

Hierarchical Multi-Label Propagation using Speaking Face Graphs for Multimodal Person Discovery

Gabriel Barbosa da Fonseca, Gabriel Sargent, Ronan Sicre, Zenilton Kleber Gonçalves Do Patrocinio, Guillaume Gravier, Silvio Jamil F. Guimarães

▶ **To cite this version:**

Gabriel Barbosa da Fonseca, Gabriel Sargent, Ronan Sicre, Zenilton Kleber Gonçalves Do Patrocinio, Guillaume Gravier, et al.. Hierarchical Multi-Label Propagation using Speaking Face Graphs for Multimodal Person Discovery. Multimedia Tools and Applications, In press, pp.1-27. 10.1007/s11042-020-09692-x . hal-02926035

HAL Id: hal-02926035
https://hal.science/hal-02926035

Submitted on 31 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hierarchical Multi-Label Propagation using Speaking Face Graphs for Multimodal Person Discovery

Gabriel Barbosa da Fonseca · Gabriel Sargent · Ronan Sicre · Zenilton K. G. Patrocínio Jr · Guillaume Gravier · Silvio Jamil F. Guimarães

Abstract TV archives are growing in size so fast that manually indexing becomes unfeasible. Automatic indexing techniques can be applied to overcome this issue, and this work proposes an unsupervised technique for multimodal person discovery. To achieve this goal, we propose a hierarchical label propagation technique based on quasi-flat zones theory, that learns from labeled and unlabeled data and propagates names through a multimodal graph representation. In this representation, we combine audio, video, and text processing techniques to model the data as a graph of *speaking faces*. In the proposed modeling, we extract names via optical character recognition and propagate them through the graph using audiovisual relationships between speaking faces. We also use a random walk label propagation and two graph clustering strategies to serve as baselines. The proposed label propagation techniques always outperform the clustering baselines on the quantitative assessments. Our approach also outperforms all literature methods tested on the same dataset except for one, which uses a different preprocessing step. The proposed hierarchical label propagation and the random walk baseline produce highly equivalent results according to the Kappa coefficient, but the hierarchical propagation is parameter-free and over 9 times faster than the random walk under the same configurations.

1 Introduction

With TV being one of the main means of communication during the past decades, the amount of content produced and stored by TV channels is extremely vast and is continuously growing in size. Although, it is irrelevant to have an extensive amount of data that is not searchable, and with that in mind many approaches for automatically indexing TV videos were developed. Indexes that represent the identity of people in these archives are essential when searching for content since human nature leads people to be very interested in other people. However, at the moment that content is created or broadcasted,

it is not always possible to predict which people will be the most relevant in the future. For this reason, it is not possible to assume that any model capable of detecting a specific individual will be present at indexing time. This combined with the impossibility of manually labeling entire databases ends up on the creation of partially, usually minimally, annotated archives. To solve such a problem, many methods to automatically index video databases are studied.

The problem of detecting and naming people on videos without supervision can be addressed as a person discovery (PD) task. No prior knowledge, such as person biometric models, should be used on PD since it is an unsupervised problem by definition. To tackle such a task, one can make use of the many sources of information present on a video, using only one channel of information to solve the PD problem (*i.e.*, using visual-only or acoustic-only sources) or using multi-channel analysis. Methods that use multiple channels of information try to take advantage of the multimodal nature of videos to improve results, and when they are applied to tackle the PD problem, we can describe the approach as a multimodal person discovery (MPD).

The first approaches for automatic person identification [9,10] used name extraction techniques based on pronounced names; while other works make use of biometric models for speaker identification [19,35,62]. However, these methods are highly impacted by poor speech transcription and poorly detected named entities. Even if the methods for assigning labels (names in this case) to speakers are good, the use of noisy labels can lead to high error rates. In addition, visual-only approaches were proposed, using overlaid text recognition for extracting name labels. Similarly to the audio-only approaches, the performance of these methods are very dependent on the quality of the extracted names [24,56,67,68]. Tuytelaars *et al.* [63] proposed an approach for naming persons in TV news by extracting names from video transcripts and using graph-based label propagation algorithms to assign names to other appearing persons. Two common obstacles found in these works are related to the use of monomodal approaches and the adoption of unsupervised name extraction strategies.

Started in 2011, the REPERE challenge aimed at supporting research on multimodal person recognition [20,26] to overcome some limitations of monomodal approaches, and annual evaluations were organized in 2012, 2013, and 2014. Much progress was achieved in either supervised or unsupervised multimodal person recognition [4,8,22,25,45,46,49,52,66,73,76].The MediaEval Person Discovery task [47] can be seen as a follow-up campaign, which focused on unsupervised person identification. In this challenge, participants were encouraged to develop multimodal approaches without using any prior biometric models. Two campaigns of the MPD task were promoted in 2015 and 2016, leading to collaborative advances on the task [30,48].

More recently, Azab *et al.* [3] proposed a multimodal optimization approach for speaker naming in movies. Although, they use information from movie subtitles to perform the naming of speakers, which is not always available in the context considered in this work. Similarly, in other works [41,54] movie scripts and video metadata are used to extract person names. Additionally,

Kakaletsis *et al.* propose a method for fast identity label propagation [27], but in this work, initial labels are manually given on the dataset. The manual labeling removes part of the automatic and unsupervised aspects of an MPD approach.

Without having any prior knowledge of person names and by automatically extracting them from visual or acoustic sources, we are prone to getting a small number of good quality labels. To deal with the label shortage, many works on multimedia processing use semi-supervised learning approaches, which build models from both labeled and unlabeled data [32,65,72]. Some of these works use semi-supervised approaches to make name-person assignments, using for example Laplacian support vector machines [66] and label propagation [76].

In this work, we extend the study of a semi-supervised hierarchical label propagation method to propagate initially extracted names to all occurrences of the same person in a video, as proposed in our previous work [14]. This label propagation approach is inspired by a seeded hierarchical image segmentation method [1,42,43]. Image segmentation is a task that consists of separating distinct objects on an image, consequently grouping perceptually coherent objects. Many image segmentation methods rely on the similarity present in the local neighborhood to define boundaries between distinct regions of an image [13,69]. Similarly to the image segmentation proposal, in the context of PD we want to separate distinct persons that appear in a video by relying on their similarities, to be able to correctly propagate labels from labeled persons to unlabeled ones.

Some image segmentation methods have qualities that are greatly desirable in other data processing domains, such as being very computationally efficient or being resistant to extensive amounts of noise present in data [70, 71,11]. With that in mind, being able to adapt image segmentation methods to general data processing can bring many benefits. We propose in this work the reformulation of the previously proposed label propagation method [14] in the form of quasi-flat-zones hierarchy theory [13], using a highly efficient algorithm for propagating labels during the creation of this hierarchy. This formulation opens doors to other adaptations of image segmentation methods to the multimedia processing domain.

To assess the quality of the proposed label propagation method we compare it to a random walk label propagation method, as an adaptation of [74]. Another common way to solve the PD problem is by clustering the appearing persons in a video and then performing cluster-name associations. Even though it looks like a more classical approach, clustering instances is vastly used for naming persons and other multimedia indexing problems [44,49,60, 64]. To compare the graph-based label propagation strategies with more classical approaches, we use two graph clustering methods as baselines for naming persons.

To create a better suited representation considering the specificities of the PD problem, we use in this work multimodal graph-based modeling for the PD task as an extended version of our previous work [14]. In this modeling, we combine feature extraction and multimedia processing techniques available

on the literature such as face detection and tracking [15,16], speech diarization [53], visual feature extraction [58,57], audio feature extraction [21], and named entity recognition [50] to create a multimodal representation for each video. By using this representation, we can effectively exploit the audio-visual relationship between persons appearing in the same video. Besides that, to deal with problems on name extraction which may lead to very few annotations, we propose a label propagation approach that learns from both labeled and unlabeled data by using the multimodal graph topology.

The main contributions presented in this work are twofold: (i) a formulation of the hierarchical propagation using quasi-flat-zone theory and fast component tree creation algorithms, leading to an efficient and parameter-free algorithm and opening doors to the adaptation of other image processing approaches to general multimedia analysis; and (ii) a deeper assessment of the multimodal graph representation and label propagation methods by analyzing how audiovisual weight fusion approaches (early, intermediate and late fusion), feature specificity level, and edge pruning intensities (light, moderate and intense pruning) impacts the whole labeling process.

The remainder of this paper is organized as follows. Section 2 contains the MPD problem formulation with graphs. In Section 3, the strategies for fusing different modalities are presented and discussed. In Section 4, the hierarchical label propagation process is fully described. In Section 5, some experiments and analysis are given to illustrate the relevance of our proposal, and finally, some conclusions are drawn in Section 6.

2 Multimodal graph modeling for MPD

The multimodal person discovery task consists in automatically tagging all shots of a set of broadcast TV videos with the names of people both speaking and appearing at the same time during each shot. As defined in the MediaEval MPD task [7], it is a completely unsupervised task as the list of appearing persons is not provided a priori. Consequently, the only way to identify person names is by extracting them via speech transcription or using optical character recognition (OCR) over video overlays.

Extracting and associating names from audio sources is not a trivial task, containing two major challenges: (i) dealing with noise resulting from the automatic speech transcription; and (ii) correctly associating a spoken name to an appearing person on a video. When using visual information one must also deal with noise on the extracted text, but the person-name association is easier since in most cases there is a temporal overlap between the appearing person and its name. This advantage highly encourages the use of OCR extracted names, but imposes other limitations. The number of shots where a person appears is much larger than the number of shots where its name is visually shown. Also, depending on the broadcasted content, overlaid names are not available, and other sources must be used to extract name information.

To produce the output containing the names of the active persons, *i.e.*, persons that are speaking and appearing at the same time on all shots of a video, one must: (i) detect the active persons on a video; (ii) extract the appearing names; (iii) associate the extracted names to the active person on the same shot; and (iv) propagate the names to other shots where the same active person appears. Many works that tackle the MPD problem propose some kind of clustering-based approach to propagate the initially associated names to other instances of the same person throughout the video. Although, these approaches are prone to limitations of clustering techniques, such as initially selecting the number of clusters for some methods, and dealing with situations where there are multiple labels in one cluster. Badly addressing these issues can lead to mistakes in person-name associations.

An alternative to avoid the clustering issues is the use of semi-supervised label propagation approaches, which use the information of both labeled and unlabeled data to propagate labels properly. Also, as stated in other works, label propagation methods excel where the number of labeled data is excessively small and supervised methods are not a viable option [63,74]. In this work, we propose a hierarchical label propagation approach, inspired by a highly efficient seeded image segmentation algorithm using quasi-flat zone hierarchies (QFZ).

Inspired by [42,43], the pipeline for performing multimodal label propagation over quasi-flat zone hierarchies can be outlined as follows: (i) transformation of the multimedia data into a graph; (ii) computation of a hierarchy from the graph (*e.g.*, quasi-flat zone hierarchy); and (iii) computation of the final label propagation from the hierarchy according to a criterion (*e.g.*, number of labels to be propagated). Here, we compute the QFZ from a graph of active persons on a video named *speaking face* graphs. In this graph, each node represents a person, and each edge represents the audio-visual similarity between two persons. The details for the *speaking face* graph creation are discussed in the following.

2.1 Speaking face graph modeling

To take advantage of the complementary information on multiple data sources and to work around the difficulty of combining several modalities, we have used a graph-based approach to merge multimodal information taking into account audiovisual data. We accomplish this by combining different well-established techniques of multimedia processing to create the multimodal graph representation. We define the creation of this representation as follows.

Given a video, we first segment it into shots. Then, a face detection and a face clustering methods are applied to each shot. This results in clusters of frames which are contiguous in time and are related to a single face. These sequences of frames are denoted face tracks. We also compute a set of speech turns, by applying a speaker diarization method on the video, creating segments of audio related to a single speaker. The set of face tracks and speech

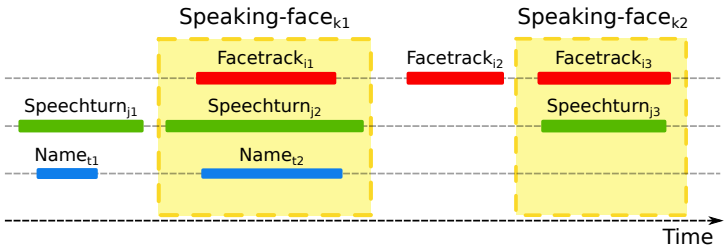


Fig. 1: Speaking face creation

turns are represented by FT and ST, respectively. A *speaking face* V_k represents an association of a face track FT_i with a co-occurring speech segment ST_j, both assumed to belong to the same person, as illustrated in Fig. 1. In particular, V_k exists if and only if the intersection of temporal spans of FT_i and ST_j is non-empty.

To create a representation that fits well on the MPD problem, a multimodal graph representation of *speaking faces* was proposed in [55]. In this modeling, a *speaking face* graph $\mathcal{G} = (V, E)$ is a graph in which each node in V represents a person who appears speaking on a video, and edges represent audiovisual relations between these nodes. Let Y be a set of names extracted from the video. In a graph of *speaking faces*, each vertex V_k can have a name Y_t assigned to it. A *speaking face* graph is illustrated in Fig. 2.

Let $G = (V, E)$ be a speaking face graph in which $V = \{V_k\}_{1 \leq k \leq N},\ N \in \mathbb{N}$ is the set of *speaking faces*. If W is a map from the edge set of G to $\mathbb{R}$, then the pair (G, W) is called a weighted *speaking face* graph. If (G, W) is a weighted *speaking face* graph, for any edge $E_{i,j} = (V_i, V_j) \in E$, with $V_i, V_j \in V$ of G, the value $W_{i,j}$ is called the weight of $E_{i,j}$ (for W), and it stands for the similarity between two *speaking faces* with value in $[0, 1]$, which can be a visual similarity, an acoustic similarity, or their combination. For instance, let v_i and v_j be a pair of *speaking faces*, the visual similarity σ^V measures the resemblance between the face tracks represented by them; while audio similarity σ^A measures the proximity between speech segments also represented by the two *speaking faces*.

2.2 Graph Pruning

On the *speaking face* graph creation, we calculate the similarities between all pairs of nodes on the graph, leading to the creation of complete graphs. However, it is not always possible to get complete graphs under real scenarios. Sometimes the number of nodes is too big, and calculating all similarities between them can be costly. In other circumstances, one can have access to an already preprocessed graph instead of the raw data, and calculating the remaining relationships can be impossible.

To mimic these circumstances and create graphs with missing relationships, we apply a graph pruning with an adaptive threshold on the *speaking face* graphs. To calculate the threshold, let (G, W) be a weighted graph in

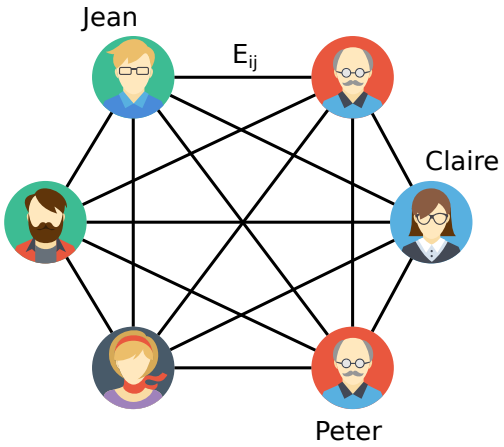


Fig. 2: *Speaking face* graph illustration, in which each node represents a speaking face. Eventually, some nodes are associated with their respectively labels (or names).

which each edge E_i has a weight W_i, and N be the total number of edges on the graph, *i.e.*, $N = |E|$. The method comprises setting a threshold,

$$T = \frac{1}{N}\sum_{i \in E} W_i - \delta \times \sqrt{\frac{1}{N}\sum_{i \in E}(std - W_i)}, \tag{1}$$

in which std is the standard deviation of the similarities of the graph and δ is a parameter that controls the intensity of the pruning.

3 Audiovisual fusion

We use the *speaking face* graph topology to propagate the initial labels to other persons on the same video, based on the audiovisual similarities between them. Multiple sources of data from distinct natures can have complementary information, thus combining these multiple sources can produce more powerful descriptions, and consequently, more discriminative similarities. Although, there are many ways to combine different modalities to produce multimodal relationships, and we study some of them in this work.

According to [17], there are three different fusion categories: data, feature, and decision level fusion, depending on the processing stage where the fusion occurs. Atrey et al. [2] claimed that the most widely used strategy is to fuse multimodal information at the feature level, which is also known as early fusion. The other approach is decision level fusion or late fusion, which combines multiple modalities in the semantic space. Atrey et al. also describe in [2] that a combination of these approaches is also practiced and called a hybrid fusion approach.

In [23], it is presented the existence of multimodal fusion (MF) techniques in diverse research areas. In this work, a class of fusion methods called early fusion is presented. It combines various modalities to generate new modalities by incorporating the latent correlated information. It is also claimed in [23] that fusion is implemented within the feature space in the early stage of multimedia analysis, generally using feature concatenation and feature space transformation to produce a high-dimension feature space. Late fusion is also described as another class of methods where detectors (or classifiers) are trained independently for each modality and then combined to obtain a more comprehensive prediction, usually implemented by combining likelihoods or scores from those detectors.

More recently in [6], multimodal fusion is also referred to as integrating multiple data modalities, their associated features, or the intermediate decisions to perform an analysis task (following the concepts described in [2]). Generally speaking, two types of fusion are considered, depending on the level at which the fusion takes place: early or feature-level fusion and late or decision-level fusion. In naïve early fusion schemes, features from different modalities are concatenated before the learning takes place. According to [2], we can divide fusion methods into the following three categories: rule-based methods, classification-based methods, and estimation-based methods.

From the literature [2,28], it has been observed that many fusion methods such as linear weighted fusion, support vector machine (SVM), and dynamic bayesian networks (DBN) have been used more often in comparison to the other methods. According to [2], this is due to the fact the linear weighted fusion can be easily used to prioritize different modalities while fusing; SVM has improved classification performance in many multimedia analysis scenarios; and the DBN fusion method is capable of handling temporal dependencies among multimodal data, which is an important issue often considered in multimodal fusion. For more extensive treatments on multimodal data fusion, the reader may refer to [2,28].

In this work, we study the impact of three different fusion approaches for audio and visual modalities, named early fusion, intermediate fusion, and late fusion. The different adopted fusion types are described hereinafter:

- *Early fusion approach*: audio and visual features are concatenated in one vector, creating an audiovisual feature, which is then used to compute similarities between nodes. For instance, the cosine similarity is chosen to calculate similarities between these vectors;
- *Intermediate fusion 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 in the range $[0, 1]$;
- *Late fusion approach*: label propagation is done for each modality, producing two labels and two confidence scores for each *speaking face*. This is equivalent to using two distinct weighted average functions: one with $\gamma = 1$; and another one with $\gamma = 0$, resulting in $\sigma_1^{AV} = \sigma^V$ and $\sigma_2^{AV} = \sigma^A$.

Then, the label with the highest confidence score is kept for each *speaking face*.

4 Label Propagation Strategy

Extracting names from either visual overlays or automatically generated speech transcripts leads to a small number of detected names and, usually, only a very small portion of the *speaking faces* is initially annotated. This highly encourages us to make use of semi-supervised graph-based label propagation approaches to label the *speaking faces* that were not initially labeled.

Semi-supervised methods stand between the unsupervised methods and the supervised ones, as they use labeled and unlabeled data together to work. For some minimally annotated datasets, the use of semi-supervised approaches has shown better results than the use of supervised ones [74]. Even though methods like the label propagation in [74] and random walks in [75] show good results in these conditions, their iterative versions consist of a series of matrix multiplications, which are computationally costly.

In this work, a novel hierarchical approach based on **quasi-flat zones** is used for propagating labels over weighted *speaking face* graph. Here, we extract a quasi-flat zone hierarchy (QFZ) from a component tree, which is constructed in a quasi-linear time [36]. After this construction, we guide the propagation of the labels by the QFZ hierarchy in which labels are assigned to every *speaking face* detected, leaving none unlabeled node at the end of the propagation. We calculate 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 weak correctness of an associated name, and 1 standing for a very strong certainty that the labeling is correct. We assume that the initial labels have a confidence score of 1, and this must not change during the label propagation phase.

Using the quasi-linear algorithm for creating a component tree, that depicts a QFZ hierarchy for the weighted *speaking face* graph, depends on the weight map which represents the audiovisual similarities. Taking into account that a label must be first propagated between nodes with high similarity, and without loss of generality, for the component tree creation, we have considered the edges into a non-increasing order of their weight values for forcing the propagation in the edges with high weight values, *i.e.*, high similarity. Contrarily, if the weight values of the edges represent dissimilarity (instead of similarity), we consider the edges in a non-decreasing order of their weight values which represent small distances.

Inspired by [42, 43] which propagate a label through the hierarchy to perform supervised image segmentations, we propose a hierarchical multi-label propagation. Before discussing the proposed method, we give some important definitions.

4.1 Preliminary concepts

Let (G, W) be a weighted *speaking face* graph in which the weight map W represents the audiovisual similarities. Following [13], the hierarchy is constructed according to a weight map that represents distances between the nodes, thus we define (G, W') as a weighted distance *speaking face* graph in which $W' = 1 - W$. We define $\mathbb{E} = [0, 1]$ as the set of values of W'.

Let λ be a number in $\mathbb{E}$. A λ-level edge set of G is the set of all edges in which the weights are smaller than λ. The λ-level graph of G is the subgraph G_λ of G which contains the edges of the λ-level edge set of G, and all vertices of G. We say that a graph G is connected if there exists a path between all pairs of nodes of G. We say that the nodes V' of a subgraph G' of G is a connected component if there is no other subgraph greater than G' in which there exists a path between all nodes of V' in G'. The set of all connected component partitions $\mathrm{C}(G_\lambda)$ induced by the λ-level graphs of G represented by the following sequence

$$QFZ(G) = (C(G_\lambda) \mid \lambda \in \mathbb{E}) \tag{2}$$

is considered a quasi-flat zone hierarchy of G.

The connected components on the lower parts of the QFZ hierarchy (*i.e.* for lower values of λ) contain elements with small distances (or high similarities) between themselves, which we can see as components composed by very similar elements. As the λ value increases, the components of higher levels of the QFZ hierarchy are unions of components of lower levels, and the average distance within these components also increases. This means that elements on a connected component of $C(G_\lambda)$ for large values of λ are not necessarily similar. Thus, we propose a hierarchical multi-label propagation algorithm based on QFZ hierarchies, in which labels are propagated over the connected component partitions on a weighted distance *speaking face* graph for propagating label between connected components on a QFZ hierarchy taking into account an extension of the propagation method proposed in [43] to cope with multi-labels. The choice of which label to propagate depends on a confidence score we compute for each node.

4.2 Hierarchical multi-label propagation

In a previous work [14], we have developed a label propagation based on minimum spanning trees; and, according to [13], a QFZ hierarchy and a minimum spanning tree (MST) are equivalent, thus a QFZ hierarchy can be computed from a component tree or one can compute an MST to produce the same hierarchy since both are equivalent. Here, we take advantage of this equivalence to create an efficient implementation of the QFZ label propagation, using the procedure for creating component trees in quasi-linear time [36].

The procedure for creating the component tree is based on three main operations:

- **Make-Set**(V_i): Create a connected component composed by the element V_i;
- **Find**(V_i): Find the connected component S_i that contains the element V_i;
- **Merge**(V_i, V_j): Create a new connected component by merging the connected components that contain the elements V_i and V_j. Two connected components S_i and S_j that contains V_i and V_j, respectively, can only be merged if they are not the same.

With these three operations, the procedure for computing QFZ_{LP} as showed in Algorithm 1, can be simplified into the following steps: (i) taking a graph (G, W') as input, create a unitary connected component S_i for each element V_i in V; (ii) then, for each edge $E_{i,j}$ taken in non-decreasing order, check if V_i and V_j are in the same connected component by applying **Find**(V_i) and **Find**(V_j); and (iii) if they belong to different connected components, apply **Merge**(V_i, V_j). Repeat steps (ii) and (iii) until all edges are visited.

To perform the hierarchical multi-label propagation, we have proposed a new operation so-called **Propagate** as described in the Algorithm 2. The **Propagate** operation is called whenever two disjoint connected components are merged (right after the operation **Merge**), considering three different situations: (i) if only one of the connected components is labeled, its label propagates to all nodes belonging to the other connected components, as illustrated in Fig. 3a; (ii) if none of the connected components is labeled, nodes of both connected components remain unlabeled, as illustrated in Fig. 3b; and (iii) if both connected components are labeled, their labels do not change, and one of the labels is taken to represent the new connected component formed (this representative label will be the one propagated to other groups when the new connected component eventually merge with another one), as illustrated in Fig. 3c. To choose the representing label for the new merged connected component in the latter case, the confidence scores of the connected components are compared, and the one with the biggest score is selected.

To calculate the confidence scores when propagating a label to an unlabeled connected component, we multiply the weight of the edge $E_{i,j}$ that merged both components by the confidence score of the labeled component.

Algorithm 1: QFZ_{LP} Algorithm

```
1 QFZ Propagation ((G, W'), in which W' = 1 − W);
  Input  : Partially labeled graph distance graph G
  Output: The set of labels L for the speaking faces
2 foreach vertex V_i ∈ G do
3     Make-Set(V_i);
4 foreach edge u = E_{i,j} in a non-decreasing order of W' do
5     if Find(V_i) ≠ Find(V_j) then
6         S_k ← Merge(V_i, V_j);
          /* The S_k label will be either S_i or S_j label, if exists label to
             propagate.                                                    */
7         Propagate(S_k,S_i,S_j,W'(u));
```

Algorithm 2: Pseudo-code for the label propagation from two connected components.

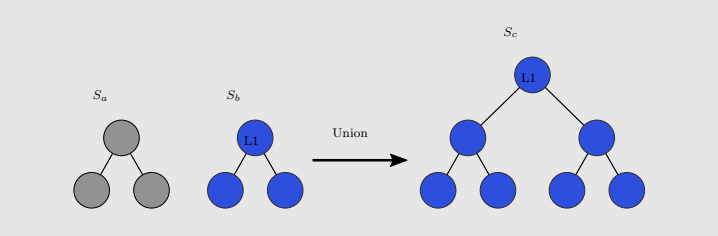
```
Propagate (S_k, S_i, S_j, w);
Input  : The connected components (with confidence scores and labels) and the
         weight value of the edge that connects the subgraphs
Output: Label and confidence of the new connected component
(L_i, C_i) ← LabelAndConfidence(S_i);
(L_j, C_j) ← LabelAndConfidence(S_j);
if L_i and L_j are unlabeled then
    L_k ← Null; C_k ← 0;
else
    if L_i and L_j are labeled then
        if C_i ≤ C_j then
            C_k ← C_j × score_function(w);
            L_k ← L_j;
        else
            C_k ← C_i × score_function(w);
            L_k ← L_i;
    else
        if L_i is labeled then
            C_k ← C_i × score_function(w);
            L_j ← L_i; L_k ← L_i;
        else
            C_k ← C_j × score_function(w);
            L_i ← L_j; L_k ← L_j;
```

It is worth to mention that the initial labels have a confidence score equal to 1. The confidence score of the new labeled connected component will be the product between the confidence score of the merged connected components and a scoring function applied to W'. For instance, the *score_function(w)* may return the similarity value between V_i and V_j, which can be simply referred as $W_{i,j}$. Hence, if a labeling occurs between similar elements, the confidence score will be high (closer to 1), and if it occurs between dissimilar elements, the resulting confidence score will be smaller, closer to 0.

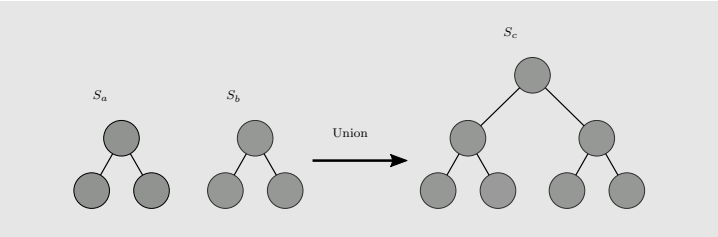
Since there is only one new operation on the union-find step for this algorithm when compared to the original algorithm, the complexity order is still the same. In this case, the complexity is $O(E \times \phi(E))$, where ϕ is a slow-growing diagonal inverse of the Ackermann's function (in [36], $\phi(10^{80}) \approx 4$).

5 Experiments

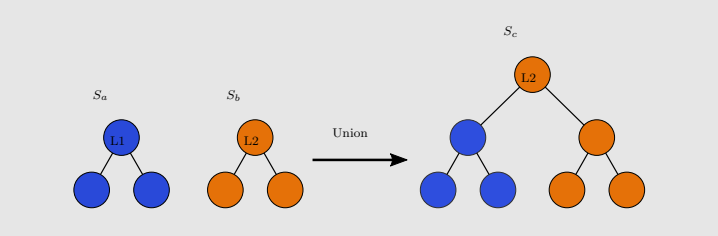
To evaluate the proposed methods, we adopted the test set of the MediaEval 2016 MPD task, which was manually annotated during the campaign of the respective year [7]. This set is divided into three parts, named as 3-24, INA, and DW. The 3-24 part comprises a Catalan TV news channel named 3-24. The second INA part is a subset from the INA dataset composed of two different French TV channels. Lastly, the DW part comprises videos downloaded from



(a) First case



(b) Second case



(c) Third case.

Fig. 3: Merging cases of the hierarchical label propagation.

the Deutsche Welle website, containing videos in English and German. The INA part represents 90 hours of content, the DW part has a total duration of 50 hours, and the 3-24 part has a duration of 13 hours of TV broadcast. The dataset did not include annotations before the MediaEval 2016 event, and it was annotated based on the participants' submissions and feedback. More details about the annotation process can be found in [7]. The final annotation version assembled on October 16, 2016, is used in this work as ground-truth. The ground-truth contains $3,431$ annotated shots, with one or more names assigned to each shot.

Along with the raw data, the MediaEval organization also provided a baseline, containing pre-processed data related to all steps of MPD. Since they give a baseline, one can select key parts of the entire process to tweak, without having to process all other steps that are not related to the improvements. The provided baseline includes:

- Segmentation of the video stream as a sequence of contiguous shots;
- Detection of face tracks within video streams;

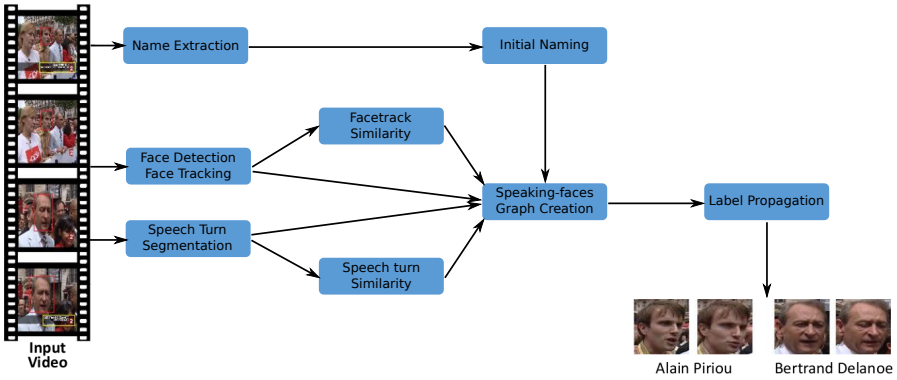


Fig. 4: Block diagram of the MPD framework.

– Detection and transcription of the overlays from video frames for finding names;
– Segmentation of the audio streams into speech segments;
– Similarity values between all high-level features; and
– Speech transcription that can be also used for name detection.

5.1 Pre-Processing and Feature Extraction

The pipeline of the MPD approach applied in this work is illustrated in Fig. 4. The provided pre-computed features are used in some of the framework steps, while in the others, we compute features either to improve results or to fill specific needs of this work. From the provided features, we use: (i) shot segmentation (notice that we discard shots whose duration is less than 1 second or over 10 seconds); (ii) text detection and recognition by IDIAP [12]; (iii) segments of speech obtained with the speaker diarization system from LIUM [53]; (iv) face tracks obtained with a histogram of oriented gradients-based detector [15] and a correlation tracker [16].

We compute two visual features 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. Previous works show how to extract generic visual descriptors from pre-trained Convolutional Neural Networks. Oquab *et al.* [39] extract features from intermediate layers to build mid-level generic visual representations for classification. Razavian *et al.* [51] similarly build descriptions for image retrieval. More recently, Tolias *et al.* [61] use convolutional layers of a pre-trained CNN to efficiently build the MAC and R-MAC descriptors for retrieval, while Sicre *et al.* [58] use both fully connected and convolutional layers outputs to build region descriptors. For calculating visual features, each face track is first represented by its central face, or key face. The image of the face is further described by one of the two descriptors:

- FaceNet: A face specific descriptor based on FaceNet network[57];
- CNN: A generic descriptor [58] based on VGG-19 [59] CNN, trained on the ImageNet dataset.

For the CNN feature, similarly to Tolias et. al. [61], the last convolutional layer of the network is extracted, then an average pooling followed by power normalization is performed, *i.e.*, signed square root and *L2*-normalization. The final descriptor is 512-dimensional and can be used to compute similarities between faces using cosine similarity. The resulting similarity σ^V takes values between 0 and 1 as these visual features are normalized. For the FaceNet descriptors, similarities are also calculated using cosine similarity between those features.

For the audio description we 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 16 components. We compute them by 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 space emphasizing speaker similarity regarding channel properties [21].

For calculating audio features, each speech segment is described by a sequence of Mel-Frequency Cepstral Coefficients (hop size 10 ms, window size 20 ms) from which is learned a Gaussian Mixture Model with 16 components. Two speech segments are compared using a normalized distance approximating of the Kullback-Liebler divergence [5]. We transform this distance into a similarity by:

$$\sigma^A_{i,j} = \exp(\alpha \delta^A_{i,j}), \tag{3}$$

where $\sigma^A_{i,j}$ and $\delta^A_{i,j}$ are the similarity and the distance between segments i and j, respectively. With i-vector descriptors, the computation of the cosine similarity between them incorporates a channel compensation processing which emphasizes again the similarity between channels [18]. In the end, all the similarities are values between 0 and 1, with 1 meaning most similar possible. Two pairs of audiovisual features are created in this work, one containing a more generic audiovisual description (the CNN-GMM combination), another being the combination of a face-specific descriptor and a state-of-the-art audio descriptor (FaceNet-iVector).

To extract initial names from the videos, we use the OCR extracted text provided by the MediaEval dataset. Then, we filter the provided text by applying a name entity detection tool, designed for the French language [50]. We use OCR instead of speech transcripts since a considerable part of the used dataset comprises TV news broadcasts, and visual overlays are regularly

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

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

available. Besides that, the initial name assignment is also easier and more reliable, considering that the names appearing on the screen are related to a person appearing at the same time. It is important to note that when using the proposed framework on a dataset where visual overlays are not available, one must choose another source to extract names.

For the initial labeling of the *speaking faces* graphs, we take the set of extracted names and check if they temporally overlap any of the created speaking faces. If a name Y_i co-occurs (in time) with a *speaking face* V_i, Y_i is assigned to V_i with a confidence score of 1. If the set of names Y is empty for a given video, meaning that no names were automatically extracted from the video, none of the nodes on the graph will be initially labeled. Without initial labels, the label propagation methods have no effect and the detected persons on the video will remain unlabeled.

5.2 Baselines

To compare the proposed label propagation method to a traditional label propagation algorithm, we adopt a random walk with absorbing states as a label propagation baseline. As discussed on [34], random walk methods have been used in tasks ranging from locomotion of animals and descriptions of financial markets to ranking systems. Label propagation can also be achieved by utilizing random walks on graphs. The classification of unlabeled data is made based on the expected random steps required for an unlabeled node to reach each labeled one. Thus, as a baseline for label propagation, we use random walks with absorbing states, adapting the proposal from [74].

To perform a random walk label propagation (RW_{LP}), we first build a probability matrix P from an input graph (G, W). We calculate the transition probabilities based on the similarities between *speaking faces*, generating high transition probabilities between similar elements. To ensure that the initial labels will not change, the initially labeled *speaking faces* are set as absorbing states on P, so the probability of a labeled element taking random steps to any other element is 0. After calculating the random walks with t steps on P, unlabeled elements receive the label from the labeled element that they have the highest probability of randomly walking to. The confidence score of a labeled element V_i that received its label from V_j is equal to the probability $P_{i,j}$. To ensure more consistency with the initial labeling as in [74], we use a slowing factor, which is set to 0.5.

A more classical approach to tackle the MPD problem is to label elements that are grouped into clusters. The usual procedure applies a clustering method on the elements and then adopts an intra-cluster labeling policy [31]. To assess the proposed label propagation approaches against more commonly used methods, but without leaving the *speaking face* graph scenario, we propose two graph clustering baselines, one using spectral clustering and another using Markov clustering.

The graph clustering baselines are identical to the proposed label propagation method up to the initial labeling part, differing only on the propagation step, in which graph clustering techniques are used to label *speaking faces* that were not initially labeled. To perform the baseline labeling, we apply one of the graph clustering methods on a weighted *speaking face* graph (G, W). We set the number of clusters as the number of distinct labels on each graph plus one, with this one extra cluster representing possible *speaking faces* which do not have a name related to them. After clustering the nodes, a cluster can contain a combination of unlabeled nodes and nodes with distinct labels. To decide which labels will be propagated, we calculate a histogram of labels for each cluster, and the label with the highest number of incidence on each cluster is used to label the unlabeled nodes on that same cluster, with a confidence score set to 0.5. Note that unlike the other propagation methods, in the graph clustering baseline methods some nodes can remain unlabeled due to clusters formed only by unlabeled nodes.

5.3 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] [7] to evaluate our proposals, as recommended by the MediaEval MPD task. To have complementary insights on the performance of the distinct methods, we also use the error and recall rates. When measuring the level of agreement of two different configurations, we apply the Kappa coefficient.

To calculate the error and recall rates, 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 labels, *i.e.*, when a symmetrized and normalized Levenshtein distance 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. We then calculate the error rate metric and recall metric by:

$$Err = \frac{N^D + N^I}{n^r}, \tag{4}$$

$$R = \frac{N^C}{n^r}. \tag{5}$$

5.4 Setup

In the experiments, there are parameter settings regarding the audiovisual similarities and graph pruning. For the pruning parameter, we manually set values

[3] We have adopted the same evaluation script written and provided by Hervé Bredin in the context of the MPD task.

of δ as 0, 1 and 2. We selected these values in a way that leads to very light pruning, to a moderate pruning, or to an intense one. On the FaceNet-iVector configuration, the light pruning removes around 1% of the total amount of edges present on the dataset; while on the intense pruning, it removes approximately 80% of the edges.

The α parameter is used for the distance-to-similarity transformation when using GMM audio features, and the γ parameter is used for doing the weighted average on the intermediate fusion of modalities. In Section 5.5 we analyze the impacts of using different value combinations for the two parameters, using the hierarchical label propagation as labeling strategy. Besides this evaluation, we tune the α and γ parameters for the remaining experiments using a 10-fold cross-validation protocol with recall as the evaluation metric and the hierarchical label propagation as labeling strategy. After the tuning, α is set as 0.3 and γ as 0.5 for the CNN-GMM configuration. For the Facenet-iVector configuration, γ is set as 0.3 (there is no α in this configuration since it is only used with the GMM distances).

5.5 Results

To analyze the impacts of the α and γ on our experiments, we test different value combinations for the parameters using the QFZ_{LP} as labeling method. Table 1 shows error and recall rates for different values of α and γ, with the best results (higher recall and smaller errors) highlighted in bold. When using γ set as 0 and 1 we have video-only and audio-only similarities, respectively. The results using monomodal similarities never achieve better results when compared to multimodal similarities. We can observe this behavior on both CNN-GMM and Facenet-iVector configurations.

On the following results, we specify each configuration by the labeling method followed by one of the fusion approaches, where EF stands for early fusion, IF for intermediate fusion, and LF for late fusion. In the first batch of experiments displayed on Table 2 one can observe error and recall rates, along with MAP@K results of the proposed method and the baselines. We highlight the two best scoring methods for each metric in bold. If there is a tie, all methods scoring the best and second-best values are highlighted.

In Table 2a one can observe that all labeling methods improve the results when compared to the initial labeling only (NoProp). This suggests that by only using OCR extracted names it is not possible to correctly name all appearing persons on a video, thus labeling techniques can help to solve this issue. The experiments also show that the label propagation methods RW_{LP} and QFZ_{LP} achieved the best scores on all metrics, using either intermediate fusion or late fusion approaches. The label propagation methods also perform better than the naïve clustering approaches, ranging from 0.500 to 0.512 against 0.412 to 0.440 on MAP@100. This shows that in our context, using semi-supervised learning algorithms leads to better results than using clustering-based labeling processes.

Table 1: Error and recall rates for different α and γ values for the two feature configurations, using QFZ_{LP} for label propagation.

CNN-GMM			
α	γ	Error	Recall
0.3	0.0	0.66	0.47
0.3	0.3	0.51	0.52
0.3	0.5	**0.49**	**0.52**
0.3	0.7	**0.49**	**0.52**
0.3	1.0	0.51	0.50
0.5	0.0	0.66	0.47
0.5	0.3	0.50	0.52
0.5	0.5	**0.49**	**0.52**
0.5	0.7	0.50	0.52
0.5	1.0	0.51	0.50
0.7	0.0	0.66	0.47
0.7	0.3	0.50	0.52
0.7	0.5	**0.49**	**0.52**
0.7	0.7	0.50	0.52
0.7	1.0	0.51	0.50

Facenet-iVector		
γ	Error	Recall
0.0	0.65	0.49
0.3	0.57	**0.51**
0.5	**0.51**	0.52
0.7	**0.51**	**0.51**
1.0	0.53	0.49

In Table 2b we tested the strategies using the combination of a face-specific image descriptor and a state-of-the-art audio descriptor, as opposed to the prior CNN-GMM configuration, which uses good but more generic descriptors. The results show that improving the quality of the features does not necessarily improve the results obtained with the proposed framework. In some cases, like the QFZ_{LP}, the scores are improved by using the FaceNet-iVector configuration, but the opposite happens for the RW_{LP} propagation.

In this work, three different fusion approaches are utilized, named early fusion, intermediate fusion, and late fusion. To assess the impact of different fusion strategies on the labeling methods, the three different fusion strategies are tested on both graph configurations. The results for all methods are shown on Fig. 5a and Fig. 5b. One can observe that the behaviors of all methods remain constant on the two graph clustering approaches concerning the fusion strategies. On the label propagation methods, *i.e.*, QFZ_{LP} and RW_{LP}, the best performing fusion strategy is the intermediate fusion, followed by the late fusion and early fusion, in that order. The behavior on the graph-clustering based baselines is different, but what is common between all methods, is that the early fusion approach was the worst-performing fusion strategy.

By analyzing these results, it is possible to assume that simply applying a generic similarity function over the concatenation of acoustic and visual features does not create better discriminant relationships. This happens since two feature vectors, not normalized and extracted from two different information channels, are combined naively. Using a better suited multimodal feature fusion and a more appropriate similarity metric for the new multimodal features could lead to improvements in the early fusion performance.

We show the results for all propagation methods using intermediate and late fusion under the different pruning levels in Fig. 6a and Fig. 6b. These

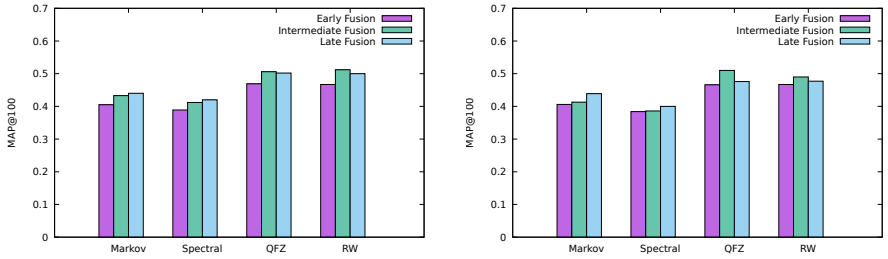
Table 2: Error and recall rates, and MAP@K results for the two graph configurations.

(a) CNN-GMM

Method	Error	Recall	MAP@1	MAP@5	MAP@10	MAP@100
NoProp	0.83	0.18	0.543	0.342	0.323	0.312
Markov_EF	0.72	0.29	0.608	0.447	0.421	0.405
Spectral_EF	0.64	0.38	0.591	0.426	0.405	0.389
QFZ_{LP}_EF	0.54	0.48	0.644	0.510	0.486	0.469
RW_{LP}_EF	0.53	0.49	0.636	0.504	0.482	0.467
Markov_IF	0.60	0.41	0.618	0.471	0.448	0.433
Spectral_IF	0.59	0.44	0.604	0.447	0.426	0.412
QFZ_{LP}_IF	**0.49**	0.52	0.658	**0.546**	**0.523**	**0.506**
RW_{LP}_IF	0.51	**0.54**	**0.671**	**0.553**	**0.531**	**0.512**
Markov_LF	0.64	0.41	0.628	0.479	0.456	0.440
Spectral_LF	0.60	0.44	0.613	0.457	0.436	0.420
QFZ_{LP}_LF	0.53	0.53	0.659	0.543	0.520	0.502
RW_{LP}_LF	**0.49**	**0.54**	**0.663**	0.539	0.517	0.500

(b) FaceNet-iVector

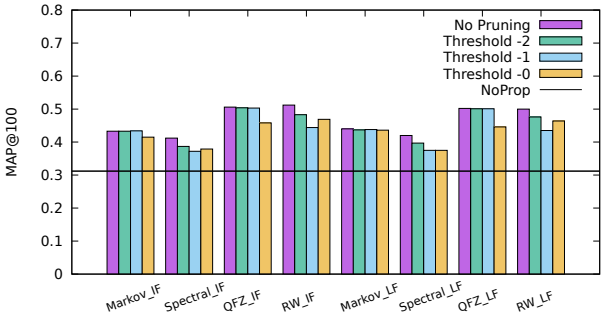
Method	Error	Recall	MAP@1	MAP@5	MAP@10	MAP@100
NoProp	0.83	0.18	0.543	0.342	0.323	0.312
Markov_EF	0.72	0.29	0.601	0.447	0.423	0.406
Spectral_EF	0.65	0.37	0.591	0.420	0.400	0.384
QFZ_{LP}_EF	0.53	0.49	0.634	0.503	0.482	0.466
RW_{LP}_EF	0.52	0.50	0.641	0.502	0.482	0.467
Markov_IF	0.62	0.42	0.604	0.443	0.426	0.413
Spectral_IF	0.68	0.42	0.594	0.417	0.398	0.386
QFZ_{LP}_IF	**0.57**	0.51	**0.669**	**0.550**	**0.528**	**0.510**
RW_{LP}_IF	0.59	**0.53**	**0.659**	**0.535**	**0.508**	**0.490**
Markov_LF	0.62	0.44	0.626	0.478	0.454	0.439
Spectral_LF	0.63	0.42	0.604	0.433	0.414	0.400
QFZ_{LP}_LF	**0.58**	**0.52**	0.653	0.515	0.493	0.476
RW_{LP}_LF	0.59	**0.52**	0.649	0.520	0.494	0.477



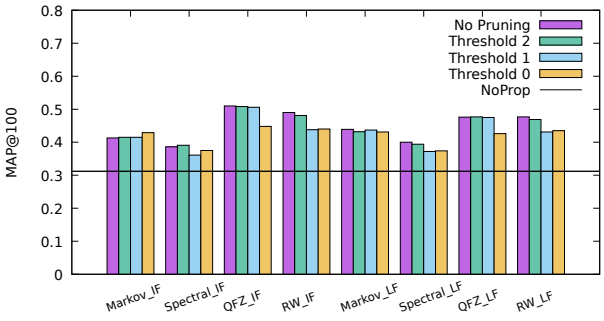
(a) CNN-GMM (b) FaceNet-iVector

Fig. 5: Comparative results between the different types of fusion on all propagation methods using the FaceNet-iVector and CNN-GMM configurations.

results are heterogeneous over the methods at a certain level, but they show in most cases that pruning is prejudicial to the propagation methods regarding the evaluation scores. Although, the score differences are not substantial, and even with a very low quantity of edges the propagation methods can still per-



(a) CNN-GMM



(b) FaceNet-iVector

Fig. 6: Comparative results between the different pruning intensities on all propagation methods. The black line represents the MAP@100 score for when there is no propagation used.

form well and improve the efficacy compared to not using any labeling method (represented by the horizontal line on the plots). This is a very interesting characteristic, meaning that the proposed method still performs well on scenarios where there is a low percentage of edges, which can decrease drastically the computational cost without suffering much loss on the resulting labeling.

Table 3 shows the comparative results of the participant teams on MediaEval MPD 2016 and the proposed propagation methods using the Facenet-iVector descriptors. The best performing method is proposed by EUMSSI team [31], 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 diarization, it can be observed that the NoProp configuration (which stands for the initial labeling only) is almost equivalent to the UPC team [33],

Table 3: Comparative results between the proposed methods using the FaceNet-iVector descriptors and the literature. For each metric, the two best performing methods are highlighted in boldface.

Method	MAP@1	MAP@5	MAP@10	MAP@100
MediaEval Participants Methods				
EUMSII [31]	**0.791**	**0.672**	**0.650**	**0.629**
GTM-UVIGO [40]	0.249	0.199	0.188	0.166
HCMUS [37]	0.100	0.091	0.089	0.086
Tokyo Tech [38]	0.254	0.173	0.157	0.147
UPC [33]	0.474	0.350	0.335	0.323
Our Methods				
NoProp	0.543	0.342	0.323	0.312
Markov_EF	0.601	0.447	0.423	0.406
Spectral_EF	0.591	0.420	0.400	0.384
QFZ_{LP}_EF	0.634	0.503	0.482	0.466
RW_{LP}_EF	0.641	0.502	0.482	0.467
Markov_IF	0.604	0.443	0.426	0.413
Spectral_IF	0.594	0.417	0.398	0.386
QFZ_{LP}_IF	**0.669**	**0.550**	**0.528**	**0.510**
RW_{LP}_IF	0.659	0.535	0.508	0.490
Markov_LF	0.626	0.478	0.454	0.439
Spectral_LF	0.604	0.433	0.414	0.400
QFZ_{LP}_LF	0.653	0.515	0.493	0.476
RW_{LP}_LF	0.649	0.520	0.494	0.477

and already better than the Tokyo Tech, HCMUS [37] and GTM-UVIGO [40] scores. When using the proposed hierarchical label propagation QFZ_{LP}, it outscores the second-best method by 0.183 on MAP@100.

To further compare the two best-performing label propagation methods (RW_{LP}_IF and QFZ_{LP}_IF on the MAP@K metric), we measure their level of agreement using the Kappa coefficient and compare their processing time. The Kappa coefficient scored a level of agreement of 0.847 between the two methods, which according to [29] can be considered as an almost perfect agreement. The processing time for the RW_{LP}_IF is 184.72 seconds, while for the QFZ_{LP}_IF it is only 19.8 seconds. The processing time comprises the average total time to process all graphs in one configuration (Facenet-iVector without pruning), ignoring only the graphs without initial labels. This shows that even though the hierarchical approach reaches slightly lower scores when compared to the random walk, its results are highly equivalent to the best performing propagation method and it achieves a speedup of 9.33 times.

6 Conclusion

In this work, we proposed the use of *speaking face* graphs along with an efficient, parameter-free, graph-based hierarchical label propagation approach to tackle the multimodal person discovery task. The hierarchical propagation was inspired by an image segmentation approach, and thanks to the theoret-

ical framework used, this work opens doors to more adaptations of efficient graph-based image processing methods to the multimedia analysis context.

We showed that the proposed label propagation method and the label propagation baseline outperform conventional graph clustering techniques for the selected database. We also showed that our proposed method performs better than the literature methods tested in the same dataset, except for one method that does not use conventional speaker and face diarization as pre-processing. We showed that using multiple modalities increases the resulting score of our propagation methods when compared to using only visual or acoustic modalities. We also showed that pruning the graphs impacts the RW_{LP} and QFZ_{LP} methods negatively, but the score losses are small, even when a large number of edges are pruned. For the modality fusion strategies, its inconclusive if the intermediate or late fusion performs better, but both outperform the early fusion strategy. Furthermore, it was statistically shown that the QFZ_{LP} with late fusion and the RW_{LP} produce highly equivalent results, but the proposed hierarchical method is more than 9 times faster.

Acknowledgements The authors thank Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – (Universal 421521/2016-3 and PQ 310075/2019-0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES – (Grant STIC-AmSUD MOTIF 001/2013 and STIC-AmSUD TRANSFORM 88881.143258/2017-01) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – (Grants PPM-00006-18, APQ-01806-13 and CEX-APQ-03195-13). This work was also partially supported by the STIC AmSud program, under the project "Unsupervised Mining of Multimedia Content", and by the Inria Associate Team program.

References

1. de Almeida, C.S.J., Cousty, J., Perret, B., do Patrocínio Jr., Z.K.G., Guimarães, S.J.F.: Label propagation guided by hierarchy of partitions for superpixel computation. In: E. Ricci, S.R. Bulò, C. Snoek, O. Lanz, S. Messelodi, N. Sebe (eds.) Image Analysis and Processing - ICIAP 2019 - 20th International Conference, Trento, Italy, September 9-13, 2019, Proceedings, Part II, *Lecture Notes in Computer Science*, vol. 11752, pp. 3–13. Springer (2019). DOI 10.1007/978-3-030-30645-8_1. URL `https://doi.org/10.1007/978-3-030-30645-8_1`
2. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6), 345–379 (2010)
3. Azab, M., Wang, M., Smith, M., Kojima, N., Deng, J., Mihalcea, R.: Speaker naming in movies. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2206–2216 (2018)
4. Bechet, F., Bendris, M., Charlet, D., Damnati, G., Favre, B., Rouvier, M., Auguste, R., Bigot, B., Dufour, R., Fredouille, C., et al.: Multimodal understanding for person recognition in video broadcasts. In: International Conference on Spoken Language Processing (ICSLP), pp. 607–611 (2014)
5. Ben, M., Betser, M., Bimbot, F., Gravier, G.: Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In: Proceedings of the 8th International Conference on Spoken Language Processing, pp. 333–444 (2004)
6. Bernal, E.A., Yang, X., Li, Q., Kumar, J., Madhvanath, S., Ramesh, P., Bala, R.: Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. IEEE Transactions on Multimedia **20**(1), 107–118 (2017)

7. Bredin, H., Barras, C., Guinaudeau, C.: Multimodal person discovery in broadcast TV at MediaEval 2016. In: Working notes of the MediaEval 2016 Workshop (2016)
8. Bredin, H., Roy, A., Le, V.B., Barras, C.: 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)
9. Canseco, L., Lamel, L., Gauvain, J.L.: A comparative study using manual and automatic transcriptions for diarization. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 415–419 (2005)
10. Canseco-Rodriguez, L., Lamel, L., Gauvain, J.L.: Speaker diarization from speech transcripts. In: International Conference on Spoken Language Processing (ICSLP), pp. 1272–1275 (2004)
11. Cayllahua-Cahuina, E., Cousty, J., Guimarães, S.J.F., Kenmochi, Y., Cámara-Chávez, G., de Albuquerque Araújo, A.: Hierarchical segmentation from a non-increasing edge observation attribute. Pattern Recognition Letters **131**, 105–112 (2020)
12. Chen, D., Odobez, J.M.: Video text recognition using sequential Monte Carlo and error voting methods. Pattern Recognition Letters **26**(9), 1386–1403 (2005)
13. Cousty, J., Najman, L., Kenmochi, Y., Guimarães, S.: Hierarchical segmentations with graphs: Quasi-flat zones, minimum spanning trees, and saliency maps. Journal of Mathematical Imaging and Vision **60**(4), 479–502 (2018). DOI 10.1007/s10851-017-0768-7
14. Da Fonseca, G.B., Freire, I.L., Patrocínio Jr, Z., Guimarães, S.J.F., Sargent, G., Sicre, R., Gravier, G.: Tag propagation approaches within speaking face graphs for multimodal person discovery. In: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI), p. 15. ACM (2017)
15. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)
16. Danelljan, M., Häger, G., Shahbaz Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference. BMVA Press (2014)
17. Dasarathy, B.V.: Decision fusion, vol. 1994. IEEE Computer Society Press Los Alamitos, CA (1994)
18. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (2011)
19. Estève, Y., Meignier, S., Deléglise, P., Mauclair, J.: Extracting true speaker identities from transcriptions. In: International Conference on Spoken Language Processing (ICSLP), pp. 2601–2604 (2007)
20. Galibert, O., Kahn, J.: The first official repere evaluation. In: First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013) (2013)
21. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of i-vector length normalization in speaker recognition systems. In: 12th Annual Conference of the International Speech Communication Association (2011)
22. Gay, P., Dupuy, G., Lailler, C., Odobez, J.M., Meignier, S., Deléglise, P.: Comparison of two methods for unsupervised person identification in tv shows. In: 12th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2014)
23. Geng, J., Miao, Z., Zhang, X.P.: Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection. IEEE Transactions on Multimedia **17**(4), 498–511 (2015)
24. Houghton, R.: Named faces: putting names to faces. IEEE Intelligent Systems and their Applications **14**(5), 45–50 (1999)
25. Hu, Y., Ren, J.S., Dai, J., Yuan, C., Xu, L., Wang, W.: Deep multimodal speaker naming. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1107–1110. ACM (2015)
26. Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., Joly, P.: A presentation of the repere challenge. In: 10th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2012)

27. Kakaletsis, E., Zoidi, O., Tsingalis, I., Tefas, A., Nikolaidis, N., Pitas, I.: Fast constrained person identity label propagation in stereo videos using a pruned similarity matrix. Signal Processing: Image Communication **67**, 199–209 (2018)
28. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: An overview of methods, challenges, and prospects. Proceedings of the IEEE **103**(9), 1449–1477 (2015)
29. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
30. Le, N., Bredin, H., Sargent, G., Lopez-Otero, P., Barras, C., Guinaudeau, C., Gravier, G., da Fonseca, G.B., Freire, I.L., Patrocínio Jr, Z., 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 (CBMI), p. 18. ACM (2017)
31. Le, N., Meignier, S., Odobez, J.M.: Eumssi team at the mediaeval person discovery challenge 2016. In: Working Notes Proceedings of the MediaEval 2016 Workshop, EPFL-CONF-223040 (2016)
32. Ma, Z., Nie, F., Yang, Y., Uijlings, J.R., Sebe, N., Hauptmann, A.G.: Discriminating joint feature analysis for multimedia data understanding. IEEE Transactions on Multimedia **14**(6), 1662–1672 (2012)
33. Martí, G., Cortillas, C., Bouritsas, G., Sayrol, E., Morros, J.R., Hernando, J.: Upc system for the 2016 mediaeval multimodal person discovery in broadcast tv task. In: Working Notes Proceedings of the MediaEval 2016 Workshop (2016)
34. Masuda, N., Porter, M.A., Lambiotte, R.: Random walks and diffusion on networks. Physics Reports **716-717**, 1 – 58 (2017). DOI https://doi.org/10.1016/j.physrep.2017.07.007
35. Mauclair, J., Meignier, S., Esteve, Y.: Speaker diarization: About whom the speaker is talking? In: IEEE Odyssey - The Speaker and Language Recognition Workshop, pp. 1–6 (2006)
36. Najman, L., Couprie, M.: Building the component tree in quasi-linear time. IEEE Transactions on Image Processing **15**(11), 3531–3539 (2006)
37. Nguyen, V.T., Nguyen, M.T.H., Che, Q.H., Ninh, V.T., Le, T.K., Nguyen, T.A., Tran, M.T.: Hcmus team at the multimodal person discovery in broadcast tv task of mediaeval 2016. In: Working Notes Proceedings of the MediaEval 2016 Workshop (2016)
38. Nishi, F., Inoue, N., Iwano, K., Shinoda, K.: Tokyo tech at mediaeval 2016 multimodal person discovery in broadcast tv task. In: Working Notes Proceedings of the MediaEval 2016 Workshop (2016)
39. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
40. Otero, P.L., Docio-Fernandez, L., Mateo, C.G.: Gtm-uvigo system for multimodal person discovery in broadcast tv task at mediaeval 2016. In: Working Notes Proceedings of the MediaEval 2016 Workshop (2016)
41. Pang, L., Ngo, C.W.: Unsupervised celebrity face naming in web videos. IEEE Transactions on Multimedia **17**(6), 854–866 (2015)
42. Perret, B., Cousty, J., Guimarães, S.J.F., Maia, D.S.: Evaluation of hierarchical watersheds. IEEE Trans. Image Processing **27**(4), 1676–1688 (2018). DOI 10.1109/TIP.2017.2779604. URL `https://doi.org/10.1109/TIP.2017.2779604`
43. Perret, B., Cousty, J., Ura, J.C.R., Guimarães, S.J.F.: 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, pp. 39–50. Springer (2015)
44. Pham, P.T., Moens, M., Tuytelaars, T.: Cross-media alignment of names and faces. IEEE Transactions on Multimedia **12**(1), 13–27 (2010). DOI 10.1109/TMM.2009.2036232
45. Pini, S., Cornia, M., Bolelli, F., Baraldi, L., Cucchiara, R.: M-vad names: a dataset for video captioning with naming. Multimedia Tools and Applications **78**(10), 14,007–14,027 (2019)
46. Poignant, J., Besacier, L., Quénot, G.: Unsupervised speaker identification in tv broadcast based on written names. IEEE Transactions on Audio, Speech, and Language Processing **23**(1), 57–68 (2015)

47. Poignant, J., Bredin, H., Barras, C.: Multimodal person discovery in broadcast TV at mediaeval 2015. In: Working Notes Proceedings of the MediaEval 2015 Workshop (2015)
48. Poignant, J., Bredin, H., Barras, C.: Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015. Multimedia Tools and Applications **76**(21), 22,547–22,567 (2017)
49. Poignant, J., Fortier, G., Besacier, L., Quénot, G.: Naming multi-modal clusters to identify persons in TV broadcast. Multimedia Tools and Applications **75**(15), 8999–9023 (2016)
50. Raymond, C.: Robust tree-structured named entities recognition from speech. In: International Conference on Acoustics, Speech and Signal Processing (2013)
51. Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications **4**(3), 251–258 (2016)
52. Rohrbach, A., Rohrbach, M., Tang, S., Joon Oh, S., Schiele, B.: Generating descriptions with grounded and co-referenced people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4979–4989 (2017)
53. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meigner, S.: An open-source state of the art toolbox for broadcast news diarization. In: INTERSPEECH, pp. 25–29 (2013)
54. Sang, J., Xu, C.: Robust face-name graph matching for movie character identification. IEEE Transactions on Multimedia **14**(3), 586–596 (2012)
55. dos Santos Jr., C.E., Gravier, G., Robson Schwartz, W.: SSIG and IRISA at Multimodal Person Discovery. In: Working Notes Proceedings of the MediaEval 2015 Workshop. Wurzen, Germany (2015)
56. Satoh, S., Nakamura, Y., Kanade, T.: Name-it: naming and detecting faces in news videos. IEEE MultiMedia **6**(1), 22–35 (1999)
57. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
58. Sicre, R., Rabin, J., Avrithis, Y., Furon, T., Jurie, F., Kijak, E.: Automatic discovery of discriminative parts as a quadratic assignment problem. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1059–1068 (2017)
59. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR) (2015)
60. Somandepalli, K., Kumar, N., Guha, T., Narayanan, S.S.: Unsupervised discovery of character dictionaries in animation movies. IEEE Transactions on Multimedia **20**(3), 539–551 (2018)
61. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. International Conference on Learning Representations (ICLR) (2016)
62. Tranter, S.E.: Who really spoke when? finding speaker turns and identities in broadcast news audio. In: 2006 IEEE ICASSP, vol. 1, pp. I–I (2006)
63. Tuytelaars, T., Moens, M.F., et al.: Naming people in news videos with label propagation. IEEE Multimedia **18**(3), 44–55 (2011)
64. Vallet, F., Essid, S., Carrive, J.: A multimodal approach to speaker diarization on tv talk-shows. IEEE Transactions on Multimedia **15**(3), 509–520 (2013)
65. Wu, J., Zhao, S., Sheng, V.S., Zhang, J., Ye, C., Zhao, P., Cui, Z.: Weak-labeled active learning with conditional label dependence for multilabel image classification. IEEE Transactions on Multimedia **19**(6), 1156–1169 (2017)
66. Xiong, C., Gao, G., Zha, Z., Yan, S., Ma, H., Kim, T.K.: Adaptive learning for celebrity identification with video context. IEEE Transactions on Multimedia **16**(5), 1473–1485 (2014)
67. Yang, J., Hauptmann, A.G.: Naming every individual in news video monologues. In: Proceedings of the 12th ACM International Conference on Multimedia, pp. 580–587. New York, NY, USA (2004)
68. Yang, J., Yan, R., Hauptmann, A.G.: Multiple instance learning for labeling faces in broadcasting news video. In: Proceedings of the 13th ACM International Conference on Multimedia, pp. 31–40. New York, NY, USA (2005)

69. Yu, H., He, F., Pan, Y.: A novel region-based active contour model via local patch similarity measure for image segmentation. Multimedia Tools and Applications **77**(18), 24,097–24,119 (2018)
70. Yu, H., He, F., Pan, Y.: A novel segmentation model for medical images with intensity inhomogeneity based on adaptive perturbation. Multimedia Tools and Applications **78**(9), 11,779–11,798 (2019)
71. Yu, H., He, F., Pan, Y.: A scalable region-based level set method using adaptive bilateral filter for noisy image segmentation. Multimedia Tools and Applications **79**(9), 5743–5765 (2020)
72. Zhang, X., Zhang, L., Wang, X.J., Shum, H.Y.: Finding celebrities in billions of web images. IEEE Transactions on Multimedia **14**(4), 995–1007 (2012)
73. Zhang, Y., Tang, Z., Wu, B., Ji, Q., Lu, H.: A coupled hidden conditional random field model for simultaneous face clustering and naming in videos. IEEE Transactions on Image Processing **25**(12), 5780–5792 (2016)
74. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in neural information processing systems, pp. 321–328 (2004)
75. Zhu, X.J.: Semi-supervised learning literature survey. University of Wisconsin-Madison Department of Computer Sciences **2** (2008)
76. Zoidi, O., Tefas, A., Nikolaidis, N., Pitas, I.: Person identity label propagation in stereo videos. IEEE Transactions on Multimedia **16**(5), 1358–1368 (2014)