Multimodal person discovery using label propagation over speaking faces graphs

Gabriel Barbosa da Fonseca*, Zenilton K. G. Patrocnio Jr*, Guillaume Gravier†, Silvio Jamil Ferzoli Guimares*
*Computer Science Department of PUC Minas,
Belo Horizonrte 31980-110, Brazil
E-mails: gbrl12@gmail.com, zenilton@pucminas.br, sjamil@pucminas.br
†CNRS, IRISA
Rennes 35042, France
E-mail: guig@irisa.fr

***Abstract*—The indexing of large datasets is a task of great importance, since it directly impacts on the quality of information that can be retrieved from these sets. Unfortunately, some datasets are growing in size so fast that manually indexing becomes unfeasible. Automatic indexing techniques can be applied to overcome this issue, and in this study, a unsupervised technique for multimodal person discovery is proposed, which consists in detecting persons that are appearing and speaking simultaneously on a video and associating names to them. To achieve that, the data is modeled as a graph of *speaking-faces*, and names are extracted via OCR and propagated through the graph based on audiovisual relations between speaking faces. To propagate labels, two graph based methods are proposed, one based on random walks and the other based on a hierarchical approach.In order to assess the proposed approach, we use two graph clustering baselines, and different modality fusion approaches. On the MediaEval MPD 2017 dataset, the proposed label propagation methods outperform all literature methods except one, which uses a different approach on the pre-processing step. Even though the Kappa coefficient indicates that the random walk and the hierarchical label propagation produce highly equivalent results, the hierarchical propagation is more than 6 times faster than the random walk under same configurations.**

I. Introduction

With TV channels broadcasting for decades, there is a huge amount of stored content on their archives, and since there are no signs that TV is going to be replaced by other means of communication anytime soon, these archives will continuously grow. The need to make these archives searcheable has led researches to devote a big effort on developing better indexing technologies. Often, the provided video indexing relies on few, and usually subjective tags and small descriptions, which makes large scale searches fairly difficult. A human interest that is not fulfilled by these descriptions is the interest in other people - metadata and annotations usually do not provide information regarding the participants of a video. Also, even when there is some information, it does not cover all appearing persons. It happens since we cannot know if someone with no interest to the public today will become a person of interest in the future. This fact combined with the impossibility of manually labeling entire databases implicate on partially, usually minimally, annotated archives. To solve such problem, many methods to automatically index video databases are studied.

One of the methods used for video indexing is automatically naming people on videos. It consists in detecting persons of interest in a video, name them, and then create a list of persons that appeared on video linked with when they appeared. When this work is performed with no prior information such as biometric models and pre-processed data, this task can be addressed as person discovery on videos. Solving this problem on a unsupervised way is facilitated by multimodal analysis, but choosing which modalities and how to use them can be quite complicated.

Many methods of naming persons on video were developed during the last decades. In a video, there are many sources for extracting information, and in each different work the authors usually focus a specific source to solve a direct task. The result is a vast gamma of strategies that use different means of name extraction, person identification and description, and name-person associations.

One of the first proposed approaches for naming persons is the one proposed in [1], where the authors name characters from the ”Buffy: The vampire slayer” series. In this work, names are extracted from scripts gotten in fan websites. The scripts are then matched with the TV subtitles, for applying temporal information to the extracted names. Finally, the detected names are assigned to detected faces that are temporally co-occurring. Although it is automatic naming process, it is made use of external human-made scripts for the name extraction, and this type of information is usually non existent on other real life scenarios.

In [2], [3], Canseco et al. proposed the first approaches to automatic person identification, with the name extraction based on pronounced names; while the use of biometric models for speaker identification appears in [4]–[6]. However, these audio-only approaches did not achieve good performance because of high error rates due to poor speech transcriptions and bad named entity detection. Similarly, visual-only approaches were very dependent on the quality of overlaid title

This work is related to the M.Sc. dissertation of Gabriel Barbosa da Fonseca.

box transcriptions [7]–[10]. In [11], the authors proposed an approach for naming persons in TV news by extracting names from video transcripts and using graph based label propagation algorithms to spread names to appearing persons.

Two common obstacles found on the works cited above are related to the use of monomodal approaches and to the unsupervised name extraction strategies. Started in 2011, the REPERE challenge aimed at supporting research on multimodal person recognition [12], [13] to overcome the limitations of monomodal approaches. 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 text overlay and speech transcripts). To assess the technology progress, annual evaluations were organized in 2012, 2013 and 2014. Much progress was achieved in either supervised or unsupervised multimodal person recognition [14]–[21]. MediaEval Person Discovery task [22] can be seen as a follow-up campaign with a strong focus on unsupervised person recognition, promoting two campaigns of the Multimodal Person Discovery task, on the years of 2015 and 2016.

To deal with the challenges found when dealing with multiple modalities and name extraction strategies, we propose a label propagation strategy over a graph with audio-visual relationships. The remainder of this document is organized as follows. On Section II, the graph modeling is described. In Section III the two graph-based label propagation approaches are formally presented. On Section IV, the experimental setups is detailed. On Section V the results are shown with a quantitative and qualitative analysis. And finally on Section VI the conclusions are presented.

II. Speaking Faces graph

Common works tend to extract names via audio transcripts or OCR from video overlays (where usually there is a description of the appearing person), and then perform a speech diarization or face clusters to create mono-modal name-cluster associations. In this work, to avoid errors that are ordinarily present in cluster based strategies, a graph based approach is chosen. This approach is a continuation of the one first presented in [23], in which the authors proposed the use of a multimodal graph, where nodes represent persons and the edges are audio-visual similarities between them. Here, this model is referred as a *speaking-face*graph, and its concepts and definitions are described as follows.

To create a representation that fits well on the MPD problem, it was proposed in [23] a multimodal graph representation of speaking persons. In this modeling, a *speaking-face* graph $\mathcal{G} = (V, E)$ is a graph in which each node in V represent a person that appears speaking on a video, and the edges represent audio-visual relations between these nodes. In this graph, each *speaking-face* V_i can have a name Y_i assigned to it. The process for creating a *speaking-faces* graph is described as follows.

First, a video is divided in a set of shots, passed through a face detection and tracking method and a speech diarization method. The set of face tracks and speech turns are represented by FT and ST respectively. Then, names are extracted from the video overlays by applying an OCR followed by an name entity recognition method. The set of names can be represented as Y. A *speaking face* is defined by V_n as the association of a face track FT_i and a co-occurring speech segment ST_j, assumed to belong to the same person. In particular, V_n exists if and only if the intersection of temporal spans of FT_i and ST_j is non-empty. Let the set of *speaking faces* be $V = \{V_n\}_{1 \leq n \leq N}$, $N \in \mathbb{N}$. After the set of *speaking faces* is set for a video, a weighted complete graph $\mathcal{G} = (V, E)$ is calculated, in which each node is a *speaking face* and every pair of nodes V_i and V_j is connected by an edge $E_{i,j} = (V_i, V_j)$ with weight $W_{i,j}$ that represent the similarity between two *speaking-faces*.

For a given pair of *speaking faces*, visual similarity σ^V evaluates the resemblance between face tracks related to it; while audio similarity σ^A measures the proximity between speech segments belonging to the same pair. Thus, audiovisual similarity σ^{AV} between *speaking faces* could be interpreted as a function of visual and audio similarities, *i.e.*, $\sigma^{AV}_{i,j} = f(\sigma^V_{i,j}, \sigma^A_{i,j}), 1 \leq i, j \leq N$.

III. Label Propagation Strategies

In the *speaking-faces* graph model, due to the sparsity of information given by the overlaid person names, usually only a very small portion of data is initially annotated. This highly encourages us to make use of semi-supervised graph based tag propagation approaches to tag the *speaking faces* that were not initially tagged, since for some minimally annotated datasets, the use of semi-supervised approaches has been shown better than the use of supervised ones [24]. In this work two methods are used for propagating tags over *speaking faces*, one as a novel hierarchical approach based on minimum spanning trees, and another as an a adaptation of a commonly utilized tag propagation approach.

In both methods tags are assigned to every *speaking face* detected, leaving none unlabeled at the end of the propagation. Also, it is set a confidence score for each labeled node, representing the level of certainty of that labeling being correct. The confidence score can take values between 0 and 1, with 0 representing a weak correlation between a name and a node, and 1 representing a very strong certainty that a tagging is correct. We assume that the initial tags have a confidence score of 1, and this must not change during the tag propagation phase.

A. Minimum spanning tree based propagation

In the first method, we make use of the Kruskal algorithm on a distance graph for propagating tags between sets hierarchically, based on the propagation proposed in [25]. The novelty in our method is the implementation of a confidence score calculation that allows the propagation to continue even when there is conflict between two different labels. The steps

to perform the MST label propagation (MST_{LP}) are described hereafter.

Algorithm 1: MST_{LP}Algorithm

```
1  MST Propagation ((G, W), where G = (V, E));
   Input  : Partially labeled graph G
   Output: Labeled graph G'
2  Sort E
3  foreach vertex V_i ∈ G do
4  |  MAKE-SET(V_i)
5  end
6  foreach edge E_{i,j} taken in nondecreasing order do
7  |  if FIND-SET(V_i) ≠ FIND-SET(V_j) then
8  |  |  UNION(S_i,S_j)
9  |  |  PROPAGATE(S_i,S_j)
10 |  end
11 end
```

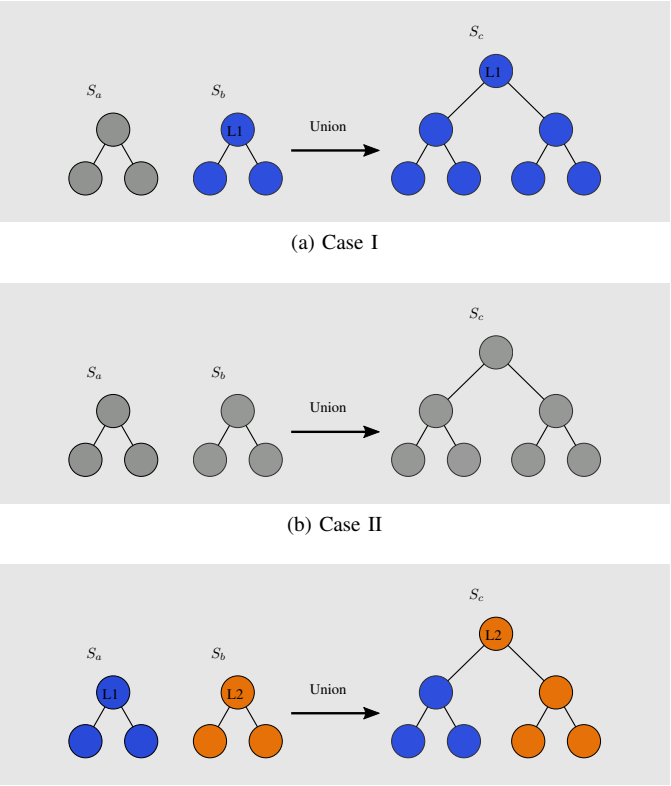
In the Kruskal's algorithm, the graph's edges are sorted in a nondecreasing way, and since the original algorithm treats edges as costs, we must apply the MST_{LP} in a graph G_{sim} where the edge weights W'_{ij} represent distances between *speaking faces*. Then, for each edge taken beginning from the one with the smallest value first, it is checked if this edge connects two different sets or not (step 6 on the Algorithm 1). If it does connect two different sets, a merging of these sets happen, and if not, the selected edge is skipped.

On the MST_{LP}, there is an extra step, and the propagation happens when two disjoint sets are merged, and in this phase, three situations can happen: (i) if only one of the sets is labeled, its label propagates to all nodes belonging to the other set, as illustrated in Figure 1a; (ii) if none of the sets is labeled, nodes of both sets remain unlabeled, illustrated in Figure 1b; and (iii) if both sets are labeled, their labels do not change, and one of the labels is taken to represent the new set formed (this representative label will be the one propagated to other groups when the new set eventually merge with another one), illustrated in Figure 1c. The choose the representing label, their confidence scores are compared, and the one with biggest score is selected. Since there is only one extra operation on the union find step for this algorithm when compared to the original Kruskal's algorithm, the time complexity is still the same. In this case, the complexity is $O(E \log E)$.

To calculate the confidence scores when propagating a label to an unlabeled set, we take into consideration the edge $E_{i,j}$ that united both sets and sets the confidence score of the propagated label based on W'_{ij}, remembering that the initial tags have a confidence score of 1. The confidence score of the new tagged elements will be the result of the product between the last confidence score and the scoring function applied on W'_{ij}.

B. Random Walk label propagation

In order to perform the random walk on a *speaking-faces* graph, the probability matrix P must be created. To do



(a) Case I

(b) Case II

(c) Case III

Fig. 1. Three different cases that can happen on the hierarchical propagation step.

that, first the degree matrix D is calculated by $D_{ii} = \sum_j W_{ij}$, where W is the weight matrix of a *speaking-face*. Than, P is initially defined as $D^{-1}W$, and can be represented in the form of 4 quadrants.

$$P \rightarrow \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$$

The sub-matrix P_{ll} represents the probability of labeled nodes walking to other labeled nodes. P_{lu} represents the probability of labeled nodes randomly walking to unlabeled nodes. P_{ul} and P_{uu} represent the probability of unlabeled nodes walking to labeled nodes and to unlabeled nodes respectively. Since we assume that the initial tags must not change, the initially tagged nodes are set as absorbing states on P, which means that the probability of a tagged node walk to any other node is 0. Thus, after setting labeled nodes as absorbing states, P is represented as follows:

$$P \rightarrow \begin{pmatrix} I & 0 \\ P_{ul} & P_{uu} \end{pmatrix},$$

in which I is an identity matrix. P_{ul} and P_{uu} remain unchanged.

The random walk with t steps is calculated by $P^t = P \times P^{t-1}$, and the number of steps should be enough for P^t reaching convergence. To achieve a random walk based labeling (RW_{LP}) that is consistent with the initial label information, a slowing factor can be applied to the walk, and in

this work it is given by ω. The final random walk with slowing factor is calculated by $P^t = (\omega \times P \times P^{t-1}) + ((1-\omega) \times P)$. As it can be observed, the core of this algorithm is a $V \times V$ matrix multiplication, which leads to a time complexity of $O(n^3)$ if we consider the basic algorithm for matrix multiplication.

With P^t calculated, the label assignment is made based on the P^t_{ul} sub-matrix probabilities. For each unlabeled node in P^t_{ul}, there are the probabilities of it randomly walking to all labeled nodes. The label from the most probable ending node will be applied to each unlabeled node. This maximum probability is also used as the confidence score for the tagging.

A variant of the random walk algorithm for multimodal environments is also proposed. In this variant, named Alternating Random Walk (AltRW), it is created one probability matrix for each modality, and these propability matrices are alternated on each step of the propagation. An AltRW with two modalities A and B is performed by alternating between $P^t = (\omega \times P_A \times P^{t-1}) + ((1-\omega) \times P_A)$ and $P^t = (\omega \times P_B \times P^{t-1}) + ((1-\omega) \times P_B)$. The core of the algorithm would still the same, having the same amount of $V \times V$ matrix multiplications, since the aural and visual matrices have the same size, hence the complexity of this variant still is $O(n^3)$.

IV. Experimental framework

In this section, each part of the proposed approach is precised, with feature extraction details, used dataset, parameters and evaluation choices, among others. A pipeline of the MPD approach is shown on Figure 2.

To evaluate the proposed methods we use the test set of the MediaEval 2016 MPD task, which was manually annotated during the campaign of the respective year [26]. This set is divided in three parts, named as 3-24, INA and DW. The 3-24 is composed by a Catalan TV news channel, named 3/24. The subset used from the INA dataset is composed by 2 different French TV channels. Lastly, the DW dataset is composed by downloaded videos from Deutsche Welle website, containing videos in English and German.

Along with the raw data, the Mediaeval organization also provided a baseline, containing pre-processed data related to all MPD's steps. The provided pre-processed data includes: segmentation of the video stream into shots; detection of the face tracks within the shots; detection and transcription of the overlays from the video frames for finding names; segmentation of the audio stream into speech segments; similarity values between all high-level features; and speech transcription that can be also used for name detection.

A. Feature extraction and preprocessing

The preprocessed data we used from the dataset were: shot segmentation - shots whose duration is less than 1 s or more than 10 s are discarded -, the text detection and recognition by IDIAP [27], the segments of speech obtained with the speaker diarization system from LIUM [28], the facetracks obtained with a histogram of oriented gradients-based detector [29] and a correlation tracker [30]. The features we computed are listed hereinafter:

Name Detection: For the name detection, the text extracted by OCR is then filtered by an name entity detection tool designed for the French language [31].

Visual Features: Two visual features are computed in this work. One is a generic convolutional neural network (CNN) based feature, and the other is also a convolutional network based descriptor, but it is specific for describing faces. The first is extracted from the last layers of a VGG-19 network trained on the ImageNet dataset [32]. The second is the face specific descriptor FaceNet [33].

Acoustic Features: For the audio features we also calculate two different features.

- GMM: For calculating the first feature, each speech segment is described by a sequence of Mel-Frequency Cepstral Coefficients from which is learned a Gaussian Mixture Model with components. Their computation is done using the SPro[1] and Audioseg[2] toolboxes.
- I-VECTOR: For the second feature, an i-vector is calculated. The i-vector for an audio segment is obtained by stacking all the mean coefficients of the GMMs in a supervector, and expressing this supervector in a reduced spaces with emphasizes speaker similarity regarding channel properties [34].

B. Audio-visual Similarities

For calculating the audio-visual similarities between *speaking-faces*, three different modality fusions are used in the present work. The different fusion types are listed hereinafter:

- A early fusion approach: visual and audio features are concatenated in one vector, creating a audio-visual feature, which is then used to calculate similarities between nodes. The cosine similarity is chosen to calculate similarities between the audio-visual feature vectors.
- A intermediate approach: visual and audio similarities are combined using a weighted average, *i.e.*, $\sigma^{AV} = f(\sigma^V, \sigma^A) = \gamma\sigma^A + (1-\gamma)\sigma^V$, in which γ is the range $[0, 1]$
- A late fusion approach: tag-propagation is done for each modality (producing two confidence scores). This is equivalent to use two distinct functions (with $\gamma = 1$ or $\gamma = 0$): $\sigma_1^{AV} = \sigma^V$ and $\sigma_2^{AV} = \sigma^A$. Then, the tag with the highest confidence score is kept for each *speaking face*.

C. Baselines

A more classical approach to tackle the MPD problem is to label elements that are grouped together into clusters. The usual framework applies a clustering method on the elements, and then applies a intra-cluster labeling policy. To assess the proposed label propagation approaches against more naive methods, but without leaving the *speaking-faces* graph

[1]https://gforge.inria.fr/projects/spro/

[2]https://gforge.inria.fr/projects/audioseg/

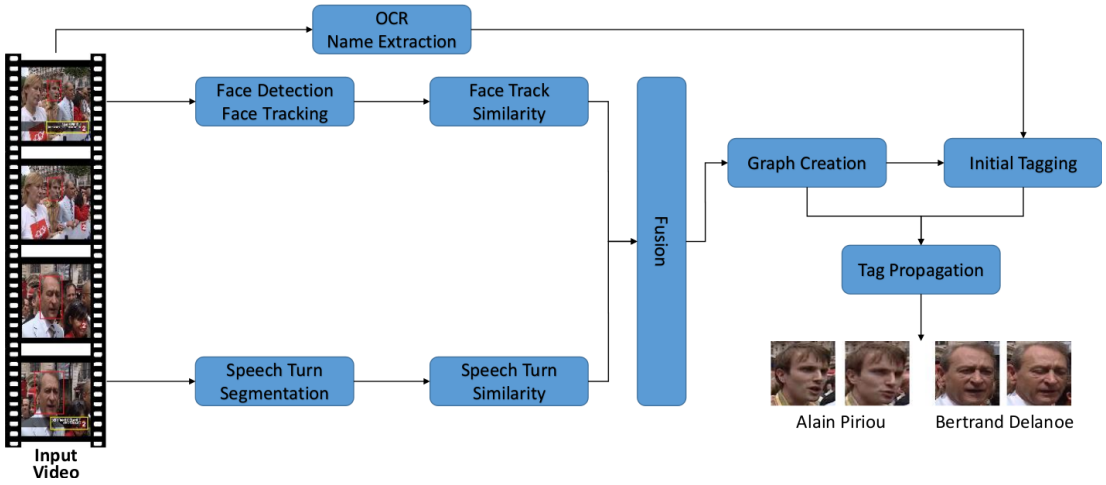


Fig. 2. Diagram illustrating all the steps of the proposed MPD framework.

scenario, two graph clustering baselines are proposed, one using spectral clustering and the other using Markov clustering [35].

The baselines are identical to the proposed methods up to the initial labeling part, differing only on the propagation step. Here graph clustering techniques are used to tag *speaking faces* which were not initially tagged. To perform the baseline tagging, one of the graph clustering methods is applied on a *speaking faces* graph $\mathcal{G}$. The number of clusters is set as the number of distinct tags on each graph plus one, where this one extra cluster represents possible *speaking faces* which do not have a name related to them. After clustering the nodes, a cluster can contain a combination of untagged nodes and nodes with different tags. To decide which tags are going to be propagated, a histogram of tags is calculated for each cluster and the tag with the highest number of incidence on each cluster is used to tag the untagged nodes on that same cluster, with a confidence score set as 0.5. Note that unlike the other propagation methods, in the baseline methods some nodes can remain untagged due to clusters formed by only untagged nodes.

D. Evaluation Metrics

Since the ground-truth of the used dataset is not fully annotated, we consider the Mean Average Precision at K (MAP@K) used in MediaEval[3] [26] to evaluate our frameworks, as if it was a recommendation task. To have complementary insights on the performance of the distinct methods we also use the error rates and recall measures. When measuring the level of agreement of two different configurations we use the Kappa coefficient.

The error rates and recall are calculated as follows: for each video document v, let n^a be the number of (name, shot) c^a couples found by the algorithm and let n^r be the number of (reference name, shot) c^r couples associated to this video. Let N^C be the size of the intersection between c^a and c^r. We allow a small tolerance for matching two tags T_n and T_m ;$1 \leq n, m \leq N$, i.e. when a symmetrized and normalized Levenshtein distance d_L between them is below 0.2. Let N^D be the number of deletions and let N^I the number of insertions to get the list of reference names of the video from the list of estimated names of the algorithm. The error rate Err, and recall R are computed as:

$$Err = \frac{N^D + N^I}{n^r}, \tag{1}$$

$$R = \frac{N^C}{n^r} \tag{2}$$

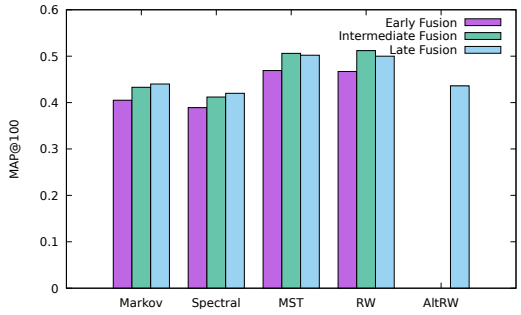
V. Results

In this work, three different fusion modalities are utilized, named early fusion, intermediate fusion and late fusion. To asses the impact of different fusion types on the labeling methods, the three different fusion types are tested on the GMM-CNN and FaceNet-iVector graph configurations. The results for all methods on both configurations are illustrated on Figure 3.

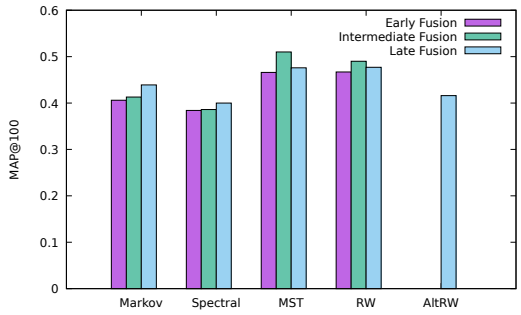
Observing the two bar charts, one can see that the behaviors of all methods remain constant on the two graph configurations with regard to the fusion types. The first observation is that the AltRW achieves the worst results compared to all other RW_{LP} fusion types, showing itself as a bad performing late fusion approach. On the proposed label propagation methods, *i.e.*MST_{LP} and RW_{LP}, the best performing fusion type is the intermediate fusion, followed by the late fusion and early fusion, in this specific order. The comportment of the fusion type results on the graph-clustering based baselines is a bit different, but what is common between all methods, with exception for the AltRW, is that the early fusion approach was the worst performing fusion type.

By analyzing the charts, one can also see that the two proposed label propagation algorithms achieve very similar results, but outperform the two graph-clustering baselines on

[3]we use the script written and provided by Hervé Bredin in the context of the MPD task



(a) CNN GMM fusion



(b) FaceNet iVector fusion

Fig. 3. All fusion strategies for the 2 configurations

all fusion types, showing that for the presented scenario the label propagation algorithms are preferable. The MST_{LP} and RW_{LP} strategies score 0.86 on the Kappa's coefficient, which according to [36] can be considered as an almost perfect agreement. The processing times of both algorithms are measured for propagating labels for the entire dataset[4]. The processing time of the RW_{LP} is of 7m12s, and for the MST_{LP} it is 1m8s, representing a speedup of 6.35 times.

A. Comparison with the state-of-the-art

The proposed methods are compared to literature methods applied to the same dataset. The propagation methods are applied on the CNN-GMM configuration, as it was used by the author in the MediaEval 2016 benchmark. The compared methods are the MST_{LP} and RW_{LP} with intermediate and late fusion variants, including the AltRW.

In Table I the comparative results of the participant teams on MediaEval MPD 2016 and the proposed propagation methods are shown. The best performing method is the one proposed by EUMSSI team [37], and it is the only one not based on speaker and face diarization. Apart from the EUMSSI team, our proposed strategy outperformed all the other literature methods by a significant margin.

When comparing the proposed methods with the ones that used speaker or face diarisation, one can see that the NoProp configuration, which stands for the initial taggin only is almost equivalent to the UPC team [41], and already top the Tokyo Tech, HCMUS [39] and GTM-UVIGO [38] scores. When using the RW_{LP}, which is the best performing of the proposed methods, it outscores the second best method by $0,189$ on MAP@100.

[4]The computational times were measured on an Intel i3-6100 CPU @ 3.70GHz with 4GB of 1333MHz DDR3 RAM

TABLE I
COMPARATIVE RESULTS BETWEEN THE PROPOSED METHODS AND THE LITERATURE. THE PROPOSED METHODS ARE EVALUATED USING THE CNN-GMM CONFIGURATION. THE TWO BEST PERFORMING METHODS ARE HIGHLIGHTED IN BOLDFACE.

Method	MAP@1	MAP@5	MAP@10	MAP@100
[37]	**0.791**	**0.672**	**0.650**	**0.629**
[38]	0.249	0.199	0.188	0.166
[39]	0.100	0.091	0.089	0.086
[40]	0.254	0.173	0.157	0.147
[41]	0.474	0.350	0.335	0.323
NoProp	0.543	0.342	0.323	0.312
MST	0.658	0.546	0.523	0.506
RW	**0.671**	**0.550**	**0.531**	**0.512**
MST_LF	0.659	0.543	0.520	0.502
RW_LF	0.663	0.539	0.517	0.500
AltRW	0.628	0.476	0.452	0.436

VI. CONCLUSION

We believed that creating person specific modeling using multimodal information might result in good data representation for the MPD task, and using semi-supervised label inference methods can work around the sparsity issues of the visually extracted names, increasing the number of indexed persons without sacrificing the labeling correctness. It is showed in this work that the proposed strategy beats all other methods also based on face and/or speaker diarizartion, which enforces the first affirmative. It is also shown that the label propagation methods outperform the graph-clustering baselines, showing that semi-supervised methods are the best choices for the presented scenario. We also presented a evaluation of different modality fusion types, showing that for the label propagation algorithms, the intermediate fusion achieved better results. Also, even if the two proposed label propagation methods produce highly similar results, the $MST_{LP}$is more than 6 times faster than the RW_{LP}.

ACADEMIC PRODUCTION

This study resulted in the following published papers:

- **PUC Minas and IRISA at Multimodal Person Discovery** [42]. In: Working Notes Proceedings of the MediaEval Workshop. 2016.
- **Towards large scale multimedia indexing: A case study on person discovery in broadcast news.** [43]. In: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI) 2017.
- **Tag Propagation Approaches within Speaking Face Graphs for Multimodal Person Discovery.** [44] Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI) 2017.

REFERENCES

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

[2] L. Canseco, L. Lamel, and J. L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, Nov 2005, pp. 415–419.

[3] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *INTERSPEECH*. ICSLP, 2004.

[4] S. E. Tranter, "Who really spoke when? finding speaker turns and identities in broadcast news audio," in *2006 IEEE ICASSP*, vol. 1, May 2006, pp. I–I.

[5] Y. Estève, S. Meignier, P. Deléglise, and J. Mauclair, "Extracting true speaker identities from transcriptions," in *INTERSPEECH 2007 – ICSLP*, 2007, pp. 2601–2604.

[6] J. Mauclair, S. Meignier, and Y. Esteve, "Speaker diarization: About whom the speaker is talking ?" in *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, June 2006, pp. 1–6.

[7] R. Houghton, "Named faces: putting names to faces," *IEEE Intelligent Systems and their Applications*, vol. 14, no. 5, pp. 45–50, Sep 1999.

[8] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: naming and detecting faces in news videos," *IEEE MultiMedia*, vol. 6, no. 1, pp. 22–35, Jan 1999.

[9] J. Yang and A. G. Hauptmann, "Naming every individual in news video monologues," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, NY, USA, 2004, pp. 580–587.

[10] J. Yang, R. Yan, and A. G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, New York, NY, USA, 2005, pp. 31–40.

[11] T. Tuytelaars, M.-F. Moens *et al.*, "Naming people in news videos with label propagation," *IEEE multimedia*, vol. 18, no. 3, pp. 44–55, 2011.

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

[13] J. Kahn, O. Galibert, L. Quintard, M. Carr, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2012, pp. 1–6.

[14] F. Bechet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille *et al.*, "Multimodal understanding for person recognition in video broadcasts." in *INTERSPEECH 2014 – ICSLP*, 2014, pp. 607–611.

[15] M. Bendris, B. Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in tv content using audio-visual sources," in *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2013, pp. 243–249.

[16] H. Bredin, A. Laurent, A. Sarkar, V.-B. Le, S. Rosset, and C. Barras, "Person Instance Graphs for Named Speaker Identification in TV Broadcast," in *Odyssey 2014, The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014.

[17] 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," *International Journal of Multimedia Information Retrieval*, 2014.

[18] P. Gay, G. Dupuy, C. Lailler, J. M. Odobez, S. Meignier, and P. Delglise, "Comparison of two methods for unsupervised person identification in tv shows," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2014, pp. 1–6.

[19] J. Poignant, L. Besacier, and G. Qunot, "Unsupervised speaker identification in tv broadcast based on written names," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 57–68, Jan 2015.

[20] J. Poignant, G. Fortier, L. Besacier, and G. Quénot, "Naming multimodal clusters to identify persons in TV broadcast," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 8999–9023, 2016.

[21] M. Rouvier, B. Favre, M. Bendris, D. Charlet, and G. Damnati, "Scene understanding for identifying persons in tv shows: Beyond face authentication," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2014, pp. 1–6.

[22] J. Poignant, H. Bredin, and C. Barras, "Multimodal person discovery in broadcast TV at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[23] C. E. dos Santos Jr., G. Gravier, and W. Robson Schwartz, "SSIG and IRISA at Multimodal Person Discovery," in *Working Notes Proceedings of the MediaEval Workshop*, Wurzen, Germany, 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01196171

[24] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[25] B. Perret, J. Cousty, J. C. R. Ura, and S. J. F. Guimarães, "Evaluation of morphological hierarchies for supervised segmentation," in *Proceedings of the 12th International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*. Springer, 2015, pp. 39–50.

[26] H. Bredin, C. Barras, and C. Guinaudeau, "Multimodal person discovery in broadcast TV at MediaEval 2016," in *Working notes of the MediaEval 2016 Workshop*, October 2016.

[27] D. Chen and J.-M. Odobez, "Video text recognition using sequential Monte Carlo and error voting methods," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1386–1403, July 2005.

[28] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meigner, "An open-source state of the art toolbox for broadcast news diarization," in *Interspeech*, 2013, pp. 25–29.

[29] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[30] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsber, "Accurate Scale Estimation for Robust Visual Tracking," in *Proceedings of the British Machine Vision Conference*. BMVA Press, September 2014.

[31] C. Raymond, "Robust tree-structured named entities recognition from speech," in *International Conference on Acoustics, Speech and Signal Processing*, 2013.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[34] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[35] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic acids research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[36] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[37] N. Le, S. Meignier, and J.-M. Odobez, "Eumssi team at the mediaeval person discovery challenge 2016," in *MediaEval Benchmarking Initiative for Multimedia Evaluation*, no. EPFL-CONF-223040, 2016.

[38] P. L. Otero, L. Docio-Fernandez, and C. G. Mateo, "Gtm-uvigo system for multimodal person discovery in broadcast tv task at mediaeval 2016," in *MediaEval*, 2016.

[39] V.-T. Nguyen, M.-T. H. Nguyen, Q.-H. Che, V.-T. Ninh, T.-K. Le, T.-A. Nguyen, and M.-T. Tran, "Hcmus team at the multimodal person discovery in broadcast tv task of mediaeval 2016." in *MediaEval*, 2016.

[40] F. Nishi, N. Inoue, K. Iwano, and K. Shinoda, "Tokyo tech at mediaeval 2016 multimodal person discovery in broadcast tv task." in *MediaEval*, 2016.

[41] G. Martí, C. Cortillas, G. Bouritsas, E. Sayrol, J. R. Morros, and J. Hernando, "Upc system for the 2016 mediaeval multimodal person discovery in broadcast tv task," in *MediaEval*, 2016.

[42] G. Sargent, G. B. de Fonseca, I. L. Freire, R. Sicre, Z. K. G. do Patrocínio Jr., S. J. F. Guimarães, and G. Gravier, "Pucminas and IRISA at multimodal person discovery," in *Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[43] N. Le, H. Bredin, G. Sargent, P. Lopez-Otero, C. Barras, C. Guinaudeau, G. Gravier, G. B. da Fonseca, I. L. Freire, Z. Patrocínio Jr *et al.*, "Towards large scale multimedia indexing: A case study on person discovery in broadcast news," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 2017, p. 18.

[44] G. B. Da Fonseca, I. L. Freire, Z. Patrocínio Jr, S. J. F. Guimarães, G. Sargent, R. Sicre, and G. Gravier, "Tag propagation approaches within speaking face graphs for multimodal person discovery," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 2017, p. 15.