PONTIFICAL CATHOLIC UNIVERSITY OF MINAS GERAIS

Graduate Program in Informatics

Gabriel Barbosa da Fonseca

MULTIMODAL PERSON DISCOVERY USING LABEL PROPAGATION OVER SPEAKING FACES GRAPHS

Belo Horizonte

2018

Gabriel Barbosa da Fonseca

MULTIMODAL PERSON DISCOVERY USING LABEL PROPAGATION OVER SPEAKING FACES GRAPHS

Dissertation presented to Graduate Program in Informatics - Data Analysis, Knowledge Discovery and Information Retrieval of Pontifical Catholic University of Minas Gerais, as partial requirement to obtain Master's degree in Informatics.

Advisor: Prof. Dr. Silvio Jamil F. Guimarães

Co-advisor: Dr. Guillaume Gravier

Belo Horizonte

2018

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Fonseca, Gabriel Barbosa da
F676m Multimodal person discovery using label propagation over speaking faces graphs / Gabriel Barbosa da Fonseca. Belo Horizonte, 2018.
52 f.: il.

Orientador: Silvio Jamil Ferzoli Guimarães
Coorientador: Guillaume Gravier
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Gravações de vídeo. 2. Processamento de imagens - Técnicas digitais. 3. Programação (Computadores). 4. Sistemas de recuperação da informação. 5. Indexação automática. 6. Teoria dos grafos. I. Guimarães, Silvio Jamil Ferzoli. II. Gravier, Guillaume. III. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. IV. Título.

CDU: 681.3.093

Ficha catalográfica elaborada por Fabiana Marques de Souza e Silva - CRB 6/2086

Gabriel Barbosa da Fonseca

MULTIMODAL PERSON DISCOVERY USING LABEL PROPAGATION OVER SPEAKING FACES GRAPHS

Dissertation presented to Graduate Program in Informatics - Data Analysis, Knowledge Discovery and Information Retrieval of Pontifical Catholic University of Minas Gerais, as partial requirement to obtain Master's degree in Informatics.

Prof. Dr. Silvio Jamil F. Guimarães

Prof. Dr. Zenilton Kleber Gonçalves do Patrocínio Júnior

Prof. Dr. Rudinei Goularte

Dr. Guillaume Gravier

Belo Horizonte, 06 de julho de 2018.

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. This phenomenon can be observed on the broadcast TV databases, that are already big and are continuously growing. 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 this objective, related works proposed frameworks based on detecting names via OCR and automatic speech transcripts, and associating these names to clusters of detected faces. Others model a graph of faces, and spread names through the graph structure. In this study, 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 methods are proposed, one based on random walks and the other based on a hierarchical approach. In order to analyze the proposed framework, it is evaluated using the MediaEval 2017 MPD database, along with graph clustering baselines and the study of different modality fusions and their impact on the label propagation techniques. The proposed propagation methods over multimodal graphs outperform all literature methods except one, which uses a different approach on the pre-processing step. It is also shown that the use of multiple modalities improves the results, although better modality fusion techniques can be studied make these improvements even more significant.

Keywords: Multimodal analysis. Graph Modeling. Label Propagation strategies.

RESUMO

A indexação de grandes bases de dados é uma tarefa de grande importância, uma vez que afeta diretamente a qualidade das informações que podem ser recuperadas desses conjuntos. Infelizmente, alguns conjuntos de dados estão crescendo tão rápido que a indexação manual torna-se inviável. Esse fenômeno pode ser observado nos bancos de dados de *broadcast* televisivo, que já são extensos e continuam a crescer. Técnicas de indexação automática podem ser aplicadas para superar esse problema e, neste estudo, é proposta uma técnica não supervisionada para descoberta multimodal de pessoas, que consiste em detectar pessoas que aparecem e falam simultaneamente em um vídeo e associar nomes a elas. Para atingir esse objetivo, trabalhos relacionados criaram estratégias baseadas na detecção de nomes através de OCR e transcrições automáticas de fala, e associando esses nomes a *clusters* de faces detectadas. Outros modelam um grafo de faces e espalham nomes através da estrutura do grafo. Neste estudo, os dados são modelados como um grafo de *faces-falantes*, os nomes são extraídos através de OCR e propagados através do grafo com base em relações audiovisuais entre faces falantes. Para propagar rótulos, são propostos dois métodos, um baseado em *random walks* e outro baseado em uma abordagem hierárquica. Para analisar o trabalho proposto, ele é avaliado usando o conjunto de dados MediaEval 2017 MPD, juntamente com trabalhos de referência baseados em agrupamento em grafos, e o estudo de diferentes fusões de modalidades e seus impactos nas técnicas de propagação de etiquetas. Os métodos propostos de propagação de etiquetas sobre grafos multimodais superam todos os métodos da literatura, exceto um, que usa uma abordagem diferente na etapa de pré-processamento. Também é mostrado que o uso de múltiplas modalidades melhora os resultados, embora melhores técnicas de fusão de modalidades possam ser estudadas afim de tornar essas melhorias ainda mais significativas.

Palavras-chave: Análise Multimodal. Modelagem de Grafos. Estratégias de propagação de etiquetas.

LIST OF FIGURES

LIST OF TABLES

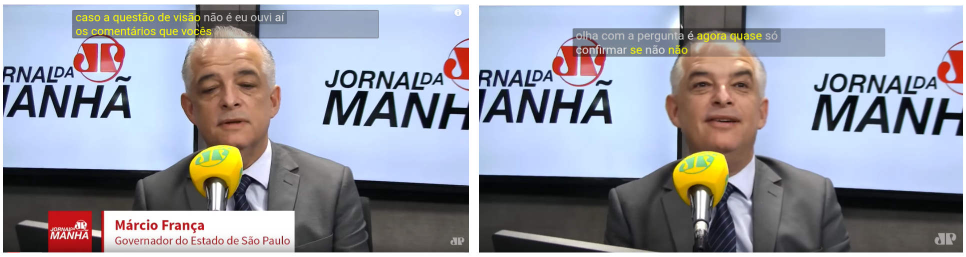
CONTENTS

1 INTRODUCTION

Television has been the main mean of communication for years, and even with the advances of the Internet, it still plays a big role on the communication world. According to a research done by SECOM in 2016, 63% of the Brazilian population use TV as their main source of information (SECOM, 2016). 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.

Figure 1 – Interviewed guest on news video.


caso a questão de visão não é eu ouvi aí
os comentários que vocês
JORNAL DA MANHÃ
Márcio França
Governador do Estado de São Paulo
olha com a pergunta é agora quase só
confirmar se não não

(a) (b)

Two shots showing the same guest with visual description on (a) and no description on (b).

Source: Youtube *

There are many different sources of information within a video as it can be seen in Figure 1, such as visual data, acoustic data, text extracted from visual sources and in the form of audio transcripts, metadata, and temporal relations. When working with multimodal person discovery on videos (MPD), using multiple types of data can ease the process if done well, but if done without care, one can be accumulating noises and propagating errors instead of stacking gains. An example of a troublesome step on the MPD task is the name extraction step, which is usually done via the analysis of automatic speech transcripts (AST) or visually overlaid texts. The first source is very noisy and hard to filter in order to get satisfying results. The second one demands less effort on its analysis, but does not cover a big amount of appearing persons, leading to sparse number of information, also noticeable on Figure 1. Considering this, methods that smartly use multiple modalities can lead to improvements on MPD performances, thus leading to better indexing techniques for large video archives.

1.1 Goals

In this section, the main and specific goals are presented.

1.1.1 Main Goal

It is stated that acquiring labeled data in some fields where unlabeled data is abundant can be quite expansive, and sometimes impossible. For the task of automatic naming persons on video, the name extraction part is rather difficult due to a high amount of noise present on data extracted via AST and by optical character recognition (OCR) (POIGNANT; BREDIN; BARRAS, 2017). Semi-supervised learning methods present good performances when dealing with low amounts of labeled data for training, and a considerable share of the semi-supervised algorithms are graph based.

Considering these assertions, the main goal of this work is to develop methods for automatically naming persons on video using a multimodal and graph-based label propagation solution. To achieve this, a multimodal graph representation and graph-based naming methods are proposed and studied.

1.1.2 Specific Goals

The specific goals of this work are:

- Use a multimodal graph modeling for the PD problem, here named as *speaking-faces*graphs.

- Propose graph-based label propagation methods to work around the sparsity of labeled data.

- Assess the proposed methods, analyzing their performances against traditional graph-based baselines and literature methods.

- Assess the impact of different modality fusion types on the proposed methods.

1.2 Justification

The Internet and TV archives are already big enough to be manually indexed, and they grow faster each day. With a consistent method for automatically indexing these databases, the retrieval of multimodal data becomes possible and more reliable. There are strong indications that graph-based label propagation methods can be successfully used on multimedia data (ZOIDI et al., 2015). So, by studying and developing a graph-based label propagation method that efficiently makes use of multimodalities to solve the MPD problem, it is improved - and sometimes made possible - the automatic indexing of important archives, saving enormous amounts of human resources at the same time.

1.3 Main Contributions

The main contribution of this work is a novel strategy for solving the multimodal person discovery problem, combining the use of *speaking-face*graphs and label propagation techniques. For propagating labels, two methods are employed, one based on the vastly utilized Random Walks algorithm, and the other as a new modified version of the hierarchical label propagation proposed in (PERRET et al., 2015). The proposed methods achieved good performances, outperforming other methods based on speaker/face diarization. Along with the formal presentation of the proposed strategy, a study is done to analyze its possible perks and flaws. For this study, the proposed method is assessed under different modality fusion types and different levels of graph pruning. For deeper comparison with other naive methods applied on the *speaking-face*environment, two graph clustering baselines are also presented in this work.

1.4 Document organization

The remainder of this study is organized in six chapters. Chapter 2 presents the basic concepts related to this work, as well as related works. Chapter 3 introduces the *speaking-faces*graph modeling along with proposed strategies for graph based label propagation. In Chapter 4 it is described the setup of the experimental framework with details on the method design choices, with the experimental results and analysis. Chapter 5 summarizes the main contributions of this study and also present ideas for possible future works.

2 THEORETICAL BACKGROUND

In this Chapter it is presented the basic concepts that support this dissertation. The Section 2.1 is dedicated to the description of semi-supervised label propagation methods, specifically the ones that are graph-based. Section 2.2 introduces the MPD problem, describing its concepts and reviewing some related works, including specific ones that also use label propagation approaches, graph based or not.

2.1 Graph based label propagation

The process of tagging can be understood as applying labels to elements on a dataset, similar to a classification problem. Tagging a dataset can be achieved in various ways, using different techniques to specific problems based on their constraints. One classic way to perform a tagging is executing a clustering technique and then assigning different tags for each group created (YEUNG et al., 1995), also known as a categorization strategy. Another well used technique is the learning of models based on pre-tagged elements for labeling others in the future (LIN; HAUPTMANN, 2002). The first approach relies only on the data information itself to work, not needing a tag-related prior knowledge, defining this approach as a unsupervised one. The second method needs the tag-related information to create the models that are used to infer tags to elements, so it is defined as a supervised method. On unsupervised methods, the tagging is done by creating a distinct tag for each group found. Even though it works well, sometimes elements are given *a priori* with specific tags that are used to tag the remaining elements of the set. For these cases, supervised learning is usually preferred.

Supervised methods are getting an increased attention over the last years, thanks to the advances on machine learning techniques. Many tasks that were once extremely difficult to solve using computers are now trivial with the use of well trained models. One limitation of these methods however is the need of a great number of examples to learn the models well enough, without hitting walls such as overfitting. One way to dodge the issue of not having enough annotated data on training sets is to make use the labeled data altogether with the unlabeled data, and the relations existing between all the elements. This type of approach is known as semi-supervised learning, and it has been shown that for minimally labeled sets, it can perform better than supervised learning methods (ZHOU et al., 2004).

The key to the success of most semi-supervised learning methods is the use of the relations between all elements to improve label inference. Since the relations between elements become so important, many well-known strategies are based on the use of graph modeling (ZHU, 2005).

2.1.1 Basic concepts on graphs

A basic graph is defined as a set of elements and its relations. These elements can be called nodes, and can be connected by edges, which represent relations between these nodes. Formally, a graph is defined as $G = (V, E)$, where V stands for the set of nodes - or vertices - and E stands for the set of edges. A graph can be directed, where the relations also have directions, meaning that an edge $E_{i,j}$ between the nodes i and j can be different than the edge $E_{j,i}$ between the same nodes but in the opposite direction. Also, a graph can be paired with a set of weights W, thus having a ordered pair (G, W) such that for each edge $E_{i,j}$ there is a correspondent weight $W_{i,j}$. When that happens, the graph is called a weighted graph.

Since graph modeling is simple in its essence and can be used to represent relationships between different types of raw data, it becomes a very powerful tool with great generalization aspects. Thanks to this, for many years graph representations are used to model multimedia data for various tasks, and methods can be adapted from one task to another.

2.1.2 Label Propagation

Label propagation algorithms are a special type of semi-supervised learning methods. Semi-supervised methods can be either transductive or inductive classifiers. The inductive classifiers can learn a generic representation of the used learning data, and can be used on initially not known data. The transductive classifiers are local classifiers that make use of all labeled and unlabeled data to perform a classification, and therefore can only be used on the available data. Label propagation algorithms fit in the transductive share of the semi-supervised learning algorithms. As it can be observed on Figure 2, using label propagation algorithms is preferred over common supervised learning methods on minimally annotated datasets.

Label propagation algorithms are methods that try to spread labels through the entire data using the structure of the data along with the initial labeling information. These strategies are similar to the graph diffusion models usually seen in social network analysis, where the opinions of the users (labels) are adopted by different users. This diffusion methods are usually based on biological (OPUSZKO; RUHLAND, 2013) or physical (WANG; KING; LEUNG, 2011) phenomenons.

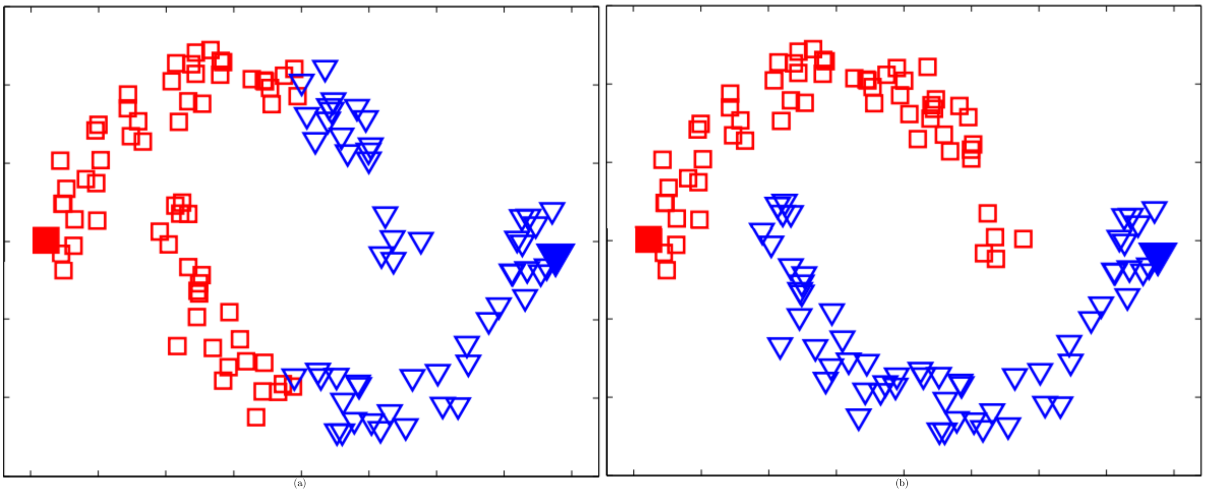
A very common approach to propagate labels through graphs is by using iterative label inference methods. In this type of method, the labels are gradually spread from labeled to unlabeled data, following the data structure and finishing when convergence is reached. One of the first label propagation methods (ZHU; GHAHRAMANI, 2002) used the following algorithm for propagating labels:

In this algorithm, T is calculated by $D^{-1}W$, where W is the affinity matrix of a input graph and $D_{ii} = \sum_j W_{ij}$. In this method, the labels are gradually spread and updated within each iteration. A similar method proposed in (ZHOU et al., 2004), uses a slowing factor to reinforce the initial labeling during the propagation and also uses a normalized laplacian transition matrix. The propagation occurs by iteratively calculating:

$$F^{t+1} = \mu(I - \widetilde{L})F^t + (1-\mu)Y \tag{2.1}$$

where F is a labeling function that applies for each node a value for each possible label. In the end, the label with greatest value is applied for each node. The use of the $\widetilde{L}$ matrix guarantees that the labeling process is applied symmetrically on the matrix.

Figure 2 – SVM and label propagation comparison



Classification of the "two moons"dataset done by SVM on the left and by label propagation on the right. In this case, the label propagation achieves the expected classification.

Source: (ZHOU et al., 2004)

There are also random walk based methods for label propagation. The random walk methods rely on labeling nodes based on the commute time between nodes on a converged probability matrix. The commute time is the expect number of steps needed from one node reaching another by taking only random steps (LOVÁSZ, 1993). The random walk can also be computed with a fixed number of steps, and if done so with a probability matrix calculates as W, the step function is equal to:

$$P^t = P \times P^{t-1} \tag{2.2}$$

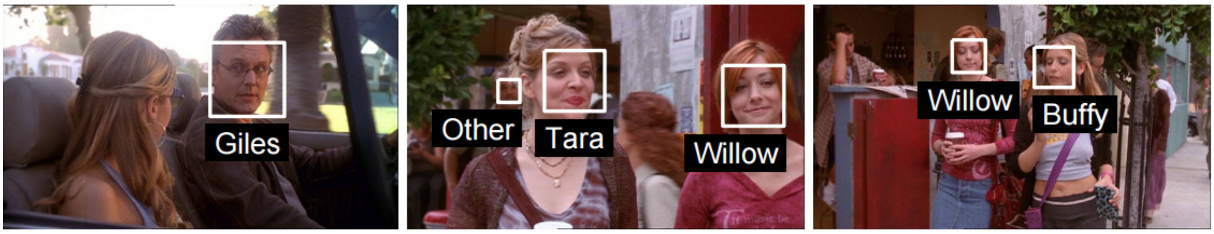
where $PD^{-1}W$. After the iterative calculation of the random walk, a labeling function F can be applied on P^t, and depending on the used function, the results can be equivalent to the ones on the method proposed by (ZHU; GHAHRAMANI, 2002). In this work, a variant of the random walk label propagation is used. The main difference in this variation is that the initially labeled nodes are set as absorbing states, to assure that their labels do not change during the propagation.

2.2 Person discovery on videos

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. In this section, there is a small revision of the related person discovery works.

One of the first proposed approaches for naming persons is the one proposed in (EVERINGHAM; SIVIC; ZISSERMAN, 2006), 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.

Figure 3 – Automatic person naming on video



Result of a person naming strategy on videos.

Source: (EVERINGHAM; SIVIC; ZISSERMAN, 2006)

In (CANSECO; LAMEL; GAUVAIN, 2005; CANSECO-RODRIGUEZ; LAMEL; GAUVAIN, 2004), 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 (TRANTER, 2006; ESTÈVE et al., 2007; MAUCLAIR; MEIGNIER; ESTEVE, 2006). 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 box transcriptions (HOUGHTON, 1999; SATOH; NAKAMURA; KANADE, 1999; YANG; HAUPTMANN, 2004; YANG; YAN;

HAUPTMANN, 2005). In (TUYTELAARS; MOENS et al., 2011), 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.

Figure 4 – Person discovery illustration

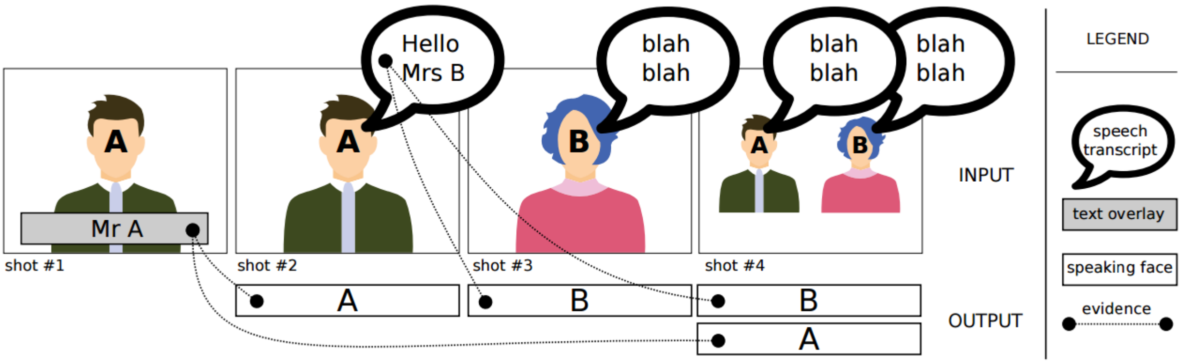


Illustration of the MPD task, as defined in the MediaEval 2016 MPD benchmark. The output boxes represent the "who appears and speaks when"

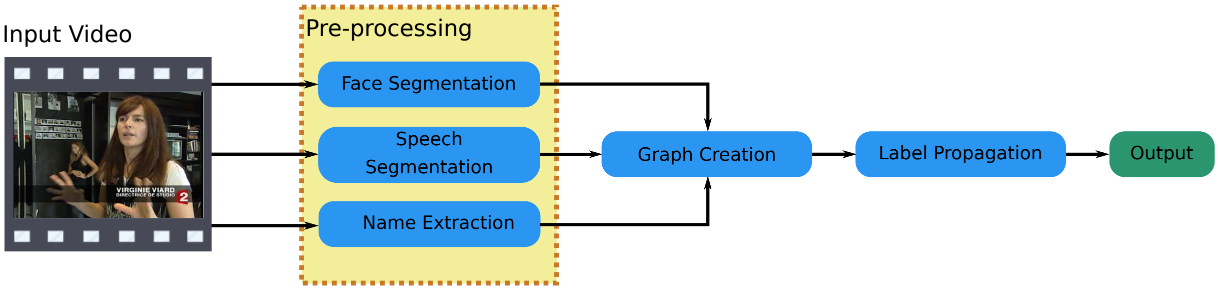
Source: (BREDIN; BARRAS; GUINAUDEAU, 2016)

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 (GALIBERT; KAHN, 2013; KAHN et al., 2012) 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 (BECHET et al., 2014; BENDRIS et al., 2013; BREDIN et al., 2014a, 2014b; GAY et al., 2014; POIGNANT; BESACIER; QUéNOT, 2015; POIGNANT et al., 2016; ROUVIER et al., 2014). MediaEval Person Discovery task (POIGNANT; BREDIN; BARRAS, 2015) 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.

3 LABEL PROPAGATION ON SPEAKING-FACES GRAPHS

In this Chapter, the methodology for tackling the MPD task is presented and formalized. It consists in using label propagation algorithms over graphs of *speaking-faces* to overcome the sparsity of names automatically extracted from videos. A basic pipeline representation for the proposed strategy is illustrated on Figure 5.

Figure 5 – Flow diagram of the proposed method



High level illustration of the steps for the proposed method.

Source: Elaborated by the author.

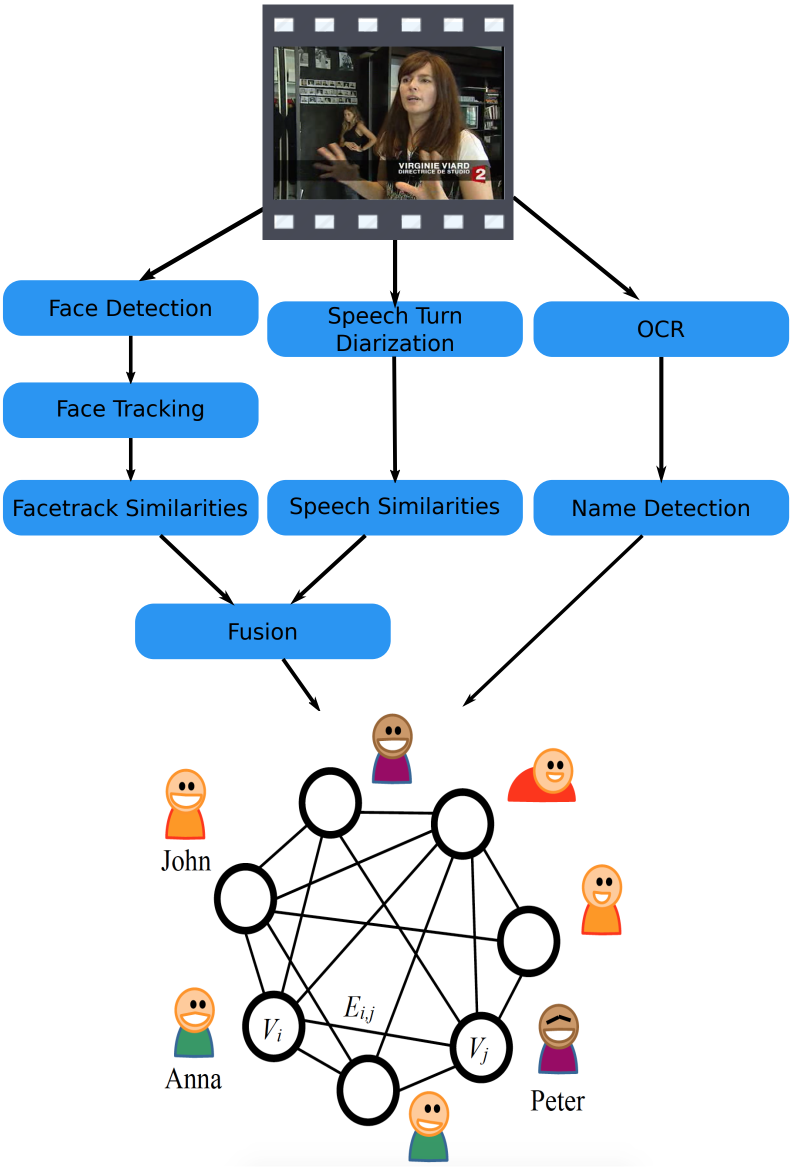
This Chapter is organized as follow: in Section 3.1 the *speaking-faces*graph modeling using in this work is formalized. In Section 3.2 the two proposed label propagation methods that are applied on the *speaking-faces*graphs are presented.

3.1 Speaking Faces graph

Common works tend to extract names via audio transcripts or OCR, 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 (JR.; GRAVIER; SCHWARTZ, 2015), 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 (JR.; GRAVIER; SCHWARTZ, 2015) a multimodal graph representation of speaking persons. In this modeling, a *speaking-face* graph $\mathscr{G} = (V, E)$ 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 illustrated in Figure 6 and described as follows.

Figure 6 – Flow diagram of the *speaking-face* graph cration



Source: Elaborated by the author.

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 $V = \{V_n\}_{1 \le n \le N}$, $N \in \mathbb{N}$. 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, which can be a visual similarity, acoustic similarity or a fusion of both (more details of the similarity calculation are described on Section 4.2.2). A *speaking-faces* graph is illustrated at the end of the flow diagram on Figure 6.

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.*, $\sigma^{AV}_{i,j} = f(\sigma^{V}_{i,j}, \sigma^{A}_{i,j}), 1 \leq i, j \leq N$. In this work, we study the impact of three different fusion approaches for audio and visual modalities. The first is an intermediate fusion, done by the weighted average of two distance values. The second is an early fusion, done by concatenating two feature vectors, thus creating a single audio-visual descriptor for calculating similarities. The last is a late fusion approach, in which the labeling methods are executed separately for each modality and then the fusion occurs on decision level.

3.2 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. Semi-supervised methods stand somewhere between the unsupervised methods and the supervised ones, as they utilize tagged and unttaged data together to work. For some minimally annotated datasets, the use of semi-supervised approaches has been shown better than the use of supervised ones (ZHOU et al., 2004).

In this work two methods are used for propagating tags over *speaking faces*, one as a novel hierarchical approach, and another as an a adaptation of a commonly utilized tag propagation approach. The propagation methods presenter hereafter are:

1 **Minimum spanning tree label propagation:** The first method uses the hierarchical tree created by applying the Kruskal's algorithm for creating a minimum spanning tree (MST) of a graph to propagate labels. The labels are propagated through the process of the MST's creation.

2 **Random walk label propagation:** The second method relies on propagating labels through applying random walks on a graph of probabilities. The labeling process relies on calculating the probabilities of a unlabeled node reaching a labeled node by randomly walking through the probability graph.

In both methods tags are assigned to every *speaking face* detected, leaving none unttaged 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.

3.2.1 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 (PERRET et al., 2015). 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.

Generating a MST using Kruskal's algorithm consists in sorting the edges of a graph and then start clustering its nodes in a agglomerative way, always taking the smallest edges possible to unite sets until there is only one set composed by all nodes and $N-1$ edges, being N the number of nodes in the graph. So we chose the Kruskal's algorithm as base for the first propagation method since the MST connects all elements of a graph with a minimal cost, which in our case represents the highest audio-visual similarities. With that in mind, the tag propagation happens through an optimal path. The steps to perform the MST label propagation (MST_{LP}) are described hereafter.

Algorithm 1: MST_{LP}Algorithm

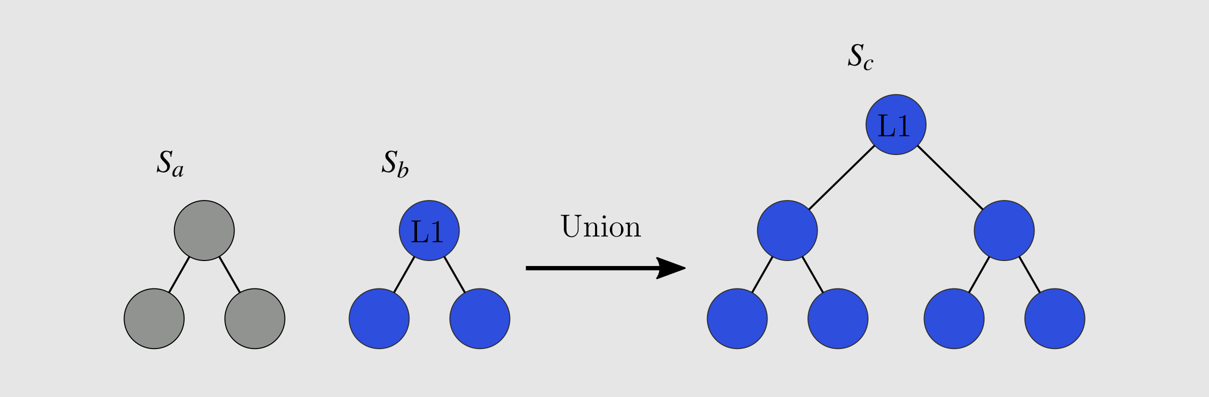
```
1 MST Propagation ($(G,W)$, where $G=(V,E)$);
  Input  : Partially labeled graph $G$
  Output: Labeled graph $G'$
2 Sort $E$
3 foreach VERTEX $V_i \in G$ do
4     MAKE-SET($V_i$)
5 foreach EDGE $E_{i,j}$ TAKEN IN NONDECREASING ORDER do
6     if FIND-SET($V_i$) $\neq$ FIND-SET($V_j$) then
7         UNION($S_i$,$S_j$)
8         PROPAGATE($S_i$,$S_j$)
```

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 7; (ii) if none of the sets is labeled, nodes of both sets remain unlabeled, illustrated in Figure 8; 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 9. 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}.

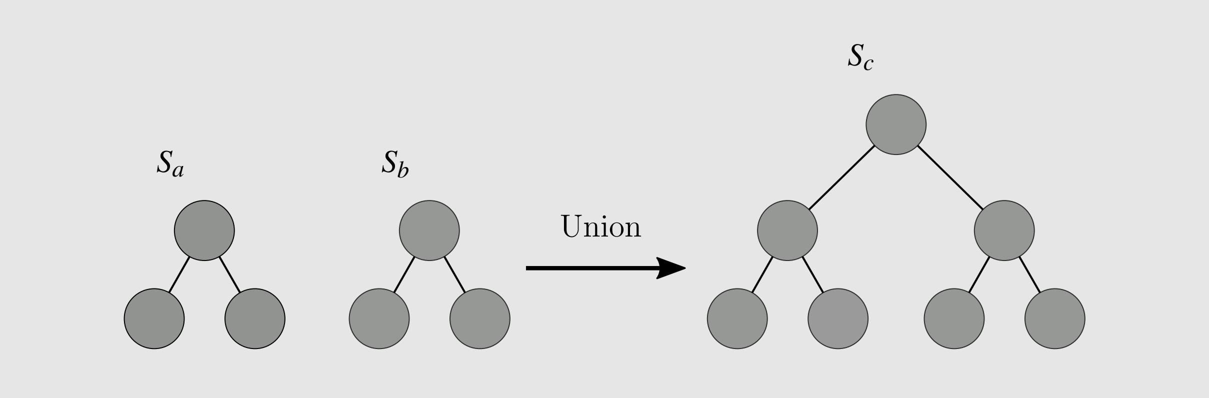
Figure 7 – MST merging - case (i)



First case of merging, when only one of the sets is tagged.

Source: Elaborated by the author

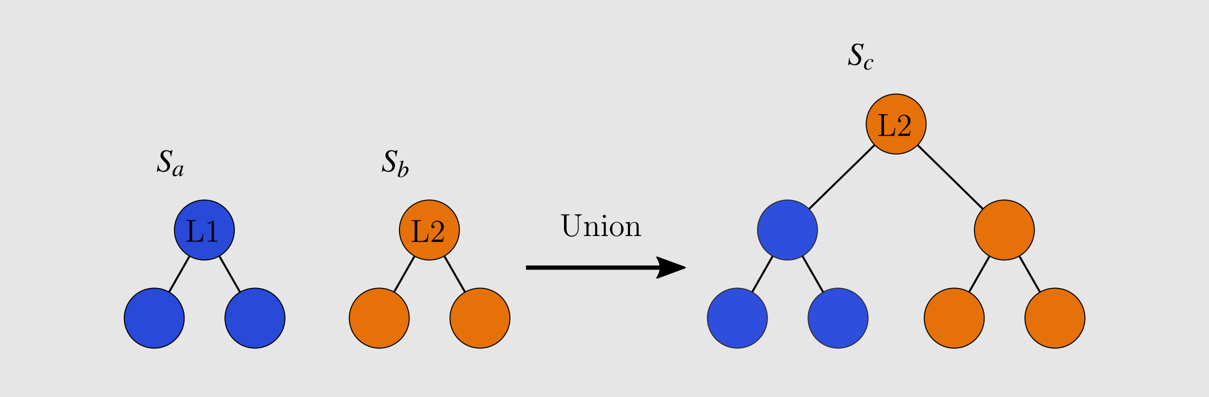
Figure 8 – MST merging - case (ii)



Second case of merging, when none of the sets is tagged.

Source: Elaborated by the author

Figure 9 – MST merging - case (iii)



Third case of merging, when both of the sets is tagged. The tags do not change and one is chosen to represent the set, in this case, the label L2.

Source: Elaborated by the author

3.2.2 Random walk based propagation

The concept of random walks has been vastly used in various fields due to its interesting theoretic aspects and practical power. Many methods that opt for a stochastic inference strategy uses random walk based modeling for their problems. As cited on (MASUDA; PORTER; LAMBIOTTE, 2017), 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 a unlabeled node to reach each labeled one. For the second propagation approach, random walks with absorbing states are used to perform the label propagation, adapting from (ZHU; GHAHRAMANI; LAFFERTY, 2003).

Given a graph, a walk with an unitary step is defined by moving from one determined node to one of its neighbors. If a walk $Wlk =< a,b,c,d...k >$ is composed by non repeated elements, *i.e.* with no cycles, it is defined as a path between the starting and the ending node. If a node is taken as a starting point, and one of its neighbors is selected at random to be walked into, that is called a Random Walk. The probability of a node walking to another can be distributed evenly based on the degree - number of connections - of this node, or based on a probability graph. On a probability graph $P = (V,E)$, the $W_{i,j}$ set represents the probability of i walking to j in one step. On $P^t = (V,E,)$, $W'_{i,j}$ represents the probability of i randomly reaching j in t steps. The concept of Random Walks on a probability graph is very similar to a finite Markov Chain (LOVÁSZ, 1993).

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

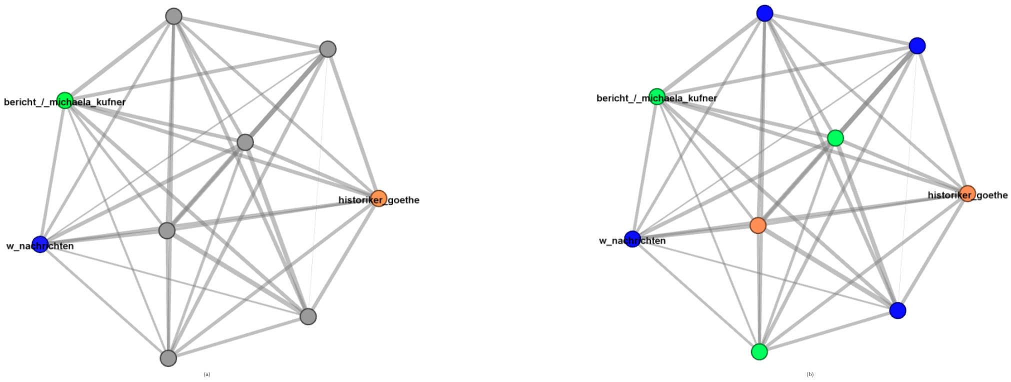
The random walk with t steps is calculated by $P^t = P \times P^{t-1}$, and the number of steps should be enough for P^t reaching convergence. To achieve a random walk based labeling (RW_{LP}) that is consistent with the initial label information, a slowing factor can be applied to the walk, and in this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

Algorithm 2: RW_{LP}Algorithm

```
RW Propagation (G);
Input  : Partially labeled graph $G$ and weight matrix $W$
Output: Probability Matrix $P^t$
Calculate $D_{ii} = \sum_j W_{ij}$
Calculate $P = D^{-1}W$
foreach NODE $V_i$ do
    if $V_i$ IS LABELED then
        Set $V_i$ as absorbing state
for $t$ IN [0..MAX_STEPS] do
    $P^t = (\omega \times P \times P^{t-1}) + ((1-\omega) \times P)$
```

Figure 10 – Example of RW propagation



Graph example before propagation on the left and after RW_{LP}on the right.

Source:Research data

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 $O(n^3)$.

4 EXPERIMENTAL FRAMEWORK

In this chapter, the details on the framework setup are presented, along with the datasets used to validate this work as well as the evaluation setup and metrics used for assessment. The main objective of the experimental set is to analyze the aspects of the proposed method on solving the MPD problem, with a focus on the proposed label propagation methods applied over the speaking-face graphs.

The dataset used for the assessment is described on Section 4.1. The framework setup describing the sub-processes involved on the task is presented in Section 4.2. Two graph clustering methods are proposed to serve as baselines to the label propagation methods, and its details are described in Section 4.3. The metrics used on the evaluation of this work are detailed in Section 4.4. At last, the experiments and evaluations are presented in Section 4.5.

4.1 Dataset

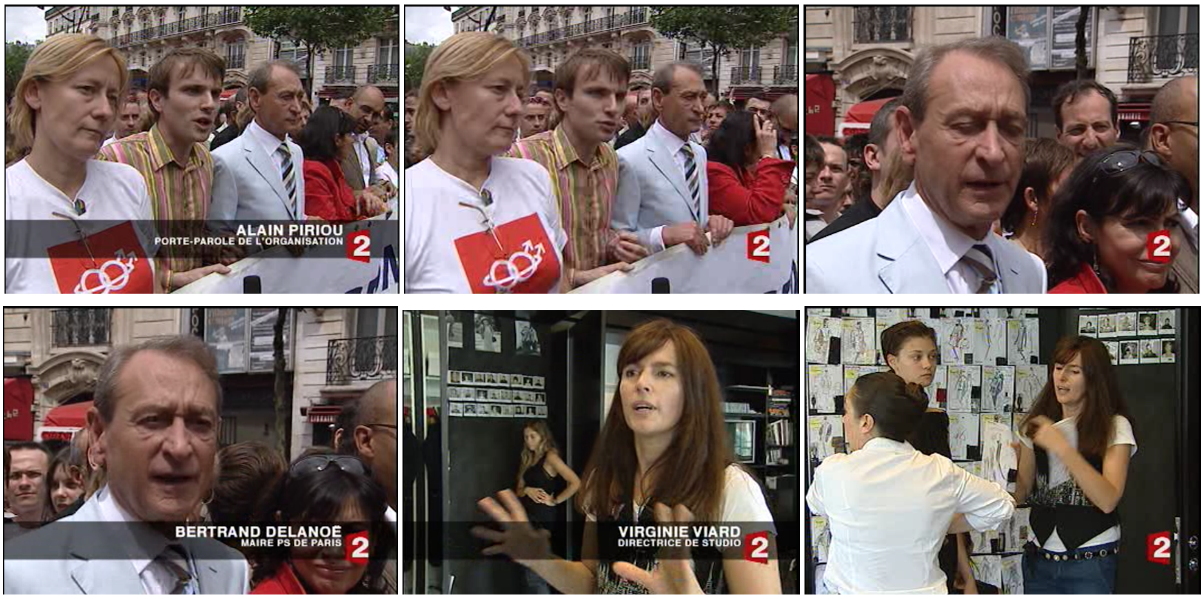
To evaluate the proposed methods we use the test set of the MediaEval 2016 MPD task, which was manually annotated during the campaign of the respective year (BREDIN; BARRAS; GUINAUDEAU, 2016). This set is divided in three parts, named as 3-24, INA and DW. The 3-24 is composed by a Catalan TV news channel, 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. The INA dataset is contains a total of 90 hours of duration, the DW has a total duration of 50 hours, and the 3/24 has a duration of 13 hours of TV broadcast. The dataset was free of annotation before the Mediaeval 2016 event, and it was annotated based on the participants submissions, more details about the annotation process can be found in (BREDIN; BARRAS; GUINAUDEAU, 2016). The final annotation was assembled at 16th of October 2016, and it is the one used in this work as ground truth. The ground truth contains 3431 annotated shots, which can have one or more names assigned to it.

Along with the raw data, the Mediaeval organization also provided a baseline, containing pre-processed data related to all MPD's steps. This baseline is given so if someone wants to change only a step and not the whole method, it is possible to do that without having to process all the other steps unrelated to the tweaked part. The provided baseline includes:

- Segmentation of the video stream as a sequence of S contiguous shots, two shots being delimited by a brief or smooth change of camera take.

- Detection of the face tracks within the video stream, a face track being a sequence of portions of frames which are contiguous in time and relate to a single face. A face track is assumed to be completely contained within a single shot.
- 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.
- Speech transcription that can be also used for name detection.

Figure 11 – Example of dataset videos



Source:Data extracted from Mediaeval MPD 2016 dataset.

4.2 Framework Setup

In this Section, each sub-process of the framework is detailed. An expanded version of the pipeline illustrated on Figure 5 is shown on Figure 12, detailing the logical order of the processes to be described. This Section is organized as follows: first, the pre-processing involved and the extracted features are detailed in Section 4.2.1. The process for calculating the audio-visual features, including the applied modality fusion types are explained in Section 4.2.2. The pruning formula and parameters are described in Section 4.2.3. Finally, the procedure to choose the number of steps for the RW_{LP} is shown on Section 4.2.4.

Figure 12 – MPD flow diagram

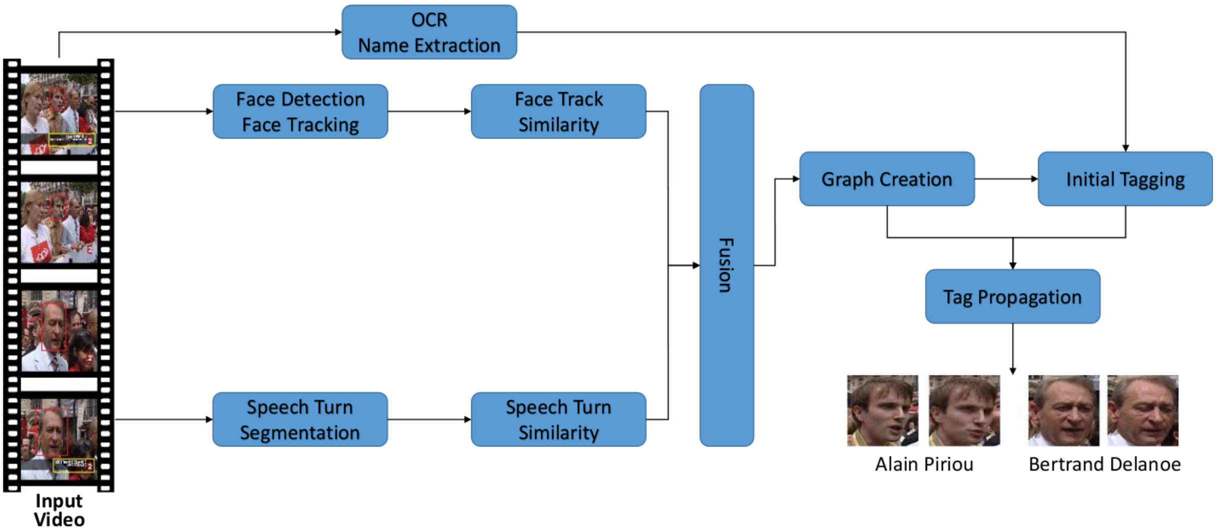


Diagram illustrating all the steps of the proposed MPD framework.

Source: Elaborated by the author

4.2.1 Pre-Processing and Feature Extraction

As mentioned in Section 4.1, the Mediaeval 2016 dataset provides some pre-processed steps to help the participants, allowing them to focus on specific parts of the problem. The pipeline of the MPD method applied in this work is illustrated in Figure 12. In some of the showed steps, it was used the provided pre-computed features, and in others, features are computed to best fill the project needs.

The provided features used are the shot segmentation - shots whose duration is less than 1 s or more than 10 s are discarded -, the text detection and recognition by IDIAP (CHEN; ODOBEZ, 2005), the segments of speech obtained with the speaker diarization system from LIUM (ROUVIER et al., 2013), the facetracks obtained with a histogram of oriented gradients-based detector (DALAL; TRIGGS, 2005) and a correlation tracker (DANELLJAN et al., 2014). 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 (RAYMOND, 2013).

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.

Previous work shows how to extract generic visual descriptors from pre-trained Convolutional Neural Networks. Oquab *et al.* (OQUAB et al., 2014) extract intermediate layers to build mid-level generic visual representations for classification. Razavian *et al.* (RAZAVIAN et al., 2016) similarly build descriptions for image retrieval. More recently, Tolias *et al.* (TOLIAS; SICRE; JÉGOU, 2016) uses convolutional layers of a pretrained CNN to efficiently build the MAC and R-MAC descriptors for retrieval while Sicre *et al.* (SICRE et al., 2016) uses both fully connected and convolutional layers output 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:

- CNN: The very deep vd-19 (SIMONYAN; ZISSERMAN, 2015) CNN trained on the ImageNet dataset.
- FaceNet: The face specific descriptor FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015).

For the CNN feature, similarly to Tolias et. al. (TOLIAS; SICRE; JÉGOU, 2016), the last convolutional layer of the network is extracted, then a 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 face using cosine similarity. The resulting similarity $\boldsymbol{\sigma}^V$ takes values between 0 and 1 as these visual features were normalized. For the FaceNet descriptors, similarities are also calculated by calculating the cosine similarities between features.

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[*] and Audioseg[†] 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 (GARCIA-ROMERO; ESPY-WILSON, 2011).

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

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

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

(BEN et al., 2004). It is turned into a similarity using the function $\sigma^A_{i,j} = \exp(\alpha\delta^A_{i,j})$, where $\sigma^A_{i,j}$ and $\delta^A_{i,j}$ are respectively the similarity and the distance between segments i and j. In the case of i-vector descriptors, the computation of the cosine similarity between them incorporates a channel compensation processing which emphasizes again the similarity between channels (DEHAK et al., 2011). In the end, all the similarities are values between 0 and 1, with 1 meaning most similar possible.

Two pairs of audio-visual features are created in this work, one containing a generic video descriptor along with a GMM based audio descriptor, and the other using a face specific descriptor along with a state-of-the-art audio descriptor. Apart from two configurations, the other two possible feature combinations are also produced, leading to four different graph configurations. These configurations are referred as:

- CNN-GMM
- FaceNet-iVector
- CNN-iVector
- FaceNet-GMM

4.2.2 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 $\gamma = 0$): $\sigma^{AV}_1 = \sigma^V$ and $\sigma^{AV}_2 = \sigma^A$. Then, the tag with the highest confidence score is kept for each *speaking face*.

In the framework there are two parameters regarding the similarity creation on the *speaking-faces*. These parameters are the α and γ, which relates to the weighted average when applying the intermediate fusion and the distance to similarity transformation respectively. To better assess the proposed methods, a tuning of these parameters is proposed. The tuning is based on a cross-validation scheme, in which the dataset is divided into small groups and different parameters are tested in each of these subsets. This type of strategy usually reduces the bias when choosing parameter values.

Table 1 – Alpha and gamma values for each configurations

Configurations	α	γ
CNN-GMM	0.3	0.5
CNN-iVector	n/a	0.3
FaceNet-GMM	0.3	0.7
FaceNet-iVector	n/a	0.3

Values set for α and γ parameters on each graph configuration. Since the only feature that uses distances is GMM, only the configurations containing it have a α value.

Source:Research data

In order to perform the cross-validation, first it is created an intersection of the whole dataset and the ground truth, so only the annotated videos are used. After, this first subset is divided randomly into 10 different folds, containing approximately the same number of videos in each one. Then, for each fold, its is executed a label propagation with all combinations of α and γ in the range $[0, 1]$ with a step of 0.1. To select the label propagation used in this phase, all propagation methods were tested with the α and γ set as 0.5, and the best scoring one —in this case the RW_{LP}—was selected. Each run is evaluated by its recall. The α and γ combination with the biggest mean recall is selected for each graph configuration.

On the creation of the graphs, the *speaking-faces* creation was the same for all configurations. For calculating the similarities, besides from applying the three different fusion types, they were also pruned in three different levels. The α parameter for the distance-to-similarity transformation when using GMM audio features was tuned along with the γ parameter for doing the weighted intermediate fusion of modalities. It can be observed in Table 1 that the γ values are never 0 or 1, showing that the use of audio-visual modalities is better than using only audio or only visual relations.

For each of the four configurations, a combination of α and γ values were set using the protocol described in above. The selected values are exhibited on table 1.

4.2.3 Graph Pruning

Originally, we calculate the similarities between all nodes in a graph, ending with a complete graph. However we also want to study the impact of pruning in our propagation methods. For pruning the graphs we use an adaptive method that consists in setting a threshold,

$$Threshold = m - \delta * std \tag{4.1}$$

where m is the average and std is the standard deviation of the similarities of a given graph, with delta being a negative real number. The values for the δ pruning parameter were manually set as 0, -1 and -2. These values were selected in a way that leads to a very soft pruning, a moderate pruning, and a drastic one.

4.2.4 Random Walk steps

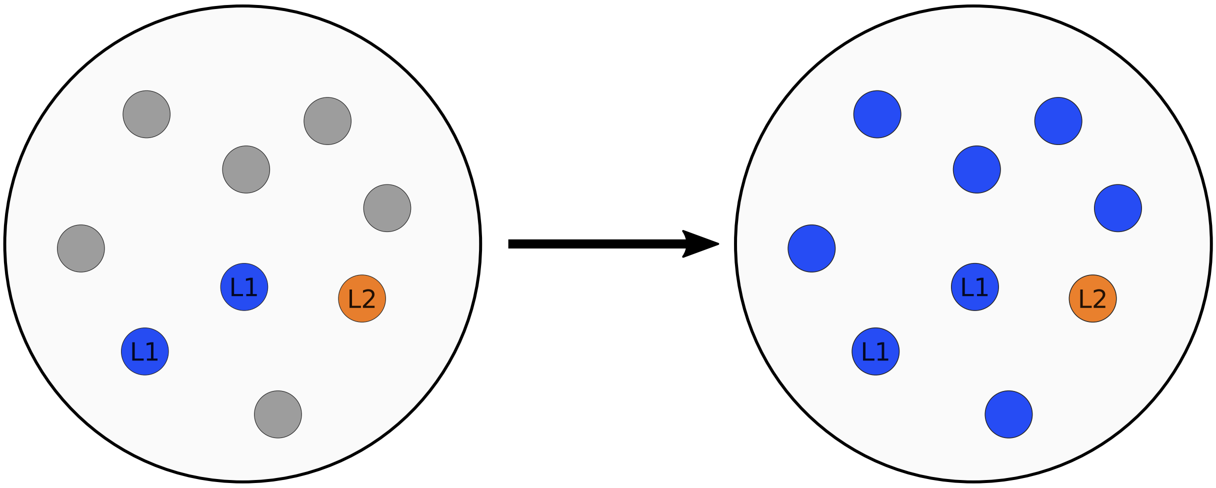
On the label propagation algorithms, only one parameter needed to be set, which is the t number of steps on the RW_{LP}. The parameter value was set heuristically, checking the number of necessary steps to achieve convergence on the development dataset of the Mediaeval MPD 2016 campaign. The 2016 development contains 106 hours of video, corresponding to 172 editions of evening broadcast news "Le 20 heures" of the French public channel "France 2" (BREDIN; BARRAS; GUINAUDEAU, 2016). The minimum tested value that achieved convergence on all used graphs was 50 steps.

4.3 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 scenario, two graph clustering baselines are proposed, one using spectral clustering and the other using Markov clustering (ENRIGHT; DONGEN; OUZOUNIS, 2002).

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 $\mathscr{G}$. 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.

Figure 13 – Baseline labeling illustration



Label spreading inside a cluster. Note that the initial labels do not change after the labeling process.

Source: Elaborated by the author

4.4 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[‡] (BREDIN; BARRAS; GUINAUDEAU, 2016) 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.

To calculate the MAP@K, let $\{q_j\}_{1\leq j\leq J}$ be a list of J "firstname_lastname" reference names. Each name q_j is assigned to the set of reference shots S_j^r where the related person appears. Then, for each q_j is returned the set of shots S_j^a which were automatically associated to a tag approximating or equaling q_j in terms of edit distance. S_j^a is ranked using decreasing confidence scores and the classical average precision value $P_{\text{av}}^K(q_j)$ is calculated on the K first elements of this ranking. Finally, the MAP@K is computed as:

$$\text{MAP@}K = \frac{\sum_{j=1}^{J} P_{\text{av}}^K(q_j)}{J} \tag{4.2}$$

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

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

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:

$$Err = \frac{N^D + N^I}{n^r}, \tag{4.3}$$

$$R = \frac{N^C}{n^r} \tag{4.4}$$

The Kappa coefficient is a metric that measures the level of agreement between two sets of results, as if they were decisions made by different judges. To calculate the coefficient, two results A and B must be given as inputs. For each query, the answers of A and B are matched, and set as C_AC_B if both methods are correct, C_AF_B if only A is correct, F_AC_B if only B is correct, and F_AF_B if both are wrong. Then, it is possible to calculate p_o, which is the relative observed agreement among the two judges, and p_e, which is the hypothetical probability of chance agreement. They are computed as:

$$p_o = (C_AC_B + F_AF_B)/(C_AC_B + C_AF_B + F_AC_B + F_AF_B) \tag{4.5}$$

$$p_c = (C_AC_B + C_AF_B + C_AC_B + F_AC_B)/(2*(C_AC_B + C_AF_B + F_AC_B + F_AF_B)) \tag{4.6}$$

$$p_f = (F_AF_B + C_AF_B + F_AF_B + F_AC_B)/(2*(C_AC_B + C_AF_B + F_AC_B + F_AF_B)) \tag{4.7}$$

$$p_e = {p_c}^2 + {p_f}^2 \tag{4.8}$$

Finally, the Kappa coefficient can be computed by:

$$Kappa = p_o - p_e/1 - p_e \tag{4.9}$$

4.5 Experiments

The experiments done to assess the proposed approaches on solving the MPD task are described in this section, along with the experimental setup. In this section we study the impact of our proposed tag-propagation approaches with respect to cases where no propagation is performed or where graph-clustering techniques are used to spread the initial tags. Also, it is studied the impact of different modality fusions and graph pruning on the proposed methods. The strategies defined in this work are also compared to literature methods applied on the same dataset.

The remainder of this section is as follows. First, it is described and discussed the quantitative results obtained through different experiments Section 4.5.1, including the consequences of using different fusion strategies and the impact of different levels of pruning on the label propagation. On Section 4.5.3, a qualitative analysis of the label propagation methods is presented.

4.5.1 Quantitative assessment

The quantitative analysis of the framework is presented and discussed in this section. The main objective of this analysis is to validate the characteristics of the proposed label propagation methods when applied to the MPD problem solving. The proposed methods are assessed with regard to the proposed baseline methods, to observe if label propagation techniques really outperform the naive clustering baselines.

Two pairs of different features -two audio features and two visual features- are used to describe *speaking-faces*, and three different fusion strategies are used to merge these modalities, namely early fusion, intermediate fusion and late fusion. Every propagation method is evaluated under all configurations, and hence they are referred in this work as:

- NoProp: Only the initial tagging, with no propagation applied.
- MST: Hierarchical Label propagation with intermediate fusion.
- MST_LF: Hierarchical Label propagation with late fusion.
- MST_EF: Hierarchical Label propagation with early fusion.
- RW: Random Walk Propagation with intermediate fusion.
- RW_LF: Random Walk Propagation with late fusion.
- RW_EF: Random Walk Propagation with early fusion.
- Markov: Markov Clustering with intermediate fusion.
- Markov_LF: Markov Clustering with late fusion.
- Markov_EF: Markov Clustering with early fusion.
- Spectral: Spectral Clustering with intermediate fusion.
- Spectral_LF: Spectral Clustering with late fusion.
- Spectral_EF: Spectral Clustering with early fusion.
- AltRW: Alternate Random Walk.

In the first batch of experiments displayed on Tables 2, 3, 5, and 4 we can observe the error rates, recall and MAP@K results of the methods proposed in this work. The two best scoring methods for each metric are highlighted in bold. If there is a tie, all methods scoring best and second best values are highlighted.

Observing Table 2 one can observe that all labeling methods improve the results when compared to the initial taggin only (NoProp). This suggests that by only using

OCR extracted names it is not possible to correctly name all appearing persons on a video, and labeling techniques can help to solve this issue.

Table 2 – CNN-GMM Results

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	0.60	0.41	0.618	0.471	0.448	0.433
Spectral	0.59	0.44	0.604	0.447	0.426	0.412
MST	**0.49**	0.52	0.658	**0.546**	**0.523**	**0.506**
RW	0.51	**0.54**	**0.671**	**0.553**	**0.531**	**0.512**
MST_LF	0.53	0.53	0.659	0.543	0.520	0.502
RW_LF	**0.49**	**0.54**	**0.663**	0.539	0.517	0.500
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
AltRW	0.68	0.36	0.628	0.476	0.452	0.436

Source:Research data

Table 3 – FaceNet-iVector Results

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	0.62	0.42	0.604	0.443	0.426	0.413
Spectral	0.68	0.42	0.594	0.417	0.398	0.386
MST	**0.57**	0.51	**0.669**	**0.550**	**0.528**	**0.510**
RW	0.59	**0.53**	**0.659**	**0.535**	**0.508**	**0.490**
MST_LF	**0.58**	**0.52**	0.653	0.515	0.493	0.476
RW_LF	0.59	**0.52**	0.649	0.520	0.494	0.477
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
AltRW	0.71	0.39	0.611	0.455	0.431	0.416

Source:Research data

Table 4 – FaceNet-GMM Results

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	0.65	0.36	0.623	0.467	0.445	0.430
Spectral	0.57	0.46	0.606	0.447	0.428	0.416
MST	**0.48**	**0.53**	0.644	**0.536**	**0.515**	**0.498**
RW	**0.50**	**0.55**	**0.666**	**0.550**	**0.528**	**0.508**
MST_LF	0.56	**0.53**	**0.659**	0.532	0.508	0.492
RW_LF	0.57	0.52	0.653	0.526	0.500	0.484
Markov_LF	0.61	0.46	0.623	0.473	0.450	0.434
Spectral_LF	0.57	0.47	0.618	0.460	0.437	0.423
AltRW	0.62	0.43	0.626	0.484	0.461	0.444

Source:Research data

One can also observe that the proposed label propagation methods $\boldsymbol{RW_{LP}}$ and $\boldsymbol{MST_{LP}}$ achieved the best scores on all metrics, with intermediate fusion on some and

Table 5 – CNN-iVector Results

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	0.60	0.42	0.598	0.447	0.430	0.414
Spectral	0.66	0.40	0.596	0.430	0.409	0.395
MST	**0.51**	**0.52**	**0.661**	**0.541**	**0.515**	**0.497**
RW	0.57	0.51	**0.681**	**0.557**	**0.532**	**0.511**
MST_LF	0.54	**0.53**	0.649	0.530	0.508	0.491
RW_LF	**0.52**	**0.52**	0.659	0.527	0.504	0.488
Markov_LF	0.64	0.41	0.616	0.470	0.446	0.430
Spectral_LF	0.64	0.40	0.599	0.434	0.413	0.399
AltRW	0.73	0.33	0.613	0.455	0.431	0.418

Source:Research data

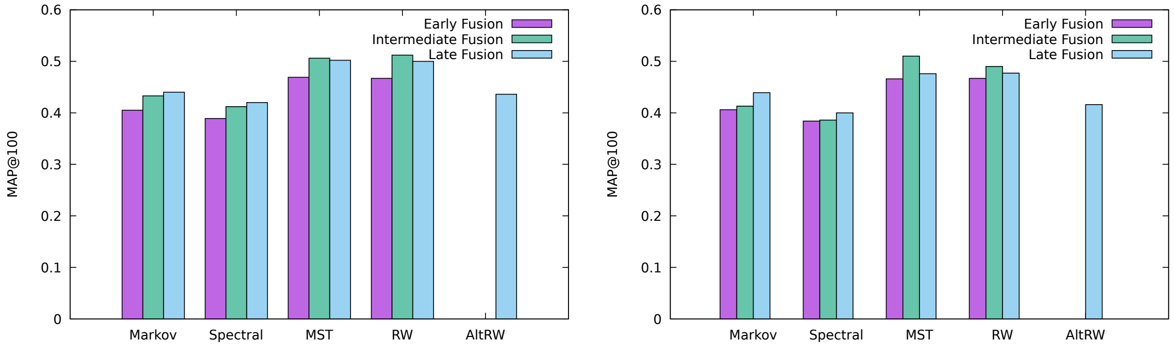
late fusion on others. The alternate $\boldsymbol{RW_{LP}}$ achieved worst results than the traditional $\boldsymbol{RW_{LP}}$, by a considerable margin. The label propagation methods also perform better than the naive clustering approaches, ranging from 0.500 to 0.512 against 0.412 to 0.440 on MAP@100. This shows that using semi-supervised learning algorithms leads to better results than only using clustering based labeling processes.

On Table 3 the strategies were tested using the combination of a face specific image descriptor and a state-of-the art audio descriptor, opposed to the prior CNN-GMM configuration, which uses good but generic descriptors. The results on 3, show that improving the quality of the features does not necessarily improve the results obtained when using the proposed framework. In some cases, like the MST, the scores are improved by using the FaceNet-iVector configuration, but the opposite happens for the RW propagation. On Tables 5 and 4, there is the combination of the CNN-iVector descriptors and FaceNet-GMM descriptors. Like in the other configurations, the observed behaviours remain constant.

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 14.

Observing the two bar charts, one can observe 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 $\boldsymbol{RW_{LP}}$ fusion types, showing itself as a bad performing late fusion approach. On the proposed label propagation methods, *i.e.*$\boldsymbol{MST_{LP}}$ and $\boldsymbol{RW_{LP}}$, 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 different. On them, the best performing fusion type is the late fusion, followed by the intermediate

Figure 14 – Fusion Strategies.



(a) GMM CNN features

(b) FaceNet Ivector features

Source:Research data

fusion and early fusion at last. 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 these results is possible to assume that simply applying a generic similarity function over the concatenation of aural 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 in a naive way. Using a better suited multimodal feature fusion and more appropriate similarity metric for the new multimodal features could lead to improvements on the early fusion performance.

Table 6 – Number of edges per each pruning intensity

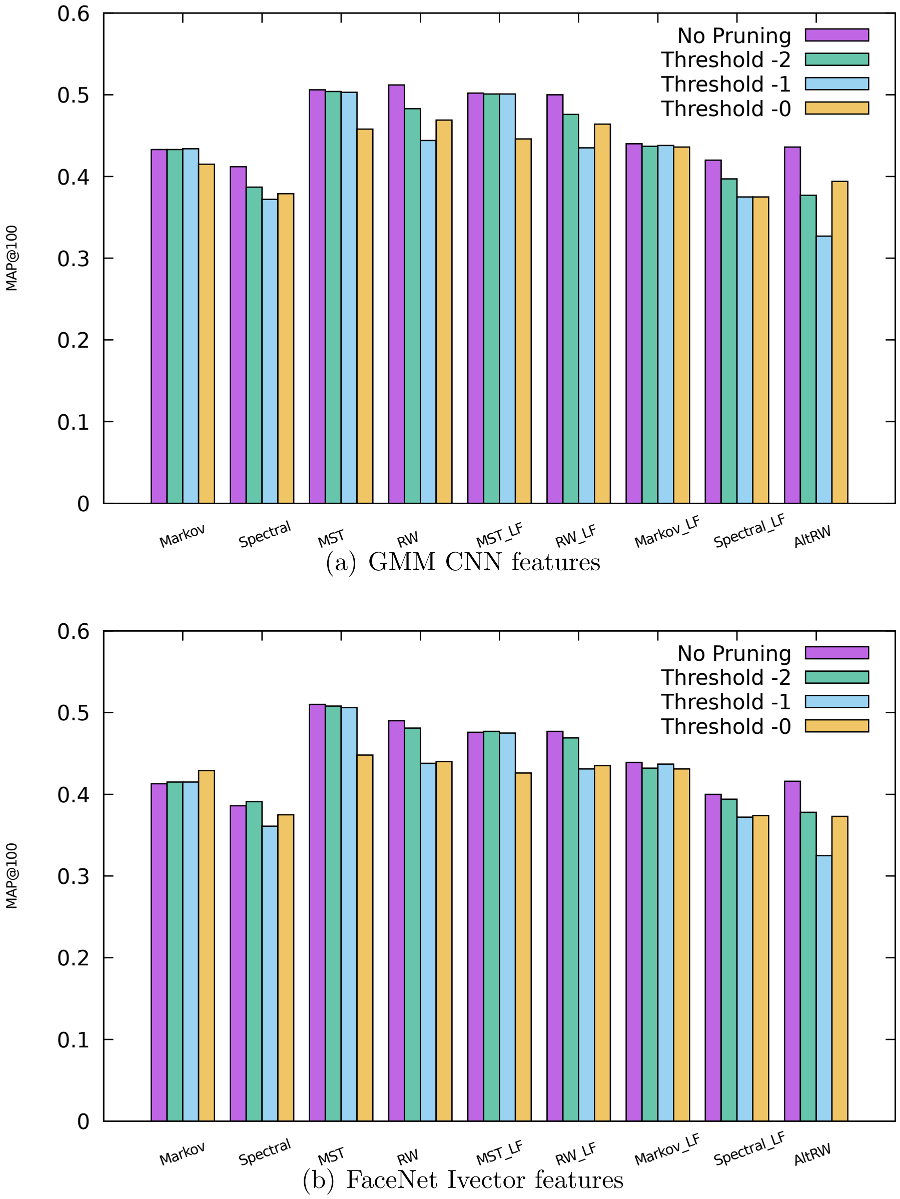
	Total number of edges			
Configurations	No Pruning	δ = -2	δ = -1	δ = 0
CNN-GMM	56.578.108	54.343.228	42.349.534	13.794.879
FaceNet-iVector	56.578.108	56.012.823	43.338.496	11.567.599

Source:Research data

To understand the impact of graph pruning on the used labeling methods, three levels of pruning were applied to the graphs. The graph configurations used to analyze the impact were the CNN-GMM and FaceNet-iVector configurations. The total number of edges on each graph configuration before and after the pruning are exhibited on Table 6. On the FaceNet-iVector configuration, the smoothest pruning leaves 99% of the total number of edges, and the hardest one leaves only 20%.

The results for all propagation methods using intermediate and late fusion under the different pruning levels are shown in Figure 15. Once again, the behaviors of each method on the two different graph configurations are very similar. On most cases, pruning the graphs results on diminish the scores, with a few exceptions. On the Markov propagation pruning improved the results. For the FaceNet-iVector configuration, pruning also improved results for the Spectral and MST_LF labeling methods. For the rest, the

Figure 15 – Impact of pruning edges.



(a) GMM CNN features

(b) FaceNet Ivector features

Source:Research data

act of pruning is always prejudicial, but not always proportionally to the level of pruning applied.

These results are heterogeneous along the methods at a certain level, but showing on most cases that pruning is prejudicial to the propagation methods when 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 perform well and improve the efficacy with regard to not using label propagation. This is a very interesting characteristic, meaning that the proposed methods still perform well on scenarios where there is a low percentage of edges, which can diminish drastically the computational cost without suffering much loss on the resulting labeling.

4.5.2 Comparison with the state-of-the-art

In this section, 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.

Table 7 – MAP@K comparison against literature methods.

Method	MAP@1	MAP@5	MAP@10	MAP@100
(LE; MEIGNIER; ODOBEZ, 2016)	**0.791**	**0.672**	**0.650**	**0.629**
(OTERO; DOCIO-FERNANDEZ; MATEO, 2016)	0.249	0.199	0.188	0.166
(NGUYEN et al., 2016)	0.100	0.091	0.089	0.086
(NISHI et al., 2016)	0.254	0.173	0.157	0.147
(MARTÍ et al., 2016)	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

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.

Source:Research data

In Table 7 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 (LE; MEIGNIER; ODOBEZ, 2016), 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 (MARTÍ et al., 2016), and already top the Tokyo Tech, HCMUS (NGUYEN et al., 2016) and GTM-UVIGO (OTERO; DOCIO-FERNANDEZ; MATEO, 2016) 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.

4.5.3 Qualitative analysis of results

In order to enrich the assessment of the proposed methods, in this section two main analysis are presented. The first analysis displays some more detailed information on the dataset and its annotations, and how they can impact the evaluations. In the second analysis, the ***MST***$_{LP}$ and ***RW***$_{LP}$ methods are put under a paired comparison, exploiting some of their perks and flaws.

As mentioned in Section 4.1, the used dataset is not fully annotated. The ground truth was built by a collaborative effort, and in its final form, it contained 3431 annotated shots. On the graph creation step, the number of detected *speaking-faces* is 179.905. After the initial automatic naming, 11.267 *speaking-faces* are labeled, which represents 6.7% of the total number of *speaking-faces*. This fact confirms the expectations extracting names from visual overlays would end in a sparsely labeled set. After temporally gathering the *speaking-faces* into their respective shots, the total number of shots to be evaluated is 94193. The ground truth contains 3431 annotated shots, therefore only a portion of approximately 3.5% of the annotated shots can be evaluated. This is not an optimal scenario, and having more annotated data would help to improve the evaluation of the proposed methods.

From the 3431 annotated shots in the ground truth, 811 were named after the initial tagging phase, *i.e.* NoProp configuration. From the 811 initially tagged nodes, 207 labels are wrong, resulted by errors in the pre-processing and graph creation parts. This leaves a portion of 2620 shots to evaluate the effects of the proposed label propagation methods.

Table 8 – Relation of correct and wrong propagation

	Propagated Labels	
Configurations	Correct	Wrong
MST$_{LP}$	1207	677
RW$_{LP}$	1261	623

Source:Research data

It can be observed on Table 8 that both methods have similar results. On a total of 2620 unlabeled shots after the *speaking-face* graph creation, 677 remain unlabeled after propagation. This happens because the *speaking-face* related to this shots were not found during the pre-processing, hence these entities cannot be labeled. From the 1884 propagated labels, both methods correctly propagate around 65% of it, with ***RW***$_{LP}$ excelling ***MST***$_{LP}$ by 54 hits. In this work, the label propagation methods leave no unlabeled nodes in the graphs, and by selectively leaving unlabeled nodes, the methods would be able continue propagating correct labels without propagating noise and false labels through the graphs.

The ***MST***$_{LP}$ and ***RW***$_{LP}$ strategies score 0.86 on the Kappa's coefficient, which according to (LANDIS; KOCH, 1977) can be considered as an almost perfect agreement. If we analyze both algorithms, the ***MST***$_{LP}$ has a smaller time complexity, which makes it more scaleable. The processing times of both algorithms are measured for propagating labels for the entire dataset[§]. The processing time of the $\boldsymbol{RW_{LP}}$ is of 7m12s, and for the $\boldsymbol{MST_{LP}}$ it is 1m8s, representing a speedup of **6.35** times and corroborating with the difference of complexity between both algorithms.

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

5 CONCLUSION

In this chapter, the main contributions and conclusions are presented, along with the possible future extensions for this work. Also, it is presented the published papers regarding the label propagation over *speaking-face* graphs for multimodal person discovery strategies.

5.1 Contributions

This study tackled the multimodal person discovery problem. This task has gained a lot of attention over the years, but many of the related studies use mono-modal clustering based strategies to solve the problem. It is presented in this work a multimodal modeling for the problem, along with two methods for propagating labels over the proposed *speaking-face* graph model.

In the following, it is addressed the main scientific contributions of this work:

- The formal definition of a multimodal graph representation, named ***speaking-face* graph**. It is shown that the use of multiple modalities is superior than using one only modality for calculating similarities between the detected speaking persons. It is also shown that the use of different acoustic and visual features, from generic to case-specific features, do not change the behavior of methods applied on the *speaking-face* graphs significantly, but that behavior could be affected by the use of a similarity measure that is not perfectly suited for the used representations.

- Two semi-supervised, **graph based label propagation algorithms** to expand the initially named entities on the *speaking-face* graphs. One of the proposed methods is an novel hierarchical label propagation strategy, using confidence scores decisions to leave no unlabeled nodes. The second is an adaptation of existing methods, based on random walks. Both methods improve the labeling by propagating names through the *speaking-face* graphs. The MST_{LP} and RW_{LP} methods produce highly equivalent results, according to the Kappa coefficient. The MST_{LP} method however is approximately 6.35 times faster than the RW_{LP} method. Both methods outperformed all diarization and clustering based literature methods applied to the same dataset.

- **Two graph-clustering baselines** to study the impact of the label propagation algorithms against more naive methods applied on *speaking-faces* graphs. The proposed label propagation methods achieve better results than the baselines on all graph configurations. This suggests that semi-supervised learning methods are best suited for this specific environment.

- The study of **how graph manipulations impact the proposed label propagation algorithms**. Three fusion types were studied to create multimodal similarities between *speaking-faces*. The intermediate fusion performed better when using the proposed label propagation methods, but when using the clustering baselines, the best results are achieved by using the late fusion strategy. Early fusion is the worst choice for both labeling types. Also, it was studied the effects of graph pruning on the labeling methods. On the greater part of the experiments, graph pruning shows itself bad for the label propagation, but the achieved recall is not proportional to the intensity of the pruning on all cases. Nevertheless, even if the pruning is not beneficial, it does not invalidate the proposed methods, showing that they can still work well in cases where there is a low percentage of edges, giving good results with diminished computational cost for creating the graphs and propagating the labels.

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 increase the number of correctly labeled faces, with regard to the initially labeled *speaking-face* graphs. Additionally, the label propagation methods outperform the graph-clustering baselines, showing that semi-supervised methods are the best choices for the presented scenario.

5.2 Future work

The proposed work opens possibilities for novel studies, such as:

- Using metric learning to create better similarity values on initially labeled graphs.
- Applying the *speaking-face* graph modeling for tackling other multimodal tasks;
- Using the hierarchical label propagation as an alternative semi-supervised learning method on different areas;
- Study the extraction of audio-visual features to create better *speaking-faces* similarities;
- Study better suited modality fusion techniques;
- Limiting the hierarchical propagation by applying cuts on the tree, as a way of avoiding wrong propagation; and

- Using mathematical morphological operations to improve the hierarchical propagation.

5.3 Published papers

This study resulted in the following published papers:

- **PUC Minas and IRISA at Multimodal Person Discovery** (SARGENT et al., 2016). In: Working Notes Proceedings of the MediaEval Workshop. 2016.
- **Towards large scale multimedia indexing: A case study on person discovery in broadcast news.** (LE et al., 2017). 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.** (FONSECA et al., 2017) Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI) 2017.
- **Tag Propagation Approaches within Speaking Face Graphs for Multimodal Person Discovery.** Journal in preparation for the IEEE Transactions on Multimedia.

REFERENCES

BECHET, F. et al. Multimodal understanding for person recognition in video broadcasts. In: INTERSPEECH 2014 – ICSLP. [S.l.: s.n.], 2014. p. 607–611.

BEN, M. et al. 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. [S.l.: s.n.], 2004. p. 333–444.

BENDRIS, M. et al. Unsupervised face identification in tv content using audio-visual sources. In: 2013 11TH INTERNATIONAL WORKSHOP ON CONTENT-BASED MULTIMEDIA INDEXING (CBMI). [S.l.: s.n.], 2013. p. 243–249.

BREDIN, H.; BARRAS, C.; GUINAUDEAU, C. Multimodal person discovery in broadcast TV at MediaEval 2016. In: WORKING NOTES OF THE MEDIAEVAL 2016 WORKSHOP. [S.l.: s.n.], 2016.

BREDIN, H. et al. Person Instance Graphs for Named Speaker Identification in TV Broadcast. In: ODYSSEY 2014, THE SPEAKER AND LANGUAGE RECOGNITION WORKSHOP. Joensuu, Finland: [s.n.], 2014.

BREDIN, H. et al. 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, Springer-Verlag, 2014.

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, 2005. [S.l.: s.n.], 2005. p. 415–419.

CANSECO-RODRIGUEZ, L.; LAMEL, L.; GAUVAIN, J.-L. Speaker diarization from speech transcripts. In: ICSLP. INTERSPEECH. [S.l.], 2004.

CHEN, D.; ODOBEZ, J.-M. Video text recognition using sequential Monte Carlo and error voting methods. Pattern Recognition Letters, v. 26, n. 9, p. 1386–1403, July 2005.

DALAL, N.; TRIGGS, B. Histograms of Oriented Gradients for Human Detection. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. [S.l.: s.n.], 2005. v. 1, p. 886–893.

DANELLJAN, M. et al. Accurate Scale Estimation for Robust Visual Tracking. In: PROCEEDINGS OF THE BRITISH MACHINE VISION CONFERENCE. [S.l.]: BMVA Press, 2014.

DEHAK, N. et al. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, IEEE, v. 19, n. 4, p. 788–798, 2011.

ENRIGHT, A. J.; DONGEN, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. Nucleic acids research, Oxford University Press, v. 30, n. 7, p. 1575–1584, 2002.

ESTÈVE, Y. et al. Extracting true speaker identities from transcriptions. In: INTERSPEECH 2007 – ICSLP. [S.l.: s.n.], 2007. p. 2601–2604.

EVERINGHAM, M.; SIVIC, J.; ZISSERMAN, A. Hello! my name is... buffy–automatic naming of characters in tv video. 2006.

FONSECA, G. B. D. et al. Tag propagation approaches within speaking face graphs for multimodal person discovery. In: ACM. Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing. [S.l.], 2017. p. 15.

GALIBERT, O.; KAHN, J. The first official repere evaluation. In: First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013). [S.l.: s.n.], 2013.

GARCIA-ROMERO, D.; ESPY-WILSON, C. Y. Analysis of i-vector length normalization in speaker recognition systems. In: Twelfth Annual Conference of the International Speech Communication Association. [S.l.: s.n.], 2011.

GAY, P. et al. Comparison of two methods for unsupervised person identification in tv shows. In: 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI). [S.l.: s.n.], 2014. p. 1–6.

HOUGHTON, R. Named faces: putting names to faces. IEEE Intelligent Systems and their Applications, v. 14, n. 5, p. 45–50, Sep 1999.

JR., C. E. dos S.; GRAVIER, G.; SCHWARTZ, W. R. SSIG and IRISA at Multimodal Person Discovery. In: Working Notes Proceedings of the MediaEval Workshop. Wurzen, Germany: [s.n.], 2015. Disponível em: <https://hal.archives-ouvertes.fr/hal-01196171>.

KAHN, J. et al. A presentation of the repere challenge. In: 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI). [S.l.: s.n.], 2012. p. 1–6.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. biometrics, JSTOR, p. 159–174, 1977.

LE, N. et al. Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In: ACM. Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing. [S.l.], 2017. p. 18.

LE, N.; MEIGNIER, S.; ODOBEZ, J.-M. Eumssi team at the mediaeval person discovery challenge 2016. In: MediaEval Benchmarking Initiative for Multimedia Evaluation. [S.l.: s.n.], 2016.

LIN, W.-H.; HAUPTMANN, A. News video classification using svm-based multimodal classifiers and combination strategies. In: ACM. Proceedings of the tenth ACM international conference on Multimedia. [S.l.], 2002. p. 323–326.

LOVÁSZ, L. Random walks on graphs. Combinatorics, Paul erdos is eighty, v. 2, n. 1-46, p. 4, 1993.

MARTÍ, G. et al. Upc system for the 2016 mediaeval multimodal person discovery in broadcast tv task. In: MEDIAEVAL. [S.l.: s.n.], 2016.

MASUDA, N.; PORTER, M. A.; LAMBIOTTE, R. Random walks and diffusion on networks. Physics Reports, v. 716-717, p. 1 – 58, 2017. ISSN 0370-1573. Random walks and diffusion on networks. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0370157317302946>.

MAUCLAIR, J.; MEIGNIER, S.; ESTEVE, Y. Speaker diarization: About whom the speaker is talking ? In: 2006 IEEE ODYSSEY - THE SPEAKER AND LANGUAGE RECOGNITION WORKSHOP. [S.l.: s.n.], 2006. p. 1–6.

NGUYEN, V.-T. et al. Hcmus team at the multimodal person discovery in broadcast tv task of mediaeval 2016. In: MEDIAEVAL. [S.l.: s.n.], 2016.

NISHI, F. et al. Tokyo tech at mediaeval 2016 multimodal person discovery in broadcast tv task. In: MEDIAEVAL. [S.l.: s.n.], 2016.

OPUSZKO, M.; RUHLAND, J. Impact of the network structure on the sir model spreading phenomena in online networks. In: PROCEEDINGS OF THE 8TH INTERNATIONAL MULTI-CONFERENCE ON COMPUTING IN THE GLOBAL INFORMATION TECHNOLOGY (ICCGI'13). [S.l.: s.n.], 2013.

OQUAB, M. et al. Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR. [S.l.: s.n.], 2014.

OTERO, P. L.; DOCIO-FERNANDEZ, L.; MATEO, C. G. Gtm-uvigo system for multimodal person discovery in broadcast tv task at mediaeval 2016. In: MEDIAEVAL. [S.l.: s.n.], 2016.

PERRET, B. et al. Evaluation of morphological hierarchies for supervised segmentation. In: SPRINGER. PROCEEDINGS OF THE 12TH INTERNATIONAL SYMPOSIUM ON MATHEMATICAL MORPHOLOGY AND ITS APPLICATIONS TO SIGNAL AND IMAGE PROCESSING. [S.l.], 2015. p. 39–50.

POIGNANT, J.; BESACIER, L.; QUéNOT, G. Unsupervised speaker identification in tv broadcast based on written names. IEEE/ACM Transactions on Audio, Speech, and Language Processing, v. 23, n. 1, p. 57–68, Jan 2015.

POIGNANT, J.; BREDIN, H.; BARRAS, C. Multimodal person discovery in broadcast TV at mediaeval 2015. In: WORKING NOTES PROCEEDINGS OF THE MEDIAEVAL 2015 WORKSHOP. [S.l.: s.n.], 2015.

POIGNANT, J.; BREDIN, H.; BARRAS, C. Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015. Multimedia Tools and Applications, Springer, v. 76, n. 21, p. 22547–22567, 2017.

POIGNANT, J. et al. Naming multi-modal clusters to identify persons in TV broadcast. Multimedia Tools Appl., v. 75, n. 15, p. 8999–9023, 2016.

RAYMOND, C. Robust tree-structured named entities recognition from speech. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. [S.l.: s.n.], 2013.

RAZAVIAN, A. S. et al. Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications, The Institute of Image Information and Television Engineers, v. 4, n. 3, p. 251–258, 2016.

ROUVIER, M. et al. An open-source state of the art toolbox for broadcast news diarization. In: INTERSPEECH. [S.l.: s.n.], 2013. p. 25–29.

ROUVIER, M. et al. Scene understanding for identifying persons in tv shows: Beyond face authentication. In: 2014 12TH INTERNATIONAL WORKSHOP ON CONTENT-BASED MULTIMEDIA INDEXING (CBMI). [S.l.: s.n.], 2014. p. 1–6.

SARGENT, G. et al. Pucminas and IRISA at multimodal person discovery. In: WORKING NOTES PROCEEDINGS OF THE MEDIAEVAL 2016 WORKSHOP. [S.l.: s.n.], 2016.

SATOH, S.; NAKAMURA, Y.; KANADE, T. Name-it: naming and detecting faces in news videos. IEEE MultiMedia, v. 6, n. 1, p. 22–35, Jan 1999.

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. [S.l.: s.n.], 2015. p. 815–823.

SECOM. PESQUISA BRASILEIRA DE MÍDIA 2016 HÁBITOS DE CONSUMO DE MÍDIA PELA POPULAÇÃO BRASILEIRA. 2016. [Online; accessed 28-March-2018]. Disponível em: <http://pesquisademidia.gov.br/?utm_term=Informe+Semanal+-+Edicao+no+287+-+06.01.2017&utm_campaign=LISTA+GLOBAL&utm_source=e-goi&utm_medium=email&eg_sub=626a9a8fe4&eg_cam=e2dc0b091f6057705ff9b4c43a45c57c&eg_list=13#/Geral/details-917>.

SICRE, R. et al. Automatic discovery of discriminative parts as a quadratic assignment problem. CCV Workshops–CEFRL, 2016. Disponível em: <http://arxiv.org/abs/1611.04413>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. ICLR, 2015.

TOLIAS, G.; SICRE, R.; JÉGOU, H. Particular object retrieval with integral max-pooling of cnn activations. ICLR, 2016.

TRANTER, S. E. Who really spoke when? finding speaker turns and identities in broadcast news audio. In: 2006 IEEE ICASSP. [S.l.: s.n.], 2006. v. 1, p. I–I.

TUYTELAARS, T.; MOENS, M.-F. et al. Naming people in news videos with label propagation. IEEE multimedia, IEEE Computer Society, v. 18, n. 3, p. 44–55, 2011.

WANG, D.; KING, I.; LEUNG, K. S. “like attracts like!”– a social recommendation framework through label propagation. In: . [S.l.: s.n.], 2011.

YANG, J.; HAUPTMANN, A. G. Naming every individual in news video monologues. In: PROCEEDINGS OF THE 12TH ANNUAL ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA. New York, NY, USA: [s.n.], 2004. p. 580–587.

YANG, J.; YAN, R.; HAUPTMANN, A. G. 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: [s.n.], 2005. p. 31–40.

YEUNG, M. M. et al. Video browsing using clustering and scene transitions on compressed sequences. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. IS&T/SPIE'S SYMPOSIUM ON ELECTRONIC IMAGING: SCIENCE & TECHNOLOGY. [S.l.], 1995. p. 399–413.

ZHOU, D. et al. Learning with local and global consistency. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. [S.l.: s.n.], 2004. p. 321–328.

ZHU, X. Semi-supervised learning literature survey. world, Computer Sciences, University of Wisconsin-Madison, v. 10, p. 10, 2005.

ZHU, X.; GHAHRAMANI, Z. LEARNING FROM LABELED AND UNLABELED DATA WITH LABEL PROPAGATION. [S.l.], 2002.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In: PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML-03). [S.l.: s.n.], 2003. p. 912–919.

ZOIDI, O. et al. Graph-based label propagation in digital media: A review. ACM Computing Surveys (CSUR), ACM, v. 47, n. 3, p. 48, 2015.