Unsupervised Face Identification in TV Content using Audio-Visual Sources

Meriem Bendris[1], Benoit Favre[1], Delphine Charlet[2], Géraldine Damnati[2],
Grégory Senay[3], Rémi Auguste[4], Jean Martinet[4]
[1]Aix Marseille Université, [2]OrangeLabs, [3]Université d'Avignon[4], Université Lille 1
{firstname.lastname}@{[1]lif.univ-mrs.fr,[2]orange.com,[3]univ-avignon.fr,[4]lifl.fr}

Abstract

Our goal is to automatically identify faces in TV content without pre-defined dictionary of identities. Most of methods are based on identity detection (from OCR and ASR) and require a propagation strategy based on visual clusterings. In TV content, people appear with many variation making the clustering very difficult. In this case, identifying speakers can be a reliable link to identify faces. In this work, we propose to combine reliable unsupervised face and speaker identification systems through talking-faces detection in order to improve face identification results. First, OCR and ASR results are combined to extract locally the identities. Then, the reliable visual associations are used to propagate those identities locally. The reliable identified faces are used as unsupervised models to identify similar faces. Finally speaker identities are propagated to the faces in case of lip activity detection. Experiments performed on the REPERE database show an improvement of the recall of $+5\%$ compared to the baseline, without degrading the precision.

1 Introduction

The recent abundance of multimedia content requires the development of technologies to navigate through these data. Persons are central to videos and indexing their presence and mentions could enable new efficient ways of browsing for interesting content. Most visual indexing methods are based on face detection and recognition. Those methods require large databases of facial models trained to recognize each person who could appear in a video. However, the variability of face appearance in TV content (pose, facial expressions, lighting, occlusions) makes identification using facial models very unreliable. In addition, maintaining up-to-date large dictionaries of face models is prohibitively expensive. In this paper, we are interested in unsupervised methods for naming faces in TV content.

Unsupervised person identification methods are often performed in two steps: (1) names are extracted from a range of sources and (2) an association-propagation strategy assigns each detected name to a person. In the first step, the identities can be extracted from speech (using Automatic Speech Recognition [1, 2]), image (with Optical Character Recognition [3] on overlaid text) and text content (such as scripts and subtitles [4]). In the second step, the extracted identities are propagated via clustering methods [3, 5]. This step is the focus of our paper. Figure 1 illustrates that process on a debate video from the *REPERE*[1] corpus [6].

We propose to directly associate OCR and speech detected names with current faces and speakers, and then propagate that information within and cross modalities with face and speaker similarities and talking face detection. This paper is organized as follows: Section 2 describes related work; Section 3 describes person name acquisition from *OCR* and *ASR* output; Section 4 similarity measures for speaker and face clustering; Section 5 presents our identity propagation method based on direct and indirect association. Finally, Section 6 presents the *REPERE* corpus, results of experiments and a discussion.

FIGURE **1. The *REPERE* corpus. The identity appears in multiple sources.**

2 Related work

Several studies have addressed the problem of association-propagation strategies for unsupervised face identification. Name-it [7] proposed to find face-name

1. Reconnaissance de PERsonnes dans des Emissions Audiovisuelles : www.defi-repere.fr

associations by maximizing the co-occurrence between similar faces and names extracted from OCR output. [8] proposed to name faces in images using a graphical model for face clustering. Nodes represent detected faces and edges are weighted by SIFT-based similarity. Then, for each name detected in OCR, greedy search is applied to find the sub-graph that maximizes face similarities within the set of faces associated to the name. However, this approach cannot identify faces if no name is detected in the image. In [9], authors proposed to identify faces in *TRECVID* news videos using training data obtained automatically from Google image search. Names were extracted from both OCR and ASR output. In [4], authors proposed to align detected faces with names from the script and used rules based on lip activity and gender detection to resolve ambiguities. In [5], names are extracted from movie scripts and subtitles and associated to faces according to lip activity. Identities are then propagated using face-level and clothes-level similarities. Although preliminary results are promising, face and clothes variability (pose, expression, color...) hamper the robustness of the similarity measure. In this case, audio information can be used in addition to visual cues to associate names to faces through speaker identity. In fact, in TV content speaker diarization appears to be more robust than face clustering [10]. [1, 2] proposed to extract names using *ASR* output and associated them to speakers using lexical rules on speaker clusters. In [3], names are extracted from OCR output and propagated to speaker clusters in order to maximize co-occurrence.

3 Name acquisition

This section describes how names are detected in video and speech.

From overlaid text: For person identification, a multi-stage system has been specifically designed in order to detect and recognize Overlaid Person Names (OPN). Text detection is achieved on each frame with a convolutional neural network approach described in [11]. It is limited to a predefined area of interest (usually at the bottom of the frame) which contains text directly related to the show. Each text region is tracked on consecutive frames using bounding box overlap. A confusion network is built by aligning character sequence hypotheses of consecutive frames resulting in a character sequence hypothesis for the text track.

Two open source video OCR engines are used: GOCR [2] and Tesseract [3]. Their resulting character sequence hypotheses for a given track are merged to form confusion networks from which the most probable sequence is extracted. A rule-based classifier distinguishes Overlaid Person Names (OPN) from any other text. It relies on a diverse set of rules including the number of words, vicinity of another box (e.g. the person's job), etc. Finally, the character sequence hypothesis associated to a detected OPN is submitted to a normalization module which consists in finding the most similar name in a large dictionary of person names (e.g. recognized sequence "Valérie Péciesse" in Figure 1 is approximated and normalized as "Valerie_PECRESSE"). The dictionary is obtained from the WEB. It covers 90% of people in the REPERE corpus.

From spoken content: Uttered Person Name (UPN) extraction in spoken content is achieved by two modules which are based on the output of an Automatic Speech Recognition (ASR) [12] which transcribes audio signal to text. The first module extracts names from the transcript generated by ASR. However, the transcription task is difficult because words are often misrecognized and the lexicon from which the system generates its output is limited. This is especially problematic for UPNs as proper names are infrequent and therefore often out of vocabulary (OOV).

In order to find misrecognized or OOV person names, the second module mines the phoneme confusion network [13] formed from the output of the ASR system. Unlike pure phonetic decoding output, this representation has the advantage of integrating linguistic constraints which make it much more suitable for mining names. For instance first names are likely to be in the lexicon and well recognized by the system while family names are more likely to be OOV. We perform a Levenshtein distance between the phonetic representation of candidate UPNs and the confusion networks. The alignment score of a UPN is defined as the sum of the phoneme probabilities from the aligned phonemes normalized by the number of phonemes in the name.

The output of the UPN detection system is the union of the UPNs found in the ASR transcript and those detected in the phoneme confusion network with a score higher than 0.85. The list of names used as candidates is the same as for OCR-extracted names.

4 Audio and visual diarization

The task of diarization aims at determining for each pair of (visual or acoustic) frames whether it contains the same person. This task is often referred to as clustering.

Speaker diarization: The speaker diarization system is based on [14]: First, agglomerative clustering of speech segments is performed based on Bayesian Information criterion (each cluster is modelled with a single full covariance matrix). Then, that initial set of clusters is modeled with Gaussian mixtures in order to more accurately compare voices using a cross-likelihood criterion for another pass of agglomerative clustering. At each iteration, Viterbi decoding

2. http://www.jocr.sourceforge.net
3. http://code.google.com/p/tesseract-ocr/

is performed to re-segment the speech data into speaker turns given the new clusters.

Face diarization: Faces are detected using OpenCV's cascade classifier [15] for frontal and profile faces. The resulting detections are tracked until shot boundaries using bounding box overlap. Then, the upper body is detected using a background subtraction algorithm based on Grabcut [16], initialized with detected face. The background subtraction algorithm yields a very accurate silhouette of the person, even in presence of a dynamic background. Each extracted person is then modelled using a space-time color histogram [17]. This model stores color along with geometric and time information. It allows to retain the aspect of the person as it moves throughout the shot. A similarity matrix is build between person tracks using a combination of Bhattacharyya coefficient and Mahalanobis distance [17].

Finally, agglomerative clustering is performed on the similarity matrix using the Ward criterion. The algorithm stops on a show-specific *a priori* number of clusters determined on a development set.

5 Propagation strategy

The problem of associating faces with names can be cast as bipartite matching [18] where a face is linked to at most one name. This approach is appealing but assumes that face clustering results are very accurate because it cannot resolve ambiguity. When processing unconstrained videos, high accuracy is difficult to reach because of pose and appearance variability of faces. Alternatively one can work on face tracks before clustering and perform multi-modal clustering where names and face tracks are grouped in clusters using inter and intra modal similarities. Unfortunately, most multi-modal clustering techniques assume that the similarities are comparable, a constraint difficult to enforce in heterogeneous modalities (names and faces). Combinatorial random Markov field clustering [19] or correlational clustering [20] are good contender multi-modal clustering approaches that can account for differences between similarity spaces but both assume that all modalities are always observed whereas identities can only be directly bound to faces on a subset of frames. Therefore, while keeping the idea of heterogeneous clustering, we subdivide the problem in two tasks: *direct identification* when direct evidence of face-name association is available and *indirect identification* when evidence comes from previous clustering decisions. Each type of identification is performed on a single modality couple, effectively removing the similarity scale problem. Even though we loose optimality, making decisions in high confidence regions yields reasonable performance in practice.

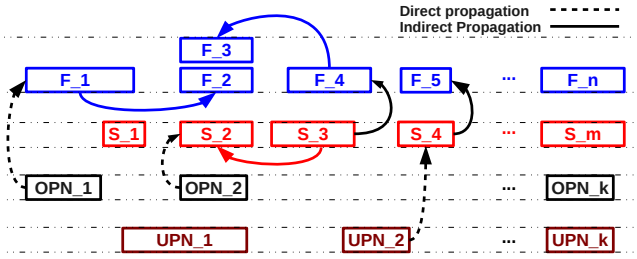

FIGURE **2. Direct and indirect identifications.**

In the following, we detail the identity association and propagation for all pairs of modalities used in our method. Four types of objects are considered: two name sources (*OPN* for Overlaid Person Names and *UPN* for Uttered Person Names) and two person instance modalities (*Face* and *Speaker*). Direct identification is obtained by local propagation of *OPNs* and *UPNs* to *Faces* and *Speakers*. Indirect identification is performed through *Face* and *Speaker* similarities and lip activity detection. Figure 2 shows an example of direct and indirect identification.

5.1 Direct identification

OPN $\rightarrow$ Face: We make the assumption that most OPNs occur while the corresponding face appears on the screen. Statistics on the *REPERE* corpus presented in Table 1 corroborate this idea, showing that 98.5% of the annotated frames containing an OPN also contain the corresponding face. Consequently in unambiguous shots where only one face is detected, we locally propagate the OPN to the face track.

Then, for ambiguous shots where multiple faces could be identified by an OPN, we make a global decision using bipartite matching. For a given OPN, potential face sets are formed by gathering all face tracks that do not co-occur in the same shot. Then, that name is associated to the purest cluster containing all shots it occurs in.

Let N be the set of all OPNs, F_n be the set of faces that co-occur with name $n \in N$ and G_n the set of face clusters from F_n that span all shots where n occurs but do not contain faces from the same shot.

$$G_n = \left\{ \begin{array}{l} g \in \mathcal{P}(F_n) \\ \exists f \in g : \text{shot}(f) = s \, \forall s \in \text{shots}(n) \\ \text{shot}(f_1) \neq \text{shot}(f_2) \, \forall f_1, f_2 \in g^2 \end{array} \right\}$$

where $\text{shot}(f)$ is the shot of face track f and $\text{shots}(n)$ is the set of shots where name n appears. We define the following dispersion measure:

$$D_1(g) = \sum_{f_i \in g} \sum_{f_j \in g} d(f_i, f_j)$$

where $d()$ is the distance between two face-tracks as defined in Section 4. Then, the face cluster $\hat{g}_n$ bound to name n is the one

which minimizes dispersion:

$$\hat{g}_n = \underset{g \in G_n}{\operatorname{argmin}} D_1(g)$$

OPN → Speaker: As for direct face identification, the assumption that OPNs occur when the corresponding speaker talks is used to identify speakers (statistics in Table 1 show that 80.4% of the annotated frames containing an OPN also contain the corresponding speaker). For a given speaker, we associate it with the OPN that overlaps the most with it. If there is no overlap, the speaker remains anonymous.

UPN → Face: Generally speaking, it is very difficult to guess whether an uttered person name (UPN) identifies one of the displayed faces. Even with proper understanding of the whole spoken content, the task of predicting visual presence only from uttered names remains hard, even for human annotators [21]. In this study, we restrict the $UPN \rightarrow Face$ association to the identification of static faces from photographs. Such photographs are usually inserted to illustrate the spoken content leading us to identify it among Uttered Person Names. The static face detector combines a measure of head movement and a lip activity detector. For each face track *f*, we compute a dispersion ratio as the number of different positions of the detected face within the track. *f* corresponds to a photograph if the dispersion ratio is below a given ratio *and* if the lip activity value is above a given threshold. It is hence associated to the UPN that occurs the most during the track.

UPN → Speaker: Uttered names rarely refer to the current speaker, but rather to other speakers. A method of determining whether UPNs identify the previous, current or next speaker or a third party is described in [22]. It was shown to perform well on radio broadcasts where speech is the only modality but it is less suited for processing videos where speakers rely on multimodal cues for introducing other people. Instead of searching for a speaker to be linked to a given UPN, we search UPNs that can correspond to the current speaker. As an approximation we consider that potential candidates may be found in a time window immediately preceding or following the current speaker turn. The closest name in this window is chosen.

5.2 Indirect identification

Face → Face: While direct identification can recover the name of faces which co-occur with OPNs and UPNs, indirect identification propagates this information to other faces. Two methods of face-based name propagation are explored. First, a method relying on face-clustering is evaluated as a baseline: each face cluster (results of the process described in Section 4) is named from the OPN which it co-occurs most with. Then, after the direct OPN propagation, each unnamed face gets the name given to the cluster it belongs to. This approach is called "*Cluster-based*" in the rest of the paper. Secondly, we propose an alternative to face clustering, based on the principle that directly-named faces are very reliable, and can be considered as *model* in an open-set face identification paradigm. Let $\hat{g}_n$ be the set of faces directly associated to name n, for each face f with no direct naming, a distance D_2 is defined between this face and $\hat{g}_n$. The name $\hat{n}(f)$ given to face f is the name for which the distance is minimal, if the distance $< \theta_1$.

$$\hat{n}(f) = \begin{cases} \operatorname{argmin}_{n \in N} D_2(\hat{g}_n, f) \text{ if } D_2(\hat{g}_n, f) < \theta_1 \\ \emptyset \text{ otherwise} \end{cases}$$

with $D_2(g, f) = \frac{1}{|g|} \sum_{f_i \in g} d(f, f_i)$. This approach is called "*Similarity-based*" in the rest of the paper

Speaker → Speaker: Each speaker cluster is named from the OPN which it co-occurs most with. Then, each speaker segment that was not already named locally gets the name associated to the cluster it belongs to.

Speaker → Face: At the end of the $Face \rightarrow Face$ propagation, there remain faces that are not named. Either because no OPN was present in the face cluster they belongs to in the face-clustering based propagation, or because their distance always exceed θ_1 in the open-set identification-based propagation. For those unidentified faces, speaker identities are used. In order to propagate identities between speaker segments and face tracks, we detect whether the face is talking at the same time as the speaker is uttering speech. Talking face detection is performed by measuring lip activity in the following manner: the lower region of consecutive face detections is aligned, we then measure the entropy of the pixel movement direction (optical flow) on that region. This method is described in [23]. If the average lip activity over the face track $> \theta_2$, the speaker identity at that time is propagated to the current face.

6 Results and discussion

6.1 Corpus

We used the TV recording corpus from the *REPERE* Challenge [6]: news, political debate and talk-shows of 7 french programs (2 from BFMTV and 5 from LCP) aired in 2011. In these experiments, we use 2 subsets provided by the campaign: C_1 contains 135 videos for a total duration 48h (known as "phase1_train" in the campaign) and C_2 contains 25 videos for a duration of 13h (known as "phase0_test" or "phase1_dev"). The reference annotation covers 24h of C_1 and 3h of C_2. It consists of 8624 annotated frames for C_1 and 1107 for C_2 (about 1 every 10s). For each keyframe, annotations cover three modalities: text (overlaid text, person names in the text), speech (speaker identity, speech transcript and names in the transcript) and video (face outline, person name, occlusions and attributes). The annotated keyframes give a total amount of 9748 faces to be named for C_1 and 1150 for C_2. Statistics on the $C2$ show that 33.3% of keyframes contain more than one face. People can appear in both C_1 and C_2. In our experiments, we used the C_1 as development corpus (for model selection and parameter tuning) and C_2 as the held-out test corpus.

Table 1 shows statistics on C_1 that shed light on potential inter-modality propagations. If a name is overlaid, there is a 98.5% chance that the correspond face is visible on the same keyframe and 80.4% of chances that it identifies the current speaker. This validates our direct propagation strategy for OPN→Face and

Modality A	**Modality** B	$A \Rightarrow B$	$B \Rightarrow A$
Overlaid name	Face	98.5	10.0
Face	Speaker	42.1	63.4
Overlaid name	Speaker	80.4	12.3

TABLE **1. Co-occurrence statistics on reference keyframes in C_1 in % of keyframes.**

OPN→Speaker. If we exclude one of the programs containing split screens, overlaid names apply to 90% of speakers. Finaly, 42.1% of the time a face is visible the person is also speaking, while 64.4% of the time the speaker is also visible on screen. This justifies our intuition that the speaker modality is a potential identity propagation channel.

6.2 Evaluation protocol

The usual metrics precision (P), recall (R) and F-score (F) are used in addition to *EGER* (Estimated Global Error Rate), the official metric of *REPERE* defined as follows:

$$\text{EGER} = \frac{\#\text{inserted} + \#\text{missed} + \#\text{confused}}{\#\text{references}} \quad (1)$$

where #references is the number of named people in the reference keyframes, #inserted, #missed and #confused are the number of errors in each category. A lower EGER means better performance.

6.3 Experiments

We have defined in Section 5 a set of propagation rules. We combine them in a sequential way. In the following, ruleA $\oplus$ ruleB means : apply ruleA then apply ruleB on remaining unnamed utterances. For speaker naming, the process is the following: OPN→Speaker$\oplus$UPN→Speaker $\oplus$ Speaker→Speaker.

For face naming, different combination rules are evaluated:

1. *Direct Face* (DF) : OPN→Face $\oplus$ UPN→Face.
2. *DF+Clu* (baseline): OPN→Face $\oplus$ UPN→Face $\oplus$ Face→Face (Cluster).
3. *DF + Sim* : OPN→Face $\oplus$ UPN→Face $\oplus$ Face→Face (Similarities).
4. *DF+Clu+Lip* : OPN→Face $\oplus$ UPN→Face $\oplus$ Face→Face (Clu)$\oplus$ Speaker→Face
5. *DF+Sim+Lip* : OPN→Face $\oplus$ UPN→Face $\oplus$ Face→Face (Sim) $\oplus$ Speaker→Face.

6.4 Results and discussion

Tables 2 and 3 summarize the performance of the unsupervised face identification systems in the C_1 and C_2. As expected, face identification based only on local propagation of the detected names ($DirectFace$) is reliable (high precision) but not sufficient to identify all sequences (R= 20%). On the baseline system, we notice that even if the face-clustering method is basic, it preformed reasonable results (better *EGER* than [10] on the same data). The

System	**P**	**R**	**F**	**EGER**
DirectFace	**91.3**	20.4	33.4	79.8
DF+Clu	73.6	48.1	58.2	55.4
DF+Sim	86.5	35.1	49.9	66.0
DF+Clu+Lip	68.8	53.6	60.2	51.7
DF+Sim+Lip	76.0	**53.7**	**63.0**	**49.3**

TABLE **2. Performance of unsupervised face identification systems on C_1 for all shows.**

System	**P**	**R**	**F**	**EGER**
DirectFace	**93.5**	21.4	34.8	78.9
DF+Clu	71.3	47.3	56.9	55.9
DF+Sim	84.2	32.8	47.2	68.4
DF+Clu+Lip	65.1	**52.8**	58.3	**52.1**
DF+Sim+Lip	68.7	52.0	**59.2**	**52.1**

TABLE **3. Performance of unsupervised face identification systems on C_2 for all shows.**

indirect propagation of the reliable associations improved the recall (R= 47%) introducing some errors. The origin of errors can be faces not detected by our system, face-clustering or face identification errors. Missed faces (not detected) impacts directly the recall of our methods (on parliamentary debates in particular). However, the face detector evaluated on the C_1 for all shows obtained 80% of precision with R=73% (recall of the oracle propagation system). The degradation of the precision is less important on $DirectFace + Sim$ explained by the fact that unsupervised training models are more reliable than our face-clustering, allowing the system to select similar faces with high confidence.

Both systems based on speaker identity propagation DF+Clu+Lip and DF+Sim+Lip performed an improvement of the EGER measure. Propagating the speaker identity improves the recall with an important degradation on the precision introduced by errors. Those errors can come from speaker identification errors or lip activity detection. The speaker identification system has a precision of 69% a recall of 66% on C_2. The talking face detector is based on the hypothesis that a lip movement means a talking face presence. This is not always verified on the REPERE corpus. In particular, on debate programs where multiple speakers appear on the screen, propagating the speaker identity degraded the EGER (+2.6%). In addition, this hypothesis dos not take into account report sequences with voice over.

Table 4 show the origin of the face identification provided by the system $DF + Sim + Spk$. % Test is the proportion of test utterances named with the application of the corresponding rule. % correct is the proportion of correct among this subset of test utterances. The correct rate of the local propagation of the OPN (OPN→Face) is 93.5%. This confirm the reliability of the local associations in TV content. The indirect propagation based on similarity measure (OPN→Face $\oplus$ Face→Face) performed 70.3% of correct identification (63% on the clustering based one). Concerning the 3 combination of rules that propagate the speaker identity, the system performed 76.5% of precision when the speaker-name

Identity origins	% Test	% Correct
OPN→Face	30.1	93.5
UPN→Face	0.6	20.0
OPN→Face ⊕ Face→Face	20.6	70.3
OPN→Spk ⊕ Spk→Face	11.2	76.5
OPN→Spk ⊕ Spk→Spk ⊕ Spk→Face	18.4	62.7
UPN→Spk ⊕ Spk→Spk ⊕ Spk→Face	19.0	28.7

TABLE **4. Error analysis of $DF + Sim + Spk$ on C_2 for all shows.**

is obtained locally (OPN→Spk ⊕ Spk→Face) and 62.7% when it is obtained by speaker clustering (OPN→Spk ⊕ Spk→Spk ⊕ Spk→Face). Propagating the UPN does not seems efficient in our systems (only 28.7% correct for the UPN→Spk ⊕ Spk→Spk ⊕ Spk→Face).

7 Conclusion and perspectives

Unsupervised identification of faces in TV-Content is a challenging problem. In this paper, we propose to use the speaker information in addition to the face, speech and OCR in order to improve unsupervised face identification. Names are detected from the speech and OCR and locally propagated to speakers and faces (direct identification). Then, the identities are propagated using tow speaker and face diarization methods (indirect identification). Then, speaker identities are propagated to faces when a talking face is detected. Results show improvement of the face identification in the REPERE corpus ($+4.8\%$ of the F measure in c_1 and $+2.3$ in c_2 compared to the baseline). Using The speaker identity to identify faces in TV-Content seems very promising. However, our methods showed limits on the improvement because of missed faces and talking-face detection errors. A way of improvement is to add other sources that may identify face (detecting visual concepts as report, dialogues, journalists, etc). Promising results have been observed by applying show-dependent rules. Also, scene detection and 3D modelling can be used to visualize the position of people on the camera (even if the face is not detected). Finally, we want to investigate general multi-modal identity propagation.

In the proposed method, the propagation rules are combined sequentiality. In future work, we want to study methods that generalize the multi-modal identity propagation in order to make it more interactive between modalities.

Références

[1] S. E. Tranter. Who really spoke when? finding speaker turns and identities in broadcast news audio. *ICASSP*, 2006.

[2] Feifan Liu and Yang Liu. Identification of soundbite and its speaker name using transcripts of broadcast news speech. *ACM*, 2010.

[3] Johann Poignant, Hervé Bredin, Viet-Bac Le, Laurent Besacier, Claude Barras, and Georges Quénot. Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. *Interspeech*, 2012.

[4] Timothee Cour, Chris Jordan, and Ben Taskar. Learning from ambiguously labeled images. *CVPR*, 2009.

[5] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image Vision Comput.*, 2009.

[6] Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carré, Aude Giraudel, and Philippe Joly. A presentation of the repere challenge. In *CBMI*, 2012.

[7] Shin'ichi Satoh and Takeo Kanade. Name-it: Association of face and name in video. *CVPR*, 1997.

[8] Derya Ozkan and Pynar Duygulu. A graph based approach for naming faces in news photos. *CVPR*, 2006.

[9] Chunxi Liu, Shuqiang Jiang, and Qingming Huang. Naming faces in broadcast news video by image google. *ACM*, 2008.

[10] H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V.B. Le, T. Napoléon, G. Hua, C. Barras, S. Rosset, L. Besacier, et al. Fusion of speech, faces and text for person identification in tv broadcast. *IFCVCR*, 2012.

[11] Manolis Delakis and Christophe Garcia. Text detection with convolutional neural networks. *VISAPP*, 2008.

[12] G. Linarès, D. Massonié, P. Nocera, and C. Lévy. The lia speech recognition system : from 10xrt to 1xrt. *Text, Speech and Dialogue : 10th International Conference*, 2007.

[13] L. L. Mangu. Finding consensus in speech recognition. *Computer Speech and Language*, 2000.

[14] C. Barras, X. Zhu, S. Meignier, and J-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[15] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2002.

[16] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 2004.

[17] R. Auguste, A. Aissaoui, J. Martinet, and C. Djeraba. Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés. *CORESA*, 2012.

[18] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Automatic face naming with caption-based supervision. *CVPR*, 2008.

[19] Ron Bekkerman and Jiwoon Jeon. Multi-modal clustering for multimedia collections. *CVPR*, 2007.

[20] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. *CVPR*, 2008.

[21] Benoit Favre Geraldine Damnati Frederic Bechet. Detecting Person Presence in TV Shows with Linguistic and Structural Features. In *ICASSP, Kyoto (Japan)*, 2012.

[22] Vincent Jousse, Simon Petit-Renaud, Sylvain Meignier, Yannick Esteve, and Christine Jacquin. Automatic named identification of speakers using diarization and asr systems. *ICASSP*, 2009.

[23] Meriem Bendris, Delphine Charlet, and Gérard Chollet. Talking faces indexing in TV-Content. *CBMI*, 2010.