face attributes and scene co-occurrence [11, 14] in images from photo albums.

In the video context, improvements have been obtained using the audio modality and talking head detection to propagate identification from speakers to faces [16, 5, 8].

In this paper, we identify faces by alternating between a clustering step and a naming step of those clusters. Each step is performed by a dedicated CRF. First, unlike previous works that do speaker and face clustering separately [16, 5], we use a joint CRF clustering of face tracks and speaker segments already proposed in [17]. We extend this model to benefit from the OCR/OPNs information. This is achieved by computing a local face visual background (LFB) around each face track, clustering them and assigning to each face track a signature which characterizes the level of recurrence of its LFB in the data. Intuitively, recurrent LFB characterizes people who are important and can be seen as a soft role assignment distinguishing faces to be named from faces of figurative people and thus concretely encourages faces tracks with recurrent LFB to join named clusters, i.e. overlapping an OPN. Secondly, a second CRF performs the join naming of all person clusters, thus allowing to account for uniqueness constraints and co-occurrence statistics between clusters and OPNs. Experiments on a comprehensive public dataset show the benefit of the modelling elements.

2. METHOD

2.1. Pre-processing

First, faces are detected [18] and tracked within each shot, resulting in a set of face tracks denoted as $V = \{V_i, i = 1 \dots N^V\}$, where each face track is characterized by a set of visual features $\mathbf{x}_i^{surf}$ (sets of Speeded-Up-Robust features extracted in up to 9 images of the face track which is the representation chosen in [13]), talking head detection features $\mathbf{x}_i^{av}$, and a boolean x_i^{lfb} indicating whether V_i corresponds to a recurrent LFB. Second, OCR [19] and Named entity detection techniques based on string matching against external resources (predefined lists, freebase database, google hits,...) are applied to extract the set $O = \{O_i, i = 1 \dots N^O\}$ of OPNs. Each O_i is characterized by its name $x_i^{opn} \in M$ where $M = \{n_i, i = 1 \dots N^{Na}\}$ denotes the set of unique names extracted from 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 $\mathbf{x}_i^a$.

2.2. Audio-visual (AV) person diarization

Problem formulation The clustering of face tracks and utterances consists in estimating the label field $E = \{e_i^a, i = 1..N^A, e_j^v, j = 1..N^V\}$ such that 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. We then seek to maximize the CRF posterior probability $P(E|A, V, O) =$

$$\frac{1}{Z(A,V,O)} \times exp\left\{ \sum_{i=1}^{6} \sum_{Clique \in G_i} \lambda_i f_i(Clique) \right\} \quad (1)$$

where each triplet (f_i, G_i, λ_i) is composed of a feature function f_i, the set G_i of cliques where this function is defined and its CRF weight λ_i learned at training time. This model is summarized on the right part of Fig 2 and the 6 feature functions will be described in the next section.

Diarization Model components Taking the first 3 functions would correspond to the model in [17] and the last 3 ones are introduced in this paper to account for OPN information and unicity constraints.

The association function $f_{av}(A_i, V_j, e_i^a, e_j^v)$ favours the association of talking heads to utterances. It is based on an SVM classifier and a set of measures $\mathbf{x}_i^{av}$ (lips activity, average absolute and relative face size, etc.).

The visual biometric feature functions $f_v(V_i, e_i^v)$, defined for all face tracks $V_i \in V$, indicates how likely the visual features $\mathbf{x}_i^{surf}$ of V_i should be labelled with the person index e_i^a. This is a face modelling task where for each label e_i, we need to define a visual model that is learned from the data currently associated to the label. In practice, f_v computes as score between V_i and a label e_j^v the average pair-wise SURF vector distances between $\mathbf{x}_i^{surf}$ and the SURF features of the current face-tracks associated with this label [13].

The acoustic biometric function $f_a(A_i, e_i^a)$ defined over all utterances $A_i \in A$ is the audio equivalent of f_v. We choose a

Fig. 3. *The persons on the left are actually announced by an OPN, whereas the persons on the right are non-talking figurative people. (see Fig 1 for the full images)*

GMM-UBM model following [20].

The LFB feature function is driven by the assumption that faces inside a recurrent LFB are likely to correspond to a person announced by an OPN. To this end, we focus on an area around each face track V_i to capture the background context of this face as illustrated in Fig 3. We do not consider full images as the same image might include different face visual contexts (see Fig 1b). We then characterized each area with SURF features and cluster them using a hierarchical clustering approach [13]. Then, we set x_i^{LFB} to $true$ if face track V_i belongs to a cluster whose number of elements is higher than a threshold T_{lfb}. The Fig 3 shows examples of obtained recurrent and non-recurrent patterns.

To favour the 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,

$$f_{LFB}(V_i, e_i^v) = \begin{cases} 1 & \text{if } e_i^v \in P^{opn} \text{ AND } x_i^{LFB} = true \\ 0 & \text{otherwise} \end{cases}$$

where P^{opn} is the set of person clusters co-occurring with an OPN.

The OPN feature functions favour segments (face tracks or utterances) co-occurring with an OPN O_j to be assigned to a person cluster likely to belong to the name x_j^{OPN}. For each co-occurring couple (V_i, O_j) for which V_i is alone in the image we define:

$$f_{opn\_alone}(V_i, O_j, e_i^v) = p(x_j^{opn}|e_i^v)$$

with $p(x_j^{opn}|e_i^v)$ the probability of a name given a cluster as defined in Sec 2.3. Similarly, we use f_{opn_multi} if V_i co-occurs with other faces and f_{opn_audio} for each co-occurring couple (A_i, O_j). Differentiate those 3 cases enables to learn specific λ weights so that the model behaviour is adapted to each situation.

The uniqueness feature function enforces two faces that co-occur in the same shot to have different labels [3, 10]. For such pair V_i, V_j:

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

2.3. Person cluster identification

The previous diarization step provides a set of audiovisual clusters $C = \{C_i, i = 1..N^C\}$. The naming step consists in estimating the label field $E^N = \{e_i^c, i = 1..N^C\}$ such that the label e_i^c corresponds to the name of the cluster C_i. e_i^c takes value from the set of names M augmented by a $anonymous$ label which should be assigned to anonymous persons. As illustrated in the left factor graph of Fig 2, the CRF posterior probability uses 6 feature functions and is expressed as:

$$P(E^N|C,O) = \frac{1}{Z(C,O)} \times exp\left\{ \sum_{i=1}^{6} \sum_{Clique \in G_i} \lambda_i f_i^{naming}(Clique) \right\}$$

This naming model exploits four different co-occurrence statistics between clusters and OPNs. First, for each triplet (e_i^c, C_i, O_j), for which O_j co-occurs visually only with C_i, we define:

$$f_{alone}^{naming}(e_i^c, C_i, O_j) = \begin{cases} 1 & \text{if } x_j^n = e_i^c \\ 0 & \text{otherwise} \end{cases}$$

As for the OPN diarization model components, we define similarly two other functions f_{multi}^{naming} and f_{audio}^{naming} to account for multi-face images and co-occurrences with the audio data. Eventually, we follow the assumption that a person does not appear or speak before the first apparition of his name in an OPN and define $f_{before}^{naming}(e_i^c, C_i, O)$ which returns how many segments from cluster C_i occur before the first apparition of the name e_i^c. We also introduce prior knowledge over the $anonymous$ label by defining a fifth feature function $f_{ano}^{naming}(e_i^c, C_i)$ which returns 1 if e_i^c is the $anonymous$ label. Lastly, we define a uniqueness function $f_{uniq}^{naming}(e_i^c, C_i, e_j^c, C_j)$ over visually overlapping clusters just as in the diarization step.

2.4. Optimization

The CRF inference is conducted by applying the following steps: i) The labels are initialized by first performing separately audio and video clustering (see Sec. 2.5) 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 and the AV association clues as in [17]. ii) For each resulting person label $p_i \in P$, biometric models are learned from their associated data and naming probabilities for each label $P(E^N|C,O)$ are estimated by running the Loopy Belief Propagation algorithm on the naming CRF. iii) Given these models, we get the most probable segment labels E according to the diarization CRF by solving $E = \arg\max_E P(E|A,V,O)$. Steps ii) and iii) can be iterated in a Expectation-Maximization style. Faces are eventually identified using the naming probabilities $P(E^N|C,O)$.

Experiments

2.5. Data and experimental protocol

Data and metrics. Experiments are done using the corpora *dev2* and *test1* from the REPERE evaluation campaign [21]. *dev1* is used to train the CRFs and optimise the LFB threshold T_{lfb} and *test1* for evaluation. Each set consists in 3 hours of annotated data extracted from 7 different shows which include TV news, reports, debates and talk shows (see Fig 1). The evaluation metric chosen to measure identification performance is the official REPERE Estimated Global Error Rate (EGER). This metric is defined as follow:

$$EGER = \frac{\#false + \#miss + \#conf_{identity}}{\#total}$$

where $\#total$ is the total number of faces to be detected, $\#conf_{identity}$ the number of faces wrongly identified, $\#miss$ the number of missed faces and $\#false$ the number of false alarms. We also study the behaviour of the clustering error rate $CER = \frac{\#conf_{label}}{\#total}$ where $\#conf_{label}$ is the number of faces assigned to the wrong cluster. It is measured before the naming step.

Initialisation. The initial face clustering uses a bottom-up algorithm which combines SURF descriptors and statistical models [22]. The initial speaker diarization is performed using a bottom-up approach with an ILP formulation and i-vector representation [23].

Model comparisons To analyse the contribution of the audio modality, we evaluate two versions of the naming model. The first one CRF_{AV}^{na} is as described in Sec 2.3. The second one, denoted as CRF_{V}^{na} takes into account co-occurrences with the visual data only. Moreover, we provide results with an oracle which is the best obtainable result with a perfect face-name association and propagation system. Thus, oracle errors are faces annotated in the reference but not announced by an OPN and miss face detections. Results have been reported on this dataset in [15]. His approach use the same cues as in this paper (talking head, co-occurrence with OPNs, uniqueness constraint) and combine them in a hierarchical person clustering framework.

2.6. Results

The results are shown in Table 1. The system (1) consists only in the visual biometric models and the monomodal naming model CRF_{V}^{na}. It obtains an EGER of 42.6 % which is comparable to [15].

The system (2) is the same as (1) augmented with the use of the LFB patterns and OPN feature functions. In this case, the CER improves from 6.7% to 5.2%. As mentioned above, the use of LFB recurrence encourages tracks labelled as recurrent to join named clusters to enable their identification in the naming step. We observe that it improves the diarization first by merging small isolated clusters corresponding to different views or poses of a person into that person clusters which co-occur with OPNs. Secondly, this approach enable to avoid wrong assignments of recurrent LFB tracks to figurative person clusters which appear during report and off-voice shot. On the other hand, wrong assignments happen when a face track of a figurative person is labelled as recurrent, but this is less frequent. The contribution of the OPN feature function is less significant, which makes sense given the sparsity of OPN presence. Regarding identification results, it seems that the naming model CRF_{V}^{na} is not able to profit from the clustering improvement since the EGER remains constant.

Table 1. *Face identification in EGER and clustering results in CER for the different systems.*

System	EGER	CER
[15]	46.2	-
(1): f_v+ CRF_V^{na}	42.6	6.7
(2): (1) + f_{opn}+f_{lfb}	42.6	5.2
(3): f_v + f_a+ f_{av} + CRF_{AV}^{na}	40.8	6.2
(4): (3) + f_{opn}+f_{lfb}	36.8	5.4
oracle	32.5	0

With the system (3), the addition of the audio modality seems beneficial to the naming performance. The EGER improves from 42.6% to 40.8%. We found that improvements mainly come from the naming step where the model CRF_{AV}^{na} successfully uses co-occurrences between audio data and OPNs to solve ambiguities in the name-face associations.

The best identification results are achieved by combining all the components with an EGER of 36.8%. It combines the strengths of a person diarization dedicated to identification and an AV naming model able to exploit multimodal co-occurrences and uniqueness constraints. The face diarization obtained with the CRF enables to build more relevant co-occurrence statistics and a better name propagation, and the AV naming model is able to exploit them.

3. CONCLUSION

In this paper, a method is proposed to unsupervisely identify faces using text overlays in broadcast news. It leverage on two CRF models used for person diarization and person cluster identification. The models combine multiple cues from audio and visual sources. Among them, the characterisation of faces using recurrent LFB patterns optimise the face clustering for the naming task, while the use of the audio modality enables to improve face-name association.

Considering the possible extensions of this work, more faces could be identified by integrating predefined biometric models. Such models can be learned in a supervised or unsupervised manner [24]. Face recognition in open sets is a difficult problem, but name entity detection from ASR could provide hypothesis on the faces likely to appear.

4. REFERENCES

[1] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: Naming and detecting faces in news videos," in IEEE *Trans. on Multimedia*, 1999.

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

[3] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth, "Names and faces in the news," in *Proc. of CVPR*, 2004.

[4] M. Moens, T. Tuytelaars, and all, "Naming persons in news video with label propagation," in *Proc. of ICME*, 2010.

[5] 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," in *Proc. of ACM Multimedia*, 2013.

[6] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition," in *Proc. of ICLRE*, 2012.

[7] J. Poignant, L. Besacier, V. Le, S. Rosset, and G. Quénot, "Unsupervised naming of speakers in broadcast tv: using written names, pronounced names or both?," in *Proc. of InterSpeech*, 2013.

[8] J. Poignant, "identification de personnes dans des flux televises," in *Phd thesis*, 2013.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Multiple instance metric learning from automatically labeled bags of faces," in *Proc. of ECCV*, 2010.

[10] K. Deschacht, T. Tuytelaars, M. Moens, et al., "Naming persons in video: Using the weak supervision of textual stories," in *JVCIR*, 2013.

[11] Y. Zhang, W. Wu, Y. Li, C. Jin, X. Xue, and J. Fan, "Automatic name-face alignment to enable cross-media news retrieval," in *Proc of ICAI*, 2013.

[12] D. Anguelov, K. Lee, S. Gokturk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *Proc. of CVPR*, 2007.

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

[14] L. Zhang, D. Kalashnikov, and S. Mehrotra, "A unified framework for context assisted face clustering," in *Proc. of ICMR*, 2013.

[15] H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. Le, T. Napoleon, H. Gao, C. Barras, S. Rosset, L. Besacier, et al., "Qcompere@ repere 2013.," in *Proc. of SLAM*, 2013.

[16] M. Bendris, B. Favre, D. Charlet, G. Damnati, R. Auguste, J. Martinet, G. Senay, et al., "Unsupervised face identification in tv content using audio-visual sources," in *Proc. of CBMI*, 2013.

[17] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise, "A conditional random field approach for audio-visual people diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014.

[18] P. Viola and M. Jones, "Robust real-time face detection," *IJCV*, 2004.

[19] D. Chen and J.-M. Odobez, "Video text recognition using sequential monte carlo and error voting methods," *Pattern Recogn. Lett.*, vol. 26, no. 9, pp. 1386–1403, 2005.

[20] 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.

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

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

[23] 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.

[24] C. Liu, S. Jiang, and Q. Huang, "Naming faces in broadcast news video by image google," in *Proc. of ACM Multimedia*, 2008.