# PUC Minas and IRISA at Multimodal Person Discovery

Gabriel Sargent, Gabriel Barbosa de Fonseca, Izabela Lyon Freire, Ronan Sicre, Zenilton Do Patrocínio Jr., Silvio Guimarães, Guillaume Gravier

# PUC Minas and IRISA at Multimodal Person Discovery

Gabriel Sargent[1], Gabriel Barbosa de Fonseca[2], Izabela Lyon Freire[2], Ronan Sicre[1], Zenilton K. G. Patrocínio Jr[2], Silvio Jamil F. Guimarães[2] and Guillaume Gravier[1],
[1]IRISA & Inria Rennes , CNRS and Univ. Rennes 1, Rennes, France
[2]Computer Science Department - PUC de Minas Gerais, Belo Horizonte, Brazil

## ABSTRACT

This paper describes the systems developed by PUC Minas and IRISA for the person discovery task at MediaEval 2016.

We adopt a graph-based representation and investigate two tag-propagation approaches to associate overlays co-occurring with some speaking faces to other visually or audio-visually similar speaking faces.

Given a video, we first build a graph from the detected speaking faces (nodes) and their audio-visual similarities (edges). Each node is associated to its co-occurring overlays (tags) when they exist. Then, we consider two tag-propagation approaches, respectively based on a random walk strategy and on Kruskal's algorithm.

## 1. INTRODUCTION

The task of multimodal person discovery in TV broadcast consists in identifying persons of a video corpus which both speak and are visible at the same time, in an unsupervised way [2]. Most approaches to the task use clustering, either of face tracks or of voice segments (or both) before finding a good match between text in overlays and clusters [6, 4]. While this type of approaches worked well in 2015, we believe that the clustering steps involved are error prone. Indeed, errors in the clustering step cannot be undone afterwards in the naming stages. In 2015, IRISA and UFMG proposed a graph-based approach in which each node corresponds to a speaking face and edges to the similarity between its vertices [3]. The similarity can be computed at the visual level, the voice level or both. Names can be associated to nodes based on co-occurrences of a speaking face and names overlays. However, only a small fraction of the nodes can be tagged by this method. Hence, in 2016, we studied tag propagation algorithms that take advantage of the graph structure to assign tags to nodes with no overlapping overlays, thus potentially improving recall. Tab. 1 recaps the different configurations submitted.

## 2. GRAPH GENERATION

Each video is modeled by a graph where each node represents a speaking face, and each edge quantifies the visual or audiovisual similarity between two speaking faces. A speaking face is defined as the association of a facetrack (sequence of faces related to the same person in adjacent video frames) with the speech segment for which the overlap is maximum and at least 60 %. The facetracks and speech segments are the ones provided by MediaEval, the latter being extracted from the speaker diarization result disregarding the arbitrary speaker number.

| Submission | Similarity | | Tag propagation |
|---|---|---|---|
| | audio | video | |
| primary ($p$) | binary | CNN | hierarchical |
| contrast 1 ($c1$) | GMM | CNN | random walk |
| contrast 2 ($c2$) | – | CNN | hierarchical |
| contrast 3 ($c3$) | GMM | CNN | hierarchical |
| contrast 4 ($c4$) | – | – | – |

**Table 1: Components of the systems at the origin of our 5 submissions.**

### 2.1 Audiovisual similarities

We consider three weighting schemes for the edges in the graphs, resulting from the combination of different strategies to combine visual similarity and voice similarity.

The visual similarity $S^V_{ij}$ between two facetracks $i$ and $j$ is calculated as follows. A key face is selected from the central frame of each facetrack, from which a generic image descriptor is computed by applying a very-deep convolutional neural network pre-trained on the ImageNet dataset [8]. Specifically, we extract the last convolutional layer [9] and perform average pooling and "power normalization", *i.e.*, square-root compression followed by L2-normalization. Finally, $S^V_{ij}$ is calculated as the cosine similarity between the descriptors of the two key face images.

Voice similarity can be computed two ways. A simple binary audio similarity is derived from the speaker diarization provided by MediaEval, where the similarity is 1 if the two segments are labeled with the same speaker in the diarization. Alternately, the audio similarity $S^A_{ij}$ between two segments can be calculated as follows. Each speech segment is modeled with a 16-Gaussian mixture model (GMM) over Mel cepstral features. The distance $D^A_{ij}$ is computed using the Euclidean-based approximation of the KL2 divergence between the two GMMs [1], and turned into a similarity according to $S^A_{ij} = \exp(\log(\alpha)\ D^A_{ij})$, where $\alpha = 0.25$ in the experiments here.

Fusion of the visual and voice similarities is done by a weighted average, $S^{AV}_{ij} = \beta S^V_{ij} + (1-\beta) S^A_{ij}$. We experimentally set $\beta = 0.85$ in the case of binary voice comparison and $\beta = 0.5$ for the GMM-based comparison.


MediaEval 2016 Workshop, Oct. 20–21, 2016, Hilversum, Netherlands.

## 2.2 Tag initialization

Initially, each node in the graph is tagged using the overlay for which the overlap with the facetrack is maximum. We used the overlay detection and name recognition provided (output from the OCR system 2), which we filtered using the named entity detector NERO [7], keeping only words tagged as "pers" by the named entity recognition. Note that this approach is rather aggressive as NERO was initially designed for the speech transcription in the French language. In practice, many nodes are not tagged as they do not overlap with a valid overlay (Sets T15 and T16, introduced in Section 4, show respectively 25.5% and 6.6% of nodes initially tagged). This is why tag propagation is required.

# 3. TAG PROPAGATION APPROACHES

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

## 3.1 Random walk tag propagation

In a graph where transition probabilities between nodes are known, the probability of ever reaching node $j$ starting from node $i$ can be calculated using a random walk strategy with absorbing states [10]. Let $n$ be the number of nodes of the graph, we define a symmetrical weight matrix $\mathbf{W} = \{\mathbf{W}_{ij}\}_{1\leq i,j\leq n}$, where $\mathbf{W}_{ij}$ is the similarity between nodes $i$ and $j$, and a diagonal *degree matrix* $\mathbf{D} = \{\mathbf{D}_{ij}\}_{1\leq i,j\leq n}$, where $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. The transition probability matrix $\mathbf{P}^0 = \{\mathbf{P}^0_{ij}\}_{1\leq i,j\leq n}$, where $\mathbf{P}^0_{ij}$ is the probability of reaching node $j$ from node $i$ in one step, is given by $\mathbf{P}^0 = \mathbf{D}^{-1}\mathbf{W}$. Tagged nodes are set as *absorbing states* in $\mathbf{P}$, according to

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & 0 \\ \mathbf{P}_{ul} & \mathbf{P}_{uu} \end{pmatrix},$$

where $l$ is the set of tagged nodes, $u$ is the set of untagged nodes, $\mathbf{I}$ is an identity matrix of size $|l| \times |l|$, $\mathbf{P}_{ul}$ contains probabilities of untagged nodes ending their walk on tagged nodes, and $\mathbf{P}_{uu}$ contains probabilities of untagged nodes getting to other untagged nodes. We denote $\mathbf{P}^t$ the transition probability after $t$ iterations. The random walk iteration is performed according to $\mathbf{P}^{t+1} = (1-\gamma)\ \mathbf{P}^0\ \mathbf{P}^t + \gamma\ \mathbf{P}^0$, where $\gamma$ is a parameter enforcing the consistency of the initial state (here, $\gamma = 0.4$). Once the random walk has converged ($\sum_{i,j} |\mathbf{P}^{t+1}_{i,j} - \mathbf{P}^t_{i,j}| < 10^{-9}$), each untagged node is associated to the tagged one on which it has the highest probability to end its walk, i.e., each row index of $\mathbf{P}_{ul}$ is matched with the column index with maximal probability. This maximal probability is kept as the confidence score.

## 3.2 Hierarchical tag propagation

This method is based on the computation of a minimum spanning tree (MST) from an undirected weighted graph, using Kruskal's algorithm. The MST establishes a hierarchical partition of a set [5]. A connected graph $\mathcal{G}$ is given (see Section 2), where edge weights represent distances (functions of their respective similarities $S^{AV}$). To propagate the initial tags, we start from a null graph $\mathcal{H}$ on $\mathcal{G}$'s nodes, and the following process is repeated, until all edges of $\mathcal{G}$ are examined: from $\mathcal{G}$, the unexamined edge $e$ corresponding to the smallest distance is chosen. If it does not link different trees in $\mathcal{H}$, skip it; otherwise, it links trees $T_1$ and $T_2$ (thus forming $T_3$), and $e$ is added to the minimum spanning forest $\mathcal{H}$ being created; three cases are possible: **I.** None of $T_1, T_2$ is tagged: $T_3$ will not be tagged **II.** Only $T_1$ is tagged, with confidence score $C_{T_1}$: $T_1$'s tag is assigned to the entire $T_3$ (*i.e.,* to all its unlabelled nodes), with a confidence score $C_{T_3} = C_{T_1} \times (1 - w_e)$, where $w_e$ is the weight of $e$ in $\mathcal{G}$. **III.** Both $T_1$ and $T_2$ are tagged: one of the tags (of $T_1$ or of $T_2$) is picked (at random), and assigned to $T_3$ with confidence scores as in case II.

| | MAP@1 | | MAP@10 | | MAP@100 | |
|---|---|---|---|---|---|---|
| | T15 | T16 | T15 | T16 | T15 | T16 |
| primary ($p$) | 87.9 | 64.4 | 82.1 | 49.3 | 81.9 | 47.8 |
| contrast 1 ($c1$) | 87.9 | 64.4 | 79.8 | 48.4 | 79.6 | 46.7 |
| contrast 2 ($c2$) | 87.9 | 62.9 | 81.7 | 46.2 | 81.5 | 44.8 |
| contrast 3 ($c3$) | 87.9 | 63.6 | 80.2 | 49.3 | 80.0 | 47.5 |
| contrast 4 ($c4$) | 87.9 | 56.8 | 79.7 | 36.1 | 79.5 | 35.1 |
| $(p - c4)/c4$ | 0.0 | 13.4 | 3.0 | 36.6 | 3.0 | 36.2 |

**Table 2: Mean average precision at different ranks (in %) for the 5 submissions. Last row gives the relative improvement of the primary run over the no-propagation baseline ($c4$).**

| | T15 | T16 |
|---|---|---|
| MAP@1 | $p = c1 = c2 = c3 = c4$ | $p = c1, c3, c2, c4$ |
| MAP@10 | $p, c2, c3, c1, c4$ | $p, c2, c3, c1, c4$ |
| MAP@100 | $p, c2, c3, c1, c4$ | $p, c3, c1, c2, c4$ |

**Table 3: Ranking (best first) of the submissions.**

# 4. RESULTS

Tab. 2 reports the results obtained on the 2015 and 2016 test data (T15=development data for 2016, and T16, respectively). For T16, the reference annotation dump of 2016/09/14 is used. The rank of the submissions is shown in Tab. 3. All tag propagation approaches improve over the no-propagation baseline ($c4$), the interest of tag propagation being much clearer on T16. The baseline highlights noticeable differences between T15 and T16. In T15, propagation was almost useless as most nodes could be tagged in the initial stage. This is not the case in T16 where tag propagation yields significant gain. The hierarchical tag propagation on graphs combining CNN visual similarity and binary voice similarity (primary) consistently outperforms other combinations, showing the interest of combining audio and visual similarities. Comparing approaches, $c3$ usually (except for T16, MAP@1) performs better than $c1$, indicating that the hierarchical tag propagation performs better than the random walk, at least with GMM-CNN audiovisual similarities. The comparison of $c3$ and $c1$ shows the weakness of the GMM-based voice comparison over the state-of-the-art approach used for diarization. Finally, the comparison of $c3$ and $c2$ gives mixed feelings. The use of the GMM-based voice comparison decreases performance in most cases except on T16 at $K = 1, 100$.

# 5. ACKNOWLEDGEMENTS

Work supported by FAPEMIG/INRIA/MOTIF (CEX-APQ 03195-13), FAPEMIG/PPM (CEX-PPM-6-16) and CAPES (064965/2014-01).

## 6. REFERENCES


[1] M. Ben, M. Betser, F. Bimbot, and G. Gravier. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *Proceedings of the 8th International Conference on Spoken Language Processing*, pages 333–444, October 2004.

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

[3] C. E. dos Santos Jr, G. Gravier, and W. R. Schwartz. SSIG and IRISA at Multimodal Person Discovery. In *Working notes of the MediaEval 2015 Workshop*, September 2015.

[4] N. Le, D. Wu, S. Meignier, and J.-M. Odobez. EUMSSI team at the MediaEval Person Discovery Challenge. In *Working notes of the MediaEval 2015 Workshop*, September 2015.

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

[6] J. Poignant, L. Besacier, and G. Quénot. Unsupervised speaker identification in TV broadcast based on written names. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(1):57–68, 2015.

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

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the 2016 International Conference on Learning Representations*, 2016.

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