Scene understanding for identifying persons in TV shows: beyond face authentication

Mickael Rouvier, Benoit Favre, Meriem Bendris
Aix Marseille Université, CNRS, LIF UMR 7279
firstname.lastname@lif.univ-mrs.fr

Delphine Charlet, Geraldine Damnati
OrangeLabs
firstname.lastname@orange.fr

***Abstract*—Our goal is to automatically identify people in TV news and debates without any predefined dictionary of people. In this paper, we focus on the problem of person identification beyond face authentication in order to improve the identification results and not only where the face is detectable. We propose to use automatic scene analysis as features for people identification. We exploit two features: scene classification (studio and report) and camera identification. Then, people are identified by propagation strategies of overlaid names (OCR results) and speakers to scene classes and specific camera shots. Experiments performed on the REPERE corpus show improvement of face identification using scene understanding features (+13.9% of F-measure compared to the baseline).**

I. Introduction

The proliferation of multimedia content makes necessary the development of technologies that facilitate browsing and searching through these data. Person indexing in videos can help users locate specific sequences featuring a given person. Classical approaches for person identification rely on face detection followed by face authentication from a dictionary of face models trained beforehand.

Unfortunately, collecting and maintaining those models is very expensive due to the number of people that may appear in TV content. In addition, the variety of conditions encountered, with far, occluded or back-facing heads, might harm the performance of face detection and identification. Most multi-modal person identification approaches that don't rely on prior biometric models are based on identity detection and propagation strategies. The identity can be detected from speech [1], overlaid names (using OCR techniques [2]) or subtitles when available [3]. The detected identities are propagated to visual clusters (comparing faces [2], or clothing [3]). However those methods are affected by the same challenges as face identification.

In this paper, we focus on using contextual information for identifying people in TV broadcast news and talk shows. In particular, we try to take advantage of rudimentary understanding of scenes in order to supplement face identification with extra constraints on who might be shown on screen. We propose two sources of scene-level cues:

- Scene classification which aims at detecting coherent segments of studio or field report footage. For instance it is unlikely that the same person appears in the studio and reports during the same show.

978-1-4799-3990-9/14/$31.00

Fig. 1. Face identification is ineffective when dealing with back-facing, partially hidden or cropped heads (image from the REPERE corpus).

- Camera identification which uses background features in order to pinpoint which angle of a recurring studio is being shot and infer from its layout the persons being shown. This takes the assumption that same cameras are used cross different episodes of the same show (verified in all shows of the corpus REPERE).

Those two sources of information are leveraged in the context of a multi-modal person identification system which propagates identities from OCR results with overlaid names and recognized speaker identities.

After presenting the related work (Section II), we briefly describe the REPERE corpus on which experiments are carried, and discuss the motivation for performing scene understanding (Section III). Then, we present the extraction of scene-level descriptors and how they are included in our multi-modal person identification system (Sections IV and V). Finally, we present a set of experiments run on the REPERE corpus and discuss them (Section VI).

II. Related work

In [4] the authors proposed to identify characters in TV-series by modelling each person by a Markov Random Field (MRF) integrating face, speech and body features, which showed promising results. However, the number of persons must be fixed *a priori*.

A large variety of approaches have been proposed for face authentication. Most of them rely on a holistic representation of the face and make use of subspace or manifolds, such as Kernel PCA Plus LDA [5] and Local Region PCA [6].

One of the most challenging problem in face authentication is the problem of session variability (changes in illumination, environment, expression, pose, etc), which cause mismatch between images of the same person. Different methods have been proposed for addressing this problem like modeling inter-session variability [7] or probabilistic linear discriminant analysis [8].

In the literature, there are several works that study scene analysis in sports videos, CCTV (Closed-Circuit Television) and news programs [9], [10], [11]. In [10], authors noticed 13 categories of shots in TV-news (interview, anchor, commercial, etc) and trained a decision tree to classify shots, then a Hidden Markov Model (HMM) to correct classification errors. Features used were shot durations, motion measurements, text and face detections. In [11] authors proposed to identify anchor shots using face position, size and clothes. In [9], anchor shots are detected using rules on motion estimation, shot duration and text captions.

III. Motivation

This work was conducted as part of the REPERE evaluation campaign [12]. The main purpose of the challenge is to identify people in TV news and debates. Targets are both television professionals (journalists, anchors, etc) and guests (experts, politicians celebrities). The test corpus consists in 7 hours of French TV from BFM and LCP channels. Persons are annotated on keyframes approximately every 10 seconds. The task focuses on *person identification*, a superset of face identification (silhouette, back-facing identification, etc). Figure 2 shows the different sources of people identification on the REPERE corpus.

Fig. 2. Multiple sources of people identification on the REPERE corpus.

Table I presents the percentage of identifications that have to be performed in difficult conditions such as back-facing or small heads. These conditions are challenging because (1) no face feature is visible, (2) the size of the head is smaller than that of the window used for feature extraction for face identification. For instance, [7] reported that the error rate for face id doubles on small faces. In the REPERE corpus about 10% of persons cannot be identified because of these factors.

Next comes the problem of biometric model dictionaries. Only about 50% of the person occurrences on the test corpus can be identified using models from the REPERE training corpus (assuming perfect identification if a model is available).

TABLE I. Percentage of person occurrences in keyframes that are either back-facing the camera (only their hair is visible) or for which the head is small ($< 2000px^2$).

Show	%Back-facing	%Small head
BFMStory	4.56	9.86
CaVousRegarde	2.52	16.80
LCPInfo	17.11	11.14
EntreLesLignes	12.45	22.28
PileEtFace	9.26	17.13
Total	9.30	14.24

Table II details this oracle result on a show by show basis. This table also shows the maximum performance of a system that would garner OCR results to associate overlaid names with persons using face clustering to propagate the identities to detected faces without corresponding overlaid text. Even when both prior models and overlaid names are accounted, for the performance ceiling is a recall of 75%. In addition, if the face id system works perfectly on small heads, it can achieve 86% of identifications with OCR and this ramps up to 95% if it can identify back-facing heads as well. The 5% remaining persons are neither present in training data nor identified by overlaid text.

TABLE II. Percentage of person occurrences in keyframes which can be identified with face id, overlaid names or both. Note that these figures already account for back-facing and small heads.

Show	%Face id	%OCR	%Both
BFMStory	33.72	77.68	79.71
CaVousRegarde	65.54	30.25	70.58
LCPInfo	55.29	70.24	70.38
EntreLesLignes	61.37	70.05	70.28
PileEtFace	74.85	75.56	75.56
Total	51.72	72.93	75.00

This study of the corpus shows that we cannot only rely on OCR results and face identification for successfully identifying people in the REPERE challenge. We will now show how to take advantage of scene understanding in order to infer person identities in most challenging conditions.

IV. Scene understanding

Scene understanding involves describing automatically what is happening in the scene (events, conditions, location, etc). In this paper, we take advantage of automatic scene analysis in order to predict the identity and the number of people in specific scenes and propagate this information to other scenes. We propose to extract two pieces of information from scenes: scene segmentation and categorization, and camera identification.

A. Scene segmentation and classification

In TV broadcast news, scenes are characterized by coherent sequences of shots of the same location or on the same topic. We first perform shot-level classification and then aggregate consecutive shots in order to form scenes.

Fig. 3. Example of scene segmentation and classification on the REPERE corpus.

We have manually annotated about 50k shot-central frames (middle frame of each shot) from the training corpus with the following categories: studio, field report, static picture, computer-generated graphics, other. For predicting labels at test time, a show-specific linear classifier is trained to predict a label for each shot-central frame. We extract the histogram of oriented gradients (HOG features) on a resized frame (128×64), stacked with 16×8 RGB features, and feed them to liblinear [13].

Using shot-level classification results, we segment each show into sequences of studio and report shots called scenes. The basic principle is that a report is the longest sequence of consecutive shots with a majority of shot-central frames labelled as report. In particular, all consecutive report shots are concatenated. Then, reports separated by less than a given interval are merged (< 3s). Finally, all report sequences longer than a certain duration (> 10s) constitute report scenes, all other sequences are considered as studio scenes. Figure 3 shows an example of scene segmentation and classification on the REPERE corpus.

B. Cameras Identification

A TV broadcast is produced using several cameras where each one is directed towards a specific view (anchor, guests, journalists, etc). For some broadcasts where the number of people is fixed, knowing which camera is used can help deduce the identity of the filmed people without knowing their exact location. In our experiments, we have annotated different types of cameras and the role of the persons being shot. For instance: Camera_1 is "Anchor, Guest 1, Guest 2", Camera_2 is "Guest 2, Guest 3", etc. Then, given a new video frame, we use a linear SVM classifier [13] to identify the camera. The input features are RGB color histograms extracted from resized frames (127×100).

Table III shows the number of cameras annotated per show. Notice that we did not perform camera identification on BFMStory videos due to the varying number of guests appearing in the show.

TABLE III. NUMBER OF CAMERAS PER SHOW. NOTE THAT WE DIDN'T APPLY THIS STRATEGY FOR BFMSTORY.

Shows	Nb Cameras
LCPInfo	12
LCP CaVousRegarde	28
LCP EntreLesLignes	18
LCP PileEtFace	12

V. ARCHITECTURE OF THE SYSTEM

This section first details the baseline person identification system which uses OCR, speaker id and face clustering for person identification and then we explain how scene-level evidence can be integrated.

A. Overlaid Person Name

The Overlaid Person Name (OPN) Recognition module corresponds to the succession of three sub-modules: Text Recognition, Person Name spotting and Person Name recognition. Text Recognition is performed with the Video OCR technology developed at Orange Labs based on a Character Recognition Convolutional Neural Network [14]. Amongst overlaid text boxes that can be detected in a video, we need to distinguish OPNs (that present a person on screen) from other texts (such as titles, locations, time, etc). To this purpose, a rule-based detection module has been designed relying on relative box position, character size and the number of words. Even if the text recognition module performs well in terms of Character Error Rate, we need to identify Person Names, and a single character miss-recognized can be harmful for the whole process. To overcome these OCR errors we use large name lists as a language model to correct the OCR hypotheses. The lists are obtained through the combination of general static list coming from knowledge databases such as Wikipedia and dynamic ones that are specifically built from the web according to the date of the show.

B. Shot boundary detection

Shot boundary detection is performed by computing the average Manhattan distance between the HSV histograms of 4 frames before the boundary and 4 frames after the boundary. An adaptive threshold is computed as two times the median boundary score over a 72 frame window. Since most cuts in the studied corpus are hard cuts and fast cross-fading, this simple algorithm performs well.

C. Speaker ID

We adopt a hierarchical combination of OPN-based speaker identification and acoustic-based speaker identification, both methods using speaker clustering. The whole process is explained in details in [15], replacing GMMs with better performing I-Vectors [16]. It basically relies on the assumption that an OPN detected in the video corresponds to the current speaker. In fact, on the REPERE corpus, 80.4% of the annotated OPNs appear with their corresponding speakers. Thus,

the process of multi-modal speaker identification is performed following those rules:

1) Acoustic speaker ID if the confidence score is sufficient.
2) OPN that maximizes the temporal overlap with the speaker turn.
3) OPN that maximizes the temporal overlap with the speaker cluster.

D. Face detection and clustering

In TV-broadcast, clustering faces is difficult because of the variability of face appearance. Assuming that within a single show there is a bijection between people and their clothes, we chose to cluster facetracks using the signature of clothes colors. First, face detection is performed using OpenCV's frontal and profile cascade detectors on every frame of the video. Then, the clothing area is deduced by taking the rectangle under the detected face proportional to its size ($\times$2.5 face width and $\times$1.5 face height). We removed the fixed information banner that appears on BFMTV shows from the search region. Then, the HSV colour histogram determines the features vector and a cosine-based distance is used to measure the similarity between facetracks. Figure 4 shows an example of clothing regions for detected faces.

Fig. 4. Clothes regions detection

E. Multi-modal face identification

The baseline algorithm associates names to facetracks by combining speaker identification, OPNs, and similarity between facetrakcs (described in details in [17]). The algorithm is based on two assumptions: first, an OPN corresponds to the person who is visible at the same time, which is correct in 98.5% of cases in the REPERE corpus. The second assumption is that when someone is speaking, he/she is visible, which is true in 80.4% of the time. The process follows 3 steps:

- identify facetracks with the OPN that maximizes temporal overlap. In case of multiple facetracks having similar overlap, a disambiguation process is used to attribute the OPN to one of the facetracks (see [17]). The set of facetracks attributed to the OPN n is called $\hat{g}_n$.
- for each unnamed facetrack f, a distance D is defined between f and $\hat{g}_n$. The name $\hat{n}(f)$ given to facetrack f is the name for which the distance is minimal, if that distance does not exceed a fixed threshold θ.

$$D(g,f) = \frac{1}{|g|}\sum_{f_i \in g} d(f,f_i)$$

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

- identify the remaining unnamed facetracks with the speaker identity that maximizes temporal overlap each facetrack.

The motivation of step 2 in the process is that the facetracks named thanks to a direct temporal overlap with OPN are probably correctly named, and thus, can be considered as model in an unsupervised open-set face identification paradigm. Face identification in TV shows is a difficult task, and is relatively limited in our system since it is a simple similarity measure between clothes. Thus, we propose to exploit scene understanding to improve the face identification process of step 2.

1) Scene constrained identification:

In TV broadcasts, different people appear in different environments, making the identity propagation very hard. Within a type of scene, the variability and occurrences of a person's appearance is significantly reduced and it is also more probable to detect another occurrence of a person already named with an OPN in the same scene. We propose to constrain the set of potential models $\hat{g}_n$ for each facetrack f to $\hat{g}_n(Scene(f))$, the set of potential models that belong to the same scene type as f. Hence, it is proposed to modify step 2 in the following way:

$$\hat{n}(f) = \begin{cases} \text{argmin}_{n\in N}\, D(\hat{g}_n(\text{Scene}(f)), f) \\ \quad \text{if } D(\hat{g}_n(\text{Scene}(f)), f) < \theta \\ \emptyset \text{ otherwise} \end{cases}$$

In our experiments, we consider as distinct studio and reports scenes. For the studio scenes, we can either regroup them all or consider them as distinct scenes. The optimal choice depends on the type of the show: BFMStory is a news talk-show where different studio sequences imply that different people are present (different interviewees, anchor, etc) whereas the same people are visible throughout all studio sequences for LCPInfo.

2) Camera-Id-constrained person identification:

In studio scenes of the type of shows we are processing, protagonists are identified within close-ups with overlaid names (OPNs). Then, the idea is to propagate this information to other views using the layout of the set and the relation between cameras. Unfortunately, it is very difficult to distinguish a close-up camera from another close-up camera, therefore we can't identify with certainty the position of the guest around the central table of the set. To address this problem, we gather all close-up shots of the show and cluster them with K-means by setting K to the known number of protagonists (using RGB histogram distance). Names are then propagated to views with more than one person by splitting the view in

TABLE IV. PRECISION (P), RECALL (R) AND F-MEASURE (F) PER SHOW ON THE REPERE CORPUS.

Shows	Baseline			Scene-based			Camera-ID		
	P	R	**F**	P	R	**F**	P	R	**F**
BFMStory	72.7	56.1	**63.4**	77.9	56.0	**65.2**	77.9	56.0	**65.2**
LCPInfo	60.0	47.2	**52.8**	51.1	60.8	**55.9**	74.0	67.5	**70.6**
CaVousRegarde	73.6	59.3	**65.7**	68.2	53.1	**59.7**	89.7	77.0	**82.9**
EntreLesLignes	83.4	39.6	**53.6**	83.4	39.6	**53.6**	91.3	74.8	**82.2**
PileEtFace	92.6	60.2	**73.0**	92.6	60.2	**73.0**	91.2	88.9	**90.1**
All	75.3	51.7	**61.3**	80.7	52.2	**63.4**	83.3	68.7	**75.3**

regions and comparing each region with the close-up models. This approach is extremely effective in this kind of closed-set identification in studio but it cannot be applied to report shows.

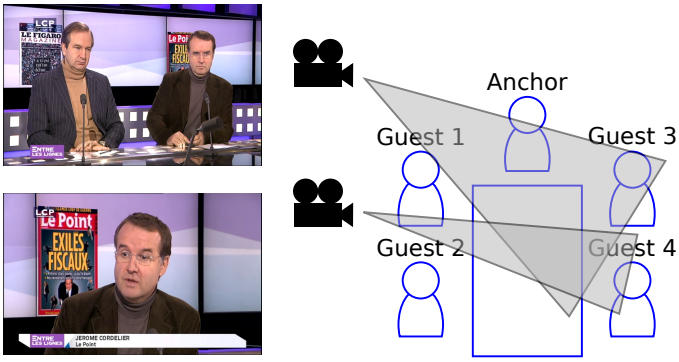

Fig. 5. Given an identified camera, names can be propagated to cameras known to film the same persons.

VI. EXPERIMENTS

To evaluate the identification system, 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 the *REPERE* challenge, defined as follows:

$$\text{EGER} = \frac{\#\text{inserted} + \#\text{missed} + \#\text{confused}}{\#\text{references}}$$

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.

We performed experiments on the phase2-test *REPERE* corpus using different features of scene analysis. The following variants are evaluated:

- **Baseline:** OPNs propagation to visual clusters and speakers propagation to unknowns.
- **Scene-based:** *Baseline* system and use the scene-constrained person identification.
- **Camera-ID:** OPNs propagation to camera clusters. Note that we didn't perform camera identification on BFMStory shows due to the varying number of guests appearing in the show.

Tables IV and V show the results obtained by our identification systems based on scene understanding. The *Baseline* system obtains an EGER of 42.6% with 51.7% in recall explained by the face detector misses. Improvements of the EGER, Precision and Recall are observed on the *Scene-based* system. No improvement is observed for the shows EntreLesLignes and PileEtFace, which is explained by the fact that those specific shows occur in studio in one long scene.

TABLE V. THE EGER PER SHOW ON THE REPERE CORPUS.

Shows	Baseline	Scene-based	Camera-ID
BFMStory	44.4	**42.1**	42.1
LCPInfo	55.9	48.4	**41.6**
CaVousRegarde	37.2	42.5	**27.4**
EntreLesLignes	43.3	43.3	**21.0**
PileEtFace	25.6	25.6	**8.9**
All	42.6	40.4	**30.9**

Using camera identification in addition to the chapter constrained propagation allowed us to improve significantly the EGER measure (-11.7% compared to the baseline). Major gains are obtained on the shows CaVousRegarde, EntreLesLignes and PileEtFace, because they are very well structured with separable camera views.

VII. CONCLUSION

In this paper we propose to use scene understanding in order to improve unsupervised people identification in TV-broadcast. News and debate programs are very well structured. we proposed to exploit this structure in order to improve face identification. The proposed method uses scene classification (studio/report) and camera identification features and restraint overlaid names detected from the OCR propagation to specific scenes and cameras. The scene understanding applied to unsupervised face identification obtained promising results (a gain of 13.9 absolute in F-measure and 11.7 in EGER). Limits of using scene analysis is the use of a priori knowledge on the structure of the TV-programs.The proposed method can identify faces without detecting them. This orthogonal approach could be combined with more traditional face identification.

In future work, we want to go further on automatic scene analysis. For instance, we can introduce topics detection, dialogues analysis, split up the studio class into interviews, news, debates, etc.

VIII. KNOWLEDGE

We would like to thank PERCOL consortium participants for providing their subsystem outputs. This work is funded by ANR under project PERCOL 2010-CORD-102-01.

REFERENCES

[1] Feifan Liu and Yang Liu, "Identification of soundbite and its speaker name using transcripts of broadcast news speech.," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, no. 1, 2010.

[2] Johann Poignant, Herv Bredin, Viet Bac Le, Laurent Besacier, Claude Barras, and Georges Qunot, "Unsupervised speaker identification using overlaid texts in tv broadcast.," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*. 2012, ISCA.

[3] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, apr 2009.

[4] Makarand Tapaswi, Martin Buml, and Rainer Stiefelhagen, ""knock! knock! who is it?" probabilistic person identification in tv-series.," in *Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 2658–2665, IEEE.

[5] Jian Yang, Alejandro F Frangi, Jing-yu Yang, David Zhang, and Zhong Jin, "Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[6] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O'Toole, David S Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *International Conference on Automatic Face & Gesture Recognition and Workshops*. IEEE, 2011, pp. 346–353.

[7] Roy Wallace, Mitchell McLaren, Christopher McCool, and Sebastien Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–8.

[8] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[9] Marco Bertini, Alberto Del Bimbo, and Pietro Pala, "Content-based indexing and retrieval of tv news.," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 503–516, 2001.

[10] Lekha Chaisorn and Tat seng Chua, "The segmentation and classification of story boundaries in news video," in *In IEEE International Conference on Multimedia and Expo*, 2002, vol. 216, pp. 95–109.

[11] Chien-Chuan Ko and Wen-Ming Xie, "News video segmentation and categorization techniques for content-demand browsing," *Congress on Image and Signal Processing*, vol. 2, pp. 530–534, 2008.

[12] Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carr, Aude Giraudel, and Philippe Joly, "A presentation of the repere challenge.," in *Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*. 2012, pp. 1–6, IEEE.

[13] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[14] Khaoula Elagouni, Christophe Garcia, Franck Mamalet, and Pascale Sbillot, "Text recognition in multimedia documents: A study of two neural-based OCRs using and avoiding character segmentation," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–13, jun 2013.

[15] Delphine Charlet, Corinne Fredouille, Graldine Damnati, and Grgory Senay, "Improving speaker identification in tv-shows using person name detection in overlaid text and speech.," in *Proceedings of the International Speech Communication Association (Interspeech)*. 2013, pp. 2778–2782, ISCA.

[16] A. Larcher, J.-F. Bonastre, and B. et al. Fauve, "Alize3.0 - open source toolkit for state-of-the-art speaker recognition," *Proceedings of the International Speech Communication Association (Interspeech)*, 2013.

[17] Meriem Bendris, Benoit Favre, Delphine Charlet, Géraldine Damnati, Rémi Auguste, Jean Martinet, and Gregory Senay, "Unsupervised Face Identification in TV Content using Audio-Visual Sources," in *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, Veszprém, Hongrie, 2013.