Towards large scale multimedia indexing: A case study on person discovery in broadcast news

Nam Le, Hervé Bredin, Gabriel Sargent, Miquel India, Paula Lopez-Otero, Claude Barras, Camille Guinaudeau, Guillaume Gravier, Gabriel Barbosa da Fonseca, Izabela Lyon Freire, et al.

Towards large scale multimedia indexing: A case study on person discovery in broadcast news

Nam Le[1], Hervé Bredin[2], Gabriel Sargent[3], Miquel India[5], Paula Lopez-Otero[6], Claude Barras[2], Camille Guinaudeau[2], Guillaume Gravier[3], Gabriel Barbosa da Fonseca[4], Izabela Lyon Freire[4], Zenilton Patrocínio Jr[4], Silvio Jamil F. Guimarães[4], Gerard Martí[5], Josep Ramon Morros[5], Javier Hernando[5], Laura Docio-Fernandez[6], Carmen Garcia-Mateo[6], Sylvain Meignier[7], Jean-Marc Odobez[1]
[1] Idiap Research Institute & EPFL, [2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, [3] CNRS, Irisa & Inria Rennes, [4] PUC de Minas Gerais, Belo Horizonte, [5] Universitat Politècnica de Catalunya, [6] University of Vigo, [7] LIUM, University of Maine
nle@idiap.ch,bredin@limsi.fr,gabriel.sargent@irisa.fr,miquel.india@tsc.upc.edu,plopez@gts.uvigo.es

ABSTRACT

The rapid growth of multimedia databases and the human interest in their peers make indices representing the location and identity of people in audio-visual documents essential for searching archives. Person discovery in the absence of prior identity knowledge requires accurate association of audio-visual cues and detected names. To this end, we present 3 different strategies to approach this problem: clustering-based naming, verification-based naming, and graph-based naming. Each of these strategies utilizes different recent advances in unsupervised face / speech representation, verification, and optimization. To have a better understanding of the approaches, this paper also provides a quantitative and qualitative comparative study of these approaches using the associated corpus of the Person Discovery challenge at MediaEval 2016. From the results of our experiments, we can observe the pros and cons of each approach, thus paving the way for future promising research directions.

ACM Reference format:
Nam Le[1], Hervé Bredin[2], Gabriel Sargent[3], Miquel India[5], Paula Lopez-Otero[6], Claude Barras[2], Camille Guinaudeau[2], Guillaume Gravier[3], Gabriel Barbosa da Fonseca[4], Izabela Lyon Freire[4], Zenilton Patrocínio Jr[4], Silvio Jamil F. Guimarães[4], Gerard Martí[5], Josep Ramon Morros[5], Javier Hernando[5], Laura Docio-Fernandez[6], Carmen Garcia-Mateo[6], Sylvain Meignier[7], Jean-Marc Odobez[1] [1] Idiap Research Institute & EPFL, [2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, [3] CNRS, Irisa & Inria Rennes, [4] PUC de Minas Gerais, Belo Horizonte, [5] Universitat Politècnica de Catalunya, [6] University of Vigo, [7] LIUM, University of Maine. 2017. Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proceedings of CBMI, Florence, Italy, June 19-21, 2017,* 6 pages.
https://doi.org/10.1145/3095713.3095732

CBMI, June 19-21, 2017, Florence, Italy

ACM ISBN 978-1-4503-5333-5/17/06…$15.00
https://doi.org/10.1145/3095713.3095732

1 INTRODUCTION

As the retrieval of information on people in videos is of high interest for users, algorithms indexing identities of people and retrieving their respective quotations are vital for searching archives. This practical need leads to research problems on how to index people presence in videos. Started in 2011, the REPERE challenge aimed at supporting research on multimodal person recognition [4, 13]. Its main goal was to answer the two questions *"who speaks when?"* and *"who appears when?"* using any available source of information including pre-existing biometric models and person names extracted from the videos. Thanks to this challenge and the associated multimodal corpus [13], significant progress was achieved in either supervised or unsupervised multimodal person recognition [2, 6, 12, 26, 29].

However, when a content is created or broadcast, it is not always possible to predict which people will be the most important to find in the future and biometric models may not yet be available at indexing time. Under real world conditions, this raises the challenge to index people in the archive when there is no pre-set list of people to index. This makes the task completely unsupervised. To successfully tag people with the correct identities, names must first be detected from audio-visual sources such as automatic transcripts (ASR) or optical character recognition (OCR). Then one must find a way to assign a name correctly to a presence of the corresponding person, and that name must also be propagated to all the shots during which that person appears and speaks.

A standard approach to solve this is first based on face/speech clustering to partition a videos into homogeneous segments corresponding to identities, followed by the assignment of names to segments appropriately. Although commonly used in state-of-the-art systems [21, 26], it has several drawbacks such as potential errors of face/speech clustering or the lack of straightforward way to combine audio-visual streams. In order to alleviate these drawbacks of clustering-based naming two alternative strategies are proposed based on verification and graph optimization. All these three strategies share some common building blocks such as face/speech representation, person diarization, or audio-visual (AV) verification. Though each of these blocks has been well studied within its respective context [3, 23, 31, 32], they have never been fully investigated and compared as whole systems in multimedia indexing context

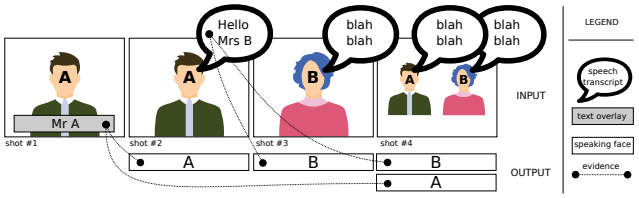

Figure 1: For each shot, participants have to return the names of every speaking face. An evidence is also returned for annotation process.

before. Thus in this paper, we aim to investigate these approaches with variations in their components using the medium scale multimedia dataset associated to the "Multimodal Person Discovery in Broadcast TV" task [5, 25]. The benchmarking results allow the analysis of all three approaches to understand their pros and cons to draw lessons for good practice in large-scale person discovery in broadcast news.

The next Section introduces more details about the Person Discovery challenge, its corpus and evaluation protocol. Then Section 3 gives an overview about our approaches while Sections 4 to 7 describe the methodologies in more details. Section 8 presents experiments and analysis, while Section 9 concludes the paper with further discussions.

2 PERSON DISCOVERY CHALLENGE

The goal of this challenge is to address the indexing of people in archives under real-world conditions when no pre-existing labels or biometric models exist.

Task overview. Participants are provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot $s \in \mathbb{S}$ has to be automatically tagged with the names of people both speaking and appearing at the same time during the shot: this tagging algorithm is denoted by $\mathcal{L} : \mathbb{S} \mapsto \mathcal{P}(\mathcal{N})$.

The list of persons is not provided *a priori*, and person biometric models (neither voice nor face) cannot be trained on external data. The only way to identify a person is by finding their name $n \in \mathcal{N}$ in the audio (*e.g.* using ASR) or visual (*e.g.* using OCR) streams and associating them to the correct person (Fig. 1). We denote by $\mathbb{N}$ the set of all possible person names in the universe, correctly formatted as `firstname_lastname`, while $\mathcal{N}$ is the set of hypothesized names.

Datasets and annotation. The test set is divided into three sets: INA, DW, and 3/24. The INA dataset contains a full week of broadcast for two TV french channels (total duration of 90 hours). The DW dataset [14] is composed of video downloaded from the Deutsche Welle website, in English and German for a total duration of 50 hours. The last dataset contains 13 hours of broadcast from the 3/24 Catalan TV news channel.

Partial annotation was performed to tag each shot with the names of people who appear and speak within that shot using the following approach. From all participant submissions to the challenge, a set of hypotheses were generated for each shot. Then participants also engaged in an interactive annotation process. Detected names were first annotated with thumbnails which were then used to verify whether people appeared and talked in a particular shot. This annotation process yielded 3431 shots with 619 identities annotated (see Tab. 1 for details).

Table 1: Number of identities and corresponding shots where people appear and speak in each set of the corpus.

	DW	INA	3/24	Total
# shots	950	2250	231	3431
# identities	344	232	44	619

Metrics. The task is evaluated indirectly as an information retrieval task. For each query $q \in \mathbb{Q} \subset \mathbb{N}$, returned shots are first sorted by the edit distance between the hypothesized person name and the query q and then by confidence scores. The average precision $\mathrm{AP}(q)$ is then computed based on the list of relevant shots (according to the groundtruth) and the sorted list of shots. Finally, the mean average precision (MAP) is computed as follows:

$$\mathrm{MAP} = \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} \mathrm{AP}(q)$$

Video OCR-NER. As the task we aim at is fully unsupervised, the names of people have to be found in the audio or visual streams. Person identification from automatic ASR transcripts usually deteriorates performance. Meanwhile, video text can be reliably extracted using OCR and names from overlaid texts often coincide temporally with the people visible and speaking. Thus, in this work we use only names coming from OCR segments.

For OCR recognition, we relied on the approaches described in [7]. In brief, first the video is preprocessed with a motion filtering to reduce false alarms, and individual frames are processed to localize the text regions. Then, multiple image segmentations of the same text region are decoded, and all results are compared and aggregated over time to produce several hypotheses. The best hypothesis is used to extract people names for identification. Then MITIE open library[1] is used to perform named entity recognition (NER). To improve the raw MITIE results, a rule-based step identifies names not corresponding to introduced people (e.g. editorial staff, based on their roles like cameraman or writer) since they do not appear within the video.

3 OVERVIEW OF OUR APPROACHES

Conventional approaches for person recognition rely on face and/or voice biometric models. Thus, a very large amount of trained models is needed to cover only a decent percentage of all the people in TV shows. In addition, it is not always possible to predict which people will be the most important to find in the future. To solve these problems, detected people names are assigned to faces and voices following the basic principle that occurrences of similar faces and voices should have the same name. Below, we briefly introduce the 3 different paradigms used in of this paper to solve the task, which have different characteristics (generative vs. discriminative models, pairwise verification vs. global optimization, etc.), while later sections provides more details about them.

Clustering-based naming (CBN). This is the most common approach. Face/speech tracks are first aggregated into homogeneous clusters according to person identities. Then each cluster is tagged

[1] https://github.com/mit-nlp/MITIE

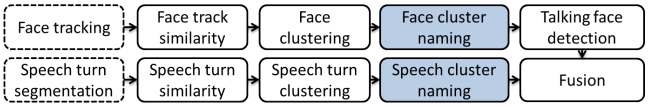

Figure 2: Clustering-based naming process. Light blue boxes are when names are combined with clusters.

with the most probable person name (Fig.2). This approach heavily depends on the clustering quality and granularity: a large number of clusters can significantly reduce the indexing recall, while a too small number may produce false alarms and affect the indexing precision (*i.e.* over-clustering).

Verification-based naming (VBN). To overcome the weakness of CBN, VBN puts higher priority on detected names, and proceed in two main steps (Fig. 3). A person enrolment step relying on face/speech tracks reliably associated with OCR names, and a verification step on all other face/speech segments, which implictly ranks them according to the identity.

Graph-based naming (GBN). VBN propagates names based on a one-one distance while in CBN, all the distances are globally considered. Graph-based naming is thus proposed as an hybrid approach between them. A graph is built using face/speech tracks as a nodes and AV similarities between nodes as edge weights. As in VBN, some nodes are initially tagged with the names, and this information is then propagated along the edges within the graph (Fig. 4).

4 CLUSTERING-BASED NAMING (CBN)

Two tested systems followed this approach (LIMSI and EUMSSI) in which, roughly speaking, a video is first segmented into homogeneous clusters according to person identity using face clustering and speaker diarization, and then clusters are combined with the OCR names to find an optimal assignment (Fig. 2).

4.1 Face clustering

Given the video shots, face clustering consists of (i) face detection, (ii) face tracking (extending detections into continuous tracks), and (iii) face clustering, grouping tracks with the same identity into clusters.

4.1.1 LIMSI system. Face tracking-by-detection is applied within each shot using a detector based on histogram of oriented gradients [8] and the correlation tracker proposed by *Danelljan et al.* [9]. Each face track is then described by its average *FaceNet* embedding and compared with all the others using Euclidean distance [31]. Finally, average-link hierarchical agglomerative clustering is applied. Source code for this module is available in *pyannote-video*[2].

4.1.2 EUMSSI system. A fast version of deformable part-based model (DPM) [11] is first applied. Then tracking is performed using the CRF-based multi-target tracking framework [15], which relies on the unsupervised learning of time sensitive association costs for different features. The detector is only applied 4 times per second and an explicit false alarm classifier at the track level is learned [19]. Each face track is then described using a combination of keypoint matching distances and total variability modeling (TVM) [17, 32].

[2]http://pyannote.github.io

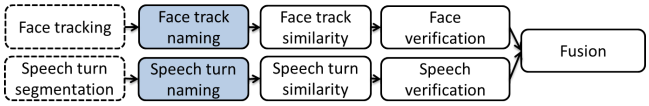

Figure 3: Verification-based naming process. Light blue boxes are when names are combined with face tracks and speech turns to create enrollment models.

4.2 Speaker diarization

The speaker diarization system ("who speaks when?") is based on the LIUM Speaker Diarization system [28], freely distributed[3]. Music and jingle regions are first removed using a Viterbi decoding with 8 GMMs. Then, the diarization system first applies an acoustic Bayesian Information Criterion (BIC)-based segmentation step followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. A Viterbi decoding step re-segments the signal using GMMs for each cluster. In a second step, the background environment information contribution is removed from each GMM cluster through feature gaussianization, and a clustering based on i-vector representation and Integer Linear Programming (ILP) is applied [30].

4.3 Name assignment

After obtaining homogeneous clusters during which distinct identities speak or appear, one needs to assign each name from NER module to the correct clusters. We use a direct naming method [26] to find the mapping that maximizes the co-occurrences between clusters and names. Names are propagated on the outputs of face clustering and speaker diarization independently. A name coming from face naming is ranked based on the talking score of the segment within that shot using lip motion and temporal modeling with LSTM [20].

5 VERIFICATION-BASED NAMING (VBN)

Two systems (GTM-UVigo and UPC) were built on this paradigm which, as summarized in the overview, can be divided in an enrollment and a search (verification) stage (see (Fig. 3) as described below.

5.1 Enrollment

For each identified name in the set of OCR-NER output, the enrollment consists of finding the *speaker segments/face tracks* which best overlap with the temporal occurrence of the OCR name. These tracks/segments are the data used to create a biometric model for the named person. The systems mainly differ in the identification of the associated track and the voice and face representations.

5.1.1 GTM-UVigo system. Given the interval $(t_{\text{start}}, t_{\text{end}})$ associated with the OCR name occurrence (or the set of segments in the case a given name appeared several times), the person speech enrollment segment was extracted by using the whole interval and iteratively extending it in the past and future by 10*ms* step until a change point is detected using the BIC algorithm for speaker segmentation and using standard audio features (19 MFCCs plus

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

delta, acceleration and energy). On the video side, the LIMSI approach was used to detect and track faces, and the track which overlapped most with the OCR temporal segment $(t_{\text{start}}, t_{\text{end}})$ was considered as enrollment data and associated to the voice. Faces were represented with normalized DCT features [1].

Given the audio and video enrollment data, speech segments and face tracks were represented using an i-vector [10] extracted for each modality using the Kaldi toolkit [27]. In case of speech, speech activity detection (SAD) was performed beforehand.

5.1.2 UPC system. Speaker segments/face tracks that overlap with the OCR name segments were obtained as enrollment data. Speaker modelling was implemented using the Alize toolkit[18] by extracting a 400-dimension i-vector [10] (20 MFCCs plus delta and acceleration). Note that OCR names with less than 3s speaker turn enrollment data were discarded.

Regarding video, activations from the last fully connected layer of VGG-face [23] convolutional neural network (CNN) were used to train a triplet network architecture [31] using the FaceScrub and LFW datasets [16, 22]. An autoencoder was used to reduce the dimensionality of the VGG vectors to 1024. The features from each of the detected faces in each track were extracted and then averaged to obtain a single feature vector.

5.2 Search/verification

5.2.1 GTM-UVigo system. To decide which speaker was present in a shot, speech and face detection were first performed. A logistic regression approach was used to classify audio segments as speech or non-speech. For the video, face tracks within the shot were identified, and the one that appeared in more frames (if any) was chosen. Then, the same procedure as in the enrollment stage was performed: features were extracted from the shot and an i-vector was extracted for each modality (after SAD for audio). Given the speech ad face i-vectors of the shot, cosine scoring with the enrollment i-vectors were computed, and the person names that achieved the highest score for each modality were assigned to the shot, provided the scores were greater than a threshold.

5.2.2 UPC system. For the speech modality, target i-vectors were extracted from 3*s* segments with a 0.5*s* shift. The identification was performed evaluating the cosine distance of the i-vectors with each query i-vector. The query with the lowest distance was assigned to the segment. A global distance threshold was previously trained with the development database to discard assignations with high distances.

For the video modality, using the set of named tracks from the full video corpus, a Gaussian Naive Bayes (GNB) binary classifier model was trained, using the euclidean distance between pairs of samples from the named tracks. Then, for each specific video, each unnamed track was compared with all the named tracks of the video, computing the Euclidean distance between the respective feature vectors of the tracks. This value was classified using the GNB to either being a intra-class distance (both tracks belong to the same identity) or an inter-class distance (the tracks are not from the same person). The probability of the distance being intra-class was used as the confidence score. The unnamed track was assigned the identity of the most similar named track. A threshold on the confidence score was used to discard tracks not corresponding to any named track.

6 GRAPH-BASED NAMING

In this approach, all the *speaking faces* of a video are the nodes of a complete and undirected graph $\mathcal{G}$, and each edge between two nodes is weighted by the similarity between their respective voices and/or the face tracks. An initial tagging is done by associating to each face track the co-occurring name(s). Then propagation is performed according to the weights of the graphs using two different strategies, namely MOTIF-RW and MOTIF-MST (Fig. 4).

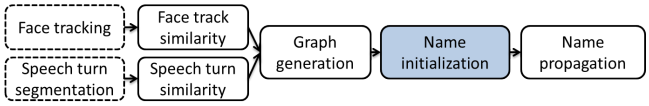

Figure 4: Graph-based naming process. Light blue boxes are when nodes in graph are initiated with names.

6.1 Graph generation details

A node is created for every *speaking face* detected, namely when a face track temporally overlaps a speech segment by at least 60%. If several speech segments overlap it, the face track is associated the one with the most overlapping one. Edges between nodes are weighted using a measure of similarity deriving from the voice and/or face track similarities.

We compute the visual similarity σ_{ij}^V as the cosine between the FaceNet embedding vectors v_i and v_j related to the face tracks of two nodes N_i and N_j: $\sigma_{ij}^V = 1/2 + \frac{v_i \cdot v_j}{||v_i|| * ||v_j||}$, where $\cdot$ is the dot product and $||.||$ is the L2 norm.

The similarity σ_{ij}^A between the speech segments of two nodes N_i and N_j is computed as follows. Each speech segment is modelled with a 16-GMMs over MFCC features. An Euclidean-based approximation of the KL2 divergence, noted δ_{ij}^A, is then computed between the two GMMs [3], and turned into a similarity according to $\sigma_{ij}^A = \exp(\log(\alpha)\ \delta_{ij}^A)$, where $\alpha = 0.25$. The way two modalities can be combined is described in Sec. 7

6.2 Name propagation

Two different approaches are considered for the propagation of the initial tags: a random walk approach and a hierarchical one based on Kruskal's algorithm. In both cases, every node is associated a particular tag with a confidence score at the end of the propagation phase.

6.2.1 Random walk (RW). This method implements a random walk algorithm with absorbing states, adapting [33]. Let n be the number of nodes of $\mathcal{G}$, we compute the probability transition matrix $\mathbf{P}^0$ between all the nodes as $\mathbf{P}^0 = \mathbf{D}^{-1}\mathbf{W}$ where $\mathbf{D}$ is the diagonal *degree matrix* where $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, $1 \leq i \leq n$. Nodes which are already tagged in $\mathbf{P}^0$ are set as *absorbing states*, *i.e.* if i is a tagged node, $\mathbf{P}_{ii}^0 = 1$ and $\mathbf{P}_{ij}^0 = 0$. The random walk iteration is performed according to $\mathbf{P}^{t+1} = (1-\gamma)\ \mathbf{P}^0\ \mathbf{P}^t + \gamma\ \mathbf{P}^0$, where γ is a parameter enforcing consistency with the initial state and slows down the walk (here $\gamma = 0.5$). When the random walk has converged, let T be the

final number of iterations. Each untagged node u is then associated a tagged one l^*, where $l^* = \arg\max_l \mathbf{P}^T_{ul}$. $\mathbf{P}^T_{ul^*}$ is considered as the confidence score related to the tagging of node u.

6.2.2 Minimum spanning tree (MST). This method is based on the computation of a minimum spanning tree, using Kruskal's algorithm. The MST establishes a hierarchical partition of a set [24]. A new connected graph $\mathcal{G}'$ is derived from $\mathcal{G}$ with the same nodes but edge weights representing distances between them (functions of their respective similarities σ^{AV}). To propagate the initial tags, we start from a null graph $\mathcal{H}$ consisting in the nodes of $\mathcal{G}'$ only, and the following process is repeated, until all edges of $\mathcal{G}'$ are examined: from $\mathcal{G}'$, the unexamined edge e corresponding to the smallest distance is chosen. If it does not link different trees in $\mathcal{H}$, skip it; otherwise, it links trees T_1 and T_2 (thus forming T_3), and e is added to the minimum spanning forest $\mathcal{H}$ being created. Three cases are possible: **I.** None of T_1, T_2 is tagged: T_3 will not be tagged **II.** Only T_1 is tagged, with confidence score C_{T_1}: T_1's tag is assigned to the entire T_3 (*i.e.,* to all its unlabelled nodes), with a confidence score $C_{T_3} = C_{T_1} \times (1 - w_e)$, where w_e is the weight of e in $\mathcal{G}'$. **III.** Both T_1 and T_2 are tagged: one of the tags (of T_1 or of T_2) is picked (at random), and assigned to T_3 with confidence scores as in case II.

7 MULTIMODAL FUSION

As names are propagated based on the outputs of face and speech processing modules independently in CBN and VBN systems, we employed a fusion strategy to aggregate the results. Meanwhile, it is more straightforward to combine 2 modalities as joint similarity in GBN.

Late fusion ranking. Within each shot, $\{N_i^F, f(N_i^F)\}$ is the set of names returned by face naming and the corresponding talking scores and $\{N_i^A, s(N_i^A)\}$ is the set of names returned by speaker naming. The final set is the union of N_i^F and N_i^A. The names which the two methods agree on are ranked highest. The same late fusion strategy is applied to both CBN and VBN but with different ranking strategies. For the disjoint names, VBN systems ranks them based on the scores. Meanwhile, for CBN names from talking face naming are ranked higher than speaker naming because we found that face naming is more reliable in empirical experiments.

Audiovisual similarity. For the graph-based approach, the audio and visual modalities can be combined straightforwardly into one similarity. Thus the similarity is extended to multi-modality by using a linear combination of the audio and visual similarities defined in section 6.1: $\sigma_{ij}^{AV} = \beta\sigma_{ij}^{V} + (1-\beta)\sigma_{ij}^{A}$. β is experimentally set to 0.5.

8 EXPERIMENTS

First, contrastive experiments with various configurations are performed for each approach. Then, we conduct a comparative study of the three approaches. All figures are reported using Person Discovery benchmark dataset and the metrics is MAP@K ($K \in 1, 10, 100$). MAP@10 is used as the primary number for comparison.

Baseline. This is when there is no name propagation, *i.e* names are only associated to the most overlapped face / voice. The baseline achieves 55.9%, 33.8%, and 32.8% of MAP@K respectively.

Table 2: MAP@K results of clustering-based naming systems.

	LIMSI			EUMSSI		
	@1	@10	@100	@1	@10	@100
A	29.9	26.2	25.2	29.9	26.2	25.2
V	65.8	46.0	45.0	62.3	50.3	49.2
V-Talking	66.3	46.3	45.4	69.3	57.0	55.8
AV	67.8	47.4	46.4	73.6	59.8	57.9

Table 3: MAP@K results of verification-based naming systems.

	UVigo			UPC		
	@1	@10	@100	@1	@10	@100
A	44.1	36.9	35.9	40.1	35.1	34.7
V	40.9	37.1	35.7	56.7	42.5	41.9
AV	45.6	38.4	37.0	54.8	45.8	45.1

8.1 Contrastive Results

Clustering-based naming. Tab. 2 shows the results using CBM with different settings. The system based solely on speaker diarization (A), which is common for both LIMSI and EUMSSI, is far behind the baseline (29.9% vs. 55.9%) because speech turns are wrongly over-clustered due to dubbing and voice-over. When comparing 2 face clustering methods, LIMSI (V) outperforms EUMSSI (V) at MAP@1 while being slightly behind in MAP@10. This can be explained by the more robust detector used in EUMSSI (V) which detects faces at multiple poses while LIMSI (V) only detects frontal faces which has higher precision. This also explains why after applying talking face detection, EUMSSI (V-talking) has a significant increase while LIMSI (V-talking) only has a minor improvement (6.7% vs. 0.3%). People appearing in frontal faces often are those who talk as well. Finally when AV results are fused, we can observe a substantial improvement in both systems.

Verification-based naming. Tab. 3 shows the results achieved with UVigo and UPC systems. UVigo systems perform better on audio domain than on visual one because the face system only verifies the most dominant face of each shot. Meanwhile for UPC systems, the one based on face verification works better than that of speech processing. UPC face system also has problem when multiple individuals are associated with a single text name. Similarly to CBN, speech processing system cannot be used individually to perform in this task and must be combined with other face system. Multimodal systems slightly improved the performance of monomodal approaches.

Graph-based naming. Tab. 4 gathers the performances obtained by the graph-based systems. We see that the propagation step increases the MAP@K from 7% to 15%. The best performance is obtained by the AV version of RW ($\beta = 0.1$), which outperforms the audio-only ($\beta = 0$) and video-only ($\beta = 1$) versions. The MST system gives the highest result when only vision is considered, which is in favor of a better tuning of β in the audio-visual case.

8.2 Comparative analysis of three approaches

Comparing the MAP@10 of the best configurations, CBN still remains state-of-the-art (59.8%), followed by GBN (57.4%) and VBN

Table 4: MAP@K obtained by graph-based naming systems.

	MOTIF-RW			MOTIF-MST		
	@1	@10	@100	@1	@10	@100
A	67.3	51.6	50.1	62.9	50.1	48.6
V	69.3	53.8	52.1	70.5	56.0	54.3
AV	71.3	57.4	55.5	68.9	55.4	53.6

(45.8%). This shows the possible drawbacks of VBN. The verification models are trained using only one track, which does not contains enough variation. Moreover, this approach is affected more by the quality of OCR-NER as false names can be spread to multiple shots. In the future, some early clustering can help to increase the size of training data while some text filtering can increase the precision of enrolment. On the other hand, GBN requires a face track and a speech turn to be sufficiently overlapped before assigning a name, thus reducing the effect of false texts. The combination of AV similarities also implicitly performs talking face detection, which achieves higher precision in tagging people appearing and speaking. However, discriminative talking detection model still outperforms when applied in CBN systems. Therefore, using this talking face detector in GBN is an interesting future work. VBN can be used to learn more discriminative similarity for GBN edges. Lastly, the effectiveness of combining with audio and visual results is still not as significant as other improvements. This requires further experiments in the future to fully exploit the potential of multimodal processing.

9 FUTURE WORKS

We have presented three different methodologies to perform unsupervised person identification in broadcast news. The quantitative analysis was done on the associated corpus of the Multimodal Person Discovery challenge of MediaEval 2016. In this challenge, person discovery is benchmarked as an index retrieval problem, in which indices represent shots when a person appears and speaks. From the experiments, we can observe that clustering-based methods still achieve better accuracy than the alternatives. The results also suggest potential directions to improve verification-based and graph-based methods by increasing the quality of OCR, hyper parameter tuning, or discriminative talking face detection. On the other hand, these two approaches have many interesting improvements such as discriminative models or unified audio-visual similarity, which can be exploited by combining them with clustering-based methods. Our results also emphasize the importance of multimodal processing, which is a future direction of our work.

Acknowledgement. This work was supported by the EU project EUMSSI (FP7-611057), ANR project MetaDaTV (ANR-14-CE24-0024) project, Camomile project (PCIN-2013-067), and the projects TEC2013-43935-R, TEC2015-69266-P, TEC2016-75976-R, TEC2015-65345-P financed by the Spanish government and ERDF.

REFERENCES

[1] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM MM*, pages 1449–1452. ACM, 2012.

[2] F. Bechet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, G. Linarès, J. Martinet, G. Senay, and P. Tirilly. Multimodal Understanding for Person Recognition in Video Broadcasts. In *Interspeech*, 2014.

[3] 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 *Interspeech*, 2004.

[4] G. Bernard, O. Galibert, and J. Kahn. The First Official REPERE Evaluation. In *SLAM-Interspeech*, 2013.

[5] H. Bredin, C. Barras, and C. Guinaudeau. Multimodal person discovery in broadcast TV at MediaEval 2016. In *MediaEval*, 2016.

[6] H. Bredin, A. Roy, V.-B. Le, and C. Barras. Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identification in TV broadcast. In *IJMIR*, 2014.

[7] D. Chen, J.-M. Odobez, and H. Bourlard. Text detection and recognition in images and video frames. *Pattern Recognition*, 37(3):595–608, 2004.

[8] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[9] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate Scale Estimation for Robust Visual Tracking. In *BMVC*, 2014.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[11] C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013.

[12] P. Gay, G. Dupuy, C. Lailler, J.-M. Odobez, S. Meignier, and P. Deléglise. Comparison of Two Methods for Unsupervised Person Identification in TV Shows. In *CBMI*, 2014.

[13] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The REPERE Corpus : a Multimodal Corpus for Person Recognition. In *LREC*, 2012.

[14] J. Grivolla, M. Melero, T. Badia, C. Cabulea, Y. Esteve, E. Herder, J.-M. Odobez, S. Preuss, and R. Marin. EUMSSI: a Platform for Multimodal Analysis and Recommendation using UIMA. In *COLING*, 2014.

[15] A. Heili, A. Lopez-Mendez, and J.-M. Odobez. Exploiting long-term connectivity and visual motion in crf-based multi-person tracking. *IEEE Transactions on Image Processing*, 23(7):3040–3056, 2014.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, Uni. of Massachusetts, Amherst, 2007.

[17] E. Khoury, P. Gay, and J.-M. Odobez. Fusing Matching and Biometric Similarity Measures for Face Diarization in Video. In *ACM ICMR*, 2013.

[18] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, H. L. Christophe Lévy, J. S.D, Mason, and J.-Y. Parfait. ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition. In *Interspeech*, 2013.

[19] N. Le, A. Heili, D. Wu, and J.-M. Odobez. Temporally subsampled detection for accurate and efficient face tracking and diarization. In *International Conference on Pattern Recognition*. IEEE, Dec. 2016.

[20] N. Le and J.-M. Odobez. Learning multimodal temporal representation for dubbing detection in broadcast media. In *Multimedia*. ACM, 2016.

[21] N. Le, D. Wu, S. Meignier, and J.-M. Odobez. Eumssi team at the mediaeval person discovery challenge. In *MediaEval Workshop*, 2015.

[22] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*. IEEE, 2014.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[24] B. Perret, J. Cousty, J. C. R. Ura, and S. J. F. Guimarães. Evaluation of morphological hierarchies for supervised segmentation. In *ISMM*, 2015.

[25] J. Poignant, H. Bredin, and C. Barras. Multimodal Person Discovery in Broadcast TV at MediaEval 2015. In *MediaEval 2015*, 2015.

[26] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *Interspeech*, 2012.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[28] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech*, Lyon (France), 25-29 Aug. 2013.

[29] M. Rouvier, B. Favre, M. Bendris, D. Charlet, and G. Damnati. Scene understanding for identifying persons in TV shows: beyond face authentication. In *CBMI*, 2014.

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

[31] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015.

[32] R. Wallace and M. McLaren. Total variability modelling for face verification. *Biometrics, IET*, 1(4):188–199, 2012.

[33] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.