Segmentation et identification audiovisuelle de personnes dans des journaux télévisés

Paul Gay

Université du Maine

SEGMENTATION ET IDENTIFICATION AUDIOVISUELLE DE PERSONNES DANS DES JOURNAUX TÉLÉVISÉS.

THÈSE

présentée et soutenue publiquement le 25 mars 2015

pour l'obtention du

Doctorat de l'Université du Maine

(spécialité informatique)

par

GAY PAUL

Composition du jury

Rapporteurs :	M. Frédéric Jurie	Professeur des Universités	GREYC, Université de Caen
	Mme. Régine André-Obrecht	Professeur des Universités	IRIT, Université de Toulouse
Encadrant :	M. Sylvain Meignier,	Maitre de conférences	LIUM, Université du Maine
Examinateurs :	M. Laurent Besacier	Professeur des Universités	LIG, Université de Grenoble
	M. Jean-François Bonastre	Professeur des Universités	LIA, université d'Avignon
Directeurs de thèse :	M. Paul Deléglise	Professeur des Universités	LIUM, Université du Maine
	M. Jean-Marc Odobez	Maître d'Enseignement et de Recherche	IDIAP Research Institute EPFL

Résumé

Cette thèse traite de l'identification des locuteurs et des visages dans les journaux télévisés. L'identification est effectuée à partir des noms affichés à l'écran dans les cartouches qui servent couramment à annoncer les locuteurs. Puisque ces cartouches apparaissent parcimonieusement dans la vidéo, obtenir de bonnes performances d'identification demande une bonne qualité du regroupement audiovisuel des personnes. Par regroupement, on entend ici la tâche de détecter et regrouper tous les instants où une personne parle ou apparaît. Cependant les variabilités intra-personnes gênent ce regroupement. Dans la modalité audio, ces variabilités sont causées par la parole superposée et les bruits de fond. Dans la modalité vidéo, elles correspondent essentiellement à des variations de la pose des visages dans les scènes de plateaux avec, en plus, des variations de luminosité (notamment dans le cas des reportages). Dans cette thèse, nous proposons une modélisation du contexte de la vidéo est proposée afin d'optimiser le regroupement pour une meilleure identification. Dans un premier temps, un modèle basé sur les CRF est proposé afin d'effectuer le regroupement audiovisuel des personnes de manière jointe. Dans un second temps, un système d'identification est mis en place, basé sur la combinaison d'un CRF de nommage à l'échelle des classes, et du CRF développé précédemment pour le regroupement. En particulier, des informations de contexte extraites de l'arrière plan des images et des noms extraits des cartouches sont intégrées dans le CRF de regroupement. Ces éléments permettent d'améliorer le regroupement et d'obtenir des gains significatifs en identification dans les scènes de plateaux.

Mots-clés: Identification, journaux télévisés, Champ conditionnel aléatoire, regroupement audiovisuel des personnes

Abstract

This Phd thesis is about speaker and face identification in broadcast news. The identification is relying on the names automatically extracted from overlaid texts which are used to announce the speakers. Since those names appear sparsely in the video, identification performance depends on the diarization performance i.e. the capacity of detecting and clustering together all the moments when a given person appears or speaks. However, intra-person variability in the video signal make this task difficult. In the audio modality, this variability comes from overlap speech and background noise. For the video, it consists in head pose variations and lighting conditions (especially in report scenes). A context-aware model is proposed to optimize the diarization for a better identification. Firstly, a Conditional Random Field (CRF) model is proposed to perform the diarization jointly over the speech segments and the face tracks. Secondly, an identifcation system is designed. It is based on the combination of a naming CRF at cluster level and the diarization CRF. In particular, context information extracted from the image background and the names extracted from the overlaid texts are integrated in the diarization CRF at segment level. The use of those elements enable us to obtain better performances in diarization and identification, especially in studio scenes.

Keywords: Identification, broadcast news, Conditional Random Field, Audiovisual person diarization

Sommaire

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Identification dans les journaux télévisés : motivations et description de la problématique

L'objectif de cette thèse est de développer un programme permettant d'identifier automatiquement les personnes parlant et apparaissant à l'intérieur de vidéos de journaux télévisés. Le résultat d'un tel programme est illustré sur la figure 1.1. Pour une vidéo donnée, le programme doit être capable de fournir tous les instants correspondant à chaque locuteur et à chaque visage.

Depuis plusieurs dizaines d'années, des chercheurs mettent au point des programmes pour l'extraction automatique d'informations dans des vidéos afin de valoriser des archives audiovisuelles et des collections dont la taille et le nombre augmentent régulièrement. Les technologies impliquées incluent la transcription automatique de la parole, l'indexation du contenu, l'identification des personnes présentes et la catégorisation des types de scènes (reportage, publicité, scènes d'extérieur/intérieur). Ces dernières années, des projets comme QUAERO[1], AXES [van der Kreeft 2014] et NewsRover [Jou 2013] continuent d'améliorer et de combiner ces différents domaines. Ils proposent des systèmes aidant les utilisateurs à exploiter les archives

[1] http ://www.quaero.org

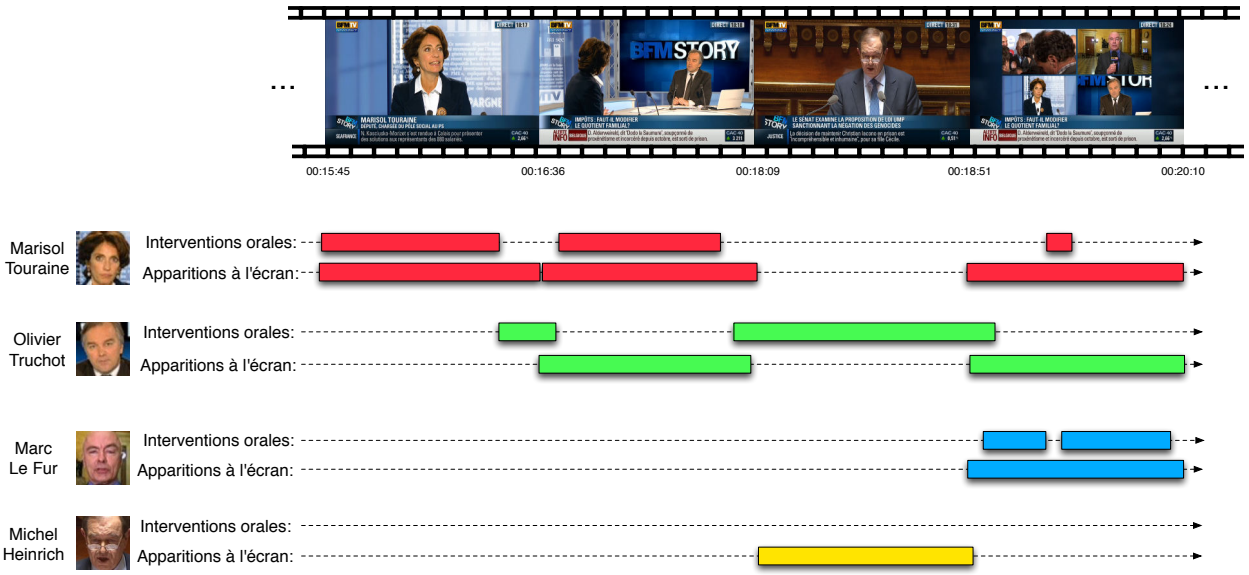


FIG. 1.1: Le haut de la figure représente une vidéo. Le bas de la figure représente la sortie attendue du système d'identification.

audiovisuelles pour répondre aux 5 questions décrivant les différentes facettes d'un sujet d'actualité : *Qui ? Où ? Quand ? Quoi ? Pourquoi ?* À cette fin, le projet GDELT [Kwak 2014] de l'entreprise Google, ambitionne de *suivre tous les médias du monde dans presque tous les coins de tous les pays [. . .] dans plus de 100 langues à chaque moment de chaque jour.*

Parmi les différentes problématiques liées à ces travaux, cette thèse traite de l'identification des personnes. Ce problème tient une place importante car, dans les sujets d'actualité, les idées et les thèmes peuvent souvent être vues par le prisme de leurs principaux acteurs. De plus, l'accès aux moments d'intervention de chaque personne constitue un moyen d'accès à la structure de la vidéo : une personne correspond généralement à une scène ou un sujet.

L'identification des locuteurs et des visages dans des vidéos peut être supervisée à partir de ressources externes. Est alors appris un modèle biométrique de la personne que l'on souhaite retrouver. Cependant, les ressources nécessaires à un tel apprentissage (exemples annotés manuellement) sont généralement coûteuses à obtenir ce qui limite le nombre de personnes identifiables avec cette stratégie. C'est pourquoi de nombreuses approches utilisent en complémentarité des sources de nommage déjà présentes dans la vidéo. Par exemple, elles extraient automatiquement les noms prononcés ou ceux présents dans les cartouches superposés à l'image qui sont utilisés pour annoncer les locuteurs.

Dans un cas comme dans l'autre, il est nécessaire de détecter les visages et les locuteurs, puis de regrouper les détections qui correspondent à la même personne. Dans le cadre de cette thèse, on considère que la détection est déjà effectuée. Nos travaux se concentrent sur les étapes de regroupement. La reconnaissance du locuteur et celle du visage sont des thématiques de recherche populaires depuis plusieurs dizaines d'années ([Pruzansky 1963, Bledsoe 1966]) mais il existe encore des facteurs de variabilité intra-personnes que même les représentations récentes ont du mal à compenser. La présence et l'importance de ces facteurs varient selon le type de vidéos ; par exemple selon que l'on traite des reportages en extérieur ou des débats en studio. La section suivante présente les corpus et les protocoles d'évaluation utilisés dans nos expériences. Cela permettra en outre d'illustrer la problématique et ses difficultés selon les situations.

1.2 Présentation du corpus et des métriques

Cette thèse s'est déroulée dans le cadre de la campagne d'évaluation REPERE (REconnaissance de PERsonnes des Emissions audiovisuelles) [Kahn 2012] qui a duré de 2011 à 2014. La tâche principale de cette campagne est précisément la reconnaissance multimédia de personnes dans des documents télévisuels. Ce cadre nous a donné les moyens d'effectuer les expériences

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIG. 1.2: Exemples des différentes situations présentes dans le corpus REPERE :

- images a, b : *Entre Les Lignes* et *Pile Ou Face* : débats politiques en plateau
- images c, d, e, f : *BFMStory* et *LCP Info* : journaux télévisés
- images g, h : *Top Questions* : séances de l'assemblée nationale
- image i : *Culture Et Vous* : magazine People

nécessaires à l'évaluation de nos systèmes. De plus, la participation de trois consortiums autorise la comparaison avec plusieurs approches, à l'exception de la tâche de regroupement en classes de visages où l'absence de soumissions nous oblige à utiliser un autre corpus pour nous comparer à l'état de l'art. La description du corpus présentée dans la section suivante illustre les difficultés de l'identification en fonction des différents types d'émissions. Ensuite, la présentation des métriques permettra de discuter de l'évaluation des systèmes d'identification.

1.2.1 Description des émissions composant le corpus

Le corpus REPERE est composé de journaux télévisés extraits des chaînes LCP et BFMTV. Les vidéos contiennent différents types de situations incluant des débats, des reportages et des interviews en plateau (voir Fig. 1.2).

Les débats sont caractérisés par une interaction importante entre les locuteurs avec des interruptions, de la parole spontanée et de la parole superposée rendant le regroupement plus difficile. Côté vidéo, les principales variations de l'apparence sont dues au fait que le visage d'une même personne peut apparaître sous différentes poses quand la personne se tourne vers les différents participants et différents points de vue si plusieurs caméras sont utilisées.

Les journaux télévisés sont des émissions au contenu plus varié. Ils incluent des interviews en plateau ou en duplex et des reportages avec des voix off. Du point de vue de la difficulté des traitements, les interviews en plateau s'apparentent aux débats. Certaines émissions comme *BFMStory* intègrent les flux de plusieurs caméras sur la même image ce qui augmente le nombre de visages et complique le regroupement audiovisuel entre locuteurs et visages (figure 1.2, image d). Les reportages peuvent contenir des conditions sonores gênant le regroupement en locuteur. Ainsi, un présentateur peut parler en voix off pendant le reportage et être présent à l'image plus tard sur un plateau. Les conditions visuelles ont également une plus grande variabilité dans les reportages. Cependant, l'impact sur le regroupement est limité car la plupart des personnes n'y apparaissent que dans une seule scène, voire sur un seul plan. Enfin, la présence de public peut contribuer à augmenter la confusion pour le regroupement et l'identification (figure 1.2, images c, d et g).

L'émission *TopQuestions* (figure 1.2, images g et h) est un cas à part et correspond aux enregistrements des questions posées au gouvernement par les membres de l'Assemblée Nationale. Elle est constituée de longues interventions les unes à la suite des autres. La présence des députés de l'hémicycle provoque un fond sonore (applaudissements, huées) et le public apparaît fréquemment avec le locuteur.

Enfin, l'émission *CultureEtVous* (figure 1.2, image i) se distingue par la plus grande variabilité de toutes les émissions du corpus tant au niveau audio que vidéo puisqu'elle multiplie les jingles, les fonds sonores, les extraits de films et de concerts et les plans en travelling.

Le tableau 1.1 donne la répartition des émissions dans le corpus par durée. Seules figurent les parties du corpus utilisées dans les expériences.

Annotations disponibles pour REPERE

Les annotations incluent la transcription manuelle ainsi que l'identité de chaque locuteur. Les visages sont également identifiés sur une image-clé par plan (ou une image-clé toutes les

Genre	DEV0	TEST0	DEV2	TEST2	Émission
Débats politiques	5h16 *(0h45)*	6h14 *(0h50)*	12h34 *(1h24)*	9h36 *(2h32)*	*ÇaVousRegarde PileEtFace EntreLesLignes*
Journaux télévisés	6h12 *(1h37)*	6h15 *(1h15)*	11h42 *(3h06)*	15h46 *(6h26)*	*BFMStory RuthElkrief LCPInfo LCPActu*
Questions au gouvernement	1h02*(0h23)*	0h50*(0h30)*	1h28 *(1h00)*	1h07 *(0h30)*	*TopQuestions*
Magazine people	1h56*(0h15)*	2h07*(0h15)*	2h57 *(0h30)*	10h38 *(0h48)*	*CultureEtVous*
Total	14h26*(3h00)*	15h26*(3h00)*	28h41 *(6h00)*	37h16 *(10h16)*	92h40 *(30h20)*

TAB. 1.1: Répartition en durée des données du *dry-run* (PHASE0) et de la phase finale (PHASE2) du corpus REPERE. Les durées correspondent à la taille totale des vidéos avec entre parenthèses la taille de la zone annotée.

FIG. 1.3: Exemple d'une image-clé illustrant l'annotation des images dans REPERE. Les visages détourés en bleu sont annotés dans les références. La tête dans le cercle rouge n'a pas été annotée car sa taille est inférieure au seuil. Les textes apparaissant à l'écran détourés par les cadres verts ont aussi été transcrits.

10 secondes pour une partie du corpus). Ces annotations comportent le nom de la personne et sa position dans l'image. Les instants d'apparition et de disparition de ce visage sont précisés. Les visages d'une taille inférieure à un seuil (une aire d'environ 10000 pixels, la taille du visage étant déterminée visuellement par l'annotateur) n'ont pas été annotés. Les textes apparaissant à l'écran ont aussi été transcrits, notamment celui des cartouches qui contiennent le nom du locuteur courant. Un exemple d'image annotée est donné sur la figure 1.3.

1.2.2 Métrique pour l'identification : l'EGER

La métrique officielle de la campagne pour évaluer les performances en identification est l'*Estimated Global Error Rate* (EGER) :

$$EGER = \frac{\#false + \#miss + \#conf}{\#total} \tag{1.1}$$

où $\#total$ est le nombre de personnes qui doivent être détectées, $\#conf$ le nombre de personnes auxquelles a été attribuée la mauvaise identité, $\#miss$ le nombre de personnes qui n'ont pas été détectées et $\#false$ le nombre de fausses alarmes, c'est à dire si le système détecte une personne qui n'est pas présente. Un extrait d'un fichier d'évaluation est donné ci-dessous.

```
BFMTV_BFMStory_2012-01-10_175800 frame: 2407.777
ref: 0 0 Arno_KLARSFELD Jacques_CHIRAC
hyp: 0 0 Arno_KLARSFELD
```

Arno Klarsfeld apparaît dans la référence et est bien détecté dans l'hypothèse. En revanche, Jacques Chirac est manqué. La valeur EGER correspondante à ce cas est de $\frac{1}{2}$. La valeur de l'EGER finale sera la moyenne des EGER sur toutes les images du corpus.

La métrique principale compte les détections des locuteurs et des visages. Sur une image, une personne qui parle et qui apparaît compte pour deux et doit être identifiée dans chaque modalité pour conserver un EGER égal à 0. Il est aussi possible de calculer un EGER séparément pour chaque modalité audio et vidéo ou de l'utiliser pour évaluer la détection des noms écrits.

Il faut aussi noter que les réponses évaluées correspondent à des listes de personnes pour chaque image-clé. Ainsi, il n'y a pas de correspondances spatiales au niveau des visages entre la référence et la sortie automatique. Le nombre de fausses alarmes correspond donc à l'excédent de réponses du système par rapport au nombre de personnes présentes dans la référence, les autres erreurs sont comptées comme des confusions.

La métrique EGER peut être discutée selon plusieurs points de vue.

– L'EGER évalue les réponses du système pour chaque image annotée. Pendant l'évaluation, les systèmes doivent fournir tous les instants où les personnes apparaissent car la position des images annotées est inconnue. Cette détection dense sous-entend un cadre applicatif particulier. Dans d'autres systèmes d'indexation, savoir si la personne est présente ou non dans la vidéo est suffisant.

– La présence d'inconnus (des personnes dont le visage est annoté mais dont l'identité n'a pas été trouvée par l'annotateur) est ignorée par la métrique EGER. Ainsi, les systèmes

ne sont pas pénalisés par la non-détection de ces personnes. De plus, nommer un inconnu ne sera pas compté comme une confusion mais comme une fausse alarme.

– L'EGER évalue l'identification des locuteurs par image-clé, c'est à dire de manière discrète. Cependant, ils sont annotés de manière continue. Il serait donc possible de tenir compte précisément de la taille des segments dans l'évaluation. Cependant, comme les visages sont annotés de manière discrète, le choix a été fait de conserver une évaluation discrète des locuteurs afin de garder la cohérence entre les différentes modalités.

1.2.3 Métriques pour le regroupement : DER, rappel, précision, F-mesure

Diarization Error Rate (DER)

La mesure usuelle pour évaluer le regroupement en locuteur est le *Diarization Error Rate* [NIST 2003] (DER). C'est aussi la métrique choisie pour évaluer le regroupement en locuteur dans le cadre de la campagne REPERE. Nous donnons sa définition ci-dessous :

$$DER = Confusion + Non\_détection + Fa \tag{1.2}$$

avec $Confusion$ correspondant à la durée relative attribuée au mauvais locuteur.

$$Confusion = \frac{\sum_{seg \in Segs} dur(seg) * (min(N_{Ref}(seg), N_{Sys}(seg)) - N_{Correct}(seg))}{\sum_{seg \in Segs} dur(seg).N_{ref}(seg)} \tag{1.3}$$

où le fichier audio est divisé en segments continus à chaque changement de locuteur et :

– $dur(seg)$ est la durée du segment seg

– $N_{Ref}(seg)$ le nombre de locuteurs actifs pendant le segment seg

– $N_{Sys}(seg)$ le nombre de locuteurs détectés par le système

– $N_{Correct}(seg)$ le nombre de locuteurs correctement détectés par le système. Le calcul de ce terme nécessite d'apparier les classes des sorties automatiques avec les locuteurs de la référence. Cet appariement est calculé de façon à minimiser le DER final. Dans la pratique, l'algorithme hongrois peut être utilisé en maximisant la durée de co-occurrence [Galibert 2013].

Les autres parties du DER correspondent aux zones de parole non détectées :

$$Non\_détection = \frac{\sum_{seg \in Segs} dur(seg) * max(0, (N_{Ref}(seg) - N_{Sys}(seg)))}{\sum_{seg \in Segs} dur(seg).N_{ref}(seg)} \tag{1.4}$$

et aux zones détectées à tort comme de la parole, c'est à dire les fausses alarmes :

$$Fa = \frac{\sum_{seg \in Segs} dur(seg) * max(0, (N_{Sys}(seg) - N_{Ref}(seg)))}{\sum_{seg \in Segs} dur(seg).N_{ref}(seg)} \tag{1.5}$$

Nous utilisons aussi cette mesure pour évaluer le regroupement en classes de visages et le regroupement audiovisuel des personnes. Dans le cas des visages, chaque segment vidéo correspond à une détection avec une durée égale à l'intervalle entre deux images. Toutefois, l'utilisation du DER n'est pas commune dans la communauté image. Nous verrons que les travaux de l'état de l'art utilisent d'autres mesures pour évaluer le regroupement comme la précision ou l'effort nécessaire pour corriger les sorties du système en terme de nombre de clics de souris. Pour le cas des personnes, une classe rassemble les segments audios et vidéos. Le calcul du DER s'effectue alors sur tous ces segments.

Précision, Rappel et F-mesure pour le regroupement

Comme précédemment, l'utilisation des mesures va d'abord être donnée pour le regroupement en locuteur avant d'être généralisée à la modalité vidéo et au cas multimodal. Comme pour le DER, il est supposé que le fichier audio est divisé en segments continus. Ensuite, un appariement est également effectué entre les classes et les locuteurs de la référence. À partir de ces correspondances, les mesures suivantes sont créées :

Précision et **Rappel** :

$$Pr = \frac{T^{correct}}{T^{Auto}}, \qquad Rap = \frac{T^{correct}}{T^{Ref}} \tag{1.6}$$

où $T^{correct}$ représente la durée totale des segments attribués à la bonne classe, T^{Auto} représente la durée totale des segments détectés par le système et T^{Ref} représente la durée totale des segments présents dans la référence.

La **F-mesure** résume les deux mesures précédentes :

$$F_m = \frac{2 \times Pr \times Rap}{Pr + Rap} \tag{1.7}$$

Comme pour le DER, ces trois dernières mesures peuvent aussi être utilisées pour évaluer le regroupement en classes de visages et le regroupement audiovisuel des personnes. Elles ont l'avantage d'être largement reconnues et sont donc souvent utilisées.

1.3 Contributions

Plusieurs approches utilisent le contexte dans lequel les personnes apparaissent en complémentarité des représentations monomodales. Par exemple, si l'on sait quel locuteur correspond à quel visage à l'écran, il est sans doute plus facile de les identifier de manière jointe. Autre exemple, si un visage est identifié dans un plan d'une scène de plateau télévisé, il est probable qu'il apparaîtra à nouveau sur d'autres plans de cette scène. Or, les journaux télévisés présentent une structure relativement codifiée avec un découpage en sujet, une catégorisation en scènes de plateau et en scènes de reportage et des rôles pour la plupart des intervenants. Il est donc possible d'utiliser ces informations. Les travaux présentés dans cette thèse proposent de tirer parti de cette structure afin d'améliorer l'identification dans les journaux télévisés. Les différents éléments de contexte déjà utilisés dans l'état de l'art sont étudiés et de nouveaux sont proposés.

La première contribution est un travail effectué pendant les premiers mois de la thèse sur la représentation des visages pour le regroupement vidéo. La représentation proposée combine les avantages de deux mesures de similarité complémentaires : (i) une comparaison directe de descripteurs locaux permettant de regrouper avec précision des visages provenant de scènes similaires (ii) une similarité fondée sur les rapports de vraisemblance par rapport à des modèles de mélanges de gaussiennes (GMM) qui permettent de modéliser la distribution de caractéristiques extraites de manière dense sur tout le visage. Cette deuxième modélisation permet de mieux gérer les variations d'apparence.

Notre deuxième contribution est un nouveau système pour le regroupement audiovisuel des personnes. Le problème est formulé à l'aide du cadre des Champs Conditionnels Aléatoires (désignés par la suite avec leur acronyme anglais : CRF). Le modèle proposé effectue conjointement l'association audiovisuelle et le regroupement dans chaque modalité. L'idée est que l'information d'association entre les voix et les visages peut servir à guider et à améliorer le regroupement dans chaque modalité. Cette stratégie a été utilisée par plusieurs méthodes de l'état de l'art, mais dans des contextes relativement différents ou en utilisant des systèmes qui ne bénéficient pas des dernières avancées de l'état de l'art. Nous appliquons notre modèle au cas des journaux télévisés en le comparant avec les derniers systèmes de regroupement monomodaux.

Ensuite, nous étudions l'identification à l'échelle des classes au moyen des noms extraits des cartouches en testant deux méthodes automatiques. Dans les deux cas, l'information de nommage vient des co-occurrences entre les classes et les cartouches. La première méthode est

fondée sur des règles. La seconde exploite le formalisme des CRF et inclut des relations entre les différentes classes afin d'améliorer la précision. La comparaison avec un oracle permet de constater que le potentiel d'identification des cartouches n'est pas complètement exploité par ces deux méthodes automatiques. Les erreurs de confusion dans les regroupements monomodaux expliquent une partie de cet écart entre l'oracle et le système automatique.

Afin d'améliorer l'identification, le modèle CRF proposé pour le regroupement est repris et enrichi avec des éléments de contexte afin de l'optimiser pour la tâche d'identification. Ces éléments incluent l'information de nommage extraite des cartouches et une caractérisation du type de scènes fondée sur la récurrence de l'arrière plan. Le contexte utilisé ici modélise principalement l'arrière-plan des visages afin de distinguer à l'échelle des segments les personnes identifiables par un cartouche par opposition aux anonymes présents dans la vidéo mais ne participant pas à l'émission. L'utilisation conjointe des deux CRF permet d'améliorer les performances par rapport à l'utilisation du CRF de nommage seul dans la majorité des émissions, en particulier pour les débats et les émissions en plateau.

1.4 Plan de la thèse

Les deux premiers chapitres de cette thèse concernent l'état de l'art. Le premier s'intéresse au regroupement audiovisuel des personnes. Il passe en revue les difficultés existantes pour cette tâche. Après une description des approches monomodales de regroupement, l'accent est mis sur l'utilisation du contexte. En particulier, il souligne l'intérêt des approches multimodales qui effectuent les regroupements des locuteurs et des visages de manière jointe.

Le deuxième chapitre traite de l'identification non supervisée des personnes. Il s'agit d'extraire les noms de la vidéo et de les associer à toutes les apparitions ou interventions des personnes correspondantes. Dans le contexte de la campagne REPERE, il a été montré que les noms extraits automatiquement des cartouches constituent la meilleure source de nommage. Ensuite, nous montrons que les tâches de regroupement et d'identification sont liées. La plupart des stratégies existantes tendent d'ailleurs à effectuer ces deux tâches de manière jointe.

Le chapitre 4 présente les travaux sur le regroupement en classes de visages et la nouvelle représentation visuelle proposée. Le chapitre 5 décrit le premier système CRF pour la tâche de regroupement audiovisuel des personnes. L'identification proprement dite est traitée dans le chapitre 6. Dans un premier temps, les méthodes identifiant les classes de visages et de locuteurs sont introduites. La dernière partie présente le système final combinant regroupement à l'échelle des segments et identification au niveau des classes. Ce système inclut les éléments décrits dans les chapitres précédents.

Première partie

État de l’art

Chapitre 2

Regroupement des visages et des locuteurs

Sommaire

Ce chapitre dresse l'état de l'art sur la segmentation et le regroupement audiovisuel des personnes. Le but de cette tâche est d'annoter une vidéo en fonction des personnes. Comme l'illustre la figure 2.1, l'annotation contient pour chaque personne les moments où elle parle et les moments où elle apparaît à l'écran. Il faut noter qu'il ne s'agit pas encore d'identifier les personnes avec un nom : les interventions de chaque personne sont regroupées dans une classe identifiée par un numéro. Cependant, cette tâche a souvent été étudiée pour elle même. En effet, de nombreuses applications peuvent tirer partie de ces annotations pour l'indexation de vidéos, la navigation ou l'aide à l'annotation manuelle.

Dans ce chapitre, l'accent est mis sur l'intérêt d'effectuer le regroupement audiovisuel des personnes de manière jointe en tirant parti au mieux des informations de l'audio et de la vidéo. Dans les sections 2.1 et 2.2, les tâches de regroupement en locuteur et de regroupement en classes de visages sont introduites séparément. Si, en audio, le contexte n'est que rarement

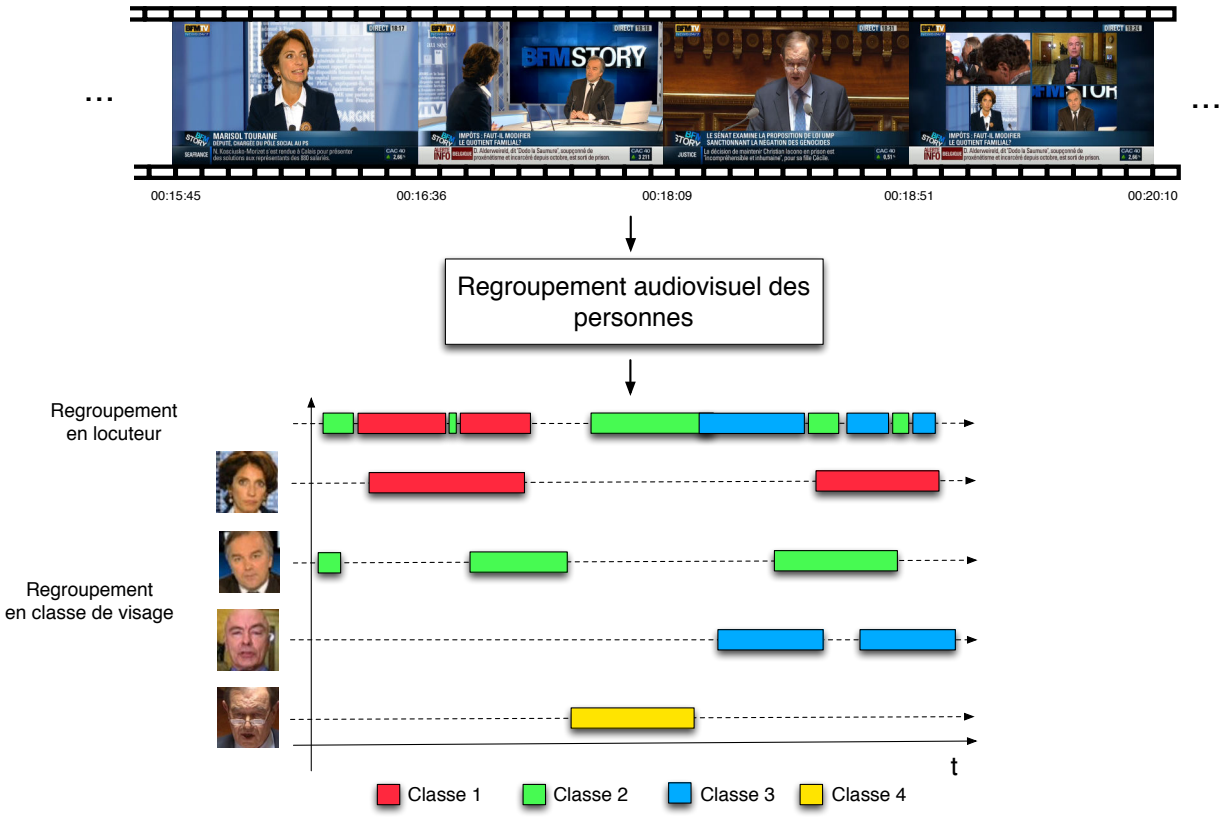


FIG. 2.1: Le haut de la figure représente la vidéo en entrée du système de regroupement : les images et l'enregistrement audio. Le bas représente la sortie attendue. Il doit être noté qu'il s'agit d'effectuer un regroupement en locuteur, un regroupement en classes de visages ET d'associer les locuteurs et les visages entre eux.

utilisé, nous verrons que son utilisation a été étudiée pour le regroupement en classes de visages avec l'exploitation de l'arrière-plan de l'image, la co-occurrence des visages entre eux, etc. Enfin, la section 2.3 présente des approches effectuant le regroupement audiovisuel des personnes de manière jointe. Nous expliquons comment peuvent être associés les locuteurs et les visages et quelles sont les techniques qui ont été imaginées afin d'utiliser l'association audiovisuelle pour améliorer le regroupement. À notre connaissance, cette dernière section présente les travaux les plus proches de notre contribution pour le regroupement audiovisuel des personnes, contribution présentée dans le chapitre 5.

2.1 Regroupement en locuteur monomodal

Dans cette section, la tache de regroupement en locuteur est présentée et le système qui sera utilisé dans nos expériences est détaillé. Ce système a été développé par le laboratoire LIUM pour la campagne d'évaluation REPERE. Il servira d'étalon pour mesurer l'apport de la multimodalité pour le regroupement audio.

Le regroupement en locuteur consiste à annoter tous les instants d'un enregistrement audio avec des étiquettes de classe pour répondre à la question *qui parle quand ?* Pour atteindre cet objectif, la modélisation du locuteur a donné lieu à de nombreux travaux et le regroupement en locuteur lui-même est un sujet de recherche depuis presque 25 ans [Gish 1991]. Ces travaux se sont attachés à rendre le regroupement robuste à des facteurs qui masquent les informations propres au locuteur comme les bruits de fond ou la parole superposée. Une autre difficulté rencontrée vient des tours de parole courts (d'une durée de quelques secondes) car les modèles statistiques utilisés disposent alors de peu de données pour être appris.

La présence et l'importance de ces facteurs de difficulté diffèrent en fonction du type d'enregistrement. Les approches se sont donc spécialisées par type de données : conversations téléphoniques, émissions de radio/télévision (cadre de cette thèse) et enregistrements de réunions. Les principales différences entre ces 3 situations concernent le nombre de locuteurs, la durée des enregistrements et le degré de spontanéité de la parole. Comparativement aux deux autres cas, les émissions de radio/télévision contiennent d'avantage de locuteurs (souvent plus de 10). Le silence, le bruit et la musique sont aussi plus nombreux. Le degré de spontanéité est variable. Il est faible dans le cas des documentaires et élevé dans les débats où il est du même ordre que celui des réunions.

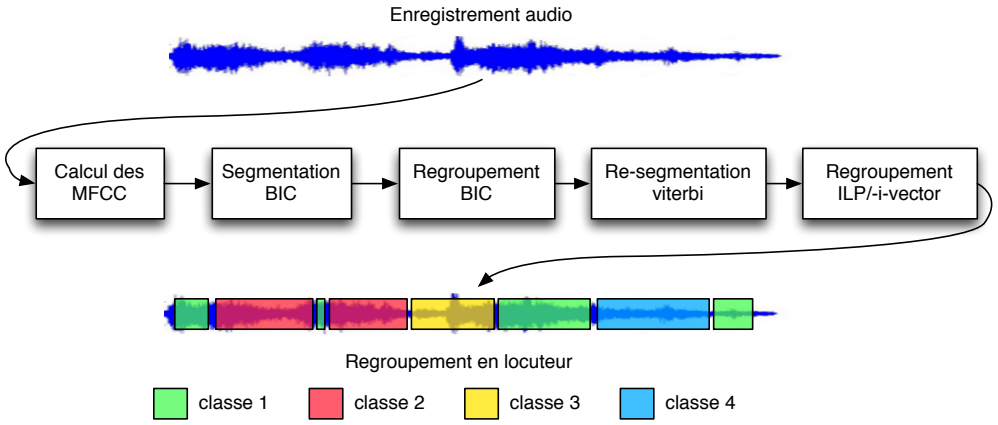


FIG. 2.2: Schéma énumérant les principales étapes du système de regroupement en locuteur détaillé dans [Rouvier 2013].

Le système présenté ici est celui de la bibliothèque *LIUM_SpkDiarization* [Rouvier 2013] et a obtenu des performances à l'état de l'art dans la campagne d'évaluation REPERE [Giraudel 2012]. Il est dédié aux émissions de télévision et de radio.

Ses différentes étapes sont énumérées dans la figure 2.2 et vont être à présent brièvement résumées.

1. **Calcul des MFCC :** les Mel-frequency cepstrum coefficients (MFCC) [Davis 1980] sont couramment utilisés en traitement de la parole et en reconnaissance du locuteur. Ils forment des vecteurs de caractéristiques extraits du signal sur des fenêtres glissantes dont la taille est de l'ordre de quelques millisecondes. Les différents modèles et distributions utilisés dans les étapes suivantes sont estimés dans l'espace de ces coefficients MFCC.

2. **Segmentation BIC :** dans cette première étape de segmentation, le signal est initialement découpé en segments d'au moins 2,5 secondes. Ensuite, deux segments consécutifs sont fusionnés s'ils présentent les mêmes caractéristiques acoustiques. La distance entre deux segments utilise le critère d'information bayésien [Chen 1998] (BIC). Elle consiste à tester si cette paire de segments est mieux modélisée par 2 distributions gaussiennes au lieu d'une seule.

3. **Regroupement BIC :** c'est une classification ascendante hiérarchique (CAH) où la distance entre chaque paire de classes est une distance BIC. Le modèle pour chaque classe est une gaussienne à matrice de covariance pleine. Ce regroupement est stoppé précocement afin d'obtenir des classes pures. C'est à dire qu'il faut que chaque classe ne contienne que les données d'un seul locuteur. À cause de cet arrêt précoce, il arrive qu'un locuteur soit divisé en plusieurs classes.

4. **Re-segmentation Viterbi :** un modèle *Hidden Markov Model* (HMM) est appliqué sur le signal audio où chaque état correspond à une classe. Cette étape permet d'affiner les frontières des segments issues du découpage initial de l'étape 2.

5. **Regroupement ILP/i-vector :** à ce stade, la quantité de données présente dans chaque classe est suffisamment élevée pour utiliser des modèles plus complexes et plus robustes que les gaussiennes utilisées dans l'étape 3. En particulier, le système décrit dans [Rouvier 2012] introduit l'utilisation des i-vectors pour le regroupement en locuteur. Dans cette représentation, chaque classe de locuteur est caractérisée par un vecteur de dimension réduite (entre 50 et 400). Un modèle de mélange de gaussiennes est tout d'abord estimé sur les données de chaque classe par adaptation d'un modèle générique. Ensuite, les moyennes des gaussiennes de ce mélange sont concaténées pour former un super-vecteur. Pour finir, ce super-vecteur est projeté dans un espace de dimension réduite optimisé suivant une technique probabiliste [Dehak 2011]. Le but est de ne conserver que la variabilité propre au locuteur et de retirer les variabilités dues à des facteurs externes (bruit de fond, canal...). Une fois obtenus, les i-vectors sont regroupés en se plaçant dans le cadre de la programmation linéaire en nombres entiers (ILP). Cet algorithme permet de trouver le nombre minimal de classes avec la contrainte d'une distance maximale entre le centre de la classe et ses éléments. La distance utilisée entre deux i-vectors est une distance de Mahalanobis apprise sur un corpus d'apprentissage.

Alors que les trois premières étapes de traitement sont généralement similaires d'un système à l'autre, la réalisation de la quatrième concentre l'essentiel des recherches actuelles. Dans la campagne REPERE, les deux autres consortiums ont utilisé un regroupement hiérarchique où chaque locuteur est modélisé par un GMM avec une distance fondée sur le *Cross Likelihood Ratio* (CLR) [Barras 2006]. Par ailleurs, il est surprenant de remarquer que les travaux de regroupement pour les journaux télévisés sont principalement issus de laboratoire français.

Les performances de ce système seront discutées plus amplement dans le chapitre 5 où il sera évalué sur les données REPERE. Il est cependant notable que, comme la plupart des systèmes orientés vers les émissions de radio/journaux, la parole superposée ne soit pas gérée car jugée peu fréquente et difficile à caractériser. De plus, ce système est purement monomodal. Le cadre de cette thèse rend la modalité vidéo disponible et les expériences montreront dans quels cas un système multimodal peut obtenir de meilleures performances. Par ailleurs, l'exploitation des visages pour le regroupement en locuteur sera présentée dans la section 2.3 de cet état de l'art.

2.2 Regroupement des visages

Le regroupement en classes de visages est le pendant vidéo du regroupement en locuteur. Il consiste à détecter et à regrouper les visages de sorte que chaque classe corresponde à une personne. Dans le cas d'une vidéo, cela revient à répondre à la question : *qui apparaît quand ?*

2.2.1 Difficultés et état de l'art en représentation des visages

Un visage contient de nombreuses marques distinctives comme la forme, la couleur des yeux ou de la peau. Toutes ces caractéristiques peuvent en principe être utilisées conjointement pour reconnaître une personne. Cependant, le regroupement doit être robuste aux variations d'apparence qui proviennent de changements de pose de la tête ou du plan de la caméra, des différentes expressions du visage (grimaces, sourire,...), d'occlusion d'une partie du visage par un autre objet, de changement des conditions d'illumination, de changement d'attributs (coiffure, lunettes, vêtements ; il faut cependant noter que ce cas est rare dans le cadre des vidéos de journaux télévisés), du flou engendré par des mouvements de la tête, et enfin des changements de résolution. Pour ces raisons, représenter un visage est encore une problématique non résolue.

Ces problèmes ont été largement étudiés dans le contexte de la reconnaissance du visage. De nombreux descripteurs optimisés ont été proposés comme les filtres de Gabor [Liu 2002], les *Scale Invariant Feature Transform* (SIFT) [Lowe 2004] ou les *Local Binary Pattern* [Ahonen 2006]. Ces descripteurs peuvent être extraits autour de points d'intérêt propres à chaque visage et choisis pour leurs particularités visuelles, sur des points précis du visage (les yeux, la bouche...) comme dans le cas des *facial landmarks* afin d'y associer une information sémantique, ou ils peuvent être extraits de manière dense sur toute l'image (voir illustration de la figure 2.3).

Comme pour le regroupement en locuteur, les méthodes utilisées dépendent du contexte applicatif. De nombreuses bases de données ont été créées pour évaluer l'état de l'art sur différentes problématiques (pose, résolution, niveau de contrôle des conditions). Afin d'avoir une idée des capacités actuelles des représentations de visage dans le contexte des émissions télévisuelles, il est intéressant d'observer les performances obtenues sur la base de données *Labeled Face in the wild* (LFW) [Huang 2007] qui est la plus proche de ce contexte. Au contraire d'autres bases de données où les conditions expérimentales sont soigneusement contrôlées, les visages y ont été collectés tels quels sur le web à partir d'articles de *Yahoo news !* La tâche associée à cette base est une tâche de vérification de visages : le système doit répondre à la question : *Est ce que cette paire de visages correspond à la même personne ?* Les méthodes utilisant des réseaux de neurones profonds publiées en 2014 obtiennent une précision de 97%

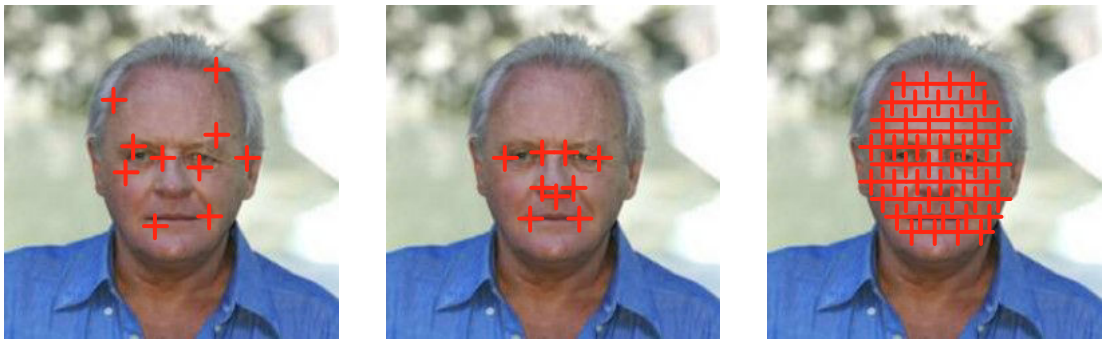


FIG. 2.3: Illustration de différentes techniques d'extraction des descripteurs. Chaque croix correspond à une zone de l'image où un descripteur est extrait.

ce qui est proche des capacités humaines [Fan 2014, Yaniv 2014, Sun 2014] (voir Fig 2.4 pour une illustration des résultats). Une autre approche populaire consiste à apprendre une métrique de manière supervisée [Simonyan 2013] ou semi-supervisées [Bhattarai 2014]. Je reviendrai sur les approches tirant partie des noms présents dans les données textuelles associées aux images dans la sections 3.2.1 et 3.2.2.

Il faut noter qu'une pré-traitement important consiste à effectuer un réalignement du visage afin d'obtenir une vue frontale. Dans le cas des vidéos, le suivi de visages sur des images consécutives permet de les associer de manière fiable. Les variabilités à l'intérieur de ces séquences peuvent être ensuite exploitées afin de mieux modéliser la variabilité des apparences de la classe. Selon le domaine d'application, il est généralement possible d'utiliser des informations additionnelles de celles pouvant être extraites de la fenêtre de détection du visage. Dans la suite de ce chapitre, nous aborderons l'utilisation du contexte visuel comme l'arrière-plan, les vêtements et la co-occurrence dans la section 2.2.2, l'utilisation conjointe des modalités audio et vidéo dans la section 2.3.

2.2.2 L'intégration du contexte

Plutôt que de chercher à obtenir des représentations de visage discriminantes en se fondant uniquement sur la fenêtre de détection, certains travaux cherchent plutôt à caractériser un visage à partir des éléments situés autour de lui. Cette approche est une alternative intéressante pour représenter des personnes vues de profil ou de dos ; c'est à dire des cas où il est difficile d'obtenir une vue frontale nécessaire aux approches basées sur la représentation du visage présentées dans le chapitre précédent. Les travaux de [Zhang 2013] sont particulièrement illustratifs de cette idée. 5 types d'éléments sont exploités :

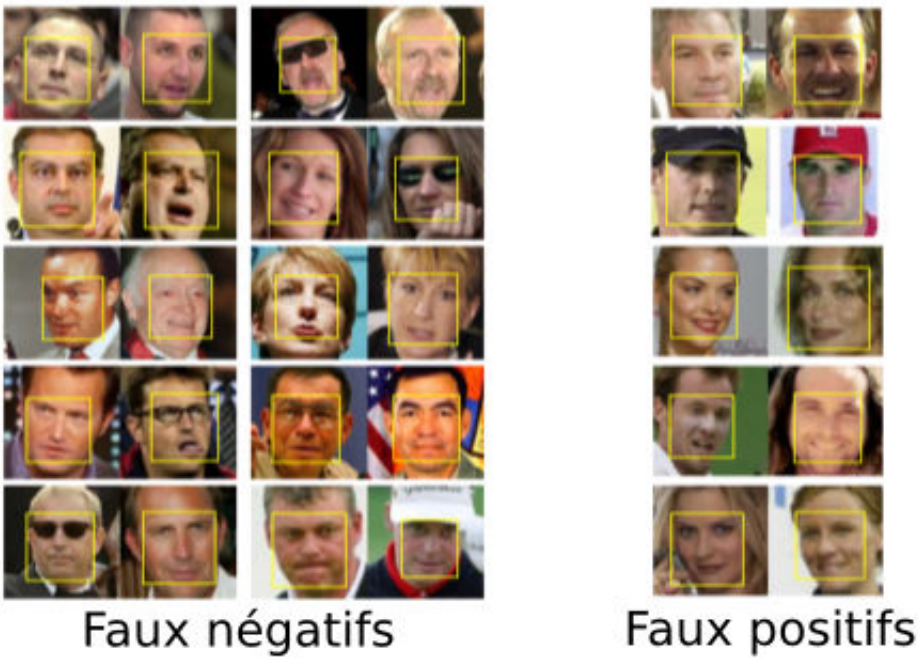


FIG. 2.4: Illustrations des résultats de [Fan 2014] sur la base de données LFW. Les paires dans la double colonne de gauche correspondent à la même personne mais ont été étiquetées comme différentes. La colonne de droite contient des images appartenant à des personnes différentes mais étiquetées comme semblables. Ces erreurs commises par le système sont aussi des cas difficiles à résoudre pour un humain. Figure extraite de [Fan 2014].

- **les vêtements :** la zone située en dessous du visage est considérée comme contenant le corps de la personne et est représentée par un histogramme de couleur.
- **la scène :** dans leur article, la scène fait référence à l'arrière-plan visuel, donc à un lieu. L'hypothèse est faite que 2 visages présents dans la même scène correspondent probablement à la même personne.
- **la co-occurrence :** le fait que 2 personnes apparaissent fréquemment ensemble permet de supposer la présence de l'une si l'autre est observée.
- **les attributs :** ceci fait référence à la détection de caractéristiques de haut niveau comme la couleur des cheveux, la forme du nez, la présence de lunettes. La détection de ce type d'attributs utilise des classifieurs spécifiques pré-entraînés.
- **contrainte d'unicité :** cette contrainte traduit l'hypothèse qu'une personne ne peut pas apparaître deux fois sur la même image.

Les éléments exploités ici sont intéressants par leur complémentarité : ils sont naturellement plus robustes aux variations affectant les représentations des visages comme par exemple celles de la pose et de l'expression.

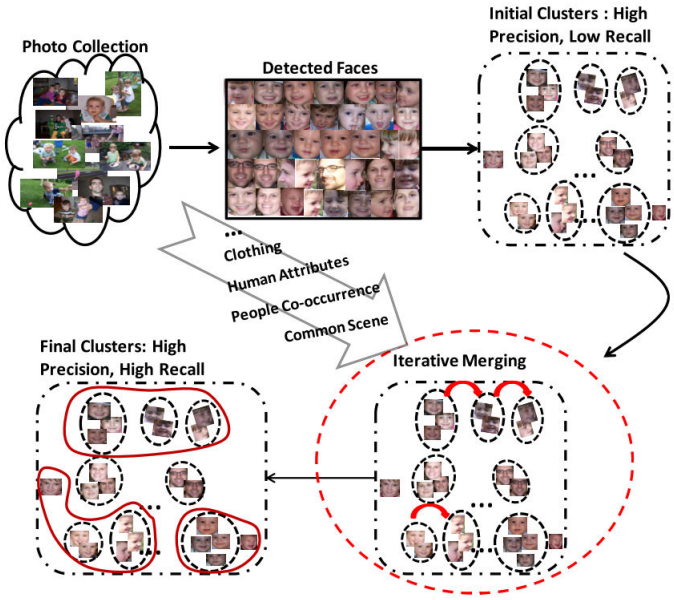


FIG. 2.5: Image extraite de [Zhang 2013] donnant une vue globale du système.

L'intégration de ces éléments est illustré sur la figure 2.5. L'approche commence par un regroupement hiérarchique fondé uniquement sur l'apparence visuelle du visage de sorte que la précision soit grande au détriment du rappel. Chaque classe ne contient donc que des visages de la même personne, mais une même personne peut être divisée en plusieurs classes.

Les caractéristiques de contexte sont alors utilisées dans un classifieur répondant à la question *"Ces deux classes doivent-elles être fusionnées ou non ?"* Afin d'obtenir des données d'apprentissage de manière non supervisée, des paires de visage d'une même classe sont sélectionnées en supposant qu'ils appartiennent à la même personne (hypothèse réaliste dans la mesure où le regroupement initial a une grande précision) et des paires de visage n'appartenant pas à la même personne sont obtenues en sélectionnant des visages qui apparaissent sur la même image. Ces paires positives et négatives forment ainsi un corpus d'apprentissage qui est exploité pour apprendre un classifieur d'association basé sur les autres caractéristiques contextuelles.

Une mesure de similarité entre classes est construite à partir d'arbres de classification appris sur ce corpus. Cette similarité est enfin exploitée par un regroupement hiérarchique classique. Les données utilisées dans les expériences sont des collections de photos de famille [Gallagher 2008], des vidéos de surveillance, et une collection de photos d'un mariage extraites de *Picasa*.

D'autres approches ont proposé des modèles graphiques afin d'intégrer ces mêmes éléments de contexte pour le regroupement des visages. Contrairement à l'architecture de [Zhang 2013]

qui utilise un regroupement hiérarchique, la décision est fondée sur une optimisation globale et non une série de décisions locales. En revanche, comme de tels modèles donnent à chaque visage une étiquette parmi un ensemble d'étiquettes prédéfinies, le nombre maximal de classes doit être fixé au préalable.

Parmi ces travaux, nous pouvons citer [Wu 2013] où des champs de Markov permettent d'exploiter la contrainte d'unicité et de propager les similarités entre visages. L'approche proposée dans [Du 2012] est un Champ conditionnel aléatoire (CRF) qui prend en compte la position dans l'image, la pose de la tête, la couleur des vêtements et la contrainte d'unicité. L'avantage d'utiliser un CRF par rapport à un modèle de Markov est la possibilité d'apprendre automatiquement des paramètres de pondération entre les différents éléments. Le système étant itératif, il peut faire varier le nombre d'étiquettes (le nombre de visage présents dans chaque image) à chaque itération. Le modèle est utilisé dans les expériences pour effectuer un suivi de visage à l'intérieur d'un plan sur des vidéos amateurs.

2.3 Traitement joint des locuteurs et des visages

Les sections précédentes ont permis d'aborder le regroupement en locuteur et le regroupement en classes de visages. Dans la mesure où le but final est d'obtenir pour chaque personne tous les instants où elle parle ET tous les instants où elle apparaît tel qu'illustré dans la figure 2.1, il est nécessaire de pouvoir associer les visages et les locuteurs. La section 2.3.1 passe en revue les techniques existantes pour obtenir cette association.

Par ailleurs, les liens d'association audiovisuels ont été utilisés avec succès dans plusieurs travaux pour corriger des erreurs venant des regroupements monomodaux. Pour chaque publication, l'importance des améliorations dépend de la stratégie de correction proposée mais également de plusieurs autres facteurs.

- Afin de corriger les erreurs des regroupements monomodaux, elles doivent survenir dans des zones où l'association audiovisuelle est fiable.
- Afin que les modalités audio et vidéo soient complémentaires, il vaut mieux que les erreurs de chaque modalité n'apparaissent pas aux mêmes endroits.
- Plus les systèmes monomodaux sont performants, plus il est difficile d'apporter des améliorations.

Ainsi, pour un type de vidéo donné, il est difficile de prévoir l'intérêt d'effectuer le regroupement audiovisuel de manière jointe sans connaître les performances des regroupements monomodaux et la fiabilité de l'association audiovisuelle. Pour chaque méthode que nous

présentons, ces points seront donc précisés dans la mesure où l'information est dans la publication. Dans la section 2.3.2, il sera présenté plusieurs approches utilisant cette association audiovisuelle pour améliorer le regroupement en locuteur. La section 2.3.3 décrit des systèmes effectuant le regroupement en locuteur et en classe de visages de manière jointe.

2.3.1 Association des locuteurs et des visages

Dans ce paragraphe, nous étudions les différentes possibilités pour retrouver le visage du locuteur dans l'image s'il y est. Dans une étude publiée dans [Bendris 2009] et portant sur les débats télévisés, il a été mesuré qu'un locuteur n'est visible que 60% du temps et qu'un visage ne parle que durant 30% de son temps d'apparition. Cette co-occurrence temporelle est déjà aisément exploitable afin d'associer les locuteurs et les visages à l'échelle des classes comme nous le verrons avec l'utilisation de l'algorithme hongrois. Cependant, elle ne suffit pas à résoudre les ambigüités sur un segment isolé. Les cas ambigus peuvent être classés dans les trois catégories suivantes, par ailleurs illustrées sur la figure 2.6.

- Présence de plusieurs visages dans l'image.
- Locuteur présent dans l'image, mais dont le visage est difficile à détecter.
- Locuteur absent de l'image et présence d'autres visages à l'écran. Il est important de noter que ce cas ne correspond pas simplement à la situation d'un intervenant en situation d'écoute. En effet, de nombreux programmes montrent des visages en train de parler pendant qu'une autre personne est audible. Par exemple, on peut imaginer un reportage où un journaliste en voix off commente les images d'un politicien prononçant un discours.

Détection de l'activité des lèvres

La plupart des méthodes résolvent le problème de détection du locuteur en mesurant l'activité des lèvres. Idéalement, il faudrait estimer la synchronie entre le mouvement des lèvres et le signal de parole. Plusieurs approches ont vu le jour et suivent généralement les étapes ci-dessous :

1. **Détection** : dans [El Khoury 2012b], la détection des lèvres est effectuée en supposant qu'elles se situent à une position fixée à l'avance par rapport au visage. Les travaux décrits dans [Bendris 2010] utilisent un modèle de visage standard [Milborrow 2008] qui apporte une plus grande précision mais demande une bonne résolution d'image pour être utilisé. On notera de plus que cette précision est sensible aux variations de pose.

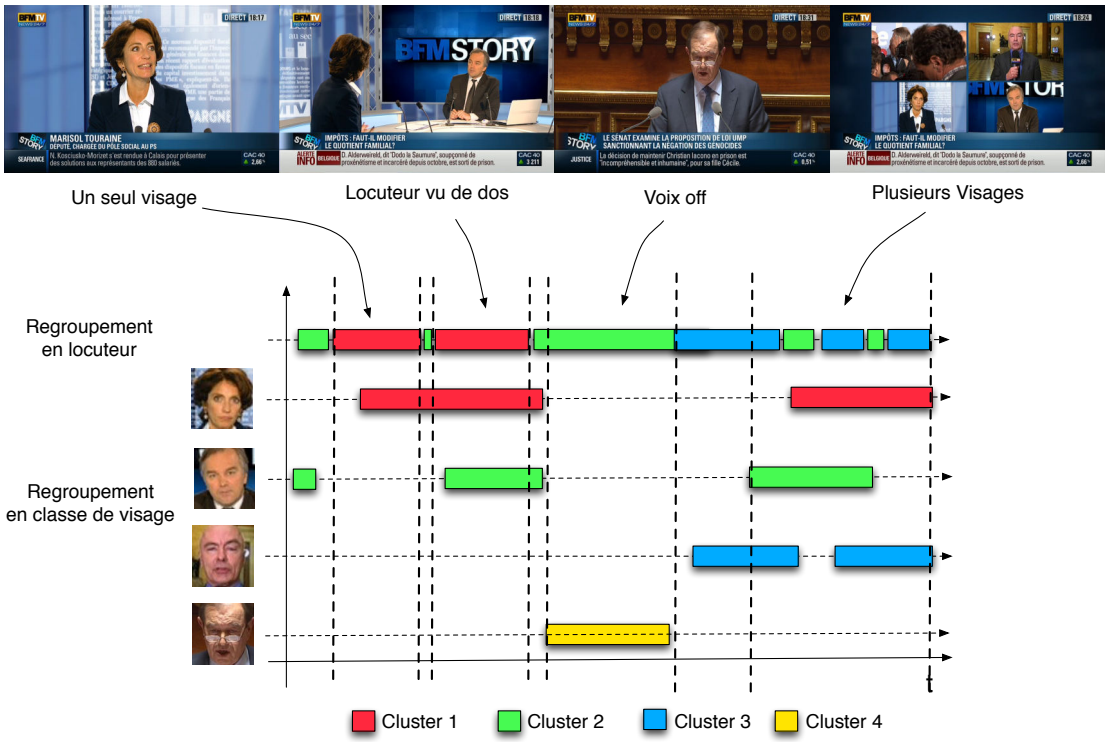


FIG. 2.6: Illustration des difficultés de l'association voix/visages : les images avec plusieurs visages et les cas où le visage du locuteur n'est pas détecté.

2. **Alignement :** afin de compenser des imprécisions dans la détection, les positions successives des lèvres sont alignées afin de s'assurer qu'elles correspondent à la même zone sur le visage.

3. **Mesure d'activité :** cette mesure est généralement calculée entre 2 images consécutives. Les travaux de [El Khoury 2012b, Everingham 2009] utilisent la différence moyenne d'intensité des pixels, ceux de [Bendris 2010] estiment le flux optique (le mouvement à l'intérieur de l'image à l'échelle des pixels) et se fonde sur l'entropie de la direction de ce flux. Le flux optique est aussi utilisé dans [Vallet 2013] où celui mesuré sur la tête est comparé à celui des lèvres. Dans [Noulas 2010], c'est l'information mutuelle entre les signaux visuel et audio qui est utilisé, des réseaux de neurones profonds sont également testés.

Peu de travaux évaluent la précision du module d'association voix/visages en raison du manque d'annotations. À titre d'indication, la précision de l'association voix/visages mesurée dans [Bendris 2010] est de 82.48% sur les vidéos de l'émission de divertissement de type plateau *"on a pas tout dit !"*.

Enfin, pour pallier les difficultés de la mesure de la synchronie des lèvres avec la parole, certains auteurs ont considéré d'autres éléments, comme la taille de la tête [El Khoury 2012b], la durée du plan [Vallet 2013], la position dans l'image, la pose et la durée de co-occurrence entre la séquence de visage et le tour de parole [Poignant 2013a].

2.3.2 Regroupement en locuteur assisté par l'information visuelle

Dans cette première partie, les méthodes présentées améliorent le regroupement en locuteur grâce à des informations extraites de la vidéo. Il existe également de nombreux autres travaux à l'instar de [Friedland 2009] sur le cas de réunions filmées par plusieurs caméras et enregistrées par plusieurs micros. Cependant, ces travaux ne seront pas traités ici car ils correspondent à des conditions trop différentes de celles utilisées dans les journaux télévisés.

Afin de s'assurer de la fiabilité de l'association visage/locuteur, [Vallet 2013] sélectionne les plans correspondant à des monologues avec un seul visage à l'écran en se fondant sur des critères de durée, de taille des visages et d'activité des lèvres. Les personnes détectées à l'intérieur de ces plans peuvent être précisément regroupées en se fondant sur des histogrammes de couleur des vêtements. Ceci permet d'améliorer l'initialisation du regroupement en locuteur en apprenant un modèle visuel et audio pour chaque classe. Des expériences sur l'émission de variétés *"Le Grand Échiquier"* présentent une amélioration de la mesure DER (voir la section 1.2.3 pour une définition du DER) de 38% à 32% par rapport à un système monomodal de type *Extended-Hidden Markov Model* décrit dans [Fredouille 2009]. Bien que les améliorations soient significatives, le système monomodal utilisé pour la comparaison n'est pas à l'état de l'art. Ainsi, pour les données REPERE, les DER obtenus pour les émissions de débat avec des systèmes fondés sur les i-vectors sont de l'ordre de 15%.

Le travail de [Noulas 2012] intègre les informations monomodales et multimodales de manière jointe dans un cadre probabiliste avec un modèle de Markov caché factoriel. L'attribution d'un tour de parole à une classe se fonde sur le score biométrique du tour avec le modèle de la classe et sur les liens d'association audiovisuels avec les visages apparaissant en même temps. Une représentation graphique est donnée dans la figure 2.7.

Pour chaque intervalle de temps t correspondant à une trame audio d'une durée de 40 ms, des observations $Y_t = (A_t, V_t, N_t^f, J_t)$ sont extraites de la vidéo.

- A_t correspond aux caractéristiques dérivées du signal audio.
- V_t correspond aux descripteurs extraits des visages détectés.
- N_t^f correspond au nombre de visages détectés.

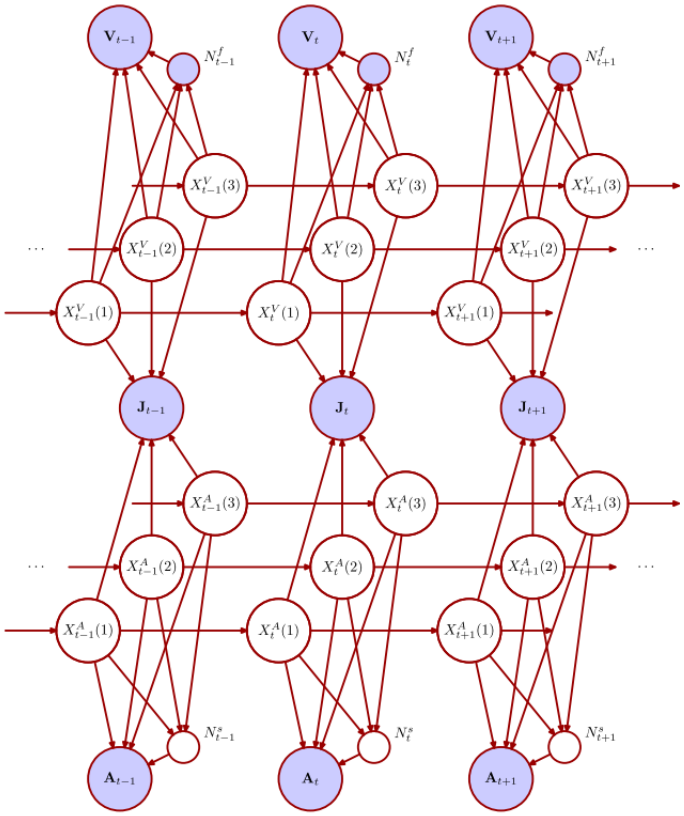


FIG. 2.7: Représentation du modèle utilisé par [Noulas 2012]. Les nœuds bleutés représentent les variables observées, les autres nœuds représentent les variables cachées. Le nœud N^s est une variable additionnelle servant à modéliser la parole superposée. Les autres nœuds ont été introduits dans le texte. Figure extraite de [Noulas 2012].

– J_t est une mesure de l'activité des lèvres de chaque visage.

Ce qui doit être estimé par le modèle est le vecteur d'état X_t correspondant à l'état des différents intervenants. Ce vecteur d'état peut être divisé en deux sous-vecteurs binaires X^V_t et X^A_t chacun étant de taille N. N est le nombre d'intervenants qui est supposé connu. L'élément $X^V_t(i)$ est un booléen indiquant si le $i^{ème}$ intervenant apparaît et l'élément $X^A_t(i)$ est un booléen indiquant si le $i^{ème}$ intervenant parle. En utilisant ces notations, la probabilité conditionnelle d'une réalisation des observations est alors :

$$P(y_t|x_t) = p(v_t, a_t, j_t, n_t|x_t) = p(v_t|x^v_t)p(n_t|x^v_t)p(j_t|x_t)p(a_t|x^a_t) \tag{2.1}$$

– Le calcul de $p(v_t|x^v_t)$ utilise une représentation des visages fondée sur la technique des *Bag Of Words* [Csurka 2004] avec des descripteurs SIFT [Lowe 2004].

FIG. 2.8: Quelques images extraites du journal télévisé utilisé dans les expériences. Figure extraite de [Noulas 2012].

– $p(n_t^f|x_t^a)$ permet d'inclure un *a priori* sur le nombre de locuteurs apparaissant dans l'image :

$$p(n_t^f|x_t) = \begin{cases} 0.9 & \text{si } \sum_i^N x_t(i) = N_t^f \\ 0.1/(N-1) & \text{sinon} \end{cases}$$

– $p(a_t|x_t^a)$ correspond à la vraisemblance des caractéristiques MFCC x_t^v par un modèle GMM appris sur les données associées au locuteur t.

– $p(j_t|x_t^a, x_t^v)$ est une mesure de la synchronie entre les mouvement des lèvres de chaque visage et le signal audio. Elle est calculée à partir de l'information mutuelle entre les signaux visuel et audio. Ce terme va encourager le système à décider que le locuteur courant apparaît à l'image si la synchronie mesurée est significative.

Comme un HMM, le modèle inclut aussi des probabilités de transition indépendamment pour chaque personne afin d'encourager la continuité des états.

– La probabilité $p(x_t|x_{t-1})$ est factorisé en $p(x_t^a|x_{t-1}^a) * p(x_t^v|x_{t-1}^v)$. Cette distribution est exprimée par $4N$ paramètres rangés dans deux matrices A^{na} et A^{nv}. L'élément A_{ij}^{na} (respectivement A_{ij}^{nv}) représente la probabilité du locuteur na (respectivement du visage nv) de passer de l'état i vers l'état j.

Les expériences sont effectuées sur des vidéos provenant de 3 réunions et d'un journal télévisé d'une des campagnes TRECVID (voir figure 2.8). Les performances sont mesurées en terme de précision : il s'agit de mesurer le taux de visages ou de trames acoustiques assignés à la bonne personne (ou éventuellement à aucune d'entre elles s'il s'agit d'une fausse alarme). Du point de vue de l'audio, le système multimodal permet une augmentation de la précision par rapport au système monomodal dédié aux vidéos de réunions décrit dans [Wooters 2008] (voir tableau 2.1). Il faut noter que le regroupement monomodal des visages est particulièrement fiable avec une précision moyenne de 98%. Cette fiabilité facilite l'utilisation de la vidéo pour améliorer les performances du regroupement en locuteur.

Méthode	IDIAP A	IDIAP B	Edinburgh	Journal télévisé
[Wooters 2008]	70%	70%	76%	77%
[Noulas 2012]	84%	77%	89%	94%

TAB. 2.1: Résultats en terme de précision reportés par [Noulas 2012] pour les trois meetings (IDIAP A, IDIAP B et Edinburgg) et le journal télévisé.

Comme dans les approches de [Du 2012, Wu 2013] décrites précédemment, le modèle graphique proposé ici permet de modéliser les interactions entre variables au sein d'une optimisation globale. En revanche, le nombre maximum de locuteurs présents doit être fixé à l'avance.

2.3.3 Systèmes de regroupement audiovisuel des personnes

Les méthodes présentées précédemment cherchent à améliorer le regroupement en locuteur. À présent, nous allons explicitement aborder la question de la tâche de regroupement audiovisuel des personnes. Les premières approches de regroupement joint effectuent généralement deux regroupements séparés des visages et des voix dans un premier temps, puis se fondent sur la co-occurrence entre classes pour l'association audiovisuelle [Dielmann 2010, Liu 2007, El Khoury 2007]. D'autres approches plus récentes cherchent à utiliser l'association audiovisuelle à l'échelle des segments pour guider les regroupements monomodaux.

La figure 2.9 illustre l'architecture d'un tel système qui est décrit dans [El Khoury 2012b]. Deux regroupements hiérarchiques monomodaux sont d'abord effectués séparément (étape 1 sur la figure). Ces regroupements hiérarchiques sont stoppés prématurément afin d'obtenir des classes pures. Ils permettent d'obtenir un premier ensemble de classes audios et vidéos et deux matrices de similarité S_a et S_v. Ensuite, une matrice de similarité audiovisuelle notée m est calculée où m_{ij} représente la similarité audiovisuelle entre le locuteur i et la classe de visage j (étape 2 sur la figure). Cette similarité est fondée sur la co-occurrence, la taille des têtes et l'activité des lèvres. Ces mesures de similarité sont intégrées par combinaison linéaire aux similarités monomodales (étape 3). Suivant ces nouvelles similarités, les classes les plus proches sont fusionnées. Le processus est itéré jusqu'à ce que les distances entre classes soient au-dessus d'un seuil (étape 4).

Des expériences sont effectuées sur des vidéos de différents types : débats (*Le grand journal, C'est dans l'air...*), journaux télévisés (*CNN, France2, LBC news...*) et films (*Les choristes, Virgins Suicide...*). Les regroupements monomodaux utilisés correspondent pour l'audio, à un système GMM/CLR [Barras 2006] et pour la vidéo à une version antérieure du système décrit dans le chapitre 4 où seulement la représentation à base de descripteurs SURF est utilisée. Par

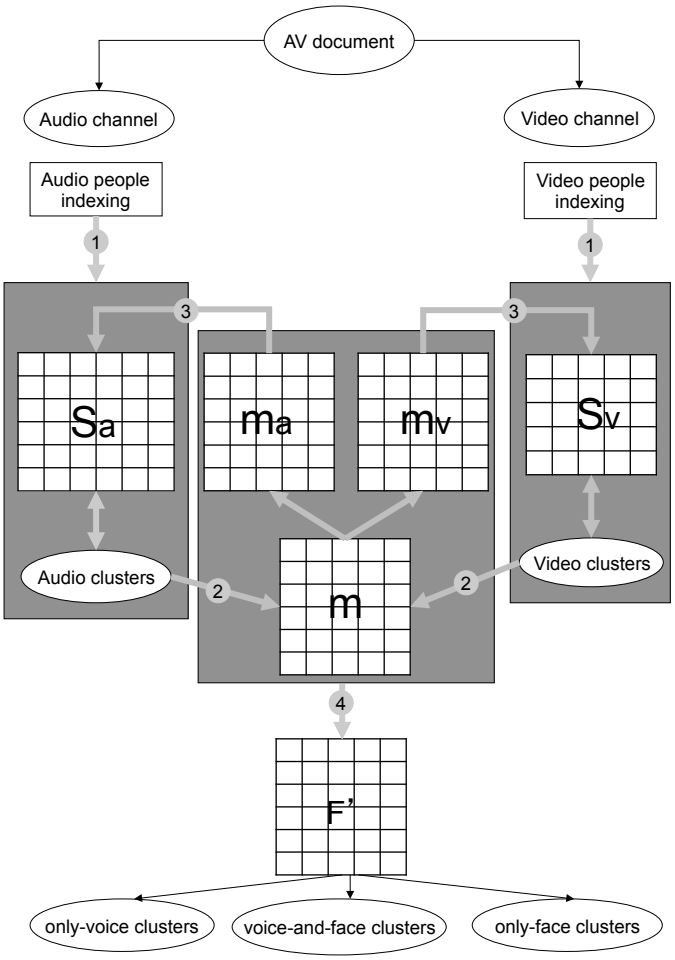


FIG. 2.9: Image extraite de [El Khoury 2010a] résumant le système de regroupement audiovisuel des personnes décrit dans la publication. Les matrices de similarités monomodales S_a et S_v sont mises à jour itérativement à partir de la matrice de similarité audiovisuelle m.

rapport au système monomodal audio, le DER passe de 18.68% à 15.85% pour les journaux télévisés et de 25,96% à 14,89% pour les débats. Les améliorations pour la vidéo sont reportées avec uniquement la partie confusion du DER. La baisse est de 9,10% à 7,64% pour les journaux et 15,73% à 12,41% pour les débats. Dans les deux modalités, il est notable que le bénéfice de l'association audiovisuelle est plus important pour les vidéos de débat.

L'approche décrite dans [Bendris 2011] commence également par effectuer deux regroupements séparés. Ensuite, les classes de voix et de visage sont associées avec un algorithme glouton en maximisant la co-occurrence entre les voix et les visages. Enfin, une dernière étape permet de raffiner le regroupement en remettant en cause les classes initiales. Premièrement, des

modèles biométriques sont appris sur les données de ces classes. Ces modèles servent à remettre en cause les couples de segments visage/voix se chevauchant temporellement mais n'étant pas assignés à la même classe :

1. un test est effectué pour chacun de ces couples. Il se fonde sur l'activité des lèvres et permet de décider si les deux segments appartiennent à la même personne.
2. Si le test est positif, l'un des segments endosse l'étiquette de l'autre de façon à maximiser les scores biométriques.

Les expériences sont effectuées sur des vidéos de l'émission de divertissement *On n'a pas tout dit* afin de valider l'intérêt de la dernière étape. C'est la qualité du regroupement audiovisuel des personnes qui est mesurée, c'est à dire qu'il faut que les numéros de classe attribués à un visage et un tour de parole appartenant à la même personne soient les mêmes. Il est reporté que la dernière étape permet d'augmenter la *F-mesure* (voir la section 1.2.3 pour une définition) de 72,7 à 77,3. La méthode utilisée pour associer les visages et les voix est décrite dans [Bendris 2010] et a été citée dans la section 2.3.1. Elle permet de décider si un visage et une voix correspondent à la même personne avec une précision de 82,48% sur ce type de données.

2.4 Conclusions

Ce premier chapitre a permis de passer en revue les techniques de regroupement audiovisuel des personnes. Dans chaque modalité, il semble que les systèmes utilisent d'abord des modèles simples pour détecter et regrouper les segments proches et similaires. Une fois que des classes de taille conséquente sont obtenues, des modèles plus complexes peuvent ensuite être appris. Dans cette thèse, l'accent a été mis sur cette dernière étape, en insistant sur les aspects multimodaux et sur l'utilisation du contexte.

Les techniques de regroupement en classes de visages qui ont été présentées ici nous donnent des indications utiles sur le type de contexte qu'il est possible d'intégrer (contrainte d'unicité, description des vêtements, arrières-plans, co-occurrence...). Cet examen sera poursuivi avec la revue des techniques d'identification du visage dans le chapitre 3. Les conclusions qui découlent de ces observations seront exploitées dans le chapitre 6.

Concernant le regroupement audiovisuel des personnes, les premières techniques associent les classes de visages et de locuteurs obtenues à partir de deux regroupements monomodaux indépendants. Plus récemment, des travaux intègrent l'information d'association directement

dans le regroupement. Ce dernier est alors effectué de manière jointe sur les segments audio et vidéo. Ce processus permet d'améliorer la qualité du regroupement audiovisuel des personnes. Il faut noter que la qualité des regroupements monomodaux est aussi améliorée en général. En particulier, des améliorations de performances en audio sont reportées par rapport à un système monomodal sur des vidéos de débats et de journaux télévisés.

Parmi les autres techniques présentées, celle de [Noulas 2012] à base de modèles graphiques semble particulièrement intéressante car :

- Il montre qu'une intégration de l'information multimodale est possible au sein d'un seul modèle graphique. Cela permet une optimisation globale du regroupement plutôt qu'une série de décisions prises localement.
- Ce type de modèles graphiques peut être facilement étendu ensuite pour intégrer d'autres éléments. Or, dans la mesure où il a été montré que le regroupement en classes de visages peut intégrer d'autres éléments que l'audio, cela nous semble une piste intéressante à suivre. En outre, l'objectif de nos travaux est également d'intégrer de l'information permettant d'identifier les personnes.

Notons cependant que ces expériences ont été effectuées dans un cadre un peu différent du notre car les données traitées correspondent à des vidéos de réunions et de journaux télévisés pour lesquelles le regroupement en classes de visages est presque parfait. Un modèle de ce type pour le regroupement audiovisuel sur les données REPERE sera proposé dans le chapitre 5. Avant cela, la deuxième partie de l'état de l'art va présenter une étude sur le nommage des personnes.

Chapitre 3

Identification des visages et des locuteurs

Sommaire

Ce chapitre donne un état de l'art sur l'identification non supervisée des visages et des voix dans les flux télévisuels. La tâche est étendue par rapport au chapitre précédent. À présent, les étiquettes à attribuer ne sont plus des classes anonymes mais des noms. Deux sources de nommage illustrées dans la figure 3.1 vont principalement être considérées ici : les noms issus des textes incrustés dans les cartouches et dans une moindre mesure les noms extraits de la transcription. Par non supervisée, nous entendons que les sources de nommage ne prendront pas en compte d'éventuels modèles biométriques de personnes appris *a priori*. Si le lecteur souhaite avoir une autre vue sur ce même sujet, il peut consulter la thèse de Johann Poignant [Poignant 2013a] dont l'état de l'art correspond peu ou prou aux deux premières sections de ce chapitre. La section 3.1 traite de l'identification des locuteurs et montre l'intérêt de l'utilisation des noms extraits des cartouches par rapport à ceux de la transcription. La section 3.2 aborde l'identification des visages et permet de positionner notre problème par rapport à d'autres travaux traitant d'images ou de vidéos de série de fiction. La dernière section 3.3 présente les travaux réalisés pendant la campagne REPERE, avec notamment l'intégration de la multimodalité et l'exploitation du contexte à travers la compréhension de scènes.

FIG. 3.1: Le personnage interviewé dans l'image de droite (Jean ARTHUIS) peut être identifié de 3 manières : par le cartouche où est écrit son nom et sa fonction, par la transcription du présentateur qui annonce qui va parler, par une correspondance avec des modèles biométriques de personne entraînés sur un corpus d'apprentissage.

FIG. 3.2: Noms prononcés et noms écrits provenant des cartouches dans des journaux télévisés. Image extraite de [Poignant 2013b].

3.1 Identification des locuteurs

Cette section traite du nommage des locuteurs. Une première discussion portera sur les avantages à utiliser les noms extraits de la transcription ou des cartouches. Ensuite, les méthodes d'identification à partir de la transcription, seront abordées, puis nous étudierons celles qui recourent aux cartouches.

3.1.1 Sources de nommage non supervisées : transcriptions *vs* cartouches

Concernant l'identification des locuteurs dans les émissions de télévisions, l'étude de Poignant [Poignant 2013a] à partir des données de la campagne REPERE donne les conclusions suivantes :

- Les techniques de l'état de l'art permettent d'extraire les noms issus des cartouches avec plus de précision que les noms cités à l'oral grâce aux progrès de la reconnaissance optique des caractères (OCR).
- De [Poignant 2013a] :

  > *Les noms cités à l'oral proposent un plus grand nombre d'occurrences ainsi qu'un plus grand nombre de personnes différentes citées. En contrepartie, la probabilité que les personnes correspondant à ces noms soient présentes dans les vidéos est plus faible.*

  Ainsi, dans l'exemple de la figure 3.2, *Philippe Salvador* et *Roland Vienne* qui apparaissent à l'écran sont tous deux cités à l'oral et à l'écrit. En revanche, *Nathalie Kosciusko-Morizet* qui est citée à l'oral n'apparaît pas.

Dans la suite de cet état de l'art, la plupart des travaux récents privilégient les cartouches sur la transcription comme source de nommage car les noms y sont plus faciles à extraire et à associer. Il est déjà possible d'entrevoir pourquoi en considérant la figure 3.2. Sur les deux

images du milieu, il suffit d'associer le nom présent dans le bandeau bleu au locuteur courant alors que l'association des noms issus de la transcription demande une analyse sémantique de la parole transcrite.

Il reste pourtant à effectuer l'analyse sémantique du texte pour déterminer si il sert à annoncer une personne ou non. Par exemple, ce serait une erreur de chercher à associer les noms présents dans les bandeaux blancs car ces personnes n'apparaissent pas à l'image. Cependant, il est généralement supposé que le rôle des différents bandeaux est connu *a priori*.

3.1.2 Utilisation de la transcription

Dans [Canseco-Rodriguez 2004, Tranter 2006] et [Jousse 2009], les auteurs étudient l'utilisation de la transcription pour l'identification du locuteur. L'objectif est de collecter des évidences locales venant de chaque nom prononcé pour ensuite les combiner au niveau global et prendre une décision au niveau de la classe. La première et la deuxième référence utilisent des règles linguistiques fondées sur des modèles N-grammes. La troisième utilise des classifieurs apprenant à partir de données étiquetées la probabilité qu'un nom prononcé appartienne au locuteur précédent, courant ou suivant. Sur le corpus d'émissions radiophoniques de la campagne ESTER1, [Jousse 2009] note un rappel de 83% pour une précision d'identification de 90% avec une transcription manuelle. Les performances sont de 18% pour le rappel et de 43% pour la précision avec un système complètement automatique (Le DER du regroupement en locuteur est de 11% et le Taux d'erreur mot du système de transcription est de 20.5%). Leurs commentaires mentionnent que l'augmentation des erreurs d'identification observée avec l'utilisation du regroupement automatique vient principalement d'erreurs de segmentation aux frontières des segments. De plus, il est important que la transcription soit performante dans la reconnaissance des noms propres afin qu'ils puissent être utilisés pour nommer les locuteurs. Dans [El Khoury 2012a], ce système est étendu en y ajoutant des modèles biométriques.

Avec la motivation d'utiliser une grande variété de caractéristiques hétérogènes, Chenguyan [Ma 2007] utilise un modèle de type *Maximum d'Entropie* (MaxEnt) et améliore les performances obtenues dans [Tranter 2006]. Les caractéristiques utilisées sont les règles linguistiques décrites dans [Tranter 2006], la compatibilité des genres entre le nom prononcé et le locuteur, la position du nom par rapport au tour de parole du locuteur et d'éventuels scores de modèles biométriques.

En règle générale et à l'image des résultats reportés dans [Jousse 2009], ces travaux obtiennent de bons résultats avec une transcription et un regroupement en locuteur manuels, mais les

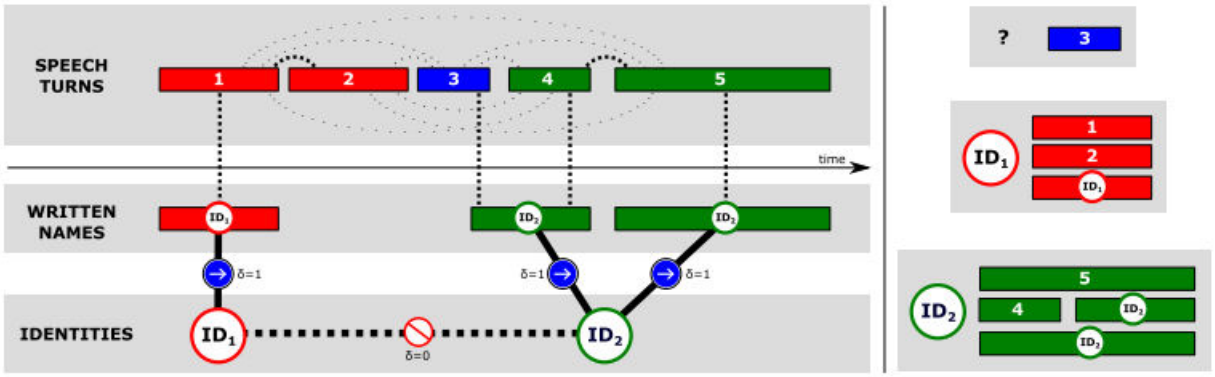


FIG. 3.3: À gauche : example de graphe probabiliste multimodal incluant les tours de parole, les cartouches et les variables d'identités. À droite, les résultats attendus : les classes de locuteurs et les identités assignées. Image extraite de [Bredin 2013].

performances se dégradent vite avec des systèmes automatiques. Enfin, il est probable que les performances de ces systèmes testées sur des émissions de radio se dégradent sur des vidéos. En effet, les personnes sont moins souvent citées à la télévision qu'à la radio car l'image suffit au spectateur pour retrouver l'identité du locuteur.

3.1.3 Utilisation des noms écrits

La plupart des systèmes d'identification de la campagne REPERE [Poignant 2013a, Bendris 2013] utilisent en priorité l'information venant des cartouches. Ces systèmes d'identification des locuteurs et des visages seront décrits dans la section 3.3.

Concernant des travaux évalués sur l'identification des locuteurs seuls, [Bredin 2013] construit un graphe complet (voir figure 3.3) entre les tours de parole et les cartouches. Chaque arc entre deux tours de parole est pondéré par une mesure de similarité fondée sur la distance BIC [Chen 1998]. Les similarités entre les tours de parole et les cartouches sont calculées à partir des fréquences sur un ensemble d'apprentissage. Une variable d'identité ID est ajoutée pour chaque nom extrait des cartouches. Ensuite, l'algorithme de regroupement ILP (similaire à celui utilisé par le système de regroupement en locuteur présenté dans la section 2.1) est utilisé sur ce graphe. Des contraintes sont ajoutées dans l'algorithme pour empêcher deux variables d'identités portant des noms différents d'apparaître dans la même classe. Ainsi le regroupement obtenu peut être utilisé directement pour le nommage car une classe n'est associée qu'à un seul nom. Ce système est évalué sur la partie TEST0 des données REPERE, il obtient une précision de 90.6% et un rappel de 58.2%. Ces résultats sont intéressants, mais nous verrons dans la section 3.3 que des systèmes intégrant d'avantage d'éléments de contexte obtiennent de meilleures performances.

3.2 Identification des visages

Un des premiers travaux associant visages et noms est le système *name-it* de [Satoh 1999]. Par rapport à l'identification des locuteurs, l'apparition de plusieurs visages sur une même image crée des ambiguïtés supplémentaires. Cela est particulièrement vrai pour les journaux télévisés depuis le passage au 16/9 et à la HD : en de nombreux endroits, plusieurs images sont incrustées sur un même plan (par exemple : image d, figure 1.2). Dans ce contexte, des simplifications comme celles utilisées par [Yang 2004] où seuls les monologues avec un visage sont traités deviennent vite restrictives.

Par ailleurs, alors qu'il est difficile de construire un modèle robuste à partir d'un segment de parole de quelques secondes, une image d'un visage est suffisante pour construire une représentation intéressante. Les méthodes d'identification de visages peuvent donc plus volontiers travailler à une granularité plus faible, au niveau du segment plutôt qu'au niveau de la classe.

Cependant, quelle que soit la méthode d'association entre le visage et le nom issu du cartouche, il est évident que cette information est parcimonieuse. De même que pour l'identification des locuteurs, la qualité du regroupement conditionne donc la capacité à propager les noms à toutes les occurrences d'une personne.

Dans la suite de cette section, plusieurs approches visant à identifier des visages de manière non supervisée vont être présentées. Bien qu'elles ne soient pas toutes appliquées aux journaux télévisés, elles partagent le fait que l'information de nommage présente est ambigüe. Nous allons donc voir les différentes stratégies qui ont été mises en place pour résoudre ces ambiguïtés. L'idée générale est qu'il est possible d'obtenir un *a priori* sur l'identité de quelques visages à partir des co-occurrences entre visages et noms. Ces *a priori* servent alors à guider le reste du regroupement ou à construire des représentations visuelles plus discriminantes. Les sources de nommage dans cette section sont des transcriptions et des textes accompagnant les images de personnes. L'utilisation des cartouches pour l'identification des visages sera traitée dans la section 3.3.

3.2.1 Regroupement contraint par l'information venant des noms

Les auteurs de [Berg 2004] introduisirent des données composées d'articles du site *Yahoo !News*. Ce sont des articles de presse qui sont chacun accompagnés d'une image illustrative contenant des visages (figure 3.4). Souvent, les visages correspondent à des noms cités dans

FIG. 3.4: Exemple d'images accompagnées de textes de la base *Labeled Yahoo ! News*. Étant donnés les images et les textes, la tâche associée à ces données est de nommer les visages. Image extraite de [Berg 2004].

le texte. Ainsi, ces données donnèrent lieu à une importante littérature sur l'identification des visages : l'enjeu est d'identifier les visages à partir des noms extraits des textes.

L'approche proposée dans [Berg 2004] consiste à adapter un mélange de modèles gaussiens où chaque composante représente un nom. Ces gaussiennes modélisent la distribution de descripteurs extraits des images de visage. Le descripteur utilisé s'inspire de [Yang 2002] : Il est obtenu en appliquant deux transformations à l'image originale : Une analyse dite *kernel principal component analysis* est utilisée afin de réduire sa dimension, et son potentiel discriminant est augmenté avec la méthode *Linear Discriminant Analysis*. Dès lors, les paramètres de ce mélange peuvent être estimés avec un algorithme de type *Expectation-Maximization* (EM) maximisant la vraisemblance. Cette maximisation inclue des caractéristiques et des contraintes issues du texte :

– L'initialisation pour l'algorithme EM utilise une règle simple sur les co-occurrences entre les visages et les noms. Un nom et des visages qui présentent plusieurs co-occurrences sont associés. Le nombre de gaussiennes correspond au nombre de noms.

– L'attribution des visages à chaque composante lors de la mise à jour des paramètres tient compte des co-occurrences entre les noms et les visages. Enfin, une contrainte d'unicité empêche 2 visages de la même image d'avoir le même nom.

Les auteurs de [Guillaumin 2012] améliorent les performances obtenus dans [Berg 2004]. Il propose de calculer des similarités visuelles entre chaque paire de visages, puis de maximiser la somme des similarités des paires de visages associés à un nom tout en respectant les contraintes suivantes :

1. Un visage ne peut être attribué qu'à un seul nom.
2. Ce nom doit provenir du texte associé à l'image où le visage apparaît.
3. Dans une même image, un nom ne peut être associé qu'à un seul visage.

La contrainte numéro 2 permet de profiter des co-occurrences entre noms et visages. Cependant, il est probable que ce type de contraintes soit inadéquat pour les journaux télévisés où il y a beaucoup moins de noms que dans le cas de ces paires image/texte. Les co-occurrences contiennent donc moins d'information.

3.2.2 Apprentissage à partir de données faiblement étiquetées

Dans le contexte de l'apprentissage statistique, le terme de données faiblement étiquetées désigne simplement le fait que seul un *a priori* sur les étiquettes des données est disponible. Le but des approches présentées ici est d'adapter les méthodes d'apprentissage supervisé à ce nouveau cadre.

Les premiers travaux de ce type appliqués à des journaux télévisés sont ceux de [Yang 2005] qui utilisent le concept de *Multiple Instance Learning* (MIL [Dietterich 1997]). L'idée est illustrée sur la figure 3.5. Les exemples d'apprentissage sont regroupés en sacs. Il est supposé que chaque exemple peut avoir deux étiquettes : une positive, une négative. L'étiquetage n'est cependant pas connu au niveau de chaque exemple, mais uniquement au niveau du sac. Si pour un sac donné, il y a au moins un exemple avec une étiquette positive, le sac est étiqueté comme positif ; sinon, il est négatif. L'apprentissage consiste alors à localiser la zone dans l'espace des caractéristiques qui inclut une partie de chaque sac étiqueté comme positif, mais qui exclut tous les éléments venant des sacs étiquetés comme négatifs. Dans ce cadre, les auteurs de [Yang 2005] définissent un exemple comme étant une paire Nom/Visage et un sac comme étant un visage et les noms potentiels qui y sont associés. Un sac est alors positif si l'un des noms correspond au visage. Il utilise alors l'approche MIL pour apprendre un classifieur qui réponde, pour chaque exemple, à la question : *Est ce que ce nom correspond à ce visage ?* Ce classifieur est appris sur des caractéristiques dérivées des noms extraits de la transcription, des noms extraits des cartouches et sur la correspondance de genre entre le nom et le visage.

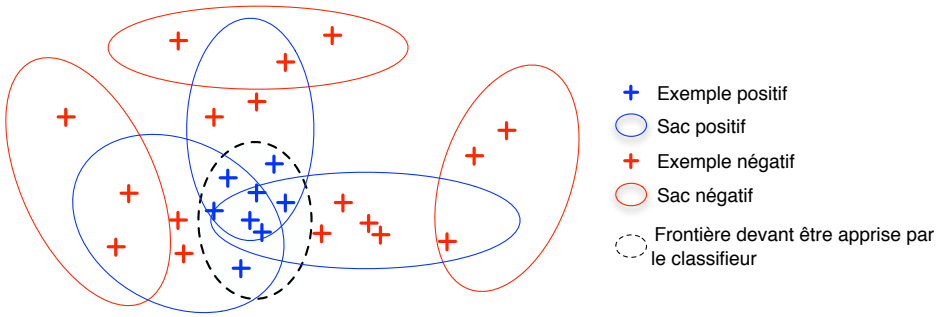


FIG. 3.5: Illustration du problème d'apprentissage MIL. Le classifieur doit être appris pour séparer les points négatifs (rouges) des points positifs (bleus). Pour cela, il doit déterminer la zone commune aux sacs étiquetés positifs en excluant la zone des sacs négatifs. Pendant l'apprentissage, l'algorithme n'a accès qu'aux étiquettes des sacs.

De nombreux travaux ont eu lieu sur l'identification dans des séries télévisées. Ce sujet a été popularisé par [Everingham 2006] avec des vidéos de la série *Buffy the vampire slayer*. Dans ce contexte, illustré par la figure 3.6, les noms sont extraits des sous-titres et du script ce qui permet de connaître l'identité du locuteur (avec quelques erreurs dues à l'imprécision des bornes des sous-titres). Grâce à un module de détection visuelle du locuteur, le nom du locuteur peut être propagé vers son visage. Cependant, tous les visages ne peuvent pas être nommés par le script. Le but est alors d'apprendre un classifieur pour chaque personnage à partir des visages déjà identifiés afin de propager leur identité vers les visages restants. Là encore, les tâches de regroupement et d'identification sont liées.

Les travaux décrits dans [Wohlhart 2011, Kostinger 2011] tentent de résoudre ce problème avec une approche de type MIL. Pour chaque personnage, un ensemble de sacs positifs et négatifs est généré. Les sacs positifs correspondent à une séquence de visages identifiée par propagation de l'identité d'un locuteur. Un sac négatif pour ce personnage est une séquence de visages détectée comme silencieuse alors que le script annonce que ce personnage parle. Une probabilité est aussi estimée afin de prédire si un personnage apparaît à l'écran à partir de caractéristiques dérivées de la transcription (présence répétée de son nom entre autres). Cette information est incluse dans l'apprentissage par l'intermédiaire de sacs contenant tous les visages d'une scène et ne possédant pas d'étiquette binaire mais seulement une distribution sur la valeur de cette étiquette. L'évaluation est conduite à partir de séquences de visages obtenues automatiquement et étiquetées ensuite manuellement avec des noms. Elle porte donc sur la capacité à identifier les visages détectés automatiquement en ignorant le fait que certains visages ont pu ne pas être détectés. Les résultats publiés dans [Wohlhart 2011] montre une précision d'identification de 68% avec l'utilisation des sous-titres seuls.

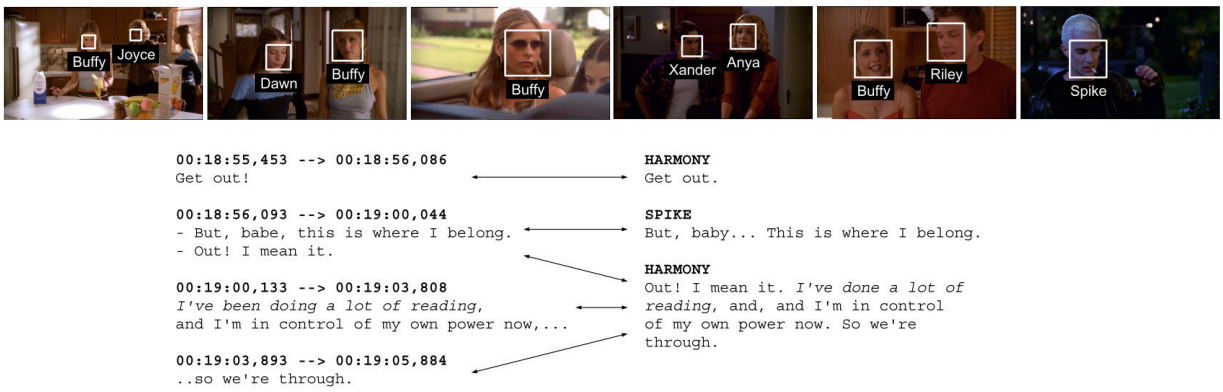


FIG. 3.6: Les images du haut montrent des extraits de la série avec des visages identifiées. Le texte en bas est un extrait du script et des sous-titres qui permet d'identifier les locuteurs. Images extraites de [Everingham 2006].

Les travaux présentés dans [Cour 2009, Cour 2011] se placent dans le même contexte d'identification dans des séries télévisées (la série *Lost*) à partir de scripts. Le but est d'apprendre un classifieur dans des conditions appelées *étiquetage ambigu* : chaque exemple est fourni avec plusieurs étiquettes et une seule d'entre elles est correcte. Le but est là aussi d'apprendre un classifieur de type un contre tous pour chaque nom. Cette fonction intègre des caractéristiques additionnelles à la similarité entre visages : la détection du locuteur par mesure du mouvement des lèvres et une hypothèse stipulant que deux visages dans deux plans consécutifs sont différents.

[Bauml 2013] utilise un cadre d'apprentissage semi-supervisé. Le script et les sous-titres permettent de nommer quelques séquences de visages qui correspondent alors aux données étiquetées. Un classifieur de type régression logistique multinomiale est appris à partir des données étiquetées et non-étiquetées en optimisant une fonction de coût. Afin de tenir compte des visages non-étiquetés, un terme dans cette fonction encourage les frontières du classifieur à se placer dans des zones de l'espace des caractéristiques de faible densité. La fonction intègre également des contraintes d'unicité. Les effets de ces deux termes sont illustrés sur la figure 3.7. L'incertitude des étiquettes est gérée par une pénalité de régularisation qui, en empêchant les frontières de décision complexes, permet d'être robuste aux points aberrants. Les expériences sont effectuées sur deux séries. La précision d'identification est de 79% pour *Big Bang Theory* (BBT) et 66% pour *Buffy The Vampire Slayer* (BF). Cependant, il est difficile de conclure sur la différence de performances entre les deux émissions. En effet, elles peuvent venir des variations

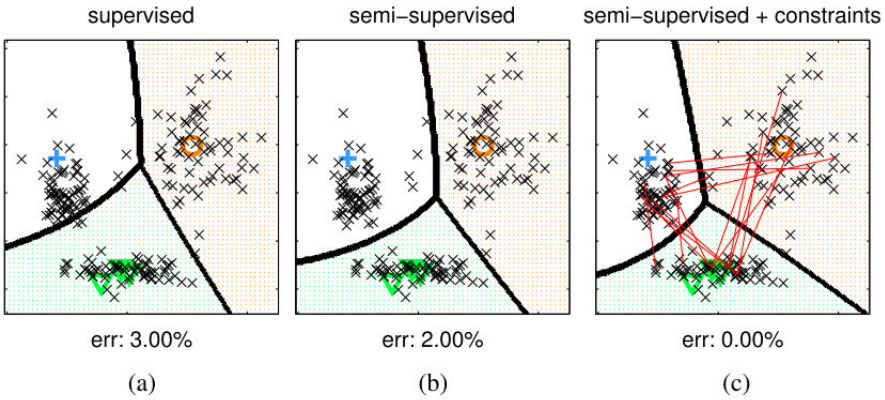


FIG. 3.7: Visualisation des effets des différentes parties de la fonction de coût sur des données synthétiques. (a) Apprentissage à partir des données étiquetées. (b) Ajout des données non-étiquetées dans l'apprentissage, ce qui permet d'adapter les frontières à la distribution des données. (c) Avec les contraintes d'unicité, les erreurs de classification sont encore réduites. Image extraite de [Bauml 2013].

visuelles des données (illumination, changement d'apparence) comme de la différence de quantité d'exemples étiquetés qu'il est possible d'obtenir par les ressources textuelles. Néanmoins, il a été remarqué que la contrainte d'unicité est plus utile pour BBT, car le nombre de personnes est plus faible (5 versus 12) et le nombre moyen de visages par image est plus élevé. Là aussi, il faut noter que cette précision est mesurée sur les visages déjà détectés.

La comparaison de l'identification dans des séries de fiction par rapport aux journaux télévisés fait ressortir plusieurs points. Certaines séries ont des variations visuelles relativement importantes dans le cas de changement de décor, d'apparence des personnages, etc. De plus, les scripts sont des sources de nommage plus riches que la transcription ou les cartouches. Enfin, le nombre de visages est généralement fixé pour une série donnée en se concentrant sur les personnages principaux (12 pour la série buffy, 5 pour la série *Big Bang Theory*). Dans les vidéos REPERE, le nombre de visages à identifier dépend de la durée de la vidéo car de nouveaux visages apparaissent à chaque nouveau sujet dans un journal télévisé. Il dépend aussi du type de scène. Une vidéo de débat contient quelques intervenants (3 ou 4 personnes maximum), tandis qu'un journal télévisé d'une demi-heure peut contenir les visages d'une vingtaine de personnes différentes.

3.2.3 Apprentissage de modèles biométriques de manières non supervisée et utilisation de données externes

Les approches de la section précédente permettent de nommer des visages présents dans un document à partir des ressources textuelles liées à celui-ci. D'autres travaux ont cherché à apprendre un modèle de visages à partir de données faiblement étiquetées obtenues à partir de ressources externes.

Les auteurs de [Liu 2008] utilisent le moteur de recherche *Google Image*. Le nom de la personne dont ils souhaitent obtenir un modèle est utilisé comme requête. La personne doit être suffisamment célèbre pour que les résultats de la requête contiennent des visages de cette personne. Même ainsi, certains visages correspondent à d'autres personnes. Les auteurs mesurent que 86% des visages obtenus correspondent à la personne cible. Dés lors, une sélection des visages est effectuée en cherchant la partie la plus dense dans un graphe de similarité. Cette approche a déjà fait l'objet d'une application dans l'état de l'art avec [Guillaumin 2012](section 3.2.1,page 41). Les données collectées servent ensuite à apprendre un modèle de visage. Ce modèle peut ensuite être utilisé pour l'identification dans de nouveaux documents. Les expériences sont faites sur des vidéos de journaux télévisés de la compétition TRECVID. Les vidéos choisies concernent des affaires internationales, les personnes à identifier sont ainsi suffisamment connues.

Les travaux publiés dans [Parkhi 2012] utilisent le même principe afin de répondre à des requêtes sur une collection de vidéos. Dans ce contexte, un utilisateur souhaite obtenir les vidéos où une certaine personne nommée dans une requête textuelle apparaît. Un modèle biométrique de la personne cible est appris en ligne en se servant des données retournées par *Google*. Ce modèle sert alors à retrouver les vidéos où cette personne apparaît. Ce travail a été inclus dans le système AXES PRO [McGuinness 2013] qui a pour but de développer des outils pour la gestion d'archives audiovisuelles.

La même motivation a conduit [Ozkan 2006] et [Song 2004] à utiliser des collections de vidéos avec la transcription comme source faible de nommage. Leur application vise à identifier des personnes célèbres.

Dans tous les cas, un point important est que la quantité de données externes soit suffisamment grande ou que les personnes cibles soient suffisamment connues afin d'obtenir un nombre important de co-occurrences de la paire visage/nom recherché. L'extension et l'évaluation de ces méthodes pour des personnes peu connues n'a pas été traitée dans la littérature.

3.3 Identification jointe des visages et des locuteurs

Dans cette dernière section, le but des systèmes présentés est de regrouper et d'identifier les visages et les locuteurs. Ces systèmes peuvent donc utiliser l'association audiovisuelle visage/locuteur ainsi que des informations venant des noms pour prendre la meilleure décision.

3.3.1 Soumissions de la campagne REPERE

3.3.1.1 Regroupement multimodal contraint

Johann Poignant a effectué sa thèse sur l'identification des visages et des locuteurs dans des flux vidéos [Poignant 2013a, Poignant 2012]. Notre cadre de travail est identique au sien. Il propose un système d'identification multimodal de manière jointe et non-supervisée. La source de nommage utilisée est l'ensemble des noms extraits des cartouches. Plusieurs approches sont discutées : dans la première, les noms sont intégrés tardivement après avoir effectué le regroupement ; à l'inverse, dans la deuxième méthode, l'information de nommage est exploitée de manière précoce à l'intérieur de l'opération de regroupement. Le nommage tardif est utile si le système de regroupement est une boîte noire. En revanche, s'il est possible d'y accéder et de le modifier, le nommage précoce permet d'obtenir de meilleures performances. Ce dernier système est illustré sur la figure 3.8 et nous le décrivons plus en détail.

Dans une première étape, les noms sont propagés de manière locale vers les visages et les tours de parole en se fondant sur la co-occurrence. Ensuite, un regroupement hiérarchique multimodal contraint est effectué sur ces segments. Pour cela, des matrices de distance sont calculées entre les paires de visage D_{visage}, les paires de tours de parole S_{voix}, et les couples visage/tour de parole $P_{voix/Visage}$. Le score d'association des tours de parole et des visages est calculé avec un perceptron multi-couches appris sur diverses caractéristiques décrites dans la section 2.3.1, page 26. Ces différentes distances sont homogénéisées sous forme de probabilités. Elles sont ensuite combinées linéairement pour former la distance finale entre deux classes multimodales. La matrice de distance qui en résulte est utilisée dans un regroupement hiérarchique contraint. Les contraintes sont les suivantes :

1. Deux classes ne sont pas fusionnées si les noms attribués par l'étape de nommage précoce sont différents.
2. Deux classes ne sont pas fusionnées si elles apparaissent visuellement dans la même image.

La première contrainte donne la priorité à l'information venant du nom extrait des cartouches et suppose donc que cette source est la plus fiable. Le fait d'effectuer le regroupement multimodal

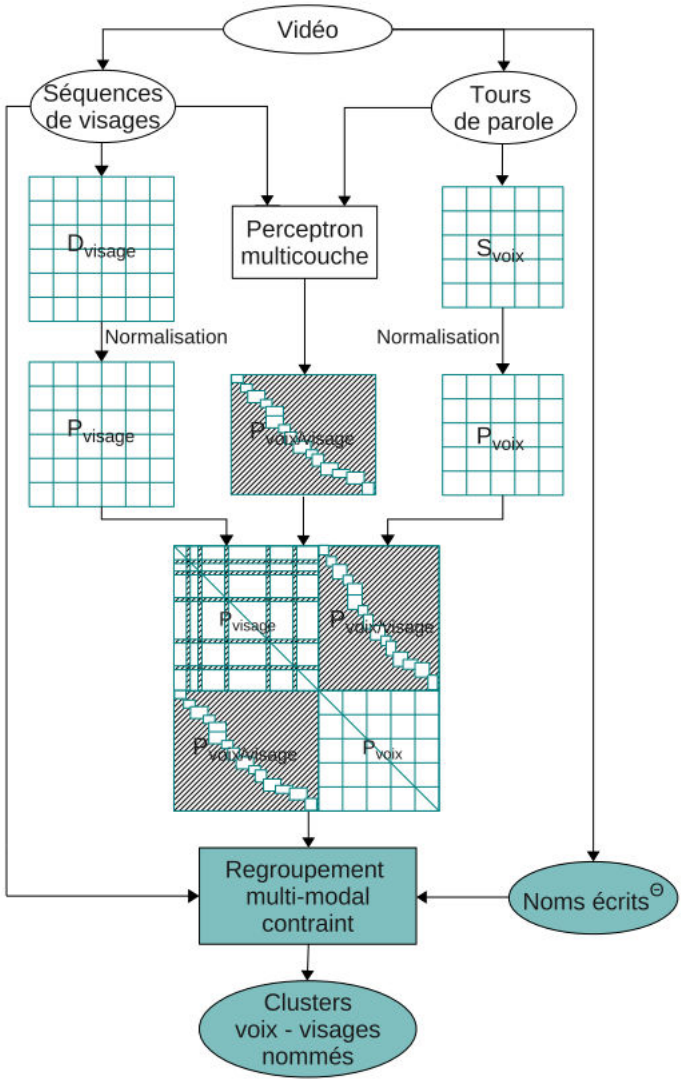


FIG. 3.8: Schéma du système de nommage précoce. Les matrices de similarité audio, vidéo et audio-vidéo sont rassemblées dans une matrice de distance multimodale. Cette matrice est alors utilisée pour le regroupement multimodal contraint. Image extraite de [Poignant 2012].

de manière jointe permet de propager des identités d'une modalité à l'autre et donc d'augmenter le rappel.

L'évaluation porte sur la partie PHASE1 des données de la campagne REPERE. Les résultats sont reportés en taux d'erreur EGER qui est la métrique officielle de la campagne. Le nommage tardif obtient un EGER de 32.1% pour les locuteurs. Le nommage précoce appliqué sur des classes monomodales obtient un EGER de 29.9% pour les locuteurs et 49.8% pour les visages. La méthode finale de nommage précoce avec la matrice multimodale obtient des performances légèrement meilleures pour les locuteurs avec 29.2% et améliore significativement

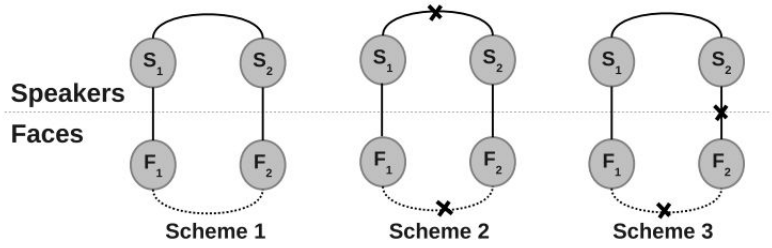


FIG. 3.9: Illustration des liens qu'il est possible de déduire entre une paire de visage (notés F1 et F2) à partir des associations avec les tours de parole (notés S1 et S2). Les liens solides représentent les correspondances observées. Les liens en points tillés sont les liens déduits. Par exemple, dans le schéma de gauche, le lien entre S1 et S2 indique qu'ils appartiennent à la même classe. Les liens S1-F1 et S2-F2 indiquent une association audiovisuelle. Ces correspondances permettent de déduire que les visages F1 et F2 devraient être regroupés dans la même classe. Image extraite de [Bendris 2014].

l'identification des visages avec un EGER de 43.8%, principalement grâce à un meilleur rappel. Les résultats des expériences montrent que l'identification des visages est plus difficile que celle des locuteurs. Ceci est dû à des visages non frontaux qui sont difficiles à détecter ou conduisant à l'extraction de descripteurs peu fiables. Ainsi, pour une précision autour de à 85-83%, le rappel de nommage des locuteurs est de 69% alors que celui des visages n'est que de 49%.

Le travail décrit dans [Bendris 2014] reprend l'idée d'utiliser des contraintes dérivées à partir des associations audio/vidéo et des noms, mais plutôt qu'un regroupement hiérarchique, c'est une formulation globale de type ILP qui est utilisée. Tout d'abord, un regroupement en locuteur est effectué et les noms présents dans les cartouches sont extraits. Ensuite, des liens d'association sont générés entre les tours de parole et les séquences de visage : Un visage est considéré comme lié au tour de parole courant si :

– l'activité des lèvres indique que le visage parle,

– un cartouche est présent et le visage est seul à l'écran.

De plus, un lien est créé entre deux tours de parole si ils sont associés au même nom.

Ces correspondances permettent de déduire de nouveaux liens (illustrés sur la figure 3.9) indiquant si deux visages doivent être regroupés dans la même classe ou non. Sur la partie TEST0 de la campagne REPERE, les performances obtenues pour l'identification des visages en terme d'EGER sont de 52.1%.

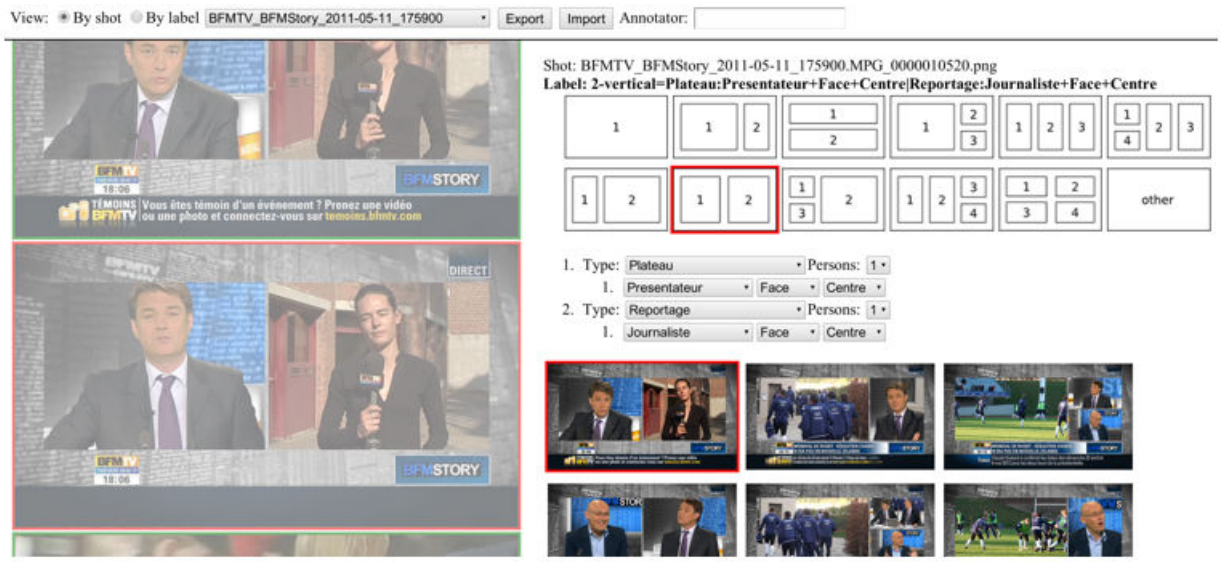


FIG. 3.10: Capture d'écran de l'interface utilisée pour réaliser les annotations. Pour chaque plan, la composition de l'image est annotée avec le nombre de personnes dans chaque cadre et le rôle qu'ils occupent. Image extraite de la page de Benoit Favre[2].

3.3.1.2 Compréhension multimodale des scènes

De manière originale, les auteurs de [Bechet 2014] explorent l'utilisation du contexte pour obtenir une meilleure compréhension des scènes et ainsi améliorer grandement l'identification des personnes présentes. Le système identifie premièrement les locuteurs, puis propage les noms des locuteurs vers les visages grâce à la modélisation de la scène et la compréhension multimodale. Outre les représentations monomodales, les caractéristiques incluent :

- une liste de personnes susceptible d'apparaître extraite de l'ensemble d'apprentissage. Cette liste est considérée comme une méta-donnée,
- l'attribution d'un rôle à chaque locuteur (présentateur, journaliste, invité...),
- une segmentation de la vidéo en chapitres à partir de la détection des logos des émissions,
- une caractérisation des scènes : type de plans (plateau, reportage, mixte, autre), rôle des visages, identifiant des caméras dans certains cas.

Un modèle spécifique est appris pour chaque émission. Pour apprendre ces modèles, des annotations ont été effectuées sur le corpus (voir figure 3.10). Dans le cas des débats, ce modèle inclut la position des caméras sur le plateau. Ceci permet de prédire automatiquement la présence de locuteurs à l'écran (voir figure 3.11), une fois l'identifiant de la caméra détectée.

[2] `http ://pageperso.lif.univ-mrs.fr/ benoit.favre/scene-description/`

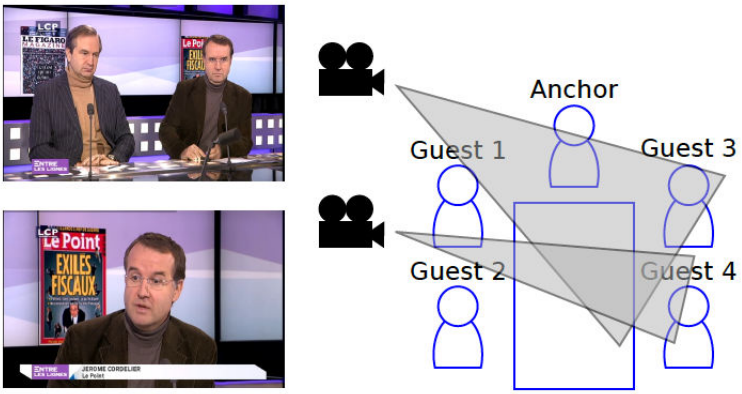


FIG. 3.11: Schéma illustrant la dépendance entre le type de caméra utilisé et les visages présents. Image extraite de [Rouvier 2014].

Comme le visage en lui même n'est pas pris en compte, cette méthode peut identifier des visages vus de dos, de profil ou occlus (10% des personnes), cas pour lesquels la détection même du visage est inopérante.

L'utilisation de ces caractéristiques permet à ce système d'obtenir les meilleures performances sur la partie PHASE2 de la campagne REPERE avec un taux d'EGER de 30.9% pour les locuteurs et 39.4% pour les visages. Cependant, il a été nécessaire d'apprendre un modèle à partir d'annotations pour les cas particuliers de chaque émission. De plus, il faut que le nom de l'émission à traiter soit connu *a priori* pour savoir quel modèle appliquer. Une extension directe de ce travail qui obtiendrait cette compréhension multimodale de manière automatique semble difficile à réaliser. Toutefois, elle demeure très prometteuse.

3.3.2 Application de l'identification des locuteurs pour un système d'indexation de l'actualité

Pour la dernière section de cet état de l'art, nous présentons le système d'identification décrit dans [Jou 2013]. Il fait partie intégrante d'un système de navigation dans des contenus de presse multimédia développé à l'université de Columbia et baptisé NewsRover. Ces contenus incluent des articles, des vidéos, et des messages de Twitter. Le but est de permettre à l'utilisateur de trouver facilement les réponses aux questions *Qui ?, Quoi ?, Quand ?, Où ?* pour un sujet d'actualité donné.

Concernant l'identification, un graphe est construit où les nœuds sont les tours de parole et les séquences de visage. Pour cela, des similarités audios, vidéos et audio-visuelles sont définies

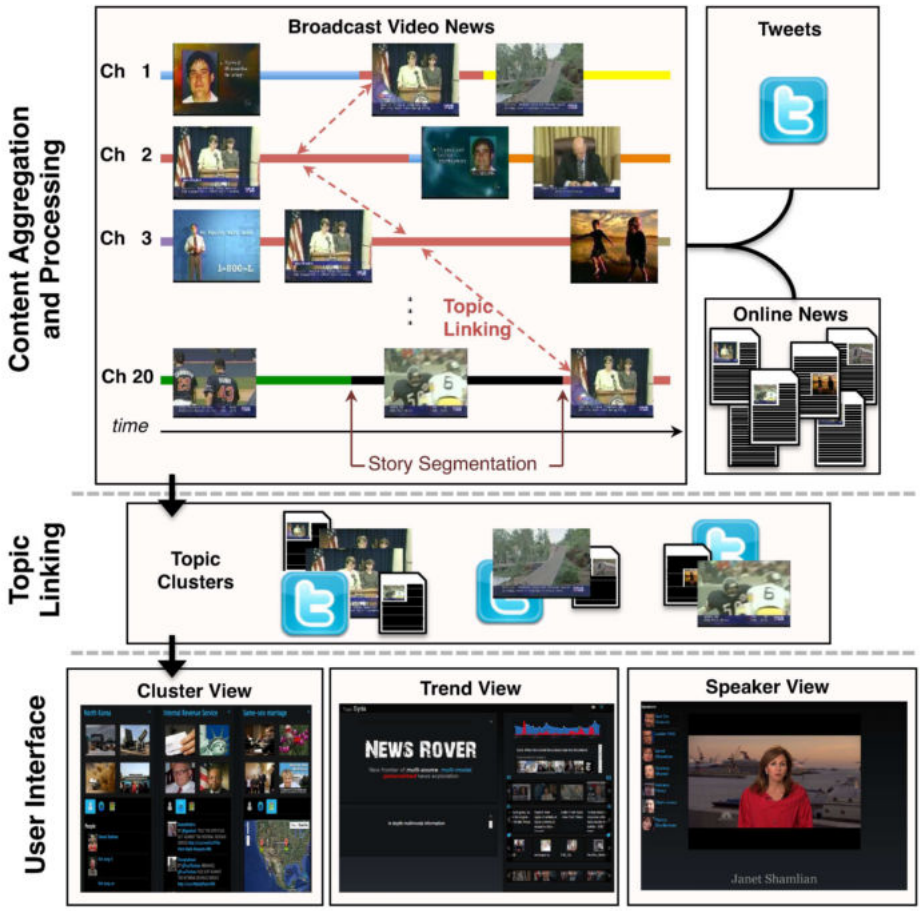


FIG. 3.12: Image extraite de [Jou 2013] présentant une vue globale du système.

entre ces nœuds. Les noms sont extraits des sous-titres et des cartouches et représentent les étiquettes. Un algorithme d'apprentissage semi-supervisé est utilisé pour assigner une étiquette à chaque tour de parole.

Combinée à la transcription et à une structuration automatique par sujet, l'identification permet alors de répondre à la question *Qui a dit quoi ?* et de parcourir un événement en fonction de ses différents acteurs ainsi qu'il est illustré sur la figure 3.12.

3.4 Conclusions

Dans ce deuxième chapitre, les techniques d'identification non supervisées des locuteurs et des visages ont été passées en revue. De cet état de l'art ressortent plusieurs éléments.

Tout d'abord, en ce qui concerne les sources de nommage, on peut voir que lorsque les cartouches sont disponibles, ils sont utilisés en priorité et considérés comme la source d'information la plus fiable (à condition que la résolution des images soit suffisante). L'utilisation de la transcription peut permettre d'augmenter le rappel mais elle est plus difficile à manier. En particulier, dans le cas des visages qui ne sont pas des locuteurs, une expérience décrite dans [Bechet 2012] montre que cette tâche est difficile même pour des humains. En effet, deux fois sur trois, l'annotateur estime qu'il ne sait pas si le visage d'une personne nommée est présent.

Enfin, de nombreux travaux utilisent avec succès les sous-titres fournissant une transcription manuelle de qui dit quoi dans la vidéo. Cependant, ces derniers ne sont pas toujours disponibles, et c'est précisément ce cadre dans lequel nous nous plaçons.

Les sources de nommage comme la transcription et les cartouches sont par nature locales en ce sens qu'elle permettent d'annoncer le visage courant ou le locuteur courant. Afin de nommer toutes les apparitions et prises de parole d'une personne, la propagation du nom vers les autres parties de la vidéo dépend de la qualité du regroupement des locuteurs et des visages. Nous avons vu que la réalisation conjointe des opérations de nommage et de regroupement peut s'avérer bénéfique.

L'identification des visages est généralement plus difficile que celle des locuteurs. D'une part, l'association est plus ambiguë avec les cas où plusieurs personnes apparaissent sur la même image. D'autre part, les détecteurs de visages peuvent échouer dans les cas de profil ou de visages vus de dos générant ainsi un faible rappel.

La compréhension du contexte de la scène est très utile pour le nommage des émissions télévisées et peut pallier les difficultés inhérentes à la détection, au regroupement monomodal et à l'association nom/visage. En suivant cette voie, le prochain défi sera d'arriver à construire les modèles de contexte de manière peu onéreuse, c'est à dire en évitant d'avoir recours à des annotations. Pour cela un point important est de trouver des modélisations généralisables à l'ensemble des émissions à traiter.

Deuxième partie

Contributions : Modèles CRF pour le regroupement et l'identification de personnes dans les journaux télévisés

Chapitre 4

Représentation de visages pour le regroupement

Sommaire

Le système de regroupement des visages qui va être présenté dans ce chapitre a été publié dans [Khoury 2013] et a été développé au début de cette thèse principalement par Elie Khoury. Il s'inspire fortement de ses travaux de thèse. Plus précisément, cette partie décrit une amélioration de la représentation de visage utilisée dans l'étape finale du regroupement. Cette représentation est une combinaison de deux représentations complémentaires.

La première de ces deux représentations est fondée sur les descripteurs *Speeded Up Robust Features* (SURF) [Bay 2006] qui sont capables de regrouper avec précision des visages provenant du même contexte. Par exemple, le regroupement sera correct pour des visages provenant de la même scène ou du même épisode d'une série. Cependant, quand les variabilités augmentent à cause de changement de pose, d'illumination ou de la coupe de cheveux, le pouvoir discriminant de ces descripteurs diminue au fur et à mesure que les distances entre deux visages d'une même personne deviennent aussi grandes que celles de deux visages venant de personnes différentes. En bref, ces approches n'ont pas une capacité de généralisation suffisante.

Pour résoudre ce problème de généralisation, il est proposé ici d'utiliser des modèles biométriques [Wallace 2012] dont le but est spécifiquement de gérer ce type de variations. Plutôt que de comparer directement une paire de visages par leurs descripteurs, chaque classe de visage est représentée par un modèle statistique dont le calcul des paramètres utilise, en plus des visages rattachés à cette classe, des *a priori* sous formes de statistiques calculées sur des milliers d'autres visages grâce à la technique de l'adaptation *Maximum a Posteriori* (MAP). Bien que l'utilisation de modèles statistiques soit parfois moins précise que la comparaison directe de descripteurs, spécialement si il y a peu de données pour apprendre les paramètres, les variations sont mieux gérées.

Le reste du chapitre est organisé comme suit : la section 4.1 présente le système dans sa globalité avec les étapes préliminaires de détections de visage et de suivi. La section 4.2 décrit la méthode, et la section 4.3 compare ses performances avec d'autres méthodes de l'état de l'art sur le corpus public de la série *Buffy The Vampire Slayer*. Enfin, une évaluation est effectuée sur les données REPERE.

4.1 Description globale de la chaîne de traitement

Les différentes étapes sont résumées sur la figure 4.1 et sont décrites brièvement ci-après.

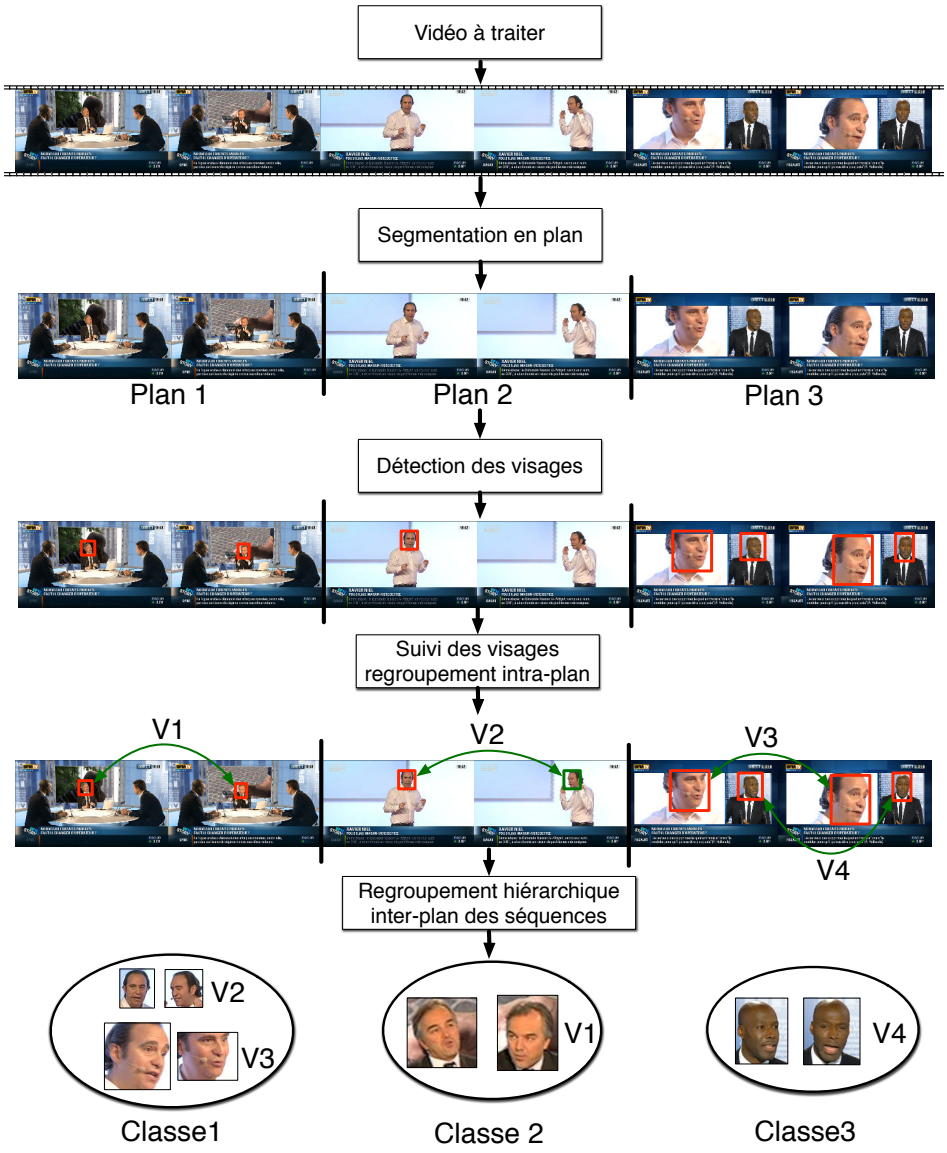


FIG. 4.1: Les différentes étapes du système de regroupement en classes de visages développé à l'Idiap Research Institute et décrit dans [Khoury 2013].

– **Segmentation en plans :** cette segmentation décrite dans [Khoury 2010] est un prétraitement qui facilite le suivi de visages dans l'étape suivante. En effet, chaque plan est considéré comme une unité homogène quant aux personnes présentes.

– **Détection des visages :** la détection utilise l'algorithme de détection frontale de Viola& Jones [Viola 2004]. Il est appliqué à chaque image. Des filtres basés sur la couleur de peau sont appliqués pour retirer les fausses alarmes.

– **Suivi des visages :** cette opération consiste à associer les visages successivement détectés à l'intérieur d'un même plan en séquences de visages. De nouvelles détections sont

éventuellement ajoutées en cherchant dans les frames voisines des régions visuellement similaires aux détections déjà obtenues.

– **Regroupement :** un algorithme de Classification hiérarchique Ascendante (CAH) des séquences obtenues précédemment est effectué en combinant deux représentations de visage. La représentation de visage utilisée dans cette étape correspond à la contribution présentée dans ce chapitre. Elle est détaillée dans la section suivante.

4.2 Combinaison d'une représentation de visage par un descripteur et un modèle statistique

L'hypothèse sous-jacente à l'approche proposée est que la représentation optimale dépend de la variabilité des données et de la quantité de données disponible pour apprendre les paramètres de cette représentation. D'un côté, la comparaison directe de descripteurs locaux est optimale si les variabilités entre visages sont faibles (moments proches dans le temps, même scène). Elle a aussi l'avantage de fonctionner même avec des classes de petite taille car elle ne nécessite pas d'apprendre un modèle. Du point de vue d'un regroupement hiérarchique, elle est plus adaptée au début de l'algorithme. D'un autre côté, les modèles statistiques sont optimaux quand une certaine quantité de données par classe est disponible et quand les variabilités sont importantes grâce à leur relative insensibilité à ces facteurs. Son utilisation dans le regroupement hiérarchique est donc plus adéquate tardivement. Il est proposé ici de combiner ces deux représentations en fonction de ces propriétés.

4.2.1 Comparaison directe de descripteurs locaux

Descripteurs SURF

Comme annoncé dans l'introduction du chapitre, des descripteurs SURF sont utilisés. Chaque descripteur SURF est calculé localement autour d'un point d'intérêt. Ils sont connus pour être peu sensibles aux variations d'échelle, de rotation, et d'illumination. Ils ont été initialement développés pour l'indexation visuelle d'images en général. Ils ont aussi montré leur intérêt pour le regroupement de visages [Zhao 2008] grâce à leur capacité à comparer avec précision deux visages acquis dans un même contexte. Pour un visage donné F, la première représentation est donnée par les descripteurs SURF associés : $Surf(F) = \{f_i^{surf}, i = 1 \ldots, N_F^{surf}\}$.

Comparaison de deux visages

La distance entre deux visages F_1 et F_2 que nous utilisons est la *Average N-Minimal Pair Distance* (ANMPD) qui a été proposée dans [El Khoury 2010b]. Elle sera notée $d_s(F_1, F_2)$. Comme son nom le suggère, cette distance retourne la moyenne des N plus petites distances choisies parmi les distances entre toutes les paires p de vecteur SURF possibles entre les deux visages. Soit D_{p_i} la $i^{ème}$ plus petite distance, la distance entre F_1 et F_2 s'écrit alors :

$$d_s(F_1, F_2) = \frac{1}{N^{anmpd}} \sum_{i}^{N^{anmpd}} D_{p_i} \tag{4.1}$$

Les deux visages sont donc considérés comme similaires si la distance entre eux est faible. Notons que le fait de sélectionner uniquement les distances les plus faibles par rapport à l'utilisation de la moyenne donne implicitement de la robustesse aux variations de pose. Dans nos expériences, nous avons utilisé $N^{anmpd} = 6$, ce qui correspond à la valeur qu'Elie Khoury avait sélectionné pendant sa thèse.

Comparaison de deux classes de visages

Soient deux classes de visages C_i et C_j avec leurs ensembles de visages $\{F_a^i, a = 1, \ldots, N_i\}$ et $\{F_a^j, a = 1, \ldots, N_j\}$. Elles sont comparées avec la distance suivante :

$$D_f(C_i, C_j) = \frac{1}{N_i N_j} \sum_{a=1}^{N_i} \sum_{b=1}^{N_j} d_s(F_a^i, F_b^j) \tag{4.2}$$

Par définition, cette mesure favorise la création de classes compactes où tous les visages sont comparés les uns aux autres. Si après une itération du regroupement hiérarchique, C_i et C_j sont fusionnées dans la classe C_i', la distance entre C_i' et une autre classe C_k est calculée par :

$$D_f(C_i', C_k) = \frac{N_i \times D_f(C_i, C_k) + N_j \times D_f(C_j, C_k)}{N_i + N_j} \tag{4.3}$$

4.2.2 Approche biométrique avec un modèle statistique

Modélisation de la distribution des descripteurs denses

Détecter des points d'intérêt est intéressant pour rechercher des correspondances. Malheureusement, les localisations de ces points ne contiennent pas d'information sémantique sur leur position dans le visage. Une alternative est d'extraire les descripteurs en des points du visage prédéfinis (les yeux, la bouche…), qui sont connus sous le nom de *Facial landmarks* (voir

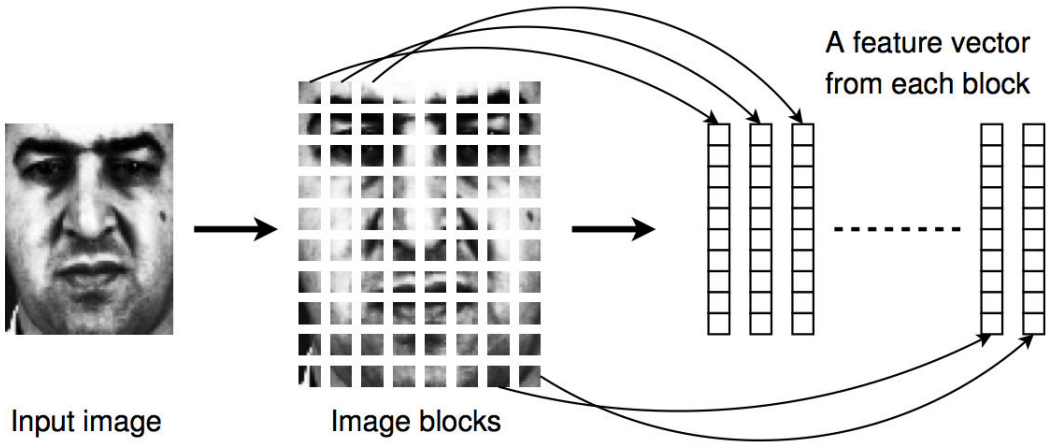


FIG. 4.2: Illustration du processus d'extraction des coefficients DCT par bloc densément échantillonnés. Dans la pratique les blocs se chevauchent plutôt que d'être côte à côte comme sur la figure. Image extraite de [Wallace 2012]

figure 2.3). Cependant, localiser ces points n'est pas toujours trivial dans le cas de visages qui ne sont pas parfaitement frontaux ou avec une faible résolution. Les descripteurs extraits peuvent alors être affectés par la précision de la détection. Une solution à ce problème consiste à extraire les caractéristiques densément sur une grille d'échantillonnage. En biométrie, l'utilisation de descripteurs extraits densément en combinaison avec des modèles statistiques a montré de bonnes performances, même si la localisation de la position du visage n'est pas parfaite [Wallace 2012]. Nous choisissons une stratégie similaire pour notre seconde représentation. Pour chaque visage F, l'ensemble des descripteurs DCT est noté $Dct(F) = \{f_i^{dct}, i = 1, \ldots, N^{dct}\}$. Étant donnée l'image du visage, la position des yeux est d'abord estimée. À partir de cette position, la taille du visage est normalisée pour former une image de 80 pixels par 64. Cette image est prétraitée par la méthode de *Tan and Triggs* [Tan 2007] qui permet de réduire les variations d'illumination. Ensuite, la Transformée en Cosinus Discret (DCT) est appliquée sur des blocs carrés de 8 pixels de côté extraits densément sur l'image. Les blocs sont extraits avec un pas de un pixel et se chevauchent donc sur 7 pixels. Pour chaque bloc, seuls les 28 premiers coefficients de la transformée sont conservés. La figure 4.2 résume l'ensemble du processus d'extraction de ces coefficients. Ces différents choix de paramètres correspondent à ceux fournis dans la publication [Wallace 2012].

Les modèles statistiques sont des outils puissants pour représenter la variabilité des visages. Dans ce travail, nous utilisons des modèles GMM qui ont déjà été utilisés dans ce contexte et qui autorisent l'estimation de paramètres par la technique du *Maximum a Posteriori*. La

vraisemblance de chaque vecteur de caractéristiques DCT est exprimée par :

$$p(f^{dct}|\Lambda) = \sum_{i=1}^{N_g} \omega_i \mathcal{N}(f^{dct}; \mu_i, \Sigma_i) \tag{4.4}$$

où $\Lambda = \{\omega_i, \mu_i, \Sigma_i, i = 1...N_g\}$ représente les paramètres du modèle GMM (dans ce travail $N_g = 200$). Pour une classe C_i, contenant les visages $\{F_a^i = a = 1 \dots N_i\}$, l'ensemble des caractéristiques est défini comme l'union des caractéristiques extraites sur tous les visages $Dct(C_i) = \cup_a Dct(F_i^a)$. La vraisemblance des données de la classe pour un modèle Λ est alors :

$$L(Dct(C_i|\Lambda)) = \prod_{f^{dct} \in Dct(C_i)} p(f^{dct}|\Lambda) \tag{4.5}$$

Dans la pratique, plutôt que par maximum de vraisemblance, les paramètres du GMM sont appris par adaptation MAP des paramètres d'un modèle du monde généralement appelé *Universal Background Model* (UBM). L'UBM est un GMM appris sur un grand nombre de visages venant d'un corpus d'apprentissage. L'adaptation permet d'éviter un sur-apprentissage potentiel compte tenu du grand nombre de paramètres et du nombre potentiellement faible de visages de chaque classe. Initialement, cette technique a été introduite pour la reconnaissance du locuteur dans [Reynolds 2000].

Similarité entre deux clusters : *Cross-likelihood ratio*

La comparaison de deux classes C_i et C_j étant donnés leurs modèles respectifs Λ_i et Λ_j est fondée sur le *Cross-likelihood ratio* (CLR) défini ainsi :

$$CLR(C_i, C_j) = log\frac{L(Dct(C_i)|\Lambda_j)}{L(Dct(C_i)|\Lambda_{ubm})} + log\frac{L(Dct(C_j)|\Lambda_i)}{L(Dct(C_j)|\Lambda_{ubm})} \tag{4.6}$$

Le CLR est une mesure de similarité symétrique qui est élevée si deux classes sont similaires. Elle capture à quel point les données d'une classe sont bien représentées par le modèle de l'autre classe. Cette vraisemblance est normalisée par la vraisemblance calculée à partir du modèle UBM qui sert alors de référence. Cette normalisation permet de distinguer les cas où la vraisemblance est faible parce que le modèle explique mal les données des cas où cette faiblesse est due à un bruit propre aux données. Cette mesure a déjà été utilisée dans [Barras 2006] pour le regroupement en locuteur.

Au début de leur utilisation dans le regroupement hiérarchique, la similarité $S_m(C_i, C_j)$ pour chaque paire de classes C_i et C_j est fixée à $S_m(C_i, C_j) = CLR(C_i, C_j)$. Ensuite, si après une itération du regroupement hiérarchique, C_i et C_j sont fusionnées dans la classe C_i', la distance

entre C'_i et une autre classe C_k est calculée par :

$$S_m(C'_i, C_k) = \frac{N_i \times S_m(C_i, C_k) + N_j \times S_m(C_j, C_k)}{N_i + N_j} \tag{4.7}$$

Une autre alternative serait d'apprendre un nouveau modèle $\lambda_{i'}$ à partir des données associées à la classe C'_i. Cependant, il est possible que ce modèle soit contaminé par des données de plusieurs personnes, par exemple si le regroupement a réalisé une fusion erronée. Avec la fusion des scores proposée, différents modèles appris sur des classes pures sont conservés. Il a été observé expérimentalement que cette dernière méthode permet d'obtenir de meilleures résultats.

4.2.3 Combinaison des deux représentations

La distance à base de descripteurs SURF D_f et la similarité fondée sur les modèles GMM S_m satisfont différents objectifs. La première est efficace pour comparer avec précision des visages provenant de contextes similaires (même scène ou même épisode). La deuxième nécessite des classes contenant d'avantage de données afin apprendre des modèles GMM représentatifs, mais elle modélise mieux les variabilités, parfois au prix d'une précision plus faible. Dans la perspective d'un regroupement hiérarchique, la première est plus adaptée au début de l'algorithme, quand les classes sont petites, alors que la seconde peut être appliquée plus tard. Nous adoptons donc la stratégie suivante :

1. Effectuer le regroupement seulement avec la distance D_f.
2. Une fois qu'un seuil est atteint, c'est à dire que $\forall(i,j), D_f(C_i, C_j) > T_f$, la distance suivante est utilisée :

$$D_c(C_i, C_j) = D_f(C_i, C_j) - \alpha S_m(C_i, C_j) \tag{4.8}$$

où $\alpha > 0$ est un paramètre de pondération entre les deux mesures.

4.3 Expériences d'évaluation

4.3.1 Evaluation sur la série *Buffy the Vampire Slayer*

Présentation des données Buffy

Notre avons évalué notre système sur les données annotées de la série *Buffy the vampire Slayer* fournies par les auteurs de [Guillaumin 2009]. Ce premier corpus a aussi été utilisé dans [Cinbis 2011]. Il contient 327 séquences de visages sélectionnées des épisodes 9, 21 et 45.

Chaque épisode appartient à une saison différente pour introduire d'avantage de variabilités. Ce corpus contient donc généralement une plus grande variabilité visuelle d'un épisode à l'autre qu'à l'intérieur d'un même épisode (voir figure 4.3). Les séquences ont été obtenues avec un système automatique et les faux positifs ainsi que les visages n'appartenant pas à l'un des 8 acteurs principaux ont été retirés.

FIG. 4.3: Visages détectés pour le personnage "Joyce" : les 3 exemples de gauche sont extraits de l'épisode 9, et les trois de droite sont extraits de l'épisode 21. Il est visible que la variabilité de l'apparence entre épisodes est plus grande que la variabilité au sein d'un même épisode.

Mesure d'évaluation : le nombre de clics

Pour être comparable à l'état de l'art, l'évaluation utilise la métrique proposée dans [Guillaumin 2009] et ré-utilisée dans [Cinbis 2011]. Cette métrique calcule le nombre de clics de souris nécessaire pour corriger manuellement la sortie automatique du regroupement. Plus précisément, il est supposé qu'il faut un clic pour associer le nom correct à une classe et un clic pour donner le bon nom à une séquence de visages mal classée.

Résultats

Les résultats de trois méthodes de l'état de l'art ont été rapportés pour ce corpus dans [Guillaumin 2009, Cinbis 2011] et sont affichés sur la figure 4.4. Elles sont toutes les trois fondées sur un regroupement hiérarchique ascendant et l'utilisation d'un ensemble de descripteurs SIFT calculés autour des *facial landmarks*. La différence entre chaque approche est la métrique utilisée pour comparer deux visages.

- La courbe notée $L2$ sur la figure correspond à la distance euclidienne.
- La courbe LFW correspond à une distance de Mahalonobis apprise sur le corpus *Labeled Face in the Wild* (voir section 2.2.1, page 21).
- La courbe UML (*Unsupervised Metric Learning*) est une métrique apprise en optimisant une fonction de coût discriminante de manière non supervisée. Les annotations nécessaires à l'apprentissage de la métrique sont obtenues ainsi : deux visages de la même séquence forment un exemple positif, et deux visages de séquences différentes mais provenant de la même image forment un exemple négatif.

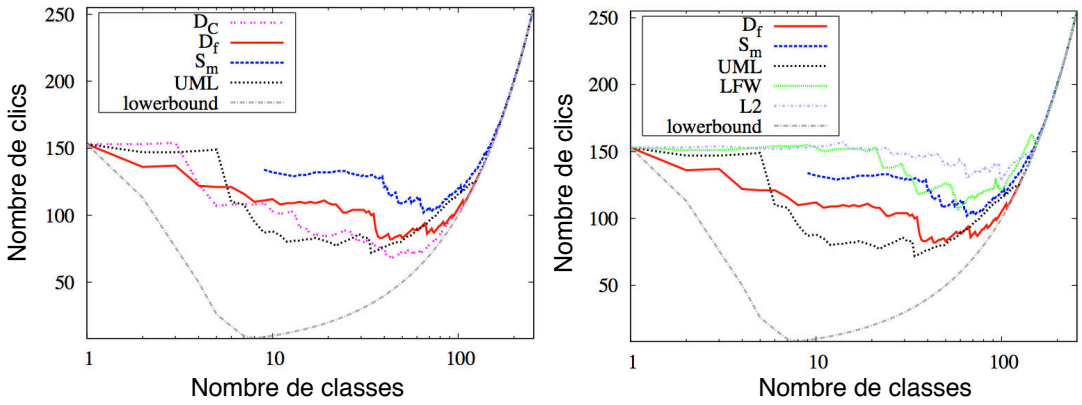


FIG. 4.4: Performances en nombre de clics en fonction du nombre de classes. Le nombre de classes diminue à cause des fusions opérées à chaque étape par le regroupement hiérarchique. La figure se lit donc chronologiquement de droite à gauche. Pour une meilleure visualisation, l'échelle logarithmique est utilisée. La figure de gauche compare les deux mesures présentées dans nos travaux avec leur combinaison et la métrique UML. La figure de droite compare ces deux mesures à trois approches de l'état de l'art présentées dans [Guillaumin 2009, Cinbis 2011].

La figure présente également les résultats obtenus avec les deux mesures S_m et D_f décrites précédemment et leur combinaison D_C. La courbe *lowerbound* représente la meilleure performance possible.

Les courbes montrent l'évolution du nombre de clics en fonction du nombre de classes. Premièrement, la méthode à base de descripteurs locaux D_f obtient de meilleures performances que la méthode fondée sur les modèles statistiques S_m. C'est probablement du au manque de données d'apprentissage au début du regroupement. Ensuite, la distance D_f obtient de meilleures performances par rapport à la méthode UML au début du regroupement jusqu'à un nombre de classes plus grand que 60. Par la suite, la méthode UML devient la meilleure. Ceci accrédite l'hypothèse que la comparaison directe de descripteurs locaux est capable de regrouper avec précision des visages issus des mêmes conditions visuelles, mais n'est plus suffisante quand la variabilité augmente. La combinaison D_C est meilleure que chacune des deux approches D_f et S_m. Elle est aussi comparable à l'approche UML. Plus précisément, le nombre minimum de clics atteint par les deux courbes est de 72 pour UML et 68 pour D_C.

Illustration du comportement de nos deux représentations

Afin d'illustrer les différences de comportement de nos deux mesures, les distances D_f et $(-S_m)$ ont été calculées pour toutes les paires de visages inter et intra-épisodes du personnage principal (Joyce). Les histogrammes cumulés des distances ont été affichés sur la figure 4.5. Qualitativement, il est visible que la distance D_f est plus adaptée que la similarité S_m pour

comparer les visages à l'intérieur d'un épisode. Cette remarque est illustrée par le fait que 26% des distances intra-épisodes sont plus faibles que les distances inter-épisodes. En comparaison, seulement 12% des similarités S_m intra-épisodes sont plus élevées que toutes les similarités inter-épisodes.

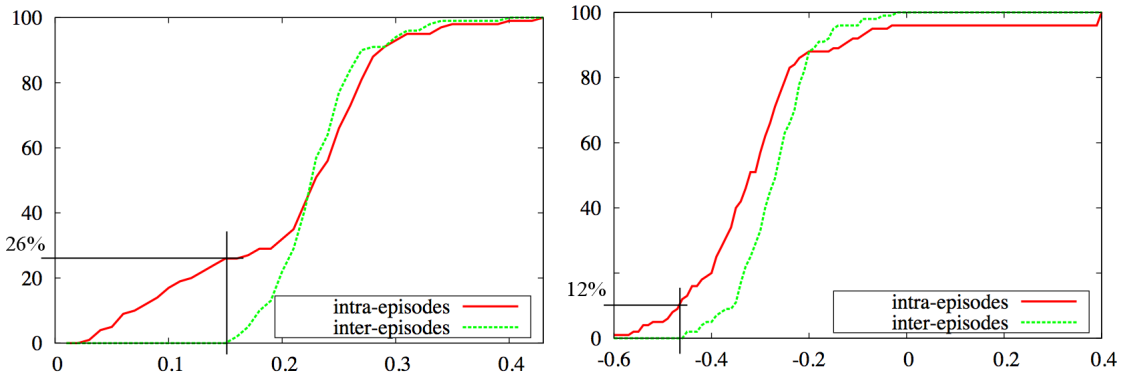


FIG. 4.5: Histogramme cumulé des distances entre les paires de visages du personnage Joyce pour la distance D_f à gauche et la distance $(-S_m)$ à droite. Les distributions des scores entre les paires intra et inter-épisodes illustrent notre hypothèse sur les propriétés de nos deux mesures.

4.3.2 Evaluation sur les données REPERE

Au moment où ces expériences ont été effectuées, seule la partie *dry-run* du corpus REPERE était disponible (les données de l'évaluation finale de la campagne correspondant à la partie PHASE2 seront utilisées dans le chapitre suivant). Elle consiste en 38 vidéos extraites des mêmes émissions que pour le reste du corpus et est divisée en une partie développement (DEV0) de 28 vidéos et une partie test (TEST0) de 10 vidéos. Les durées ont été reportées dans le tableau 1.1.

Les séquences utilisées ont été obtenues à partir du système automatique décrit dans la section 4.1. Comme pour le corpus *Buffy*, les fausses alarmes ont été filtrées afin d'évaluer uniquement la tâche de regroupement. C'est à dire que les erreurs de fausse alarme et de non-détection ne sont pas comptabilisées, le DER ici ne compte donc que la partie de confusion. Après le nettoyage, 1076 séquences appartenant à 264 personnes sont disponibles. Les résultats sont reportés dans le tableau 4.1.

Les algorithmes utilisent différents paramètres : pour l'utilisation des distances $(-S_m)$ et D_f seules, ce sont les seuils d'arrêt du regroupement T_f et T_m. Pour l'approche combinée, il y a le coefficient α pondérant la contribution de chaque distance, et le seuil d'arrêt T_C. Ces paramètres sont optimisés sur le corpus de développement DEV0.

Méthod	cross-DER(DEV0)	min-DER (DEV0)	DER (TEST0)
D_f	6.41%	5.13%	8.33%
S_m	6.68%	6.68%	8.21%
D_C	**5.49%**	**4.37%**	**5.28**

TAB. 4.1: Évaluation du regroupement en classes de visages sur les données REPERE en comparant les deux mesures et leur combinaison.

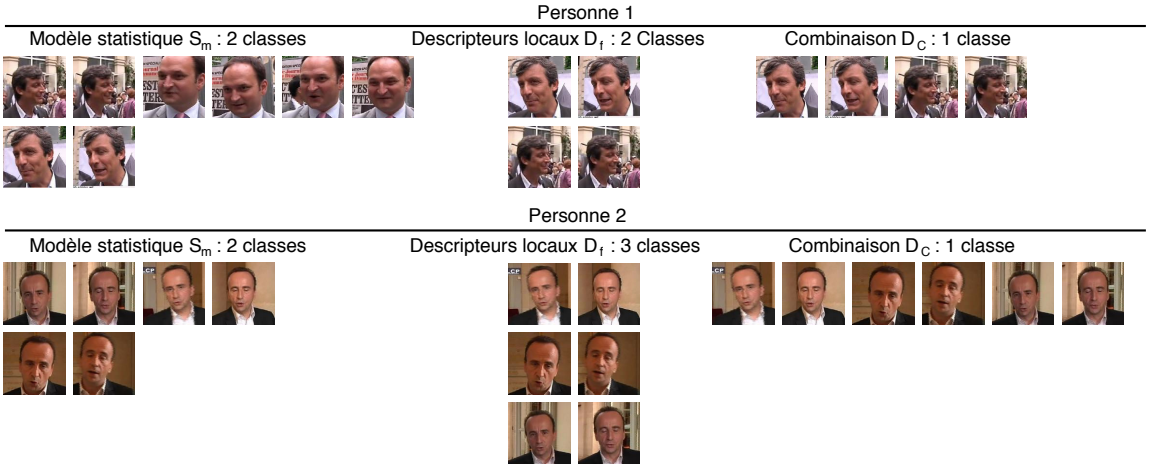


FIG. 4.6: Illustration des sorties obtenues par chaque méthode S_m, D_f et D_C pour deux personnes. Chaque ligne correspond à une classe.

Le résultat **min-DER** correspond au meilleur DER atteint sur le DEV0, et **cross-DER** correspond au DER sur le DEV0 en optimisant les paramètres par validation croisée. La comparaison entre ces deux colonnes donne une idée de la stabilité des performances par rapport à des variations des paramètres. Il est visible que la similarité S_m est plus stable que la distance D_f. C'est peut être dû à l'utilisation de l'*a priori* venant de l'UBM. La combinaison des deux similarités est meilleure que chaque similarité prise séparément sur les corpus DEV0 (14% de gain relatif par rapport au **cross-DER**) et TEST0 (35% de gain relatif).

La figure 4.6 représente les sorties des 3 approches pour deux personnes. Sur ces exemples, les méthodes S_m et D_f créent plus de classes qu'il n'y a de personnes. Pour la personne 1, la méthode S_m fait en plus une erreur de confusion. Cependant, la combinaison des deux méthodes permet de corriger ces erreurs.

FIG. 4.7: Illustration des erreurs de confusion restantes du regroupement vidéo monomodal avec la méthode D_c où quatre personnes sont séparées en plusieurs classes. Pour chaque personne, chaque ligne correspond à des exemples extraits d'une classe. L'exemple le plus à gauche correspond à un reportage, les trois autres sont extraits de scènes de plateau.

Limites de notre système de regroupement dans le cas de REPERE.

Plus généralement, même avec la méthode D_C le système a tendance à créer des classes pures au risque de diviser une personne en plusieurs classes où chacune correspondra à un contexte d'illumination ou une pose de visage particulière. La majorité des erreurs survenant dans les reportages proviennent de plans qui illustrent le sujet. Ils sont donc filmés dans différents lieux avec à chaque fois des conditions d'illumination et de résolution qui leur sont propres comme illustré dans l'exemple de gauche de la figure 4.7. Pour les émissions avec plateau, les conditions d'illumination sont homogènes. Cependant, il existe des variations de pose et d'échelle dues à l'utilisation de plusieurs caméras et au fait qu'une personne se tourne pour parler à ses différents interlocuteurs. De par la composition des données REPERE, le deuxième type d'erreurs survenant sur les scènes de plateaux est plus fréquent et concerne des visages qui sont aussi des locuteurs. Nous verrons dans le chapitre 6 comment corriger ces erreurs, notamment avec l'utilisation de l'arrière-plan.

4.4 Conclusions

Dans ce chapitre, nous avons présenté une nouvelle mesure de similarité pour le regroupement en classes de visages. Elle est fondée sur une analyse des propriétés des descripteurs

SURF et d'une approche utilisée en biométrie fondée sur des modèles statistiques. La comparaison directe de descripteurs est optimale pour décrire des images de visages proches dans le temps et dans l'espace alors que l'approche biométrique est plus efficace pour modéliser des variations plus importantes. La combinaison des deux approches utilisée dans un regroupement hiérarchique nous permet d'obtenir de meilleurs résultats qu'avec chacune des deux approches prises séparément.

Ce chapitre nous a également permis d'étudier les erreurs que peut commettre notre système sur les données REPERE, notamment, qu'une personne peut être divisée en plusieurs sous-classes. Nous avons distingué deux situations : la première correspond à des reportages où la variation inter-visages est élevée. La deuxième correspond à des scènes de plateau où le principal facteur de variation vient de l'expression et de la pose.

Le cadre expérimental utilise des détections de visages "propres" car les fausses alarmes ont été manuellement retirées. Il semble intéressant d'étudier les performances avec une chaîne de traitement entièrement automatique. Je suppose que les systèmes éviteront de confondre les fausses alarmes avec les visages d'autres personnes. En revanche, les méthodes seront beaucoup plus sensibles à des visages détectés dans des poses de profils ou masquées par d'autres éléments (vêtements, autre personnes). Bien qu'ils ont été ignorés dans notre évaluation, ces visages représentent une part non-négligeable des apparitions de personnes.

Chapitre 5

Regroupement joint des visages et des locuteurs

Sommaire

Ce chapitre présente le modèle CRF proposé pour le regroupement audiovisuel des personnes. Les travaux cités dans l'état de l'art suggèrent que les performances sont meilleures quand cette tâche est effectuée de manière jointe sur les segments audios et vidéos : l'association audiovisuelle permet de guider les regroupements monomodaux dans des cas difficiles à traiter si une seule modalité est utilisée.

Notre modèle suit une stratégie similaire. Le regroupement des segments est optimisé globalement au sein d'un CRF qui prend en compte i) les liens d'association entre les séquences de visage et les tours de parole ii) la similarité monomodale entre un segment et sa classe qui est calculée avec un modèle biométrique. Les expériences sur les données REPERE permettent de valider la capacité de ce modèle à intégrer des informations multimodales. Par ailleurs, elles permettent d'étudier l'intérêt d'effectuer le regroupement de manière jointe compte tenu des récents progrès de l'état de l'art sur les systèmes de regroupement monomodaux. Après une brève introduction sur les CRF, la section 5.1 présente la formulation du modèle. Ensuite, chaque partie du modèle est détaillée. La section 5.2 contient les expériences sur l'association et le regroupement audiovisuel des personnes.

5.1 Modèle de regroupement audiovisuel des personnes

5.1.1 Introduction aux Champs Conditionnels Aléatoires (CRF)

Cette introduction aux modèles graphiques et aux CRF a pour but de faciliter la description des modèles présentés dans cette thèse. Elle s'inspire principalement du tutoriel écrit dans [Sutton 2006] et de sa version étendue [Sutton 2010]. Un excellent chapitre d'introduction aux modèles graphiques se trouve dans [Bishop 2006].

Supposons qu'il faille prédire un vecteur y à partir d'un vecteur d'observations x. L'une des difficultés est de modéliser les dépendances entre les différentes variables. Les modèles graphiques représentent une distribution probabiliste de ces variables sous la forme d'un produit de facteurs locaux. Chaque facteur opère sur un sous-ensemble de variables. Par exemple, supposons que $y = \{y_1, y_2\}$ et $x = \{x_1, x_2\}$. Nous avons donc 4 variables : y_1, y_2, x_1, x_2. La distribution de ces quatre variables peut s'écrire comme le produit de facteurs :

$$p(y, x) = \Psi_1(y_1)\Psi_2(y_1, x_1)\Psi_3(y_1, y_2)\Psi_4(y_2)\Psi_5(y_2, x_2) \tag{5.1}$$

Le choix des variables que contient chaque facteur permet de modéliser les dépendances entre elles. Dans la pratique ce choix est souvent un compromis : un modèle prenant en compte de

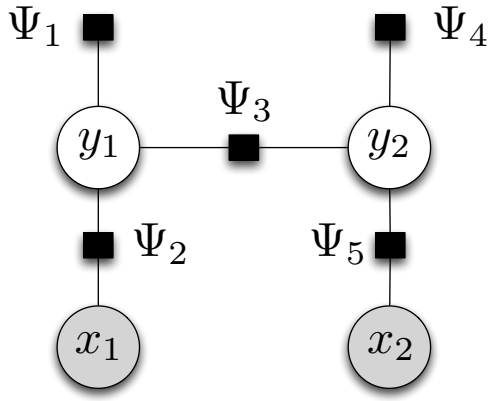


FIG. 5.1: Graphe de facteurs.

nombreuses dépendances peut potentiellement faire de meilleures prédictions, mais cela peut le rendre plus difficile à utiliser et à optimiser.

Chaque distribution peut être formellement décrite avec un graphe de facteurs. Le graphe de facteurs correspondant à la distribution de l'équation 5.1 est représenté sur la figure 5.1. Dans la terminologie usuelle, les variables dont on cherche à prédire la valeur sont appelées les variables cachées, par opposition aux variables observées. Pour les distinguer sur le graphe, les variables observées sont représentées par des nœuds grisés. Afin d'estimer la valeur des variables cachées, on cherche généralement à estimer la valeur y^{opt} qui maximise la vraisemblance. Cette opération s'appelle le décodage :

$$y^{opt} = \arg\max_{y} p(y, x) \tag{5.2}$$

Nous avons déjà vu l'utilisation de modèles graphiques dans l'état de l'art avec [Noulas 2012, Ma 2007, Du 2012]. Vu leur capacité à modéliser les interdépendances entre variables, ils sont particulièrement utiles pour la modélisation du contexte.

Les CRF sont un cas particulier de modèles graphiques. La définition d'un CRF est la suivante : soient G un graphe de facteurs sur des vecteurs de variables aléatoires dont les réalisations sont notées x et y et $\lambda = \{\lambda_k\} \in R^K$ un vecteur de K paramètres. Alors, la distribution $p(y|x)$ est un CRF si quelles que soient les réalisations x et y, elle factorise selon G. Soit $F = \{\Psi_a\}$ l'ensemble des facteurs. La distribution conditionnelle du CRF s'écrit alors :

$$p(y|x) = \frac{1}{Z(x)} \prod_{a=1}^{A} \Psi_a(y_a, x_a) \tag{5.3}$$

dans laquelle :

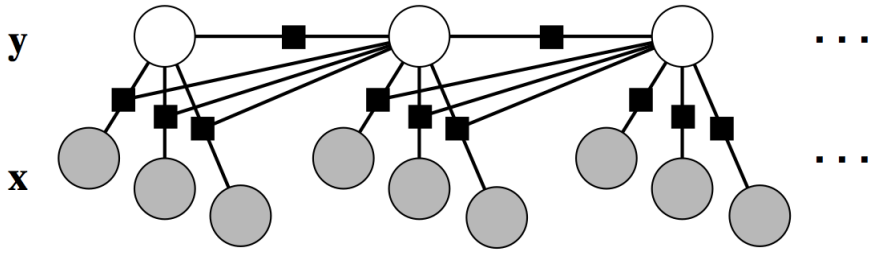


FIG. 5.2: Graphe de facteurs d'un CRF linéaire. Dans cette vue, la variable cachée de chaque tranche du modèle correspond à l'état d'un mot dans la séquence. Image extraite de [Sutton 2006].

- x représente les observations et y l'ensemble des variables cachées dont on souhaite prédire la valeur.
- A est le nombre de facteurs dans le graphe et $\{y_a, x_a\}$ est la réalisation du sous-ensemble des variables présentes dans le facteur Ψ_a.
- $\Psi_a(y_a, x_a)$ correspond au facteur d'index a et est une fonction exponentielle de la forme : $\Psi_a(y_a, x_a) = exp\{\sum_{k=1}^{K(a)} \lambda_a^k f_a^k(y_a, x_a)\}$.
- Les paramètres λ_a^k sont appris sur des données d'entraînement généralement par maximisation de la log-vraisemblance des données. Ces paramètres peuvent être vus comme une pondération entre les différents facteurs.
- $Z(x)$ est une constante de normalisation afin d'assurer que la somme des probabilités sur tous les états soit égale à 1 : $\sum_i p(y_i|x) = 1$.

Une forme de CRF couramment utilisée est le CRF linéaire introduit dans l'article [Lafferty 2001] pour étiqueter des données séquentielles et illustré sur la figure 5.2. Dans cet article, l'application consiste à donner une étiquette grammaticale à chaque mot d'un texte. Les variables cachées y correspondent donc ici aux étiquettes à trouver pour chaque mot. Leur valeur est estimée en fonction des observations x qui correspondent à des séquences de mots. Le CRF linéaire peut être vu comme la version discriminante d'un HMM. On peut noter que les CRF ont été utilisés pour de nombreuses applications comme la segmentation d'image, et y compris pour le regroupement en classes de visages [Du 2012] (section 2.2.2, page 22).

5.1.2 Formulation du modèle pour notre problème

Soient $A = \{A_i, i = 1 \dots N^A\}$ un ensemble de tours de parole et $V = \{V_i, i = 1 \dots N^V\}$ un ensemble de séquences de visages détectés dans une vidéo[3]. Le problème de regroupement

[3]Une séquence de visages est en pratique une série de visages détectés sur des images successives dans un plan vidéo (voir section 4.1, page 58)

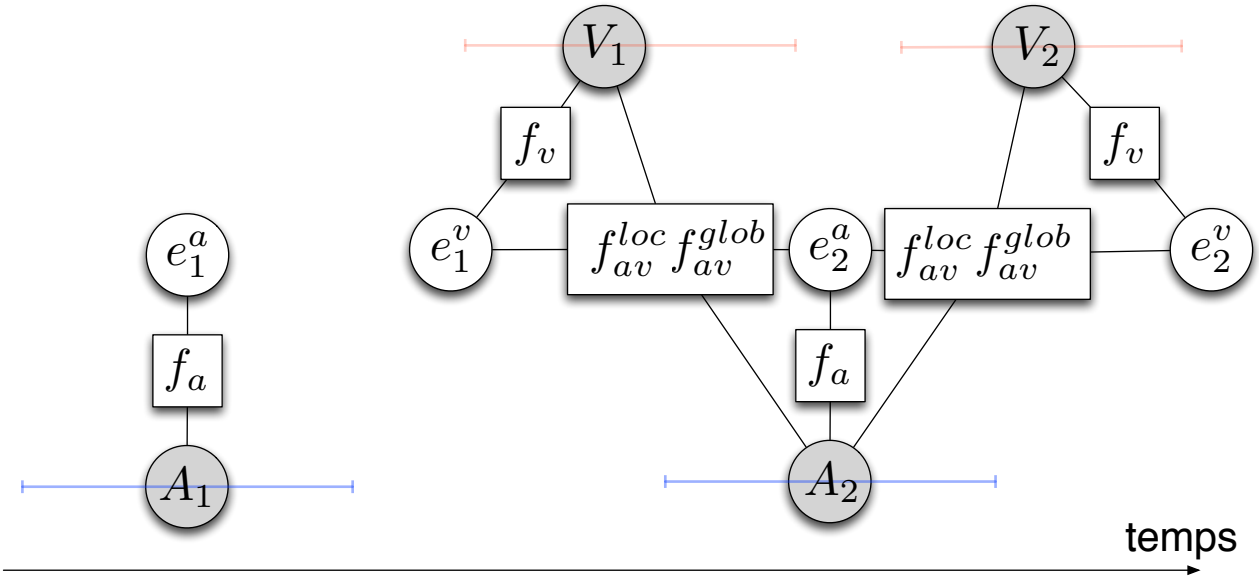


FIG. 5.3: Exemple d'un graphe de facteurs représentant le modèle déployé sur 4 segments. La séquence de visages V_1 apparaît en même temps que le tour de parole A_2. C'est pourquoi la vraisemblance de leurs étiquettes est liée à travers les 2 fonctions d'association f_{av}^{loc} et f_{av}^{glob}. Le tour de parole A_1 est isolé car aucun visage n'apparaît à l'écran en même temps. Ce cas correspond par exemple à une voix off dans un reportage. Sa vraisemblance ne dépend donc que de son label e_i^a à travers la fonction biométrique acoustique f_a.

audiovisuel des personnes est alors vu comme l'estimation du champ d'étiquettes $E = \{e_i^a, i = 1 \dots N^A, e_j^v, j = 1 \dots N^V\}$ en maximisant la distribution *a posteriori* $P(E|A,V)$. Le but est que la même valeur soit attribuée à e_i^a et e_j^v si le tour de parole A_i et la séquence de visages V_j correspondent à la même personne. Les étiquettes e_i^a et e_j^v sont définies sur l'ensemble des indices de personnes possibles qui sera noté $\mathcal{E}$. Nous exprimons alors la probabilité des étiquettes E par :

$$P(E|A,V) = \frac{1}{Z(A,V)} \times exp\Bigg\{ \lambda_a \sum_{i=1}^{N^A} f_a(A_i, e_i^a) + \lambda_v \sum_{i=1}^{N^V} f_v(V_i, e_i^v) + \lambda_{av}^{loc} \sum_{(i,j)\in G_{av}} f_{av}^{loc}(A_i, V_j, e_i^a, e_j^v) + \lambda_{av}^{glob} \sum_{(i,j)\in G_{av}} f_{av}^{glob}(A_i, V_j, e_i^a, e_j^v) \Bigg\} \tag{5.4}$$

où $Z(A,V)$ est la constante de normalisation et f_{av}^{loc},f_{av}^{glob}, f_a et f_v sont respectivement les deux fonctions caractéristiques d'association audiovisuelle, la fonction biométrique acoustique, et la fonction biométrique visuelle. Un exemple de représentation graphique du modèle déployé sur quelques segments est donné dans la Figure 5.3. Dans la section suivante, nous décrivons plus en détail chacune des fonctions caractéristiques.

5.1.3 Description des composantes du modèle

Mesure de l'association voix/visages à l'échelle des segments

Afin d'intégrer l'information d'association audiovisuelle dans le CRF, il faut être capable de détecter si le locuteur est présent à l'écran et à quel visage il correspond. Les caractéristiques que nous utilisons pour l'association AV sont similaires à celles relevées dans l'état de l'art (voir section 2.3.1, page 26). Elles sont extraites pour chaque couple de segments (A_i, V_j)(un tour de parole et une séquence de visages) présentant une durée de co-occurrence non nulle.

1. **Mesure de l'activité des lèvres.** Elle est calculée selon la méthode décrite dans [El Khoury 2012b] que nous rappelons brièvement ici. Soit $x_j^{lad}(k)$ l'activité de la séquence V_j à l'image k. Cette activité est calculée à partir de la différence d'intensité entre les images k et $k+1$ dans des zones prédéfinies correspondant à la région des lèvres. La valeur finale décrivant l'activité de la séquence V_j par rapport à l'utterance A_i est la moyenne des activités sur les images co-occurrant avec A_i :

$$x_{ij}^{lad} = \frac{1}{\#N_{ij}} \sum_{k \in N_{ij}} x_j^{lad}(k) \tag{5.5}$$

   où N_{ij} correspond à l'ensemble des indices des images où A_i et V_j sont présents.

2. **Activité relative par rapport aux autres visages présents.** La mesure d'activité que nous cherchons n'est pas nécessairement une valeur absolue. Le but est de faire ressortir le visage qui semble plus parler que les autres. Pour cette raison, nous utilisons la mesure relative suivante :

$$x_{ij}^{ladr} = \frac{1}{\#N_{ij}} \sum_{k \in N_{ij}} \frac{x_j^{lad}(k)}{\sum_{l \in V^k} x_l^{lad}(k)} \tag{5.6}$$

   où V^k représente l'ensemble des indices de séquences ayant un visage dans l'image k.

3. **Taille du visage.** Cette caractéristique exploite l'hypothèse que le visage du locuteur est plus grand que les autres. Si la fenêtre de détection a une hauteur h_k et une largeur w_k sur l'image k, la taille $x_j^h(k)$ de la séquence V_j sur l'image k est définie comme :

$$x_j^h(k) = \sqrt{h_k^2 + w_k^2} \tag{5.7}$$

   La valeur retenue au final correspond à la moyenne sur l'ensemble des images de la séquence V_j co-occurrant avec le tour de parole A_i.

4. **Taille relative du visage par rapport aux autres visages présents.** De la même manière que pour l'activité, la taille relative est ajoutée aux autres caractéristiques.

$$x_{ij}^{hr}(k) = \frac{1}{\#N_{ij}} \sum_{k \in N^{ij}} \frac{x_j^h(k)}{\sum_{l \in V^k} x_l^h(k)} \tag{5.8}$$

Ces caractéristiques sont utilisées avec un classifieur de type Séparateur à Vaste Marge (SVM) avec un noyau gaussien. Le classifieur est entraîné sur des couples de tour de parole/séquence de visages afin de répondre à la question : *Ce couple correspond-il à la même personne ?* Ses performances seront évaluées dans la section 5.2.1.

Fonction caractéristique d'association voix/visage à l'échelle des segments : $f_{av}^{loc}(A_i, V_j, e_i^a, e_j^v)$

À présent, il reste à intégrer les décisions du SVM dans le CRF. La fonction caractéristique f_{av}^{loc} est définie sur l'ensemble G_{av} des couples de tour de parole/séquence de visages se chevauchant temporellement : $G_{av} = \{(i,j)/t(A_i, V_j) \neq 0\}$ où $t(x,y)$ est la durée pendant laquelle les segments x et y se chevauchent. Le but est de produire un score élevé si le tour de parole A_i et la séquence de visages V_j correspondent à la même personne.

$$f_{av}^{loc}(A_i, V_j, e_i^a, e_j^v) = \begin{cases} t(A_i, V_j)h(A_i, V_j) & \text{si } e_i^a = e_j^v \\ -t(A_i, V_j)h(A_i, V_j) & \text{sinon} \end{cases} \tag{5.9}$$

où $h(A_i, V_j)$ représente la sortie binaire du classifieur SVM :

$$h(A_i, V_j) = \begin{cases} 1 & \text{si le SVM considère que } A_i \text{ et } V_j \text{ correspondent à la même personne} \\ -1 & \text{sinon} \end{cases} \tag{5.10}$$

Association voix/visage à l'échelle des classes : $f_{av}^{glob}(A_i, V_j, e_i^a, e_j^v)$

Le classifieur SVM peut fournir des informations erronées pour l'association. Or, la co-occurrence temporelle à l'échelle de la classe est une caractéristique complémentaire de celles utilisées dans le SVM. Pour cette raison, une fonction f_{av}^{glob} est introduite afin de favoriser les couples d'étiquettes ayant une grande co-occurrence temporelle. Un ensemble $\mathcal{H}$=$\{(A_i, V_i)\}$ de couples de segments est construit à partir d'une association qui maximise la co-occurrence entre les classes de locuteurs et de visages. La fonction suivante est alors définie pour tous les

couples de segments $(i, j) \in \mathcal{H}$:

$$f_{av}^{glob}(A_i, V_j, e_i^a, e_j^v) \begin{cases} 1 & \text{si } e_i^a = e_j^v \text{ ET } (A_i, V_j) \in \mathcal{H} \\ 0 & \text{sinon} \end{cases} \tag{5.11}$$

Les couples $\mathcal{H}$ sont choisis de manière à maximiser la co-occurrence temporelle entre les classes de locuteur et de visage. Premièrement, une matrice de similarité M est construite, où chaque élément M_{ij} de la matrice correspond à la durée de chevauchement entre la classe de visage i et le locuteur j. À partir de cette matrice, l'algorithme hongrois [Kuhn 1955] est utilisé afin d'obtenir les couples locuteur/visage maximisant la durée de co-occurrence.

$$\mathcal{C} = \arg\max_{C} \sum_{(i,j)\in C} M_{ij} \tag{5.12}$$

avec la contrainte sur C qu'un visage ou un locuteur ne peut appartenir qu'à un seul couple : $\nexists (i, j) \in C, (i, k) \in C, j \neq k$. L'algorithme hongrois permet de trouver une solution globalement optimale à ce problème en un temps polynomial. L'ensemble $\mathcal{H}$ correspond alors à tous les couples de segments (A_i, V_j) tels que les A_i et V_j proviennent d'un couple d'étiquettes présent dans $\mathcal{C}$:

$$\mathcal{H} = (\{A_i, V_j), (e_i^a, e_j^v) \in \mathcal{C}\} \tag{5.13}$$

Fonction biométrique acoustique : $f_a(A_i, e_i^a)$

Cette fonction estime la vraisemblance des caractéristiques acoustiques pour le tour de parole A_i étant donné qu'il est étiqueté avec le locuteur e_i^a. C'est un problème de reconnaissance du locuteur. Il faut donc construire un modèle acoustique pour chaque étiquette $e \in \mathcal{E}$ et l'apprendre à partir des données attribuées à cette étiquette quand c'est possible. L'attribution des données pour l'apprentissage du modèle sera détaillée dans la section 5.1.4.

Le modèle choisi est de type GMM-UBM à matrice de covariance diagonale, similaire à celui adopté dans [Reynolds 2000]. Les caractéristiques sont des MFFC avec les dérivées du premier ordre. Deux normalisations sont appliquées. La première est à base de *feature warping* [Pelecanos 2001] sur des fenêtres de 3 secondes. La seconde consiste à centrer et réduire les données à l'échelle de la classe selon la méthode *Mean and Variance Substraction*. Le modèle GMM de chaque locuteur i est appris par adaptation des moyennes de l'UBM. Le score de la fonction $f_a(A_i, e_i^a)$ correspond à la vraisemblance du segment A_i par rapport au modèle associé à l'étiquette e_i^a après une normalisation de type *z-norm*. Le choix de la méthode GMM-UBM a été privilégiée sur celui d'une approche à base de *i-vector* en raison de la faible durée des

segments. Toutefois, une modélisation de type *Joint Factor Analysis* ([Kenny 2007]) aurait pu être utilisée.

Fonction biométrique visuelle : $f_v(V_i, e_i^v)$

De manière similaire, la fonction f_v estime la vraisemblance des caractéristiques visuelles pour la séquence de visage V_i étant donné qu'elle est étiquetée par le visage e_i^v. Le modèle visuel s'inspire de la distance décrite dans la section 4.2.1, page 60. La distance entre deux séquences de visages est une moyenne pondérée de l'ensemble des distances entre paires de descripteurs SURF. La différence avec nos travaux précédents est que la distance entre une séquence V_i et une étiquette e_i^v est le $10^{ème}$ quantile (au lieu de la moyenne) de l'ensemble des distances entre V_i et les séquences attribuées à l'étiquette e_i^v. Dans nos expériences, l'utilisation du quantile plutôt que la moyenne a permis d'être plus robuste aux variations de pose.

5.1.4 Initialisation et optimisation

L'utilisation du CRF pour étiqueter de nouvelles données, illustrée sur la figure 5.4, est conduite de la manière suivante.

1. Effectuer séparément un regroupement en locuteur et un regroupement en classes de visages.
2. Associer les classes des locuteurs et des visages pour initialiser l'ensemble $\mathcal{E}$ des étiquettes.
3. Pour chaque étiquette, apprendre à partir des données qui y sont associées son modèle acoustique et son modèle vidéo.
4. Effectuer le décodage : obtenir les étiquettes les plus probables.

L'association de l'étape 2 utilise l'algorithme hongrois comme dans la section 5.1.3. Cette fois, la matrice de similarité M est construite de telle sorte que chaque entrée M_{ij} est la somme des scores de la fonction d'association f_{av}^{loc} sur tous les couples audiovisuels de segments associés au $i^{ème}$ locuteur et au $j^{ème}$ visage.

Il faut noter que le modèle CRF n'a pas la possibilité de créer des étiquettes une fois celles-ci initialisées. En revanche il peut n'attribuer aucun segment à une étiquette. Il est donc supposé que les regroupements monomodaux ont produit au moins autant d'étiquettes que de personnes présentes.

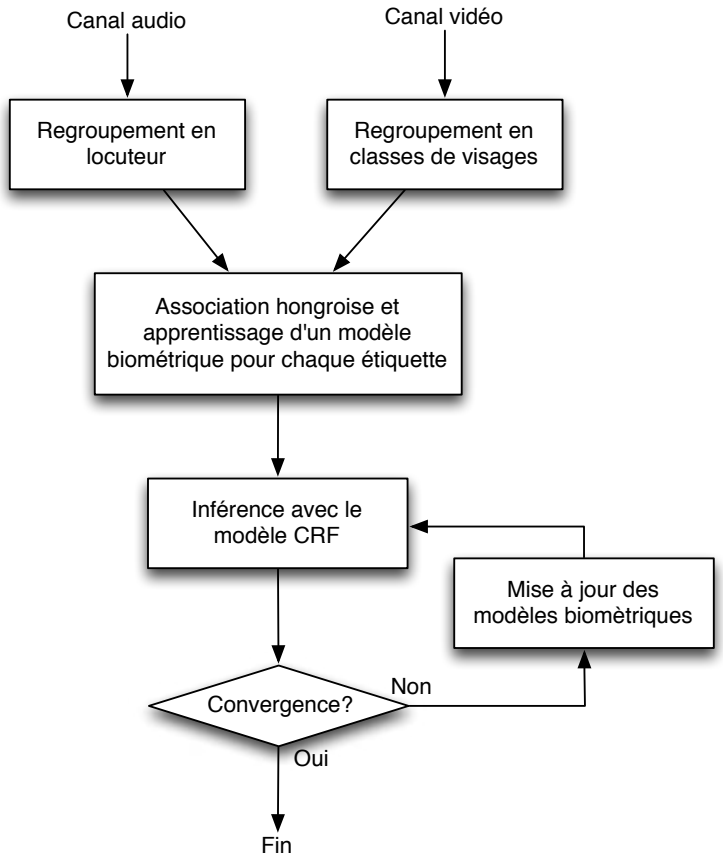


FIG. 5.4: Diagramme représentant l'utilisation du modèle CRF de regroupement audiovisuel des personnes.

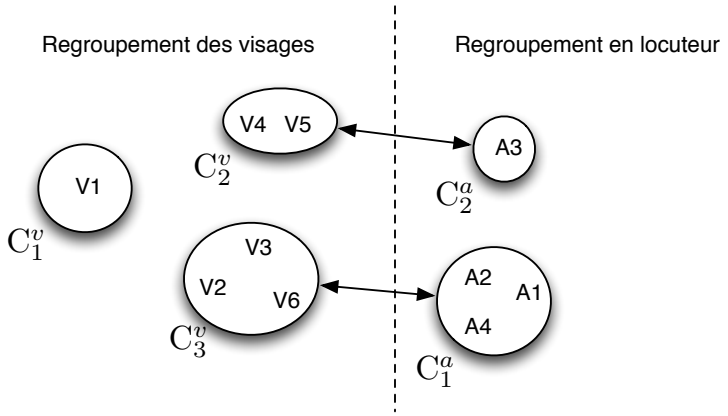
Pour chaque étiquette obtenue, des modèles biométriques sont appris à partir des données associées. L'initialisation des étiquettes et des modèles biométriques est illustrée sur la figure 5.5. Sur cette figure, le modèle acoustique de l'étiquette $e_1(C_3^v - C_1^a)$ est ainsi appris sur les tours de parole A1, A2 et A4.

À partir de ces modèles, l'algorithme d'inférence *Loopy belief propagation* [Murphy 1999](LBP) est utilisé dans l'étape 4 pour obtenir les étiquettes E^{opt} les plus probables en résolvant :

$$E^{opt} = \arg\max_{E} P(E|A, V) \tag{5.14}$$

L'algorithme LBP a été choisi car il permet de traiter des graphes contenant des cycles.

Les étapes 3 et 4 peuvent être itérées comme dans un algorithme d'Expectation-Maximization [Moon 1996] en alternant les mises à jour des modèles biométriques et les étapes d'inférence. Si les étiquettes des segments obtenues par l'inférence du CRF sont les mêmes



Étiquettes audiovisuelles		Segments utilisés pour l'apprentissage des modèles visuels (gauche) et acoustiques (droite)	
e_1 :	$C_2^v - C_2^a$	V4, V5	A3
e_2 :	$C_3^v - C_1^a$	V2, V3, V6	A1, A2, A4
e_3 :	C_1^v	V1	$\varnothing$ (modèle neutre)

FIG. 5.5: Ce schéma représente les classes obtenues par les deux regroupements monomodaux. Les flèches représentent les associations obtenues avec l'algorithme hongrois ($C_3^v - C_1^a$ et $C_2^v - C_2^a$). Le tableau liste les segments qui sont utilisés pour apprendre chaque modèle avec le cas spécial de l'étiquette $e_3(C_1^v)$ qui se voit attribuer le modèle neutre car aucune donnée audio ne lui est associée.

que celles de l'itération précédente, l'algorithme a convergé. Cette convergence n'est cependant pas garantie dans notre cas car la vraisemblance n'est pas définie de la même manière dans les étapes 3 et 4. En pratique, un nombre fixe n d'itérations est effectué.

Cas des visages muets et des voix offs.

À cause de la présence de visages muets et de voix offs, certaines étiquettes sont attribuées uniquement à des visages ou uniquement à des tours de parole. Ce cas est représenté sur la figure 5.5 où l'étiquette e_3 n'est associée qu'avec des visages. Cependant, lors de l'inférence, le modèle doit évaluer la vraisemblance d'un tour de parole pour cette étiquette. Il faut donc obtenir un modèle acoustique pour cette étiquette mais il n'y a pas de données associées pour l'apprendre. Afin de contourner ce problème, un modèle biométrique *neutre* est défini pour chaque modalité. Ces modèles sont alors utilisés pour la modalité manquante. Concrètement, le score d'une observation pour le modèle neutre est le seuil où le taux de fausse alarme est égal au taux de mauvaises détections. Ces taux ont été calculés séparément dans des expériences de reconnaissance du locuteur et de visage.

5.1.5 Entraînement des paramètres λ_i

Le modèle CRF est paramétré par les poids λ_i qui permettent de pondérer la contribution de chaque fonction f_i. Ces paramètres sont appris en utilisant des données d'entraînement (des vidéos pour lesquelles les références du regroupement audiovisuel des personnes sont disponibles).

Afin d'apprendre les λ_i efficacement, le comportement des fonctions f_i doit refléter autant que possible le comportement qu'elles auront dans les conditions de test. Pour cette raison, lors de l'apprentissage, elles sont calculées en utilisant les résultats des regroupements monomodaux automatiques. Les modèles biométriques appris pour chaque étiquette utilisent les classes produites automatiquement et reflètent donc le bruit présent au moment du test.

En d'autres termes, le CRF est entraîné en prenant en compte la fiabilité des algorithmes de regroupement mono-modaux utilisés pour l'initialisation.

5.1.6 Comparaison avec d'autres approches de l'état de l'art

Contrairement aux approches détaillées dans [El Khoury 2012b, Bendris 2011], le modèle présenté ici utilise une optimisation globale de la fonction de coût plutôt qu'une minimisation par une série de règles locales. Le modèle décrit dans [Noulas 2012], (section 2.3.2, page 28) est le plus similaire au nôtre. La vraisemblance des données audios et vidéos y est aussi exprimée de manière jointe : la similarité monomodale entre un segment et sa classe est calculée à partir d'un modèle biométrique de la classe et la similarité audiovisuelle utilise des scores d'association locaux.

D'un autre côté, ses travaux sont orientés principalement vers des enregistrements de réunions alors que nous traitons des journaux télévisés. Cela implique des différences au niveau des choix de conception du système, du protocole expérimental, de la qualité de l'association audiovisuelle qu'il est possible d'obtenir, et des performances des regroupements monomodaux auxquels on peut se comparer. Il reste donc intéressant d'étudier les performances d'une telle approche sur les données REPERE. Plus précisément, les principales différences sont les suivantes :

– Le modèle de Noulas considère que la trame audio (40 ms) est l'unité temporelle de base. L'état de chaque trame est lié à l'état de la trame précédente comme dans une chaine de Markov cachée d'ordre 1. Le modèle réalise à la fois la segmentation en tours de parole et le regroupement en locuteur. Dans notre modèle, l'unité est le tour de parole (ou la séquence de visages pour la vidéo) et il n'y a pas systématiquement de lien entre deux segments consécutifs. L'inconvénient est la non-remise en cause des frontières des segments. En revanche, la suppression de ces liens rend l'inférence moins coûteuse en

temps de calcul. De plus, ce niveau de granularité est plus logique pour calculer certaines caractéristiques d'association (comme l'activité des lèvres moyenne) et de nommage (comme la durée de co-occurrence qui sera utilisée dans le chapitre 6). Enfin, les scores biométriques calculés sur toutes les données d'un segment plutôt que sur une seule trame les rendront plus fiables.

– Dans ses expériences, le nombre de locuteurs présents dans l'enregistrement est connu *a priori* et est relativement faible (4 à 5 locuteurs). Dans le cadre de REPERE, le nombre de locuteurs est inconnu et certains enregistrements contiennent jusqu'à 20 locuteurs. Comme nous l'avons vu dans la section 5.1.4, l'initialisation des étiquettes dans notre système se fonde sur les sorties des regroupements monomodaux.

– Concernant l'association audiovisuelle, les vidéos utilisées dans ses travaux ne contiennent pas de voix off. Le locuteur est donc présent à l'écran sauf si il sort du champ de la caméra. Le nombre de visages par image semble être également plus réduit. Ces facteurs incitent à penser que l'association audiovisuelle est plus facile que dans le cas de REPERE.

– À l'inverse de notre système, l'approche de Noulas permet de gérer la parole superposée. Cependant, c'est une hypothèse courante pour les systèmes traitant de journaux télévisés de supposer que celle-ci est trop peu fréquente pour nécessiter des traitement spécifiques.

– Les données utilisées par Noulas semblent favorables à un regroupement des visages performant, tout particulièrement sur la vidéo de journal télévisé où son regroupement ne fait aucune erreur. Cela permet d'utiliser la modalité vidéo pour améliorer le regroupement locuteur avec succès.

– Enfin, le modèle de Noulas n'est pas discriminant mais génératif. Cela implique de normaliser chaque modalité sous la forme d'une probabilité afin de les combiner directement. Le CRF proposé ici apprend des paramètres λ sur des données d'apprentissage de façon à optimiser cette combinaison.

5.2 Expériences pour le regroupement audiovisuel des personnes

Cette section présente une évaluation du modèle proposé pour le regroupement audiovisuel des personnes. Cette évaluation est conduite sur la partie PHASE2 du corpus REPERE (voir section 1.2.1, page 5) qui a été utilisée pour l'évaluation finale de la campagne. La partie DEV2 a servi à régler les paramètres des systèmes de regroupement mono-modaux, à valider le

choix des fonctions caractéristiques utilisées dans le CRF et à décider du nombre d'itérations n. L'ensemble TEST2 est notre ensemble de test. Les paramètres λ du CRF ont été appris sur les données correspondant à la partie TEST0 qui contient 3h de vidéos annotées provenant de fichiers différents mais des mêmes émissions. Les vidéos du corpus TRAIN2 auraient pu être utilisées, cependant, il a été constaté que 3h de données reste suffisant pour apprendre les paramètres.

5.2.1 Évaluation du module d'association voix/visage

Le module d'association entre les tours de parole et les séquences de visages est très important pour l'efficacité du regroupement joint. C'est pourquoi, dans un premier temps, le SVM h (présenté dans la section 5.1.3) est évalué séparément sur la tâche de classification des couples tour de parole/séquence de visage en fonction du fait qu'ils appartiennent à la même personne ou pas. Comme pour les paramètres du CRF, le TEST0 a été utilisé pour apprendre le classifieur. Les résultats sont reportés dans le tableau 5.1 avec les matrices de confusion. Une analyse des erreurs a permis de tirer les conclusions suivantes :

- Les performances en terme de précision moyenne sont équivalentes entre le DEV2 et le TEST2 (77% vs 76%).
- Les locuteurs en voix off ont tendance à être associés à tort avec les visages apparaissant à l'écran pendant qu'ils parlent.
- Sur le TEST2, la majorité des erreurs sont des couples qui devraient être classés comme n'appartenant pas à la même personne. Cela correspond à des cas où l'image est divisée en plusieurs parties et où chaque partie contient un visage. Par exemple, dans l'image de droite de la figure 5.8, les 3 personnages principaux ont une activité des lèvres car celui en haut à droite est animé par son discours [4], les autres par leur conversation. De plus, ce type de cas contredit la préférence que le visage du locuteur à associer au tour de parole est normalement le plus grand sur l'image. Estimer la synchronie sur de telles séquences est hors de portée de l'état de l'art actuel étant donné la petite taille des visages et leurs mouvements.
- Dans le cas des débats télévisés, une difficulté réside dans le fait que les caméras utilisées changent durant le même tour de parole, montrant tour à tour les différents participants. Ce procédé est utilisé pendant les tours de parole relativement longs (>10 sec) afin de retenir l'attention du spectateur. Ceci peut introduire des cas de co-occurrences locuteur/visage qui pourtant ne doivent pas être associés.

[4] La personne est montrée dans un interview, pour illustration, mais son discours n'est pas entendu. Ce cas est fréquent dans les données.

– Lorsque les interactions entre locuteurs sont fréquentes (par exemple quand les personnages se coupent la parole dans un débat), les plans ne suivent pas toujours le locuteur. Cette perte de synchronisation audiovisuelle provoque des erreurs dans l'association.

		DEV2 ($Prec = 0.77$)		TEST2 ($Prec = 0.76$)	
		prédictions		prédictions	
		loc. = visage	loc. ≠ visage	loc. = visage	loc. ≠ visage
Références	loc. =visage	1147 (40%)	345 (12%)	1787 (39%)	673 (15%)
	loc. ≠ visage	328 (11%)	1062 (37%)	423 (9%)	1663 (37%)

TAB. 5.1: Résultats de classification de couples tour de parole/séquence de visages sur les corpus DEV2 et TEST2. Les matrices de confusion sont présentées avec le taux de bonne classification $Prec$.

Heureusement, dans le CRF, les similarités des données monomodales vont avoir tendance à empêcher les associations erronées. Afin de résoudre ces cas au niveau du classifieur, des *a priori* à partir du type de scène pourraient être utilisés, par exemple en détectant les reportages où le locuteur est un journaliste en off. Il est aussi possible de comparer les scores au niveau de l'image et les contraindre de sorte que seul un visage peut être rattaché au locuteur courant.

5.2.2 Évaluation du regroupement audiovisuel des personnes

Performances des systèmes monomodaux

Comme décrit dans la figure 5.4, le modèle CRF proposé est utilisé à la suite de deux regroupements monomodaux. Le système monomodal utilisé pour le regroupement en locuteur [Rouvier 2013] a été détaillé dans le premier chapitre de l'état de l'art. Celui effectuant le regroupement en classes de visages [Khoury 2013] a été l'objet du chapitre précédent. Leurs performances sur les données PHASE2 sont présentées dans le tableau 5.2. La métrique d'évaluation utilisée est le DER dont la définition a été rappelée dans la section 1.2.3. Concernant

Corpus	Locuteurs				Visages	
	DER	*Confusion*	*Non-détection*	*Fa*	*Confusion*	*Non-détection*
DEV2	13.54%	7.9%	5.4%	0.3%	5.9%	36.8%
TEST2	12.97 %	7.4%	5.4%	0.2%	5.2%	47.7%

TAB. 5.2: Évaluation du regroupement en locuteur obtenu avec le système de [Rouvier 2013] et du regroupement en classes de visages obtenu avec le système de [Khoury 2013].

le regroupement en locuteur, le tiers des erreurs correspond à de la non-détection de parole. Ces non-détections sont dues à l'emploi fréquent de jingles qui servent à retenir l'attention du

	DEV2			TEST2		
	Audio	Vidéo	AV	Audio	Vidéo	AV
Regroupement initiaux	7.9%	5.9%	9.2% (27.1%)	7.4%	5.2%	8.7%(22.2%)
CRF-Monomodal	7.7%	6.3%	9.7%	7.3%	6.4%	8.9%
CRF+Association Manuelle	6.6%	4.8%	5.9%	6.9%	4.3%	6.2%
CRF proposé	7.5%	6.2%	9.1%	7.4%	6.1%	8.5%

TAB. 5.3: La partie *Confusion* du DER est reportée pour les tâches de regroupement en locuteur (Audio), de regroupement en classes de visages (Vidéo) et de regroupement audiovisuel des personnes (AV). Les lignes correspondent à différents systèmes et conditions expérimentales décrits dans le texte.

spectateur. Les erreurs de confusion correspondent à des bruits de fond (par exemple bruit des voitures dans le cas d'un reportage dans la rue ou présence d'une musique de fond pendant l'émission) ou à des segments contaminés par de la parole superposée.

Côté vidéo, seuls les taux de confusion et de non-détection ont été reportés car il n'est pas possible de calculer un taux de fausse alarme avec précision. En effet, les références ne contiennent pas les visages en dessous d'une certaine taille. Or, ces visages sont malgré tout détectés ce qui génère, à tort, des fausses alarmes. Concernant les résultats du tableau, il y a beaucoup plus d'erreurs de non-détection que d'erreurs de confusion (36.8% vs 5.9% pour le DEV2 et 47.7% vs 5.2% pour le TEST2). Ce sont des personnes de profils ou de dos présentes dans les annotations mais que les détecteurs frontaux ne parviennent pas à récupérer. Concernant les erreurs de confusion, il a été annoncé dans la section 4.3.2, page 67 que le système a tendance à créer des classes pures mais qu'une personne peut être divisée en plusieurs classes.

Performances de la méthode CRF proposée

Dans le paragraphe suivant, plusieurs résultats vont être présentés pour expliquer les effets des différentes parties du modèle CRF sur le regroupement joint par rapport aux regroupements monomodaux initiaux. Ces résultats sont regroupés dans le tableau 5.3.

Comme le DER n'est pas disponible pour la vidéo, les résultats sont reportés uniquement en terme de taux de confusion (partie *Confusion* du DER) afin d'avoir la même mesure pour l'audio et la vidéo. Cela n'enlève rien à la validité des comparaisons car quels que soient les systèmes, la segmentation est fixée et seules les étiquettes des segments changent. Par conséquent, le nombre de fausses alarmes et de non-détections reste constant.

La première ligne correspond aux systèmes de regroupement monomodaux dont les résultats ont déjà été présentés dans le tableau 5.2. L'association audiovisuelle évaluée sur cette ligne est

celle obtenue avec l'algorithme hongrois. C'est à dire, la dernière étape avant l'utilisation du CRF (voir figure 5.4). Le score entre parenthèses est obtenu sans l'association en conservant toutes les étiquettes monomodales. La troisième ligne présente les résultats du système CRF où la décision du SVM utilisée dans la fonction d'association f_{av}^{loc} (section 5.1.3) a été remplacée par la vérité terrain. La comparaison entre la première et la troisième ligne montrent des améliorations significatives avec globalement un gain relatif autour de 15% pour les tâches de regroupement monomodales et un gain relatif de 30% pour la tâche audiovisuelle. Pour vérifier que ce gain provient de l'utilisation de l'information multimodale, les performances ont été évaluées en retirant les fonctions d'association audiovisuelles f_{av}^{loc} et f_{av}^{glob} du CRF. Les résultats, présentés dans la deuxième ligne *CRF-Monomodal*, confirment le rôle bénéfique de l'association audiovisuelle car le système monomodal ne présente pas de gain par rapport aux regroupements initiaux. Au contraire, pour la vidéo, les performances sont dégradées. La représentation semble moins bien adaptée pour le CRF qu'avec le regroupement hiérarchique. Ces premiers résultats nous permettent cependant de vérifier que notre modèle est capable de combiner les différentes informations.

Les résultats de la quatrième ligne, qui correspondent au système CRF avec l'association audiovisuelle automatique, sont globalement similaires par rapport aux regroupements initiaux ce qui est surprenant. L'analyse détaillée des résultats permet de relever les situations suivantes : dans le cas de l'audio, les segments sont mal classés à cause soit de la présence de parole superposée, soit de la présence de bruit de fond. Les cas de parole superposée sont des zones d'interaction fréquente entre locuteurs. Dans ce type de scène, les changements de plan successifs rendent l'association audiovisuelle difficile (voir figure 5.6). Ce manque de fiabilité de l'association explique que le CRF n'arrive pas à les corriger.

En revanche, des erreurs de confusion du système monomodal dues à des bruits de fond dans des zones où l'association audiovisuelle est fiable sont corrigées. Un exemple est présenté sur la figure 5.7. Ceci est aussi illustré par la légère amélioration observée sur le DEV2 (7.9 vs 7.5). Il faut aussi noter que des améliorations en DER plus significatives par rapport à un système monomodal (12.6% vs 16.9%) avaient été obtenues sur les données du dry-run (PHASE0) de la campagne REPERE [Gay 2014b]. Le système monomodal utilisé à l'époque correspond aussi à un système ILP/i-vector. La différence avec le système actuel (qui obtient un DER de 13.3% sur les données de la PHASE0) réside dans l'optimisation du choix des données d'entraînement de l'UBM. Comme le système actuel est plus performant, il est plus difficile de l'améliorer.

Nous avons vu que l'utilisation de la vérité terrain pour l'association permet d'améliorer le regroupement vidéo. Le fait que l'association automatique ait tendance à rejeter l'association quand plusieurs visages sont présents explique que ces gains soient alors perdus (voir

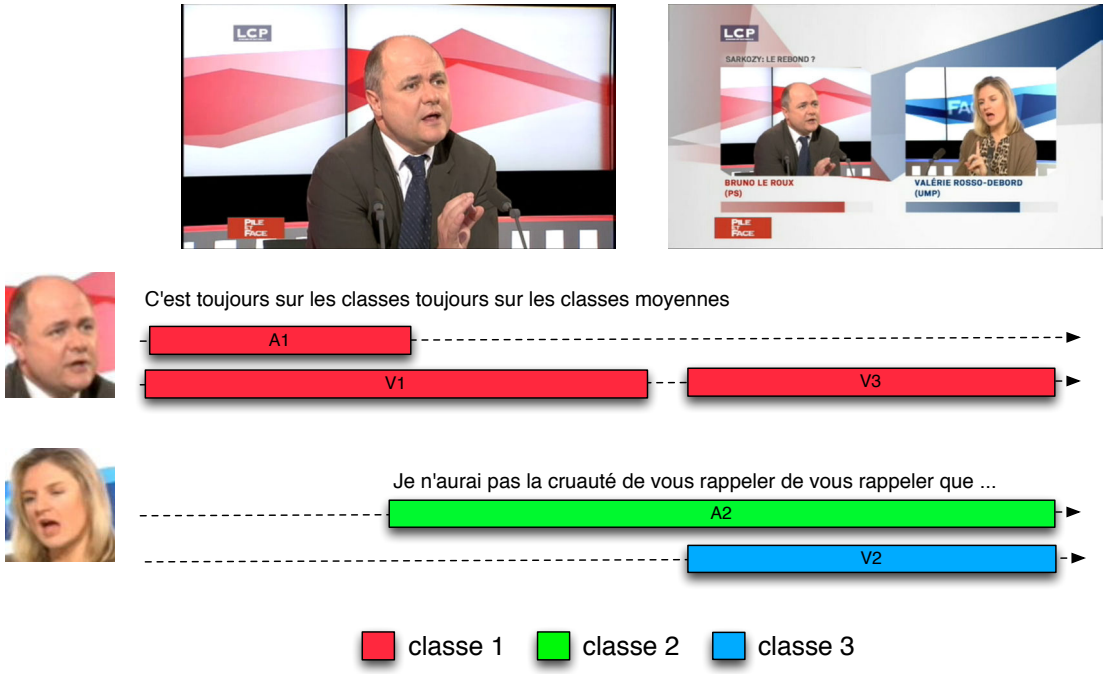


FIG. 5.6: Afin d'améliorer le DER, le tour de parole A2 devrait être ré-attribué à la classe 3. Une partie de A2 est superposée avec A3. La fonction d'association du CRF f_{av}^{loc} est définie sur les couples A2-V1 et A2-V2. Comme le second plan contient deux visages de petite taille, le score d'association est favorable, à tort, à l'association entre A2 et V1.

figure 5.8). Dans ce cas, il serait peut être judicieux de s'assurer qu'au moins une association ait lieu. Cependant il faudrait être capable de détecter automatiquement ce type de situations et les distinguer des reportages avec une voix off.

La dégradation de performance en vidéo sur le TEST2 (5.2% vs 6.1%) par rapport au système initial peut être mise en parallèle avec la dégradation observée en utilisant le CRF de manière monomodale (5.2% vs 6.4%). Dans ces expériences, le système monodal utilise comme représentation des visages une combinaison d'une distance fondée sur les descripteurs SURF et d'un modèle GMM modélisant la distribution de caractéristiques DCT (cette représentation est l'objet du chapitre 4). Le CRF utilise plus simplement une distance dérivée des descripteurs SURF. Il est possible que le système monomodal bénéficie d'une meilleur représentation visuelle. Cependant, des expériences additionnelles où le CRF a été doté d'une représentation utilisant une combinaison similaire n'ont pas permis d'améliorer les performances. Il semble donc que le regroupement hiérarchique monomodal est meilleur que l'utilisation des mêmes types de représentation sous forme de modèles biométriques dans le CRF.

Finalement, l'utilité d'effectuer un regroupement joint est à envisager en fonction des systèmes monomodaux et en fonction de la confiance qu'il est possible d'avoir avec l'association

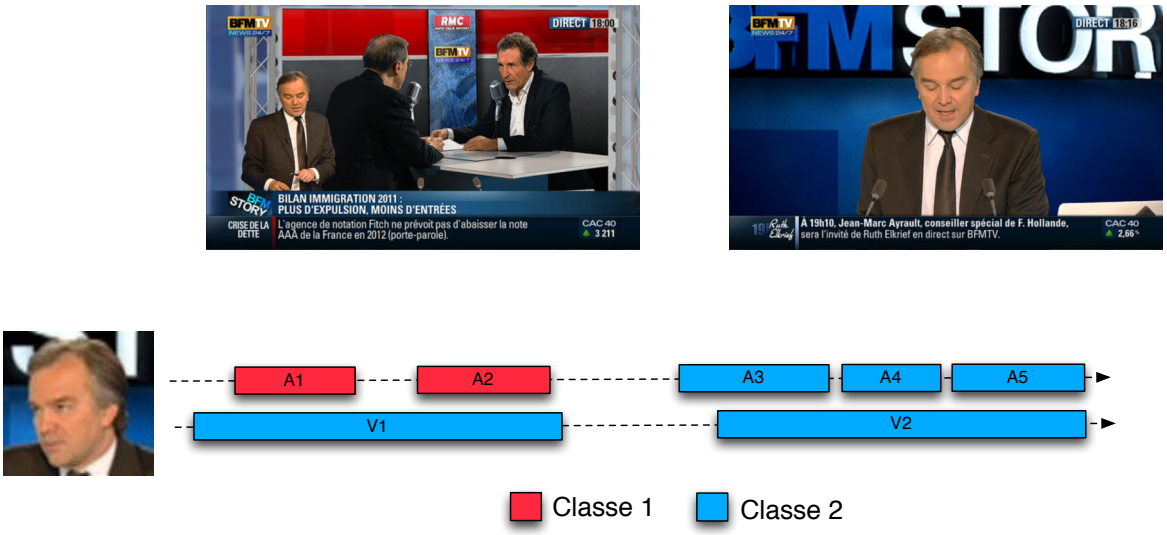


FIG. 5.7: Dans cet exemple, la classe du présentateur est divisée en deux. Cette scission est due à la présence d'une musique de fond au début de l'émission quand le présentateur résume les sujets qui vont être traités. Dans ce cas, le CRF permet de ré-attribuer les tours de parole A1 et A2 à la classe 2 et d'améliorer ainsi le regroupement audio.

audiovisuelle. Dans le cas des données REPERE, cela se traduit par des améliorations du côté audio dans le cas de musique de fond ou d'arrière fond sonore, car la vidéo constitue alors un bon adjuvant.

5.3 Conclusions

Ce chapitre nous a permis d'introduire un modèle à base de CRF pour le regroupement audiovisuel des personnes. Il permet d'améliorer le regroupement en locuteur quand le regroupement initial monomodal comporte des erreurs de confusion dues aux bruits de fond. Cependant les améliorations globales sont limitées par le fait que le regroupement initial comporte peu d'erreurs de ce type et par la difficulté de la tâche d'association des visages et des voix.

Comme décrit dans l'état de l'art, le regroupement peut être enrichi d'un contexte plus varié en incluant les vêtements, l'arrière plan ou les co-occurrences. Dans le cadre des journaux télévisés, ce contexte inclut également les informations venant des cartouches, ainsi que le rôle et le type de scène. Dans le chapitre suivant, le modèle CRF qui vient d'être présenté sera enrichi

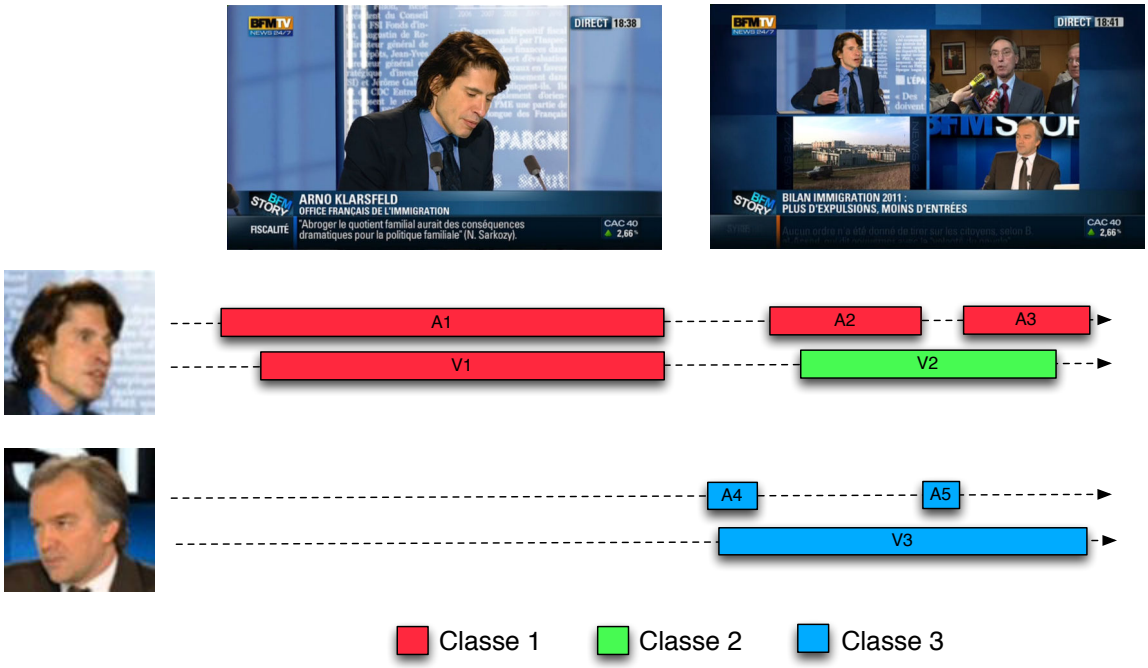


FIG. 5.8: Pour améliorer le regroupement vidéo, la séquence de visage V2 devrait passer dans la classe 1. Cependant, à cause de sa petite taille et de la présence de plusieurs visages ayant aussi une activité des lèvres, le SVM de la fonction f_{av}^{loc} décide, à tort, que les couples A2-V2 et A3-V2 ne correspondent pas à la même personne.

avec de tels éléments. Ceci permettra d'améliorer le regroupement audiovisuel et d'obtenir une meilleure identification des visages et des personnes.

Chapitre 6

Identification des visages et des locuteurs

Sommaire

Ce chapitre présente notre méthode pour l'identification des locuteurs et des visages à partir des noms présents dans les cartouches. Tout d'abord, le module de détection des noms dans les cartouches est décrit dans la section 6.1. Ensuite, nous présentons dans la section 6.2 deux méthodes d'identification à l'échelle des classes : la première, développée au sein du consortium SODA, est à base de règles, la seconde constitue notre contribution propre et exploite le formalisme des CRF. Des expériences sur l'identification des personnes évaluent le potentiel des cartouches pour l'identification et les performances des deux méthodes. Dans la section 6.3, le modèle CRF de regroupement décrit dans le chapitre précédent est revu et corrigé. Un attribut fondé sur la répétition de l'arrière-plan aux propriétés intéressantes pour l'identification est ajouté. Puis, des fonctions exploitant la co-occurrence entre les cartouches et les segments (tours de parole ou séquences de visages) sont introduites afin de guider le regroupement de ces derniers. Le système final d'identification combine le CRF de nommage à l'échelle des classes et le CRF de regroupement à l'échelle des segments. Les expériences de la section 6.4 montrent les bénéfices de cette combinaison selon le type d'émission.

6.1 Détection et normalisation des noms

Ce module a été réalisé par les membres du projet SODA et est décrit dans [Gay 2014a]. L'OCR est produit par l'Idiap Research Institute et la détection d'entités nommées par le laboratoire LIUM. Il a servi à extraire des vidéos les noms utilisés par les méthodes d'identification décrites dans ce chapitre. Le but est d'extraire les noms propres à partir des cartouches présents dans la vidéo. Comme le but final est l'identification des personnes présentes, il ne faut garder que les noms annonçant une personne parlant ou apparaissant. Pour extraire le texte incrusté dans les images, le système d'OCR détaillé dans [Chen 2005] et amélioré par Elie Khoury est utilisé. Vient ensuite l'étape de reconnaissance des noms à l'intérieur de ces textes. Elle est assurée par un système de règles imaginé par Sylvain Meignier qui va à présent être décrit en détail.

Hypothèse est faite que les personnes sont annoncées sur deux lignes, la première doit contenir uniquement son nom, la deuxième la fonction qu'elle occupe. Les règles s'appuient sur des ressources externes à la vidéo correspondant à trois listes de noms.

- *Cibles* : cette première liste contient 7345 identités. La plupart correspondent à des journalistes présents dans le corpus, des personnalités politiques, des athlètes et des artistes qui apparaissent souvent dans l'actualité française.
- *freebase* : la deuxième liste est composée de plus de 1,7 millions d'identités extraites du site *freebase* [Bollacker 2008]. Seules les identités des personnes nées après 1900 ont été extraites, ainsi que celles pour lesquelles la date de naissance est inconnue.

– *Prénoms courants* : la dernière liste correspond à 17000 prénoms extraits d'un site web spécialisé.

La détection du nom est résumée dans les 4 étapes suivantes. Elle est seulement appliquée sur la première ligne des textes transcrits qui sera notée par c.

1. Étape de rejet : c est rejetée si le nombre de caractères est inférieur à 3 ou supérieur à 30, ou si c est composée de plus de 10 mots, ou si c contient un mot clé issu d'une liste pré-définie. Cette liste contient principalement des verbes. Cette étape vise à rejeter des lignes ne contenant pas d'identités.

2. c est acceptée si elle correspond à un élément de la liste *Cibles*. La recherche dans la liste tolère une différence : l'identité de la liste est acceptée pour la ligne c si $s(c) > 0,8$ avec :

$$s(c) = \min_{n \in Cibles} (1 - \frac{d(c,n)}{max(l(c), l(n))}) \tag{6.1}$$

   où $d(c,n)$ est la distance de Levenshtein entre c et n et $l(n)$ est la longueur n.

3. c est acceptée si elle correspond exactement à une entrée de la liste *freebase*.

4. Finalement, c est acceptée si elle commence avec un nom présent dans la liste *Prénoms courants*, et si un test sur la célébrité de cette personne utilisant le moteur de recherche Google est concluant. Ce test est une analyse de la page web retournée par le moteur de recherche en réponse à la requête c. Les critères d'analyse se fondent sur la fréquence de c dans la page, la présence d'images associées à c, la présence sur la droite de la page d'une fenêtre décrivant un individu, et la proposition d'une orthographe alternative.

6.2 Identification des visages et des locuteurs à l'échelle des classes

Cette section présente les deux méthodes d'identification à l'échelle des classes. Elles ont été développées au sein du projet SODA mais en parallèle et de manière indépendante. La première méthode, appelée méthode directe, a été imaginée au sein du laboratoire LIUM. La deuxième, fondée sur les CRF, est une contribution de cette thèse. Elle sera combinée dans la section 6.3.4 avec un système de regroupement à l'échelle des segments pour former le système d'identification final. Les expériences permettent de valider le choix de la méthode CRF pour le système final. Les comparaisons avec un oracle évaluent le potentiel d'identification des cartouches et les améliorations qui restent à apporter.

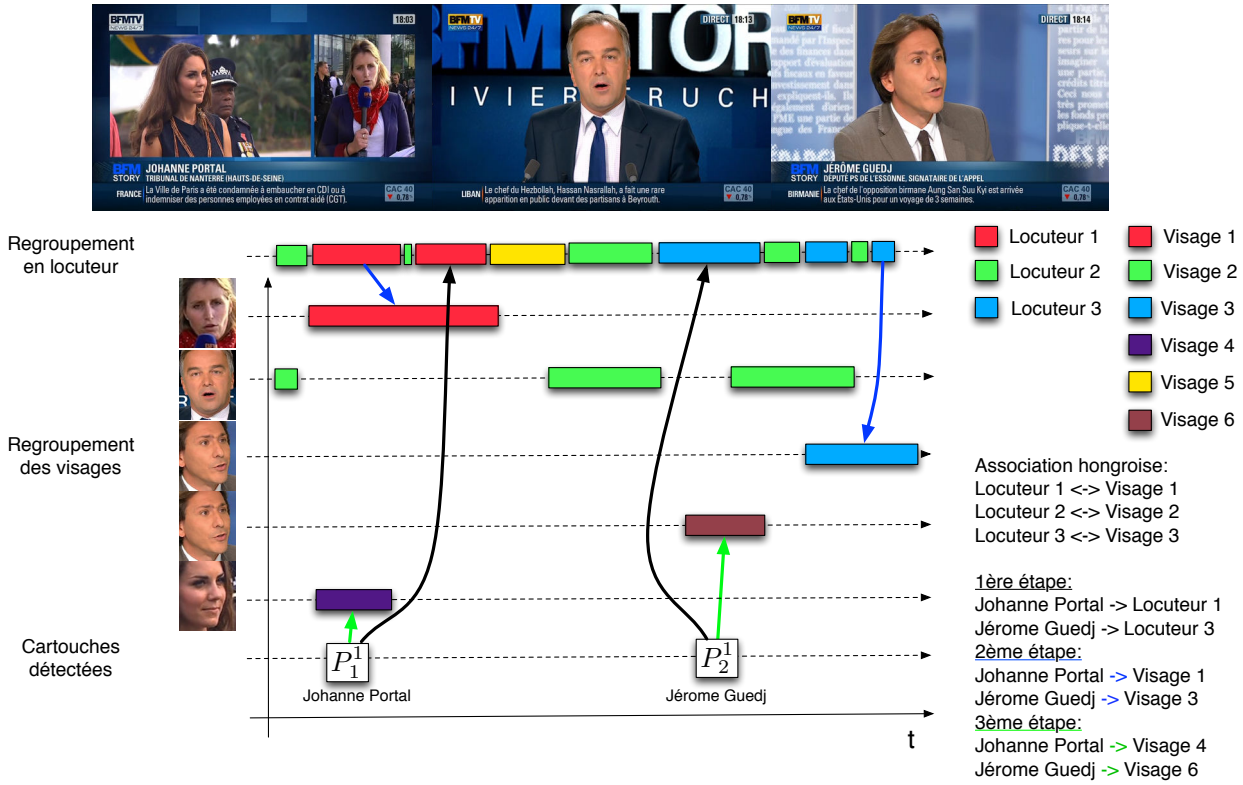


FIG. 6.1: Illustration de la méthode d'identification à base de règles.

6.2.1 Méthode directe de nommage à base de règles

Dans un premier temps, les visages et les locuteurs sont associés en maximisant la co-occurrence temporelle grâce à l'algorithme hongrois. Cela correspond à l'ensemble noté $\mathcal{C}$ dont l'obtention a été décrite dans la section 5.1.3.

La méthode se décompose en 3 étapes, par ailleurs illustrées sur la figure 6.1.

1. Les classes de locuteur sont d'abord identifiées en maximisant la co-occurrence avec les noms des cartouches. Cette étape est expliquée dans le paragraphe suivant.
2. L'identité attribuée à un locuteur est ensuite propagée vers la classe de visage qui lui est associée s'il y en a une.
3. Les classes de visage restées sans nom sont identifiées en maximisant la co-occurrence avec les noms des cartouches de manière similaire à l'étape 1.

Voici comment sont identifiés les locuteurs durant l'étape 1. Soit $P = \{P_j, j = 1..N^P\}$ l'ensemble des noms extraits des cartouches et P_j^k l'occurrence k du nom P_j avec d_j^k la durée d'apparition de cette occurrence. La durée de co-occurrence entre les segments de la classe de locuteur C_i et le cartouche P_j^k est notée $\delta t(C_i, P_j^k)$. La classe C_i est alors identifiée par le nom

P_j donnant la plus grande durée de co-occurrence relative :

$$Nom(C_i) = \arg\max_{P_j \in P} \sum_{k=1}^{K_j} \frac{\delta t(C_i, P_j^k)}{d_j^k} \tag{6.2}$$

La méthode directe n'applique pas la contrainte d'unicité sur les visages (qui oblige deux visages présents sur la même image à avoir des noms différents). Ainsi sur la figure 6.1, le nom *Johanne Portal* est donné à tort aux deux classes de visage 4 et 1. Cependant, comme la métrique d'évaluation EGER (voir section 1.2.2, page 8 pour une définition) ignore la position spatiale des visages, cette situation ne sera pas comptée comme une erreur. En fait, il a été trouvé expérimentalement que le non-respect de cette contrainte est optimal pour l'EGER car cela permet d'augmenter le rappel. En revanche, les apparitions de la personne correspondant au visage 1 seront comptées comme des fausses alarmes si *Johanne Portal* n'est pas présent à ce moment.

6.2.2 Méthode de nommage fondée sur des CRF

L'étape de nommage est vue comme l'estimation des étiquettes $E^N = \{e_i^c, i = 1..N^C\}$ de telle sorte que l'étiquette e_i^c corresponde au nom de la classe C_i. Chaque classe C_i de l'ensemble $C = \{C_i, i = 1..N^C\}$ correspond ici à une classe audiovisuelle. Elle peut donc contenir des tours de parole et/ou des visages. L'étiquette e_i^c est définie sur l'ensemble des noms possibles M, augmenté par l'étiquette $anonyme$. Chaque cartouche P_j se voit attribué une étiquette fixe e_j^p correspondant au nom présent dans ce cartouche. Utilisant le même formalisme que dans la section 5.1, la probabilité *a posteriori* s'exprime par :

$$P(E^N|C,P) = \frac{1}{Z(C,P)} \times exp\left\{ \sum_{i=1}^{n} \sum_{Clique \in G_i} \lambda_i f_i^{ident}(Clique) \right\} \tag{6.3}$$

où n est le nombre de fonctions caractéristiques f_i utilisées. Pour ce système de nommage, $n = 6$. Quatre d'entre elles exploitent des statistiques de co-occurrence entre les classes et les cartouches, et les deux autres servent respectivement à exprimer la contrainte d'unicité et à traiter l'étiquette *anonyme*. À l'exception de celle concernant la contrainte d'unicité, elles sont définies sur les cliques formées des triplets (e_i^c, C_i, P_j^k) pour tous les couples (i, j) tels que $(i, j) \in [1..N^C] \times [1..N^P]$.

La première fonction prend en compte les co-occurrences entre les tours de parole et les cartouches. En reprenant les notations de la section précédente utilisées pour la méthode directe,

elle est définie par :

$$f^{ident}_{audio}(e^c_i, C_i, P^k_j) = \begin{cases} \frac{\delta t^a(C_i, P^k_j)}{d^k_j} & \text{si } e^c_i = e^p_j \\ 0 & \text{sinon} \end{cases} \tag{6.4}$$

où $\delta t^a(C_i, P^k_j)$ retourne la durée de co-occurrence entre le cartouche P^k_j et les tours de parole de C_i.

De la même manière, deux autres fonctions sont définies pour prendre en compte les co-occurrences avec les segments vidéos de C_i :

$$f^{ident}_{vis,seul}(e^c_i, C_i, P^k_j) = \begin{cases} \frac{\delta t^{vis,seul}(C_i, P^k_j)}{d^k_j} & \text{si } e^c_i = e^p_j \\ 0 & \text{sinon} \end{cases} \tag{6.5}$$

où $\delta t^{vis,seul}(C_i, P^k_j)$ retourne la durée de co-occurrence avec les séquences de visage de C_i uniquement quand la classe C_i est seule à l'écran.

$$f^{ident}_{vis,multi}(e^c_i, C_i, P^k_j) = \begin{cases} \frac{\delta t^{vis,multi}(C_i, P^k_j)}{d^k_j} & \text{si } e^c_i = e^p_j \\ 0 & \text{sinon} \end{cases} \tag{6.6}$$

où $\delta t^{vis,multi}(C_i, P^k_j)$ retourne la durée de co-occurrence uniquement si la classe C_i n'est pas le seul visage dans l'image.

Dans la quatrième fonction, l'hypothèse est faite que l'apparition du premier cartouche coïncide avec le premier tour de parole de la personne qu'il annonce. Pour décourager les associations ne respectant pas cette hypothèse, une fonction comptant les tours de parole survenant avant la première occurrence du cartouche est ajoutée :

$$f^{ident}_{prem}(e^c_i, C_i, P_j) = \begin{cases} -\sharp\{A_i \in C_i, fin(A_i) < min_k(debut(P^k_j))\} & \text{si } e^c_i = e^p_j \\ 0 & \text{sinon} \end{cases} \tag{6.7}$$

où l'opérateur $debut(seg)$ retourne le début du segment seg et l'opérateur $fin(seg)$ retourne la fin du segment seg.

Un coût sur l'attribution de l'étiquette anonyme est inclus par la fonction suivante :

$$f^{ident}_{anonyme}(e^c_i, C_i) = \begin{cases} 1 & \text{si } e^c_i = anonyme \\ 0 & \text{sinon} \end{cases} \tag{6.8}$$

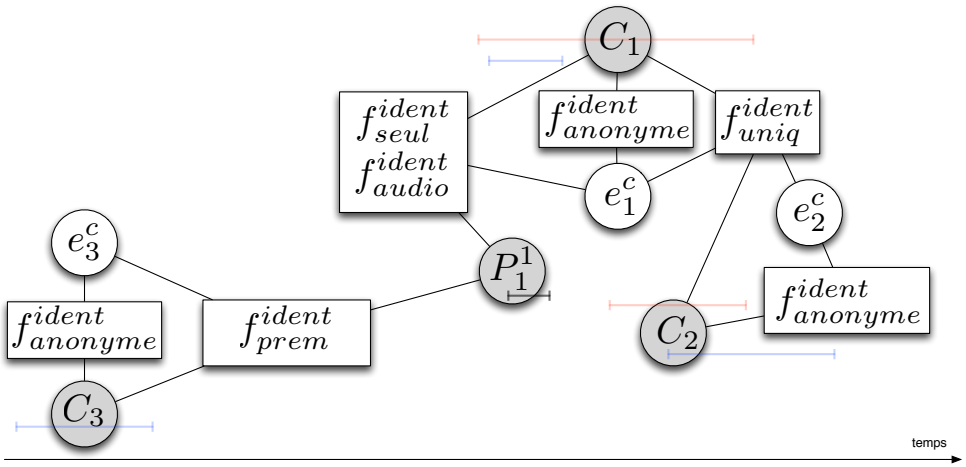


FIG. 6.2: Graphe de facteurs représentant le CRF utilisé pour le nommage. La classe C_1 apparait avec le cartouche P^1_1. Les différentes fonctions de nommage sont donc non nulles entre ces variables. Les classes C_1 et C_2 apparaissent en même temps, la contrainte d'unicité s'applique donc entre ces 2 classes.

Enfin, la contrainte d'unicité est introduite pour chaque paire de classe (C_i, C_j) possédant des séquences de visages apparaissant ensemble.

$$f^{ident}_{uniq}(e^c_i, C_i, e^c_j, C_j) = \begin{cases} -\infty & \text{si } e^c_i = e^c_j \\ 0 & \text{sinon} \end{cases} \tag{6.9}$$

Un graphe de facteurs de ce nouveau CRF est présenté sur la figure 6.2. Au contraire de la méthode directe, ce modèle permet d'obtenir une distribution de probabilités des noms pour chaque classe. Ces probabilités permettront d'inclure l'information de nommage dans le regroupement des segments (section 6.3.3).

6.2.3 Expériences : évaluation du potentiel d'identification des cartouches et des méthodes CRF et directe

Les résultats pour l'identification des voix et des visages sont présentés dans le tableau 6.1. Les sous-tableaux 1 et 4 correspondent à un oracle simulant les meilleurs résultats possibles avec un système de regroupement et d'identification parfait, tout en gardant une segmentation en tours de parole et des détections de visage et de cartouches automatiques. Concernant les erreurs sur l'identification des locuteurs, elles sont principalement dues à des journalistes en voix off, qui ne sont pas annoncés par des cartouches. Pour les visages, les détections manquées sont réparties entre des visages non frontaux et des visages qui ne sont pas annoncés par des cartouches. Les quelques confusions et fausses alarmes sont dues au manque de précision des

Corpus	EGER	#conf	#non-détection	#fa
(1) DEV2 : Oracle				
Locuteur	25.6%	54 (2.4%)	512 (22.6%)	15 (0.6%)
Visage	35.4%	4 (0.2%)	934 (35.2%)	1 (0.0%)
Locuteur + visage	30.9%	58 (1.2%)	1446 (29.4%)	16 (0.3%)
(2) DEV2 : Système "Direct"				
Locuteur	31.5%	280 (12.2%)	417 (18.3%)	17 (1.0%)
Visage	41.5%	118 (4.4%)	947 (35.7%)	34 (1.3%)
Locuteur + visage	36.9%	398 (8.1%)	1364 (27.7%)	51 (1.0%)
(3) DEV2 : Système "CRF"				
Locuteur	30.5 %	214 (9.4%)	463 (20.4%)	15 (0.7%)
Visage	41.7%	62 (2.3%)	1023 (38.6%)	19 (0.7%)
Locuteur + visage	36.5%	276 (5.6%)	1486 (30.2%)	34 (0.7%)
(4) TEST2 : Oracle				
Locuteur	28.9%	138 (3.2%)	1087 (25.4%)	13 (0.3%)
Visage	44.2%	21 (0.4%)	2244 (43.8%)	1 (0.0%)
Locuteur + visage	37.2%	159 (1.7%)	3331 (35.4%)	14 (0.1%)
(5) TEST2 : Système "Direct"				
Locuteur	35.3%	632 (14.7%)	857 (20.0%)	25 (0.6%)
Visage	54.3%	283 (5.5%)	2336 (45.6%)	162 (3.2%)
Locuteur + visage	45.7%	915 (9.7%)	3193 (33.9%)	187 (2.0%)
(6) TEST2 : Système "CRF"				
Locuteur	33.1%	546(12.7%)	846(19.7%)	25(0.6%)
Visage	54.5%	287(5.6%)	2302(44.9%)	200(3.9%)
Locuteur + visage	44.7%	833(8.9%)	3148(33.4%)	225(2.4%)

TAB. 6.1: EGER mesuré pour les tâches de l'identification du locuteur et des visages pour un oracle et les méthodes directe et CRF. Sont aussi reportés le nombre d'erreurs de confusion (#conf), de non-détections (#non-détection) et de fausses alarmes (#fa).

frontières des segments (par exemple, si un segment audio recouvre en partie deux tours de parole consécutifs appartenant à deux personnes). L'oracle fournit une borne pour les résultats d'identification et est révélateur du potentiel d'identification des cartouches.

Les sous-tableaux 2 et 5 présentent les résultats pour la méthode à base de règles. La diminution du rappel par rapport à l'oracle est principalement due à des erreurs dans le regroupement quand une personne est divisée en plusieurs classes. En effet, certaines classes plus petites ne vont pas avoir de co-occurrences avec un cartouche et ne pourront donc pas être nommées. La parcimonie des cartouches renforce la dépendance de la qualité de l'identification à la qualité du regroupement.

Considérant les comparaisons entre la méthode directe et la méthode à base de CRF, 2 différences peuvent être soulignées.

1. Le CRF tend à donner de meilleures performances pour l'identification des locuteurs avec une réduction de l'EGER de 31.5% à 30.6% sur le DEV2 et de 35.3% à 33.1% sur le TEST2. Ces améliorations viennent principalement d'une réduction du nombre de confusions (280 vs. 218 pour DEV2 et 632 vs 546 pour TEST2). Ceci peut s'expliquer par le fait que le CRF utilise plus de caractéristiques que la simple co-occurrence temporelle.
2. Deuxièmement, considérant l'identification des visages, la méthode directe tend à produire plus de noms au prix d'une confusion globalement plus grande sur le DEV2. En effet, cette méthode essaie d'identifier le plus possible de visages, jusqu'à nommer deux visages apparaissant dans la même image avec le même nom. Ce non-respect de la contrainte d'unicité a des conséquences limitées par rapport à la métrique EGER car la position du visage n'est pas incluse dans la métrique. Le CRF, en respectant la contrainte, est plus précis au prix d'un rappel inférieur.

6.3 Intégration de la compréhension des scènes et des informations de nommage dans le regroupement

Il a été régulièrement avancé dans le chapitre 3 de l'état de l'art que les performances en regroupement et en identification sont liées. C'est tout particulièrement vérifié si la source de nommage est parcimonieuse, comme c'est le cas des cartouches. En effet, si une personne est séparée en plusieurs classes, seules les classes co-occurrentes avec les cartouches peuvent être nommées. Or, il a été noté dans la section 4.3.2, page 67 que le système de regroupement en classes de visages produit plus de classes que de personnes, principalement à cause de variations de la pose des visages dans les scènes de plateau. La différence de performances entre l'oracle et le système automatique dans le tableau 6.1 est une autre raison de penser que le regroupement pourrait être amélioré pour permettre une meilleure identification. La section suivante présente notre utilisation des cartouches et de l'arrière plan des images afin d'améliorer le regroupement et l'identification.

6.3.1 Classification des visages en acteurs et figurants fondée sur la répétition de l'arrière-plan

La section 2.2.2, page 22 de l'état de l'art mentionne l'utilisation de contexte tels que l'arrière-plan, les vêtements, la co-occurrence, le type de scène ou les cartouches pour améliorer

le regroupement. Gardant en tête ces différents éléments, il semble intéressant de chercher quelles informations de contexte pourraient être utilisées dans les journaux télévisés. Dans ce cadre, nous remarquons notamment les tendances suivantes.

- Les personnes apparaissent très majoritairement dans le même type de scène : un invité en plateau n'apparaîtra que très rarement dans un reportage et vice versa.
- L'apparition d'une personne est généralement limitée à une partie de l'émission pendant laquelle est traité le sujet la concernant. Bien sur, ceci ne s'applique pas aux présentateurs qui sont présents tout le temps.
- Dans le corpus REPERE, les cartouches sont révélateurs du nombre de locuteurs de l'émission (à l'exception de certains journalistes).
- Du point de vue applicatif, la notion de scène participe également à la structuration et à l'indexation des vidéos. Elle constitue donc une information intéressante à avoir, d'un point de vue applicatif.

Par conséquent, la segmentation en scène et l'utilisation de l'information venant des cartouches semblent être des éléments intéressants pour le regroupement.

Le problème de la segmentation en scène est généralement traité comme un regroupement des plans. De nombreuses approches ont été proposées dans la littérature [Odobez 2003, Parkhi 2013], en utilisant notamment des combinaisons de caractéristiques visuelles, acoustiques, textuelles ou extraites du regroupement des personnes. Une approche exploitant ces différentes caractéristiques est sans doute préférable, cependant, ne disposant pas de tels outils, nous nous limiterons à l'utilisation de caractéristiques visuelles.

La notion de scène est dépendante de l'application. Dans notre cadre, il serait souhaitable qu'une scène soit une partie de la vidéo homogène en terme de personnes présentes. Afin d'obtenir ce type d'information, un regroupement des plans a été effectué en fonction de leurs similarités visuelles. La distance fondée sur les SIFT (décrite dans la section 4.2.1, page 60) a été utilisée dans un regroupement hiérarchique. Dans notre cadre, nous avons intérêt à ce que chaque cluster[5] de plan obtenu corresponde à un seul lieu afin qu'il soit homogène en terme de personnes présentes. Comme certains journaux télévisés intègrent dans la même image plusieurs plans qui correspondent à différents lieux, (voir figure 6.3), nous traitons séparément les régions situées autour de chaque visage. Une illustration des résultats du regroupement obtenu est présenté sur la figure 6.4. Il est remarquable que les clusters contenant plusieurs

[5]Pour le regroupement d'arrières-plans, il sera question de *clusters* et non de *classes* afin d'éviter la confusion avec les classes de personnes

FIG. 6.3: Exemple d'une image divisée en deux plans. La partie de gauche correspond à des illustrations du sujet traité. Afin de ne considérer que l'arrière-plan propre à chaque visage, la région d'intérêt sur laquelle est effectué le regroupement pour le visage détecté en bleu est restreinte au cadre rouge.

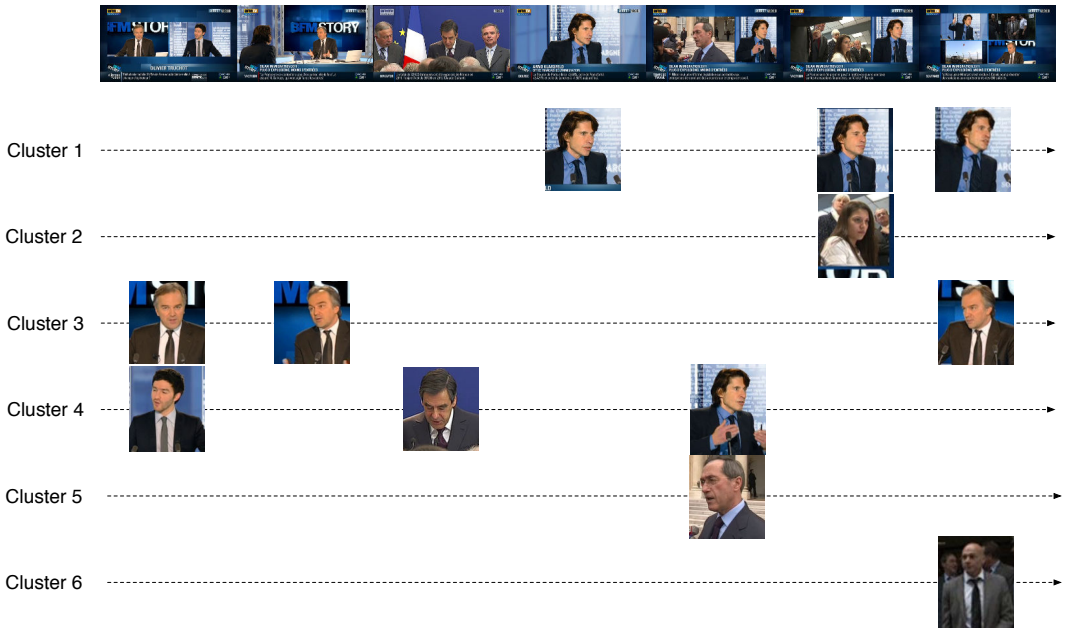


FIG. 6.4: Illustration des résultats du regroupement des arrières-plans de chaque visage. Les images du dessus correspondent aux images originales. Chaque ligne correspond à un cluster.

éléments (clusters 1, 3 et 4) correspondent généralement à des locuteurs (généralement car, par exemple, ce n'est pas cas du visage du milieu sur le cluster 4). Sachant que dans les données REPERE, les locuteurs apparaissant à l'écran sont généralement annoncés par un cartouche, nous formulons l'hypothèse de récurrence de l'arrière-plan des locuteurs (notée par la suite HRAL) suivante :

Hypothèse HRAL. *Les visages dont l'arrière-plan est étiqueté comme récurrent correspondent à des locuteurs annoncés par un cartouche.*

Ici, la récurrence d'un arrière-plan est caractérisé par la taille du cluster auquel il appartient. Simplement, il est considéré comme récurrent quand la taille de son cluster est supérieure à un seuil k.

6.3.2 Intégration de l'arrière-plan dans le regroupement audiovisuel des personnes

À présent, l'objectif est d'ajouter des fonctions au CRF de regroupement décrit dans l'équation 5.4 pour prendre en compte l'hypothèse HRAL décrite dans la section précédente. Afin d'encourager les séquences de visages classées comme récurrentes à rejoindre des classes susceptibles d'être nommées, la fonction suivante est définie pour chaque séquence V_i :

$$\forall V_i \in V,\ f^k_{recv}(V_i, e^v_i) = \begin{cases} 1 & \text{si } e^v_i \in \mathcal{E}^{opn} \text{ ET } \#(R_i) > k \\ 0 & \text{sinon} \end{cases} \tag{6.10}$$

avec :

- R_i correspondant au cluster issu du regroupement des arrières-plans dans laquelle apparaît la séquence V_i.
- $\#(R_i)$ correspondant au nombre d'éléments dans le cluster R_i.
- $\mathcal{E}^{opn}$ étant l'ensemble des indices de classes de personnes qui apparaissent seules avec un cartouche ou possèdent un tour de parole coïncidant avec un cartouche. Cet ensemble est fixé avant d'effectuer le décodage du CRF.

Une fonction similaire est utilisée pour les tours de parole si ils apparaissent en même temps qu'un arrière-plan récurrent :

$$\forall A_i \in A, f^k_{reca}(A_i, e^a_i) = \begin{cases} 1 & \text{si } \exists V_j, \#(R_j) > k, t(A_i, V_j) > 0, e^a_j \in \mathcal{E}^{opn} \\ 0 & \text{sinon} \end{cases} \tag{6.11}$$

avec $t(A_i, V_j)$ la durée de co-occurrence entre le tour de parole A_i et la séquence de visages V_j.

Ces fonctions ont plusieurs propriétés intéressantes pour le regroupement.

- Elle introduisent une notion de rôle (intervenant *versus* figurant) pour chaque segment sans avoir besoin de calculer des statistiques à l'échelle de la classe.
- Elle permettent d'influer sur le nombre de classes de personne nommées. Si la contrainte devait être stricte, le nombre de classes des segments concernés par ces fonctions correspondrait au nombre de noms différents extraits des cartouches. Ici, le CRF permet de relâcher la contrainte et de pondérer sa contribution à travers les paramètres λ. En pratique, plusieurs valeurs de k sont utilisées en même temps et un paramètre λ est appris pour chaque valeur de k.

Il faut noter qu'il est supposé que les cartouches ont été extraits sans erreur. Cependant, les données REPERE offrent des vidéos de suffisamment bonne résolution pour que ce soit le cas

avec les techniques d'OCR actuelles. Si ce n'est pas le cas, il est probable que les paramètres λ du CRF auront tendance à minimiser le poids de ces fonctions.

6.3.3 Intégration des cartouches dans le regroupement

Le but ici est de favoriser l'attribution des segments apparaissant avec une occurrence du nom P_j vers des classes susceptibles de correspondre à ce nom. Bien qu'il soient peu nombreux, il est important que ces segments soient bien classés car ce sont eux qui contiennent l'information de nommage. Il faut rappeler que les cartouches auraient pu être introduits plus précocement : dans [Poignant 2013a], il est optimal de les introduire dés l'étape de regroupement monomodal (voir section 3.3, page 48 pour une description du système). Le cadre dans lequel nous nous plaçons suppose que le système de regroupement est une boîte noire et que le CRF proposé permet d'en améliorer les sorties en post-traitement.

Trois cas sont distingués de la même manière que pour le modèle de nommage (section 6.2.2, page 98). Comme pour cette section, chaque cartouche P_j se voit attribuer une étiquette fixe correspondant à son nom e_j^p. Pour chaque couple (V_i, P_j^k) pour lequel V_i est le seul visage à l'image, la fonction est définie par :

$$f_{seul}^{regr}(V_i, P_j^k, e_i^v) = p(e_i^c = e_j^p | C, P) \tag{6.12}$$

où $p(e_i^v = e_j^p | C, P)$ est la probabilité que le nom e_j^p corresponde à e_i^c étant donné le regroupement C et les cartouches P, avec e_i^c l'étiquette de nommage de la classe e_i^v. Cette probabilité est calculée suivant le modèle CRF de nommage présenté dans l'équation 6.3, le regroupement C correspond au regroupement initial avant d'effectuer le décodage du CRF. Cette utilisation conjointe des deux CRF sera d'avantage détaillée dans la section suivante.

Pour chaque couple (V_i, P_j^k) pour lequel plusieurs visages sont présents en même temps que V_i, la fonction suivante est définie :

$$f_{multi}^{regr}(V_i, P_j^k, e_i^v) = p(e_i^c = e_j^p | C, P) \tag{6.13}$$

Enfin pour chaque couple (A_i, P_j^k) pour lesquels les segments A_i et P_j^k se chevauchent :

$$f_{audio}^{regr}(A_i, P_j^k, e_i^a) = p(e_i^c = e_j^p | C, P) \tag{6.14}$$

La fonction f_{multi}^{regr} favorise les multiples trajectoires de visages apparaissant en même temps qu'un cartouche P_j^k donné à rejoindre la classe liée au nom du cartouche e_j^p. Le risque est alors

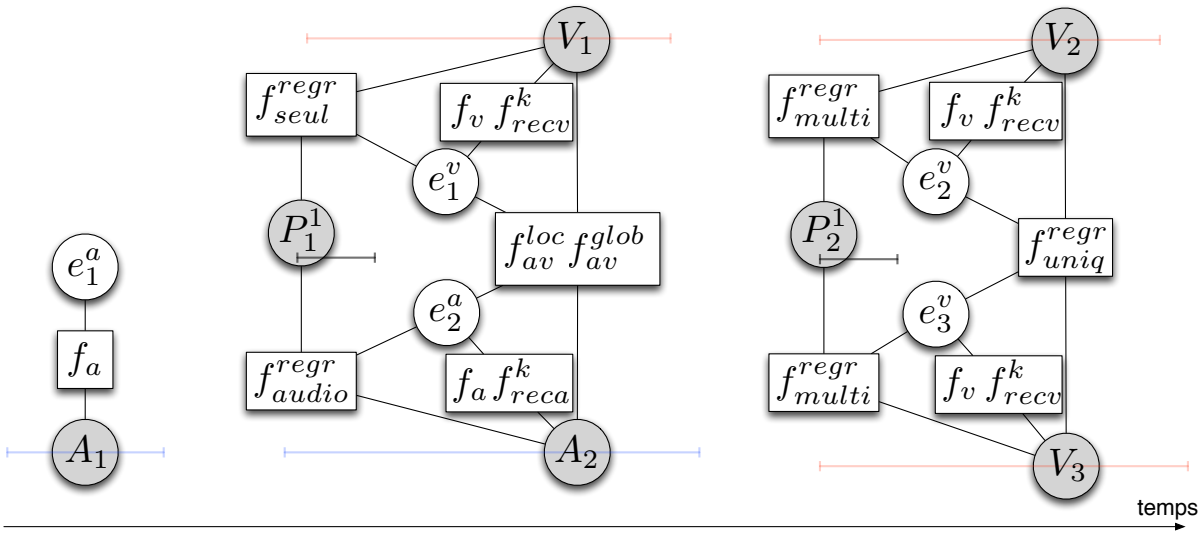


FIG. 6.5: Graphe de facteur montrant le modèle CRF intégrant les cartouches et l'arrière-plan pour le regroupement. Comme sur la figure 5.3, les segments A_2 et V_1 sont dépendants à travers les fonctions d'association audiovisuelle car ils se recouvrent. À présent, leur vraisemblance dépend aussi du cartouche P^1_1. À droite, les séquences de visages V_2 et V_3 sont liées au cartouche P^1_2. Comme ces deux séquences se recouvrent, la contrainte d'unicité f^{regr}_{uniq} est appliquée.

que ces visages sur la même image rejoignent la même classe. Pour cette raison, à ce stade, il est utile de rajouter la contrainte d'unicité. Elle est implémentée par la fonction f^{regr}_{uniq} définie pour toutes les paires co-occurrentes (V_i, V_j) si ces deux séquences co-occurrent avec un cartouche.

$$f^{regr}_{uniq}(V_i, V_j, e^v_i, e^v_j) = \begin{cases} -Inf & \text{if } e^v_i = e^v_j \\ 0 & \text{otherwise} \end{cases} \tag{6.15}$$

Il a été considéré d'appliquer la contrainte d'unicité pour tous les cas où il y a plus de deux visages sur l'image. Cependant, le graphe devenait plus complexe à traiter et l'algorithme de décodage ne convergeait plus. Des expériences ont été effectuées où la contrainte était appliquée *a posteriori*, cependant, les améliorations n'ont pas été significatives.

Le modèle CRF pour le regroupement avec ses nouvelles fonctions est illustré sous la forme d'un graphe de facteur dans la figure 6.5.

6.3.4 Système d'identification final : utilisation conjointe des deux CRF pour l'identification

Nous disposons à présent de deux CRF. Le premier concerne le regroupement audiovisuel des personnes. Il a été introduit dans le chapitre 5 et vient d'être enrichi de plusieurs fonctions

afin de prendre en compte l'arrière-plan et les cartouches. Le second est utile pour l'identification des classes de visage et de locuteur et a été présenté dans la section 6.2.2.

Il a été décrit dans la section 5.1.4 que l'utilisation du CRF pour le regroupement suit un processus itératif alternant étiquetage des segments et mise à jour des modèles biométriques. À présent, les probabilités d'identification d'une classe sont utilisées dans le modèle de regroupement par les fonctions traitant les cartouches. Le diagramme 6.6 résume l'utilisation conjointe des deux CRF.

1. Effectuer séparément un regroupement en locuteur et un regroupement en classes de visages.
2. Associer les classes des locuteurs et des visages avec l'algorithme hongrois pour initialiser l'ensemble $\mathcal{E}$ des étiquettes du CRF de regroupement.
3. – Pour chaque étiquette, apprendre le modèle biométrique à partir des données qui y sont associées.
   – Pour chaque étiquette $e \in \mathcal{E}$ et chaque nom extrait $P_j \in P$, utiliser le CRF de nommage pour estimer les probabilités d'identification $P(P_j|e, P))$.
4. Obtenir les étiquettes les plus probables pour les segments $\{A, V\}$ avec le CRF de regroupement.

Il est alors possible d'itérer en alternant la mise à jour des modèles des classes et la mise à jour de leurs probabilités d'identification (étape 3) avec l'étiquetage des segments (étape 4). Une fois que l'algorithme a convergé, c'est à dire quand le ré-étiquetage ne change plus les étiquettes des segments, les probabilités d'identification peuvent être utilisées pour attribuer un nom à chaque classe. Encore une fois, la convergence n'est pas garantie théoriquement et un nombre fixe d'itérations n est utilisé dans les expériences.

6.4 Expériences : Effet de l'optimisation du regroupement pour le nommage des visages et des voix

Afin de mesurer les performances en identification, la métrique officielle EGER de la campagne REPERE décrite dans la section 1.2.2, page 8 est à nouveau utilisée. Quatre systèmes peuvent à présent être comparés. Trois d'entre eux ont été décrits dans la section 6.2 : la méthode de nommage directe à base de règles, la méthode CRF de nommage à l'échelle des classes et un oracle sur le regroupement et l'identification. Le quatrième système est l'utilisation combinée du CRF de nommage et du CRF pour le regroupement.

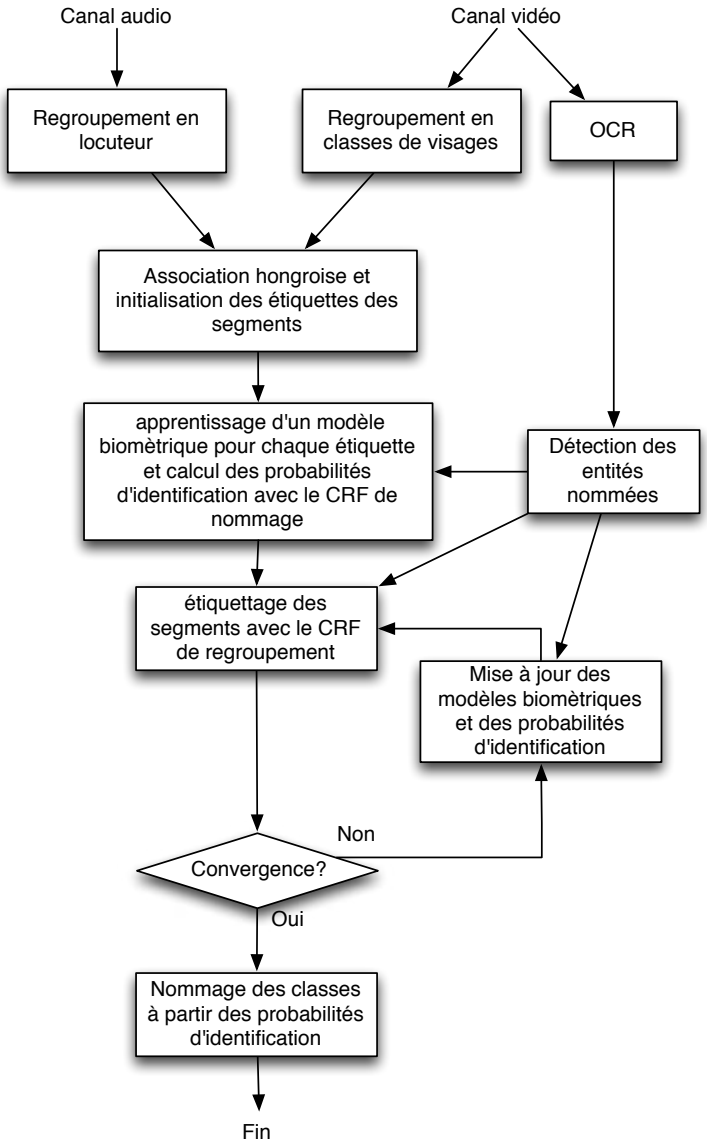


FIG. 6.6: Diagramme représentant l'utilisation conjointe des deux CRF pour l'identification des visages et des locuteurs.

Le protocole expérimental est semblable au chapitre précédent. Les paramètres λ du CRF sont appris sur le corpus TEST0. L'optimisation des hyper-paramètres est faite sur le DEV2. Ici aussi, le nombre d'itération n est fixé à 3. Les nouveaux paramètres à optimiser sont le seuil d'arrêt du regroupement des arrières-plans et les valeurs de k correspondant aux niveaux de récurrence des fonctions f^k_{rec} (section 6.3.2). Dans ces expériences, 3 valeurs de k sont utilisées dans le CRF.

	DEV2			TEST2		
	Locuteurs	Visages	Locuteurs + Visages	Locuteurs	Visages	Locuteurs + Visages
Méthode directe	31.5%	41.5%	36.9%	35.3%	54.3%	45.7%
CRF Nom.	30.5%	41.8%	36.5%	33.1%	54.5%	44.7%
CRF Nom. + Reg.	29.7%	40.1%	35.3%	31.4%	52.2%	42.7%
Oracle	25.6%	35.4%	30.9%	28.9%	44.2%	37.2%

TAB. 6.2: Performances d'identification des locuteurs et des visages pour les différences systèmes mesurées en EGER.

Les performances globales en EGER pour les différents systèmes sont résumées dans le tableau 6.2. Le reste de cette section détaille plus précisément l'apport du CRF de regroupement.

Les résultats reportés dans le tableau 6.3 comparent l'utilisation du CRF de nommage seul et l'utilisation conjointe des deux CRF. L'ajout de l'étape de regroupement permet de gagner globalement entre 1 et 2 point d'EGER par rapport à l'utilisation du CRF de nommage seul. Cependant, les différences de performances entre les deux systèmes varient largement d'une émission à l'autre.

Les améliorations les plus importantes sont mesurées dans les émissions de débat comme *EntreLesLignes, CaVousRegarde* et *PileEtFace*. Dans ces émissions, l'arrière-plan varie peu. De plus, le présentateur et tous les invités sont annoncés par un cartouche. Ainsi, l'hypothèse HRAL est valable, et de nombreux visages et tours de parole sont étiquetés comme récurrents. Par conséquent, les fonctions f_{rec}^{k} ont un impact significatif et positif sur le regroupement. Elles permettent en particulier de résoudre les principales erreurs de confusion du regroupement vidéo initial survenant dans les scènes de plateau qui, comme il a été vu dans la section 4.3.2, page 67, sont dues aux variations de pose des visages. La figure 6.7 reprend deux exemples de confusion survenus avec le regroupement initial qui sont maintenant résolus grâce au CRF de regroupement.

Afin d'appuyer cette interprétation, il semble pertinent de comparer les performances en identification mesurées avec la métrique EGER et les performances en regroupement mesurées avec la métrique DER. Or, il a bien été observé que la ré-attribution des segments par le CRF de regroupement permet d'améliorer l'identification et le regroupement pour les émissions de débats. La comparaison des performances est affichée sur la figure 6.8. Il est clair que la majorité des vidéos de débats, correspondant aux points bleus est dans le cadran nord-est du graphique. Concernant les journaux télévisés et les questions à l'assemblée nationale, la corrélation n'est

corpus	émission	Locuteurs		Visages		Locuteurs + Visages	
		Nom.	Nom. + Reg.	Nom.	Nom. + Reg.	Nom.	Nom. + Reg.
DEV2	BFMStory	28.5	**27.3**	39.6	**38.8**	34.3	**33.3**
	CaVousRegarde	25.9	**24.9**	26.9	**22.5**	26.5	**23.5**
	CultureEtVous	**84.2**	85.1	**90.2**	91.2	**87.1**	88.0
	EntreLesLignes	11.1	**8.6**	30.0	**26.2**	22.2	**19.0**
	LCPInfo	47.4	**47.4**	53.5	**52.1**	50.5	**49.8**
	PileEtFace	17.7	**16.0**	19.1	**17.0**	18.5	**16.5**
	TopQuestions	8.0	**7.2**	43.0	**41.6**	28.1	**26.9**
	TOUS DEV2	30.6	**29.7**	41.8	**40.1**	36.6	**35.3**
TEST2	BFMStory	31.8	**30.8**	65.9	**65.4**	49.7	**49.0**
	CaVousRegarde	**44.6**	**44.6**	65.5	**62.8**	55.6	**54.2**
	CultureEtVous	**85.6**	85.8	**83.9**	86.6	**84.8**	86.2
	EntreLesLignes	17.7	**15.1**	52.2	**46.9**	39.4	**35.1**
	LCPActu	**17.9**	**17.9**	38.8	**32.6**	28.9	**25.1**
	LCPInfo	31.9	**31.4**	50.4	**46.1**	41.3	**38.9**
	PileEtFace	13.8	**7.8**	26.3	**23.1**	21.0	**16.8**
	RuthElkrief	38.7	**38.0**	42.4	**41.9**	40.9	**40.3**
	TopQuestions	11.3	**8.7**	62.2	**59.6**	40.6	**38.0**
	TOUS TEST2	33.1	**31.4**	54.5	**52.2**	44.7	**42.7**

TAB. 6.3: Résultats d'identification en terme d'EGER mesurée sur les locuteurs, sur les visages, et sur les locuteurs et les visages ensemble. Pour chacune de ces trois tâches, la première colonne correspond au CRF de nommage, la deuxième correspond à l'utilisation jointe des deux CRF.

pas aussi évidente. La présence d'anonymes et de journalistes en voix off implique qu'une différence en DER ne se répercute pas forcément sur l'EGER. Ainsi, la distribution des points noirs et verts est plus uniforme, bien que centrée globalement dans le cadran nord-est. Par ailleurs, le regroupement CRF a clairement tendance à dégrader l'identification et le regroupement pour les magazines People. Comme nous l'avons montré, le regroupement CRF est performant dans les scènes de type plateaux. Or, ce type de scène est peu présente dans ces émissions qui sont essentiellement composées de scènes de reportages. D'une manière générale, la conclusion réside dans le fait que les améliorations en DER et en EGER sont d'autant plus conséquentes dans une vidéo quand celle-ci contient une plus grande proportion de scènes de plateaux. Une information similaire pourrait être récupéré en segmentant et en discriminant au sein de chaque image les zones correspondant à des interviews et les zones correspondant à des reportages. Une telle approche pourrait être une alternative à la méthode que nous avons mise en place.

FIG. 6.7: Ces exemples sont des erreurs de confusion extraits de deux vidéos différentes où les visages d'une personne sont séparés à tort en deux classes à cause de variations de poses. Étant donné le grand nombre de plans de ce type dans chaque vidéo, les arrières-plans de ces séquences de visage sont étiquetés comme récurrents. Elles sont donc encouragées à rejoindre les classes chevauchant des cartouches. Pour l'exemple du haut (respectivement du bas) la séquence $V2$ (respectivement les séquences $V1$ et $V4$) est ré-attribuée à la classe 1 (respectivement sont ré-attribuées à la classe 2).

Des améliorations sont aussi également observables dans l'identification des locuteurs. Comme pour les visages, elles sont globalement plus prononcées dans les émissions de débats. D'ailleurs, la figure comparant l'influence du regroupement CRF sur l'EGER et le DER pour les locuteurs présente les mêmes schémas que pour les visages, bien que de manière moins prononcée (voir figure 6.9). Il y a deux possibilités pour expliquer cette amélioration : la modélisation de contexte semble non seulement avoir des effets positifs, mais le fait que regroupement et identification soient effectués de manière jointe sur les visages et les locuteurs semble également jouer un rôle. Ainsi, l'amélioration du regroupement d'une modalité permet d'améliorer l'identification de l'autre.

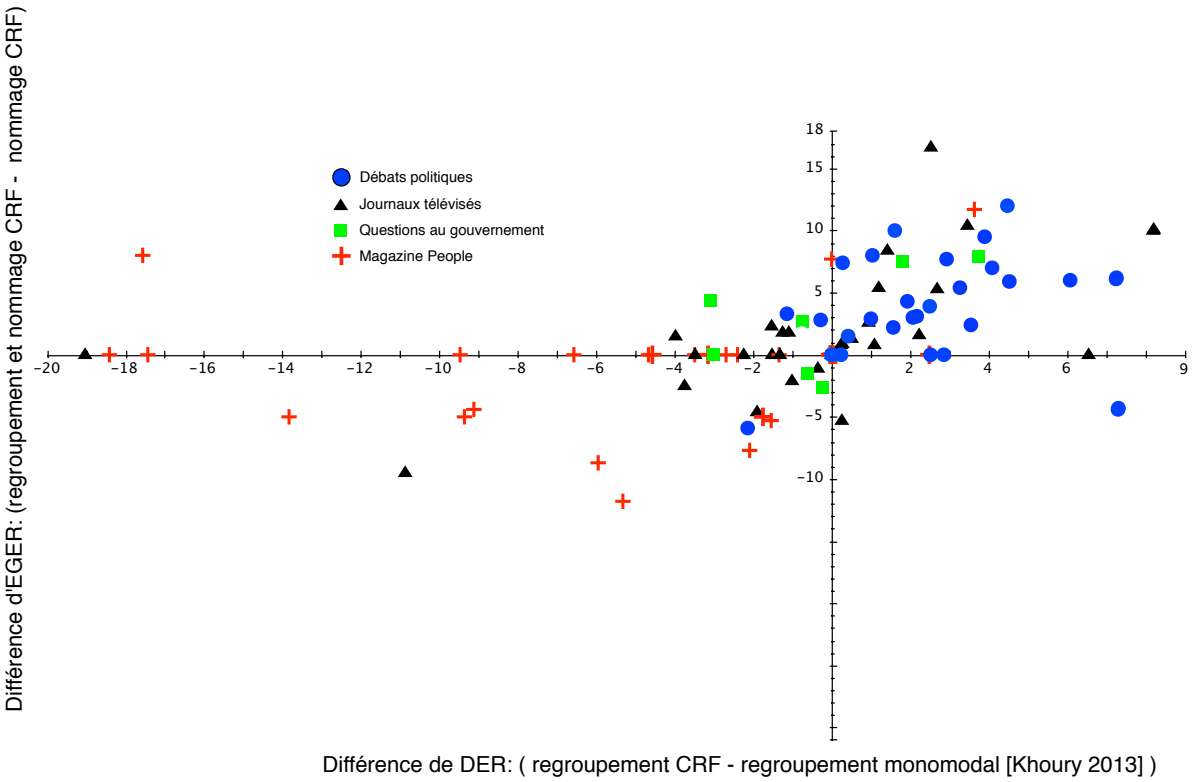


FIG. 6.8: Comparaison des effets du CRF de regroupement sur le DER et l'EGER pour les visages. Chaque point correspond à une vidéo des corpus TEST2 et DEV2. Un code couleur illustre le type de l'émission. L'abscisse correspond à la différence en terme de DER entre le modèle de regroupement CRF et le regroupement produit par le système monomodal publié dans [Khoury 2013]. L'ordonnée correspond à la différence en terme d'EGER entre l'utilisation conjointe des deux CRF et l'utilisation du CRF de nommage seul appliqué directement au regroupement monomodal.

Il faut enfin noter qu'il existe des cas où l'hypothèse n'est pas valable : nous pensons notamment au cas où des visages dans des arrières-plans récurrents ne correspondent pas à des locuteurs. Des émissions comme *Top Questions* où le public est filmé à de nombreuses reprises constituent à cet égard de bons exemples. Le modèle peut alors être conduit à nommer à tort ces visages du public, générant une perte de précision. Ceci souligne l'importance d'utiliser un modèle comme les CRF capable d'apprendre la pondération le contexte et la similarité des données (ici la similarité visuelle) entre le segment et la classe. Le cas de *CultureEtVous* est à part car les cartouches y fournissent très peu de noms. De plus, cette émission contient très peu de plateau, la majorité des scènes étant constituée de reportages et d'extraits de films ou de concerts, ce qui rend le regroupement difficile. Peu d'efforts de modélisation ont été faits pour améliorer le traitement de cette émission à cause de sa singularité par rapport aux autres émissions et de sa faible représentation dans le corpus.

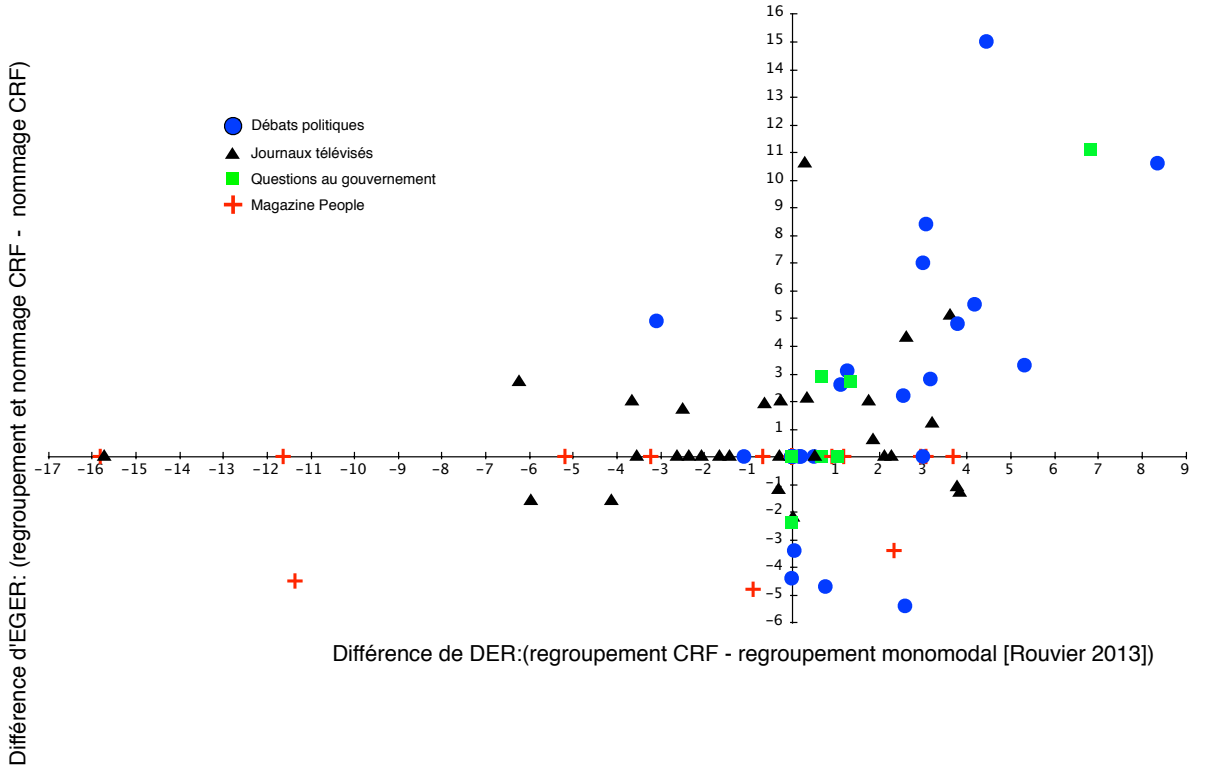


FIG. 6.9: Comparaison des effets du CRF de regroupement sur le DER et l'EGER pour les locuteurs. De manière similaire à la figure 6.8, chaque point correspond à une vidéo. Un code couleur illustre le type de l'émission. L'abscisse correspond à la différence de DER entre le regroupement produit par le système monomodal publié dans [Rouvier 2013] et le modèle de regroupement CRF. L'ordonnée correspond à la différence en termes d'EGER entre l'utilisation jointe des deux CRF et le CRF de nommage seul.

Afin de quantifier les progrès qui peuvent encore être faits par l'amélioration du regroupement et du processus d'identification, les performances en EGER par émission ont été mesurées avec un oracle. Cet oracle utilise une détection automatique des segments et des cartouches mais un regroupement et un processus d'identification utilisant les références (c'est le même que celui utilisé dans la section 6.2.3). Les conclusions sur la comparaison des résultats rejoignent celles formulées précédemment : la plupart des erreurs dans les émissions de débat sont corrigées mais de plus grandes améliorations peuvent encore être obtenues dans le cas des reportages. Ces résultats ont été reportés avec ceux des méthodes automatiques dans la figure 6.10.

Les améliorations obtenues en intégrant le contexte et les cartouches dans le regroupement sont conséquentes étant donné la simplicité des fonctions utilisées (par rapport à l'utilisation

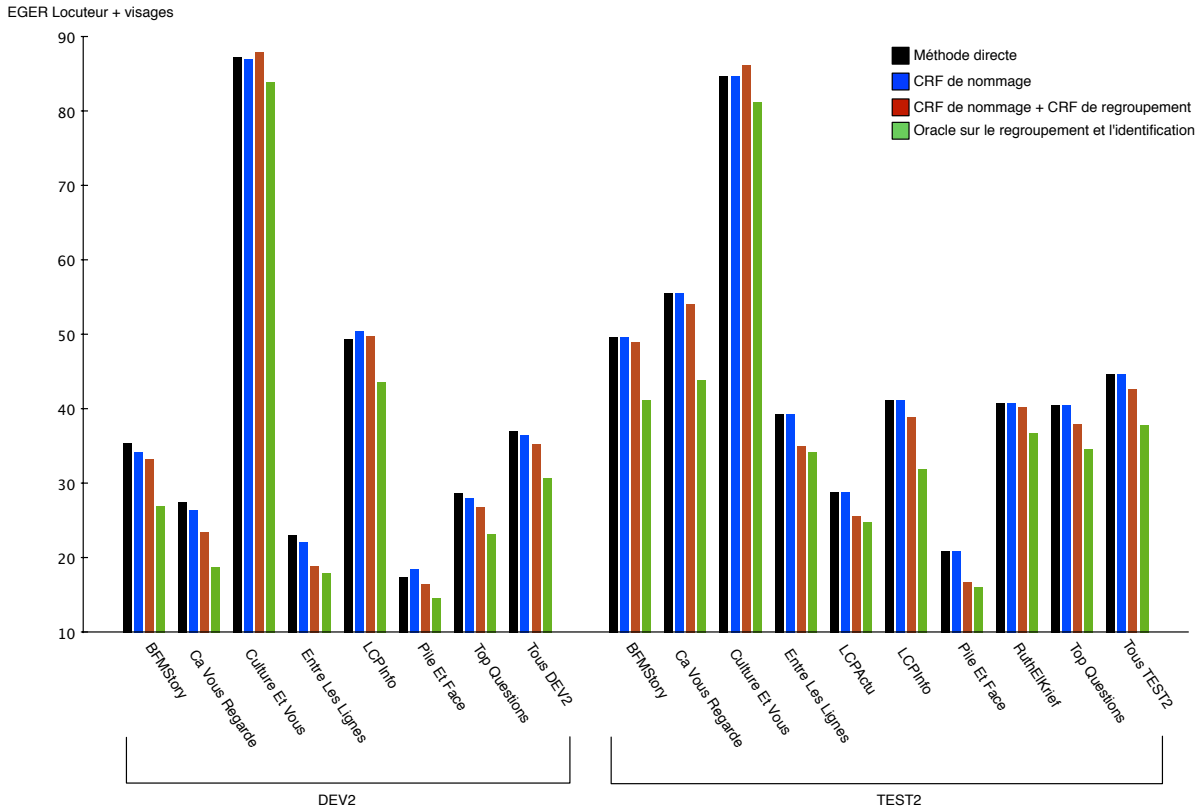


FIG. 6.10: Résultats d'identification des visages et des locuteurs en terme d'EGER pour l'utilisation de la méthode directe affichée en noir, du système d'identification avec le CRF de nommage seul en bleu, l'utilisation jointe avec le CRF pour le regroupement en rouge, et l'oracle sur le regroupement et la décision utilisé dans le tableau 6.1, en vert.

d'une segmentation complète en scène et en chapitre). De plus, les hypothèses formulées sur les données sont suffisamment génériques pour qu'il ne soit pas nécessaire de connaître à l'avance le type d'émission traitée. À l'inverse, le système proposé dans [Bechet 2014] obtient de meilleures performances, mais apprend à l'avance un modèle pour chaque type d'émission. Le score EGER obtenu est de 30.9% pour les locuteurs et 39.4% pour les visages. Les performances pour les visages sont même supérieures à notre oracle. L'amélioration par rapport à l'oracle peut venir de l'utilisation de la transcription et de listes de noms par émission, ce qui permet de nommer les journalistes en voix-off. De plus, l'apprentissage de modèles de plateau et de composition leur permet de prédire la présence de visages. Ainsi il n'y a pas de problème de non-détection de visages de profil et le rappel est augmenté.

6.5 Conclusions

Dans ce chapitre, une méthode d'identification des locuteurs et des visages à partir des cartouches a été proposée. Sur les données REPERE, cette source de nommage permet à un oracle de nommer la plupart des personnes, à part les voix-off des journalistes et les visages apparaissant en arrière-plan dans les reportages.

L'identification à l'échelle des classes a été tout d'abord étudiée. Une méthode de nommage à base de CRF a été comparée avec une autre méthode développée en parallèle dans le cadre du projet SODA. Le CRF de nommage montre des résultats compétitifs, qui sont probablement dus à la plus grande richesse des caractéristiques qu'il utilise, ce qui nous permet de valider son utilité.

Les travaux portent ensuite sur l'amélioration du regroupement audiovisuel des personnes à l'échelle des segments (tours de parole et séquences de visage) afin de faciliter l'identification effectuée *a posteriori* sur les classes. Le modèle CRF de regroupement proposé dans le chapitre 5 est enrichi avec des éléments extraits des cartouches et du contexte visuel. Le contexte qui est utilisé ici est une caractérisation de la récurrence de l'arrière-plan des visages. L'hypothèse est faite que les arrières-plans récurrents correspondent à des zones où les visages et les locuteurs ont une plus grande probabilité d'être nommés par un nom extrait d'un cartouche. La méthode d'identification finale s'appuie sur l'utilisation conjointe de ces deux CRF. Cette combinaison permet d'améliorer les performances par rapport à l'utilisation seule du CRF de nommage dans la majorité des émissions, en particulier dans les scènes se déroulant sur les plateaux. Ces résultats justifient l'utilité de la modélisation du contexte afin de pallier les difficultés inhérentes à la représentation d'un visage où d'une voix.

Conclusions et perspectives

Conclusions

La première contribution de cette thèse est présentée au chapitre 4 et propose une nouvelle représentation de visage adaptée au regroupement en classes de visages. Cette représentation combine une approche basée sur la combinaison de descripteurs et une modélisation statistique. Chaque approche est optimale dans un cas particulier qui dépend de la variabilité présente dans les données et de la taille des classes. Il a été remarqué que le regroupement final tend à proposer plus de classes que de personnes présentes. Ceci est principalement dû à des variations de pose et d'échelle dans les scènes de débat, et à des cas de reportage où chaque scène tournée en un lieu spécifique présente des points de vue et des illuminations différentes.

Dans le chapitre 5, nous introduisons un modèle exploitant l'information multimodale afin d'effectuer le regroupement audiovisuel des personnes. Alors que des améliorations en terme de regroupement sont reportées au chapitre 2 de l'état de l'art, les améliorations observées dans cette thèse sont limitées : les regroupements monomodaux initiaux comportent peu d'erreurs que l'association des visages et des voix permette de corriger. D'une part, les systèmes monomodaux ont déjà de bonnes performances sur notre corpus. D'autre part, l'association des voix et des visages est relativement difficile. Des améliorations sont malgré tout observées dans le cas d'erreurs de confusion du regroupement en locuteur dues à des bruits de fond.

Concernant l'identification, l'état de l'art du chapitre 3 a mis en évidence que l'utilisation des noms extraits des cartouches est plus fiable que l'utilisation de la transcription automatique de la parole. De plus, les stratégies les plus performantes effectuent le regroupement et l'identification de manière jointe. Un dernier point, mais non des moindres, est que la compréhension du contexte de la scène est très utile pour le nommage des émissions télévisées (type de scène, rôle...). Cependant, il n'existe pas encore un système permettant d'obtenir cette information de contexte sans une quantité non négligeable d'annotations et d'*a priori* sur les données de test.

Le chapitre 6 commence par une étude sur l'identification à l'échelle des classes au moyen des cartouches en testant deux méthodes automatiques. La comparaison avec un oracle permet de constater que le potentiel d'identification des cartouches n'est pas complètement exploité par les méthodes automatiques. Les erreurs de confusion dans les regroupement monomodaux expliquent une partie de cet écart entre l'oracle et le système automatique. Afin d'améliorer l'identification, de l'information du contexte et de nommage est intégrée dans le modèle CRF

de regroupement proposé au chapitre 5. Le contexte utilisé ici modélise principalement l'arrière-plan des visages afin de distinguer à l'échelle des segments les personnes probablement identifiables par un cartouche. L'utilisation conjointe des deux CRF permet d'améliorer les performances par rapport à l'utilisation seule du CRF de nommage dans la majorité des émissions, en particulier pour les débats et les émissions en plateau.

Perspectives

Nous présentons tout d'abord quelques pistes qui pourraient être appliquées directement sur notre modèle pour l'améliorer.

- **Intégration de détecteurs de profils et de personnes :** la majorité des erreurs d'identification des visages est due à des profils ou des occlusions qui n'ont pas été détectés. Quantitativement en terme d'EGER, il semble que ce soit la voie pour obtenir de meilleures performances. La présence de visages non frontaux est une difficulté supplémentaire pour le regroupement des visages. Cependant, la modélisation CRF proposée dans cette thèse est robuste à ce type de variations grâce à l'utilisation de l'arrière-plan.
- **Ajout de modèles biométriques appris *a priori* :** les identifications ont été effectuées dans cette thèse de manière non-supervisée. Cependant il serait relativement facile d'ajouter des modèles biométriques appris *a priori* dans le CRF de nommage. Ces modèles seront particulièrement complémentaires des cartouches dans le cas des journalistes en voix off et des visages des présentateurs qui ne sont pas annoncés, et pour qui des données d'apprentissage peuvent être obtenues à partir d'émissions précédentes.
- **Segmentation en scène multimodale :** la caractérisation utilisée se fonde sur des attributs visuels de l'arrière-plan. Il serait possible d'intégrer des notions de segmentation en chapitre et/ou par sujet à partir d'attributs textuels et audios.
- **Représentations visuelle et acoustique :** ces deux domaines font l'objet de nombreux travaux, et il est probable qu'elles pourraient être continuellement mises à jour. A l'heure où ces lignes sont écrites, une représentation des visages à base de Fisher Vector [Simonyan 2013], et des tours de parole à base de Joint Factor Analysis [Kenny 2007] pourraient apporter des améliorations.
- **Critère d'optimisation du CRF de regroupement :** ce CRF est entraîné pour minimiser l'erreur de diarization DER. Pour une tâche d'identification, il semble plus raisonnable de chercher à minimiser directement l'erreur d'identification.

Le choix du contexte et la manière de l'utiliser constitue une problématique complexe et s'inscrit dans des perspectives intéressantes à plus long terme. L'approche développée par le consortium PERCOL dans la campagne REPERE [Bechet 2014] utilise une modélisation extensive du contexte avec la détection des rôles, des types de scènes et des position des caméras. Cependant, elle reste coûteuse à mettre en place en pratique car des annotations doivent être faites pour chaque type d'émission. A l'opposé, la modélisation de contexte utilisée dans cette thèse est beaucoup plus légère car elle ne nécessite pas de tels *a priori*. En contrepartie, les performances d'identification sont moins bonnes. Ainsi, une voie prometteuse réside dans le fait de poursuivre l'élaboration de modèles capables de construire automatiquement cette compréhension du contexte d'une émission à partir d'un corpus.

Acronymes

AV	audiovisuel(le)
BIC	Bayesian Information Criterion (Critère d'information bayésien)
CAH	Classification Ascendante Hiérarchique
DCT	Discrete Cosinus Transform (Transformée en cosinus discrète)
CLR	Cross Likelihood Ratio (Rapport de vraisemblance croisé)
CRF	Conditonnal Random Field (Champ conditionnel aléatoire)
EGER	Estimated Global Error Rate (Taux d'erreur estimé global)
GMM	Gaussian Mixture Model (Modèle de mélange de gaussiennes)
HMM	Hidden Markov Model (Chaîne de Markov Cachée)
HRAL	Hypothèse de Récurrence de l'Arrière-plan des Locuteurs
ILP	Integer Linear Programming (Programmation linéaire en nombres entiers)
MFCC	Mel Frequency cepstral coefficient (Coefficient cepstral à fréquence Mél)
OCR	Optical Character Recognition (Reconnaissance optique des caractères)
REPERE	REconnaissance de PERsonnes dans des Emissions audiovisuelles
SODA	reconnaiSsance de persOnnes pour Débats et journAux télévisés
SIFT	Scale Invariant Feature Transform (Transformée en caractéristiques invariants à l'échelle)
SURF	Speeded Up Robust Features (Caractéristiques robustes et accélérées).
SVM	Séparateur à Vaste Marge
UBM	Universal Background Model (Modèle du monde)

Annexes

Annexe A

Liste des publications

– Elie Khoury, Paul Gay, Sylvain Meignier, Jean-Marc Odobez« *Fusing matching and biometric similarity measures for face diarization in video* », Proceedings of the Internation Conference on Multimedia Retrieval, 2013.

– Mickael Rouvier, Grégor Dupuy, Paul Gay, Teva Merlin Sylvain Meignier. « *An open-source state-of-the-art toolbox for broadcast news diarization* », Proceedings of the Conference of the International Speech Communication Association, 2013.

– Paul Gay, Elie Khoury, Jean-Marc Odobez, Sylvain Meignier, Paul Deléglise « *A Conditional Random field approach for audio-visual people diarization* », Proceedings of the International Conference on Acoustic, Speech and Signal Processing, 2014.

– Paul Gay, Elie Khoury, Sylvain Meignier, Jean-Marc Odobez, Paul Deléglise « *Face identification from overlaid texts using Local Face Recurrent Patterns and CRF models* », Proceedings of the International Conference on Image Processing, 2014.

– Paul Gay, Grégor Dupuy, Carole Lailler, Sylvain Meignier, Jean-Marc Odobez, Paul Deléglise « *Comparison of Two Methods for Unsupervised Person Identification in TV Shows* », Proceedings of the Content Based Multimedia Indexing workshop, 2014.

Bibliographie

[Ahonen 2006] Ahonen T., Hadid A. et Pietikainen M., Face description with local binary patterns : Application to face recognition, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2006.

[Barras 2006] Barras C., Zhu X., Meignier S. et Gauvain J., Multistage speaker diarization of broadcast news, *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5) :1505–1512, 2006.

[Bauml 2013] Bauml M., Tapaswi M. et Stiefelhagen R., Semi-supervised learning with constraints for person identification in multimedia data, dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609, 2013.

[Bay 2006] Bay H., Tuytelaars T. et Van Gool L., Surf : Speeded up robust features, dans *Computer Vision–ECCV 2006*, Springer, 2006.

[Bechet 2014] Bechet F., Bendris M., Charlet D., Damnati G., Favre B., Rouvier M., Auguste R., Bigot B., Dufour R., Fredouille C., Linares G., Martinet J., Senay G. et Tirilly P., Multimodal understanding for person recognition in video broadcasts, dans *Proceedings of InterSpeech*, 2014.

[Bechet 2012] Bechet F., Favre B. et Damnati G., Detecting person presence in tv shows with linguistic and structural features, dans *Proc. of ICASSP*, pages 5077–5080, IEEE, 2012.

[Bendris 2009] Bendris M., Charlet D. et Chollet G., Introduction of quality measures in audio-visual identity verification, dans *Proc. of ICASSP*, pages 1913–1916, 2009.

[Bendris 2010] Bendris M., Charlet D. et Chollet G., Lip activity detection for talking faces classification in tv-content, dans *International Conference on Machine Vision*, 2010.

[Bendris 2011] Bendris M., Charlet D. et Chollet G., People indexing in tv-content using lip-activity and unsupervised audio-visual identity verification, dans *Proceedings of The Content Based Multimedia Indexing Workshop*, pages 139–144, IEEE, 2011.

[Bendris 2014] Bendris M., Favre B., Charlet D., Damnati G. et Auguste R., Multiple-view constrained clustering for unsupervised face identification in tv-broadcast, dans *Proc. of ICASSP*, 2014.

[Bendris 2013] Bendris M., Favre B., Charlet D., Damnati G., Auguste R., Martinet J., Senay G. et al., Unsupervised face identification in tv content using audio-visual sources, dans *Proceedings of The Content Based Multimedia Indexing Workshop*, 2013.

[Berg 2004] Berg T. L., Berg A. C., Edwards J., Maire M., White R., Teh Y. W., Learned-Miller E. G. et Forsyth D. A., Names and faces in the news, dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 848–854, 2004.

[Bhattarai 2014] Bhattarai B., Sharma G., Jurie F. et Pérez P., Some faces are more equal than others : Hierarchical organization for accurate and efficient large-scale identity-based face retrieval, dans *Proceedings of the European Conference on Computer Vision*, 2014.

[Bishop 2006] Bishop C. M. et al., *Pattern recognition and machine learning*, volume 1, springer New York, 2006.

[Bledsoe 1966] Bledsoe W. W., The model method in facial recognition, *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 1966.

[Bollacker 2008] Bollacker K., Evans C., Paritosh P., Sturge T. et Taylor J., Freebase : a collaboratively created graph database for structuring human knowledge, dans *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, ACM, 2008.

[Bredin 2013] Bredin H. et Poignant J., Integer linear programming for speaker diarization and cross-modal identification in tv broadcast, dans *Proceedings of InterSpeech*, 2013.

[Canseco-Rodriguez 2004] Canseco-Rodriguez L., Lamel L. et Gauvain J.-L., Speaker diarization from speech transcripts, dans *Proc. ICSLP*, volume 4, 2004.

[Chen 2005] Chen D. et Odobez J.-M., Video text recognition using sequential monte carlo and error voting methods, *Pattern Recognition Letters*, 2005.

[Chen 1998] Chen S. et Gopalakrishnan P., Speaker, environment and channel change detection and clustering via the bayesian information criterion, dans *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[Cinbis 2011] Cinbis R. G., Verbeek J. et Schmid C., Unsupervised metric learning for face identification in tv video, dans *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[Cour 2009] Cour T., Sapp B., Jordan C. et Taskar B., Learning from ambiguously labeled images, dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009.

[Cour 2011] Cour T., Sapp B. et Taskar B., Learning from partial labels, *The Journal of Machine Learning Research*, 12 :1501–1536, 2011.

[Csurka 2004] Csurka G., Dance C., Fan L., Willamowski J. et Bray C., Visual categorization with bags of keypoints, dans *Workshop on statistical learning in computer vision, ECCV*, 2004.

[Davis 1980] Davis S. et Mermelstein P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1980.

[Dehak 2011] Dehak N., Kenny P., Dehak R., Dumouchel P. et Ouellet P., Front-end factor analysis for speaker verification, *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4) :788–798, 2011.

[Dielmann 2010] Dielmann A., Unsupervised detection of multimodal clusters in edited recordings, dans *IEEE International Workshop on Multimedia Signal Processing*, 2010.

[Dietterich 1997] Dietterich T. G., Lathrop R. H. et Lozano-Pérez T., Solving the multiple instance problem with axis-parallel rectangles, *Artificial intelligence*, 1997.

[Du 2012] Du M. et Chellappa R., Face association across unconstrained video frames using conditional random fields, dans *Proceedings of the European Conference on Computer Vision*, pages 167–180, 2012.

[El Khoury 2010a] El Khoury E., Unsupervised video indexing based on audiovisual characterization of persons, *These de doctorat, Université Paul Sabatier, Toulouse, France*, 2010a.

[El Khoury 2007] El Khoury E., Jaffré G., Pinquier J. et Sénac C., Association of audio and video segmentations for automatic person indexing, dans *Content-Based Multimedia Indexing, 2007. CBMI'07. International Workshop on*, pages 287–294, IEEE, 2007.

[El Khoury 2012a] El Khoury E., Laurent A., Meignier S. et Petitrenaud S., Combining transcription-based and acoustic-based speaker identifications for broadcast news, dans *Proc. of ICASSP*, pages 4377–4380, IEEE, 2012a.

[El Khoury 2010b] El Khoury E., Senac C. et Joly P., Face-and-clothing based people clustering in video content, dans *Proceedings of the international conference on Multimedia information retrieval*, pages 295–304, ACM, 2010b.

[El Khoury 2012b] El Khoury E., Sénac C. et Joly P., Audiovisual diarization of people in video content, *Multimedia Tools and Applications*, pages 1–29, 2012b.

[Everingham 2006] Everingham M., Sivic J. et Zisserman A., Hello ! my name is... buffy–automatic naming of characters in tv video, 2006.

[Everingham 2009] Everingham M., Sivic J. et Zisserman A., Taking the bite out of automated naming of characters in tv video, *Image and Vision Computing*, 2009.

[Fan 2014] Fan H., Cao Z., Jiang Y. et Yin Q., Learning deep face representation, Rapport technique, Technical Report, Face++, 2014.

[Fredouille 2009] Fredouille C., Bozonnet S. et Evans N., The lia-eurecom r' 09 speaker diarization system, dans *RT'09, NIST Rich Transcription Workshop*, 2009.

[Friedland 2009] Friedland G., Hung H. et Yeo C., Multi-modal speaker diarization of real-world meetings using compressed-domain video features, dans *Proc. of ICASSP*, pages 4069–4072, 2009.

[Galibert 2013] Galibert O., Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech., *INTERSPEECH*, pages 1131–1134, 2013.

[Gallagher 2008] Gallagher A. C. et Chen T., Clothing cosegmentation for recognizing people, dans *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, IEEE, 2008.

[Gay 2014a] Gay P., Dupuy G., Lailler C., Odobez J.-M., Meignier S. et Deléglise P., Comparison of two methods for unsupervised person identification in tv shows, dans *International Workshop on Content Based Multimedia Indexing*, 2014a.

[Gay 2014b] Gay P., Khoury E., Meignier S., Odobez J.-M. et Deleglise P., A conditional random field approach for audio-visual people diarization, dans *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, numéro EPFL-CONF-198435, 2014b.

[Giraudel 2012] Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O. et Quintard L., The repere corpus : a multimodal corpus for person recognition., dans *Proceedings of The International Conference on Language Resources and Evaluation*, pages 1102–1107, 2012.

[Gish 1991] Gish H., Siu M.-H. et Rohlicek R., Segregation of speakers for speech recognition and speaker identification, dans *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 873–876, IEEE, 1991.

[Guillaumin 2012] Guillaumin M., Mensink T., Verbeek J. et Schmid C., Face recognition from caption-based supervision, *International Journal of Computer Vision*, 96(1) :64–82, 2012.

[Guillaumin 2009] Guillaumin M., Verbeek J. et Schmid C., Is that you ? metric learning approaches for face identification, dans *Proceedings of the IEEE International Conference on Computer Vision*, 2009.

[Huang 2007] Huang G. B., Ramesh M., Berg T. et Learned-Miller E., Labeled faces in the wild : A database for studying face recognition in unconstrained environments, Rapport technique, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[Jou 2013] Jou B., Li H., Ellis J. G., Morozoff-Abegauz D. et Chang S., Structured exploration of who, what, when, and where in heterogeneous multimedia news sources, dans *Proceedings of ACM International conference on Multimedia*, 2013.

[Jousse 2009] Jousse V., Petit-Renaud S., Meignier S., Esteve Y. et Jacquin C., Automatic named identification of speakers using diarization and asr systems, dans *Proc. of ICASSP*, pages 4557–4560, 2009.

[Kahn 2012] Kahn J., Galibert O., Quintard L., Carré M., Giraudel A. et Joly P., A presentation of the repere challenge, dans *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6, IEEE, 2012.

[Kenny 2007] Kenny P., Boulianne G., Ouellet P. et Dumouchel P., Joint factor analysis versus eigenchannels in speaker recognition, IEEE *Transactions on Acoustics, Speech and Language Processing*, pages 1435–1447, 2007.

[Khoury 2013] Khoury E., Gay P. et Odobez J., Fusing matching and biometric similarity measures for face diarization in video, dans *Proceedings of International Conference on Multimedia Retrieval*, 2013.

[Khoury 2010] Khoury E., Sénac C. et Joly P., Unsupervised segmentation methods of tv contents, *International Journal of Digital Multimedia Broadcasting*, 2010.

[Kostinger 2011] Kostinger M., Wohlhart P., Roth P. et Bischof H., Learning to recognize faces from videos and weakly related information cues, dans *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, IEEE, 2011.

[van der Kreeft 2014] van der Kreeft P., Macquarrie K., Kemman M., Kleppe M. et McGuinness K., Axes-research—a user-oriented tool for enhanced multimodal search and retrieval in audiovisual libraries, dans *Proceedings of The Content Based Multimedia Indexing Workshop*, pages 1–4, IEEE, 2014.

[Kuhn 1955] Kuhn H. W., The hungarian method for the assignment problem, *Naval research logistics quarterly*, pages 83–97, 1955.

[Kwak 2014] Kwak H. et An J., A first look at global news coverage of disasters by using the gdelt dataset, dans *Journal of Social Informatics*, 2014.

[Lafferty 2001] Lafferty J., McCallum A. et Pereira F. C., Conditional random fields : Probabilistic models for segmenting and labeling sequence data, 2001.

[Liu 2008] Liu C., Jiang S. et Huang Q., Naming faces in broadcast news video by image google, dans *Proceedings of the 16th ACM international conference on Multimedia*, pages 717–720, ACM, 2008.

[Liu 2002] Liu C. et Wechsler H., Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, *IEEE Transactions on Image processing*, 2002.

[Liu 2007] Liu Z. et Wang Y., Major cast detection in video using both speaker and face information, *Multimedia, IEEE Transactions on*, 9(1) :89–101, 2007.

[Lowe 2004] Lowe D. G., Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, 2004.

[Ma 2007] Ma C., Nguyen P. et Mahajan M., Finding speaker identities with a conditional maximum entropy model, dans *Proc. of ICASSP*, volume 4, pages 261–264, 2007.

[McGuinness 2013] McGuinness K., O'Connor N. E., Aly R., De Jong F., Chatfield K., Parkhi O. M., Arandjelovic R., Zisserman A., Douze M. et Schmid C., The axes pro video search system, dans *Proceedings of The Content Based Multimedia Indexing Workshop*, ACM, 2013.

[Milborrow 2008] Milborrow S. et Nicolls F., Locating facial features with an extended active shape model, dans *Computer Vision–ECCV 2008*, pages 504–513, Springer, 2008.

[Moon 1996] Moon T. K., The expectation-maximization algorithm, *IEEE Signal processing magazine*, pages 47–60, 1996.

[Murphy 1999] Murphy K. P., Weiss Y. et Jordan M. I., Loopy belief propagation for approximate inference : An empirical study, dans *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475, Morgan Kaufmann Publishers Inc., 1999.

[NIST 2003] NIST, The rich transcription spring 2003 (rt-03s) evaluation plan, 2003, `http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf`.

[Noulas 2010] Noulas A., *Audiovisual fusion for speaker diarization*, Thèse de doctorat, Universiteit van Amsterdam, 2010.

[Noulas 2012] Noulas A., Englebienne G. et Krose B. J., Multimodal speaker diarization, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1) :79–93, 2012.

[Odobez 2003] Odobez J.-M., Gatica-Perez D. et Guillemot M., Video shot clustering using spectral methods, *International Working on Content-Based Multimedia Indexing*, 2003.

[Ozkan 2006] Ozkan D. et Duygulu P., Finding people frequently appearing in news, dans *Image and Video Retrieval*, pages 173–182, Springer, 2006.

[Parkhi 2012] Parkhi O., Vedaldi A. et Zisserman A., On-the-fly specific person retrieval, dans *International Workshop on Image Analysis for Multimedia Interactive Services*, pages 1–4, 2012.

[Parkhi 2013] Parkhi O., Vedaldi A. et Zisserman A., State-of-the-art and future challenges in video scene detection : a survey, dans *The journal of Multimedia Systems*, pages 1–4, 2013.

[Pelecanos 2001] Pelecanos J. et Sridharan S., Feature warping for robust speaker verification, 2001.

[Poignant 2013a] Poignant J., *Identification non-supervisée de personnes dans les flux télévisés.*, Thèse de doctorat, École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII), 2013a.

[Poignant 2013b] Poignant J., Besacier L., Le V. B., Rosset S. et Quénot G., Unsupervised naming of speakers in broadcast TV : using written names, pronounced names or both ?, dans *the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013b.

[Poignant 2012] Poignant J., Bredin H., Le V.-B., Besacier L., Barras C., Quénot G. et al., Unsupervised speaker identification using overlaid texts in tv broadcast, dans *Proceedings of InterSpeech*, 2012.

[Pruzansky 1963] Pruzansky S., Pattern-matching procedure for automatic talker recognition, *The Journal of the Acoustical Society of America*, pages 354–358, 1963.

[Reynolds 2000] Reynolds D. A., Quatieri T. F. et Dunn R. B., Speaker verification using adapted gaussian mixture models, *Digital signal processing*, 2000.

[Rouvier 2013] Rouvier M., Dupuy G., Gay P., Khoury E., Merlin T. et Meignier S., An open-source state-of-the-art toolbox for broadcast news diarization, 2013.

[Rouvier 2014] Rouvier M., Favre B., Bendris M., Charlet D. et Damnati G., Scene understanding for identifying persons in tv shows : beyond face authentication, dans *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, 2014.

[Rouvier 2012] Rouvier M. et Meignier S., A global optimization framework for speaker diarization, dans *Odyssey Workshop, Singapore*, 2012.

[Satoh 1999] Satoh S., Nakamura Y. et Kanade T., Name-it : Naming and detecting faces in news videos, *IEEE Multimedia*, 6(1) :22–35, 1999.

[Simonyan 2013] Simonyan K., Parkhi O. M., Vedaldi A. et Zisserman A., Fisher vector faces in the wild, dans *Proceedings of the British Machine Vision Conference*, 2013.

[Song 2004] Song X., Lin C.-Y. et Sun M.-T., Cross-modality automatic face model training from large video databases, dans *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 91–91, IEEE, 2004.

[Sun 2014] Sun Y., Wang X. et Tang X., Deep learning face representation by joint identification-verification, *CoRR*, 2014.

[Sutton 2006] Sutton C. et McCallum A., An introduction to conditional random fields for relational learning, *Introduction to statistical relational learning*, 2006.

[Sutton 2010] Sutton C. et McCallum A., An introduction to conditional random fields, 2010.

[Tan 2007] Tan X. et Triggs B., Enhanced local texture feature sets for face recognition under difficult lighting conditions, dans *Analysis and Modeling of Faces and Gestures*, 2007.

[Tranter 2006] Tranter S., Who really spoke when ? finding speaker turns and identities in broadcast news audio, dans *Proc. of ICASSP*, volume 1, pages I–I, 2006.

[Vallet 2013] Vallet F., Essid S. et Carrive J., A multimodal approach to speaker diarization on tv talk-shows, *IEEE Transactions on Multimedia*, 15(3) :509–520, 2013.

[Viola 2004] Viola P. et Jones M. J., Robust real-time face detection, *International Journal of Computer Vision*, 2004.

[Wallace 2012] Wallace R., McLaren M., McCool C. et Marcel S., Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models, dans IEEE *Transactions on Information Forensics and Security*, 2012.

[Wohlhart 2011] Wohlhart P., Köstinger M., Roth P. M. et Bischof H., Multiple instance boosting for face recognition in videos, dans *Pattern Recognition*, pages 132–141, Springer, 2011.

[Wooters 2008] Wooters C. et Huijbregts M., The icsi rt07s speaker diarization system, dans *Multimodal Technologies for Perception of Humans*, pages 509–519, Springer, 2008.

[Wu 2013] Wu B., Zhang Y., Hu B.-G. et Ji Q., Constrained clustering and its application to face clustering in videos, dans *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3507–3514, IEEE, 2013.

[Yang 2004] Yang J. et Hauptmann A. G., Naming every individual in news video monologues, dans *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 580–587, ACM, 2004.

[Yang 2005] Yang J., Yan R. et Hauptmann A. G., Multiple instance learning for labeling faces in broadcasting news video, dans *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 31–40, ACM, 2005.

[Yang 2002] Yang M.-H., Kernel eigenfaces vs. kernel fisherfaces : Face recognition using kernel methods, *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 0215–0215, 2002.

[Yaniv 2014] Yaniv T., Ming Y., MarcÁurelio R. et Lior W., Deepface : Closing the gap to human-level performance in face verification, dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[Zhang 2013] Zhang L., Kalashnikov D. V. et Mehrotra S., A unified framework for context assisted face clustering, dans *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 9–16, ACM, 2013.

[Zhao 2008] Zhao S., Precioso F., Cord M., Philipp-Foliguet S. et al., Actor retrieval system based on kernels on bags of bags, *EUSIPCO, Lausanne, Switzerland*, 2008.