Tag Propagation Approaches within Speaking Face Graphs for Multimodal Person Discovery

Gabriel Barbosa da Fonseca
Izabela Lyon Freire
Zenilton Patrocínio Jr
Silvio Jamil F. Guimarães
gbrl12@gmail.com
izabela.lyon.freire@gmail.com
zenilton@pucminas.br
sjamil@pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil

Gabriel Sargent
Ronan Sicre
Guillaume Gravier
gabriel.sargent@irisa.fr
ronan.sicre@irisa.fr
guillaume.gravier@irisa.fr
IRISA & Inria Rennes (CNRS, Univ. Rennes 1)
Rennes, France

ABSTRACT

The indexing of broadcast TV archives is a current problem in multimedia research. As the size of these databases grows continuously, meaningful features are needed to describe and connect their elements efficiently, such as the identification of speaking faces. In this context, this paper focuses on two approaches for unsupervised person discovery. Initial tagging of speaking faces is provided by an OCR-based method, and these tags propagate through a graph model based on audiovisual relations between speaking faces. Two propagation methods are proposed, one based on random walks and the other based on a hierarchical approach. To better evaluate their performances, these methods were compared with two graph clustering baselines. We also study the impact of different modality fusions on the graph-based tag propagation scenario. From a quantitative analysis, we observed that the graph propagation techniques always outperform the baselines. Among all compared strategies, the methods based on hierarchical propagation with late fusion and random walk with score-fusion obtained the highest MAP values. Finally, even though these two methods produce highly equivalent results according to Kappa coefficient, the random walk method performs better according to a paired t-test, and the computing time for the hierarchical propagation is more than 4 times lower than the one for the random walk propagation.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; *Speech recognition*; Tracking; *Visual content-based indexing and retrieval*; • **Information systems** → *Multimedia and multimodal retrieval*;

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

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

KEYWORDS

Multimedia Indexing, Multimodal fusion, Tag Propagation, Face Recognition, Speaker Recognition

ACM Reference format:
Gabriel Barbosa da Fonseca, Izabela Lyon Freire, Zenilton Patrocínio Jr, Silvio Jamil F. Guimarães, Gabriel Sargent, Ronan Sicre, and Guillaume Gravier. 2017. Tag Propagation Approaches within Speaking Face Graphs for Multimodal Person Discovery. In *Proceedings of Proc. of International Workshop on Content-Based Multimedia Retrieval, Firenze, Italy, June 19-21, 2017 (CBMI),* 5 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

TV archives are rapidly growing in size. The need for applications that make these archives searchable has led researchers to devote considerable effort to the development of technologies that create various indexes. One of the most demanded indexes is for people identification. At the moment when content is created or broadcasted, it is not always known which individuals are going to be important to find in the future, and it is impracticable to fully annotate a TV archive by hand. Also, some of the individuals appearing in the videos may not even be known at all by the archivists responsible for content annotation.

Multimodal Person Discovery (MPD) addresses the problem of indexing people in the archive, under realistic conditions. For example, a predefined list of people to be indexed may not be available. The usual approach to this problem is to segment a collection of TV broadcasts into *shots*, which are accompanied by the output of several low-level analysis components such as: speaker diarization, face detection and tracking, speech transcription, video optical character recognition (OCR), and named entity detection. Additionally, one could make use of attached textual metadata available alongside the corpus, such as subtitles, electronic program guide, textual description, etc. The final result of MPD is, for each shot, a list of names of persons whose *speaking face* appears within it. The main challenge behind this problem is that the list of persons is not provided *a priori*, and that person "models" (of their faces and voices) should not be generated using any external data. Therefore, the only way to identify persons is by extracting their names from audio and visual streams (*e.g.*, using speech transcription or OCR)

and associating them to the correct person, thus making the task completely unsupervised.

In [3], Canseco et al. proposed approaches to person identification which were based on pronounced names. The use of biometric models for speaker identification appears in [5, 14]. However, these audio-only approaches did not achieve good performance because of high errors rates, caused by poor speech transcriptions and bad named-entity detections. Similarly, video-only approaches were very dependent on the quality of overlaid title box transcriptions [16]. From 2011 to 2014, the REPERE challenge [8] encouraged research on multimodal identification of persons that looked for mechanisms to surpass the limitations of unimodal approaches. Much progress was achieved in both supervised and unsupervised multimodal person recognition [2, 7, 11]. MediaEval Person Discovery task [1] can be seen as a follow-up campaign with a strong focus on unsupervised person recognition.

The goal of Multimodal Person Discovery (MPD) consists in naming all the people which are simultaneously visible and speaking within a video document. It is a completely unsupervised task since no prior knowledge is used. As in our previous work [13], the proposed approach starts with an initial tagging followed by a graph-based tag propagation. In order to do that, we adopted for each video a segmentation of the visual stream into a sequence of contiguous shots (two shots being delimited by a abrupt or gradual change of camera take). After that, face tracks (*i.e.*, a sequence of frame regions that are contiguous in time and relate to a single face) are detected within each video shot. The audio stream is also split into speech segments. And, similarity values are calculated between all these high-level features and used to solve MPD. Potentially, speech transcriptions could be used for finding names but we have not considered that in this work. The main differences between this work and our previous paper [13] are threefold: (i) formalization of the method; (ii) comparison to two distinct baselines using graph clustering approaches (Markov and spectral clustering); and (iii) proposition of a late fusion approach for tag propagation in which visual and audio similarities are used separately and then final results are merged.

This paper is organized as follows. In Section 2, we describe some aspects of *speaking face* graph. In Section 3, we present the proposed method for tagging the *speaking faces*. Some experiments and analysis are presented in Section 4, and finally, some considerations the future directions are drawn in Section 5.

2 SPEAKING FACE GRAPH

Consider the sets of shots, face tracks, speech segments and names (tags) detected within a video. Let them be represented as $s = \{s_l\}_{1 \le l \le L}$, $F = \{F_i\}_{1 \le i \le I}$, $S = \{S_j\}_{1 \le j \le J}$ and $T = \{T_k\}_{1 \le k \le K}$, respectively, where $L, I, J, K \in \mathbb{N}$.

Let a *speaking face* V_n be the association of a face track F_i and a co-occurring speech segment S_j, assumed to belong to the same person. In particular, V_n exists if and only if the intersection of temporal spans of F_i and S_j is non-empty. Let the set of *speaking faces* be $V = \{V_n\}_{1 \le n \le N}$, $N \in \mathbb{N}$, so that a video can be modelled as a complete graph $\mathcal{G} = (V, E)$ – named *speaking face* graph, in which nodes are *speaking faces* and every pair of nodes V_n and V_m is connected by an edge $E_{n,m} = (V_n, V_m)$, with $E = \{E_{n,m}\}_{1 \le n,m \le N}$, whose weight W_{ij} represents the similarity between corresponding *speaking faces* V_n and V_m (see Fig. 1). Every tag $T_k, 1 \le k \le K$, is related to the time interval in which it appears in a specific overlay. In order to solve MPD, each node $V_n, 1 \le n \le N$ should be associated to a tag $T_k, 1 \le k \le K$, thus naming the person related to its *speaking face* track.

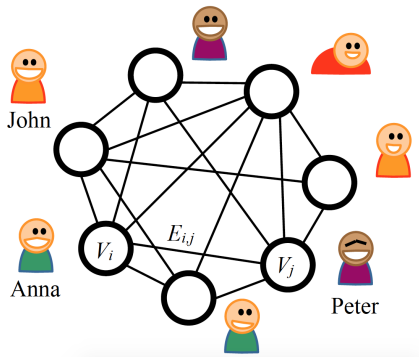

Figure 1: *Speaking face* graph representation (edges associated to a weight of zero are discarded for a more clearer visualization). Each node represents a speaking face track, and can be associated to a name.

As will be seen in Section 3, the proposed method for tagging of all the nodes consists in two main steps: (i) an initial node tagging; and (ii) a tag propagation of the initial tags to other nodes, on a graph where edges represent similarities between *speaking faces*.

In the following, some aspects of the *speaking face* graph construction are described. Feature extraction and computation of similarities are detailed.

2.1 Feature extraction

Multiple content-based features can be extracted from a multimodal dataset. Several baseline descriptors were provided to the participants in the context of the MediaEval Person Discovery task [1], in order to allow them solve the problem more easily. In this work, we have used some features provided along with the dataset, such as shot detection, text detection and recognition, speaker diarization, etc. Besides that, we have also generated other new features (or post-processed the old ones) to complement content-based data about the *speaking face*. Specific details are described, together with computational experiments, in Section 4.

2.2 Visual and audio similarities

The similarity between two *speaking faces* V_n and V_m can be based on its visual and audio content. 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, audio-visual similarity σ^{AV} between *speaking faces* can be interpreted as a function of visual and audio similarities, *i.e.*, $\sigma^{AV}_{i,j} = f(\sigma^V_{i,j}, \sigma^A_{i,j}), 1 \le i, j \le N$. An average between the visual and audio similarity values can be used as an example of this combination function.

3 METHODS FOR TAGGING SPEAKING FACES

The proposed method is composed of a two-phase tagging approach, which consists of initial node tagging followed by graph-based tag propagation. Each tagged node will have a confidence score related to its assigned tag taking values between 0 and 1.

In the initial tagging phase, we assume that the names appearing in overlays belong to persons that are both visible and speaking at the same time as the overlay is visible. We assume that the initial tags have a confidence score of 1, and this must not change during the tag propagation phase.

3.1 Tag propagation approaches

In this work, two different approaches for propagating the initial tags have been adopted, a random walk approach and a hierarchical strategy. In both of our methods, tags are assigned to every node after the propagation phase, leaving no untagged nodes at the end.

3.1.1 Random walk approach (RW). In this first method, we use random walks on graphs to perform the tag propagation, adapting from [17]. In order to perform the random walk on a graph, we first calculate a probability matrix $P = D^{-1}W$, where W is the graph weight matrix and D is a diagonal matrix defined as $D_{ii} = \sum_j W_{ij}$, in which $W_{ij} = \sigma^{AV}_{i,j}$. 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, P can be represented as follows:

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

in which I is an identity matrix, P_{ul} is the matrix of probabilities of untagged nodes walking to tagged ones, and P_{uu} is the matrix of probabilities of untagged nodes walking to other untagged ones.

After creating P, the random walk with t steps can be done by $P^t = (\omega \times P \times P^{t-1}) + ((1-\omega) \times P)$, with ω set to 0.5. The use of the ω factor guarantees a slower walk that is consistent with the initial state of the probability matrix. After the random walk is applied on P, we choose to assign to u the tag of the node l with maximal P^t_{ul} value. This value is also taken as the confidence score.

3.1.2 Hierarchical strategy (MST). This method makes use of the Kruskal algorithm for propagating tags between sets hierarchically, since a minimum spanning tree establishes a hierarchical partition of a set [10].

Given a *speaking face* graph $\mathcal{G}$, it is possible to build $\mathcal{G}' = (V, E)$ where the edge weights W'_{ij} represent distances between speaking faces (*i.e.*, $W'_{ij} = 1 - \sigma^{AV}_{i,j}$), and a null graph $\mathcal{H}$ is created with $V_{\mathcal{H}} = V_{\mathcal{G}'}$. At this point, each vertex on $\mathcal{H}$ represents a unitary set. After creating $\mathcal{H}$, the following process is repeated until all nodes of $\mathcal{H}$ belong to the same set or all edges of $\mathcal{G}'$ are visited: sort all edges on $\mathcal{G}'$. Take the unexamined edge $E_{i,j}$ with the smallest weight and check if V_i and V_j belong to the same set. If they do, skip to the next smallest edge, and if they do not, perform a union of the two sets respective to the nodes V_i and V_j. The tag propagation occurs on the merging phase, and it proceeds as follows: (i) if only one of the sets is tagged, its tag propagates to all nodes belonging to the other set; (ii) if none of the sets is tagged, nodes of both sets remain untagged; and (iii) if both sets are tagged, their tags do not change, and one of the tags is randomly taken to represent the new set formed (this representative tag will be the one propagated to other groups when the new set eventually merges with another one). When propagating a tag to an untagged set, the confidence score is based on the edge weight used on merging the two sets.

3.2 Managing multimodal information

When performing tag propagation methods (random walk and hierarchical), we consider two different ways for fusing the audio and visual modalities: (i) a "score-fusion" approach: visual and audio similarities are combined using a weighted average, *i.e.*, $\sigma^{AV} = f(\sigma^V, \sigma^A) = \lambda\sigma^V + (1-\lambda)\sigma^A$, in which λ is the range $[0, 1]$; and (ii) a late fusion approach: tag-propagation is done for each modality (producing two confidence scores). This is equivalent to use two distinct functions (with $\lambda = 1$ or $\lambda = 0$): $\sigma^{AV}_1 = \sigma^V$ and $\sigma^{AV}_2 = \sigma^A$. Then, the tag with the highest confidence score is kept for each *speaking face*.

4 EXPERIMENTS

In this section we study the impact of our proposed tag-propagation approaches with respect to the case where no propagation is performed, along with two baseline methods. First, we present our evaluation set-up, followed by a detailed description of features and similarity measures used; and, finally, we describe and discuss the results obtained through our various configurations of the methods. The experiments were run on a Intel i3-2310M CPU @ 2.10GHz with 8GB of 1333MHz DDR3 RAM.

4.1 Baselines

Two baselines were made using graph clustering approaches to propagate the initial labels over the *speaking faces*. The graph clustering techniques used were the spectral clustering (from SciPy toolkit) and Markov chain clustering by flow simulation [15]. Apart from the different clustering techniques, the two baseline methods share the same protocol.

Initially, a clustering technique is applied on a graph $\mathcal{G}$ with the number of clusters set to the total number of distinct initial tags on the video plus one. Then, a histogram of tags is calculated on each cluster, and the one with the greatest number of incidence will be used to tag the untagged nodes on that cluster. The previously tagged nodes will not have their tags changed, and some nodes might end without any tag.

4.2 Evaluation setup

Datasets. We evaluate the approaches described previously using the set of videos which were manually annotated during the MediaEval Person Discovery task in 2016 [1]. The initial catalogue of videos used is composed of 196 hours of broadcast news from two French TV channels (France2, France5), 50 hours of short documentaries from the Deutsche Welle TV channel in English and German, and 13 hours of broadcast news from the Catalan TV channel 3-24. These videos being divided automatically into shots, and a subset of 3431 shots related to 763 videos was manually annotated by the participants of that evaluation task.

Features, similarity measures and parameters. We use 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 [4], the segments of speech obtained with the speaker diarization system from LIUM [12], the face tracks obtained with a histogram of oriented gradients-based detector and a correlation tracker. We computed the following features: (i) visual feature: each face track is first represented by its central face (key face), which is then described by a generic image descriptor computed by a convolutional neural network (CNN) pre-trained on the ImageNet dataset of features; and, finally, we extract the last convolutional layer and perform average pooling and "power normalization", *i.e.*, square root compression followed by L2 normalization; (ii) audio feature: each speech segment is described by a Gaussian Mixture Model learned on mel-cepstral features and log energy with a hop size of 10 ms; and (iii) a name (tag) detection: the overlays of the dataset provided by MediaEval are filtered – only words/expressions tagged as *pers* were kept. The visual similarity σ^V between two facetracks is obtained by computing the dot product between the CNN-based features of their respective key faces. The distance between two speech segments is computed using an approximation of the Kullback-Liebler divergence. The audio similarity $\sigma^A_{i,j}$ is set to $\exp(-0.5\ \delta^A_{i,j})$, in which $\delta^A_{i,j}$ is the distance between speech segments i and j. The number of random walk steps t is set to 50 and λ is set to 0.5.

4.3 Quantitative assessment

Early versions of these systems were ranked amongst the top state-of-the-art systems in the 2016 MediaEval Person Discovery contest (see "MOTIF" team). For more details, [6] proposes a comparative study of the systems using common features, the first and the last systems obtaining MAP@1,10 and 100 of 73.6%, 59.8%, 57.9% and 45.6%, 38.4%, 37.0% respectively on a similar dataset.

The comparison between the proposed methods and the baselines are given in terms of the Mean Average Precision at K (MAP@K), as in MediaEval [1]. In Table 1, we illustrate the scores of the methods. In boldface, we highlight the two best methods for each MAP@K. We see that all methods improve the scores when compared to the use of the initial tags only. Moreover, it can be observed that the tag propagation methods (RW and MST) outperform the two clustering based baselines. Furthermore, except for MAP@1, the random walk (RW) and hierarchical approach with late fusion (MST_Late) outperform the other strategies. As shown on Table 1, the way of fusing the two modalities behave differently on our tag propagation methods. On the hierarchical propagation, late fusion performs better than the score-fusion approach, but the opposite happens when using random walks for propagation.

To compare the two best performing methods (RW and MST_Late), we measure their level of agreement by using the Kappa coefficient, perform a paired t-test and compare their processing time. The Kappa coefficient scored a level of agreement of 0.847 between the two methods, which according to [9] can be considered as an almost perfect agreement. However, according to the paired t-test with a significance level of 0.05, the random walk method is considered superior than the MST_Late. Furthermore, the processing time of RW was $7m12s$, while processing time of MST_Late was $1m42s$.

Table 1: MAP@K results of the compared methods.

	MAP@1	MAP@5	MAP@10	MAP@100
No Propagation	0.543	0.342	0.323	0.312
Spectral	0.624	0.468	0.448	0.434
Markov	0.649	0.494	0.470	0.453
MST	0.676	0.559	0.536	0.518
MST_Late	0.686	**0.565**	**0.541**	**0.523**
RW	**0.696**	**0.570**	**0.550**	**0.530**
RW_Late	**0.689**	0.561	0.538	0.521

5 CONCLUSION

In this paper we presented a graph-based approach for discovering persons on video, along with two propagation methods and two different treatments for multi-modality of information. We also proposed a comparison of our methods against two graph clustering baselines. We observed that a graph-based approach is suited for tackling such problem and that the proposed propagation methods perform better than the graph clustering baseline for this task.

Among all compared strategies, the methods based on hierarchical propagation with late fusion and random walk with *score-fusion* obtained the highest MAP values. We statistically showed that the random walk performs better than the hierarchical propagation with late fusion, even though these two methods have a very high level of agreement according to the Kappa coefficient. Moreover, even though the random walk propagation performed almost 2% better than the hierarchical one, the latter is 4.23 times faster than the first one.

It is shown that the different use of the modality fusions (late fusion) works differently for each propagation method, enhancing the scores on the hierarchical propagation, but decreasing them on the random walk. This suggests that a tuning of the modality fusion parameters could bring us more consistent results, and it should be investigated in the future.

ACKNOWLEDGEMENTS

The authors thank CNPq (Grant #477457/2013-4), CAPES (Grant STIC-AMSUD 001/2013) and FAPEMIG (Grants APQ-01806-13 and CEX-APQ-03195-13). This work was partially supported by the STIC AmSud program, under the project "Unsupervised Mining of Multimedia Content", and by the Inria Associate Team program.

REFERENCES

[1] HervÃĺ Bredin, Claude Barras, and Camille Guinaudeau. 2016. Multimodal Person Discovery in Broadcast TV at MediaEval 2016. In *Working notes of the MediaEval 2016 Workshop.*

[2] Hervé Bredin, Anindya Roy, Viet-Bac Le, and Claude Barras. 2014. Person Instance Graphs for Mono-, Cross- and Multi-Modal Person Recognition in Multimedia Data. Application to Speaker Identification in TV Broadcast. *International Journal of Multimedia Information Retrieval* (2014).

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

[4] Datong Chen and Jean-Marc Odobez. 2005. Video text recognition using sequential Monte Carlo and error voting methods. *Pattern Recognition Letters* 26, 9 (July 2005), 1386–1403.

[5] Yannick Estève, Sylvain Meignier, Paul Deléglise, and Julie Mauclair. 2007. Extracting true speaker identities from transcriptions. In *INTERSPEECH 2007 – ICSLP.* 2601–2604.

[6] Nam Le et al. 2017. Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proc. of International Workshop on Content-Based Multimedia Retrieval*.
[7] P. Gay, G. Dupuy, C. Lailler, J. M. Odobez, S. Meignier, and P. DelÃľglise. 2014. Comparison of two methods for unsupervised person identification in TV shows. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 1–6.
[8] J. Kahn, O. Galibert, L. Quintard, M. CarrÃľ, A. Giraudel, and P. Joly. 2012. A presentation of the REPERE challenge. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 1–6.
[9] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[10] Benjamin Perret, Jean Cousty, Jean Carlo Rivera Ura, and Silvio Jamil F Guimarães. 2015. Evaluation of morphological hierarchies for supervised segmentation. In *Proceedings of the 12th International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*. Springer, 39–50.
[11] Johann Poignant, Guillaume Fortier, Laurent Besacier, and Georges Quénot. 2016. Naming multi-modal clusters to identify persons in TV broadcast. *Multimedia Tools Appl.* 75, 15 (2016), 8999–9023.
[12] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meigner. 2013. An open-source state of the art toolbox for broadcast news diarization. In *Interspeech*. 25–29.
[13] Gabriel Sargent, Gabriel Barbosa de Fonseca, Izabela Lyon Freire, Ronan Sicre, Zenilton Kleber Gonçalves do Patrocínio Jr., Silvio Jamil Ferzoli Guimarães, and Guillaume Gravier. 2016. PUCMinas and IRISA at Multimodal Person Discovery. In *Working Notes Proceedings of the MediaEval 2016 Workshop*.
[14] S. E. Tranter. 2006. Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio. In *2006 IEEE ICASSP*, Vol. 1. I–I.
[15] Stijn Marinus Van Dongen. 2001. *Graph clustering by flow simulation*. Ph.D. Dissertation.
[16] Jun Yang, Rong Yan, and Alexander G. Hauptmann. 2005. 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, 31–40.
[17] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, Vol. 3. 912–919.