THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *26/09/2013* **par :**

PHILIPPE ERCOLESSI

Extraction multimodale de la structure narrative des épisodes de séries télévisées

---

JURY

PATRICK LAMBERT	Professeur d'Université	Rapporteur
BERNARD MÉRIALDO	Professeur d'Université	Rapporteur
PHILIPPE JOLY	Professeur d'Université	Directeur de thèse
HERVÉ BREDIN	Chargé de Recherche	Encadrant
CHRISTINE SÉNAC	Maître de conférence	Encadrant
SID-AHMED BERRANI	Chercheur	Examinateur
VINCENT CHARVILLAT	Professeur d'Université	Invité

École doctorale et spécialité :
MITT : Image, Information, Hypermedia

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (IRIT)

Directeur de Thèse :
Philippe Joly

Rapporteurs :
Patrick Lambert et *Bernard Mérialdo*

Remerciements

Je tiens à commencer ce manuscrit de thèse en remerciant toutes les personnes qui m'ont soutenu durant la réalisation de mon travail de recherche et qui ont rendu ce document possible.

Je tiens tout particulièrement à remercier l'ensemble des membres du jury : mon directeur de thèse, M. Philippe Joly, ainsi que mes deux encadrants que sont M. Hervé Bredin et Mme Christine Sénac, qui m'ont encouragé et soutenu dans les moments de doute et qui ont toujours été présents et d'excellent conseil lorsque j'en avais besoin. MM. Bernard Mérialdo et Patrick Lambert, qui ont accepté d'être les rapporteurs de cette thèse, et qui ont contribué à améliorer la qualité de ce manuscrit par leurs remarques et suggestions. Et enfin MM. Sid Ahmed Berrani et Vincent Charvillat qui ont accepté de participer au jury de soutenance.

Je me dois de remercier l'ensemble de l'équipe SAMOVA, que ce soit tous ceux qui sont partis, ceux qui sont arrivés en cours de route ou ceux qui y sont toujours, pour leur accueil formidable et pour l'aide technique et morale qu'ils m'ont apporté durant toutes ces années. Je remercie l'ensemble du service informatique de l'IRIT, les secrétaires ainsi que la direction pour leur soutien, leur présence et leur efficacité malgré un environnement de travail parfois difficile.

Je ne sais comment remercier tous les gens que j'ai rencontrés au laboratoire, MM. Anthony Pajot, Olivier Gourmel, Samir Torky, Mathieu Muratet, Mathieu Giorgino, Dorian Gomez, Mlles Andra Doran, Monia Ben Mlouka, et tous ceux qui m'ont appris la vie au sein de cet institut, qui ont fait que chaque jour était un délice et que je peux maintenant compter pour beaucoup comme des amis.

Cette thèse n'aurait pas été possible sans le soutien de mes amis et de ma famille. Je dois ainsi remercier M. Benjamin Amanrich et Mlle Marie Amiot chez qui j'ai pu me réfugier chaque fois que je ressentais la nécessité de m'enfuir de la rédaction de ce document. Mes remerciements vont aussi à MM. Stéphane Bezold, Georges Moulinier, Vincent Deliencourt, Julien Legrand et Loris Cathalo qui m'ont fait sortir de ma bulle plus d'une fois, ainsi qu'à Mlles Violaine Cheynier et Swanny Henry qui ont été extraordinaires pour remonter un moral qui a parfois été un peu bas.

Je dois énormément à ma mère, Annie Ercolessi, et à sa foi indéfectible en mes capacités ainsi qu'à mon père, Gilbert Ercolessi. Je pense aussi à Mlle Frédérique Benoit, qui a su me supporter dans les moments difficiles et avec qui j'ai pu partager tous les moments de joie.

Et enfin, je souhaite remercier mon frère et ma soeur, Guillaume et Marie-Émilie Ercolessi, ma marraine de cœur, Bénédicte Cottaz-Cordier, et tous ceux qui ont pris du temps pour m'aider à relire, corriger et améliorer cette thèse. J'ajoute à ces remerciements tous ceux qui m'ont aidé à annoter et analyser les données utilisées, tous ceux qui m'ont donné des idées, ou qui m'ont laissés utiliser leurs programmes, et qui sont malheureusement beaucoup trop nombreux pour pouvoir être cités sur cette page.

Table des matières

Introduction **3**

Chapitre 1
Définitions **11**

1.1 Segmentation en plans . . . 11
1.2 Segmentation en scènes . . . 13
1.2.1 Qu'est-ce qu'une scène ? . . . 13
1.2.2 Tâche de segmentation en scènes . . . 17
1.3 Le regroupement des scènes en histoires . . . 20
1.3.1 Qu'est-ce qu'une histoire ? . . . 20
1.3.2 Tâche de regroupement des scènes en histoires . . . 24
1.4 Discussion . . . 27

Chapitre 2
Etat de l'art **29**

2.1 Description et comparaison de segments de documents audiovisuels . . . 30
2.1.1 Descripteurs bas niveau . . . 31
2.1.1.1 Descripteurs visuels globaux . . . 31
2.1.1.2 Descripteurs visuels locaux . . . 34
2.1.1.3 Descripteurs audio . . . 37
2.1.2 Descripteurs haut niveau . . . 38
2.1.2.1 Concepts sémantiques . . . 38
2.1.2.2 Personnages . . . 40
2.1.3 Mesures de similarité/distance entre deux vecteurs de descripteurs . . . 43
2.2 Méthodes de regroupement de données . . . 44
2.2.1 Regroupement hiérarchique agglomératif . . . 45
2.2.2 Regroupement basé sur des graphes . . . 46

2.2.3 K-Moyennes . . . 47
2.2.4 Regroupement spectral . . . 49
2.2.5 Détermination automatique du nombre de groupes . . . 50
2.3 Travaux existant sur la segmentation en scènes . . . 50
2.3.1 Détection de frontières . . . 52
2.3.2 Regroupement de séquences . . . 55
2.3.3 Approches hybrides . . . 59
2.3.4 Évaluation de la segmentation automatique des scènes . . . 63
2.4 Regroupement de séquences traitant d'un même sujet . . . 67
2.4.1 Identification des reportages dans des journaux télévisés . . . 68
2.4.2 Regroupement de documents suivant leur ressemblance sémantique 69
2.4.3 Évaluation du regroupement de séquences traitant d'un même sujet 71

Chapitre 3
Comparaison de segments temporels de séries télévisées 75

3.1 Où ? Similarité entre lieux . . . 76
3.1.1 Similarité basée sur la couleur . . . 77
3.1.2 Limites des descripteurs de couleur . . . 79
3.2 Quoi ? Similarité entre les dialogues . . . 80
3.2.1 Similarité basée sur la transcription . . . 81
3.2.2 Limites des systèmes ASR . . . 83
3.3 Qui ? Similarité entre les personnages . . . 84
3.3.1 Similarité basée sur la présence des locuteurs . . . 86
3.3.2 Limites des systèmes de segmentation et regroupement en locuteurs 90

Chapitre 4
Segmentation en scènes 93

4.1 Protocole expérimental . . . 95
4.1.1 Corpus et annotations manuelles . . . 95
4.1.2 Métriques d'évaluation . . . 96
4.1.3 Validation croisée . . . 98
4.2 Approche monomodale basée sur les tours de parole . . . 98
4.2.1 Approche par fenêtre glissante . . . 100
4.2.2 Intérêt de l'approche . . . 102
4.2.3 Résultats expérimentaux . . . 106

4.2.4 Conclusion . . . 109
4.3 Approche multimodale par alignement de frontières . . . 110
4.3.1 Segmentation STG à partir d'histogrammes de couleur . . . 110
4.3.2 Description de l'approche de fusion . . . 111
4.3.3 Résultats expérimentaux . . . 114
4.3.3.1 À partir de tours de parole manuels . . . 114
4.3.3.2 À partir de tours de parole automatiques . . . 116
4.3.4 Conclusion . . . 118
4.4 Fusion dans le cadre du GSTG . . . 119
4.4.1 Description de l'approche . . . 119
4.4.2 Résultats expérimentaux . . . 120
4.5 Conclusion . . . 124

Chapitre 5
Regroupement des scènes en histoires **125**

5.1 Métriques d'évaluation . . . 127
5.1.1 F-Mesure . . . 127
5.1.2 (Adjusted) Rand Index . . . 128
5.1.3 Diarization Error Rate . . . 128
5.1.4 Comparaison des métriques . . . 130
5.1.5 Conclusion . . . 132
5.2 Protocole expérimental . . . 132
5.2.1 Corpus . . . 132
5.2.2 Scènes ne faisant partie d'aucune histoire . . . 133
5.2.3 Regroupement aléatoire . . . 134
5.2.4 Oracle . . . 134
5.3 Approches monomodales . . . 135
5.3.1 Méthodes utilisées . . . 135
5.3.2 Résultats et discussion . . . 137
5.4 Regroupement multimodal . . . 143
5.4.1 Méthode développée . . . 143
5.4.2 Résultats . . . 144
5.4.3 Discussion . . . 144
5.5 Sélection automatique de la meilleure méthode de regroupement . . . 146
5.5.1 Approche de sélection automatique . . . 147

5.5.2 Résultats 150
5.5.3 Discussion 151
5.6 Conclusion 153

Chapitre 6
Applications **155**

6.1 Outils de navigation de vidéo 156
6.2 Aperçu de l'interface 158
6.3 Visualisation de la structure d'un épisode 159
6.4 Évaluation des systèmes de structuration 161
6.4.1 Évaluation de la segmentation en scènes 161
6.4.2 Évaluation des histoires 162
6.5 Technologies utilisées 165
6.6 Conclusion 165

Conclusion et perspectives **167**

Annexes **173**

Annexe A Publications **173**

Bibliographie **175**

Notations

Pour l'ensemble de ce document, nous utilisons les notations suivantes :

Relatives à la structure des épisodes

p	:	**Plan**
s	:	**Séquence / Scène**
sv	:	**Scène vérité terrain**
sa	:	**Scène détectée automatiquement**
$\mathcal{S}$	:	**Ensemble des séquences (plans ou scènes) d'un document**
H	:	**Histoire**
$\mathcal{H}$	:	**Ensemble des histoires détectées automatiquement**
$\mathcal{R}$	:	**Ensemble des histoires manuellement annotées**
e	:	**Épisode**
$\mathcal{E}$	:	**Ensemble des épisodes du corpus**
t	:	**temps**
f	:	**Frontière entre deux plans/scènes**
$\mathcal{F}$	:	**Ensemble des frontières d'un épisode**

Relatives à l'évaluation

$\mathcal{L}$	:	**Métrique d'évaluation**
$\mathfrak{L}$	:	**Performance d'un algorithme**
τ	:	**Tolérance sur la position des frontières pour l'évaluation**

Relatives aux ensembles / groupes de données

$\mathcal{Q}$	:	**Modularité des communautés d'un graphe**
$\mathcal{A}$	:	**Matrice d'adjacence / Matrice d'affinité**

Relatives aux descripteurs

w	:	**Mot**
ℓ	:	**Locuteur**
To	:	**Tour de parole des locuteurs**

Relatives aux paramètres des algorithmes développés

λ	:	**Paramètres des méthodes de segmentation/regroupement**
Λ	:	**Espace de recherche des paramètres**
α	:	**Pondération des locuteurs**

Autres

$\mathbb{M}$	:	**Annotation manuelle**
$\mathbb{C}$	:	**Classification automatique**

Acronymes / Abréviations

Pour l'ensemble de ce document, nous utilisons les acronymes/abréviations suivantes :

Relatives à l'évaluation

VP	:	**Vrai Positif**
FP	:	**Faux Positif**
VN	:	**Vrai Négatif**
FN	:	**Faux Négatif**
P	:	**Précision**
R	:	**Rappel**
F_{PR}	:	**Moyenne harmonique de la précision et du rappel (F-Mesure)**
C	:	**Couverture (*Coverage*)**
O	:	**Débordement (*Overflow*)**
DED	:	**Differential Edit Distance**
RI	:	**Rand Index**
ARI	:	**Rand Index Ajusté**
DER	:	**Diarization Error Rate (Erreur de segmentation et regroupement en locuteurs)**

Relatives aux descripteurs

HSV	:	**Modalité basée sur la couleur (histogramme de couleur HSV)**
ASR	:	**Modalité basée sur les dialogues (Transcription automatique de la parole)**
SD	:	**Modalité basée sur la présence des locuteurs (Segmentation et regroupement en locuteurs)**

Introduction

Depuis l'invention de la photographie et du cinéma jusqu'à aujourd'hui, notre mode de consommation des documents audiovisuels a beaucoup évolué. Au début du $XX^{ème}$ siècle, la visualisation d'images animées était uniquement possible dans des cinémas ou des fêtes foraines. De nos jours, la vidéo est au centre de la vie quotidienne de la majorité des personnes vivant dans les pays industrialisés.

L'accroissement de l'importance de la vidéo a été rendu possible grâce à de très nombreux modes de diffusion qui ont émergé ces 50 dernières années. La vidéo est ainsi présente au cinéma bien sûr, mais aussi à la télévision, sur un ordinateur, sur des panneaux publicitaires, sur un téléphone portable ou un smartphone, et peut-être bientôt sur des lunettes (Google Glass). La vidéo est utilisée pour les loisirs, pour se former ou s'informer et pour diffuser des publicités à longueur de journées.

La quantité de documents audiovisuels diffusés chaque jour a énormément augmenté ces 20 dernières années. Quelques exemples en chiffres : aujourd'hui en France, la télévision numérique terrestre (TNT) propose 32 chaînes de télévision (hors programmes locaux) diffusant du contenu 24h/24, soit 768 heures de diffusion par jour. Certains bouquets de chaînes de télévision par satellite diffusent plus de 1000 chaînes, soit chaque jour plus de 24000 heures de contenu. Enfin, le site Youtube déclare que plus de 24h de vidéos sont mises en ligne chaque minute, soit plus de 34560 heures de contenus supplémentaires quotidiens.

Notre rapport avec la télévision a lui aussi beaucoup évolué. Nous ne sommes plus seulement des téléspectateurs regardant des programmes télévisuels au moment de leur diffusion. Dorénavant, nous voulons pouvoir choisir le programme qui nous intéresse ou revoir un programme que nous avons raté. Jusqu'à la fin du $XX^{ème}$ siècle, il était possible d'enregistrer les émissions télévisées sur cassette vidéo ou sur DVD. Aujourd'hui de nombreux services voient le jour pour permettre la visualisation des programmes sur demande. Ainsi, beaucoup de chaînes de télévision comme TF1, France 2 ou M6 proposent des services de vidéo à la demande ou des offres de « *replay* » (rediffusion).

La façon de visualiser les programmes elle-même évolue. Ainsi, des fonctionnalités permettant de contrôler les programmes diffusés à la télévision font leur apparition. Par exemple, le contrôle parental permet aux parents de filtrer les émissions qu'ils souhaitent rendre inaccessibles à leurs enfants. Dans ce cas, nous pouvons imaginer des outils permet-

tant d'identifier rapidement et efficacement les séquences violentes ou à caractère sexuel d'un film ou d'une émission d'information.

Pour que ces services soient possibles, il est nécessaire qu'un ensemble de « métadonnées » accompagne les programmes. Ainsi, il faut extraire les programmes du flux télévisuel ou des services de partage de vidéos sur internet. Il faut les identifier, les classer et les préparer pour les stocker en vue de les rediffuser. Ensuite, il est nécessaire d'annoter les programmes pour connaître leur structure, les personnes qui y apparaissent ou les thèmes abordés. Il existe deux façons de récupérer ce genre de métadonnées : manuellement ou automatiquement.

Les annotations manuelles peuvent provenir de deux sources. La première vient directement des producteurs du programme ou des chaînes de télévision. Elles sont déterminées avant la création du contenu et elles peuvent parfois être obtenues à partir du support sur lequel sont diffusés les programmes.

Par exemple, depuis la mise en place de la *Télévision Numérique Terrestre* (TNT), les chaînes de télévision diffusent des métadonnées directement disponibles sur le poste de télévision, comme la durée des programmes, l'heure à laquelle ils sont diffusés, leur nom ou un court résumé textuel. Cependant, les producteurs ou les organismes diffusant du contenu audiovisuel sont souvent réticents à diffuser gratuitement ce genre de métadonnées. En effet, elles permettraient à d'autres compagnies de fabriquer des services les utilisant, comme par exemple filtrer la publicité qui est une source de financement essentielle des chaînes de télévision.

C'est pourquoi ces métadonnées doivent être obtenues par annotation manuelle *a posteriori*. Cependant, une annotation manuelle est très coûteuse, en temps comme en argent. La question récurrente du « coût » rend cette façon de procéder impossible à cause de la quantité de vidéos qu'il est nécessaire d'annoter chaque jour.

Il est donc nécessaire de trouver des méthodes pour suppléer les annotations manuelles par des algorithmes automatiques. La structuration automatique peut être définie comme la capacité à extraire une organisation interne des documents et contenus analysés. Il s'agit généralement de dégager des sections de vidéos véhiculant une information particulière de façon automatique à l'aide de solutions logicielles.

Dans cette thèse, nous limitons notre étude sur la structuration de documents audiovisuels à des programmes télévisuels, et plus particulièrement à des épisodes de séries télévisées. Ce type de documents a pour particularité d'avoir subi une étape de « post-production ». Les différentes séquences qui le composent ont été assemblées en une suite cohérente pour assurer l'enchaînement logique du document audiovisuel. L'extraction de la structure d'un document « post-produit » peut donc être vue comme une tâche de « déconstruction » du document. Ainsi, retrouver les plans, les scènes et la façon dont ils ont été arrangés dans le document permet d'obtenir des informations sur la manière dont le récit est raconté. C'est ce que nous appelons la « **structure narrative** » des épisodes de séries télévisées.

Aperçu des travaux existant

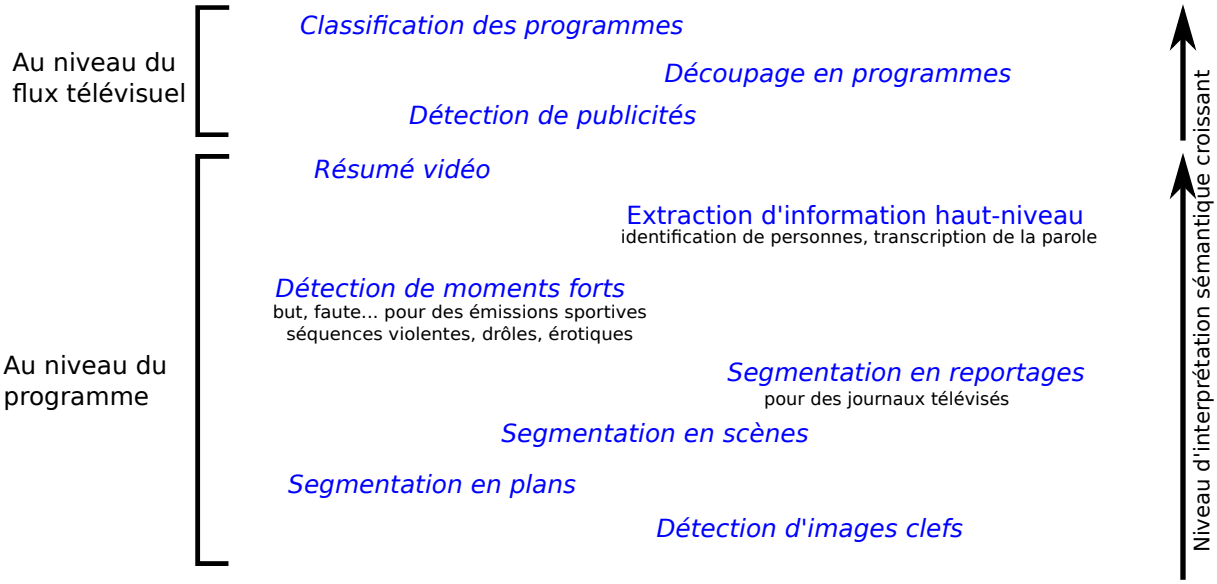

FIGURE *1 – Différents niveaux de structuration automatique de documents audiovisuel.*

De nombreuses recherches portant sur la structuration automatique de vidéos ont vu le jour ces dernières années. La Figure 1 présente de manière non exhaustive un ensemble de domaines de recherche de cette problématique. Il n'existe pas à notre connaissance de définition consensuelle de cette notion. Pour [Vallet 2011], la structuration se rapporte à toute tentative d'ordonnancement de contenu. Ainsi, les tâches de détection de changement de plans, de segmentation en scènes ou de résumé automatique peuvent être considérées comme des processus de structuration.

La problématique de la structuration est souvent présentée sous l'angle du « fossé sémantique ». Cette expression représente l'écart qui sépare la représentation numérique brute du signal audiovisuel de la représentation conceptuelle des éléments de structure du programme. Il est alors courant d'observer une hiérarchisation des techniques de structuration suivant leur niveau d'interprétation sémantique, telle qu'elle est proposée dans la Figure 1.

Les travaux sur la structuration de documents audiovisuels peuvent se distinguer selon le fait que l'on traite le flux télévisuel ou des documents isolés. La structuration du flux télévisuel consiste à le segmenter en programmes et identifier la nature de ces segments : publicité, film, journal télévisé, émission sportive, etc... C'est un double problème de segmentation du flux en segments cohérents et de classification des segments en catégories [Gros 2012].

Un document isolé correspond à un programme diffusé à la télévision, comme par exemple un film, un talk-show, ou un journal télévisé, dont il s'agit de retrouver la

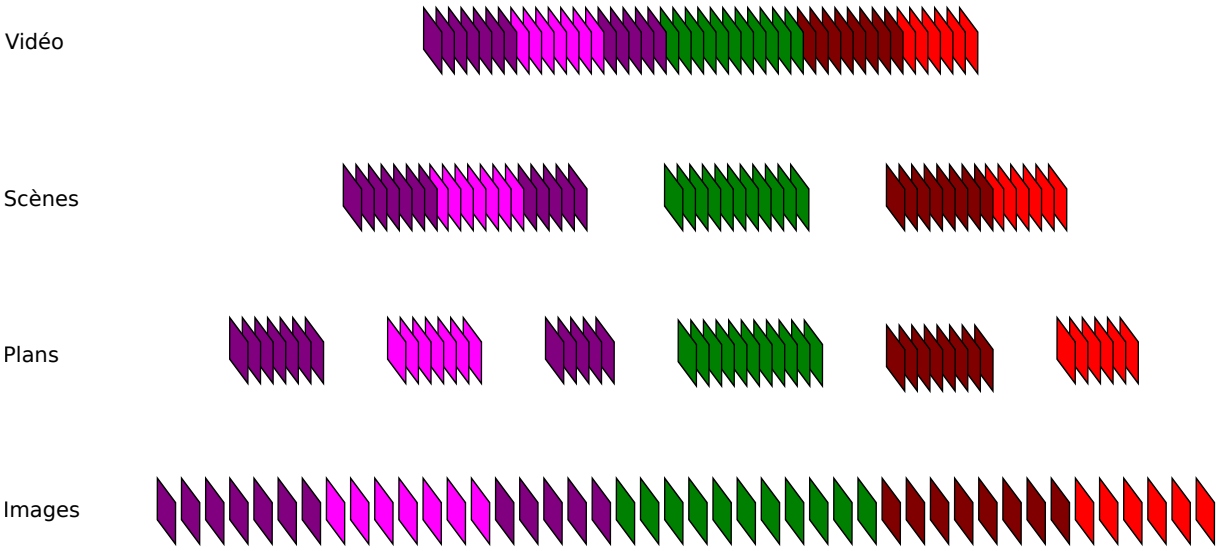

FIGURE *2 – Différents niveaux de structuration d'une vidéo*

structure et d'en identifier les passages importants. Abduraman *et al.* [Abduraman 2011] proposent un état de l'art des techniques de structuration de programmes télévisés. Ils définissent deux grandes familles d'approches : les approches spécifiques à un type de programme et les approches génériques.

Les approches spécifiques partent du principe *qu'une solution universelle pour l'analyse des vidéos est difficile, voire impossible.* Ainsi la connaissance du programme étudié doit être prise en compte par les méthodes de structuration. Abduraman *et al.* discutent de deux types de programmes. Ils traitent d'une part des émissions sportives pour lesquelles la problématique de la structuration consiste à détecter les phases de jeu ou des évènements clefs comme les buts ou les fautes lors de matchs de football [Xie 2002]. D'autre part, ils discutent de la segmentation des journaux télévisés découpés en reportages. On peut ajouter à ceux-là des outils permettant de détecter les plans violents dans des films [Penet 2012], les séquences érotiques ou pornographiques [Ulges 2012], ou encore les « chutes comiques » dans les sitcoms [Friedland 2009].

Les approches génériques considèrent qu'une structure commune à tous les documents audiovisuels « post-produits » existe. Elle peut être représentée à différents niveaux de granularité comme illustré dans la Figure 2. Le plan est souvent considéré comme l'unité la plus petite constituant la structure d'un document audiovisuel. Il est constitué d'images consécutives capturées sans interruption par une caméra. À un niveau plus élevé, on retrouve la problématique de la segmentation en scènes, qui est généralement définie comme un groupe de plans consécutifs partageant le même sens et décrivant un évènement unique.

Contributions et plan de la thèse

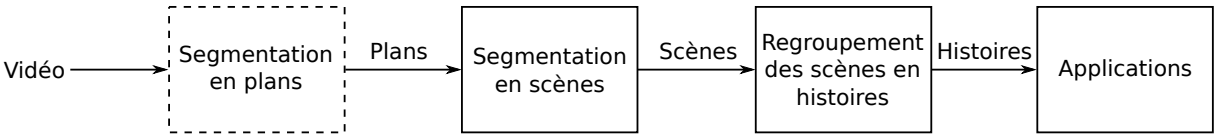

FIGURE 3 – *Contributions pour l'extraction de la structure d'épisodes de séries télévisées.*

L'objet d'étude de cette thèse concerne les épisodes de séries télévisées. L'extraction de la structure de ce type de documents comprend un grand nombre de défis. En effet, il existe un très grand nombre de formats d'une série à l'autre, autant par ce qui y est raconté (série sentimentale, série policière, ...) que par les choix de montage et d'organisation générale des épisodes (plans et scènes plus ou moins longs, durée des épisodes de 20, 40 ou 50 minutes). Nos contributions portent sur l'étude d'une **structure narrative hiérarchique d'un épisode de série télévisée** proche de la structure présentée dans la Figure 2. Elles sont résumées par la Figure 3, où les trois rectangles pleins représentent les différents thèmes abordés dans cette thèse. Le corps du document est organisé en 6 chapitres.

Les unités de structure étudiées dans cette thèse ont parfois des définitions très diverses. C'est pourquoi le Chapitre 1 propose une **définition objective des trois concepts principaux de structuration utilisés ou étudiés** : le plan, la scène et l'histoire.

Le Chapitre 2 est un **état de l'art des approches de structuration de documents audiovisuels**. Les deux premières sections proposent un aperçu des méthodes de description de segments audiovisuels et des approches classiques de regroupement de données. La troisième section propose un état de l'art des approches de segmentation en scènes. Enfin, la dernière section discute des méthodes de regroupement de séquences traitant d'un même sujet.

Dans le Chapitre 3, nous nous intéressons aux manières de **décrire des séquences de documents audiovisuels dans le but de retrouver les scènes et les histoires**. Nous étudions trois types de descripteurs qui permettent de répondre à trois questions essentielles à l'extraction de la structure narrative : où se situe l'action (où ?), qui est impliqué dans l'action (qui ?), que se passe-t-il (quoi ?).

Notre première contribution est détaillée dans le Chapitre 4. Elle consiste en une méthode de **segmentation en scènes à partir des personnages et d'une analyse de la couleur des images**. Les approches de segmentation en scènes basées sur l'analyse de la couleur des images sont très répandues dans la littérature scientifique. Cependant, il est plus surprenant que les personnages ne soient que très rarement utilisés pour retrouver les scènes. Par exemple, au théâtre, on considère qu'un changement de scène se produit lorsqu'un ou plusieurs personnages entrent ou sortent de l'espace où se joue la pièce (espace qui est lui-même appelé une scène) [Bonnabel 2000]. Nous proposons d'extraire l'infor-

mation sur la présence des personnages dans un épisode de série télévisée en détectant les tours de parole des locuteurs à partir d'un système de segmentation et regroupement en locuteurs. Nous étudions une approche de segmentation en scènes basée sur ces tours de parole seuls ou conjointement à une segmentation basée sur la couleur (publié dans WIAMIS 2011, cf. Annexe A).

Avant les années 1990, dans les séries télévisées comme *MacGyver, Columbo* ou *Magnum*, chaque épisode racontait une histoire centrée sur les deux ou trois personnages principaux et se déroulant de façon continue tout au long de l'épisode. Mais depuis le début des années 1990, il est devenu courant de voir plusieurs histoires (ou lignes d'actions) racontées en parallèle dans un même épisode. Par exemple, dans la série *Malcolm*, les épisodes sont généralement composés d'une histoire suivant le point de vue des parents, une autre celui des enfants, et au moins une troisième racontant l'histoire du grand frère expatrié. Notre deuxième contribution consiste à étudier des **méthodes de regroupement de scènes non nécessairement contiguës pour retrouver et isoler ces histoires**. Elle est présentée dans le Chapitre 5 et elle comporte trois parties principales.

La première partie concerne l'**évaluation du regroupement des scènes en histoires**. Bien qu'il soit possible d'utiliser des méthodes d'évaluation de regroupement génériques, elles ne sont pas forcément adaptées à la tâche que nous souhaitons évaluer. Nous proposons donc une comparaison de plusieurs métriques d'évaluation pour le regroupement des scènes en histoires.

Ensuite, nous proposons **une analyse de différentes approches de regroupement et de différentes méthodes de description des scènes basées sur trois modalités** : la couleur (histogrammes de couleur), les personnages (tours de parole des locuteurs) et le texte (mots reconnus par un système de transcription automatique de la parole) (publié dans CBMI 2012, cf. Annexe A). Nous proposons, en outre, **une approche de fusion permettant de tirer parti au mieux des informations fournies par ces modalités**.

Enfin, pour répondre à la variabilité de la structure narrative des épisodes de séries télévisées, nous proposons dans la dernière partie du Chapitre 5 une méthode qui s'adapte à chaque épisode. Elle permet de **choisir automatiquement la méthode de regroupement la plus pertinente parmi les différentes méthodes proposées** (publié dans AMVA 2012, cf. Annexe A).

Pour étudier les applications possibles de nos contributions, nous avons développé StoViz, un **outil de visualisation de la structure d'un épisode de série télévisée** (scènes et histoires) qui est présenté dans le Chapitre 6. Il permet de faciliter la navigation au sein d'un épisode, en montrant les différentes histoires racontées en parallèle. Il permet également la lecture des épisodes histoire par histoire, et la visualisation d'un résumé de l'épisode en donnant un aperçu de chaque histoire qui y est racontée (publié dans ACMMM 2012 et Document Numérique 2012, cf. Annexe A). StoViz est aussi un parfait compagnon du chercheur puisqu'il offre des outils d'aide à l'évaluation des approches de segmentation et de regroupement des scènes en histoires.

Pour finir, nous terminons sur un chapitre de conclusion et perspectives. Ce chapitre est l'occasion de résumer les contributions de cette thèse, et de proposer des perspectives pour les différents points qui seront développés dans nos travaux futurs.

Chapitre 1

Définitions

La définition d'un plan, d'une scène ou d'une histoire varie selon les différents domaines où apparaissent ces concepts. Ainsi, une scène a une signification différente dans une pièce de théatre ou dans un film.

Cette section a pour but de fixer une définition objective pour tous les concepts qui peuvent avoir une interprétation subjective et qui sont liés aux méthodes et expériences qui seront développées tout au long de ce document.

1.1 Segmentation en plans

Un plan est une suite d'images prises sans interruption par une caméra vidéo. Dans le domaine de la structuration automatique de vidéos, le plan est souvent considéré comme le plus petit élément de structure d'un document audiovisuel. Cependant, l'intérêt du plan varie en fonction du type de vidéo étudié. Par exemple, les caméras de vidéo surveillance filment sans interruption le même point de vue, et ainsi une séquence continue de vidéo extraite d'une caméra de surveillance n'est composée que d'un seul plan. Au contraire, pour les documents audiovisuels qui ont nécessité une étape de montage (comme les documentaires, les films, ou les séries télévisées), un plan est une courte séquence qui ne dure que quelques secondes. Les vidéos sont composées de nombreux plans agencés les uns à la suite des autres.

Cette thèse propose une étude des méthodes de structuration pour un type de document audiovisuel bien précis : les séries télévisées. La définition d'un plan est définie comme suit :

Définition d'un plan
Un plan est la plus grande suite d'images consécutives prises sans interruption par une caméra vidéo.

Deux plans consécutifs sont séparés par une transition. Il existe deux types de transitions : les transitions franches (coupures) où la première image d'un plan se situe tout de

suite après la dernière image du plan qui le précède, et les transitions progressives (parfois appelées transitions douces) où la transition implique plusieurs images (voir Figure 1.1). Cette dernière catégorie est issue d'effets de montages qui peuvent être eux-mêmes très divers : fondu enchaîné (*dissolve*), fondu à l'ouverture (*fade in*), fondu à la fermeture (*fade out*), volet (*wipe*).

Dans le domaine du cinéma en particulier, et de l'audiovisuel en général, les différentes transitions entre plans sont des effets de montage [Martin 1977]. Le montage est une opération dite de « post-production ». C'est l'assemblage des plans et des séquences de vidéo qui assure la fluidité du document audiovisuel. Les différents types de transitions entre plans sont alors utilisés pour donner des effets de liaison entre les plans ou des effets de ponctuation et de démarcation. Ainsi, les transitions franches et progressives peuvent avoir différentes significations. Par exemple, au cinéma, un fondu enchaîné marque la plupart du temps une transition entre deux scènes différentes d'un film [Martin 1977] .

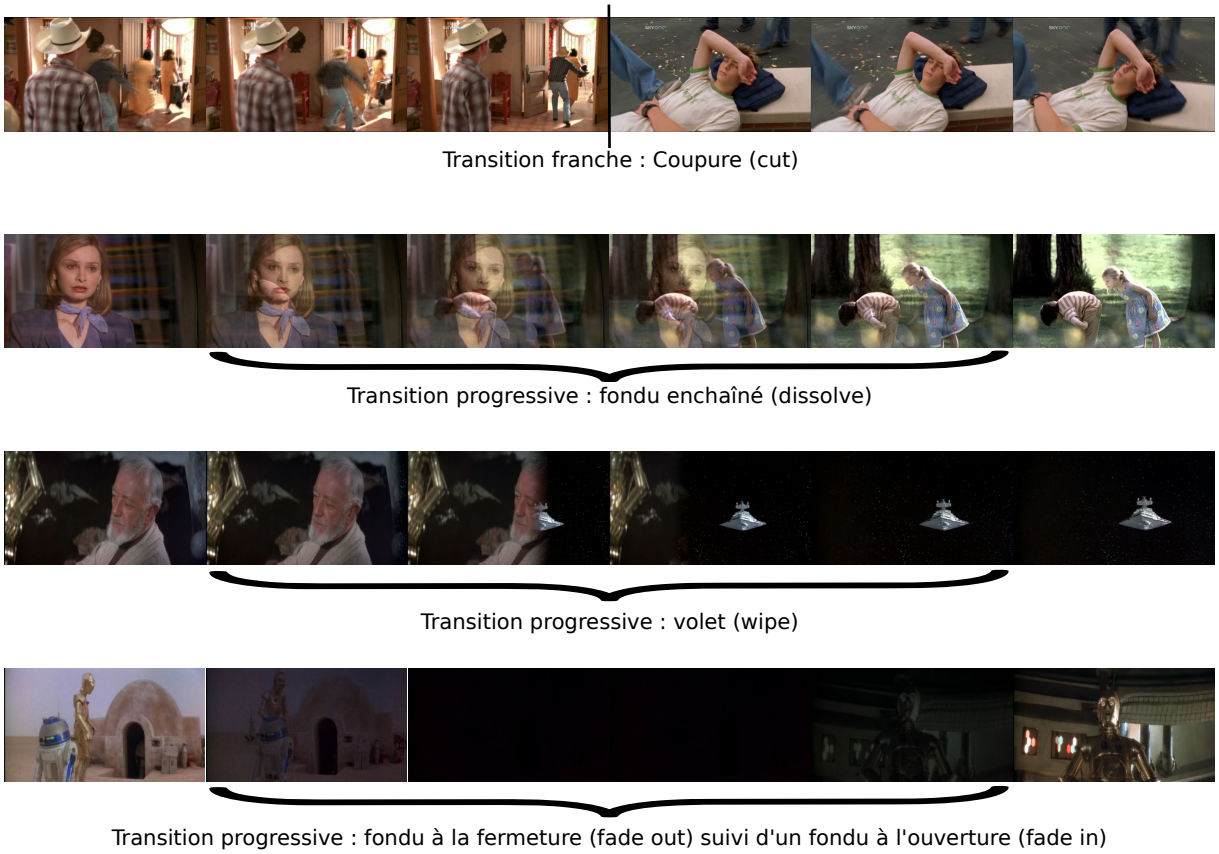

FIGURE *1.1 – Exemples de transitions entre deux plans.*

La définition de la tâche de segmentation en plans telle qu'elle sera considérée durant tout ce document respecte les critères suivants :

Segmentation en plans

- Un plan est borné par une transition de début située avant la première image du plan, et une transition de fin située après la dernière image du plan.
- Une transition franche entre deux plans p et p' est positionnée après la dernière image de p et avant la première image de p'.
- Une transition progressive étant constituée de plusieurs images, la position de la transition correspond au milieu de la transition. Dans les exemples Figure 1.1, la transition entre les plans 1 et 2 se situe toujours entre la $3^{ème}$ et la $4^{ème}$ image.

1.2 Segmentation en scènes

Un plan est souvent considéré comme le plus petit élément structurel d'une vidéo. Il est facile à définir, la position de ses bornes temporelles est bien posée, mais il est très peu porteur de « sens ». Un plan peut décrire une séquence complète d'une histoire (plan séquence). Cependant, d'un point de vue narratif, il est généralement peu représentatif de l'action en cours puisqu'il ne décrit la plupart du temps qu'une petite partie d'une action.

C'est pourquoi de nombreux travaux se sont penchés sur une notion permettant de définir une unité d'action ou des évènements dans des documents audiovisuels : la scène.

1.2.1 Qu'est-ce qu'une scène ?

Le « concept » d'unité d'action dans un film, un épisode de série télévisée, ou tout autre document audiovisuel est très subjectif. Il existe de nombreux termes pour le nommer : *paragraphe* [Wactlar 1996], *séquence* [Aigrain 1997], *unité d'histoire* [Yeung 1998], *unité logique d'histoire* [Hanjalic 1999, Vendrig 2002, Benini 2007, Sidiropoulos 2012] ou comme cela a été utilisé jusqu'ici : *scène* [Zhao 2007, Ercolessi 2011, Bredin 2012]. De même qu'il existe plusieurs façons de nommer une scène, il existe de nombreuses manières de définir ce concept.

Au théâtre, les pièces proposent toujours une structure bien définie découpée en actes et en scènes. La scène est l'unité la plus courte de la pièce, et on considère qu'un changement de scène se produit lorsqu'un ou plusieurs personnages entrent ou sortent de l'espace où se joue la pièce (espace qui est lui même appelé une scène) [Bonnabel 2000].

La notion de scène a ensuite été reprise et enrichie par le monde du cinéma et de la télévision. Ainsi, dans des ouvrages tel que *Film Encyclopedia* [Katz 1994] une scène est définie comme un segment de vidéo étant continu dans le temps et l'espace.

Dans le domaine de la recherche sur la structuration automatique de vidéos, des travaux considèrent qu'une scène est composée d'un petit nombre de plans qui sont unis par

le lieu et un événement dramatique [Yeung 1998, Benini 2007], et qui doivent respecter une continuité temporelle [Tavanapong 2004, Ercolessi 2011, Bredin 2012].

Certains de ces travaux considèrent une scène comme une suite de plans respectant une cohérence, visuelle ou auditive, dictée par des techniques cinématographiques (prise de vue, effets spéciaux, éclairage ou montage) [Sundaram 2002, Tavanapong 2004]. Pour Tavanapong *et al.* [Tavanapong 2004], une scène respecte une continuité temporelle et spatiale. La continuité temporelle signifie qu'il ne doit pas y avoir d'écart temporel dans la narration entre deux plans consécutifs d'une même scène. La continuité spatiale est dictée par des techniques cinématographiques qui garantissent une cohérence visuelle entre les plans d'une scène, et sans lesquelles la lecture de la scène serait troublante pour le spectateur [Sharff 1982].

Les règles de ces techniques cinématographiques sont présentées ci-dessous :

- **La loi des 180°**. Toutes les caméras filmant une scène sont dirigées dans la même direction et ne doivent pas franchir une ligne imaginaire appelée la ligne des 180°. La loi des 180° assure que :
  - La scène est toujours filmée dans la même direction.
  - Un gros plan sur un visage ou un objet est toujours fait suivant le même point de vue.
  - La position et le mouvement relatif des personnages sont toujours les mêmes. Les séquences A et B illustrées dans la Figure 1.2 montrent les effets du non respect de cette loi sur la relation entre les personnages lors d'un dialogue.
- **Le champ/contrechamp.** Une fois que la ligne des 180° est bien établie, une technique traditionnelle dans le montage des plans est d'utiliser un contrechamp. Un contrechamp est un plan filmé dans une direction opposée au plan précédent. Ce type de montage est particulièrement utilisé pour les dialogues entre personnages, comme illustré dans la Figure 1.2. La caméra filme le personnage qui parle, puis se concentre sur son interlocuteur, avant de revenir sur le premier personnage.
- **Vue d'ensemble/détail/vue d'ensemble.** La vue d'ensemble consiste à donner une vue globale de la scène, à montrer la position des personnages et à définir la position de la ligne des 180°. Dans la Figure 1.2, il correspond au plan de la caméra C1. Le détail consiste à montrer les détails de ce qu'il se passe dans la scène, il peut être représenté par un dialogue (champ/contrechamp), ou un mouvement de caméra suivant une action. Parfois, à la fin d'une scène, on observe de nouveau un plan large sur les personnages montrant de nouveau leur position globale : on observe de nouveau une vue d'ensemble.

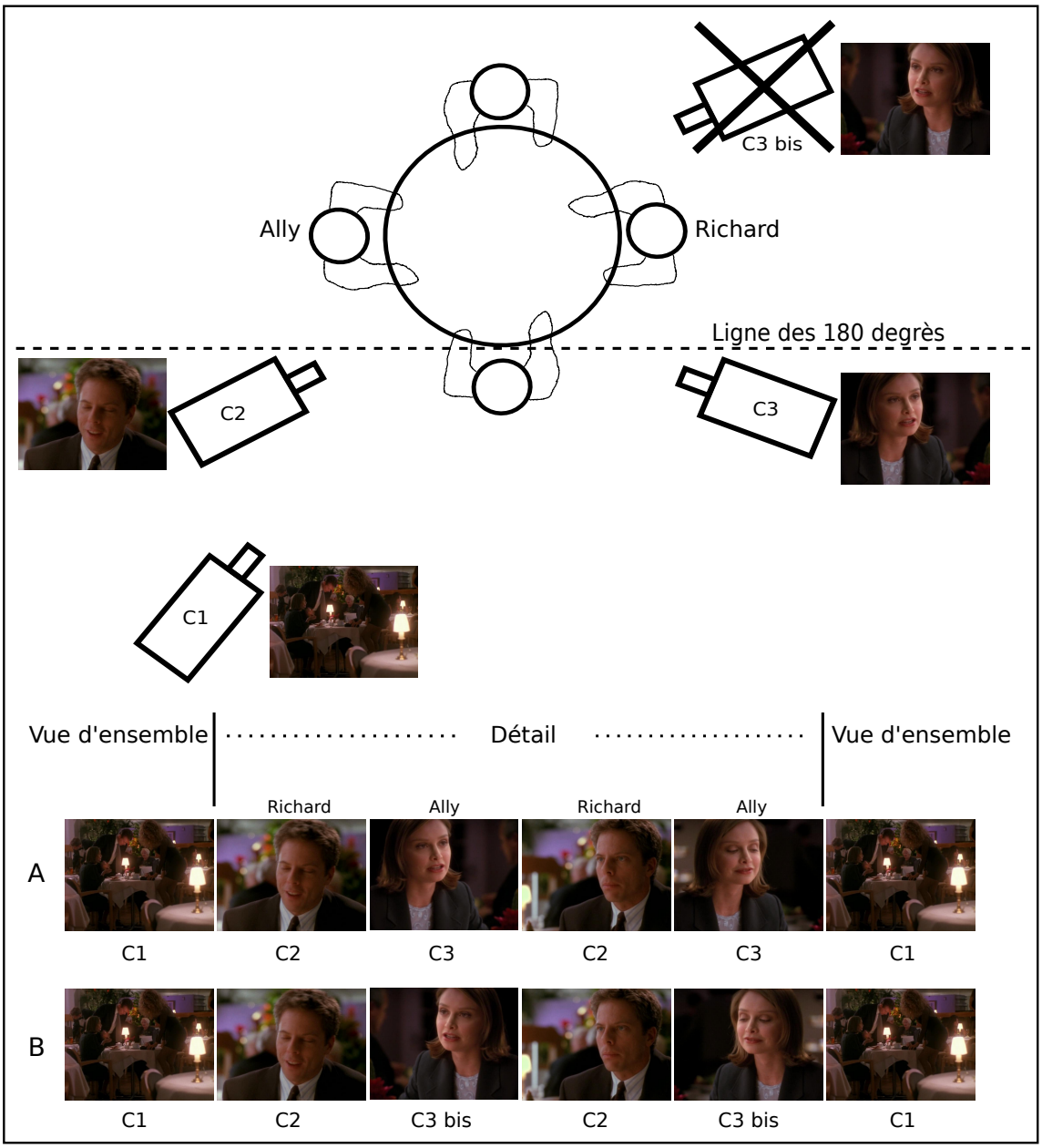

FIGURE *1.2 – Illustration des techniques cinématographiques : loi des 180°, champ/contrechamp et vue d'ensemble/détail/vue d'ensemble. Cet exemple montre un dialogue entre deux personnages représenté par la séquence s_A. Il y a 3 points de vue : une vue d'ensemble filmée par la caméra C1, et un gros plan sur les visages filmés par les caméras C2 et C3. La séquence s_B montre le résultat du non respect de la loi des 180° sur un dialogue. Changer la caméra C3 par la caméra C3bis donne l'impression qu'Ally n'est plus en face de Richard. La scène devient confuse car les deux personnages semblent ne plus parler l'un avec l'autre. L'alternance entre les plans sur Ally et Richard illustre le principe du champ/contrechamp.*

Ces différentes techniques cinématographiques permettent d'expliquer pourquoi une scène présente toujours une continuité spatiale et visuelle. La ligne des 180° implique que toutes les caméras filment dans la même direction, avec le même fond, les mêmes personnages ou objets disposés de la même façon. La technique du champ/contrechamp implique que plusieurs plans filmés exactement dans la même direction (et filmant le même sujet) se trouvent dans une même scène (mêmes fond, visage, paysage ou sujet filmé). La technique de la vue d'ensemble/détail/vue d'ensemble entraîne que plusieurs plans d'une scène peuvent montrer des plans larges avec le même fond et les mêmes personnages, assurant une continuité visuelle à la scène. Si cette continuité est brisée, c'est que la scène a changé.

D'après Truong [Truong 2002] et Tavanapong *et al.* [Tavanapong 2004], il existe des cas où la continuité spatiale d'une scène n'est pas respectée. Cela arrive par exemple lorsqu'un *flash-back* ou un *flash-forward* apparaît. Cependant, ils considèrent ce type d'événement comme un effet d'édition proche du champ/contrechamp, et donc pas considéré comme une nouvelle scène.

Dans cette thèse, nous étudions le découpage en scènes pour des épisodes de séries télévisées. Une scène est définie comme un segment temporel de vidéo ayant les caractéristiques suivantes :

Définition
- Une scène est une suite de plans consécutifs.
- Une scène décrit un unique évènement (ou aucun évènement).
- La scène respecte une continuité temporelle.
- La scène respecte une continuité spatiale dictée par des règles de montage précises.

Cette définition est fortement liée à la notion d'évènement. Un évènement est considéré comme un « fait important » de l'histoire racontée dans la vidéo. Trois classes d'évènements sont proposées et détaillées dans le Tableau 1.1 : déplacement, dialogue et interaction non verbale. Comme illustré par la Figure 1.3, les évènements racontés dans une scène peuvent appartenir à l'une de ces trois classes, ou à n'importe quelle combinaison de celles-ci. La Figure 1.4 (illustrée à la fin de cette section) montre une segmentation en scènes pour un épisode de la série télévisée *Ally McBeal*. Dans cet exemple, la scène s_7 est une combinaison d'interaction non verbale (Ally danse avec Ronnie) et de dialogue.

L'un des points de la définition est qu'une scène décrit un unique évènement ou aucun évènement. En effet, il est possible qu'il ne se passe rien de notable dans une scène, n'ayant aucun dialogue, aucune interaction, aucun déplacement. Un exemple de ce genre de scènes peut être observé dans la série *Ally McBeal*. Il arrive fréquemment qu'entre deux scènes, une courte scène de quelques secondes représentant un plan large sur une rue, ou sur la ville, soit insérée. La continuité temporelle et spatiale étant brisée, cette séquence ne peut être incluse dans la scène qui la précède ni dans celle qui lui succède. Il s'agit donc bien d'une scène, mais durant laquelle aucun évènement ne se produit.

Type d'évènement	Description
Dialogue	Dialogue impliquant un ou plusieurs personnages et traitant d'un sujet unique. Exemple : les scènes s_1, s_2 ou s_9 de la Figure 1.4.
Déplacement	Déplacement des personnages entre plusieurs lieux différents. Exemple : la scène s_8 de la Figure 1.4, où Ally et Ronnie se promènent.
Interaction non verbale	Evènement qui n'est pas un dialogue et impliquant une interaction entre plusieurs personnages ou entre un ou plusieurs personnages et des objets du décors. Exemple : la scène s_7 de la Figure 1.4 où Ally danse avec Ronnie.

TABLE *1.1 – Description des différentes classes d'évènements.*

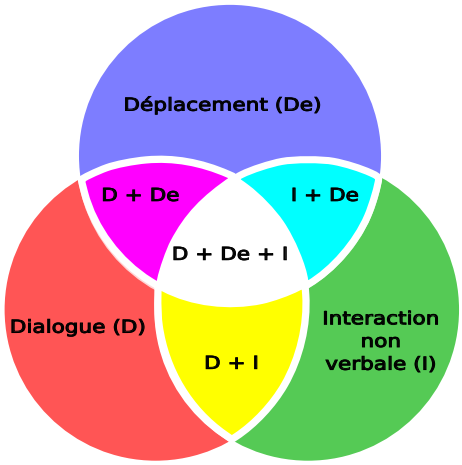

FIGURE *1.3 – Les différentes combinaisons possibles de classes d'événements.*

1.2.2 Tâche de segmentation en scènes

Même en se basant sur la définition présentée dans la section précédente, rechercher à obtenir un consensus sur la définition d'une scène est très difficile puisqu'il existe des cas particuliers sources de désaccord. Les scènes de 22 épisodes de séries télévisées ont été manuellement annotées, et voici quelques exemples de désaccords qui ont été observés entre les différents annotateurs :

- Lorsqu'un personnage entre dans une scène pour prendre part à un dialogue, certaines personnes considèrent que la perturbation introduite par l'apparition d'un nouveau personnage suffit à induire le changement de la scène. D'autres considèrent

que ce nouveau personnage était déjà présent mais « hors champ », et que son apparition n'est que la continuité logique de la scène.

- Un *flash-back* est une séquence d'un ou plusieurs plans montrant un événement passé. Ils sont souvent insérés au milieu d'une scène, soit pour montrer un souvenir d'un personnage, soit pour renforcer l'action de la scène. Un *flash-back* perturbe l'unité de temps (voire l'unité de lieu) de la scène. Pour certaines personnes, cela marque un changement de scène. Pour d'autres, un *flash-back* n'est que la visualisation de la « pensée » d'un personnage présent dans la scène et ne donne pas lieu à une scène à part puisque la continuité temporelle et de lieu est maintenue.

Pour éviter ce genre de désaccords, un ensemble de règles a été déterminé permettant de définir la tâche de segmentation en scènes de manière à ce qu'elle soit la plus objective possible. Une segmentation manuelle des scènes suivant ces règles pour un épisode de la série télévisée *Ally McBeal* est présentée à la Figure 1.4. Cet exemple montre la moitié de l'épisode 3 de la première saison d'*Ally McBeal*. Chaque image de la figure représente un plan de la vidéo. Les frontières de scènes manuellement annotées sont indiquées par des lignes verticales vertes. Les scènes sont numérotées de s_1 à s_{10}, et sont identifiées par une bande de couleur placée sous les plans, et par une courte description textuelle de l'action s'y déroulant.

Les règles suivantes sont celles qui ont été données aux annotateurs pour définir manuellement les scènes à partir desquelles les résultats des systèmes étudiés dans cette thèse seront évalués :

Règles d'annotation pour la tâche de segmentation en scènes

- Une transition entre deux scènes est aussi une transition entre deux plans. La transition entre la scène s_i et la scène s_{i+1} est la transition entre le dernier plan de s_i et le premier plan de s_{i+1}.
- Si un *flash-back* apparaît au milieu d'un ensemble de plans montrant un champ/contrechamp ou une vue d'ensemble/détails/vue d'ensemble, alors ce *flash-back* ne donne pas lieu à une nouvelle scène (exemple du *flash-back* de la scène 2 sur la Figure 1.4).
- Si un nouveau personnage entre dans une scène, mais que le sujet des dialogues ou l'événement représenté par la scène ne change pas, alors l'introduction de ce personnage ne donne pas lieu à une nouvelle scène.
- Les plans de transition, représentant des plans fixes pendant quelques secondes et utilisés pour séparer deux scènes, sont considérés comme des scènes ne décrivant aucun évènement.

FIGURE 1.4 – *Exemple de segmentation en scènes de la première moitié d'un épisode d'Ally McBeal. Chaque image représente un plan de l'épisode. Les scènes sont séparées par les lignes verticales vertes. Chaque scène est décrite par un code couleur et une courte description textuelle.*

1.3 Le regroupement des scènes en histoires

1.3.1 Qu'est-ce qu'une histoire ?

Selon le dictionnaire français [Lar 2010], une histoire telle qu'elle est racontée dans des films ou des séries télévisées, est un *récit portant sur des événements ou des personnages réels ou imaginaires, et qui n'obéit à aucune règle fixe.*

Bien qu'elle implique une certaine subjectivité, cette définition associe le mot « histoire » avec le mot « récit ». Or Chartrand expose dans son ouvrage *Grammaire pédagogique du français d'aujourd'hui* [Chartrand 1999] que la grammaire française décrit le récit comme une séquence narrative s'appuyant sur des règles préétablies. L'ensemble de ces règles, appelé « grammaire du récit », spécifie quelles sont les informations requises dans le récit et l'ordre dans lequel elles doivent être présentées.

Ainsi, une séquence narrative d'un récit, qu'il soit raconté de manière écrite ou à travers un média tel que la télévision ou le cinéma, comporte cinq parties :

- **la situation initiale :** les personnages entrent en scène, se présentent, le cadre de l'action est mis en place ;
- **l'élément déclencheur :** c'est l'évènement qui déséquilibre la situation initiale et dont découle l'ensemble des évènements suivants ;
- **le nœud (ou péripéties) :** il correspond au cœur de l'histoire, c'est une suite d'évènements qui bouleversent la continuité de l'histoire ;
- **le dénouement :** c'est l'événement ultime qui mène à une amélioration ou une dégradation de l'état des personnages ;
- **la situation finale :** elle correspond à ce qui arrive après le dénouement.

Selon *L'art des séries télé* de Vincent Colonna [Colonna 2010], chaque histoire suit un ou plusieurs personnages importants qui rencontrent des obstacles en agissant en vue d'un but. En cherchant à atteindre ce but, le héros réalise une série d'actions qui va modifier son milieu ou son état. Ainsi, contrairement aux scènes qui peuvent être décrites par un évènement unique, une histoire n'est pas une entité homogène. Elle évolue et se modifie tout au long de l'épisode. Les actions des personnages sont liées par un enchaînement causal et linéaire tel qu'illustré par la Figure 1.5, et qui est appelé : « la ligne d'actions ».

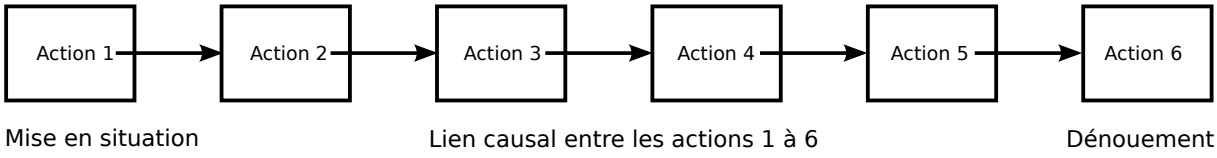

FIGURE *1.5 – Structure d'une histoire.*

Chaque ligne d'actions raconte une histoire qui doit maintenir le spectateur en haleine. Pour ce faire, l'évolution des actions épouse ce qui est appelé la structure pyramidale du

récit. D'après Gustav Freytag [Freytag 1863] *toute histoire classique, qui manifeste un procès de changement, une action aboutie, passe par cinq moments, dont trois temps forts :*

- *le nouement de l'action ;*
- *le climax ;*
- *le dénouement.*

Les cinq étapes dont fait mention Freytag sont souvent représentées visuellement sous forme d'une pyramide montrant la tension ressentie par le spectateur aux différents moments du récit (cf. Figure 1.6).

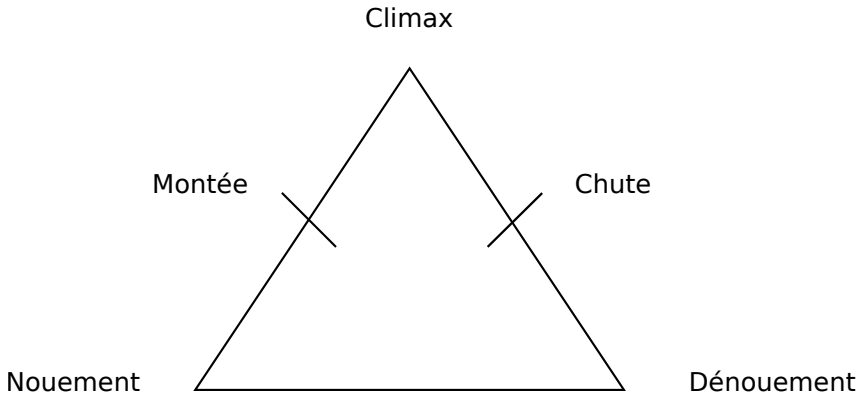

FIGURE *1.6 – Structure pyramidale du récit. La tension perçue par les spectateurs monte progressivement jusqu'à atteindre son maximum (climax), et redescend jusqu'au dénouement de l'intrigue.*

Pour faire le lien avec la structure traditionnelle du récit, le « nouement » correspond à la situation initiale et à l'élément déclencheur à partir duquel débute la montée en tension du récit. « Montée » , « climax » et « chute » sont définies par les péripéties. Le « dénouement » englobe dénouement et situation finale de la structure traditionnelle du récit.

Dans cette thèse, nous étudions une méthode permettant d'isoler différentes histoires racontées en parallèle dans des épisodes de séries télévisées. D'après Colonna [Colonna 2010], avant les années 1990, chaque épisode des séries télévisées comme *Mc Guyver, Columbo* ou *Magnum* racontait une histoire centrée sur les deux ou trois personnages principaux qui se déroulait de façon continue tout au long de l'épisode. Mais depuis le début des années 1990, la façon de raconter les fictions s'est enrichie, et dorénavant plusieurs « lignes d'action » qui sont appelées ici « histoires » se retrouvent dans chaque épisode. Ainsi de nombreuses séries télévisées, comme *Ally McBeal*, *Malcolm* ou *Le Trône de Fer*, racontent plusieurs histoires en parallèle, chacune décrivant les aventures d'un ou plusieurs personnages, ou développant différentes intrigues.

Il ressort de l'étude de Colonna sur l'évolution des séries télévisées que ces histoires peuvent être de deux sortes : celles qui sont bouclées dans un épisode, c'est à dire qui

débutent et se terminent dans l'épisode, et d'autres qui suivent ce que l'on appelle un « arc narratif ». Ce sont des histoires qui se déroulent sur plusieurs épisodes ou tout au long d'une saison d'une série.

Dans les séries télévisées américaines actuelles, on retrouve une structure dite « en ABCD ». Les épisodes sont divisés en différentes histoires qui, par convention, sont identifiées par des lettres : A, B, C, D... Ainsi, dans chaque épisode de la première saison d'Ally McBeal, l'histoire (A) suit l'arc narratif principal de la saison, c'est à dire les relations compliquées d'Ally avec ses collègues et ses amants, tandis que l'histoire (B) se concentre sur le métier d'avocat d'Ally, et on retrouve parfois une histoire (C) qui suit plus particulièrement un collègue d'Ally.

	Histoire A	**Histoire B**
Situation initiale	Ally est au bureau et prépare son rendez-vous avec Ronnie	Ally est au bureau et prépare sa journée
Élément déclencheur	Ally passe la soirée avec Ronnie, mais il ne l'embrasse pas et elle ne comprend pas sa réaction	Georgia demande à Ally de l'aider dans une affaire pour défendre une prostituée contre son rival Jack Billings
Péripéties	– Ally est perdue et ne sait plus quoi penser de Ronnie – Ally rend visite à Ronnie et lui dit ce qu'elle a sur le coeur	– Richard accepte qu'Ally aide Georgia dans l'affaire de la journaliste – Billy est réticent à ce qu'Ally et Georgia travaillent ensembles – Le procès commence et s'engage mal pour la journaliste – Ally et Georgia proposent un arrangement avec Jack Billings. Il refuse – Le procès continue
Dénouement	Ally et Ronnie se retrouve à une soirée et ils s'embrassent enfin	Ally et Georgia gagnent le procès contre toute attente
Situation finale	-	La journaliste, Ally, Georgia et leurs amis se retrouvent dans un bar pour fêter leur victoire

TABLE *1.2 – Schéma narratif des histoires présentes dans l'épisode 3 de la première saison de la série Ally McBeal.*

La Figure 1.8 montre un exemple des histoires qui sont racontées dans l'épisode 3 de la première saison de la série *Ally McBeal*. Cet exemple montre les mêmes scènes que celles présentées dans la Figure 1.4, où chaque miniature représente un plan, et une bande de couleur est présente sous les plans pour identifier ceux appartenant à la même scène. Les scènes sont numérotées de S1 à S10. On a identifié deux histoires différentes dans l'épisode présenté dans la Figure 1.8 et chacune de ces histoires suit le schéma classique du récit qui est détaillé, pour chacune d'entre elles, dans le tableau 1.2. La Figure 1.7 montre la répartition des scènes entre les deux histoires pour la totalité de l'épisode, et l'importance de chaque étape du récit.

Dans cet exemple l'histoire A présente la relation entre Ally et Ronnie. Elle correspond à un arc narratif tournant autour de la vie sentimentale d'Ally et qui rythme toute la première saison de la série. L'histoire B par contre débute lors de la deuxième scène et se termine à la fin de l'épisode. Cependant, bien que l'histoire A corresponde à un arc narratif qui débute dans un épisode précédant celui-ci, et qui se termine à la fin du dernier épisode de la saison, les différentes étapes du récit sont bien présentes dans cet épisode.

Le schéma classique du récit ne se retrouve pas toujours dans l'ensemble des histoires racontées dans un épisode, et certaines parties du récit peuvent être partagées entre les histoires. La première histoire de l'exemple présenté dans le Tableau 1.2 n'a pas de situation finale puisque l'épisode se termine sur le dénouement de cette histoire. Il arrive aussi que plusieurs histoires partagent la situation initiale, l'élément déclencheur, ou qu'il n'y ait pas de situation initiale du tout. On peut voir par exemple sur la Figure 1.7 que les deux histoires partagent plusieurs scènes lors de leurs péripéties, qui correspondent à des scènes où Ally fait un parallèle entre son état d'esprit au tribunal et sa relation compliquée avec Ronnie.

Parfois, on retrouve quelques actions isolées qui forment une ébauche d'intrigue comprenant quelques scènes et qui ne respectent pas la structure pyramidale du récit. Ces scènes, que l'on peut qualifier de pseudo-histoire, sont appelées « vignettes » par les professionnels de l'audiovisuel, et ont pour importance de préciser la vie privée de certains personnages. Dans la série *Magnum* par exemple, ces vignettes présentent la vie privée du héros et permettent de caractériser le héros, de faire aimer le héros au spectateur. Ces pseudo-histoires peuvent aussi avoir une autre importance qui est d'amorcer une histoire dans un épisode suivant.

Se contenter de définir une histoire comme une suite d'évènements respectant la structure traditionnelle du récit n'est donc pas suffisant. Cependant, on retrouve dans chaque histoire une suite d'évènements en relation les uns avec les autres et qui suivent un groupe de personnages ou un thème particulier. Ce qui caractérise le fait que deux séquences d'une même vidéo n'appartiennent pas à une même histoire, c'est le fait qu'elles montrent des évènements disjoints qui n'ont pas d'influence, ou très peu, les uns sur les autres. Les scènes composant une histoire sont ainsi liées par une idée maîtresse propre à chaque histoire. Dans notre exemple, l'une de ces idées est la relation amoureuse d'Ally avec

Ronnie, l'autre concerne le travail d'avocat d'Ally, et ces deux idées n'ont pas de relation importante l'une avec l'autre.

La définition suivante est basée sur cette notion « d'idée maîtresse » pour considérer que deux évènements sont en relation et appartiennent à la même histoire.

Définition d'une Histoire
- Une histoire est un ensemble d'évènements qui partagent une même idée maîtresse.
- Une scène décrivant un évènement unique, une histoire est donc l'ensemble de toutes les scènes qui partagent la même idée maîtresse.

1.3.2 Tâche de regroupement des scènes en histoires

La Figure 1.7 montre les scènes regroupées en deux histoires d'un épisode de la série *Ally McBeal*. On peut remarquer que les deux histoires évoluent en parallèle tout au long de l'épisode de manière « entrelacée ». L'ensemble des actions représentant une histoire ne se suivent pas directement dans un épisode. L'opération qui consiste à retrouver les histoires dans un épisode de série télévisée n'est donc pas seulement une opération de segmentation. Cette tâche est définie comme suit :

Tâche de regroupement des scènes en histoire
À partir de l'ensemble des scènes d'un épisode, la tâche de regroupement des scènes en histoire consiste à regrouper les scènes partageant la même idée maîtresse.

Pour éviter toute ambiguïté dans les annotations qui ont permis d'évaluer les résultats présentés dans cette thèse, un ensemble de règles ont été définies :

Règles d'annotation pour la tâche de regroupement des scènes en histoire
- Un ensemble de scènes est considéré comme une histoire si on y retrouve une structure proche de la structure d'un récit, comprenant au moins un élément déclencheur, un ensemble de péripéties et un dénouement.
- Si on ne retrouve pas la structure du récit pour un ensemble de scènes qui sont considérées comme des « vignettes », cet ensemble est tout de même considéré comme une histoire.
- Les scènes qui ne suivent aucune idée maîtresse, comme les scènes ne décrivant aucun évènement, ne doivent pas être inclues dans les histoires.

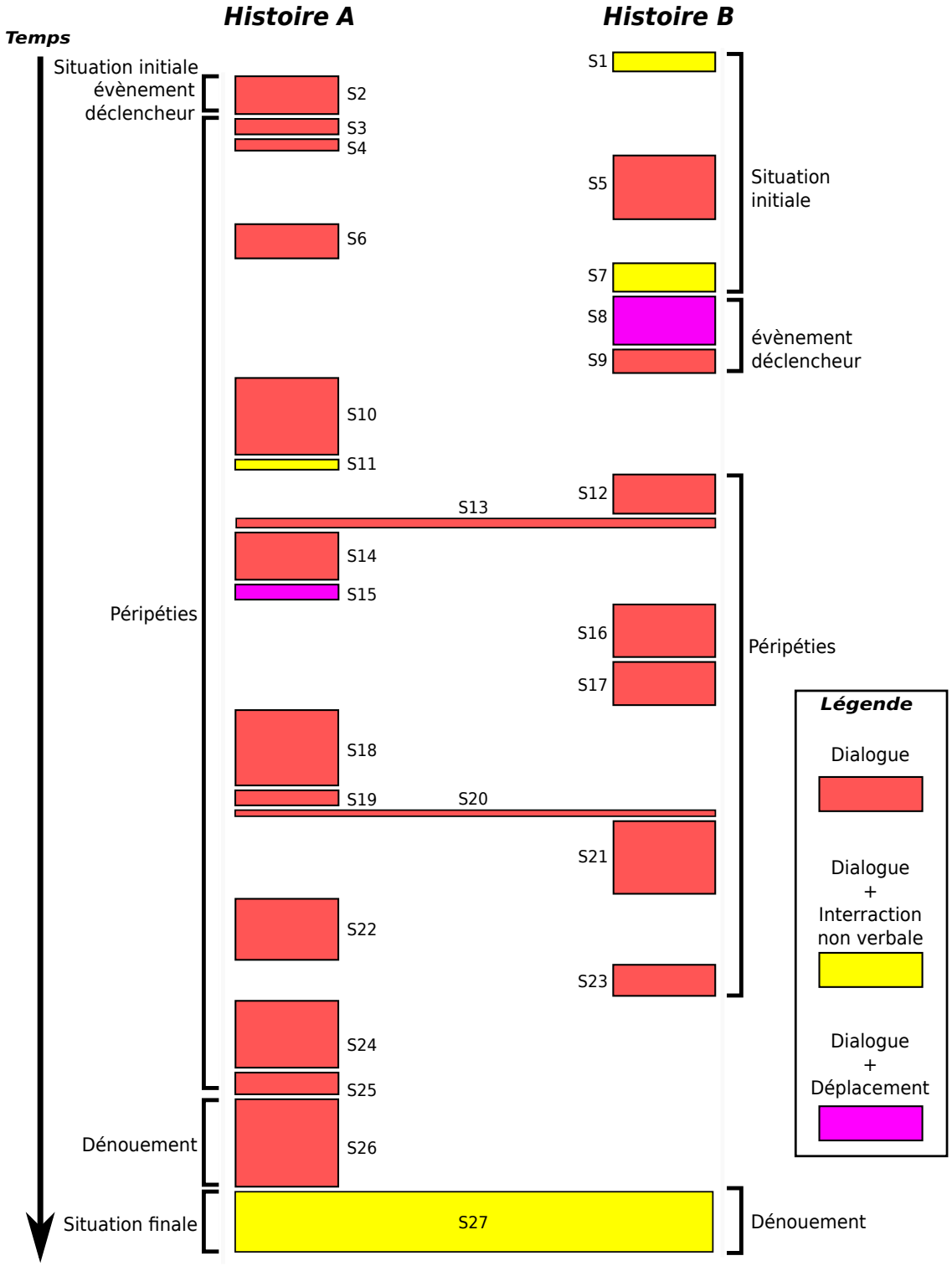

FIGURE 1.7 – *Scènes et histoires d'un épisode d'*Ally McBeal. *Chaque rectangle est une scène. La couleur associée à la scène correspond au type d'évènement décrit par la scène tel que spécifié par la Figure 1.3*

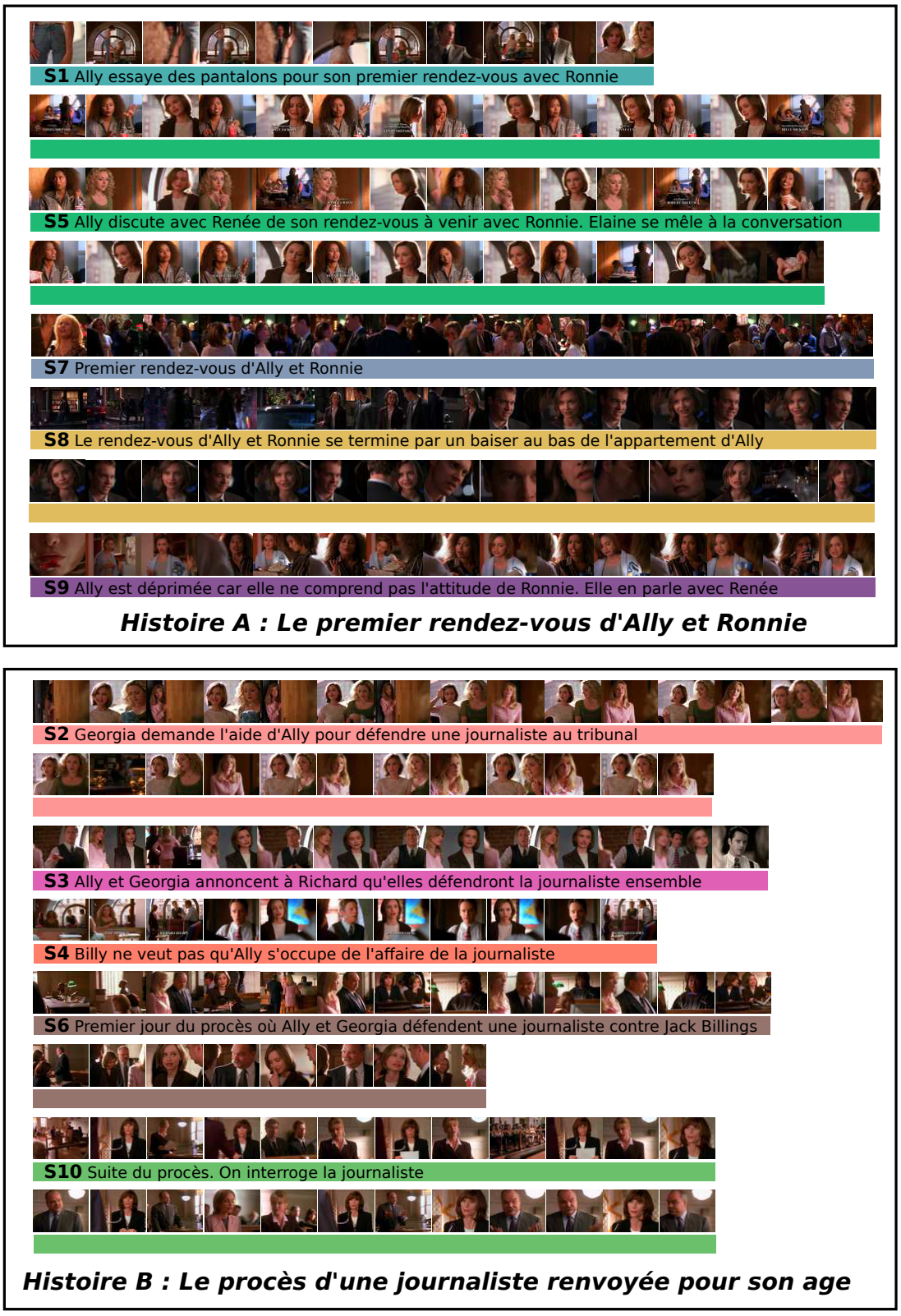

FIGURE *1.8 – Exemple de la répartition des scènes en histoires pour la première moitié d'un épisode d'Ally McBeal. Chaque image représente un plan de l'épisode.*

1.4 Discussion

La définition d'une scène ou d'une histoire varie suivant le domaine étudié. Ainsi, les définitions présentées dans ce chapitre ne sont pas universelles. Elles sont le fruit de notre expertise dans le domaine de la structuration automatique de vidéos narratives. Les définitions proposées tiennent compte de la définition d'une scène et d'une histoire dans des domaines variés comme le théâtre ou le cinéma, ainsi que de la représentation collective que l'on peut avoir de ces concepts. Elles permettent de définir les bases sur lesquelles sont construits les chapitres suivants.

Le Chapitre 4 présente les expériences menées sur le domaine de la segmentation en scènes. Le Chapitre 5 discute d'une méthode de regroupement des scènes en histoires. Pour évaluer les résultats présentés dans ces deux chapitres, nous avons annoté les scènes et les histoires de 22 épisodes de séries télévisées. 7 épisodes de la première saison de la série *Ally McBeal*, 7 épisodes de la saison 4 de *Malcolm* et 8 épisodes de la première saison de la série *Le trône de fer*.

Ces séries ont toutes un format très différent. *Ally McBeal* et *Malcolm* sont toutes deux des comédies humoristiques (sitcom). Cependant, *Ally McBeal* propose des épisodes de 40 minutes parfois composés d'arcs narratifs couvrant plusieurs épisodes, alors que la série *Malcolm* propose un format de 20 minutes par épisodes où les histoires sont essentiellement présentes dans un seul épisode. *Le Trône de Fer* est une série de type *heroic fantasy*, qui propose des épisodes de 50 minutes et des histoires s'étalant toutes sur plusieurs épisodes.

Le Tableau 5.2 résume le nombre de scènes et d'histoires annotées pour chaque série, en suivant les règles présentées dans ce chapitre. Elles permettent d'évaluer et comparer les résultats obtenus pour les différents épisodes et séries de manière claire et objective.

Série	**Nombre d'épisodes**	**Durée**	**Nombre de scènes**	**Nombre d'histoires**
Ally McBeal	7	5h ≈41min/épisode	304 moyenne de 43 scènes/épisode	19 moyenne de 2,7 / épisode
Malcolm	7	2,5h ≈22min/épisode	196 moyenne de 28 scènes/épisode	24 moyenne de 3,4 / épisode
Le Trône de Fer	8	7h ≈50min/épisode	244 moyenne de 30 scènes/épisode	34 moyenne de 4,2 / épisode

TABLE *1.3 – Description des annotations.*

Chapitre 2

Etat de l'art

Les contributions de cette thèse se font sur deux domaines de la structuration automatique de documents audiovisuels : la segmentation en scène et le regroupement des scènes en histoires ainsi que les applications possibles d'une telle structuration de vidéo.

L'une des principales difficultés du domaine de la structuration automatique de document audiovisuel concerne le « fossé sémantique » (semantic gap). Il représente l'écart observé entre les « concepts » audiovisuels créés par les utilisateurs d'un média, et leur représentation numérique directement interprétable par une machine. Les principaux défis de la structuration de document audiovisuel sont de trouver comment extraire l'information importante présente dans une vidéo, et comment l'organiser et l'utiliser pour retrouver la structure des documents sous la forme de « concepts sémantiques de haut niveau » comme les scènes ou les histoires.

Cet état de l'art est découpé en quatre parties. Les deux premières présentent des méthodes générales de description et d'organisation de données provenant de documents audiovisuels. La première partie discute des différentes manières de décrire numériquement les images et le son qui nous permettent de retrouver les « concepts » qui nous intéressent, c'est-à-dire la structure d'un épisode de série télévisée sous forme de scènes et d'histoires. La deuxième partie, s'intéresse aux outils de regroupement de données (*clustering*). En effet, les méthodes existantes de segmentation en scènes et le regroupement des scènes en histoires utilisent tous ces techniques existantes de regroupement.

La Section 2.3 s'intéresse aux méthodes existantes de segmentation en scènes de documents audiovisuels. La Section 2.4 présente les méthodes proches du principe de regroupement des scènes en histoires, qui consistent à regrouper des documents ou des séquences de documents audio-visuels selon leur thème ou leur sujet.

2.1 Description et comparaison de segments de documents audiovisuels

Un document audiovisuel peut être divisé en deux composantes : **le flux audio** et **le flux vidéo** (voir Figure 2.1). Il est facile d'isoler ces deux flux puisqu'ils sont capturés et transmis par des outils différents (caméra et écran pour la vidéo, microphone et haut-parleur pour le son). Le flux vidéo contient toutes les informations visuelles du document audiovisuel. Il est composé d'une suite d'images statiques et simule le mouvement en utilisant le principe de la persistance rétinienne découvert par Isaac Newton et le Chevalier d'Arcy [le Chevalier D'Arcy 1765] aux $XVIII^{ème}$ et $XIX^{ème}$ siècles. Le flux audio contient toutes les informations sonores : parole, musique, bruit. Un son est un phénomène physique correspondant aux vibrations mécaniques du milieu qui nous entoure. Enregistrer un son de façon numérique consiste à enregistrer et discrétiser ces vibrations, on obtient alors une information à une dimension représentant le signal sonore tel qu'illustré dans la Figure 2.1.

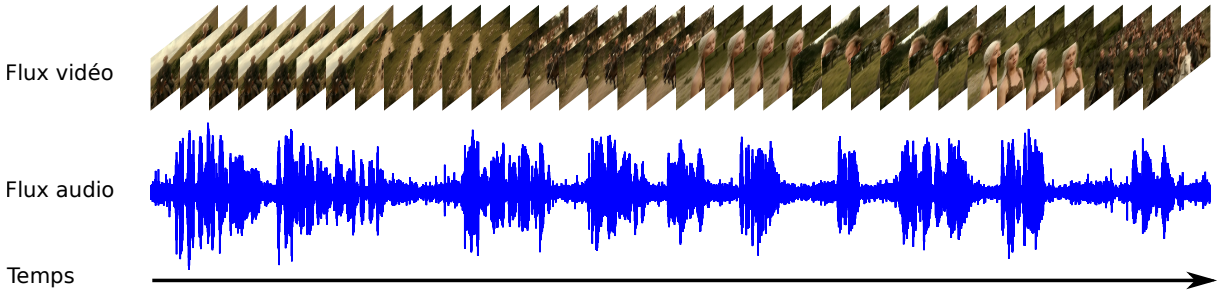

FIGURE *2.1 – Composition d'un document audiovisuel : Un flux vidéo composé d'une succession d'images statiques simulant le mouvement et un flux audio représentant les vibrations sonores.*

Les deux composantes peuvent être utilisées pour décrire un segment de document audiovisuel. On appelle descripteurs des valeurs permettant de caractériser l'information contenue dans les segments audio ou vidéo et permettant de comparer différentes séquences de documents audiovisuels. Le but de cette comparaison est de savoir si deux séquences sont proches ou non. L'idée de « proche » dépend de l'application qui nous intéresse : deux séquences peuvent être proches si elles ont été filmées dans le même lieu, si elles présentent les mêmes personnages ou des objets en commun, si la couleur globale des images est similaire, si elles ont été filmées la nuit ou le jour, si l'ambiance sonore est calme, etc. Il existe ainsi de très nombreuses façons de décrire les flux vidéo et audio d'un segment de document audiovisuel.

Que les informations proviennent des flux vidéo, audio, ou des deux à la fois, les différentes méthodes permettant de décrire un segment d'un document audiovisuel peuvent être classées suivant leur niveau d'interprétation sémantique.

Deux niveaux différents de descripteurs sont étudiés dans cet état de l'art :

- **Descripteurs bas niveau :** directement extraits du contenu brut du signal (image ou son). Ils le décrivent par un ensemble de valeurs difficilement interprétables par un être humain.
- **Descripteurs haut niveau :** ils sont orientés vers la sémantique du contenu de la scène observée. Le segment de vidéo est décrit par les objets ou les concepts sémantiques qui y sont présents (jour/nuit, visages, objets, etc.).

Quel que soit le niveau d'interprétation sémantique du descripteur, la comparaison de deux segments ne nécessite pas toute l'information présente dans un segment audiovisuel. Ainsi, pour la comparaison visuelle de deux segments audiovisuels, il est courant de sélectionner une ou plusieurs images (**images clefs**) qui serviront de base à la description du segment dans son ensemble. Dans ce cas, décrire et comparer deux segments audiovisuels revient à décrire et comparer leurs images clefs. C'est pourquoi beaucoup de méthodes présentées dans les sections suivantes ont été empruntées au domaine de l'indexation et de la recherche d'images.

Les méthodes de description présentées dans cet état de l'art sont classées suivant leur niveau d'interprétation sémantique. Les descripteurs bas niveau extraits du flux vidéo et du flux audio sont étudiés en premier lieu, avant de présenter des descripteurs haut niveaux adaptés à notre problématique. Dans chaque partie, les descripteurs et les méthodes de comparaison de séquences utilisant ces descripteurs seront étudiés.

2.1.1 Descripteurs bas niveau

Dans cette section, 3 façons de décrire un segment de vidéo sont étudiées.

- **Une description visuelle globale**, où un segment audiovisuel est décrit par une analyse globale du contenu visuel des images qui le composent.
- **Une description visuelle locale**, où seule une partie de l'information contenue dans les images est utilisée pour la description.
- **Une description de l'information acoustique**.

2.1.1.1 Descripteurs visuels globaux

Couleur

La manière la plus commune pour décrire une image est de décrire la couleur présente dans cette image. La plupart du temps, cette information est décrite sous forme d'**histogrammes de couleur**. Pour une image en niveau de gris, l'histogramme est défini comme une fonction discrète qui associe à chaque valeur d'intensité possible pour un pixel d'une image, le nombre de pixels ayant cette valeur. Il est courant de faire une quantification des histogrammes en regroupant plusieurs valeurs d'intensité en une seule

classe pour réduire la dimension du descripteur. Pour des images en couleur, on peut considérer les différentes composantes de la couleur indépendamment, ou toutes à la fois. Dans ce cas, pour des espaces de couleur à 3 composantes, on obtient un histogramme à 3 dimensions, où chaque valeur représente le nombre de pixels ayant une couleur précise.

Les histogrammes de couleur sont très utilisés dans le domaine de la recherche d'images. Swain *et al.* [Swain 1991] et Niblack *et al.* [Niblack 1993] les utilisent dans des systèmes d'indexation d'images et de recherche d'images. Dans le domaine de la segmentation en scène, de très nombreux travaux se basent sur les histogrammes de couleur pour mesurer la similarité entre deux segments de vidéo.

Utiliser des histogrammes de couleur pour décrire une image de manière globale offre des avantages très intéressants : ils sont rapides à calculer, et deux images capturées dans un même lieu et filmant un même sujet dans les mêmes conditions d'illumination ont généralement des histogrammes similaires. Cependant les histogrammes de couleur montrent quelques inconvénients listés dans les trois points suivants et illustrés dans la Figure 2.2.

- **Aucune information spatiale** : si la couleur est globalement la même entre deux images mais distribuée de façon différente, les deux images seront tout de même considérées comme similaires.
- **Sensible à la variation d'intensité des couleurs** : une image ayant subi une modification d'intensité de ses couleurs voit son histogramme se décaler. Une comparaison des histogrammes d'une image avec la même image ayant ses intensités de couleur décalées donnera une faible similarité.
- **Aucune information sur la forme et les objets** : comparer une tasse rouge avec une tasse blanche donne toujours une faible similarité alors qu'il peut s'agir de deux image proche si l'on considère le « concept » tasse. Comparer une tasse rouge avec un camion rouge donnera un résultat proche alors que ce n'est pas toujours souhaitable.

Les histogrammes de la valeur d'intensité des pixels d'une image ne sont pas les seuls moyens de décrire la couleur globale d'une image. Stricker *et al.* [Stricker 1995] proposent de décrire une image à partir de la distribution des couleurs de l'image. La distribution des couleurs est caractérisée par ce qu'ils appellent les « **moments de couleur** ». La couleur est décrite à partir de trois moments : le premier moment (couleur moyenne de l'image) et les deuxièmes et troisièmes moments (variance et asymétrie de la distribution des couleurs) pour chaque composante de la couleur. L'image est alors décrite par un vecteur composé de ces moments, qui peut être considéré comme un histogramme de moments.

Pour Keysers *et al.* [Keysers 2007] et Deselaers *et al.* [Deselaers 2008], **les pixels de l'image offrent directement une description globale de la couleur** contenant l'information spatiale de la répartition et de l'intensité des couleurs. L'idée est de réduire les images à une plus petite dimension (par exemple à une dimension de 32x32 pixels pour

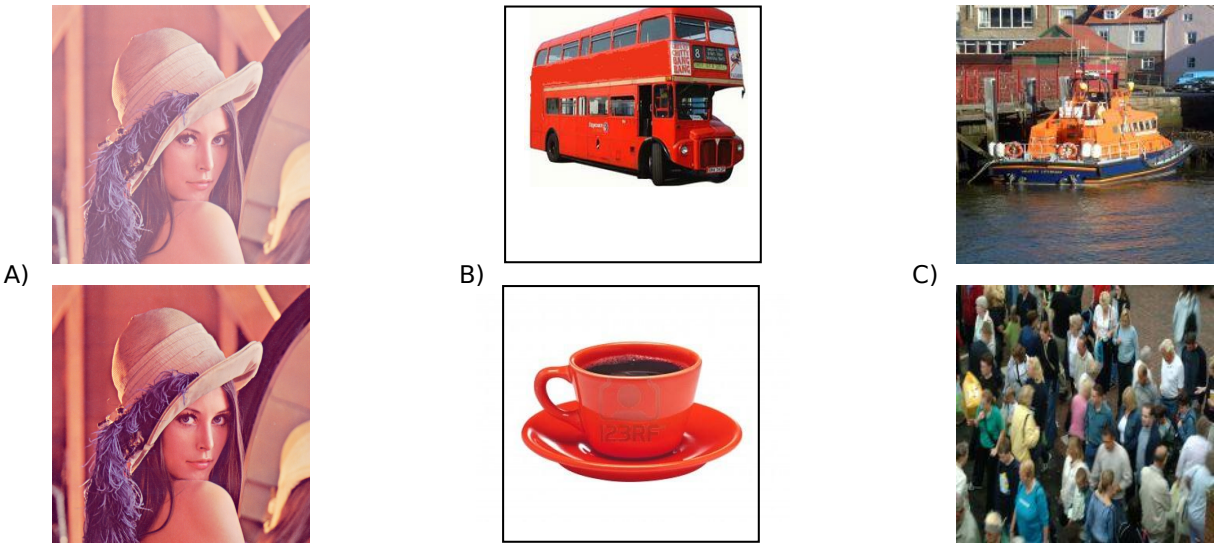

FIGURE 2.2 – *Illustration des limites des histogrammes de couleur. En A les deux images montrent une différence de luminosité, leurs histogrammes de couleur sont « décalés » et difficiles à comparer bien que les images représentent exactement le même sujet. En B et en C sont représentées deux images pouvant être décrites par des histogrammes de couleur proches bien que le sujet représenté soit très différent.*

Keysers *et al.*), permettant d'avoir une information réduite de la répartition des pixels de l'image. L'intérêt de cette méthode est de conserver l'information spatiale de la couleur dans l'image. Son inconvénient est qu'elle devient sensible à la translation des objets dans l'image : deux images représentant exactement les mêmes objets seront différentes si les objets ne sont pas à la même place.

La couleur n'est pas toujours le meilleur moyen pour décrire une image. Elle est très sensible aux différences d'illumination, et de contraste. Il est donc intéressant de décrire une image, non par la valeur des pixels qui la composent, mais par leur arrangement au sein de l'image.

Décrire cet arrangement peut se faire à l'aide de **descripteurs de texture**. D'après Smith et Chang [Smith 1996], une texture fait référence aux motifs visuels qui ont des propriétés d'homogénéité qui ne résultent pas de la présence d'une seule couleur ou d'une intensité. Haralick *et al.* [Haralick 1973] ajoutent que c'est une propriété innée pour toutes les surfaces qui contiennent une importante information à propos de l'arrangement structurel des surfaces et leurs relations avec leur environnement. De nombreux descripteurs permettent de décrire une texture : les descripteurs de Haralick [Haralick 1973], de Tamura [Tamura 1978], ou plus récemment des descripteurs utilisant l'information fréquentielle des pixels de l'image à partir du calcul de la transformée en ondelettes [Misiti 2007].

Mouvement

Les descripteurs de mouvement permettent de décrire la dimension temporelle de la vidéo. De nombreux types de descripteurs peuvent être extraits du mouvement pour représenter un segment de vidéo : l'activité du mouvement, l'activité de la caméra, la trajectoire du mouvement, etc...

L'**intensité du mouvement** est le descripteur de mouvement le plus populaire. Il capture la notion intuitive « d'intensité de l'action » ou de « rythme de l'action ». Par exemple, des scènes d'activité intense correspondent à des séquences de sport, de courses de voitures ou de bagarre, et au contraire, des séquences avec peu d'action correspondent à un dialogue ou à un présentateur annonçant un reportage.

D'après Jeannin et Divakaran [Jeannin 2001], l'activité du mouvement peut être directement extraite des informations de compression des vidéos. Dans les vidéos compressées suivant les normes MPEG, une image est divisée en macro-blocs, à partir desquels l'image suivante est calculée par un déplacement de ces blocs. Ce déplacement est représenté par des vecteurs de mouvement dont la magnitude représente la magnitude du mouvement présent entre les deux images. La Figure 2.3 montre un exemple de vecteurs de mouvements extraits entre deux images. L'intensité du mouvement peut être définie par la moyenne ou l'écart-type de ces vecteurs, et la direction est déterminée par la direction des vecteurs.

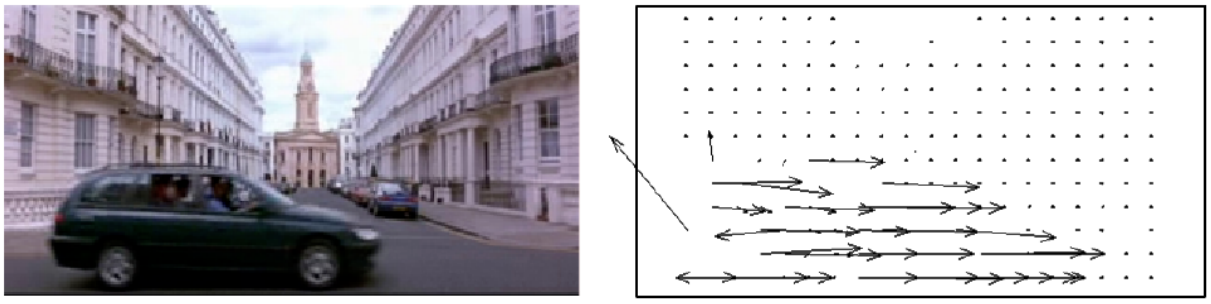

FIGURE *2.3 – Vecteurs de mouvement extraits d'une séquence vidéo*

2.1.1.2 Descripteurs visuels locaux

La section précédente présente des méthodes permettant de décrire des images de façon globale. Une autre idée pour décrire et comparer des images est d'utiliser des descripteurs locaux. Leur but est de fournir une description robuste aux changements d'illumination, d'orientation ou de point de vue.

Utiliser les descripteurs locaux pour la description d'une image se fait en général en trois étapes :

- Détection de points d'intérêt.
- Description des points d'intérêt.
- Réduction de la dimension.

Détection de points d'intérêt

La détection des points d'intérêt consiste à rechercher les points jugés « intéressants » pour décrire une image. Ce sont des points présentant des propriétés locales remarquables. Par exemple, les « coins » sont des points de l'image où le contour change brutalement de direction (comme les quatre sommets d'un rectangle). Il s'agit de points particulièrement stables et donc intéressants pour la répétabilité de l'opération de détection (ces points peuvent être détectés dans deux images différentes mais représentant la même scène). Il existe de nombreux détecteurs de points d'intérêt (détecteur de Moravec [Moravec 1981], de Shi&Tomasi [Shi 1994], SUSAN [Smith 1995], ...), mais le plus populaire reste le **détecteur de Harris** [Dorkó 2006] qui est une amélioration du détecteur de Moravec lui permettant d'être moins sensible au bruit de l'image.

Une autre idée consiste à extraire une **grille dense de points d'intérêt**. Ainsi, les points extraits sont répartis uniformément sur l'ensemble de l'image. Cela permet d'éviter que les points extraits soient tous regroupés dans une même zone de l'image, mais cette méthode perd la propriété de répétabilité des points extraits. La Figure 2.4 montre les points qui sont extraits avec un détecteur de Harris (à gauche) et à partir d'une grille dense de points d'intérêt (à droite).

FIGURE 2.4 – *Exemple de points d'intérêt extraits avec un détecteur de Harris et à partir d'une grille dense de points d'intérêt.*

Description des points d'intérêt

Une fois les points d'intérêt détectés, leur description permet de décrire les images. Il existe de nombreuses façons de décrire les points d'intérêt. Wang *et al.* [Wang 2010] proposent de calculer un **histogramme de couleur en se basant uniquement sur des régions d'intérêt des images**. Les régions d'intérêt sont tous les pixels présents aux alentours des points d'intérêt détectés. Ces pixels sont alors utilisés pour générer un histogramme de couleur représentatif de l'image. En utilisant un détecteur de points de Harris, ils considèrent que leur système résout en partie le problème spatial posé par les histogrammes de couleurs, et offre une comparaison robuste à la translation des objets et au bruit.

Les **SIFT** (*Scale-Invariant Feature Transform* ou transformation de caractéristiques visuelles invariante à l'échelle) [Lowe 2004] sont des descripteurs invariants à l'échelle, à l'angle d'observation et à l'exposition (luminosité). Deux images, représentant le même objet ou la même scène, auront beaucoup de chances d'avoir des descripteurs SIFT similaires, même si elles sont capturées selon des angles de vue différents.

De nombreux autres descripteurs locaux existent comme les *differential invariants* [Koenderink 1987], les *steerable filters* [Freeman 1991], les *shape context* [Belongie 2002], les *spin images* [Lazebnik 2003], etc. Une comparaison de ces descripteurs est proposée par Mikolajczyk *et al.* [Mikolajczyk 2005]. Leur étude est basée sur la comparaison d'images. Le but est de mettre en correspondance des images dont les différences sont des facteurs d'échelle, d'orientation, de point de vue ou d'illumination. Bien qu'elle ne prenne pas en compte des descripteurs plus récents comme les **SURF** (Speeded Up Robust Features), il ressort de cette étude que les descripteurs basés sur des SIFT donnent les meilleures performances, et particulièrement sur des images avec des textures complexes. Les SURF [Bay 2006] sont des descripteurs de points d'intérêt inspirés des SIFT. Ils ont pour vocation d'être aussi efficaces que les SIFT, mais beaucoup plus rapides à calculer.

Décrire les images par des points d'intérêt est très efficace [Mikolajczyk 2005], mais il est fréquent que plus de 1000 points soient extraits pour chaque image [Deselaers 2008], et même en réduisant la taille des descripteurs de points avec une analyse en composante principale (ACP), la complexité de la description est très grande, et donc comparer des images à partir de ce type de description est très long. C'est pourquoi des méthodes ont été développées pour utiliser efficacement les SIFT, SURF ou même les histogrammes de couleur locaux présents dans une image.

Sacs de Mots (Bag of Words ou BOW)

Pour réduire le temps de comparaison des images, il existe une méthode utilisant les descripteurs locaux et inspirée par les méthodes de classification de texte : les ***Sacs de Mots* (Bag of Words ou BOW)** parfois appelés *Sacs de points d'intérêt* (Bag of Keypoints) [Csurka 2004] ou *Sacs de mots visuels* (Bag of visual words) [Yang 2007]. Cette

méthode propose de réduire la taille des données de description d'une image en estimant la distribution des descripteurs locaux présents dans l'image. Le principe consiste à extraire des points d'intérêt d'un ensemble d'images à comparer, chaque point étant décrit par un descripteur local (SIFT, SURF ou histogramme de couleur par exemple). Ensuite, un regroupement de ces points d'intérêt est réalisé. Les groupes obtenus permettent de représenter chaque point d'intérêt par un numéro de groupe, ce qui permet de discrétiser les points d'intérêt à partir du numéro de groupe qui lui est associé. Chaque image est décrite en extrayant de l'image les descripteurs locaux, et en déterminant pour chaque descripteur local le groupe qui le représente le mieux. Un histogramme de ces groupes est alors créé comme descripteur de l'image.

L'histogramme a la même taille que le nombre de groupes proposé lors de l'étape de regroupement des descripteurs locaux. Ainsi, ce processus permet de déterminer la taille du descripteur en faisant varier le nombre de groupes. Ce processus est décrit dans la Figure 2.5.

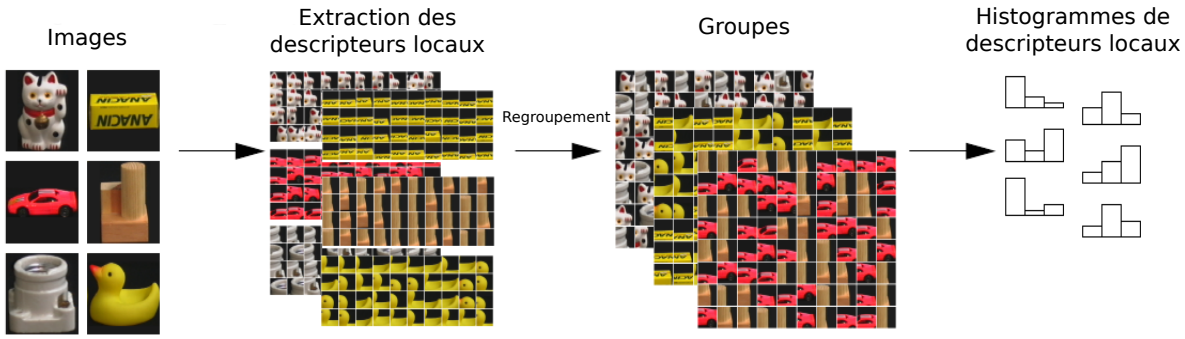

FIGURE 2.5 – *Création d'un histogramme de descripteurs locaux*

2.1.1.3 Descripteurs audio

Jusqu'à présent, les descripteurs décrits dans cet état de l'art étaient extraits du flux vidéo d'un document audiovisuel. Il existe de nombreux descripteurs extraits du flux audio, appelés descripteurs audio ou descripteurs acoustiques. Ces descripteurs peuvent être extraits directement du signal audio, ou dépendre d'une analyse fréquentielle du signal.

Les descripteurs les plus utilisés dans le domaine de l'indexation audio sont les descripteurs cepstraux, et plus particulièrement les « **Mel-Frequency Cepstrum Coefficient (MFCC)** ». Les MFCC ont été introduits par Mermelstein [Mermelstein 1976]. Ils sont souvent utilisés en reconnaissance de la parole et en identification du locuteur ou de la langue. Les paramètres issus des MFCC sont bien adaptés au signal de parole [Ganchev 2005], et ils sont utilisés pour des mesures de similarité de zones acoustiques [Müller 2007, Sundaram 2002].

Ils sont calculés selon le schéma de la Figure 2.6. La première étape consiste à appliquer un fenêtrage de Hamming du signal qui permet d'isoler un ensemble d'échantillons du signal (quelques millisecondes) à partir desquels seront calculés les MFCC. Pour chaque fenêtre, le calcul du module de la transformée de Fourier rapide (FFT) permet de modifier l'espace des données en passant du domaine temporel à un domaine fréquentiel (ou spectral). Ensuite, l'amplitude du spectre est pondérée par un banc de filtres triangulaires espacés selon l'échelle de Mel (étape de filtrage). L'échelle non linéaire Mel est connue pour rendre compte de la perception humaine [Stevens 1937]. Après l'application d'un logarithme pour accentuer les aigus (car les composantes fréquentielles aiguës sont toujours plus faibles que les graves), une transformée de Fourier inverse est calculée pour retourner dans le domaine temporel. Il en résulte un ensemble de coefficients, dits coefficients cepstraux espacés selon l'échelle de fréquences Mel.

FIGURE *2.6 – Création de descripteurs MFCC pour un signal* x

D'autres descripteurs peuvent être extraits directement du signal audio comme le taux de passage à zéro (zero-crossing rate), l'énergie du signal, la variance de l'énergie ; ou issus de l'analyse fréquentielle du signal comme le flux spectral. Tous ces descripteurs sont parfois utilisés pour décrire et comparer des segments audio [Sundaram 2002, Lu 2006], ou ils sont utilisés de manière combinée pour retrouver des informations de plus haut niveau sémantique comme par exemple des segments de parole ou de musique.

2.1.2 Descripteurs haut niveau

Les descripteurs haut niveaux visent à décrire un segment de vidéo à l'aide de notions directement interprétables par un être humain.

Cet état de l'art présente deux types d'informations haut niveau qui sont très utilisées dans le domaine de la structuration de vidéos : les concepts sémantiques, et les personnages.

2.1.2.1 Concepts sémantiques

L'utilisation de concepts sémantiques pour caractériser une séquence de vidéo est très utile pour mettre en relation deux séquences à partir de notions compréhensibles par un être humain. Les concepts recherchés peuvent être des objets (avion, voiture), des événements (marche, course, applaudissement), une description de la scène (extérieur, intérieur, forêt, désert), etc. Pour la structuration de vidéos, l'utilisation des concepts pour décrire des segments de vidéos se fait de telle sorte que deux segments de vidéos sont considérés comme proches s'ils partagent un ensemble de concepts sémantiques.

La détection de concept est principalement une tâche de classification permettant de déterminer si le segment de vidéo contient un concept sémantique donné. Le domaine de recherche associé à la détection de concepts est très actif ces dernières années, notamment grâce aux campagnes d'évaluation TRECVID [Over 2011].

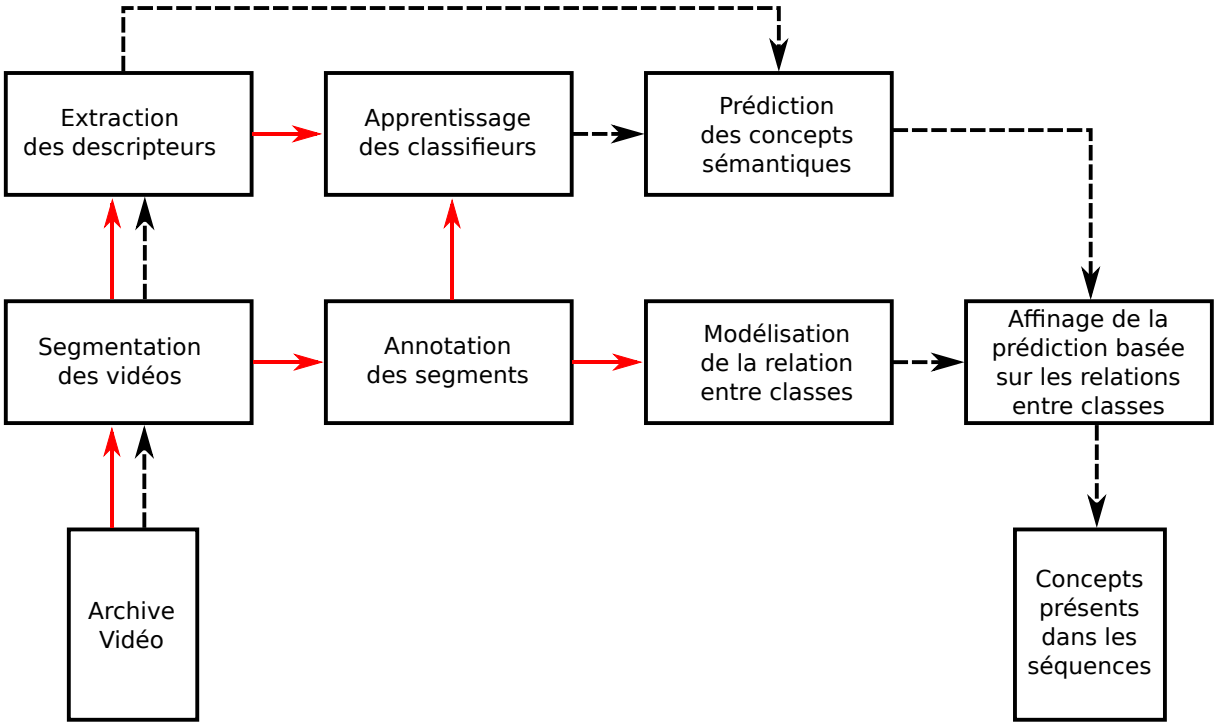

FIGURE 2.7 – *Méthode de détection des concepts dans des segments de vidéo [Weng 2008]*

Le schéma de la Figure 2.7 montre le fonctionnement d'un système de détection de concepts proposé par Weng *et al.* [Weng 2008]. À partir d'un ensemble de vidéos, une première étape consiste à les découper (en général en plans). Une partie des vidéos (ensemble d'apprentissage) servent à apprendre un classifieur à partir d'annotations manuelles et de descripteurs bas-niveau extraits des segments de vidéos tels que décrits dans les sections précédentes. Ces classifieurs sont alors utilisés pour détecter la présence ou l'absence des concepts dans des segments de vidéos. Parfois, une étape supplémentaire est rajoutée, consistant à définir une relation entre les concepts pour affiner la prédiction, en considérant que si un ou plusieurs concepts sont présents dans la séquence, alors il est probable qu'un autre concept l'est aussi.

Les méthodes proposées pour la détection de concepts diffèrent sur les descripteurs bas-niveau et le classifieur employés. Les descripteurs les plus utilisés sont les « Sacs de mots (BOW) », les histogrammes de contours, les moments de couleur, et la texture de l'image [Ngo 2009, Zha 2012]. D'autres travaux s'intéressent à la fusion des différentes modalités de la vidéo en associant les descripteurs visuels locaux à des descripteurs audio (MFCC) [Chang 2007, Jiang 2010].

Utiliser les concepts pour décrire des segments audiovisuels permet de caractériser les séquences par les objets, les actions, ou les lieux qui sont présents durant la séquence. Cependant, en ce qui concerne la structuration de vidéos narratives telles que les films et les séries télévisées, ou les émissions télévisuelles de type talk-shows, il existe un concept qui offre une information très pertinente pour retrouver la structure des vidéos et qui est l'objet de nombreux travaux de recherche : la détection des personnages.

2.1.2.2 Personnages

Savoir qu'un personnage est présent dans une séquence audiovisuelle est une information primordiale pour la structuration de vidéos. Ce sont les personnages qui vivent et créent les actions qui sont montrées au spectateur. L'information qui nous intéresse concernant les personnages consiste à savoir si un personnage est présent durant une séquence d'un document audiovisuel, et savoir dans quelle autre séquence ce même personnage est présent. Ce processus ne propose aucune notion d'identification (le système ne peut répondre à la question « Qui est ce personnage ? »). C'est donc un processus en deux étapes :

- détection des personnages ;
- regroupement des instances du même personnage.

Il existe plusieurs façons d'obtenir ce résultat dans un document audiovisuel : à partir du flux vidéo, en recherchant les visages dans les images et en mettant en correspondance les visages d'une même personne ; à partir du flux audio, en recherchant les segments de parole et en regroupant ceux appartenant aux mêmes locuteurs ; ou bien en utilisant les informations provenant de ces deux sources.

Détection et regroupement des visages dans le flux vidéo

Le principe consiste à trouver la position des visages dans le flux vidéo, et à regrouper les visages appartenant à un même personnage. De nombreuses conditions rendent la détection des visages difficile, comme la pose (visage frontal, de profil, à 45 degrés, vus de dessus), la présence ou l'absence de caractéristiques particulières (barbe, moustache, lunette), l'expression faciale, les occlusions, l'éclairage ou la résolution de l'image. La détection de visage consiste à déterminer si un visage est présent ou non dans une image, et si c'est le cas, quelle sont la position et la taille de chaque visage (cf. Figure 2.8).

D'après Robert Frischholz[1], il existe plusieurs façons de détecter des visages dans des images. Une idée est d'utiliser la **couleur typique de la peau** pour rechercher le visage [Störring 1999, Stoerring 2004, Li 2010]. Une autre idée consiste à détecter les zones en mouvement dans le cas de plans statiques (sans mouvements de caméra) et de combiner cette information avec un modèle de couleur de peau [Darrell 1998].

1. http ://facedetection.com/

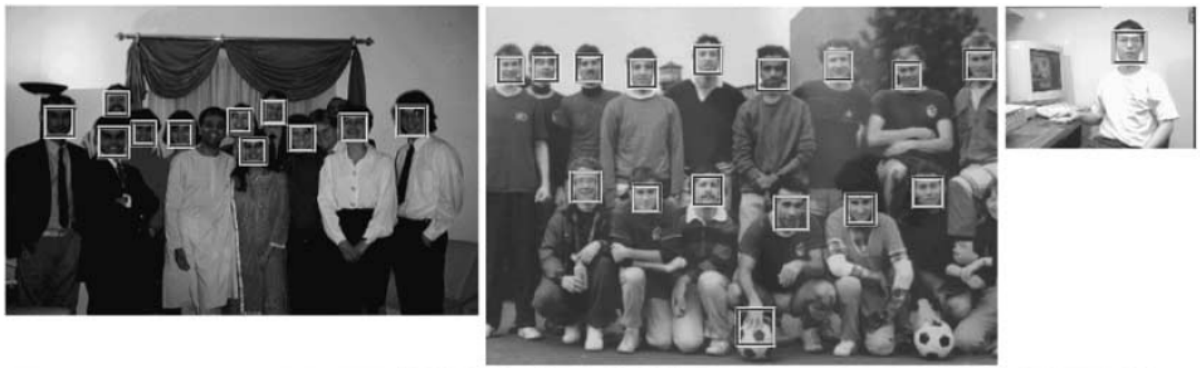

FIGURE 2.8 – *Exemple de détection des visages dans une image [Viola 2004]*

Ces méthodes sont très contraignantes et peu applicables à des épisodes de séries télévisées. On leur préférera celle proposée par Viola et Jones [Viola 2004]. Ils proposent d'utiliser une **cascade de classifieurs faibles en utilisant des descripteurs de Haar**. On retrouve une implémentation de cette approche dans OpenCV [2]. Comme elle utilise une cascade de classifieurs, une étape d'apprentissage est nécessaire. Elle permet de modéliser des visages de face aussi bien que des objets ou des visages de profil.

Sur un ensemble d'images contenant 507 visages vus de face, leur algorithme permet de retrouver correctement plus de 90% des visages avec moins de 50 faux positifs [Viola 2004]. Un exemple des résultats obtenus par cette méthode est présenté à la Figure 2.8.

Une fois le visage détecté, la seconde étape consiste à **regrouper les visages appartenant à un même personnage**. Certains chercheurs voient ce problème comme une tâche de reconnaissance des visages [Arandjelovic 2005, Bicego 2006], d'autres comme un problème de classification [Everingham 2006, Peng 2008]. La plupart du temps ces méthodes se basent sur des exemples de visages comme base du regroupement, et utilisent des critères tels que la couleur de la peau ou des cheveux, voire des vêtements comme mesure de similarité entre les visages. Comme la couleur est une information sensible aux changements d'illumination, certains travaux comme ceux proposés par Bicego [Bicego 2006] et El-Khoury [El Khoury 2010] utilisent les descripteurs SIFT pour décrire et comparer les visages.

La Figure 2.9 illustre la comparaison de visages deux à deux à partir des points d'intérêt. Une mesure de similarité entre les visages est déterminée en fonction du nombre de descripteurs SIFT similaires. Ils proposent ensuite de faire un regroupement agglomératif basé sur cette mesure de similarité entre les visages pour regrouper les visages décrivant la même personne.

2. http ://opencv.org/

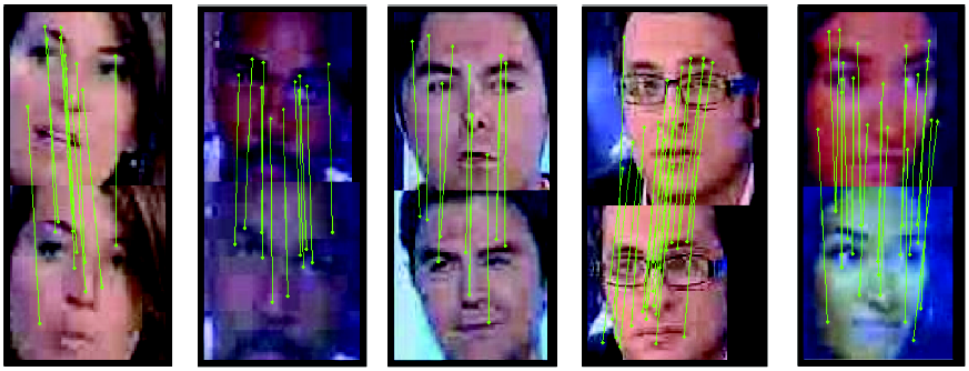

FIGURE *2.9 – Comparaison de visages à l'aide de descripteurs SIFT [El Khoury 2010]*

Segmentation et regroupement en locuteurs

Il est aussi possible de déterminer la présence de personnages à partir du flux audio d'un document audiovisuel. Dans ce cas, ce ne sont pas les visages qui sont recherchés mais les tours de parole (segments de parole d'un locuteur unique).

C'est un domaine issu de l'indexation audio appelé « segmentation et regroupement en locuteurs » (*speaker diarization*). L'architecture générale d'un système de segmentation et regroupement en locuteurs est résumée dans la Figure 2.10. Des descripteurs bas niveau sont extraits du signal audio. À partir de ces descripteurs, le signal est partitionné en segments de parole homogènes ou en segments de non-parole. Les tours de paroles appartenant à un même locuteur sont ensuite regroupés pour être en mesure de savoir quel locuteur parle à quel moment, bien que l'identité du locuteur reste inconnue.

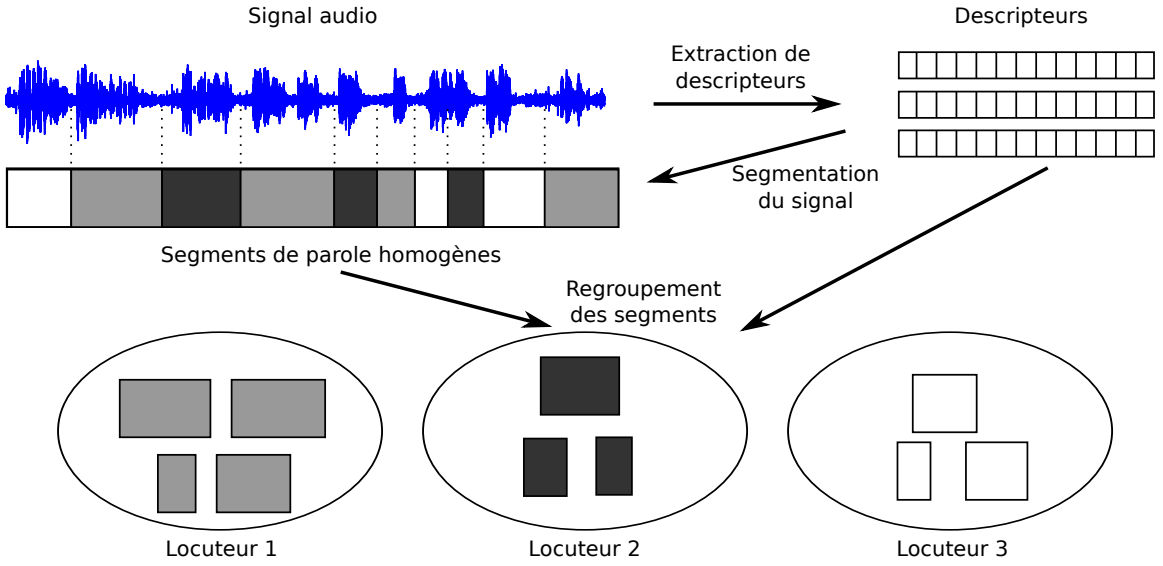

FIGURE *2.10 – Architecture générale d'un système de segmentation et regroupement en locuteurs*

Un certain nombre de contraintes rendent cette tâche difficile pour des documents audiovisuels : le nombre de personnes, le langage parlé, l'identité et le genre des locuteurs sont inconnus, le signal ne propose pas que de la parole, et les personnages peuvent parler simultanément. De plus, il existe un certains nombre de systèmes de segmentation et regroupement en locuteurs, comme celui développé par le LIMSI [Barras 2006], par IBM [Huang 2008], ou par le LIA [Bozonnet 2010], mais tous ont été optimisé pour des document de type « radio » et ne sont pas optimisés pour des documents audiovisuels de type série télévisée.

2.1.3 Mesures de similarité/distance entre deux vecteurs de descripteurs

Un segment de vidéo est décrit par un ensemble de valeurs extraites du flux vidéo ou du flux audio. Comparer deux segments audiovisuels revient à comparer leurs images clefs ou leur flux audio qui peuvent être décrits de différentes manières. Que ce soit à partir des concepts sémantiques, des locuteurs, des différentes méthodes de description par la couleur ou par la texture, la plupart des descripteurs sont représentés sous forme vectorielle. **Mesurer la distance entre deux images revient donc à mesurer la distance des vecteurs les décrivant**.

Novak *et al.* [Novak 1992] proposent une étude de la description d'images par histogrammes de couleurs. Ils concluent qu'utiliser des distances L1 (ou **distance de Manhattan**) (2.1), L2 (**distance euclidienne**) (2.2) ou la **similarité cosinus** (2.3) offrent de bons résultats pour la recherche d'images dans des bases de données. Un histogramme de couleur est un vecteur de descripteurs de couleur. Ce type de distance/similarité peut être appliqué à n'importe quel type de vecteur.

$$d_{\mathrm{L1}}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{N} |\mathbf{x}_i - \mathbf{x}'_i| \tag{2.1}$$

$$d_{\mathrm{L2}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{x}'_i)^2} \tag{2.2}$$

$$\mathrm{sim}_{\cos}(\mathbf{x}, \mathbf{x}') = \frac{< \mathbf{x}, \mathbf{x}' >}{||\mathbf{x}|| \cdot ||\mathbf{x}'||} \tag{2.3}$$

avec $\mathbf{x}$ et $\mathbf{x}'$ deux vecteurs de données de dimension N.

Les distances L1 ou L2 considèrent l'espace de données comme isotrope. Or ce n'est pas toujours le cas. Une façon de remédier à ce problème est d'utiliser **une distance euclidienne normalisée** comme celle proposée par Mahalanobis [Mahalanobis 1936]. Cette distance utilise la variance de chaque composante (estimée à priori) pour la normaliser :

$$d_{\text{Mahalanobis}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{N} \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{\sigma_i^2}} \tag{2.4}$$

Avec σ_i^2 la variance de la $i^{ème}$ composante. L'avantage de cette méthode est de tenir compte de la distribution des différents descripteurs inclus dans l'histogramme représentatif des images. Cependant, un nombre suffisant d'exemples de vecteurs descripteurs doit être disponible pour pouvoir estimer la variance.

Hanjalic *et al.* [Hanjalic 1999] proposent une méthode permettant de mesurer la distance entre deux séquences vidéo basée sur la **comparaison de blocs de couleurs**. L'idée est de décrire un segment de vidéo à partir de plusieurs images, elles mêmes découpées en blocs 4x4. Chaque bloc est décrit par un vecteur descripteur de couleur. La distance entre deux séquences s_i et s_j est alors calculée en minimisant la distance entre leurs blocs de couleur de la manière suivante :

$$d_{\text{bloc}}(s, s') = \min_{\text{toutes combinaisons de blocs entre } B \text{ et } B'} \sum_{b \in B'} d_{\text{L2}}(b, b') \tag{2.5}$$

avec B et B' l'ensemble des descripteurs des blocs des séquences s et s', et b et b' les descripteurs des blocs inclus dans B et B'. L'intérêt de cette méthode est d'être moins sensible aux déplacements de caméra dans la séquence vidéo. En effet, la caméra pouvant filmer selon différents points de vue, des blocs de couleur similaires peuvent se retrouver à des positions spatiales et temporelles différentes selon les images.

2.2 Méthodes de regroupement de données

Les problèmes du domaine de la structuration de vidéos sont souvent modélisés comme des problèmes de classification ou de regroupement (clustering) de données. Par exemple, en considérant qu'une donnée peut être une image ou un segment de vidéo, la segmentation en scène peut être vue comme un problème de regroupement de plans appartenant à la même scène. Retrouver les histoires telles qu'elles sont présentées dans le Chapitre 1 consiste aussi en un regroupement des scènes appartenant à la même histoire. Ainsi, dans cette section, nous nous intéressons aux méthodes de regroupement de données. Il existe de très nombreuses méthodes de regroupement de données. Cette section de l'état de l'art a pour but de décrire les méthodes qui sont employées dans cette thèse et traitant de la problématique du regroupement de séquences audiovisuelles. La question de la comparaison des séquences ayant été discutée dans la section précédente, cette section propose uniquement une description générale des méthodes de regroupement.

2.2.1 Regroupement hiérarchique agglomératif

Le principe du regroupement agglomératif est de regrouper de manière séquentielle les groupes les plus proches. Soit K le nombre désiré de groupes, N le nombre de données à regrouper, $\{x_i, ..., x_N\}$ l'ensemble des données à regrouper.

Algorithme 1 Regroupement agglomératif

$K' \leftarrow N$
$G_i \leftarrow \{x_i\}\ \forall i \in [\![1, N]\!]$
répéter
  $(i, j) = \arg\min_{(i,j)}\ d(G_i, G_j)$
  $G_i \leftarrow G_i \cup G_j$
  supprimer G_j
  $K' \leftarrow K' - 1$
jusqu'ï¿½ $K = K'$

Cette procédure se termine quand le nombre spécifié de groupes est atteint. Si l'on continuait jusqu'à ce que K' atteigne 1, on pourrait représenter l'ensemble du regroupement sous forme de dendrogramme, comme présenté à la Figure 2.11. A chaque niveau la « distance » entre les deux plus proches groupes donne la valeur de dissimilarité du regroupement pour ce niveau.

Cet algorithme dépend donc de cette mesure de distance entre deux groupes pour rechercher à chaque étape quels sont les deux groupes les plus proches. Nous considérons trois façons de calculer cette distance. Soit G_i et G_j deux groupes de données.

$$\text{Single-link clustering} \quad : \quad d_{\min}(G_i, G_j) = \min_{x \in G_i \ x' \in G_j} d(x, x') \tag{2.6}$$

$$\text{Complete-link clustering} \quad : \quad d_{\max}(G_i, G_j) = \max_{x \in G_i \ x' \in G_j} d(x, x') \tag{2.7}$$

$$\text{Average-link clustering} \quad : \quad d_{\text{moy}}(G_i, G_j) = \frac{1}{|G_i||G_j|} \sum_{x \in G_i} \sum_{x' \in G_j} d(x, x') \tag{2.8}$$

L'autre paramètre à prendre en compte dans l'Algorithme 1 concerne la condition d'arrêt de la boucle. Dans l'algorithme présenté ici, le regroupement continue jusqu'à ce qu'on obtienne le nombre désiré de groupes. Une autre condition d'arrêt traditionnellement utilisée est de continuer le regroupement jusqu'à ce que la distance minimum entre les deux groupes les plus proches soit plus grande qu'une valeur donnée. Cette deuxième façon de considérer la condition d'arrêt consiste donc à définir jusqu'à quel point on peut descendre dans le dendrogramme de la Figure 2.11.

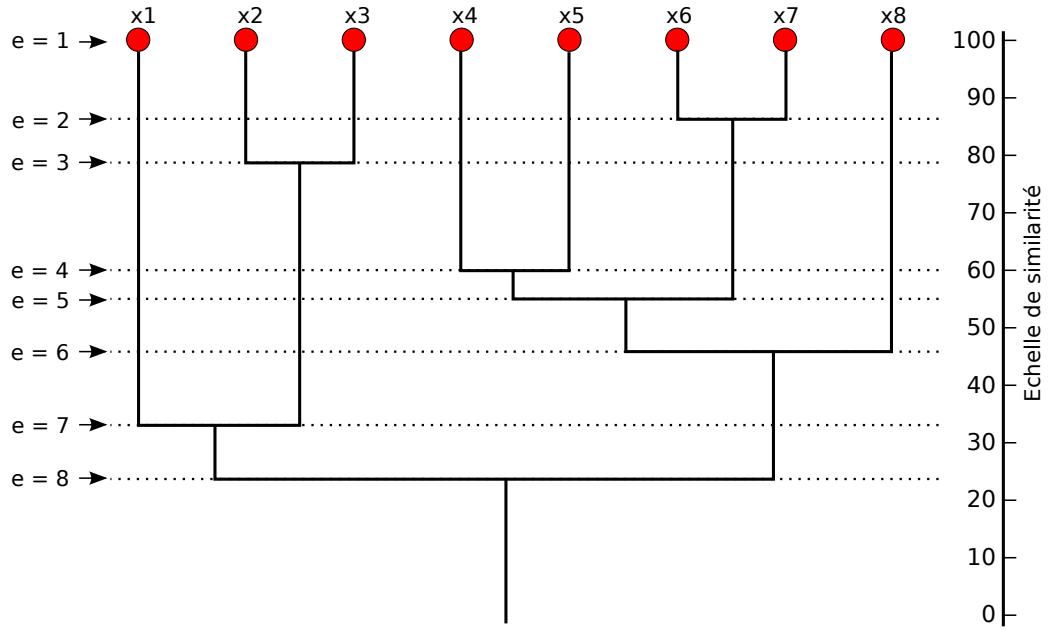

FIGURE *2.11 – Un dendrogramme représentant le résultat d'un regroupement agglomératif hierarchique. L'axe vertical montre une mesure générale de similarité entre les groupes de données. Ici à l'étape e=1, les huit points sont des groupes singletons. Les points* x_6 *et* x_7 *sont les plus similaires et sont fusionnés à l'étape e=2. À l'étape e=3 ce sont les points* x_2 *et* x_3 *qui sont fusionnés et ainsi de suite.*

2.2.2 Regroupement basé sur des graphes

Les algorithmes de regroupement de données basés sur des graphes considèrent une modélisation des données sous forme de graphe, où chaque noeud du graphe représente une donnée, et les liens entre les noeuds représentent la similarité entre ces données.

Ces algorithmes ont principalement été développés pour le domaine de la recherche de communautés au sein de réseaux ou de graphes. Le principe consiste à regrouper les noeuds du graphe qui sont proches les uns des autres et qui sont éloignés des autres noeuds. Il existe de nombreux algorithmes comme la « méthode METIS » [Karypis 1998] ou les « algorithmes de maximisation de modularité » [Newman 2006]. Ce sont ces derniers qui sont étudiés dans cette thèse et plus particulièrement la méthode de Louvain [Blondel 2008].

Il s'agit d'une méthode heuristique basée sur la maximisation de la modularité notée par :

$$\mathcal{Q} = \frac{1}{\sum_{i,j} A_{ij}} \sum_{i,j}^{N_c} \left[A_{ij} - \frac{\sum_k A_{ik} \sum_k A_{kj}}{\sum_{i,j} A_{ij}} \right] \delta_{ij} \tag{2.9}$$

où $\delta_{ij} = 1$ si les nœuds i et j appartiennent à la même communauté et 0 sinon. N_c est le nombre de données et A_{ij} est le poids de l'arête reliant les nœuds i et j.

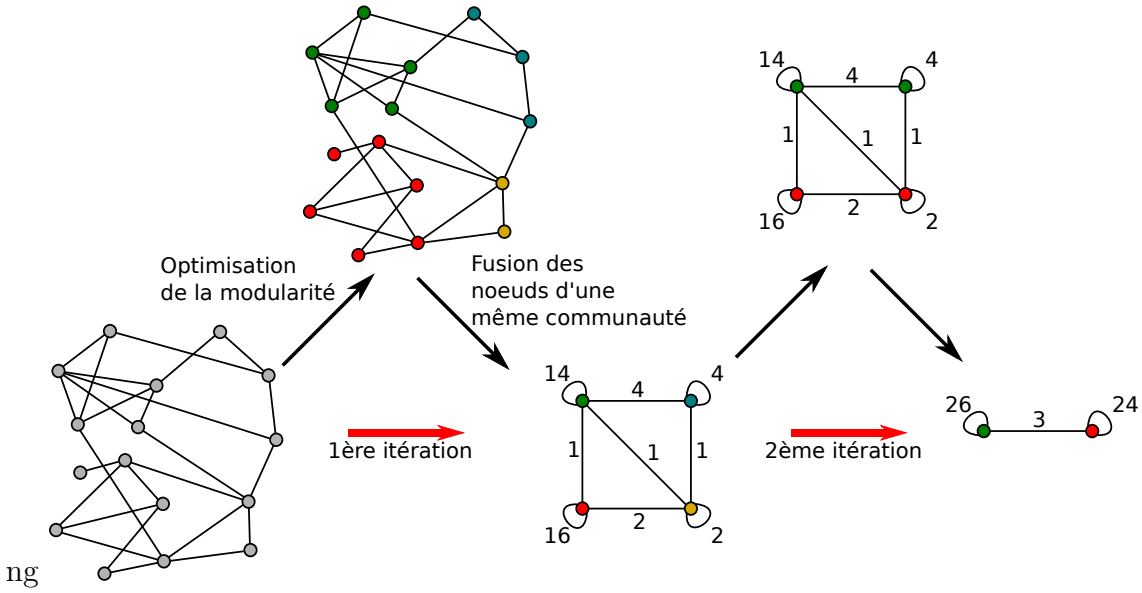

FIGURE *2.12 – Visualisation de l'algorithme de Louvain. Chaque itération est constituée de deux phases : une première phase où les nœuds sont regroupés de façon à optimiser la modularité ; une seconde phase où les nœuds d'une même communauté sont fusionnés pour former un nouveau graphe. Les étapes sont répétées itérativement jusqu'à ce que la modularité ne puisse plus augmenter. [Blondel 2008]*

$\mathcal{Q}$ peut donc être vue comme une mesure de la qualité des communautés détectées. En effet, plus les arêtes intracommunautés sont fortement valuées et plus les arêtes intercommunautés sont faiblement valuées, plus cette mesure de modularité augmente [Newman 2006].

L'algorithme de Louvain fonctionne en plusieurs itérations. Chaque itération est composée de deux étapes. La première étape consiste à regrouper les nœuds de façon à optimiser la modularité. Dans la seconde étape, les nœuds d'une même communauté sont regroupés pour former un nouveau graphe. Ces étapes sont répétées récursivement jusqu'à ce que la modularité ne puisse plus augmenter. Ce processus est illustré dans la Figure 2.12 Les groupes finaux, ou communautés finales, sont représentés par l'ensemble des nœuds qui ont été fusionnés à chaque étape de l'algorithme.

2.2.3 K-Moyennes

Le but d'un regroupement de type K-Moyennes est de regrouper les données en k différents groupes. En considérant le centroïde d'un groupe de données comme étant la moyenne des données du groupe. Soit K le nombre désiré de groupes, N le nombre de données et μ_1, μ_2, ..., μ_k les centroïdes des groupes de données, tel que μ_i est la moyenne

de toutes les données présentes dans le groupe i. L'algorithme des K-Moyennes fonctionne comme suit :

Algorithme 2 Regroupement K-Moyennes

Initialisation des μ_i $\forall i \in [\![1, K]\!]$
répéter
  Classer les N données pour qu'elles soient associées à leur plus proche μ_i
  $\mu_i \leftarrow$ moyenne des données du groupe i
jusqu'ï¿½ ce que les μ_i ne changent pas

Tout au long de l'algorithme, le déplacement des centroïdes tend à minimiser la distance moyenne des données à leur centroïde. Une étape d'initialisation est nécessaire pour positionner les centroïdes. La qualité du regroupement dépend fortement de cette étape d'initialisation. Il existe plusieurs façons d'initialiser la position de centroïdes :

- position aléatoire des centroïdes dans l'espace des données ;
- association aléatoire des données à regrouper aux centroïdes ;
- position des centroïdes de façon à ce qu'ils soient le plus éloignés possible les uns des autres [Khan 2004].

Dans le cas de données en deux dimensions, on peut représenter le partitionnement par un diagramme de Voronoï. La position des centroïdes détermine la position des segments de séparation entre les classes. Un exemple du fonctionnement des K-Moyennes est proposé dans la Figure 2.13.

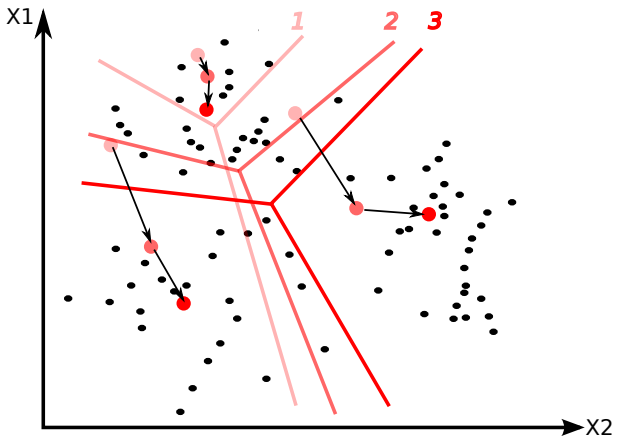

FIGURE *2.13 – Trajectoire des centroïdes lors d'un regroupement de type K-Moyennes appliqué à des données en deux dimensions. Le découpage des différentes classes à chaque étape est également représenté. Dans ce cas, la convergence vers la solution optimale est obtenue en 3 étapes*

2.2.4 Regroupement spectral

Le regroupement spectral consiste à créer, à partir d'une matrice de similarité (appelée matrice affinité) des données à regrouper, un espace de dimension réduite dans lequel les données seront regroupées à l'aide de l'algorithme des K-Moyennes. Le but d'un regroupement spectral est donc de réaliser un regroupement dans un espace différent de celui des données d'origine (espace spectral). En considérant K le nombre de groupes à trouver, l'algorithme du regroupement spectral est l'algorithme et illustré par la Figure 2.14.

Algorithme 3 Regroupement spectral

- Construction et normalisation de la matrice affinité
- Extraction des K plus grands vecteurs propres (associés aux K plus grandes valeurs propres)
- Normalisation des lignes de la matrice des K vecteurs propres
- Regroupement dans l'espace spectral par l'algorithme des K-Moyennes
- Regroupement dans l'espace d'origine directement obtenu du regroupement dans l'espace spectral (via une relation d'équivalence)

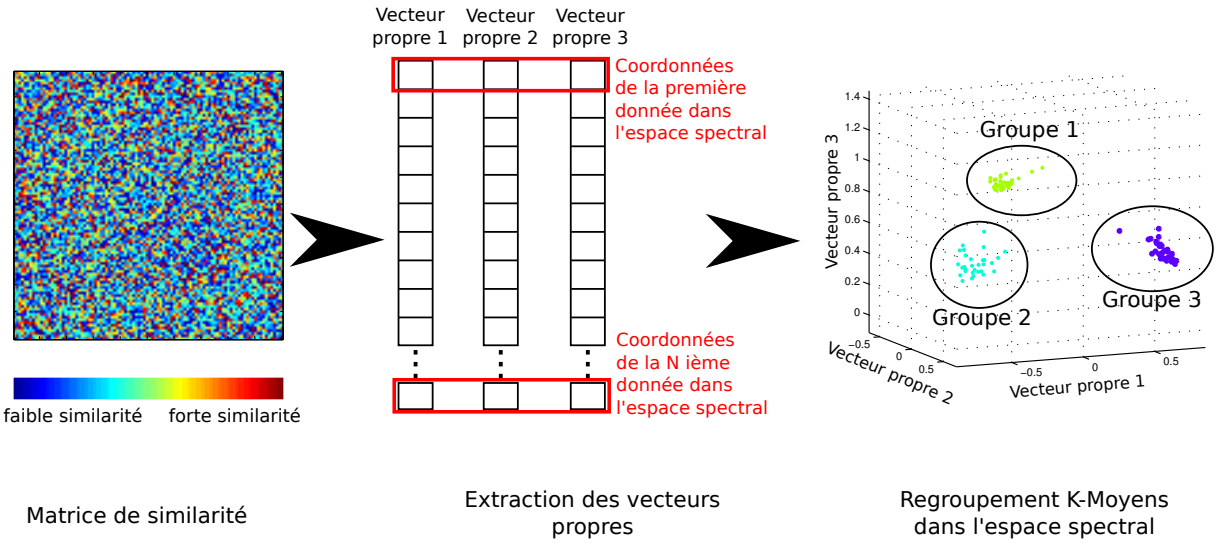

FIGURE 2.14 – *Exemple de regroupement spectral pour $K = 3$. A partir d'une matrice de similarité, les K principaux vecteurs propres sont extraits pour représenter un nouvel espace (l'espace spectral) dans lequel un regroupement de type K-Moyennes est effectué*

2.2.5 Détermination automatique du nombre de groupes

Une des limitations des méthodes des K-moyennes, du regroupement agglomératif ou du regroupement spectral est que le nombre de groupes K doit être connu à priori. Mouysset *et al.* [Mouysset 2011] proposent une solution pour déterminer automatiquement le nombre de groupes. Cette méthode est illustrée par la Figure 2.15. A partir d'une matrice de similarité, plusieurs regroupements sont effectués en faisant varier la valeur de K. Ils proposent de générer pour chaque valeur de K une matrice de similarité indexée par classe où les blocs diagonaux représentent la similarité entre les données intra-groupes, et les blocs hors-diagonaux représentent la similarité entre les données inter-groupes. L'idée est de conserver la valeur de K qui maximise la similarité moyenne intra-groupe et minimise la similarité moyenne inter-groupes.

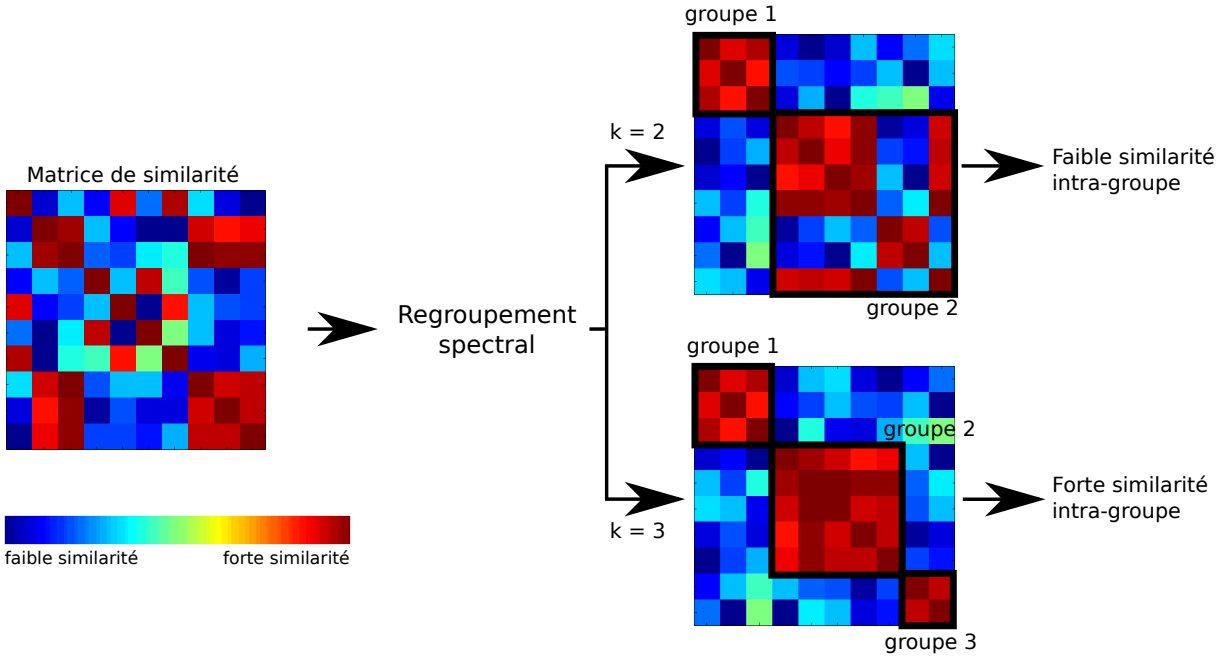

FIGURE *2.15 – Détermination automatique du nombre de groupes. Après avoir effectué plusieurs regroupements pour différentes valeurs de K, le regroupement retenu est celui dont la similarité moyenne intra-groupe est la plus grande et la similarité inter-groupes la plus faible.*

2.3 Travaux existant sur la segmentation en scènes

Dans cette section seront étudiées les différentes méthodes existantes de segmentation en scènes. Les travaux présentés ici diffèrent selon trois critères importants : la méthode employée, les descripteurs utilisés et le type de document audiovisuels pour lesquels ils ont été développés. Cette thèse traite des méthodes de structuration pour des documents audiovisuels de type narratif d'un domaine bien particulier : les séries télévisées. Certaines

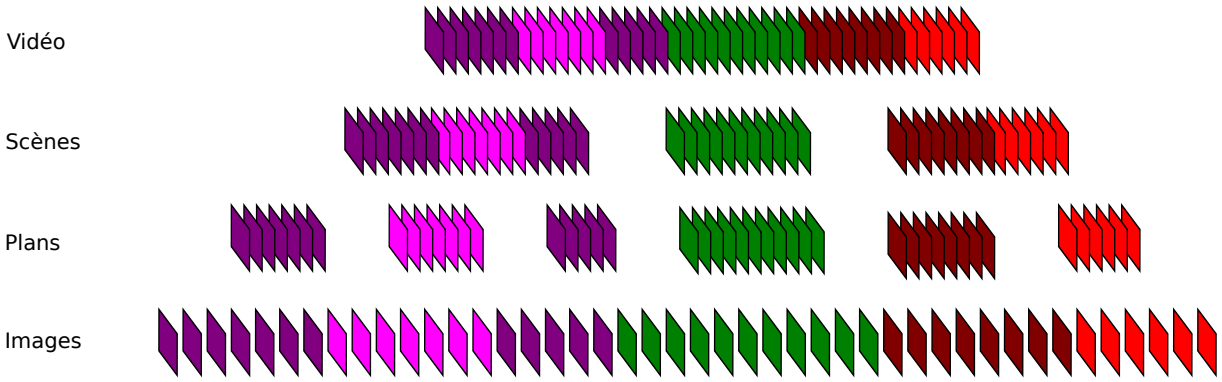

FIGURE 2.16 – *Différents niveaux de structuration d'une vidéo*

méthodes présentées dans cet état de l'art ont été développées pour d'autres types de documents, tel que des films, des talk-shows ou des vidéos personnelles.

La Figure 2.16 montre la structure hiérarchique typique d'une vidéo, présentant la vidéo comme une suite de scènes qui sont elles-mêmes composées d'une suite de plans. La définition d'une scène et de la tâche de segmentation en scènes a déjà été discutée dans la Section 1.2. Il a été exposé qu'une scène est un concept subjectif, et qu'il existe de nombreuses définitions de ce concept dans les travaux sur la segmentation en scènes pour des documents audiovisuels. Cependant, ces travaux considèrent tous une scène comme une suite de plans ou de courts segments d'un document audiovisuel liés par un même sujet.

Un état de l'art sur la segmentation en scènes a été publié récemment par Fabro *et al.* [Del Fabro 2013]. Ils proposent de classer les différentes méthodes de segmentation en 7 catégories qui utilisent 3 descripteurs bas niveau (visuel, audio et textuel) et leurs combinaisons comme illustré dans la Figure 2.17).

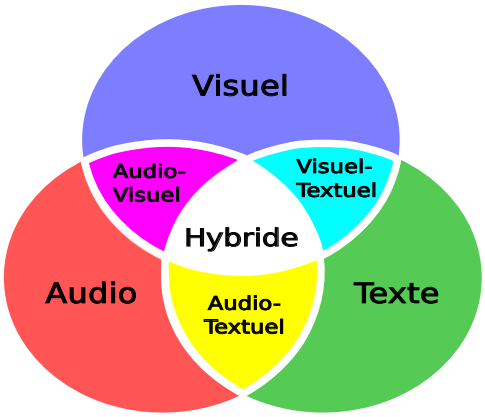

FIGURE 2.17 – *Catégories de segmentation en scènes basées sur des descripteurs bas-niveau [Del Fabro 2013]*

Comme cela est décrit dans la Section 2.1, il existe de nombreuses façons de décrire un segment de document audiovisuel. De ce fait, certains travaux se basent uniquement sur l'analyse des images pour retrouver les scènes, d'autres, plus rares, n'utilisent que la composante audio, ou même uniquement le texte à partir de métadonnées. Certains travaux proposent une analyse combinant les informations provenant de plusieurs modalités disponibles dans les documents audiovisuels.

Dans cet état de l'art, les travaux sur la segmentation en scènes sont classés suivant trois grandes familles d'approches. L'approche basée sur la détection de frontières qui consiste à rechercher les coupures dans le flux vidéo. L'approche agglomérative (regroupement de séquences) qui consiste à retrouver les scènes en regroupant les plans ou les séquences de vidéos similaires. Enfin, les approches hybrides.

2.3.1 Détection de frontières

Une des premières méthodes pour la segmentation en scènes est proposée par Kender et Yeo [Kender 1998]. Ils utilisent la notion de cohérence entre les plans pour détecter les transitions entre les scènes. Une transition de scène est présente si le plan courant est incapable de "rappeler" au spectateur ce qui a pu être vu plus tôt dans la vidéo. Ils mesurent la similarité entre deux plans comme étant la similarité maximale entre toutes les images des deux plans, cette similarité étant calculée à partir d'histogrammes de couleur. À chaque transition de plan, une fonction estime à quel point les plans suivants sont cohérents avec les plans précédents en fonction de leur similarité visuelle telle qu'illustrée par la figure 2.18. Les minima locaux de cette fonction sont alors considérés comme des transitions de scènes.

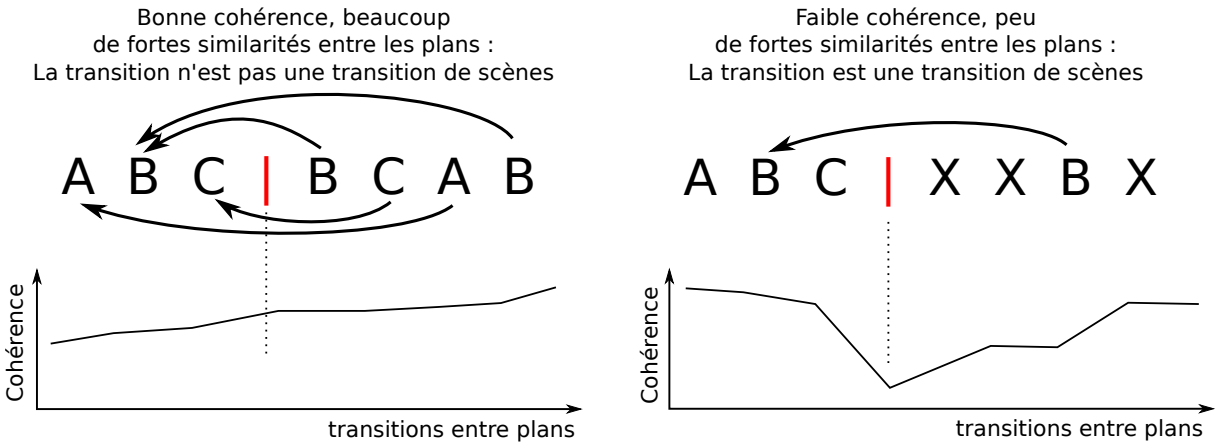

FIGURE *2.18 – Mesure de cohérence pour une transition de scène donnée. Les transitions de scènes sont retrouvées en recherchant les minimums locaux dans la fonction de cohérence.*

Liu *et al.* [Liu 1998] proposent une méthode de détection des frontières de scènes uniquement à partir de l'analyse de la bande-son. Ils commencent par classifier de petits clips

audio d'une seconde en cinq classes de programmes (publicité, match de basket, match de football, reportage, météo). Ils comparent 12 descripteurs audio bas-niveau pour faire cette classification à partir d'un réseau de neurones. Les scènes sont retrouvées en évaluant les changements entre les descripteurs de clips adjacents. Un clip est considéré comme transition de scène s'il est similaire aux clips qui le suivent directement, et différent des clips qui le précèdent. Cependant, cette méthode est contraignante puisqu'elle nécessite une étape d'apprentissage, et elle ne s'applique qu'à des vidéos du type *émissions sportives*.

Plusieurs méthodes sont basées sur des règles issues de l'analyse des techniques de montage du cinéma (aussi appelées grammaire des films) pour découper le flux vidéo en scènes. Adams *et al.* [Adams 2002] utilisent le principe du « tempo » ou « rythme » pour décrire les scènes. Le tempo correspond à la perception du rythme du film par le spectateur. Il dépend de la vitesse d'enchaînement des plans et de la quantité de mouvement que l'on peut observer dans une séquence vidéo. Dans leur article, ces auteurs [Adams 2002] considèrent que les réalisateurs de films utilisent souvent des tempos différents pour des scènes consécutives, et plus particulièrement qu'il n'y a jamais deux scènes avec un tempo très rapide qui se suivent pour ne pas perturber le spectateur. Ils déterminent une fonction décrivant le tempo en fonction du temps, et ils recherchent dans cette fonction les changements de tempo. Ils considèrent qu'un grand changement de tempo correspond à un changement de scène, alors qu'un petit changement indique un évènement à l'intérieur d'une scène.

Le tempo est aussi utilisé par Cheng et Lu [Cheng 2008] pour réduire la sur-segmentation produite par certaines méthodes de segmentation en scènes. En partant d'une segmentation en scènes existante, ils mesurent le tempo de chaque scène. Si un grand changement de tempo est observé entre deux scènes, cela indique que le contenu a changé entre les scènes et que la transition est correcte. Au contraire, si le rythme ne change pas beaucoup, alors deux cas sont possibles : si le tempo est rapide, le système fait confiance au tempo et les scènes sont fusionnées, s'il est lent, le système fait confiance à l'information visuelle et la transition est conservée.

Petersohn *et al.* [Petersohn 2009] utilisent aussi un modèle basé sur des techniques de montage cinématographiques pour retrouver les scènes. Ils recherchent le type de transition entre les plans (transitions franches ou transitions progressives). Dans un film, un changement de scène est souvent marqué par une transition progressive. Ils utilisent cette information pour déterminer la position des transitions de scènes.

Sundaram et Chang [Sundaram 2002] proposent de combiner des descripteurs audio et vidéo pour la segmentation en scènes. Ils déterminent des scènes visuelles suivant les mêmes principes de cohérences des plans que proposés par Kender et Yeo [Kender 1998]. Les scènes audio sont déterminées à partir d'un grand nombre de descripteurs audio bas niveau (taux de passage à zéro, flux spectral, énergie, MFCC, etc...). Les transitions de scènes finales sont détectées en regroupant les transitions des scènes visuelles proches des transitions des scènes audio (cf. Figure 2.19). Ils ajoutent un ensemble de règles permettant

de détecter les segments de dialogues à partir des descripteurs visuels, et les silences à partir des descripteurs audio. Ils considèrent que si un silence coïncide avec une transition visuelle, alors il y a changement de scène.

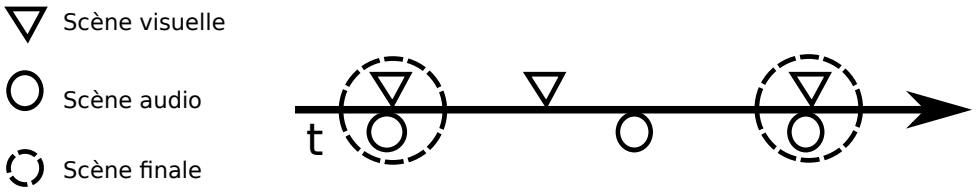

FIGURE *2.19 – Synchronisation de l'audio et de la vidéo pour la recherche des transitions entre scènes d'après Sundaram* et al. *[Sundaram 2002]. Les triangles sont des transitions visuelles. Les cercles pleins des transitions audio. Une transition de scène est détectée (cercles en pointillés) si une transition audio et une transition visuelle sont proches.*

Cour *et al.* [Cour 2008] proposent une méthode basée sur l'analyse de métadonnées textuelles. Ils ont comme hypothèse que les sous-titres et le script sont facilement disponibles pour des vidéos produites de manière professionnelle (comme les films ou les séries télévisées). Dans le script sont présentes des informations sur la structure de la vidéo, sur ce qui est dit et le personnage qui le prononce. À partir des sous-titres, ils récupèrent l'information sur ce qui est dit et à quel moment. Ils proposent une méthode d'alignement des sous-titres avec le script pour retrouver les temps de début et de fin des scènes.

Les méthodes suivant des approches basées sur une détection de frontières sont résumées dans le Tableau 2.1.

	Type de données	Taille corpus	Descripteurs	Evaluation
[Kender 1998]	Série TV	1 épisode/23 min 18 scènes (≈ 90sec / scène)	Couleur	Nombre de scènes correctes (11/18)
[Liu 1998]	Télévision	10 séquences 44 scènes	Audio	transitions : 42/44 correctes 36 fausses alarmes
[Adams 2002]	Film	4 Films/ 130 scènes	Mouvement Durée d'un plan	transitions : 122/130 correctes 8 fausses alarmes
[Sundaram 2002]	Film	1 épisode/3h 158 scènes (≈ 83sec / scène)	Couleur, Audio	Rappel (91%) Précision (100%)
[Cheng 2008]	Film	5 Films/86 scènes	Mouvement Durée d'un plan	Description du résultat
[Cour 2008]	Série TV	2 vidéos/≈40min	Texte	F-mesure (86% et 75%)
[Petersohn 2009]	Film, Série Talk-Show	7h 47min de vidéo 562 scènes (≈ 50sec / scène)	Couleur	Rappel (80.3%) Précision (83.4%) F-Mesure (81.8%)

TABLE *2.1 – Résumé des méthodes de segmentation en scènes pour les approches basées sur une détection de frontières. Les méthodes de calcul des résultats de l'évaluation sont discutées dans la Section 2.3.4.*

2.3.2 Regroupement de séquences

Cette section présente les différentes méthodes de segmentation en scènes basées sur le principe du regroupement de séquences de vidéos. Une scène étant une suite de plans traitant d'un même sujet, ces méthodes considèrent que l'on peut retrouver les scènes en regroupant entre eux les plans similaires.

Rui *et al.* [Rui 1999] proposent de retrouver les scènes en effectuant un regroupement des plans basé sur une similarité des histogrammes de couleurs de leurs images clefs, et de descripteurs de mouvement. Ils considèrent que deux plans d'une même scène doivent avoir une cohérence visuelle et une intensité de mouvement similaires. Cependant, regrouper les plans en se basant uniquement sur la similarité visuelle sans différencier le contexte n'est pas suffisant. Il peut arriver que l'on regroupe deux plans similaires visuellement, mais

appartenant à deux scènes différentes (par exemple, plusieurs scènes peuvent se dérouler dans la même pièce ou plusieurs plans montrent les mêmes personnes, portant les mêmes vêtements mais filmés dans des endroits différents). Pour résoudre ce problème, ils utilisent une méthode pour regrouper les plans appelée « time-adaptative grouping » qui permet de ne pas regrouper des plans trop éloignés temporellement, même s'ils sont très similaires visuellement.

Hanjalic *et al.* [Hanjalic 1999] proposent d'extraire une ou plusieurs images clefs pour chaque plan, et de mesurer la similarité entre deux plans utilisant la méthode des blocs de couleur présentée Section 2.1.3. Ils proposent de lier les plans dont la similarité dépasse un certain seuil. Si un groupe de plans est chevauché par au moins un lien, alors ils sont regroupés en une même scène. La Figure 2.20 montre le fonctionnement des liens qui se chevauchent et des scènes (ici appelées LSU) qui en résultent. Ce principe a été repris dans beaucoup d'autres travaux sur la segmentation en scènes. Kwon *et al.* [Kwon 2000] ont amélioré cet algorithme en utilisant des descripteurs de mouvement pour le calcul des similarités entre les plans. Zhao *et al.* [Zhao 2001] et Wengang *et al.* [Wengang 2003] utilisent le même principe mais en ajoutant une contrainte temporelle sous la forme d'une fenêtre glissante, interdisant de regrouper ensemble deux plans trop éloignés temporellement. Cette méthode a été améliorée par Mitrovic *et al.* [Mitrović 2010] qui utilisent des descripteurs de contours et des points d'intérêt pour décrire les images en plus des histogrammes de couleur. Leur méthode a été optimisée à partir de documentaires d'archives, mais ils concluent qu'elle donne des résultats similaires à la méthode proposée par Sundaram *et al.* [Sundaram 2002] en l'appliquant à des films.

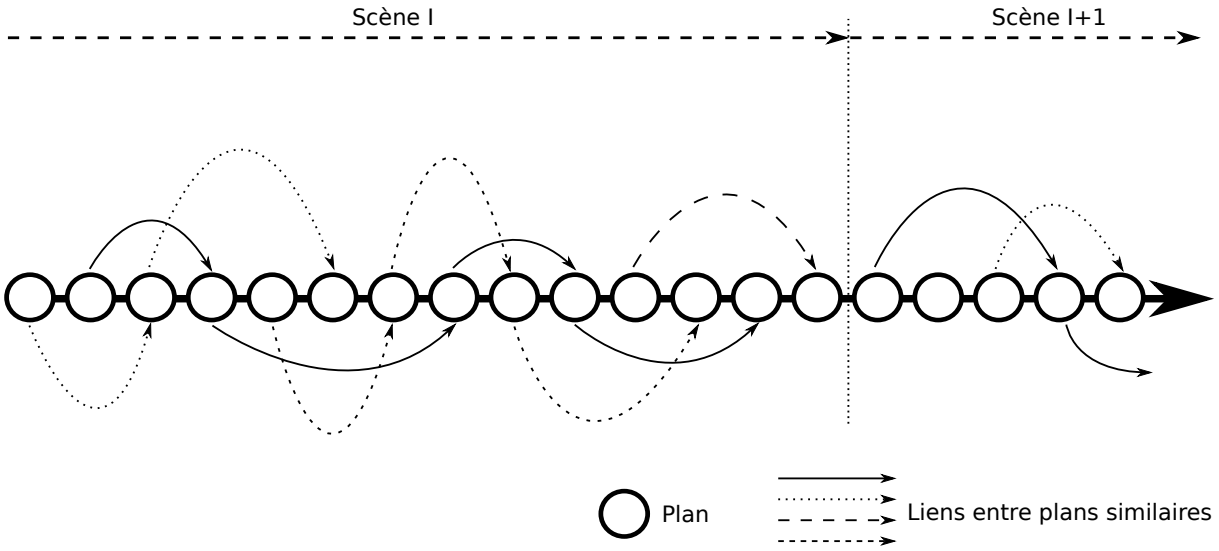

FIGURE *2.20 – Illustration de la méthode de segmentation proposée par Hanjalic* et al. *[Hanjalic 1999]. Tous les plans chevauchés par au moins un lien sont regroupés dans une même scène*

Une méthode similaire est proposée par Lu *et al.* [Lu 2006]. Elle n'utilise que la composante audio et elle est inspirée des méthodes de segmentation utilisant des descripteurs visuels. En se basant sur des descripteurs audio bas-niveau, le flux audio est divisé en segments audio (comparables à des « plans audio »). Comme pour les méthodes basées sur la couleur, les segments audio similaires sont regroupés s'ils sont suffisamment proches temporellement. Ces groupes de segments audio sont considérés comme des scènes auditives.

Pour Tavanapong *et al.* [Tavanapong 2004], la définition d'une scène et la méthode de regroupement des plans en scènes sont directement issues de l'analyse de techniques cinématographiques, telles que celles présentées Section 1.2 (loi des 180°, champ/contre-champ, vue d'ensemble/détail/vue d'ensemble). Ils proposent de décrire les images en se concentrant sur 5 zones précises : les 4 coins de l'image et la partie supérieures qui représentent le fond du décor. Ces 5 zones sont décrites par des descripteurs de couleurs, et leur algorithme regroupe les plans qui ont des zones similaires. Ils utilisent ensuite une méthode proche de celle proposée par Hanjalic *et al.* [Hanjalic 1999] pour regrouper les groupes de plans qui se chevauchent.

Odobez *et al.* [Odobez 2003] proposent de regrouper les plans en se basant sur des histogrammes de couleur RGB et leur proximité temporelle, en utilisant un regroupement spectral. Leur approche est testée sur des vidéos personnelles, sans montage, et composées de seulement quelques plans.

Rasheed et Shah [Rasheed 2005] proposent une méthode de segmentation basée sur des méthodes d'analyse de graphes. Ils construisent un graphe complet non orienté appelé « Graphe de similarité des plans » (Shot Similarity Graph ou SSG), tel que chaque noeud est un plan, et les liens entre les noeuds sont pondérés par la similarité entre les plans. La similarité est mesurée à partir de descripteurs de couleur, de mouvement et par la distance temporelle entre les plans. L'algorithme du « normalized cut » [Shi 2000] est utilisé pour partitionner le graphe en sous-graphes, où chaque sous-graphe correspond à une scène.

L'un des principaux problèmes de toutes les méthodes présentées jusqu'à présent est qu'elles sont dépendantes de divers seuils, notamment en ce qui concerne la distance temporelle entre les plans.

Les méthodes suivant des approches agglomératives sont résumées dans le Tableau 2.2.

	Type de données	**Taille corpus**	**Descripteurs**	**Evaluation**
[Rui 1999]	Film	8 Films/100 min 82 scènes (≈ 72 sec / scène)	Couleur Mouvement	transitions : 78/82 correctes 16 fausses alarmes
[Hanjalic 1999]	Film	2 Films/160 min 78 scènes (≈ 123 sec / scène)	Couleur	Description des résultats
[Kwon 2000]	Film	1 Film/≈ 1h 11 scènes (≈ 5 min / scène)	Couleur Mouvement	Description des résultats
[Wengang 2003]	Dessin animé Documentaire	4 vidéos / 59 scènes	Couleur	70/59 scènes détectées
[Odobez 2003]	Vidéos Personnelles	20 vidéos/≈400 min	Couleur	Nombre de plans mal placés (27,1%)
[Tavanapong 2004]	Film	2 Films/200 min 120 scènes (≈ 100 sec / scène)	Couleur	Rappel (51,6%) Précision (15,6%)
[Rasheed 2005]	Film, Série	5 vidéos/277 min 169 scènes (≈ 98 sec / scène)	Couleur Mouvement	Rappel (74 - 82%) Précision (50 - 85%)
[Lu 2006]	Talk-Show	1 vidéo/90 min 156 scènes (≈ 35 sec / scène)	Audio	transitions : 108/169 correctes 12 fausses alarmes
[Zhai 2006]	Film, Vidéo personnelle	7 vidéos/87 scènes 346 min (≈ 4 min / scène)	Couleur Durée d'un plan	Rappel (91.3%) Précision (84%)
[Zhao 2007]	Film, Série	4 vidéos/207 scènes	Couleur	Rappel (71%) Précision (78,2%) F-mesure (74,4%)
[Mitrović 2010]	Film Documentaire	3 vidéos/263 min 83 scènes (≈ 3 min / scène)	Couleur Texture SIFT	Rappel (75 - 93%) Précision (45 - 53%)

TABLE *2.2 – Résumé des méthodes de segmentation en scènes pour les approches basées sur le regroupement de séquences. Les méthodes de calcul des résultats de l'évaluation sont discutées dans la Section 2.3.4.*

2.3.3 Approches hybrides

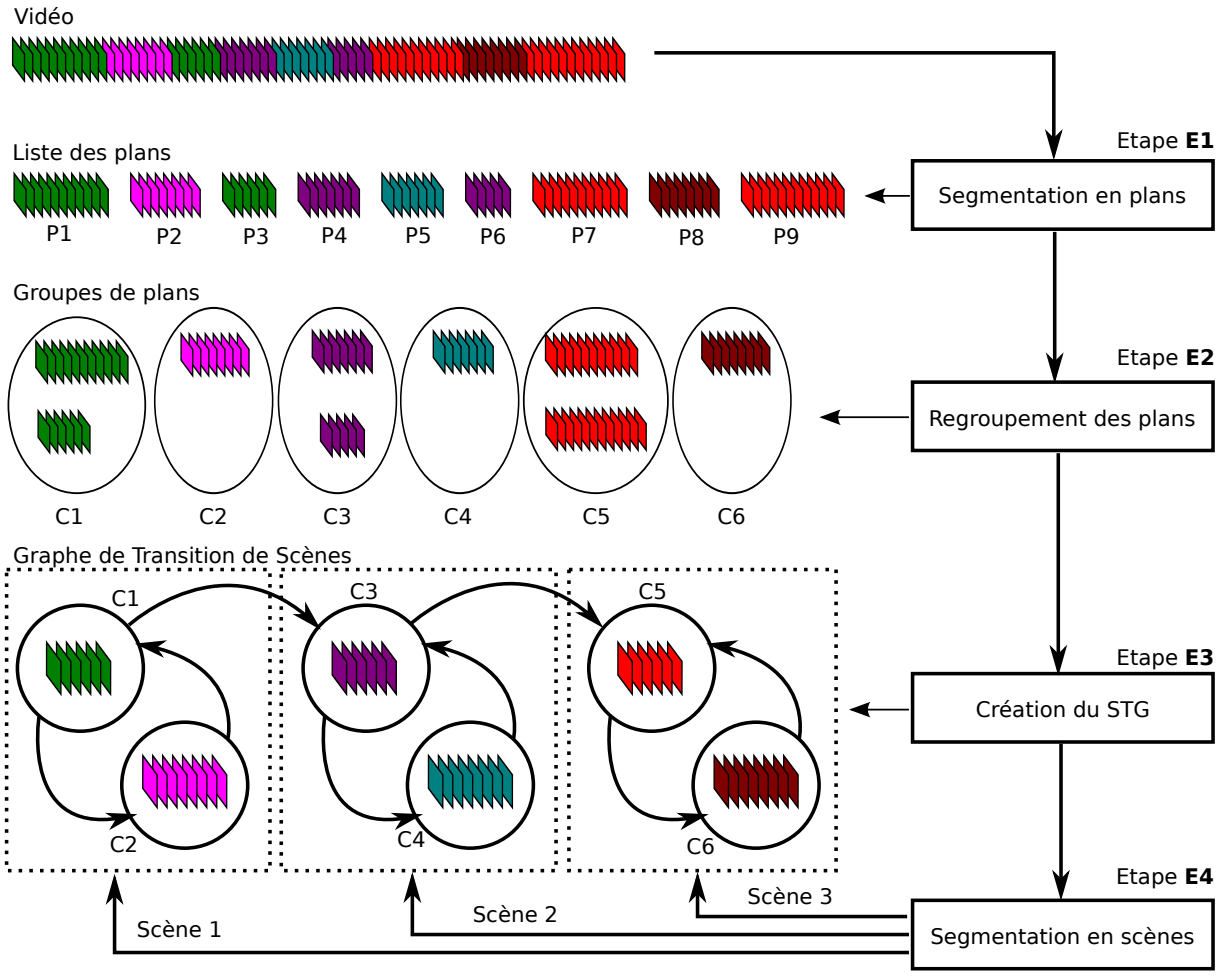

FIGURE *2.21 – Illustration de la méthode de segmentation par graphes proposée par Yeung* et al. *[Yeung 1998]*

Dans cette section, nous nous intéressons aux méthodes de segmentation en scènes basées sur une approche hybride impliquent un regroupement de séquences et une recherche de frontières.

Le premier algorithme suivant une approche hybride pour la segmentation en scènes a été proposé par Yeung *et al.* [Yeung 1998] qui utilisent la notion de Scene Transition Graph (STG) dont le principe est résumé dans la Figure 2.21. La construction du STG se déroule en 4 étapes notées **E1** à **E4** sur la figure. En partant de la vidéo, l'algorithme commence par une étape de segmentation de la vidéo en plans (étape **E1**). L'étape suivante consiste à regrouper les plans similaires visuellement (étape **E2**). Pour ce faire, un regroupement hiérarchique des plans du type complete-link est effectué en considérant qu'il est impossible de regrouper deux plans si leur distance temporelle est plus grande qu'un seuil Δt. Le regroupement se termine lorsque la distance entre deux groupes de plans est inférieure

à une mesure de similarité Δd. La mesure de similarité entre les plans est effectuée par comparaison d'histogrammes de couleur sur une image caractéristique du plan. L'étape suivante (étape **E3** de la Figure 2.21) est l'étape de création du STG. Les noeuds du STG sont les groupes de plans visuellement similaires issus de l'étape précédente, et les arcs représentent le flot temporel de l'histoire (temporal story flow). Un arc est présent entre deux noeuds seulement si un plan représenté par le premier noeud précède directement un plan représenté par le second. Les arcs de coupures (cut edges) sont recherchés dans ce STG pour le découper en plusieurs sous-graphes disjoints. Un arc du graphe est considéré comme un arc de coupure si la suppression de cet arc donne deux graphes disjoints. On considère alors que chaque sous-graphe est une scène (étape **E4**).

Le STG permet de représenter le flot temporel de l'histoire racontée dans une vidéo. La Figure 2.22 montre quelques exemples de STG que l'on peut obtenir pour différents types de vidéos. Dans ces exemples, chaque sommet du graphe représente un groupe de plans, et la lettre associée à chaque groupe correspond à un indicatif de scène annotée. En (a), on observe une histoire linéaire. En (b), l'exemple illustre deux évènements représentés par les noeuds B_1 à B_3 pour l'un et D_1 à D_3 pour l'autre. En (c), l'arrangement des groupes de plans (noeuds du graphe) montre des allers-retours entre des plans visuellement similaires. Ce type de structure illustre un enchaînement de plans représentant un champ/contrechamp (comme un dialogue). En (d), on observe la description de plusieurs évènements centrés sur un même groupe de plan (noeud A).

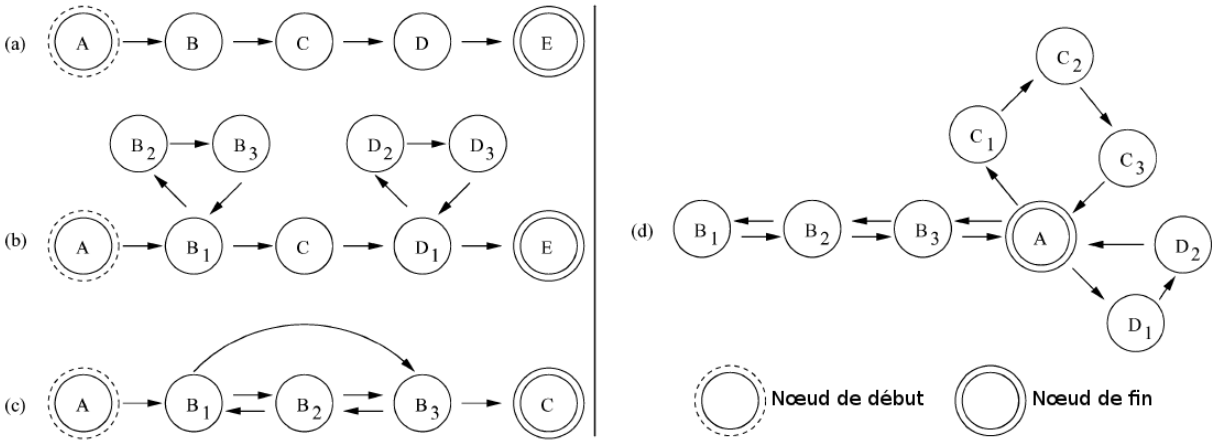

FIGURE *2.22 – Exemples de STG d'après Ngo* et al. *[Ngo 2003]*

La méthode proposée par Yeung *et al.* pour découper le graphe consiste à rechercher et supprimer les arcs de coupures du graphe. Cette méthode permet d'isoler les scènes pour les exemples (a), (b) et (c) de la Figure 2.22, mais pas pour l'exemple (d). La seule façon de résoudre ce problème est de reconstruire un STG en modifiant le seuil Δt, de manière à réduire la taille des groupes de plans similaires et donc augmenter ce nombre de groupes pour que le groupe A ne soit plus central. Ngo *et al.* [Ngo 2003] ont

repris le principe du STG, mais pour résoudre ce problème et être moins dépendants de ce seuil Δt, ils proposent d'utiliser l'algorithme « normalized cut » [Shi 2000] pour le partitionnement du graphe, et ainsi permettre d'isoler des groupes de noeuds qui ne sont pas connectés au reste du graphe uniquement par des cut-edges. Benini *et al.* [Benini 2005] ont aussi utilisé les STG pour découper des films et un journal télévisé, mais la méthode de regroupement des plans se base sur un processus de quantification vectorielle dans l'espace de couleur LUV [Benini 2006]. Ils proposent une modélisation par Modèles de Markov pour classer les scènes en trois catégories : dialogue, déplacement et hybride (déplacement + dialogue) [Benini 2008].

Sidiropoulos *et al.* [Sidiropoulos 2009] proposent deux méthodes pour améliorer l'algorithme du STG en fusionnant des descripteurs audio et vidéo. La première méthode implique l'utilisation d'une méthode de segmentation et regroupement en locuteurs de la bande audio. Après avoir réalisé un STG basé sur des descripteurs visuels, ils considèrent que si un même locuteur est détecté dans deux noeuds connectés, alors il faut fusionner ces deux noeuds. Ils proposent une deuxième approche qui est précisée dans leurs travaux plus récents [Sidiropoulos 2011] appelée « Generalized Scene Transition Graph (GSTG) ». La génération d'un STG dépend de la valeur donnée aux seuils Δt et Δd. Ils proposent de générer un grand nombre de STG à partir de différentes méthodes de calcul de similarités des plans (visuelles, audio ou basées sur la détection de concepts), et en faisant varier aléatoirement les seuils Δt et Δd. Chaque paire (Δt, Δd) associée à une méthode de calcul des similarités donne une segmentation en scènes différente. Le GSTG consiste à associer à chaque transition de plan le ratio de STG qui déterminent que cette transition est une transition de scène (cf. Figure 2.23). Les transitions finales sont déterminées en optimisant un seuil θ tel que si le ratio de STG considérant qu'une transition de plan est une transition de scène est supérieur à θ, alors cette transition de plan est une transition de scènes (lignes vertes sur la Figure 2.23).

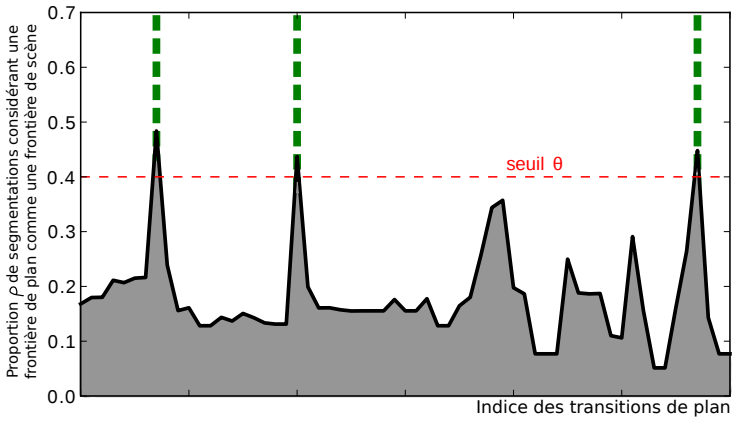

FIGURE 2.23 – *Probabilité de transition de scène d'après Bredin [Bredin 2012].*

Bredin [Bredin 2012] propose une version étendue du GSTG en utilisant un descripteur basé sur la transcription automatique de la parole. Il propose d'utiliser une grille dense de paramètres pour Δt et Δd, plutôt que de sélectionner des valeurs aléatoires de manière à rendre le GSTG déterministe.

D'après Fabro *et al.* [Del Fabro 2013], la segmentation en scènes utilisant des graphes fonctionne mieux pour des vidéos proposant des environnements réduits, et particulièrement des vidéos avec des scènes similaires qui se répètent comme pour les journaux télévisés ou les talk-shows. La précision est moins bonne pour des vidéos montrant beaucoup de mouvements. Les films proposent un environnement très dynamique, et les réalisateurs utilisent des techniques de caméra et différents effets visuels pour attirer l'attention du spectateur. Ils considèrent qu'il est plus difficile de modéliser un graphe de scène pour ce type de vidéo.

Cependant, beaucoup de méthodes présentées dans cette section ont été développées pour des films, et il n'existe pas à notre connaissance de comparaison formelle des différentes méthodes de segmentation en scènes qui pourraient étayer cette affirmation. De plus, les évaluations des différentes méthodes de segmentation sont difficilement comparables. La section suivante discute justement des différentes méthodes d'évaluation utilisées dans cet état de l'art.

Les méthodes suivant des approches hybrides sont résumées dans le Tableau 2.3.

	Type de données	**Taille corpus**	**Descripteurs**	**Evaluation**
[Yeung 1998]	Série TV	1 épisode/20 min 10 scènes (≈ 120 sec / scène)	Couleur	Description des résultats
[Ngo 2003]	Vidéos personnelles	5 vidéos/137 min 157 scènes (≈ 52 sec / scène)	Couleur Texture	Rappel (90%) Précision (87%)
[Benini 2005]	Film	4 vidéos/107 min 60 scènes (≈ 107 sec / scène)	Couleur	Couverture (85,2%) Overflow (2,2%)
[Sidiropoulos 2011]	Film Documentaire	24 vidéos/939 scènes	Couleur Concepts Audio	Couverture (88,6%) Overflow (13,2%)
[Bredin 2012]	Série	8 épisodes/5 h 306 scènes (≈ 59 sec / scène)	Couleur Audio Texte	Rappel (48,8%) Précision (62,2%) F-mesure (53,9%)

TABLE *2.3 – Résumé des méthodes de segmentation en scènes pour les approches hybrides*

2.3.4 Évaluation de la segmentation automatique des scènes

Comparer les différentes méthodes de segmentation en scènes est très compliqué pour plusieurs raisons : différence de définition d'une scène, différence des données utilisées et différence des méthodes d'évaluation. La plupart des travaux évaluent leurs résultats par une comparaison de la segmentation du système développé avec une segmentation manuelle des scènes. Or, puisqu'il n'existe pas de base de données d'annotations disponibles publiquement, ces travaux se basent sur des annotations personnelles impossibles à reproduire.

Toutefois, certains travaux proposent une évaluation utilisant des données publiques. Ainsi, Rasheed et Shah [Rasheed 2005] utilisent les chapitres inclus dans les DVD des films comme base de l'évaluation de leur méthode de segmentation. Cependant, il n'est pas acquis que ces chapitres suivent toujours une définition proche du concept de scène. C'est pourquoi ils ne sont pas utilisés dans nos travaux.

Dans les premiers travaux sur la segmentation en scènes, l'évaluation était subjective et laissée à l'appréciation du lecteur. Ainsi, Yeung *et al.* [Yeung 1998] évaluent leur méthode en décrivant les scènes obtenues sans proposer de mesure numérique de la qualité de la segmentation.

De nombreux travaux utilisent des mesures d'évaluation à partir de la comparaison de leur méthode de segmentation avec une annotation manuelle des scènes. Dans les travaux les plus anciens, les résultats étaient souvent présentés en comptant le nombre de transitions correctement détectées, le nombre de fausses alarmes et le nombre de transitions manquées.

Précision / Rappel

Actuellement, les mesures les plus utilisées sont la ***précision*** (P) (équation 2.10) et le ***rappel*** (R) (équation 2.11). Précision et Rappel dépendent de trois valeurs :

- VP le nombre de *Vrais Positifs*, c'est à dire le nombre de transitions détectées et considérées comme correctes
- FP le nombre de *Faux Positifs*, c'est à dire le nombre de transitions détectées et incorrectes
- FN le nombre de *Faux Négatifs*, c'est le nombre de transitions de la vérité terrain qui n'ont pas été détectées comme transition de scène.

Leur calcul est effectué suivant les formules suivantes :

$$\mathrm{P} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \qquad (2.10) \qquad\qquad \mathrm{R} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \qquad (2.11)$$

Une précision élevée indique que peu de faux positifs sont présents dans les transitions détectées. Le rappel au contraire est élevé lorsqu'un nombre élevé de transitions présentes

dans la vérité-terrain sont détectées. Rappel et précision donnent une information complémentaire sur la qualité de la segmentation. Ainsi, une moyenne harmonique de ces deux valeurs peut être calculée pour les prendre en compte en une seule mesure : la **F-mesure** (F_{PR}) (2.12).

$$F_{PR} = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.12}$$

Précision, Rappel et F-mesure se basent sur le principe de « transitions détectées et considérées comme correctes ». Évaluer une méthode de segmentation en scènes par une mesure de précision, rappel et F-Mesure revient à considérer la tâche de segmentation en scènes comme une tâche de détection ou de classification des frontières. Le problème est que dans le domaine de la segmentation en scènes, une scène peut être considérée comme partiellement correctement détectée. Par exemple, lorsqu'une transition détectée débute quelques secondes après le début de la scène annotée et la suivante termine quelques secondes avant la fin, la scène peut ne pas être considérée comme complètement fausse. Plusieurs travaux ont apporté leur solution à ce problème. Ainsi, Hanjalic *et al.* [Hanjalic 1999] considèrent qu'une transition de scène est correctement détectée si elle se trouve à moins de trois plans d'une transition de scène annotée. Pour Rasheed et Shah [Rasheed 2005], toute transition détectée dans un intervalle de 10 secondes avant et après chaque transition annotée est considérée correcte. Ces différences pour un même type d'évaluation rendent la comparaison des résultats exposés dans les différents articles traitant de la segmentation en scènes encore plus difficile.

Couverture / Débordement

Vendrig *et al.* [Vendrig 2002] ont développé deux nouvelles mesures qui permettent d'exprimer les taux de sur-segmentation et de sous-segmentation. Le coverage C (couverture) mesure à quel point les images appartenant à la même scène sont correctement regroupées, alors que l'overflow O (débordement) évalue le nombre d'images qui, bien que n'appartenant pas à la même scène sont regroupées de façon erronée par le système automatique. Plus précisément, la couverture est le taux de couverture moyen des scènes de la vérité terrain alors que le débordement est le taux de recouvrement moyen des scènes de la vérité terrain, comme illustré à la Figure 2.24.

Afin d'estimer la couverture et le recouvrement d'une scène sv_i de la vérité terrain, on prend en compte les scènes détectées automatiquement $sa_j, ..., sa_{j+k}$ qui recouvrent cette scène sv_i. En notant $\#s$ la durée d'un segment de vidéo (comptée en nombre de plans), la couverture est égale au recouvrement maximal divisé par la durée totale des scènes comme suit :

$$C(sv_i) = \frac{\max(\#(sv_i \cap sa_j), \#(sv_i \cap sa_{j+1}), ..., \#(sv_i \cap sa_{j+k}))}{\#sv_i} \tag{2.13}$$

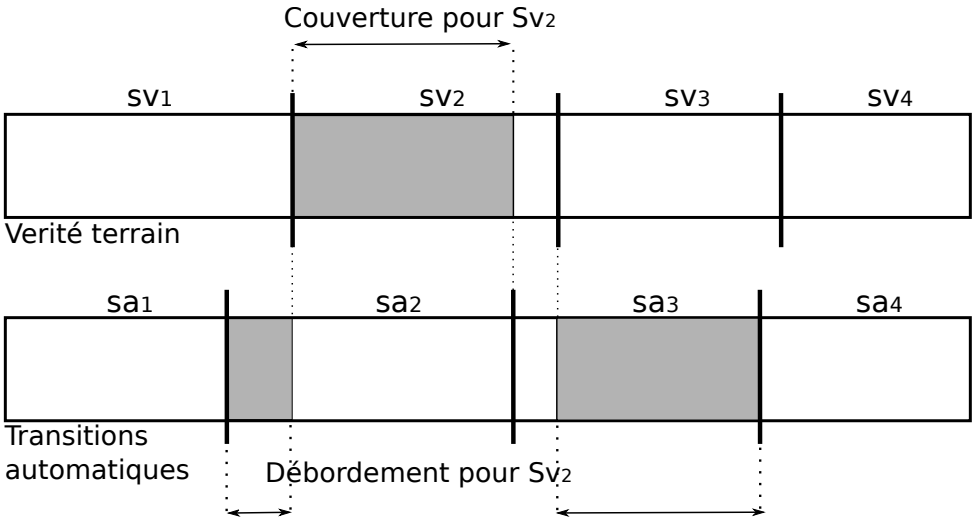

FIGURE *2.24 – Methode de calcul de la couverture et du débordement pour la segmentation en scènes.*

D'autre part, afin de calculer le taux de couverture, le débordement total de $sa_j, ..., sa_{j+k}$ avec les scènes voisines de sv_i (c'est à dire sv_{i-1} et sv_{i+1}) est estimé et divisé par la durée de ces scènes comme suit :

$$\mathrm{O}(sv_i) = \frac{\#(sv_{i-1} \cap sa_j) + \#(sv_{i+1} \cap sa_j) + ... + \#(sv_{i+1} \cap sa_{j+k})}{\#sv_{i-1} + \#sv_{i+1}} \tag{2.14}$$

Couverture et *débordement* peuvent être calculés pour une vidéo entière tel que :

$$\mathrm{C} = \frac{1}{\sum_{j=1}^{Ns} \#sv_i} \sum_{i=0}^{Ns} \#sv_i \times C(sv_i) \tag{2.15}$$

$$\mathrm{O} = \frac{1}{\sum_{j=1}^{Ns} \#sv_i} \sum_{i=0}^{Ns} \#sv_i \times O(sv_i) \tag{2.16}$$

Les valeurs optimales de la couverture et du débordement sont respectivement de 100% et 0%. Ces mesures ont été développées spécifiquement pour les méthodes de segmentation. Cependant, elles restent peu utilisées dans la littérature puisque leur appréciation est moins évidente que pour la précision et le rappel. De plus, une transition entre deux scènes qui n'est pas détectée n'a pas toujours le même impact sur le calcul de la couverture et du débordement. Ainsi, si une transition n'est pas détectée, l'erreur diffère en fonction de la durée des séquences précédent et suivant cette erreur.

Comme pour le calcul de la précision et du rappel, une moyenne harmonique de la *couverture* et du *débordement* peut être calculée pour prendre en compte ces deux mesures simultanément. Cependant, le *débordement* optimal étant de 0, la valeur de $1 - \mathrm{O}$ est utilisée à la place.

$$\mathrm{F_{CO}} = 2 \cdot \frac{\mathrm{C} \cdot (1 - \mathrm{O})}{\mathrm{C} + (1 - \mathrm{O})} \tag{2.17}$$

Differential Edit Distance

Plus récemment, Sidiropoulos *et al.* [Sidiropoulos 2012] ont proposé une nouvelle méthode : le *Differential Edit Distance* (DED) (2.18), développé pour résoudre les problèmes posés par les deux méthodes d'évaluation proposées précédemment. Ils considèrent la tâche de segmentation en scènes comme une tâche d'assignation de labels. Ainsi, si deux plans appartiennent à la même scène, ils doivent se voir attribuer le même label. Au contraire, si deux plans appartiennent à deux scènes différentes, leurs labels doivent être différents. La DED est calculée en rapport au nombre minimum de plans qui doivent changer de label pour correspondre aux labels de la vérité-terrain. Se formule peut s'écrire de la forme :

$$\mathrm{DED} = \frac{Np - N_W}{Np} \tag{2.18}$$

Avec Np le nombre total de plans de la vidéo et N_W le nombre de plans dont le label a correctement été assigné (c'est à dire le nombre de plans pour lesquels il n'était pas nécessaire de changer le label pour correspondre à la vérité-terrain). Comme il s'agit d'une distance, une segmentation parfaite donne un DED de 0.

Comparaison des approches

Ces trois méthodes d'évaluation ont des comportements très différents. La F-mesure présentée dans cette section permet d'évaluer une tâche de détection de frontières. Seules les frontières sont évaluées, et aucune information sur la durée des scènes n'est prise en compte.

Au contraire, les mesure de *couverture/débordement* et le DED évaluent une tâche de regroupement de séquences. Ces métriques évaluent à quel point deux segmentations sont similaires.

Le comportement de ces différentes méthodes d'évaluation est résumé dans le Tableau 2.4.

	Avantages	Inconvénients
$\mathbf{F_{PR}}$	• Même poids pour chaque frontière • Facilité d'implémentation	• Ne tient pas compte de la durée des scènes
$\mathbf{F_{CO}}$	• Tient compte de la durée d'un plan	• Si une transition n'est pas détectée, l'erreur diffère en fonction de la durée des séquences précédent et suivant cette erreur • Non symétrique : $\mathbf{F}_{\mathrm{CO}}(sv, sa) \neq \mathbf{F}_{\mathrm{CO}}(sa, sv)$
DED	• Tient compte du nombre de plans dans les scènes • Symétrie : $\mathrm{DED}(sv, sa) = \mathrm{DED}(sa, sv)$	• Difficulté de mise en oeuvre • Ne tiens pas compte de la durée de la scène mais du nombre de plans dans la scène

TABLE *2.4 – Avantages/inconvénients des méthodes d'évaluation de segmentation en scènes*

2.4 Regroupement de séquences traitant d'un même sujet

La Section 1.3 donne une définition d'une histoire pour des vidéos narratives, comme étant une suite de scènes traitant d'un même sujet et respectant la structure traditionnelle du récit. Une des contributions de cette thèse concerne une méthode de regroupement des scènes d'un épisode de série télévisée en histoires. Autrement dit, la méthode permet de déconstruire une vidéo qui peut contenir plusieurs histoires racontées en parallèle, et propose d'observer ces histoires séparément. La méthode proposée consiste à découper la vidéo en segments ne comportant qu'un seul évènement (une scène), et ensuite à regrouper les scènes appartenant à une même histoire.

À notre connaissance, aucune recherche n'a été menée à ce jour pour retrouver automatiquement ce type de structure dans des vidéos narratives de type séries télévisées. Cependant, certains travaux sur d'autres types de médias, comme les journaux télévisés, s'en rapprochent. Que ce soit pour des vidéos narratives ou des reportages de journaux télévisés, la difficulté de cette problématique se porte sur les méthodes permettant de **mettre en relation deux séquences de vidéos traitant d'un même sujet** mais n'étant pas forcément filmées dans un même lieu, avec les mêmes personnages et les mêmes conditions d'éclairages.

Cet état de l'art présente les méthodes permettant de regrouper des séquences de vidéos traitant d'un même sujet. Les méthodes que l'on trouve dans la littérature scientifique cherchant à résoudre cette problématique ne s'appliquent qu'aux journaux télévisés

ou écrits. Elles peuvent être séparées en deux domaines distincts : l'un s'intéresse au regroupement de séquences de journaux télévisés pour retrouver des reportages ; l'autre s'intéresse aux méthodes proposant de regrouper les reportages traitant d'un même sujet.

2.4.1 Identification des reportages dans des journaux télévisés

L'identification de reportages des journaux télévisés, aussi appelé « identification d'histoires » (*story identification*) dans la littérature, a pour but de **découper un journal télévisé en sections logiques, généralement composées de plusieurs scènes, et traitant d'un même sujet**. C'est un processus de segmentation de vidéo de haut niveau sémantique dans lequel le but est de regrouper des séquences qui peuvent être complètement différentes d'un point de vue visuel ou audio, mais liées par leur sens ou leur sujet.

Ce domaine de recherche est très actif depuis les années 2003 et 2004, poussé par la campagne d'évaluation TRECVID [Smeaton 2003] qui inclut l'identification des histoires pour les journaux télévisés dans ses tâches d'évaluation. Il est donc facile de comparer les différentes méthodes puisqu'il existe une base de données entièrement annotée.

La majorité des travaux traitant de cette problématique se fondent sur la structure fixe des journaux télévisés pour extraire les différents sujets qui y sont présents. Une idée simple est de se baser sur la détection du présentateur et de considérer qu'un reportage est présenté par celui-ci et se termine lors de sa réapparition, comme illustré à la Figure 2.25 [Goyal 2009, Ma 2009, Misra 2010].

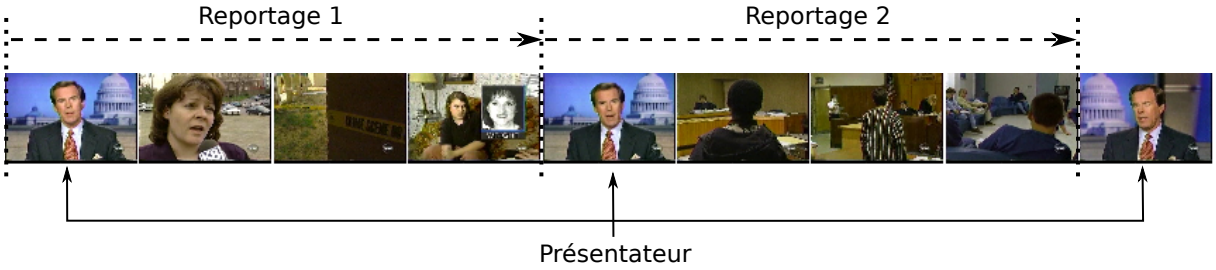

FIGURE *2.25 – Identification des histoires pour des journaux télévisés.*

Pour Chua *et al.* [Chua 2004], les méthodes basées sur des techniques d'apprentissage donnent de meilleurs résultats que celles basées sur des règles (comme celle proposée dans la Figure 2.25), avec des F-Mesures de 0.74 à 0.77 pour la méthode proposée par [Chaisorn 2003], contre moins de 0.5 pour celles basées sur des règles.

La méthode proposée par Chairsorn *et al.* [Chaisorn 2003] se base sur une segmentation en plans, et sur une catégorisation des plans en fonction des différentes catégories de séquences que l'on retrouve dans des journaux télévisés (publicité, présentateur, reportage sur le terrain, « jingles », météo, etc...). Ils proposent d'utiliser des descripteurs

visuels (histogrammes de couleur), temporels (changement de scène, changement de locuteur, catégories audio : parole, musique ou bruit, descripteurs de mouvement, ...), et des descripteurs haut niveau (visages, texte dans l'image, ...) pour classer les plans de la vidéo en 12 catégories différentes. Ils utilisent les HMM pour modéliser le changement de reportages et détecter les transitions entre les reportages à partir de ces descripteurs de plans (cf. Figure 2.26). Bien que l'on puisse considérer un reportage de journal télévisé comme une histoire (bien que ne suivant pas la même définition que celle proposée dans cette thèse), les différentes scènes ou séquences de vidéos du reportage se suivent alors que les histoires telles que proposées dans cette thèse sont composées de scènes parfois non contiguës. C'est pourquoi la section suivante s'intéresse au regroupement de séquences traitant d'un même sujet mais non contiguës dans un document (ou qui n'appartiennent pas à un même document).

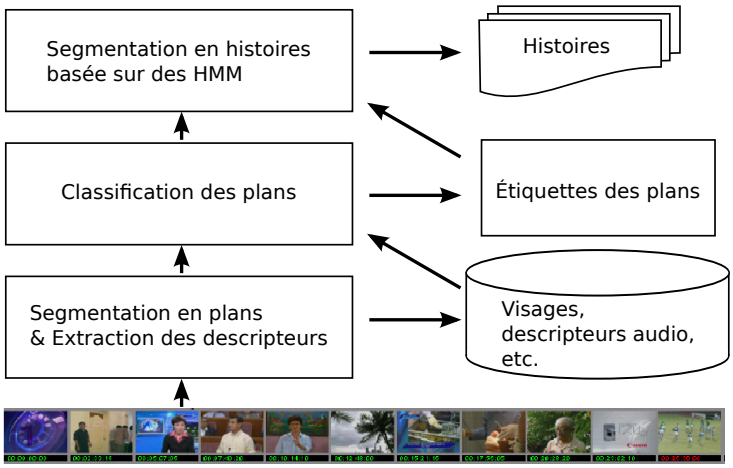

FIGURE 2.26 – *Méthode proposée par [Chaisorn 2003]*

2.4.2 Regroupement de documents suivant leur ressemblance sémantique

Cette section s'intéresse aux méthodes proposant de regrouper des reportages traitant d'un même sujet. La plupart des travaux traitant de ce sujet ont été développés pour des systèmes de navigation dans des bases de données, en proposant à l'utilisateur de consulter tous les articles traitant du sujet demandé. Certains s'intéressent aux reportages de journaux télévisés, d'autres à des articles de journaux uniquement textuels. Tous ces travaux se différencient par les données qu'ils utilisent pour mettre en relation les articles relatant la même histoire.

Pour Ide *et al.* [Ide 2005], les composants essentiels d'un reportage peuvent être résumés en se posant les questions de « quand, où, qui, quoi, pourquoi, comment ». Ainsi, Christel *et al.* [Christel 1999] proposent une méthode basée sur la question du « où ».

Ils proposent d'utiliser des systèmes de reconnaissance automatique de la parole pour extraire le texte de la vidéo, synchronisé avec les sous-titres s'ils sont disponibles. Ils utilisent ensuite une base donnée de lieux géographiques pour repérer les lieux prononcés. Chaque vidéo est décrite par l'ensemble des lieux qui y sont prononcés, et ils considèrent que deux vidéos ou reportages traitent d'un même sujet si les lieux qui les décrivent sont similaires.

Duygulu *et al.* [Duygulu 2004] utilisent un ensemble de règles extraites des techniques de mise en forme des reportages pour regrouper ceux qui relatent le même sujet. Ils ont remarqué que des séquences identiques ou très similaires apparaissent dans deux reportages traitant d'un même sujet, comme illustrées par la Figure 2.27. Ils utilisent un ensemble de descripteurs (moments de couleur, textures, position et taille des visages) pour décrire les images clefs des plans de la vidéo, et regrouper les séquences identiques provenant de l'ensemble de leur base de données de reportages. Deux reportages possédant des séquences identiques sont considérés comme présentant un même sujet. Ils proposent une deuxième méthode de regroupement des reportages en analysant les « logos » présents à l'écran. En effet, beaucoup de journaux télévisés associent un logo dans un coin de l'écran à un sujet donné. Pour Duygulu *et al.*, la mise en correspondance des reportages peut se faire en détectant et comparant les logos affichés à l'écran.

Reportage de CNN le 04/02/1998

La maison blanche demande à l'Iraq d'ouvrir ses portes à des inspecteur internationaux de l'ONU

Reportage de CNN le 11/02/1998

L'Iraq permet une nouvelle fois à un nombre limité d'inspecteurs de l'ONU d'inspecter huit sites présidentiels. Les Etats-Unis demandent un accès complet à tous les sites.

FIGURE *2.27 – Exemple de deux reportages différents présentant des séquences vidéo similaires [Duygulu 2004]*

Pour Ide *et al.* [Ide 2005], la question qui se pose pour mettre en correspondance les reportages est la question du « qui ». Ils traquent les personnages présents dans les reportages en recherchant dans le texte les noms propres à partir des particules annonçant les personnages (la base de données utilisée étant en Japonais, les noms propres sont annoncés par des particules comme « *san, sama, kun,* etc. »). Ils considèrent que deux personnes apparaissant ensemble dans un reportage sont en relation l'une avec l'autre. Ils en déduisent un graphe de relation des personnages, à partir duquel ils proposent un système de navigation dans leur base de données centré sur la relation des personnages.

Vadrevu *et al.* [Vadrevu 2011] proposent une méthode de regroupement d'articles pour un système on-line de recherche de news. Ils utilisent deux informations pour décrire les articles :

- les mots prononcés. L'article est décrit par un vecteur de données dont les valeurs sont le TF·IDF des mots prononcés durant le reportage après suppression des « mots vides » (mots communs comme « le, la, de, ça, etc. ») et une lemmatisation des mots (regroupement des mots d'une même famille sous leur forme canonique).
- les sujets wikipedia traités. Ils proposent de faire un lien entre les expressions contenues dans le texte de l'article et les pages wikipedia corresondantes. Un score est associé à chaque sujet en fonction de leur importance. Le descripteur est le tableau de ces scores.

Un vecteur descripteur de chaque article est généré comme la concaténation des deux vecteurs descripteurs. La distance entre deux articles est la distance cosinus de leurs vecteurs descripteurs, en tenant compte de la date de publication des articles (deux articles publiés à la même époque ont plus de chances d'être proches que deux articles ayant des dates de publication éloignées).

2.4.3 Évaluation du regroupement de séquences traitant d'un même sujet

Identification des reportages

L'identification des reportages dans des journaux télévisés est un processus de segmentation. L'évaluation est réalisée en comparant les segments d'histoires obtenus automatiquement avec une segmentation manuellement annotée. Ainsi, dans la campagne d'évaluation TRECVID [Smeaton 2003], l'évaluation des méthode de détection des histoires est proche de celle proposée pour la segmentation en scènes et décrite Section 2.3.4 (page 63). Elle consiste en un calcul de la précision et du rappel en considérant qu'une transition est correcte si elle se trouve à moins de 5 secondes d'une transition annotée.

Regroupement de documents suivant leur ressemblance sémantique

L'évaluation d'un regroupement de documents est sont souvent liée aux travaux pour lesquels la méthode de regroupement a été développée. Par exemple, pour Duygulu *et al.* [Duygulu 2004], l'évaluation de la méthode de regroupement de reportages est faite de manière extrinsèque :le regroupement est une première étape à une méthode de génération de résumés de reportages, et seule la qualité du résumé est évaluée.

Pour Christel *et al.* [Christel 1999] et Ide *et al.* [Ide 2005], le regroupement est évalué en présentant une interface de navigation basée sur la méthode proposée.

Vadrevu *et al.* [Vadrevu 2011] proposent une évaluation chiffrée du regroupement des reportages traitant d'un même sujet. Ils ont annoté l'ensemble des « paires de reportages » en trois catégories :

- les deux reportages doivent être liés
- les deux reportages peuvent être liés
- les deux reportages ne doivent pas être liés

Ils comparent ensuite les paires de reportages liés automatiquement par leur système à ces annotations. Ils considèrent qu'un bon système doit avoir un grand ratio de paires liés par le système et annotés par les classes « doivent être liés » et « peuvent être liés », et un faible ratio de couples liés par le système et annotés « ne doivent pas être liés ».

L'évaluation proposée par Vadrevu *et al.* [Vadrevu 2011] consiste à évaluer la qualité d'un regroupement d'objets en comparant un regroupement obtenu automatiquement (hypothèse) avec un regroupement réalisé par des experts (référence). L'évaluation de la qualité d'un regroupement est un domaine qui a longuement été étudié.

Ainsi, pour Manning *et al.* [Manning 2008], il est possible d'évaluer un regroupement d'objets en analysant les paires (i, j) d'objets, et en répondant au problème de classification suivant : *est-ce que les objets i et j font partie du même groupe ?* Quatre quantités peuvent être définies :

- **VP** : nombre de vrais positifs (objets correctement regroupés et faisant partie d'un même groupe)
- **VN** : nombre de vrais négatifs (objets ne faisant pas partie d'un même groupe et annotés comme ne faisant pas partie d'un même groupe)
- **FP** : nombre de faux positifs (objets regroupés mais ne faisant pas partie d'un même groupe dans la vérité terrain)
- **FN** : nombre de faux négatifs (objets non regroupés mais faisant partie d'un même groupe dans la vérité terrain)

À partir de ces valeurs, des mesures de **précision (P), rappel (R) ou F-mesure ($\mathbf{F_{PR}}$)** peuvent être calculées tel que

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FP}} \\ \mathrm{R} &= \frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FN}} \\ \mathrm{F_{PR}} &= 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P}+\mathrm{R}} \end{aligned}$$

De nombreuses mesures permettant de comparer à quel point deux regroupement sont similaires existent et sont basées sur le compte VP, VN, FPet FN. Ainsi, Rand [Rand 1971]

propose le **Rand Index** (RI) comme étant le pourcentage de paire d'objets pour lesquels la référence et l'hypothèse sont d'accord :

$$\mathrm{RI} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}} \tag{2.19}$$

Hubert et Arabie [Hubert 1985] proposent de normaliser le score de comparaison entre une référence et une hypothèse par une « correction pour la chance » de ce score. Leur idée est de modéliser l'espérance que peut atteindre le score d'un regroupement aléatoire (score attendu) pour qu'un regroupement proche de ce modèle ai un score égal à zéro. La forme générale du score corrigé pour la chance est présenté dans l'équation 2.20.

$$\frac{\text{score} - \text{score attendu}}{\text{score maximum} - \text{score attendu}} \tag{2.20}$$

Ainsi, ils proposent de calculer le **Rand Index Ajusté**, ou *Adjusted Rand Index* (ARI) comme étant le « Rand Index corrigé pour la chance ». Soient $\mathcal{S} = \{s_1, s_2, ...s_N\}$ un ensemble de N objets à regrouper. $\mathcal{H} = \{h_1, h2, ...h_q\}$ est l'ensemble des groupes d'objets de l'hypothèse et $\mathcal{R} = \{r_1, r_2, ...r_p\}$ est l'ensemble des groupes d'objets de la référence. La formule 2.21 montre la table de contingence entre les deux partitions qui représente les différences/similarités entre tous les groupes de $\mathcal{H}$ et $\mathcal{R}$.

$\mathcal{R}\backslash\mathcal{H}$	h_1	h_2	$\cdots$	h_q	Somme
r_1	n_{11}	n_{12}	$\cdots$	n_{1q}	a_1
r_2	n_{21}	n_{22}	$\cdots$	n_{2q}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r_n	n_{p1}	n_{p2}	$\cdots$	n_{pq}	a_p
Somme	b_1	b_2	$\cdots$	b_q	N

(2.21)

avec n_{ij} le nombre d'objets commun aux groupes r_i et h_j, $a_i = \sum_{j=1..q} n_{ij}$ et $b_j = \sum_{i=1..p} nij$.

En considérant que la valeur maximale du Rand Index est de 1, Hubert et Arabie [Hubert 1985] définissent le calcul de l'*Adjusted Rand Index* comme suit :

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \tag{2.22}$$

Chapitre 3

Comparaison de segments temporels de séries télévisées

La manière de décrire des documents audiovisuels en vue de retrouver automatiquement leur structure dépend du type de contenu étudié et du but recherché. Ce chapitre présente nos contributions dans le domaine de la description et de la comparaison de segments temporels pour les tâches de segmentation en scènes et de regroupement des scènes en histoires pour des épisodes de séries télévisées.

Nous étudions 3 types de descripteurs qui permettent de répondre à 3 questions essentielles nécessaires à l'extraction de la structure narrative : où se situe l'action (où ?), qui est impliqué dans l'action (qui ?), que se passe-t-il (quoi ?).

Pour répondre à chacune de ces questions, un descripteur est extrait du flux audio ou vidéo comme illustré dans la Figure 3.1 :

- « Où ? » : descripteurs de couleur (HSV).
- « Quoi ? » : transcription automatique de la parole (ASR).
- « Qui ? » : présence des locuteurs (SD).

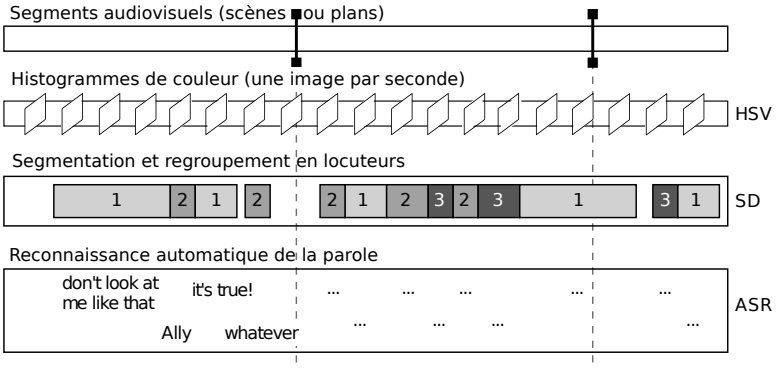

FIGURE *3.1 – Les trois principaux descripteurs de segments audiovisuels utilisés pour les tâches de segmentation en scènes et de regroupement des scènes en histoires.*

3.1 Où ? Similarité entre lieux

Selon notre définition de la scène (page 13), un changement de lieux entre deux séquences de vidéo implique un changement de scène. De plus, deux scènes d'une même histoire ont tendance à se dérouler dans un même lieu. Ainsi, la connaissance du lieu dans lequel se déroule l'action est une information très pertinente pour la structuration de documents audiovisuels narratifs.

Bien qu'il existe de multiples façons de décrire des lieux (environnement visuel ou sonore), nous les décrivons uniquement à partir de la couleur des images qui composent les séquences de vidéo. En effet, dans les séries télévisées, l'environnement sonore est masqué par la musique et les dialogues. Nous n'avons donc pas trouvé pertinent d'utiliser l'information audio pour définir un lieu. De plus, la couleur est une des informations les plus simples et les plus utilisées pour la structuration d'épisodes de séries télévisées [Yeung 1998, Kender 1998, Tavanapong 2004, Zhao 2007]. Elle donne une information pertinente sur l'environnement visuel global de la séquence dans le cas des séries télévisées.

Observation 1
Deux scènes consécutives ont généralement des couleurs globales différentes.

Cette observation est particulièrement pertinente lorsque les deux scènes sont filmées dans des lieux différents. La Figure 3.2 montre un exemple de l'utilisation de la couleur pour retrouver la transition entre deux scènes. Cet exemple montre une image caractéristique des quatre derniers plans d'une scène et des quatre premiers plans de la scène suivante. On remarque que la teinte principale des plans est similaire pour des plans appartenant à une même scène et différente pour des plans appartenant à deux scènes différentes.

FIGURE *3.2 – Illustration de l'intérêt de descripteurs de couleur pour la segmentation en scènes. Les plans des deux scènes ont des teintes principales très différentes qui permettent de différencier les scènes.*

Cependant, la couleur peut fortement varier au sein d'une même scène selon les changements d'angles de vue de la caméra. Yeung *et al.* [Yeung 1998], Hanjalic *et al.* [Hanjalic 1999] ou Tavanapong [Tavanapong 2004] utilisent les connaissances sur les règles de montage traditionnellement utilisées pour les films ou les séries télévisées pour les segmenter en

scènes. La Figure 3.3 explique pourquoi l'utilisation de la couleur est pertinente pour la segmentation en scènes dans le cas des champs/contre-champs.

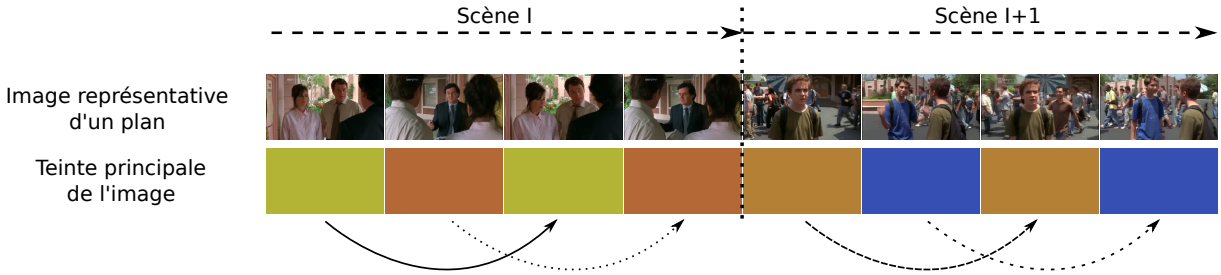

FIGURE 3.3 – *La couleur moyenne des images peut varier au sein d'une scène. La connaissance des techniques cinématographiques comme le « champ/contrechamp » permettent de retrouver la position de la transition entre les deux scènes. Les flèches lient les plans ayant une couleur proche. On remarque qu'aucun lien ne lie des plans de scènes différentes.*

Observation 2
Les scènes d'une même histoire ont généralement des couleurs similaires.

Pour certains épisodes de séries télévisées, la couleur donne une information très pertinente sur les histoires qui s'y déroulent. Dans une série comme « le Trône de Fer », deux histoires différentes se déroulent dans deux lieux différents, résultant en des couleurs globales différentes entre les histoires comme dans l'exemple de la Figure 3.4.

3.1.1 Similarité basée sur la couleur

La description de séquences de vidéo par la couleur est inspirée des méthodes de description et de comparaison d'histogrammes de couleur présentées dans la Section 2.1.3 (page 43).

Elle repose sur les histogrammes de couleur HSV (*Hue, Saturation, Value* : Teinte, Saturation, Valeur) des images composant la séquence. Les histogrammes de couleur (de dimension $10 \times 10 \times 10$) sont extraits chaque seconde et la distance d^{HSV} entre deux séquences s_i et s_j est exprimée par la distance de Manhattan (d_{L1}) minimale entre toutes les paires possibles d'histogrammes issus de ces deux séquences.

$$d^{\text{HSV}}(s_i, s_j) = \begin{cases} \dfrac{1}{|\mathbf{H}_i|} \displaystyle\sum_{h \in \mathbf{H}_i} \min_{g \in \mathbf{H}_j} d_{\text{L1}}(h, g) & \text{si } |\mathbf{H}_i| > |\mathbf{H}_j| \\ d^{\text{HSV}}(s_j, s_i) & \text{sinon} \end{cases} \tag{3.1}$$

avec $\mathbf{H}_i$ et $\mathbf{H}_j$ l'ensemble des histogrammes de couleur extraits des images des séquences s_i et s_j, et $|\mathbf{H}_i|$ et $|\mathbf{H}_j|$ le nombre de ces histogrammes.

FIGURE *3.4 – La couleur moyenne des images donne une information pertinente pour classer les scènes en histoires. Dans cet exemple, deux scènes de deux histoires différentes sont représentées. L'histoire X se déroule dans un lieu froid, les images extraites des scènes ont des teintes bleues et des couleurs froides. L'histoire Y se déroule dans un désert fait de plaines et de steppes. Les couleurs sont chaudes avec des teintes oscillant entre le jaune et le marron.*

Ce type de comparaison de séquences par la couleur a été choisi pour sa simplicité d'implémentation, au regard de la bonne performance obtenue par des systèmes de recherche d'images à partir de simples histogrammes HSV [Novak 1992]. L'utilisation de plusieurs images d'une séquence permet de tenir compte des variations au sein de la scène. Beaucoup de méthodes de structuration décrivent une séquence vidéo à partir d'une unique image clef. Cette image clef peut être la première, la dernière ou l'image centrale de la séquence [Rui 1999, Tavanapong 2004]. Cependant, elles ne sont représentatives que pour des séquences statiques ayant peu de variation de couleur tout au long de la séquence (comme des plans sans déplacement).

La Figure 3.5 illustre l'utilisation de telles distances pour les tâches de segmentation en scènes et de regroupement des scènes en histoires au regard des observations présentées précédemment. Ces graphes montrent la répartition des distances entre des séquences des épisodes de la série *Le Trône de Fer*. La distribution des distances dans l'histogramme de gauche permet de distinguer les distances entre scènes appartenant à une même histoire (en vert dans l'histogramme) de celles entre scènes d'histoires différentes (en rouge). De la même manière, la distribution des distances dans le graphique de droite permet de distinguer les distances entre les plans appartenant à la même scène. Les distances HSV

sont donc pertinentes pour la tâche de segmentation en scènes et de regroupement des scènes en histoires.

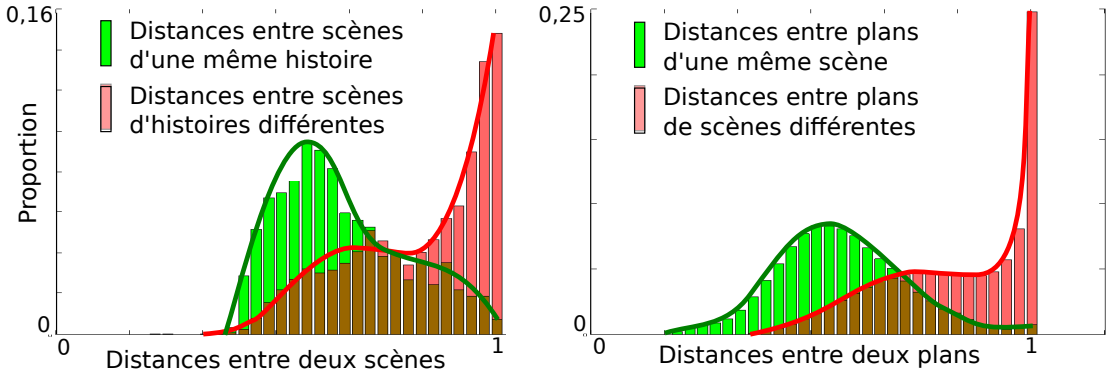

FIGURE *3.5 – Répartition des distances d^{HSV} entre séquences de vidéo. En vert, répartition des distances des scènes appartenant à la même histoire (à gauche) ou des plans appartenant à la même scène (à droite). En rouge, répartition des distances des scènes appartenant à deux histoires différentes (à gauche) ou des plans appartenant à des scènes différentes (à droite).*

3.1.2 Limites des descripteurs de couleur

Les deux observations présentées précédemment montrent que la couleur est une information pertinente pour la structuration d'épisodes de séries télévisées. Cependant, l'information colorimétrique montre un certain nombre de limites dont il faut tenir compte pour les méthodes présentées dans les chapitres suivants.

Nous présentons la couleur comme une information caractéristique du lieu où se déroule l'action. Cependant, il existe des cas où la couleur varie alors que le lieu reste constant. C'est le cas lorsque l'on observe un champ/contre-champ par exemple. Le changement d'angle de vue de la caméra provoque un changement des sujets filmés, et donc de la couleur globale de l'image, comme illustré dans la Figure 3.3 où les plans d'une même scène peuvent être très dissemblables.

À l'inverse, deux lieux différents peuvent avoir une ambiance visuelle très proche. Les histogrammes de couleur sont une information visuelle globale de l'image. Or, comme exposé dans l'état de l'art (Section 2.1.1.1, page 31), deux images représentant des objets ou des scènes différentes peuvent avoir des histogrammes similaires. De plus, la couleur d'une image ne représente pas uniquement le lieu, mais aussi les personnages présents à l'image.

Pour tenir compte de cette limite, nous utilisons l'implémentation OpenCV [Bradski 2000] de l'algorithme de Viola & Jones [Viola 2004] pour détecter les visages présents dans les images. Les personnages sont supprimés des images en masquant les pixels correspondant

aux visages détectés ainsi que la zone sous le visage correspondant au corps de la personne détectée comme illustré par la Figure 3.6. L'histogramme de couleur est calculé à partir de tous les pixels non masqués de l'image.

Aucune évaluation n'a été réalisée pour étudier la fiabilité de la détection des visages. D'après [Viola 2004], leur algorithme est en mesure de retrouver 90% des visages de leur corpus avec 10% de faux positifs. De plus, cette méthode ne permet pas de supprimer les personnages positionnés de dos, et nous utilisons deux modèles de visages : un pour les visages de face et un deuxième pour les visages de profil.

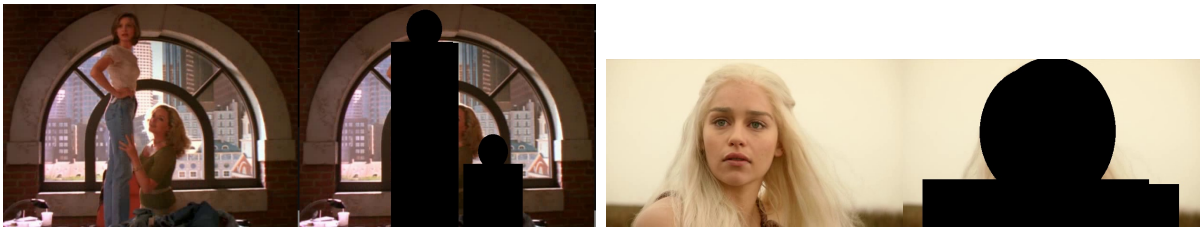

FIGURE *3.6 – Masque des personnages par détection des visages.*

3.2 Quoi ? Similarité entre les dialogues

La deuxième information jugée pertinente pour la structuration d'épisodes de séries télévisées est le texte issu de l'analyse des dialogues qui contient une grande partie de l'information sémantique de la scène.

Pour Colonna [Colonna 2010], dans les séries télévisées, la composante audio est plus importante que la vidéo. Il justifie cette affirmation par deux observations :

- Le budget des séries télévisées est beaucoup plus faible que celui d'un film. Or, produire des sons coûte moins cher que produire des images, ce qui pousse à mettre en avant les informations sonores pour des raisons financières.
- Les séries télévisées sont créées pour être visionnées dans un cadre domestique qui, contrairement aux salles de cinéma, offre un environnement dégradé. Il est fréquent de faire des activités en parallèle du visionnage d'un épisode, ce qui fait que nos yeux ne sont pas fixés sur l'écran. Ainsi, les séries télévisées retiennent l'attention du spectateur à partir de la bande-son, et notamment des dialogues, plus que par l'information visuelle.

La Figure 3.7 illustre une utilisation possible du texte pour la segmentation en scènes. Dans cet exemple, le texte extrait de chaque plan est indiqué. Il est possible de regrouper les plans appartenant à une même scène en retrouvant les mots en relation avec le sujet de la scène.

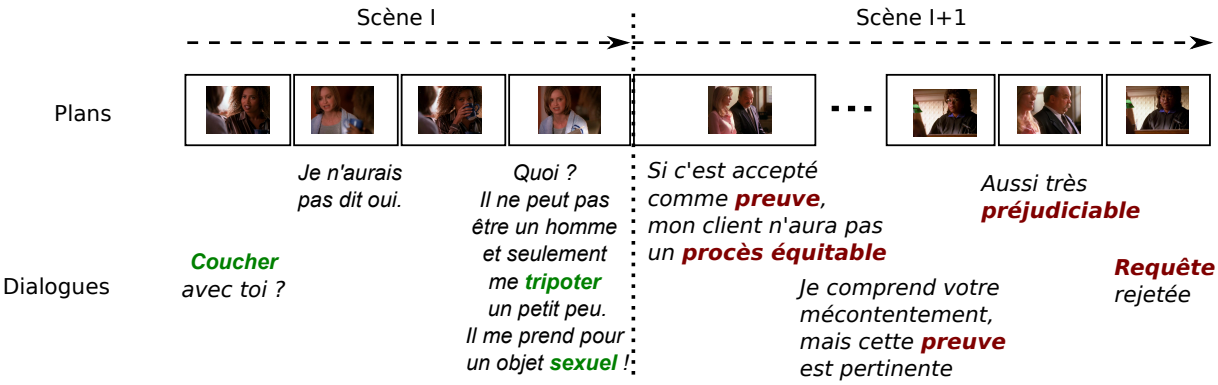

FIGURE 3.7 – *Le vocabulaire utilisé durant les plans de chaque scène est lié au sujet de la scène. Dans la scène I, le sujet traite des aventures conjugales d'Ally, on retrouve dans les dialogues des termes en relation avec la vie conjugale (mots en vert). Dans la scène I+1, le vocabulaire est propre au procès en cours (mots en rouge).*

L'étude des mots présents dans des séquences traitant d'un même sujet permet de faire l'observation suivante :

Observation 3
Le vocabulaire utilisé entre deux séquences traitant d'un même sujet est plus similaire qu'entre deux séquences traitant de sujets différents.

Cette observation a conduit à la définition d'une mesure de similarité entre deux séquences audiovisuelles basée sur la transcription de la parole.

3.2.1 Similarité basée sur la transcription

Comme l'accès à des métadonnées (script ou sous-titres) n'est pas toujours disponible, **les dialogues des personnages sont obtenus à partir d'un système de transcription automatique de la parole (ASR)** [Gauvain 2002].

Une première étape de filtrage des mots est nécessaire : à partir de la sortie de l'ASR, les **lemmes** sont extraits à l'aide de TreeTagger [Schmid 1994]. Un lemme est la forme canonique d'un mot, comme l'infinitif pour un verbe. Ainsi, les mots « est », « sommes », « êtes », « sont » sont tous regroupés sous la même forme canonique « être ».

Soit la liste des séquences $\mathcal{S}$ d'une vidéo (plans ou scènes, selon l'application), chaque séquence s de $\mathcal{S}$ est décrite par un **vecteur TF · IDF**(s) de dimension q_{ASR}, où q_{ASR} est le nombre total de lemmes uniques reconnus par le système ASR dans un épisode. Dans le corpus de test, il y a en moyenne 871 lemmes différents par épisode.

Chaque composante du vecteur représente la valeur $\text{TF} \cdot \text{IDF}_w$ d'un lemme w :

- Le terme IDF est défini par $\text{IDF}_w = \log\left(|\mathcal{S}|/M_w\right)$ où $|\mathcal{S}|$ est le nombre de séquences dans la vidéo et M_w est le nombre de séquences contenant au moins une occurrence du lemme w.
- Le terme TF est défini par $\text{TF}_w(s) = W_w(s)/W(s)$ où $W_w(s)$ est le nombre d'occurences du lemme w dans la séquence s et $W(s)$ est le nombre de lemmes présents dans la séquence s.

La distance d^{ASR}, basée sur la sortie de l'ASR, entre les séquences s_i et s_j est définie par la similarité cosinus entre leurs vecteurs $\text{TF} \cdot \text{IDF}$ respectifs.

$$d^{\text{ASR}}(s_i, s_j) = 1 - sim_{\cos}(\text{TF} \cdot \text{IDF}(s_i), \text{TF} \cdot \text{IDF}(s_j)) \tag{3.2}$$

Ce type de représentation et de comparaison est emprunté au domaine du « regroupement de textes » [Salton 1988]. Dans ce domaine, le $\text{TF} \cdot \text{IDF}$ est utilisé pour pondérer les mots ou lemmes présents dans un ensemble de documents, dans le but de regrouper les documents traitant d'un même sujet. Dans notre cas, un document correspond à une séquence de vidéo, et l'ensemble des documents est l'ensemble des séquences que l'on souhaite comparer.

Pondérer les mots par leur valeur $\text{TF} \cdot \text{IDF}$ permet de réduire l'influence des « **mots vides** » (*stop words* en anglais). Ce sont des mots très peu porteurs de sens du fait qu'ils sont très communs et présents dans la majorité des documents. Ces mots sont caractéristiques de la langue employée. En français il s'agit principalement d'articles, de prépositions ou de pronoms (e.g. « le », « la », « les », « du », « être » ou « je »). Le $\text{TF} \cdot \text{IDF}$ permet de réduire l'influence de ces « mots vides » (et plus précisément la composante IDF).

Les histogrammes de la Figure 3.8 illustrent l'utilisation de la distance d^{ASR} pour les tâches de segmentation en scènes et de regroupement des scènes en histoires. Cette répartition est très différente de ce que l'on a pu observer pour la couleur.

Il n'y a pas de séparation franche des distances entre les scènes d'une même histoire (barres vertes) et les distances entre les scènes d'histoires différentes (barres rouges). Cependant, la distribution des distances tend à montrer que les plus faibles distances sont majoritairement des distances entre scènes d'une même histoire.

La répartition des distances entre plans montre qu'elles sont essentiellement très grandes ou très petites. Plusieurs facteurs expliquent ce résultat. Comme il y a très peu de mots prononcés dans chaque plan, la quantité de mots communs à deux plans est très faible. De plus, 16% des plans de notre corpus ne comportent aucune parole. Ces plans sont décrits par des vecteurs $\text{TF} \cdot \text{IDF}$ nuls, et leur distance avec n'importe quel autre plan aura une valeur de 1. On remarque cependant que les petites distances incluent en majorité des distances entre plans d'une même scène, alors que les grandes distances incluent en majorité des plans de scènes différentes.

Ces histogrammes nous incitent à penser que l'information fournie par les dialogues est moins fiable que la couleur pour les tâches désirées. Cette différence peut s'expliquer en partie par les limites du système de transcription automatique de la parole que nous avons utilisé. L'utilisation des dialogues pour la segmentation en scènes et le regroupement des scènes en histoires est discutée dans les Chapitres 4 et 5.

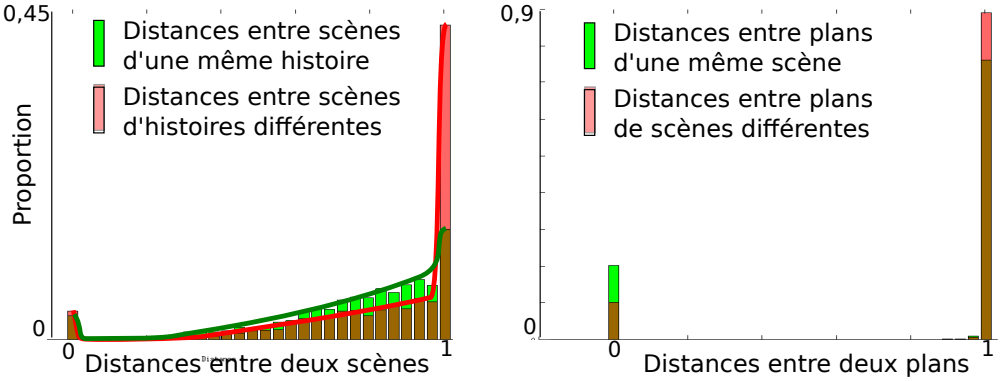

FIGURE 3.8 – *Répartition des distances d^{ASR} entre séquences de vidéo. En vert, distribution des distances entre scènes appartenant à la même histoire (à gauche) ou entre plans appartenant à la même scène (à droite). En rouge, distribution des distances entre scènes appartenant à deux histoires différentes (à gauche) ou entre plans appartenant à des scènes différentes (à droite).*

3.2.2 Limites des systèmes ASR

Le système de transcription automatique de la parole que nous avons utilisé [Gauvain 2002] a été développé pour des journaux télévisés ou radiophoniques. Les taux d'erreurs de ce système sur ce type de documents est inférieur à 20% [Gauvain 2002]. En revanche, ses performances sur des épisodes de séries télévisées sont détériorées à cause de différents facteurs : interactions très rapides entre les personnages, parole spontanée (bien que jouée) ou parole superposée.

Ainsi, de nombreuses erreurs apparaissent dans la transcription générant de ce fait des erreurs lors des étapes suivantes de la chaîne de traitement. Comme aucune transcription manuelle de la parole n'a été réalisée, il n'est pas possible de mesurer le taux d'erreurs inhérentes à la transcription automatique des épisodes de séries télévisées étudiés. Cependant, pour illustrer la qualité de la transcription automatique, le Tableau 3.1 montre un exemple du résultat de la reconnaissance automatique de la parole comparée à un court extrait d'un épisode de la série *Malcolm*.

Vérité terrain	Résultats de l'ASR
- **What** ?	- All the **what** one
- **Ok, I** ate **the cupcakes** that have been **produced** in class **last night**.	- **okay, I the cupcakes produced last night**.
- And I took dad's **license to make** a fake ID.	- back to get **license to make a** big Daddy
- And I can't return your necklace **because** I already **sold it** !	- we didn't. That was **because** art **sold it**
- **I was just going to say** your **shirt** doesn't go with your pants.	- **I was just going to say** you **shirt** in court appearance.
- Ah, pfiou (onomatopée) !	- So few
- **Would you like** more orange juice steevie ?	- **would you like** tomorrow's
- Thank **you**, Loïs.	- think **you** well.
- Steevie **stay** at the **house for a week** while his **parents** are on a ride. **They** get drinks **on** the **beach** while **he gets to watch my dad hair drying in the** kitchen.	- This **staying** in **house for a week** was **parents** and why **they** did travel with turns on a **beach**. **He gets to watch my dad air drying in the** kid
- **If everyone noticed** the way Steevie **pre-sliced the grapefruit sections. It's so nice to have a boy in this house** who **is not** a rude little **monster**.	- **if everyone noticed pre sliced the grapefruit sections. It's so nice have a boy in the house is not** a real **monster**.

TABLE *3.1 – Comparaison d'un système de transcription automatique de la parole avec la vérité-terrain correspondante. En vert sont indiqués les mots correctement reconnus.*

3.3 Qui ? Similarité entre les personnages

Le nombre de personnages communs à deux scènes consécutives est en moyenne de 10% pour une série comme *Malcolm in the Middle*, 29% pour *Ally McBeal* et 12% pour *A Game of Thrones*.

Observation 4
Les personnages présents dans deux scènes consécutives sont généralement différents. Il est donc possible d'utiliser cette information pour retrouver les transitions entre les scènes, comme illustré Figure 3.9.

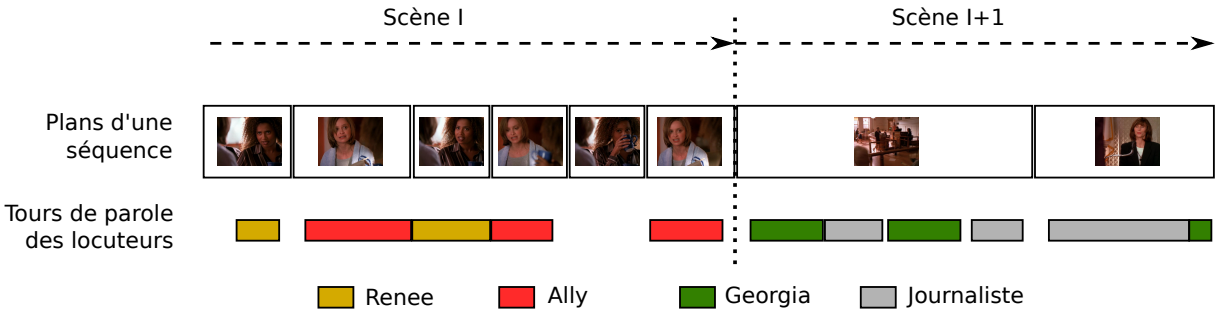

FIGURE 3.9 – *Illustration de l'intérêt de descripteurs basés sur la présence de personnages. Les deux scènes illustrées impliquent des personnages différents. Une simple analyse des personnages permet de retrouver la position de la transition de scènes.*

En outre, d'après *L'art des séries télé* de Vincent Colonna [Colonna 2010], une histoire *comprend un schéma d'action* où *un personnage important rencontre des obstacles en agissant en vue d'un but*. Les personnages sont les éléments centraux d'une histoire.

La Figure 3.10 montre les personnages présents dans les différentes histoires d'un épisode de la série *Malcolm*. Dans cet exemple, les arcs indiquent une présence simultanée et une interaction entre les personnages au sein de l'épisode. Les différentes histoires ont très peu de personnages en commun, et il y a beaucoup plus d'interaction entre les personnages appartenant à une même histoire qu'entre les personnages évoluant dans deux histoires différentes.

Observation 5
Deux histoires différentes d'un même épisode impliquent généralement des personnages différents.

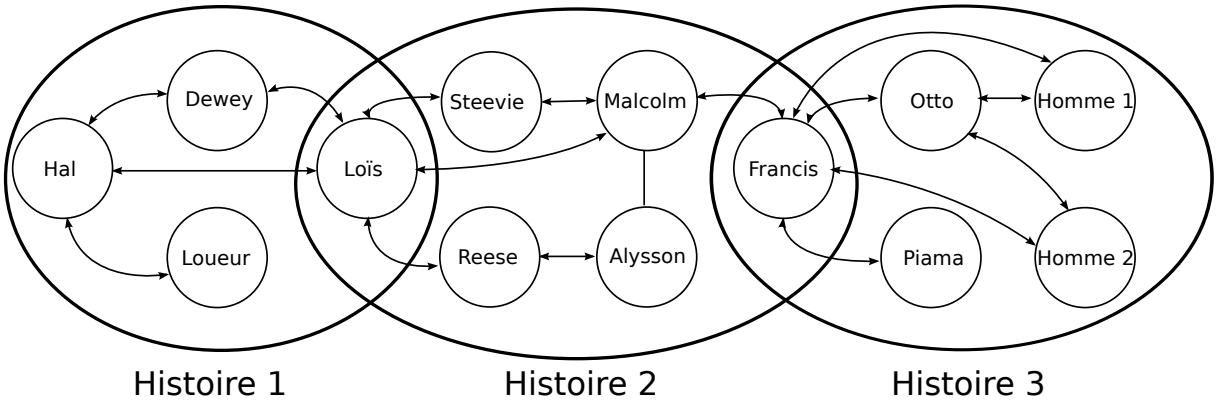

FIGURE 3.10 – *Personnages présents dans les différentes histoires d'un épisode de* Malcolm. *Les arcs indiquent une relation forte entre les personnages (présence simultanée et interaction des personnages).*

Détecter la présence de personnages peut se faire en recherchant les visages (information vidéo), en recherchant les tours de parole (information audio), ou en analysant l'information audio et vidéo simultanément.

3.3.1 Similarité basée sur la présence des locuteurs

D'après Colonna [Colonna 2010], l'information sémantique contenue dans un épisode de série télévisée passe plus par le son que par l'image. Ainsi, si un personnage important est présent à l'écran, il parle à un moment ou un autre. Bien que cette affirmation ne soit pas toujours exacte, se concentrer sur les locuteurs (personnages parlant), permet de ne pas prendre en compte les figurants et de se concentrer sur les personnages importants, Nous utilisons donc un système de segmentation et regroupement en locuteurs basé sur le flux audio seul.

Les deux observations liées aux locuteurs (4 et 5) ont conduit à définir des mesures de distance entre segments de vidéo basées sur la connaissance de la présence de locuteurs. Trois types de distances sont proposés :

- La première méthode consiste à décrire la présence des locuteurs sous forme vectorielle. Il est possible de représenter n'importe quel segment audio par un vecteur binaire $\mathbf{x}$ de dimension q, où q est le nombre de locuteurs présents dans une vidéo, tel que

$$\mathbf{x} \in \{0,1\}^q \text{ avec } \mathbf{x} = [x_1, x_2, ..., x_q]$$

$$\text{où } x_i \quad = \begin{cases} 1 & \text{si le locuteur } i \text{ parle dans le segment s} \\ 0 & \text{sinon} \end{cases} \tag{3.3}$$

La Figure 3.11 illustre le vecteur descripteur pour quatre segments d'une vidéo.

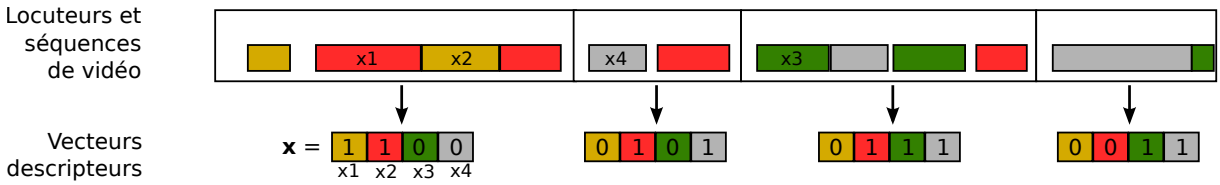

FIGURE *3.11 – Représentation des locuteurs pour des séquences d'une vidéo. Les séquences sont représentées par un vecteur binaire* $\mathbf{x}$ *de dimension* q*, où* q *est le nombre de locuteurs présents dans un vidéo. Chaque composante du vecteur indique pour chaque locuteur sa présence/absence dans le segment décrit.*

Ce type de représentation donne la même importance à tous les locuteurs. Cependant, parmi les locuteurs présents dans un épisode de série télévisée, certains sont présents dans la majorité des scènes et des histoires, alors que d'autres ne font que

des apparitions sporadiques ou ponctuelles. Il est possible que l'importance d'un locuteur pour la structuration d'une vidéo varie en fonction de son temps de parole. Ainsi, pour calculer la distance entre deux segments de vidéo, trois différentes pondérations α peuvent être appliquées à un locuteur ℓ :

– $\alpha^{=}$ / même poids pour l'ensemble des locuteurs / $\alpha_\ell = \dfrac{1}{q}$

– α^{+} / les locuteurs principaux ont un poids plus fort / $\alpha_\ell = \dfrac{\mathcal{D}_\ell}{\sum_{j=1}^{q} \mathcal{D}_j}$

– α^{-} / les locuteurs principaux ont un poids plus faible / $\alpha_\ell = 1 - \dfrac{\mathcal{D}_\ell}{\sum_{j=1}^{q} \mathcal{D}_j}$

tel que $\sum\limits_{\ell=1}^{q} \alpha_\ell = 1$

avec $\mathcal{D}_\ell$ la durée totale de parole du locuteur ℓ. La distance entre deux séquences de vidéo s_i et s_j est définie par la distance entre leurs vecteurs descripteurs $\mathbf{x}_i$ et $\mathbf{x}_j$ en fonction de la mesure de pondération désirée.

$$d_\alpha^{\alpha\text{SD}}(s_i, s_j) = \frac{1}{q} \sum_{\ell=1}^{q} \alpha_\ell \cdot |\mathbf{x}_{i\ell} - \mathbf{x}_{j\ell}| \tag{3.4}$$

- La deuxième méthode est proche de la distance d^{ASR}. L'idée est d'appliquer une pondération TF $\cdot$ IDF, en considérant un locuteur comme un mot et une séquence comme un document. En effet, dans certaines séries comme *Ally McBeal*, il arrive que des locuteurs (ici *Ally*) soient centraux à toutes les histoires et présents dans la majorité des scènes. Ces locuteurs ne sont donc pas discriminants pour déterminer qu'une scène appartient à une histoire plutôt qu'à une autre ou pour retrouver les transitions entre scènes. La pondération TF$\cdot$IDF permet de donner un poids faible à ces locuteurs qui auront donc moins d'influence sur la distance entre deux séquences. Soit $\mathcal{S}$ la liste des séquences d'une vidéo. Chaque séquence $s \in \mathcal{S}$ est décrite par un vecteur $\text{TF} \cdot \text{IDF}^{\text{SD}}(s)$ de dimension q où q est le nombre total de locuteurs ℓ dans la vidéo et $\text{TF} \cdot \text{IDF}_\ell^{\text{SD}}(s) = \text{TF}_\ell(s) \times \text{IDF}_\ell$ pour $\ell \in [\![1, q]\!]$:
  – Le terme IDF est défini par $\text{IDF}_\ell = \log\left(|\mathcal{S}|/N_\ell\right)$ où $|\mathcal{S}|$ est le nombre de séquences dans la vidéo et N_ℓ est le nombre de séquences durant lesquelles le locuteur ℓ parle.
  – Le terme TF est défini par $\text{TF}_\ell(s) = \mathcal{D}_\ell(s)/\mathcal{D}(s)$ où $\mathcal{D}(s)$ est la durée de la séquence s et $\mathcal{D}_\ell(s)$ est le temps cumulé de parole du locuteur ℓ dans la séquence s.

  La distance $d^{\text{TFIDF\_SD}}(s_i, s_j)$ entre les séquences s_i et s_j est définie comme la distance cosinus entre leurs vecteurs TF $\cdot$ IDF respectifs :

$$d^{\text{TFIDF\_SD}}(s_i, s_j) = d_{\cos}(\text{TF} \cdot \text{IDF}^{\text{SD}}(s_i), \text{TF} \cdot \text{IDF}^{\text{SD}}(s_j)) \tag{3.5}$$

  avec $\text{TF} \cdot \text{IDF}^{\text{SD}}(s)$ le vecteur TF $\cdot$ IDF de locuteurs associé à la séquence s.

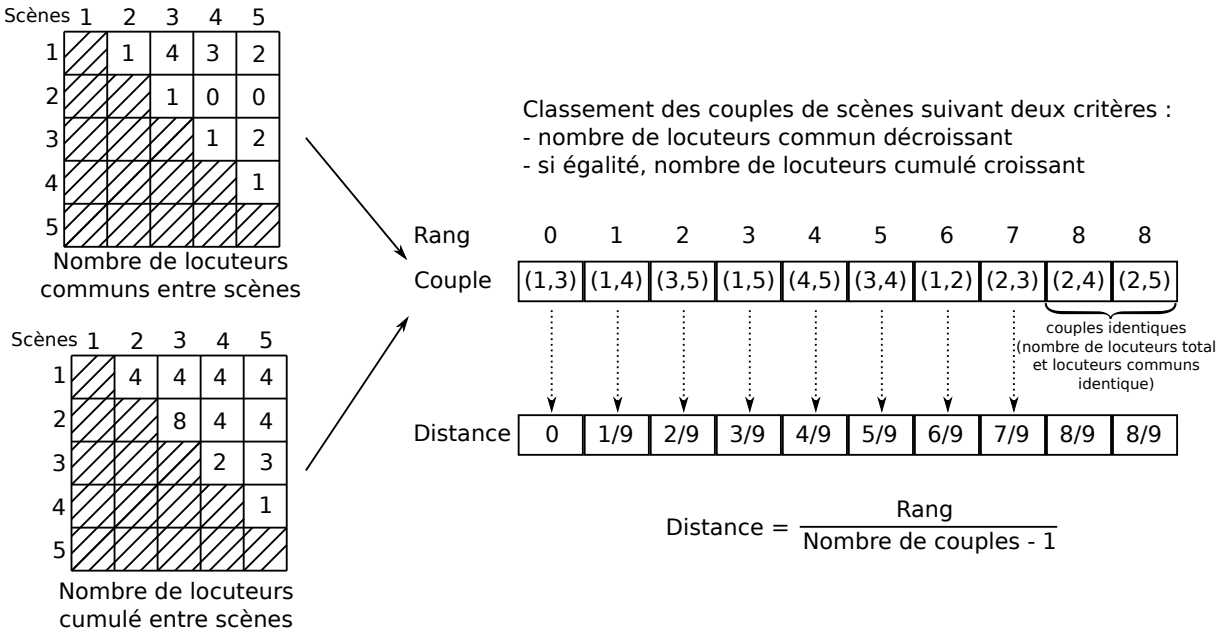

FIGURE *3.12 – Exemple de calcul de la distance d^{SD} pour 5 scènes.*

- La troisième distance d^{SD} entre deux séquences est basée sur leur nombre de locuteurs communs. Son calcul est détaillé dans la Figure 3.12. Pour chaque couple de séquences (i, j) (composé des séquences s_i et s_j), le nombre de locuteurs communs est comptabilisé. Ces couples sont triés par ordre décroissant de locuteurs communs entre les deux séquences. Ceux ayant le même nombre de locuteurs en commun sont triés par ordre croissant du nombre cumulé de locuteurs parlant dans les deux séquences. On note R_{ij} le rang du couple (i, j) dans ce tri. La distance $d^{\text{SD}}(s_i, s_j)$ entre les séquences s_i et s_j est définie par R_{ij} normalisé par le nombre de scènes N :

$$d^{\text{SD}}(s_i, s_j) = 2 \cdot \frac{R_{ij}}{N(N-1)} \tag{3.6}$$

  Cette distance a été spécifiquement développée pour le regroupement des scènes en histoires. L'idée est de considérer que deux séquences ayant plus de locuteurs en commun que deux autres doivent avoir une distance plus petite. Cependant, si deux séquences ont la totalité de leurs locuteurs en commun, nous considérons qu'elles sont plus similaires que deux séquences ayant seulement une portion des locuteurs en commun, et ce, même si le nombre de locuteurs communs est le même dans les deux cas.
  Cette distance basée sur un classement permet de tenir compte de ces deux critères à la fois. Elle est particulièrement pertinente pour un regroupement des scènes de type agglomératif, puisque les premières itérations du regroupement vont regrouper les séquences ayant beaucoup de locuteurs en commun, privilégiant la pureté des groupes de scènes si l'observation 5 est correcte.

Les histogrammes de la Figure 3.13 illustrent l'utilisation de la distance d^{SD} pour les tâches de segmentation en scènes et de regroupement des scènes en histoires.

Dans les valeurs extrêmes, la distribution des distances entre scènes d'une même histoire et celle entre scènes d'histoires différentes sont disjointes. Cependant une grande incertitude subsiste pour les valeurs intermédiaires. Ainsi, les distances très faibles nous assurent que deux scènes font partie d'une même histoire et les distances très grandes que deux scènes appartiennent à deux histoires différentes, validant l'observation 5 (page 85). Cependant, comme les distributions ne sont pas clairement disjointes, les distances basées sur le descripteur SD ne suffisent peut-être pas pour regrouper les scènes en histoires.

60% des plans de notre corpus sont décrits par un seul locuteur. Dans 16% des plans aucun locuteur n'est présent. Ainsi, dans le cas des distances $d^{\alpha\text{SD}}$ et $d^{\text{TFIDF\_SD}}$ les distances entre plans sont soit très petites (distance de 0 pour deux plans décrits par le même locuteur), soit très grandes (distance de 1 pour deux plans décrits par deux locuteurs différents). Il est donc possible que les distances proposées à partir des locuteurs ne soient pas pertinentes pour la segmentation en scènes.

Cependant, les faibles distances sont en majorité des distances entre plans d'une même scène et les grandes distances des distances entre plans de scènes différentes. L'utilisation des distances SD pour la segmentation en scènes et le regroupement des scènes en histoires est discuté dans les Chapitres 4 et 5.

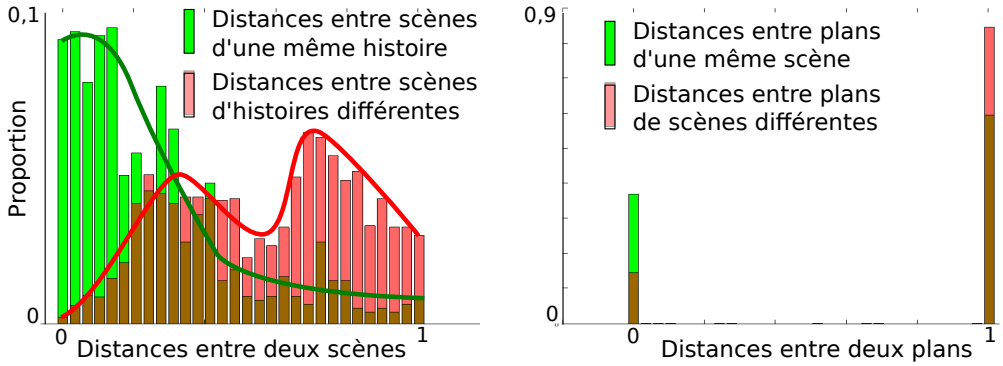

FIGURE *3.13 – Répartition des distances d^{SD} entre séquences de vidéo. En vert, répartition des distances des scènes appartenant à la même histoire (à gauche) ou des plans appartenant à la même scène (à droite). En rouge, répartition des distances des scènes appartenant à deux histoires différentes (à gauche) ou des plans appartenant à des scènes différentes (à droite).*

3.3.2 Limites des systèmes de segmentation et regroupement en locuteurs

Nous nous reposons sur le système de segmentation et regroupement en locuteurs décrit dans [Barras 2006]. Il permet de retrouver les tours de parole des locuteurs, et d'associer un identifiant unique à tous les tours de parole du même locuteur. Comme pour la distance d^{ASR}, l'utilisation des locuteurs pour mesurer une distance entre deux séquences est limitée par la qualité de la méthode utilisée pour segmenter et regrouper en locuteurs.

Pour mesurer les erreurs produites par un tel système sur des épisodes de séries télévisées, une annotation précise des tours de parole pour les quatre premiers épisodes de la série *Ally McBeal* a été effectuée. Plusieurs valeurs ont été mesurées pour calculer la qualité de la segmentation et du regroupement en locuteurs :

- **La pureté** qui mesure à quel point les tours de paroles associés à un même locuteur sont homogènes.
- **La couverture** qui mesure à quel point tous les tours de parole d'un même locuteur de la vérité-terrain ont été assignés à un même locuteur par le système automatique.
- **Le DER** (Diarization Error Rate) [Fiscus 2004] qui mesure la performance globale (taux d'erreur) d'un système de segmentation et regroupement en locuteurs. Son comportement est étudié en détail au Chapitre 5, page 125.

Les résultats obtenus par le système de segmentation et regroupement en locuteurs sur les quatre épisodes annotés sont résumés dans le Tableau 3.2.

	DER	**pureté**	**couverture**
épisode 1	78%	79%	57%
épisode 2	74%	67%	43%
épisode 3	63%	80%	55%
épisode 4	67%	80%	52%

TABLE *3.2 – Évaluation de la segmentation et regroupement en locuteurs pour quatre épisodes de la série* Ally McBeal.

Ces résultats montrent un taux d'erreur moyen très important. Cependant, la pureté est correcte (entre 67% et 80%), pour une couverture basse (entre 43% et 57%). Ce qui signifie qu'il y a sur-segmentation des locuteurs : les locuteurs de la vérité-terrain sont « découpés » en plusieurs « sous-locuteurs ».

Le système de segmentation et regroupement en locuteur utilise une première étape de détection de parole/non parole. Elle permet de déterminer quels segments du flux audio contiennent de la parole ou non. Cependant, cette étape est aussi source d'erreurs, et le taux d'erreur pour la détection de parole/non parole est en moyenne de 26%, avec 11% de fausses alarmes (segments de vidéo ne contenant pas de parole mais détectés comme tel) et 15% de détection ratée sur l'ensemble des quatre épisodes.

Comme le système utilisé pour la transcription automatique de la parole, le système de segmentation et regroupement en locuteurs que nous avons utilisé [Barras 2006] a été optimisé pour des journaux télévisés ou provenant de la radio. Ainsi, ses performances sur des épisodes de séries télévisées sont détériorées à cause des mêmes facteurs que ceux détaillés pour le système de transcription.

Chapitre 4

Segmentation en scènes

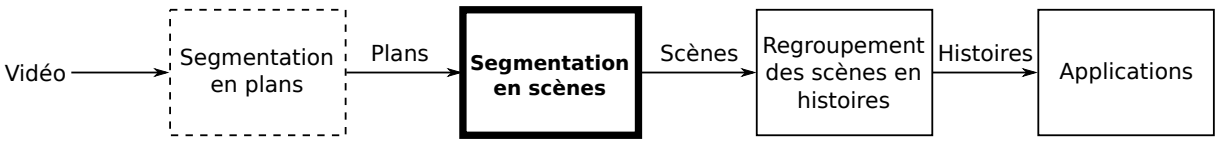

FIGURE *4.1 – Différents niveaux de structuration d'une vidéo : focus sur la segmentation en scène.*

Comme illustré dans la Figure 4.1, le niveau de structuration d'un épisode de série télévisée étudié dans ce chapitre concerne la segmentation en scènes. La définition d'une scène détaillée au Chapitre 1 (page 11) est la suivante :

Définition d'une scène
- Une scène est une suite de plans consécutifs.
- Une scène décrit un unique évènement (ou aucun évènement).
- La scène respecte une continuité temporelle.
- La scène respecte une continuité spatiale dictée par des règles de montage précises.

Cette définition explique qu'une scène est une suite de plans consécutifs. Ainsi, une frontière entre deux scènes est aussi une transition entre deux plans. La frontière f entre les scènes s_i et s_{i+1} est donc la frontière entre le dernier plan de s_i et le premier plan de s_{i+1}. La tâche de segmentation en scènes peut donc être définie comme une tâche de classification de frontières de plans.

Tâche de segmentation en scènes
Soit $\mathcal{F}$ l'ensemble des frontières de plans de la vidéo. La tâche de segmentation en scènes est une tâche de classification $\mathbb{C}$ des frontières de plans $f \in \mathcal{F}$ telle que

$$\begin{aligned} \mathbb{C} : \quad \mathcal{F} &\rightarrow \{0,1\} \\ f &\rightarrow \begin{cases} 1 & \text{si } f \text{ est une frontière entre deux scènes} \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{4.1}$$

La majorité des méthodes présentées dans l'état de l'art (Section 2.3, page 50) se focalisent sur la composante vidéo des documents audiovisuels pour réaliser la segmentation en scènes. Parmi celles qui emploient la composante audio, peu d'entre elles utilisent la connaissance sur la présence des locuteurs pour décrire des segments audiovisuels dans le but de réaliser la segmentation en scènes.

Ainsi, l'observation 1 énoncée au chapitre précédent et qui définit la couleur comme une information pertinente pour la segmentation en scènes a déjà été étudiée. Cependant, la définition de la scène est fortement liée à la présence des personnages, puisqu'elle considère qu'une scène décrit un événement unique, qui est lui même lié aux interactions des personnages avec leur environnement. Ainsi, en se basant sur l'observation 4 présentée au chapitre précédent, qui stipule que *les personnages présents dans deux scènes consécutives sont généralement différents*, il doit être possible d'utiliser l'information fournie sur les personnages pour réaliser une segmentation en scènes.

Notre contribution consiste à étudier l'information fournie par un système de segmentation et regroupement en locuteurs pour développer des méthodes de segmentation en scènes. Un tel système fournit une information sur les locuteurs (personnages parlant) de façon à savoir si c'est le même locuteur qui parle dans 2 segments de parole, appelés *Tours de parole* (To).

Ce chapitre est découpé en quatre sections :

- La première section présente le protocole expérimental utilisé pour valider nos expériences.
- La seconde section étudie l'utilisation des tours de parole détectés par un système de segmentation et regroupement en locuteurs pour la segmentation en scènes (segmentation monomodale).
- La troisième s'intéresse à la fusion de segmentations produites par la méthode basée sur les tours de parole, et une méthode de segmentation basée sur des histogrammes de couleur [Yeung 1998].
- Dans la dernière section, nous étudions l'amélioration d'une méthode de segmentation de l'état de l'art [Sidiropoulos 2011] en utilisant les tours de parole des locuteurs.

4.1 Protocole expérimental

4.1.1 Corpus et annotations manuelles

Pour évaluer les algorithmes de segmentation en scènes, nous avons annoté en scènes 22 épisodes de séries télévisées : 7 épisodes de la première saison de la série *Ally McBeal*, 7 épisodes de la saison 4 de *Malcolm* et 8 épisodes de la première saison de la série *Le Trône de Fer*.

Ces trois séries ont été sélectionnées pour leurs styles très différents (nombre de plans, de scènes ou de personnages très différents) résumés dans le Tableau 5.2. Les plans et scènes ont été annotés suivant les définitions du Chapitre 1 (page 11). La Figure 4.2 illustre les annotations sur le début d'un épisode de la série *Ally McBeal*.

Les méthodes de segmentation décrites dans ce chapitre utilisent les tours de parole des locuteurs détectés automatiquement. Afin d'étudier l'impact des erreurs de ce système, les tours de parole des locuteurs pour les 4 premiers épisodes de la série *Ally McBeal* ont été annotés manuellement.

			Annotation manuelle		
Série	**Nombre d'épisodes**	**Durée totale** (par épisode)	**Nombre de plans total** (par épisode)	**Nombre de scènes total** (par épisode)	**Nombre de personnages** par épisode
Ally McBeal	7	5h (≈41min)	4872 (696 plans)	304 (43 scènes)	15
Malcolm	7	2,5h (≈22min)	3325 (475 plans)	196 (28 scènes)	19
Le Trône de Fer	8	7h (≈50min)	8294 (1037 plans)	244 (30 scènes)	34

TABLE *4.1 – Description des annotations du corpus utilisé.*

FIGURE *4.2 – Annotation des frontières entre scènes pour le début d'un épisode d'*Ally McBeal.

4.1.2 Métriques d'évaluation

Dans la Section 2.3.4, plusieurs méthodes d'évaluation pour la segmentation en scènes sont étudiées. La tâche de segmentation en scènes est décrite comme une tâche de classification des frontières de plans. Ainsi, pour une approche de segmentation en scènes $\mathbb{C}$, chaque frontière de plans $f \in \mathcal{F}$ peut prendre comme valeur $\mathbb{C}(f) \in \{0, 1\}$, avec

- 0 : la frontière de plan n'est pas une frontière entre deux scènes
- 1 : la frontière de plan est une frontière entre deux scènes.

L'évaluation de la qualité de la segmentation est obtenue en calculant les valeurs de **précision** (P), **rappel** (R) et **F-Mesure** ($\mathrm{F_{PR}}$) en comparant une classification automatique des frontières $\mathbb{C}$ et une classification manuelle $\mathbb{M}$.

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \\ \mathrm{R} &= \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \\ \mathrm{F_{PR}} &= 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \end{aligned}$$

Dans le but de fournir une évaluation objective de nos résultats et pour que la métrique d'évaluation soit adaptée à l'optimisation des paramètres des méthodes étudiées, le calcul des *Vrais Positifs* (VP), *Faux Positifs* (FP) et *Faux Négatifs* (FN) peut être fait de trois façons différentes.

Évaluation stricte $\mathcal{L}^{strict}$

Une évaluation stricte considère qu'une frontière entre deux scènes est correctement détectée si une frontière manuellement annotée se trouve exactement à la même position. Ainsi, le comptage des VP, FP et FN suit les règles suivantes :

$$\begin{aligned} \mathrm{VP} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{C}(f) = 1 \wedge \mathbb{M}(f) = 1\} \\ \mathrm{FP} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{C}(f) = 1 \wedge \mathbb{M}(f) = 0\} \\ \mathrm{FN} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{C}(f) = 0 \wedge \mathbb{M}(f) = 1\} \end{aligned}$$

Évaluation avec tolérance $\mathcal{L}^{tol}$

Une évaluation avec tolérance suggère qu'une frontière entre deux scènes est correctement détectée si elle se situe à moins de τ secondes d'une frontière annotée. Dans ce cas, le comptage des VP, FP et FN s'effectue comme suit :

$$\begin{aligned} \mathrm{VP} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{M}(f) = 1 \wedge \exists\ f' \in \mathcal{F} / \mathbb{C}(f') = 1 \text{ avec } |\texttt{temps}(f) - \texttt{temps}(f')| < \tau\} \\ \mathrm{FP} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{C}(f) = 1\} - \mathrm{VP} \\ \mathrm{FN} &= \mathrm{card}\{f \in \mathcal{F} / \mathbb{M}(f) = 1 \wedge \nexists\ f' \in \mathcal{F} / \mathbb{C}(f') = 1 \text{ avec } |\texttt{temps}(f) - \texttt{temps}(f')| < \tau\} \end{aligned}$$

avec $\texttt{temps}(f)$ l'instant de la frontière f dans la vidéo.

Évaluation basée sur la position des tours de parole $\mathcal{L}^{To}$

Une des approches de segmentation proposée dans ce chapitre utilise uniquement les tours de parole des locuteurs pour réaliser la segmentation en scènes. Cette approche n'est pas basée sur les frontières de plans et il existe de nombreux cas, illustrés dans la Figure 4.3 où la seule connaissance des tours de parole des locuteurs ne permet pas de retrouver précisément la position de la frontière entre les scènes.

Cet exemple montre plusieurs scènes et les tours de parole des locuteurs présents dans ces scènes. La seule connaissance de ces tours de parole permet de déterminer intuitivement qu'une frontière entre deux scènes est présente entre les tours de parole To_3 et To_4. Cependant, comme plusieurs frontières de plans sont présentes entre ces deux tours de parole, il est impossible de déterminer la position exacte de la frontière entre les deux scènes.

En considérant que les frontières f_3, f_4 et f_5 sont des frontières correctes entre scènes , il est impossible d'utiliser l'évaluation $\mathcal{L}^{strict}$ pour évaluer la pertinence de la segmentation. Ainsi, le comptage des VP et FP nécessaires au calcul des précision, rappel et F-Mesure respecte les règles suivantes :

- VP = nombre de frontières détectées dans une zone sans parole dans laquelle se situe une frontière de la référence (zones d'incertitude dans la Figure 4.3).
- FP = nombre de frontières détectées qui ne sont pas des VP.

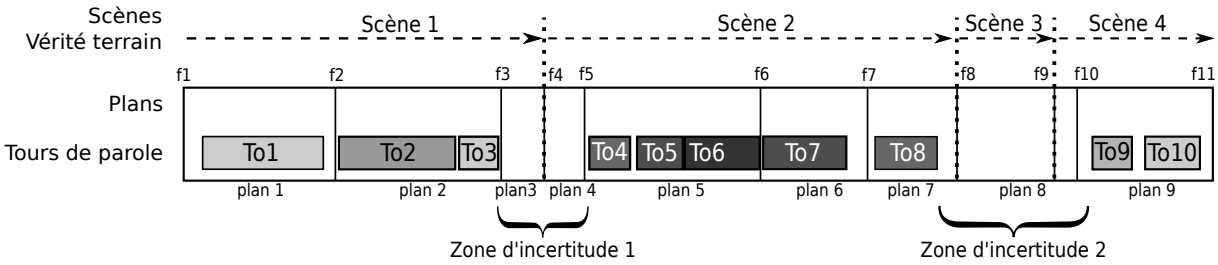

FIGURE 4.3 – f_1, f_2, ... *sont des frontières entre plans.* To_1 *à* To_{10} *sont les tours de parole des locuteurs. Dans cet exemple, la seule connaissance des locuteurs ne permet pas de déterminer quelle frontière de plans est une frontière de scènes. Il est possible de déterminer qu'une frontière entre deux scènes se trouve entre les tours de parole* To_3 *et* To_4 *ou entre* To_8 *et* To_9*. La zone d'incertitude correspond à la zone sans parole entre ces tours de parole. Il est impossible de déterminer laquelle des frontières* f_3, f_4 *ou* f_5 *est la frontière de scène dans la zone d'incertitude 1.*

4.1.3 Validation croisée

Les algorithmes proposés nécessitent d'optimiser un certain nombre de paramètres. Puisque nous n'avons que 22 épisodes annotés en scènes, la quantité de données est insuffisante pour constituer un ensemble d'apprentissage et un ensemble de test. C'est pourquoi le protocole d'évaluation suit le principe de la validation croisée (*leave-one-out cross validation*) expliqué ci-dessous.

Pour un épisode donné, les paramètres sont déterminés de façon à maximiser la métrique d'évaluation globale sur l'ensemble des autres épisodes du corpus. Soient :

- $\mathcal{E}$ l'ensemble des épisodes de test,
- $\mathbb{M}$ la référence,
- $\mathbb{C}$ un algorithme de segmentation,
- Λ l'espace de recherche des paramètres,
- $\mathcal{L}$ la métrique d'évaluation,

Les paramètres de segmentation $\lambda(e)$ pour un épisode e de $\mathcal{E}$ sont déterminés tel que

$$\lambda(e) = \underset{\lambda \in \Lambda}{\operatorname{argmax}} \frac{1}{|\mathcal{E}| - 1} \sum_{e' \in \mathcal{E} \setminus e} \mathcal{L}(\mathbb{M}(e'), \mathbb{C}_\lambda(e')) \tag{4.2}$$

Ainsi, la performance globale $\mathfrak{L}(\mathbb{C})$ de l'algorithme $\mathbb{C}$ est calculée telle que

$$\mathfrak{L}(\mathbb{C}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{L}(\mathbb{M}(e), \mathbb{C}_{\lambda(e)}(e)) \tag{4.3}$$

4.2 Approche monomodale basée sur les tours de parole

La majorité des méthodes de segmentation de l'état de l'art utilisent les plans comme unité de base du système de segmentation à partir desquels les frontières sont retrouvées. Décrire les plans à partir des locuteurs qui y sont présents est problématique. En effet, un plan est un segment très court de vidéo (3.2 secondes en moyenne dans notre corpus). Ainsi, 60% des plans de notre corpus n'incluent qu'un seul locuteur, et 16 % n'en incluent aucun.

Les méthodes de segmentation en scènes de l'état de l'art consistent à regrouper entre eux les plans similaires. Or, dans des séries télévisées comme *Ally McBeal*, un personnage comme *Ally* est présent dans la majorité des scènes.

La Figure 4.5 montre la distribution des distances entre les plans mesurées avec le calcul de distance $d^{\text{TFIDF\_SD}}$ (définie par l'équation 3.4 page 87). Les plans sont généralement soit complètement différents (distance = 1) soit identiques (distance = 0). La distribution des distances entre plans appartenant à une même scène et celle des distances entre plans de scènes différentes se recouvrent fortement et il y a peu de chances qu'elles permettent d'obtenir une segmentation en scènes correcte.

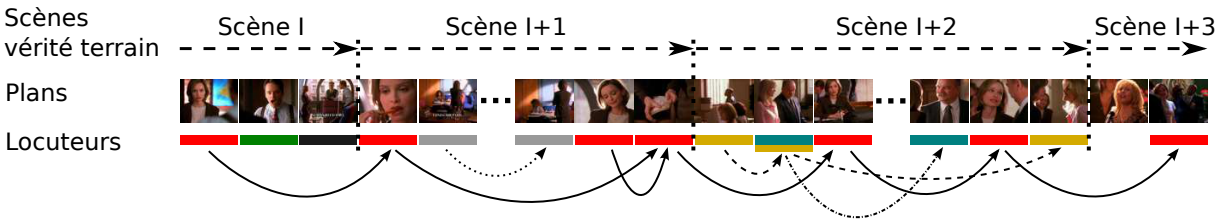

FIGURE 4.4 – *Exemple de segmentation en scènes à partir d'une description des plans par les locuteurs. Les flèches relient des plans similaires devant être regroupés. Chaque couleur représente un locuteur unique. Dans cet exemple, tous les plans sont chevauchés par au moins un lien, il est donc impossible de retrouver les frontières entre les scènes.*

Pour contrecarrer ce problème, nous proposons une méthode de segmentation basée sur la comparaison de segments dont la durée est plus grande que la durée d'un plan dans le but d'inclure plusieurs locuteurs dans chaque segment. Cette méthode est inspirée de la méthode de segmentation proposée par Kender et Yeo [Kender 1998] puisqu'elle est basée sur une mesure de cohérence de segments de vidéo successifs.

Le principe consiste à détecter les frontières à partir d'une fenêtre glissante, en considérant qu'il y a une frontière lorsque la fenêtre n'est plus cohérente avec la fenêtre initiale. Nous appelons cette méthode : *Méthode de Segmentation par Fenêtre Glissante* (MFG). La cohérence entre les fenêtres est déterminée à partir de la mesure de distance $d_{\alpha}^{\alpha\mathrm{SD}}$ (définie par l'équation 3.4 page 87), en considérant que deux segments sont cohérents si leur distance est inférieure à un seuil θ.

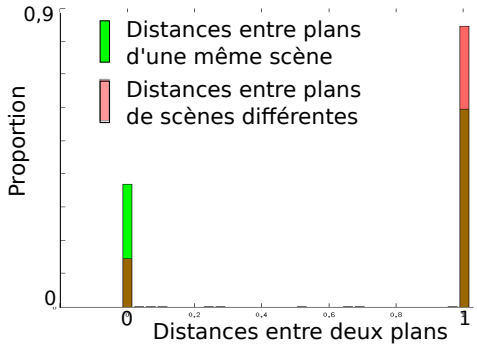

FIGURE 4.5 – *Répartition des distances $d^{\alpha SD}$ entre plans. En vert, répartition des distances entre plans appartenant à la même scène. En rouge, répartition des distances entre plans appartenant à des scènes différentes.*

4.2.1 Approche par fenêtre glissante

Soit une fenêtre glissante de longueur T et commençant au temps t de la vidéo à segmenter. La méthode de segmentation en scènes proposée est énoncée par l'Algorithme 4, et illustrée par la Figure 4.6.

Soit F_0 une fenêtre de référence débutant à l'instant t_0 de la vidéo et de longueur T, et $\texttt{temps}(f)$ la fonction qui retourne la date de la frontière $f \in \mathcal{F}$ dans la vidéo. L'algorithme consiste à déplacer séquentiellement une fenêtre F_i d'un pas de δ secondes jusqu'à ce que la distance $d_\alpha^{\alpha\text{SD}}$ entre F_0 et F_i dépasse un seuil θ. Dans ce cas, la fenêtre F_i n'est plus cohérente avec la fenêtre F_0. Une nouvelle frontière de scène est ajoutée à la liste des frontières. La frontière de scène retenue est la frontière de plan $f_j \in \mathcal{F}$ la plus proche du milieu de la fenêtre F_i. Une nouvelle fenêtre débutant à l'instant $\texttt{temps}(f_j)$ devient la nouvelle fenêtre de référence, et ce processus est répété jusqu'à la fin de la vidéo. Le comportement de cet algorithme dépend de la valeur des seuils suivants :

- **la longueur T de la fenêtre glissante** : en fonction de sa valeur il peut y avoir un délai avant qu'une frontière de scène soit détectée. Pour corriger ce problème, toute frontière détectée durant un tour de parole est déplacée soit au début soit à la fin (en fonction du plus proche) de ce tour de parole.
- **le seuil θ** a un impact direct sur le nombre de frontières de scènes détectées.
- **la pondération des locuteurs α** : le type de pondération à appliquer lors du calcul de la distance $d_\alpha^{\alpha\text{SD}}$ influe sur le résultat de la segmentation.

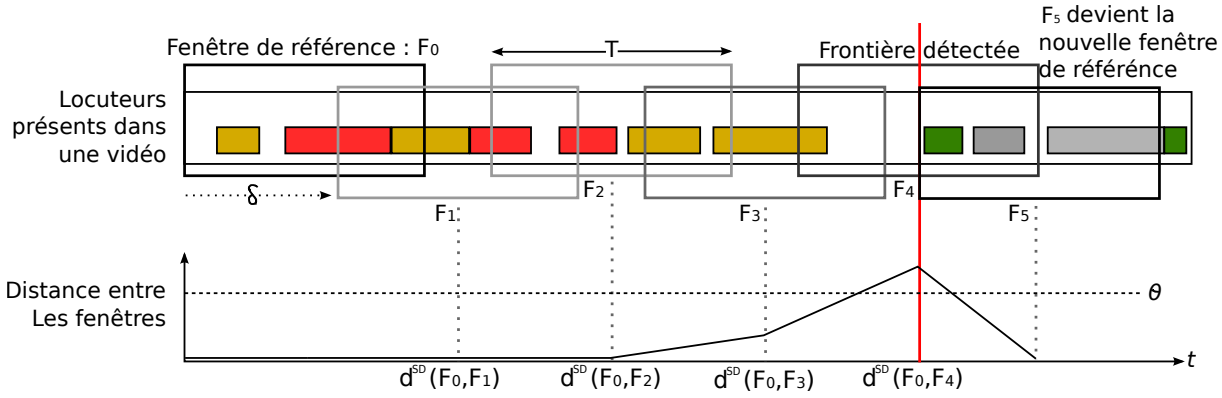

FIGURE *4.6 – Algorithme de segmentation en scènes par fenêtre glissante.*

Algorithme 4 Segmentation en scènes par fenêtre glissante.

Entrées: T, θ

$L \Leftarrow \emptyset$
$t_0 \Leftarrow 0$
$t \Leftarrow \delta$
tantque $t + T <$ durée de l'épisode **faire**
  si $d_\alpha^{\alpha\text{SD}}([t_0, t_0 + T], [t, t + T]) > \theta$ **alors**
    $f_p \Leftarrow \underset{f \in \mathcal{F}}{\text{argmin}} \, |(t + T/2) - \texttt{temps}(f)|$
    $L \Leftarrow L \cup \{f_p\}$
    $t_0 \Leftarrow \texttt{temps}(f_p)$
    $t \Leftarrow t_0$
  finsi
  $t \Leftarrow t + \delta$
fin tantque

Sorties: L

Optimisation des paramètres

On cherche à optimiser l'ensemble des paramètres $\lambda = (T, \theta, \alpha)$ de la méthode MFG. Ces paramètres sont les seuils T, θ et le type de pondération α. Il s'agit d'optimiser ces paramètres afin de minimiser l'erreur moyenne sur l'ensemble de test tel que présenté dans les formules 4.2 et 4.3. L'espace de recherche Λ de ces paramètres est défini tel que $\Lambda = \underbrace{\mathbb{R}^+}_{T} \times \underbrace{[0, 1]}_{\theta} \times \underbrace{\{\alpha^=, \alpha^+, \alpha^-\}}_{\alpha}$.

Comme seuls les tours de parole des locuteurs sont utilisés pour déterminer le vecteur descripteur des séquences de vidéo, la métrique d'évaluation utilisée pour optimiser les paramètres de la méthode MFG est la métrique $\mathcal{L}^{To}$.

Le paramètre δ a été fixé par expérimentation à 500 millisecondes. Cette durée est significativement plus courte que la durée d'un plan (3.2 secondes en moyenne), et suffisamment longue pour permettre un calcul rapide de l'algorithme.

Pour implémenter l'optimisation des paramètres, Λ doit être discrétisé. Le Tableau 4.2 décrit l'ensemble des valeurs utilisées pour l'optimisation des paramètres T et θ permettant d'obtenir les résultats proposés dans cette section.

	Min	**Max**	**Pas**
T (secondes)	2	100	1
θ	0.04	0.4	0.02

TABLE *4.2 – Ensemble des valeurs utilisées pour l'optimisation des paramètres T et θ.*

Segmentation aléatoire

Une segmentation aléatoire est utilisée comme référence de comparaison pour les résultats de notre méthode. La segmentation aléatoire consiste à positionner aléatoirement les frontières de scènes sur les frontières de plans de la vidéo. Pour simuler une sous-segmentation ou une sur-segmentation, le nombre de ces frontières est aléatoirement sélectionné entre 1 et $2 \times N$ avec N le nombre de frontières annotées de l'épisode.

La Figure 4.7 montre la répartition des F-Mesures obtenues pour 5000 segmentations aléatoires sur l'ensemble du corpus. Dans le cas d'une distribution normale, 99.9% des valeurs sont inférieures à la moyenne (μ) additionnée de 3 fois l'écart type (σ). Ainsi, les valeurs de Précision, Rappel et F-Mesure pour une segmentation aléatoire sont les moyennes des résultats obtenus par 5000 segmentations aléatoires différentes auxquelles nous avons additionné 3 fois l'écart type des 5000 résultats.

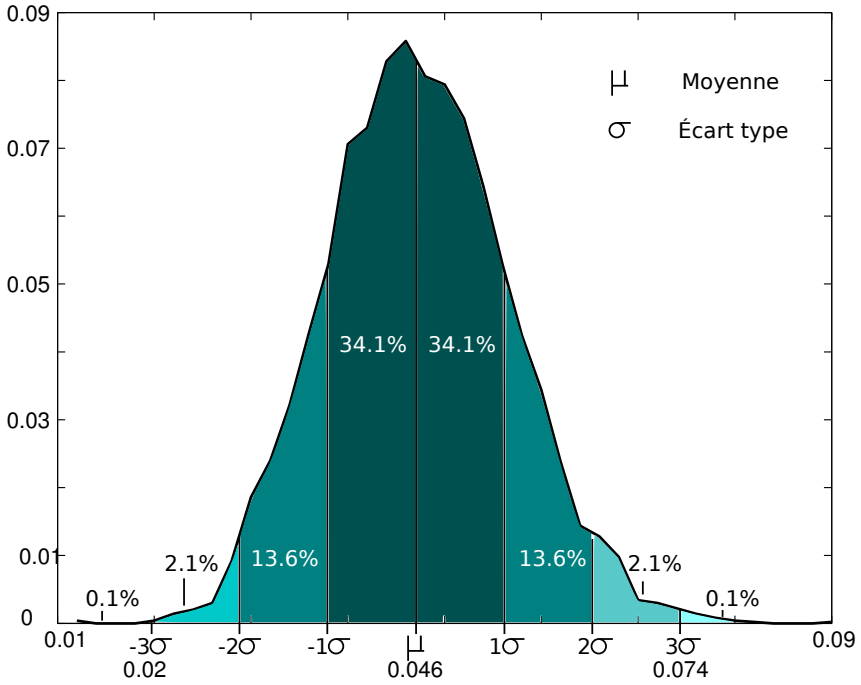

FIGURE *4.7 – Distribution des F-Mesures obtenues à partir d'une segmentation aléatoire.*

4.2.2 Intérêt de l'approche

Pour valider l'intérêt de l'approche par fenêtre glissante, nous étudions le comportement de la méthode MFG en utilisant des tours de parole des locuteurs parfaits. Pour ce faire, les tours de parole des quatre premiers épisodes de la série *Ally McBeal* ont été annotés manuellement. Les résultats présentés dans cette section ne tiennent compte que de ces 4 épisodes.

Résultats globaux

Les courbes présentées dans la Figure 4.8 montrent les résultats de F-Mesure, précision et rappel obtenus par la méthode MFG comparé à une segmentation aléatoire. Elles rendent compte de l'évolution de la métrique d'évaluation si l'on considère une tolérance τ variable en secondes sur la position de la frontière détectée par rapport à une frontière de la référence.

En augmentant cette tolérance, La F-Mesure de la méthode MFG augmente beaucoup plus rapidement que pour une segmentation aléatoire. C'est le cas autant pour la précision que pour le rappel jusqu'à une tolérance de 8 secondes. Cette forte augmentation du rappel pour une tolérance faible (environ 2 fois la durée moyenne d'un plan), suggère que la méthode MFG permet de retrouver des frontières de scènes proches de frontières annotées manuellement, mais décalées de quelques secondes.

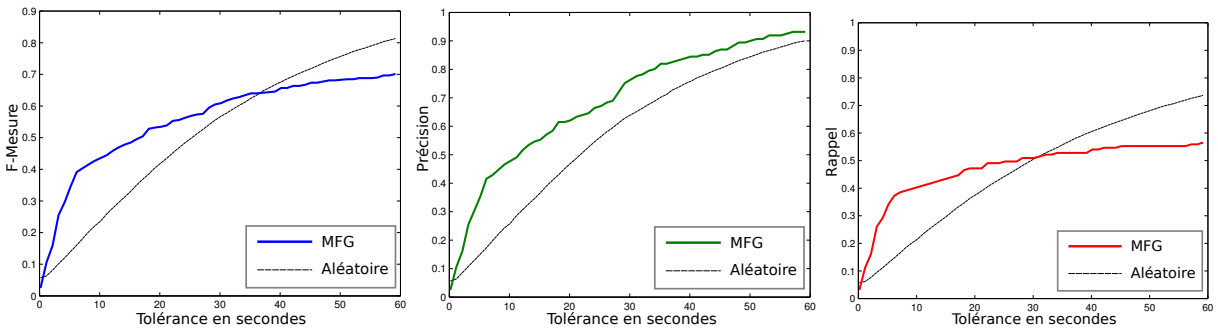

FIGURE *4.8 – Résultats de la méthode MFG en fonction d'une tolérance de durée variable, comparée à une segmentation en scènes aléatoire.*

Analyse détaillée du comportement de la méthode de segmentation

La Figure 4.11 (présente à la fin de cette section) montre le résultat de la segmentation pour la première moitié d'un épisode d'*Ally McBeal*. Les frontières de scènes détectées par l'algorithme MFG sont indiquées par les barres de couleur verticales (les Vrais Positifs en vert, les Faux Positifs en rouge et les Faux Négatifs en jaune). Les Figures 4.9 et 4.10 montrent en détail le comportement de l'algorithme aux environs de certaines de ces frontières.

Beaucoup d'erreurs sont dues à des insertions de frontières (par exemple les frontières $f4$, $f5$, $f11$ ou $f12$ de la Figure 4.11). La Figure 4.9 montre le comportement de l'algorithme MFG au niveau des transitions $f11$ et $f12$. Ces transitions apparaissent durant une scène composée d'un dialogue entre trois personnages. L'un de ces personnages n'intervient qu'au milieu de la scène, et son apparition provoque une coupure incorrecte (transition $f11$). La transition $f12$ est aussi incorrecte et elle est provoquée par le « départ » de deux personnages de la scène. En effet, la fenêtre est ici trop petite, et un long monologue d'un locuteur implique que seul ce locuteur est inclus dans la fenêtre.

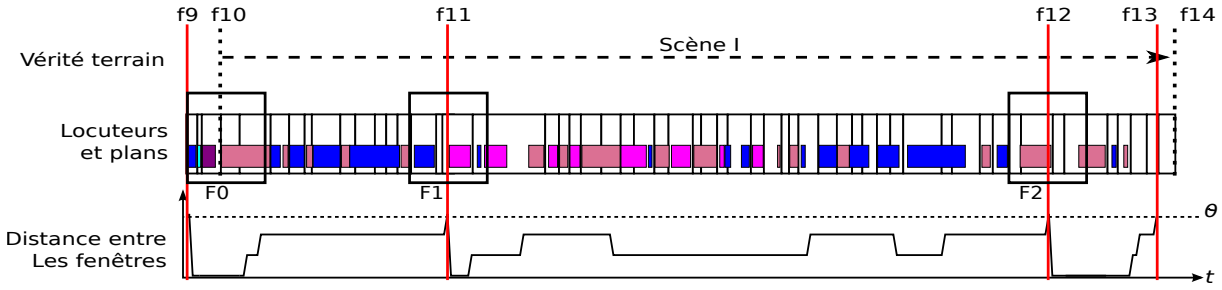

FIGURE *4.9 – Détail du comportement de la méthode de segmentation par fenêtre glissante pour les transitions $f11$ et $f12$. L'initialisation de la scène en $F0$ inclut plusieurs locuteurs de la scène $I-1$. La scène I montre un dialogue entre 3 personnages. L'un de ces personnages n'apparaît qu'au milieu de la scène. Son apparition (fenêtre $F1$) provoque une coupure (transition $f11$). Au niveau de la fenêtre $F2$, le locuteur rose parle seul. Les locuteurs bleus et fushia sont considérés comme sortis de la scène ce qui provoque une coupure (transition $f12$).*

La Figure 4.11 montre que certaines erreurs sont dues à la position incorrecte de nombreuses transitions détectées très proches (seulement quelques plans) d'une transition de référence. La Figure 4.10 permet d'illustrer pourquoi on observe un tel décalage pour beaucoup de transitions de scènes.

Dans cet exemple, la coupure est provoquée par l'arrivée du premier locuteur de la scène I+1 dans la fenêtre glissante, alors que l'un des locuteurs de la scène I ne s'y trouve plus. Les locuteurs décrits dans cette fenêtre ne sont plus cohérents avec la fenêtre d'origine, et à cause de la taille de la fenêtre glissante, la coupure est positionnée un plan avant la coupure réelle. Ce décalage est observable dans de nombreux autres exemples comme pour les transitions $f2$ et $f3$.

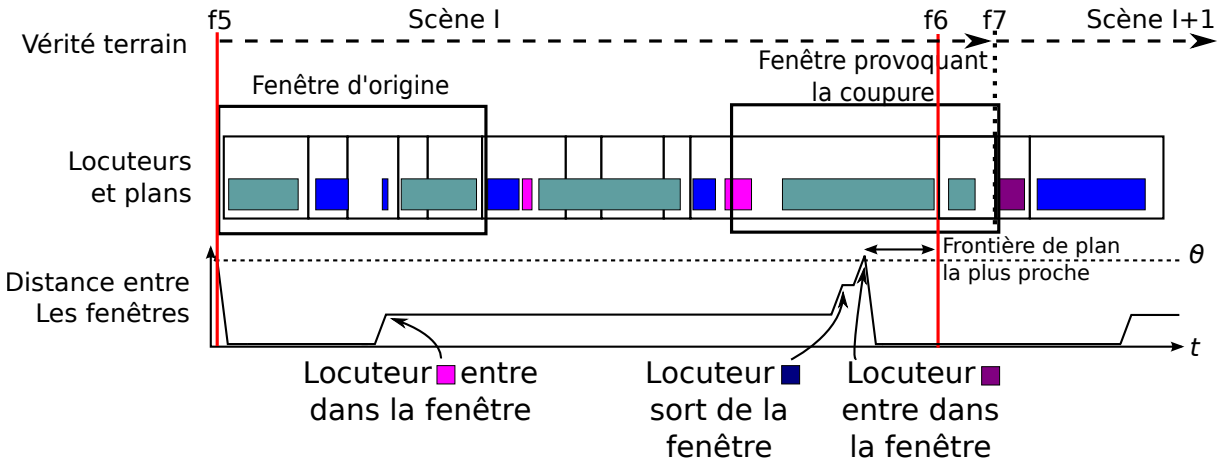

FIGURE *4.10 – Détail du comportement de la méthode de segmentation par fenêtre glissante pour les transitions $f6$ et $f7$.*

FIGURE *4.11 – Résultat d'une segmentation en scènes basée sur la méthode MFG pour un épisode de la série* Ally McBeal *à partir de tours de parole des locuteurs manuellement annotés.*

4.2.3 Résultats expérimentaux

Étiquetage manuel *vs.* automatique des tours de parole

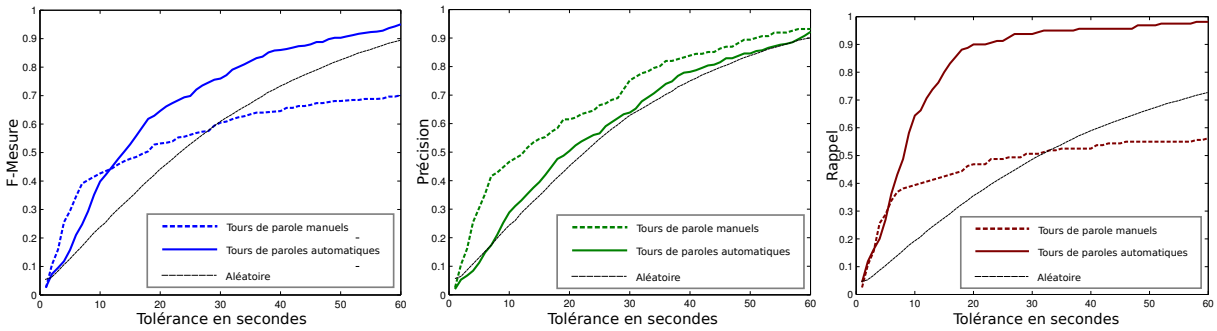

FIGURE *4.12 – Résultats de la méthode MFG en fonction d'une tolérance de durée variable. Comparaison de l'utilisation de tours de parole des locuteurs détectés automatiquement ou manuellement.*

Les tours de parole détectés automatiquement par un système de segmentation et regroupement en locuteurs comportent beaucoup d'erreurs. Pour étudier l'impact de ces erreurs sur la segmentation en scènes, la Figure 4.12 permet de comparer les résultats obtenus par la méthode de segmentation par fenêtre glissante, en utilisant des tours de parole détectés automatiquement ou manuellement, en entrée du système de segmentation. Dans les deux cas, seuls les 4 épisodes dont les tours de parole des locuteurs ont été annotés manuellement sont évalués.

Une annotation manuelle des locuteurs permet de retrouver un nombre de scènes proche du nombre de scènes annotées (165 frontières détectées pour 160 annotées), alors que la segmentation basée sur des locuteurs automatiques induit une forte sur-segmentation (361 frontières détectées). Ainsi, en considérant une tolérante τ dans l'évaluation, la segmentation obtenue à partir des tours de parole automatiques obtient un rappel qui augmente beaucoup plus rapidement que la segmentation aléatoire. Cependant, contrairement à l'utilisation des tours de parole manuels grâce auxquels le rappel augmente très peu pour $\tau > 8$ secondes, l'utilisation des tours de parole automatiques conduit à un rappel qui continue d'augmenter fortement pour $\tau > 8$ secondes, dû à la sur-segmentation de l'épisode.

La précision obtenue par une segmentation basée sur des tours de parole automatiques est légèrement supérieure à la précision aléatoire. Comme l'utilisation des tours de parole manuels conduit à une segmentation beaucoup plus précise que l'aléatoire, la F-Mesure est beaucoup plus forte avec un τ petit si l'on utilise les tours de parole manuels (0,40 à $\tau = 8$ secondes) plutôt que les tours de parole automatiques (0,25 à $\tau = 8$ secondes).

Pour un $\tau > 10$ secondes, la sur-segmentation introduite par l'utilisation de tours de parole automatiques implique une forte augmentation du rappel tout en conservant une

précision correcte. La F-Mesure augmente donc fortement et elle est bien au dessus de l'aléatoire.

Cette sur-segmentation peut s'expliquer par le fait que le nombre de locuteurs détectés automatiquement est supérieur au nombre réel de locuteurs présent dans les vidéos. Ainsi, il y a en moyenne 24 locuteurs détectés dans chaque épisode, pour 14 locuteurs annotés. Ces erreurs sur le regroupement des locuteurs induisent de grandes distances erronées entre les fenêtres, ce qui conduit à cette sur-segmentation.

Influence de la pondération α

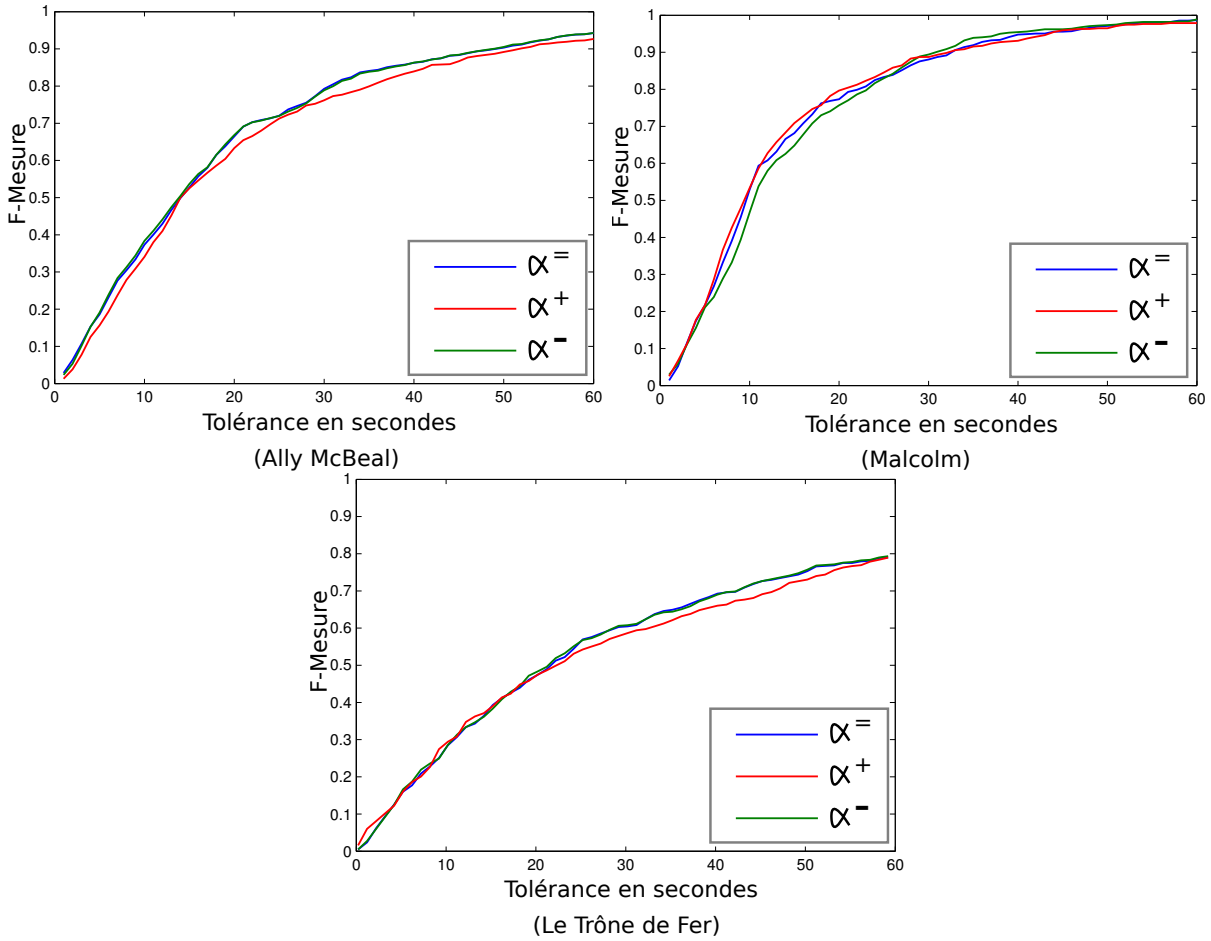

FIGURE *4.13 – Résultats de la méthode MFG en fonction d'une tolérance de durée variable. Comparaison des valeurs de pondération $\alpha^{=}$, α^{+} et α^{-} pour les trois séries du corpus.*

Un des paramètres de la méthode MFG à optimiser est la pondération α. En fonction de α, on considère que les locuteurs qui ont le plus de poids lors du calcul de la distance entre deux fenêtre sont :

- les locuteurs qui parlent le plus (α^{+}) ;
- les locuteurs qui parlent le moins (α^{-}) ;
- ou que tous les locuteurs ont le même poids ($\alpha^{=}$).

La Figure 4.13 permet de comparer les F-Mesures obtenues en utilisant ces trois types de pondérations sur les trois séries annotées. Les courbes illustrent la F-Mesure moyenne obtenue par tous les épisodes d'une même série en fonction d'une tolérance de τ secondes sur la position des frontières entre les scènes par rapport à une référence manuelle.

Pour la série *Ally McBeal*, en considérant une faible tolérance ($\tau < 10$ secondes), il y a peu de différences entre les segmentations utilisant une pondération $\alpha^=$ et α^-. Par contre, la courbe obtenue avec la pondération α^+ montre une F-Mesure toujours inférieure de 2 à 6% à celles relatives aux deux autres pondérations. Ce comportement peut s'expliquer par le format de cette série. En effet, un personnage particulier, *Ally*, est présent dans la majorité des scènes de chaque épisode. Une pondération α^- permet de lui donner moins de poids et de se focaliser sur les autres locuteurs pour la recherche des frontières entre les scènes.

Cependant, il est difficile de tirer des conclusions sur le comportement de la pondération α sur les autres séries. Les courbes de F-Mesures sont toutes très proches. Ainsi, cette pondération est peu pertinente pour la tâche de segmentation en scènes en utilisant la méthode de segmentation MFG.

C'est pourquoi sa valeur est fixée à $\alpha^=$ pour toutes les expériences qui seront présentées par la suite dans ce chapitre.

Taille de la fenêtre glissante T

	Malcolm	**Ally McBeal**	**Le Trône de Fer**
Taille optimale de T	23 sec	40 sec	52 sec
Durée moyenne d'une scène	45 sec	51 sec	1 min 25 sec

TABLE *4.3 – Taille optimale T de la fenêtre glissante.*

La taille $T \in \mathbb{R}^+$ optimale de la fenêtre glissante obtenue par validation croisée des épisodes est en moyenne de 40 secondes. Cependant, l'optimisation des paramètres sur un ensemble de test composé uniquement d'épisodes appartenant à une même série permet d'observer les différences sur la valeur optimale des paramètres spécifique à chaque série.

Ainsi, la taille de fenêtre optimale est de 40 secondes pour la série *Ally McBeal*, 23 secondes pour la série *Malcolm* et 52 secondes pour *le Trône de Fer*. Le Tableau 4.3 montre qu'il semble y avoir une corrélation entre la valeur de T et la durée moyenne des scènes.

Nous avons anticipé ce comportement puisque pour obtenir une granularité fine dans la détection des frontières entre scènes, il est nécessaire que la fenêtre utilisée soit plus petite que l'objet à détecter (ici, les scènes).

4.2.4 Conclusion

Les résultats présentés dans cette section montrent que la connaissance des tours de parole des locuteurs est une information pertinente pour la segmentation en scènes. La méthode de segmentation proposée obtient des résultats bien plus performants qu'une segmentation aléatoire. Cependant, la position des frontières entre scènes détectées manque de précision.

De plus, les erreurs introduites par l'utilisation de tours de parole automatiques en entrée du système de segmentation en scènes conduisent à de nombreuses erreurs de segmentation comme illustré par la Figure 4.12.

Ainsi, l'utilisation des tours de parole en entrée du système de segmentation, bien que pertinente, n'est pas suffisante. L'information visuelle ayant déjà prouvé son efficacité dans ce domaine, nous proposons donc deux méthodes de segmentation basées sur la fusion de l'information visuelle et des tours de parole des locuteurs :

- **fusion par alignement** : fusion du résultat d'une segmentation basée sur la couleur [Yeung 1998] et d'une segmentation MFG.
- **fusion dans le cadre du GSTG** : nous utilisons une approche de segmentation de l'état de l'art, le GSTG [Sidiropoulos 2011], avec laquelle nous testons la fusion des distances d^{HSV} et d^{SD}.

4.3 Approche multimodale par alignement de frontières

Afin d'améliorer les résultats précédents, nous proposons d'étudier une méthode de segmentation basée sur la fusion de la sortie du système de segmentation MFG avec une méthode de segmentation de l'état de l'art basée sur l'analyse de descripteurs visuels. Cette dernière est un méthode proposée par Yeung *et al.* [Yeung 1998] utilisant l'algorithme « Scene Transition Graph » (STG).

4.3.1 Segmentation STG à partir d'histogrammes de couleur

L'approche STG proposée par Yeung *et al.* [Yeung 1998] est détaillée dans la Section 2.3.3 (page 59). Il propose une segmentation basée sur une analyse de la couleur. La segmentation par STG fonctionne en quatre étapes (illustrées Figure 4.14).

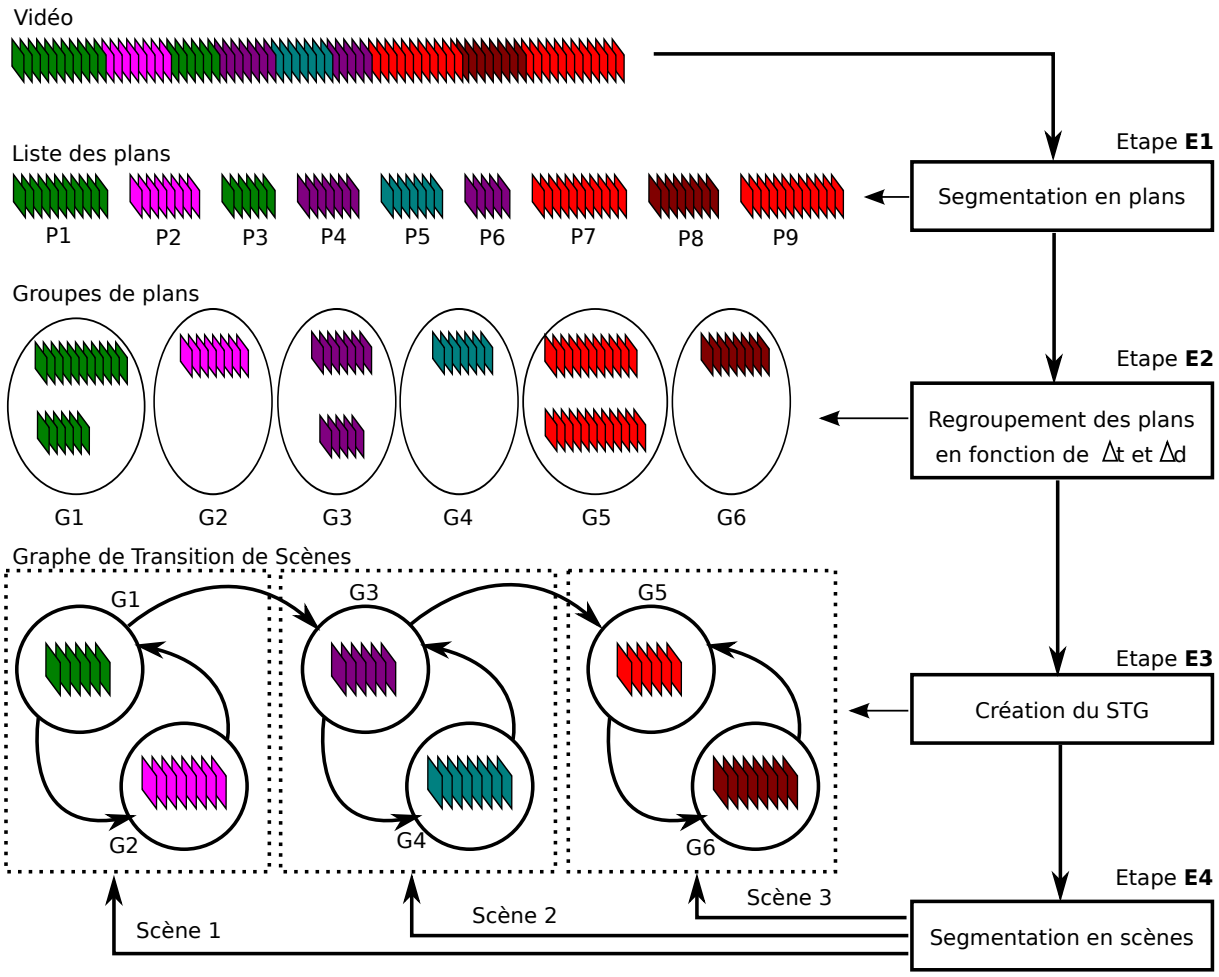

FIGURE *4.14 – Méthode de segmentation STG*

Après une étape de segmentation en plans (**E1**), on réalise une étape de regroupement des plans similaires (**E2**). La troisième étape consiste en la création d'un graphe à partir de ce regroupement (**E3**). Enfin une étape de segmentation par analyse du graphe permet de retrouver les frontières entre scènes (**E4**). Le regroupement des plans (étape **E2**) est réalisé par un regroupement hiérarchique agglomératif présenté Section 2.2.1 (page 45). La distance entre deux plans est mesurée par la distance d^{HSV} (équation 4.4).

Soient $\mathbf{H}_i$ et $\mathbf{H}_j$ l'ensemble des histogrammes de couleur extraits des images des plans p_i et p_j (une image par seconde), et $|\mathbf{H}_i|$ et $|\mathbf{H}_j|$ le nombre de ces histogrammes, la distance entre les plans p_i et p_j est définie telle que :

$$d^{\text{HSV}}(p_i, p_j) = \begin{cases} \dfrac{1}{|\mathbf{H}_i|} \displaystyle\sum_{h \in \mathbf{H}_i} \min_{g \in \mathbf{H}_j} d_{\text{L1}}(h, g) & \text{si } |\mathbf{H}_i| > |\mathbf{H}_j| \\ d^{\text{HSV}}(p_j, p_i) & \text{sinon} \end{cases} \tag{4.4}$$

Les plans sont regroupés séquentiellement suivant le modèle du regroupement hiérarchique agglomératif jusqu'à ce que la distance minimum entre deux groupes de plans dépasse un seuil Δ_d. Une contrainte supplémentaire interdit de regrouper deux plans dont la distance temporelle dans la vidéo est supérieure à un seuil Δ_t.

4.3.2 Description de l'approche de fusion

Notre méthode de fusion, décrite par l'Algorithme 9 et illustrée par la Figure 4.15, consiste à décaler chaque transition provenant d'une segmentation par fenêtre glissante (MFG) sur la transition la plus proche provenant d'une segmentation suivant la méthode de [Yeung 1998].

Algorithme 5 Segmentation en scènes basée sur la fusion des segmentations MFG et *STG*.

$L^{\text{F}} \Leftarrow \emptyset$
$L^{\text{MFG}} \Leftarrow$ transitions obtenues par la méthode MFG
$L^{\text{STG}} \Leftarrow$ transitions obtenues par la méthode [Yeung 1998]
pour Toutes les transitions f^{MFG} de L^{MFG} **faire**
  $f \Leftarrow \underset{f^{\text{STG}} \in L^{\text{STG}}}{\text{argmin}} |\texttt{temps}(f^{\text{MFG}}) - \texttt{temps}(f^{\text{STG}})|$
  si $f \notin L^{\text{F}}$ **alors**
    $L^{\text{F}} \leftarrow L^{\text{F}} \cup \{f\}$
  finsi
fin pour

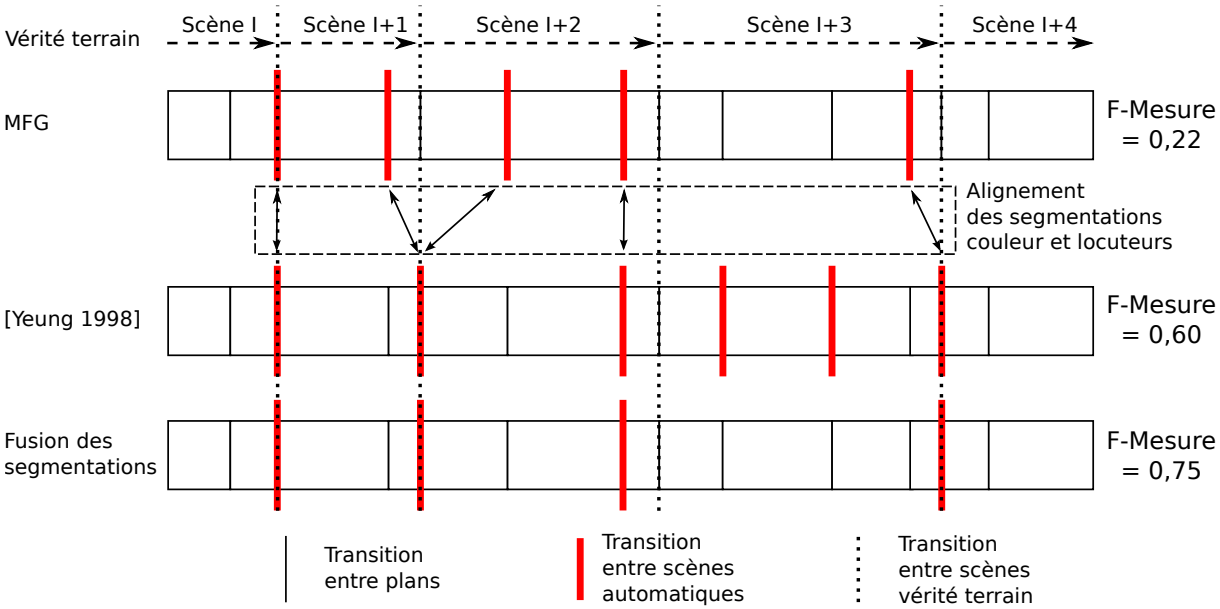

FIGURE *4.15 – Méthode de fusion de la segmentation MFG et la segmentation [Yeung 1998]. La fusion est effectuée par alignement des transitions détectées par une segmentation MFG avec les transitions détectées par une segmentation [Yeung 1998].*

Les deux algorithmes de segmentation ont plusieurs paramètres qui doivent être optimisés :

- **Paramètres pour la méthode de la fenêtre glissante**
  - taille de la fenêtre T
  - seuil de coupure θ
- **Paramètres pour la méthode [Yeung 1998]**
  - condition d'arrêt du regroupement Δ_d
  - distance maximum entre deux plans Δ_t

Optimisation des paramètres

On cherche à optimiser l'ensemble des paramètres $\lambda = \{T, \theta, \Delta_t, \Delta_d\}$ des deux méthodes de segmentation (MFG et [Yeung 1998]). L'espace de recherche Λ de ces paramètres est défini tel que $\Lambda = \Lambda^{\text{MFG}} \times \Lambda^{\text{STG}}$,

avec $\Lambda^{\text{MFG}} = \underbrace{\mathbb{R}^+}_{T} \times \underbrace{[0,1]}_{\theta}$ l'espace de recherche pour les paramètres de la méthode par fenêtre glissante,

et $\Lambda^{\text{STG}} = \underbrace{\mathbb{R}^+}_{\Delta_t} \times \underbrace{[0,1]}_{\Delta_d}$ l'espace de recherche pour les paramètres de la méthode de [Yeung 1998] (méthode STG).

	Paramètre	Min	Max	Pas
Λ^{MFG}	T (secondes)	2	100	1
	θ	0.04	0.4	0.02
Λ^{STG}	Δ_t (secondes)	5	200	5
	Δ_d	0.1	0.9	0.1

TABLE 4.4 – *Ensemble des valeurs utilisées pour l'optimisation des paramètres.*

Pour implémenter l'optimisation des paramètres, Λ doit être discrétisé. Le Tableau 4.4 décrit l'ensemble des valeurs utilisées pour l'optimisation des paramètres permettant d'obtenir les résultats proposés dans cette section.

Les paramètres de la méthode de segmentation STG sont optimisés de manière à maximiser la F-Mesure en utilisant une évaluation stricte ($\mathcal{L}^{strict}$). La méthode MFG se basant sur les tours de parole des locuteurs pour réaliser la segmentation en scènes, ses paramètres sont optimisés de manière à maximiser la F-Mesure obtenue à partir d'une évaluation $\mathcal{L}^{To}$.

Soit $\text{MFG}_\lambda(e)$ la segmentation obtenue par la méthode segmentation MFG en utilisant les paramètres $\lambda(e)$ pour un épisode e, $\text{STG}_\lambda(e)$ une segmentation obtenue par la méthode [Yeung 1998] et $\text{F}_\lambda(e)$ le résultat de la fusion des deux segmentations. Deux façons d'optimiser les paramètres λ sont explorées.

- Une optimisation séparée des paramètres :

$$\lambda^{\text{MFG}}(e) = \underset{\lambda\in\Lambda^{\text{MFG}}}{\text{argmax}} \frac{1}{|\mathcal{E}|-1} \sum_{e'\in\mathcal{E}\setminus e} \mathcal{L}^{To}(\mathbb{M}(e'), \text{MFG}_\lambda(e')) \tag{4.5}$$

$$\lambda^{\text{STG}}(e) = \underset{\lambda\in\Lambda^{\text{STG}}}{\text{argmax}} \frac{1}{|\mathcal{E}|-1} \sum_{e'\in\mathcal{E}\setminus e} \mathcal{L}^{strict}(\mathbb{M}(e'), \text{STG}_\lambda(e')) \tag{4.6}$$

  Dans ce cas, l'ensemble des paramètres $\lambda(e)$ est formé par la concaténation des paramètres $\lambda^{\text{MFG}}(e)$ et $\lambda^{\text{STG}}(e)$

- Une optimisation conjointe des paramètres :

$$\lambda(e) = \underset{\lambda\in\Lambda^{\text{MFG}}\times\Lambda^{\text{STG}}}{\text{argmax}} \frac{1}{|\mathcal{E}|-1} \sum_{e'\in\mathcal{E}\setminus e} \mathcal{L}^{strict}(\mathbb{M}(e'), \text{F}_\lambda(e')) \tag{4.7}$$

Dans les deux cas, la performance globale $\mathfrak{L}(F)$ de l'algorithme F est calculée telle que

$$\mathfrak{L}(\text{F}) = \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \mathcal{L}^{strict}(\mathbb{M}(e), F_{\lambda(e)}(e)) \tag{4.8}$$

avec $\mathbb{M}$ la référence et Λ l'espace de recherche des paramètres.

4.3.3 Résultats expérimentaux

4.3.3.1 À partir de tours de parole manuels

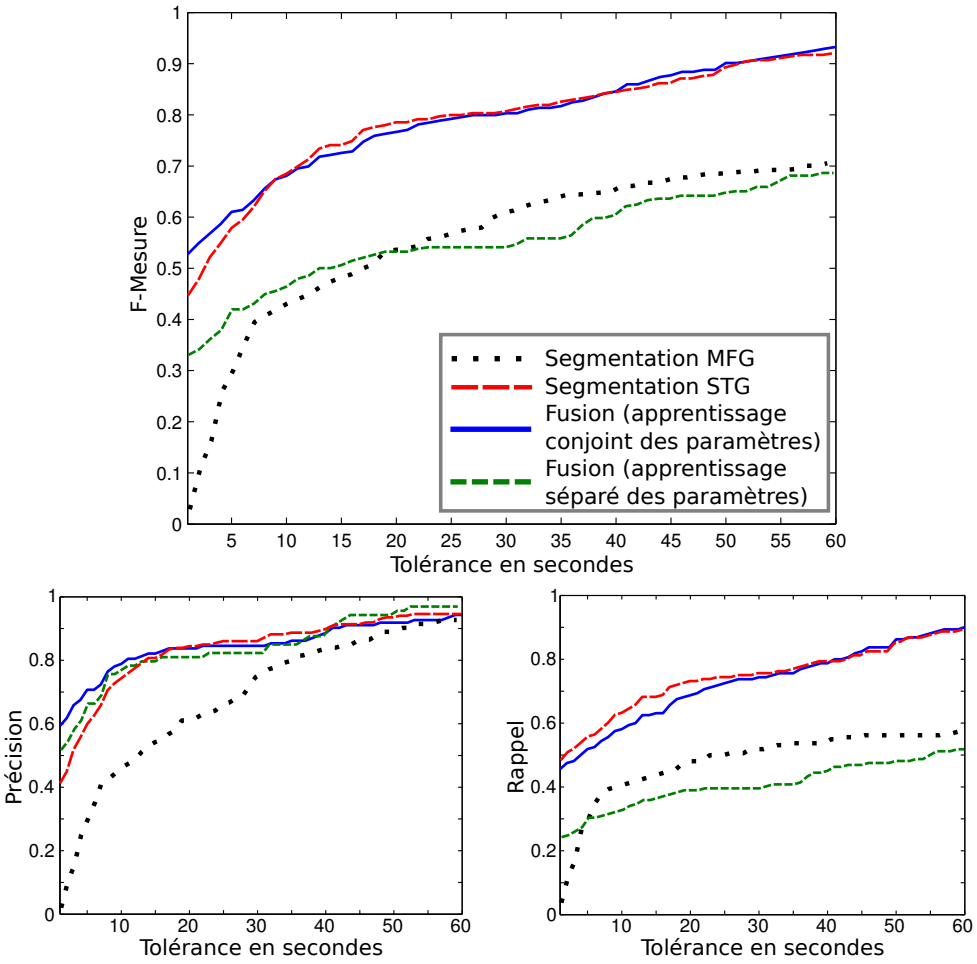

FIGURE *4.16 – Résultats des approches de fusion comparés aux méthodes de segmentation MFG et STG en fonction d'une tolérance de durée variable. L'ensemble de test utilisé est composé des quatre premiers épisodes de la série* Ally McBeal.

Les courbes de la Figure 4.16 montrent les résultats de précision, rappel et F-Mesure obtenues par les différentes approches monomodales (MFG et [Yeung 1998]) et les deux approches de fusion (avec apprentissage conjoint ou séparé des paramètres). Ces courbes montrent l'évaluation des frontières entre scènes automatiquement détectées en fonction d'une tolérance τ sur leur position par rapport aux frontières manuellement annotées.

Les deux approches de fusion obtiennent une meilleure précision que la segmentation de [Yeung 1998] basée uniquement sur la couleur. En n'autorisant aucune tolérance ($\tau = 0$), la précision obtenue par l'approche de fusion augmente de 2% si les paramètres sont appris séparément et de 16% si les paramètres le sont conjointement, comparé à l'approche de [Yeung 1998]. L'approche de fusion proposée tend à supprimer des frontières détectées par

la méthode de [Yeung 1998] si aucune frontière n'est détectée aux alentours de celles-ci par la méthode MFG. L'augmentation de la précision prouve que ces suppressions concernent majoritairement des faux positifs.

Cependant, le rappel est plus bas pour les deux approches de fusion que pour l'approche [Yeung 1998] (-3% et -24% respectivement si l'apprentissage des paramètres est conjoint ou séparé). Cela montre que les suppressions de frontières lors de la fusion ne sont pas toutes correctes. De plus, la fusion produit une sous-segmentation de l'épisode : pour 160 scènes annotées, l'approche [Yeung 1998] détecte 191 frontières, la fusion à partir d'une optimisation conjointe des paramètres produit 127 frontières et l'optimisation séparée des paramètres donne seulement 107 frontières.

Cette sous-segmentation conduit à une F-Mesure favorable à la méthode de segmentation [Yeung 1998] si l'on considère une tolérance $\tau > 5$ secondes. Sans tolérance, l'approche de fusion par optimisation conjointe des paramètres améliore la F-Mesure obtenue par l'approche [Yeung 1998] de 3%. Les paramètres des différentes méthodes de segmentation étant optimisés pour une tolérance $\tau = 0$ seconde, notre approche de fusion est donc pertinente.

Analyse détaillée de la fusion par apprentissage conjoint des paramètres

	MFG	**[Yeung 1998]**	**Fusion**
Nombre de frontières détectées	1179	130	124
Précision	0.09	0.57	0.61
Rappel	0.65	0.46	0.46
F-Mesure	0.16	0.51	0.52

TABLE *4.5 – Aperçu des segmentations MFG et [Yeung 1998] obtenues avec les paramètres conjointement appris pour optimiser la fusion.*

La fusion par apprentissage conjoint des paramètres consiste à optimiser conjointement les valeurs optimales des paramètres $\lambda = \{T, \theta, \Delta_t, \Delta_d\}$ des méthodes de segmentation MFG et [Yeung 1998], de manière à optimiser la F-Mesure globale de l'approche de fusion.

Le Tableau 4.5 montre les résultats obtenus pour les trois segmentations (MFG, [Yeung 1998] et Fusion) en utilisant ces paramètres conjointement optimisés pour la fusion. On remarque que l'approche MFG tend à sur-segmenter la vidéo : elle détecte 1179 frontières de scènes, alors que le corpus correspondant n'en contient que 160. Aligner les frontières provenant de la segmentation MFG sur les frontières de la méthode de [Yeung 1998] permet de réduire significativement ce comportement (de 1179 à 124). Le nombre de frontières détectées par la méthode de [Yeung 1998] passe de 130 à 124. L'augmentation de la précision et la stagnation du rappel indique que les 6 frontières supprimées étaient toutes des faux positifs.

4.3.3.2 À partir de tours de parole automatiques

Résultats pour les 4 premiers épisodes de la série Ally McBeal

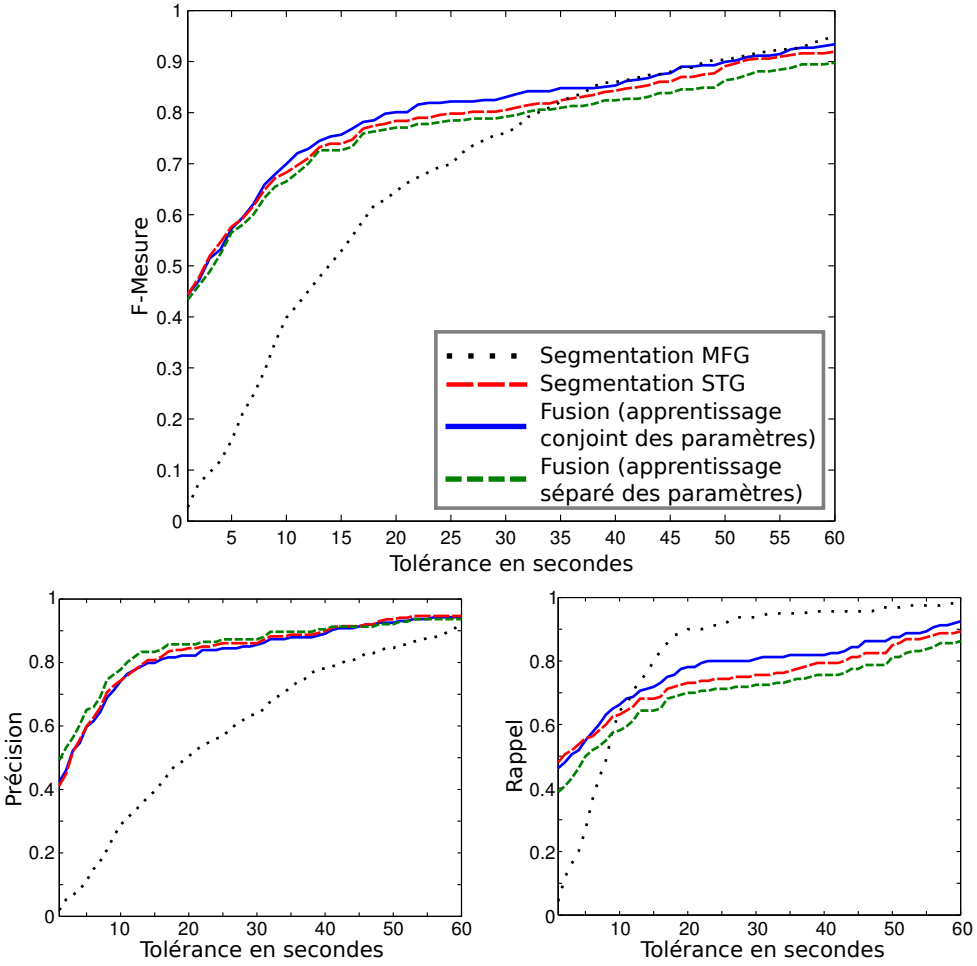

FIGURE *4.17 – F-Mesure des approches de fusion comparée aux méthodes de segmentation MFG et STG en fonction d'une tolérance dans la position des frontières détectées par rapport aux frontières annotées manuellement. L'ensemble de test utilisé est composé des quatre premiers épisodes de la série* Ally McBeal.

La Figure 4.17 permet d'observer les résultats obtenus par les approches de fusion dans le cas où l'on utilise les tours de parole détectés automatiquement en entrée de nos systèmes de segmentation en scènes. Ces résultats ont été calculés à partir des quatre premiers épisodes de la série *Ally McBeal*. Ainsi, il est possible de les comparer avec les résultats présentés dans la Figure 4.16, obtenus en utilisant des tours de parole manuellement annotés.

On observe un comportement très différent pour les approches de fusion en utilisant des tours de parole manuels ou automatiques. Nous avons vu précédemment que les tours de parole des locuteurs détectés automatiquement comportent beaucoup d'erreurs. Ces erreurs introduites en entrée de la méthode de segmentation par fenêtre glissante pro-

voquent une forte sur-segmentation (361 frontières détectées pour 160 annotées). Cette sur-segmentation a un impact dans la segmentation produite par les approches de fusion. Ainsi, l'utilisation de tours de parole automatiques permet de retrouver 178 frontières en utilisant la méthode de fusion avec optimisation conjointe des paramètres (127 avec des tours de parole manuels) et 130 frontières dans le cas d'une optimisation séparée (107 avec des tours de parole manuels).

Si l'on autorise une tolérance $\tau < 5$ secondes, les approches de fusion ne permettent pas d'améliorer le F-Mesure comparé à la segmentation de [Yeung 1998]. Cependant, on remarque que pour $\tau > 5$ secondes, l'approche de fusion par optimisation conjointe des paramètres donne un rappel supérieur à la méthode de [Yeung 1998] pour une précision similaire. Beaucoup des frontières détectées par l'approche de fusion sont donc seulement décalées de quelques secondes par rapport aux frontières annotées.

Résultats pour l'ensemble du corpus

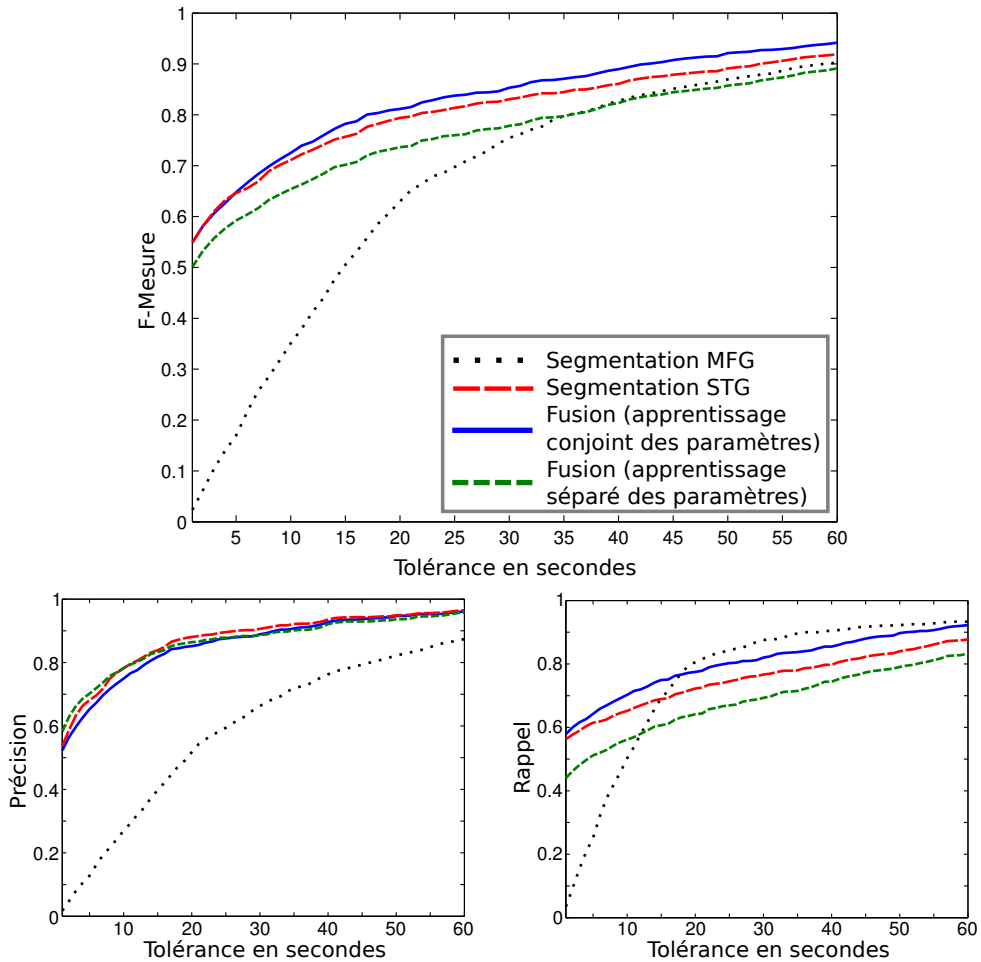

FIGURE *4.18 – F-Mesure des approches de fusion comparé aux méthodes de segmentation MFG et STG en fonction d'une tolérance dans la position des frontières détectées par rapport aux frontières annotées manuellement. L'ensemble de test utilisé est composé des 22 épisodes du corpus (les séries ont toutes un comportement similaire).*

Les mêmes observations peuvent être faites sur l'ensemble du corpus étudié. La Figure 4.18 compare les résultats obtenus par nos approches de fusion et les méthodes de segmentation MFG et [Yeung 1998] sur l'ensemble du corpus de test.

On observe que les deux approches de fusion donnent une précision similaire à l'approche de [Yeung 1998]. Cependant, la fusion par optimisation séparée des paramètres donne un rappel inférieur, et la fusion conjointe des paramètres montre un rappel supérieur, surtout en considérant une tolérance élevée lors de l'évaluation sur la position des frontières détectées. Il en résulte une F-Mesure légèrement supérieure pour notre approche de fusion si l'on autorise une tolérance $\tau > 5$ secondes.

4.3.4 Conclusion

Les approches de segmentation proposées se basent sur deux modalités, extraites du flux audiovisuel, pour retrouver les frontières entre les scènes : la couleur (modalité HSV) et les personnages ou locuteurs (modalité SD). Cependant, le Chapitre 3 propose une troisième modalité basée sur l'analyse de la transcription de la parole (modalité ASR). De nombreuses expériences ont été menées pour utiliser cette dernière modalité. Par exemple, il est possible d'utiliser une méthode proche de la méthode de segmentation MFG en décrivant la fenêtre avec les mots prononcés par les personnages. Les résultats obtenus n'étant pas convaincants, il a été décidé de ne pas discuter de cette modalité dans ce chapitre sur la segmentation en scènes.

Concernant les approches de segmentation proposées, considérons le cas où aucune tolérance n'est acceptée lors de l'évaluation sur la position des frontières détectées. Les expériences utilisant les tours de parole détectés automatiquement montrent que la fusion d'une segmentation MFG et d'une segmentation [Yeung 1998] ne permet pas d'améliorer la F-Mesure globale comparé à la segmentation [Yeung 1998] seule. Cette fusion permet d'augmenter le rappel de 2% au détriment de la précision (-2%).

Cependant, dans le cas de tours de parole manuellement annotés des locuteurs, la précision obtenue par l'approche de fusion augmente significativement comparé à la segmentation [Yeung 1998] (+16%). Toutefois, cette augmentation est faite au détriment du rappel (-9%), dû à des frontières qui ne sont pas détectées par la méthode de segmentation MFG. La F-Mesure obtenue par l'approche de fusion est ainsi supérieure de 3% à l'approche [Yeung 1998] n'utilisant que l'information sur la couleur.

Les résultats étudiés dans cette section montrent que la fusion d'une segmentation basée sur des histogrammes de couleur et une segmentation basée sur les tours de parole des locuteurs permet une amélioration des segmentations seules. Cependant, cette amélioration n'est pas significative. C'est pourquoi, dans la section suivante nous présentons une approche de segmentation de l'état de l'art, le GSTG [Sidiropoulos 2011] dans le but de fusionner les modalités HSV et SD pour améliorer la segmentation en scènes.

4.4 Fusion dans le cadre du GSTG

4.4.1 Description de l'approche

Chaque paire de valeurs $\lambda = \{\Delta_t, \Delta_d\}$ utilisées pour la méthode du STG [Yeung 1998] conduit à un ensemble différent de frontières entre scènes. Soit $\mathcal{F}$ l'ensemble des frontières de plans f d'une vidéo v. En fonction des valeurs de λ, chaque frontière f peut prendre comme valeur $\mathrm{STG}_\lambda(f) \in \{0, 1\}$, avec $\mathrm{STG}_\lambda(f) = 1$ si f est une frontière entre deux scènes, ou $\mathrm{STG}_\lambda(f) = 0$ sinon.

Les valeurs optimales (*i.e.* celles produisant le meilleur ensemble de frontières) sont dépendantes de la vidéo. Une manière élégante de simplifier l'apprentissage a été proposée par Sidiropoulos *et al.* [Sidiropoulos 2011] grâce à l'introduction du STG généralisé (GSTG). L'idée est de générer un grand ensemble de segmentations en sélectionnant des valeurs aléatoires pour Δ_d et Δ_t. Ensuite, pour chaque frontière de plan $f \in \mathcal{F}$ on calcule la proportion ρ de segmentations qui détectent f comme une frontière entre deux scènes. Soient $\mathfrak{D}_d$ et $\mathfrak{D}_t$ l'ensemble des valeurs sélectionnées pour Δ_d et Δ_t,

$$\rho(f) = \frac{1}{|\mathfrak{D}_t| \times |\mathfrak{D}_d|} \sum_{\Delta_d \in \mathfrak{D}_d} \sum_{\Delta_t \in \mathfrak{D}_t} \mathrm{STG}_{\{\Delta_d, \Delta_t\}}(f) \tag{4.9}$$

Comme illustré sur la Figure 4.19, les frontières de plans obtenant une proportion ρ supérieure à un seuil θ sont marquées comme frontières entre scènes (lignes verticales en pointillé).

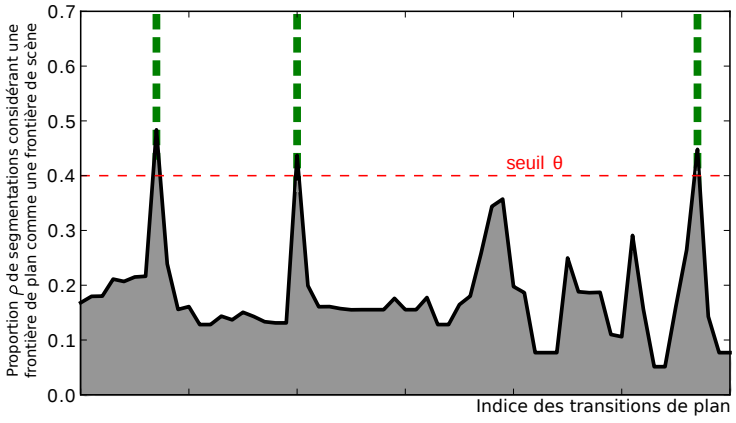

FIGURE *4.19 – Proportion de segmentations qui considèrent qu'une frontière de plan est une frontière de scène. Si cette proportion dépasse un seuil θ, la frontière de plan est marquée comme une frontière de scène (lignes vertes verticales). Cet exemple montre une partie d'un épisode durant laquelle trois frontières entre scènes sont détectées.*

Sidiropoulos *et al.* ont trouvé que cette approche donnait de meilleures performances que l'approche de segmentation STG et nos expériences préliminaires ont confirmé cette observation. De plus, il est plus aisé d'apprendre un seul paramètre (le seuil θ) au lieu de deux (Δ_d et Δ_t).

Plutôt que de sélectionner les valeurs aléatoirement pour Δ_d et Δ_t, notre implémentation génère un ensemble exhaustif de STGs en utilisant toutes les paires possibles de valeurs pour Δ_d et Δ_t (dans une grille 2D prédéfinie). L'intérêt de cette approche est d'être déterministe et donc de conduire à des résultats reproductibles. Les résultats présentés dans cette section sont obtenus en utilisant des valeurs pour Δ_d et Δ_t identiques à celles proposées dans le Tableau 4.4 page 113.

Fusion des modalités

La fusion des modalités consiste à effectuer une combinaison linéaire des proportions ρ des frontières de scènes générées par des STG basés sur les modalités précédentes (HSV et SD). Pour chaque frontière de plan, sa probabilité d'être une frontière de scène est définie par :

$$\rho(f) = w \cdot \rho^{\text{HSV}}(f) + (1 - w) \cdot \rho^{\text{TFIDF\_SD}}(f) \ \forall f \in \mathcal{F} \tag{4.10}$$

avec $\rho^{\text{HSV}}(f)$ la proportion de segmentations basées sur la modalité HSV qui détectent la frontière f comme frontière entre deux scènes, et $\rho^{\text{TFIDF\_SD}}(f)$ la proportion de segmentations basées sur la modalité SD qui détectent la frontière f comme une frontière entre deux scènes. Les poids w permet de pondérer l'influence des deux modalités pour le calcul de la probabilité finale. Le nombre de segmentations générées pour chaque modalité dépend de la grille des paramètres Δ_d et Δ_t. La grille utilisée est identique pour toutes les modalités.

La principale différence entre notre approche avec celle proposée par Sidiropoulos *et al.* [Sidiropoulos 2011] concerne la manière dont les graphes STG basés sur les locuteurs sont construits. Sidiropoulos et al. utilisent des informations audio bas-niveau pour définir des segments audio qui sont décrits par les locuteurs présents dans ces segments. Notre approche utilise les mêmes segments (des plans) comme base pour construire les graphes STG basés sur la modalité HSV et sur la modalité SD. De plus, nous proposons une description des locuteurs issue du domaine de la « fouille de texte » sous forme d'un calcul de pondération TF $\cdot$ IDF.

4.4.2 Résultats expérimentaux

Comparaison des approches monomodales et multimodales

La Figure 4.20 résume les performances d'un GSTG pour des approches monomodales (HSV et SD) et la fusion des modalités (HSV + SD) sur l'ensemble du corpus.

L'approche basée sur la modalité SD tend à détecter beaucoup trop de transitions (1136 transitions de scènes détectées pour 750 scènes annotées). Ainsi, en considérant une tolérance $\tau = 0$ sur la position des frontières entre scènes, ce comportement produit une précision très basse (0.18).

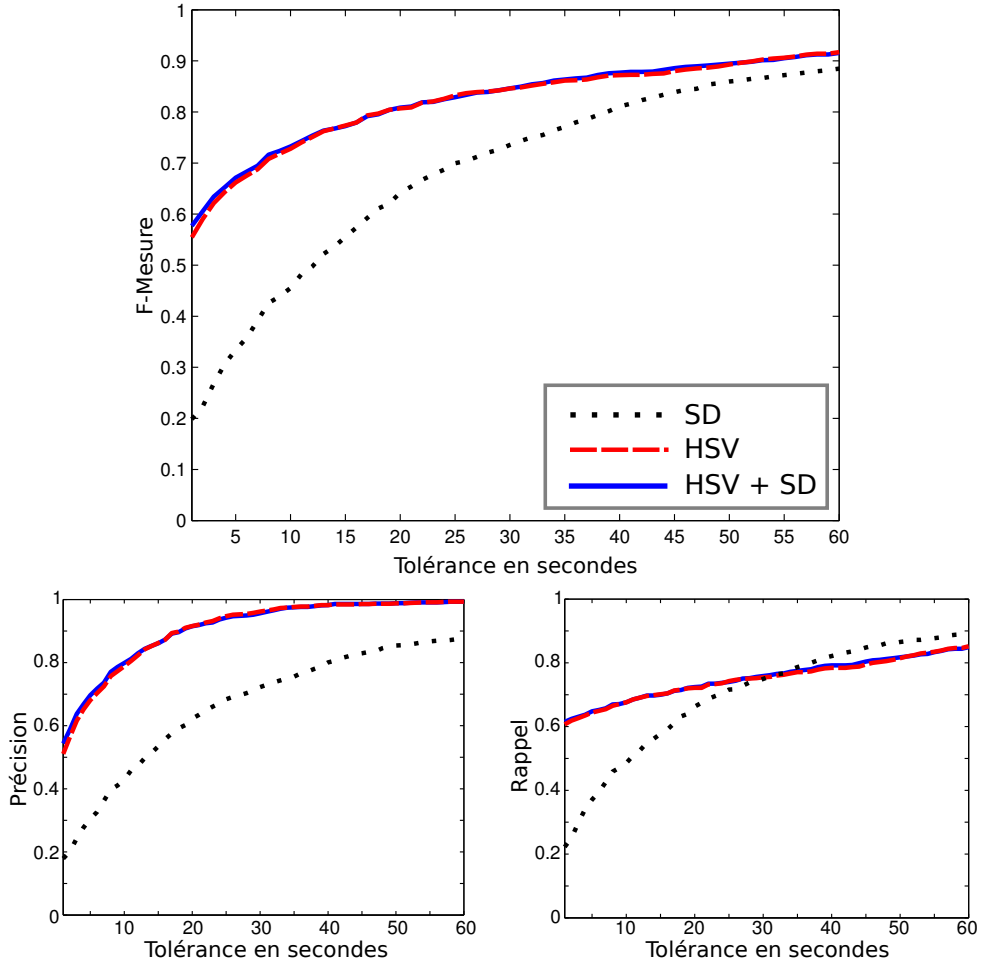

FIGURE *4.20 – Résultats de l'approche GSTG en fonction d'une tolérance de durée variable. Les courbes permettent de comparer le comportement du GSTG en fonction des modalités utilisées pour le calcul des distances entre les plans. L'ensemble de test utilisé est composé des 22 épisodes du corpus.*

Ces résultats peuvent être expliqués par les erreurs introduites par les tours de parole des locuteurs détectés automatiquement. De plus, comme un plan est généralement décrit par un seul locuteur, la distance $d^{\text{TFIDF\_SD}}$ entre deux plans est est très souvent égale à 0 (les deux plans sont identiques) ou égale à 1 (il n'y a aucun locuteur commun entre les deux plans). Ainsi, l'étape de regroupement agglomératif des STG s'arrête très tôt quel que soit le seuil Δ_d, ce qui résulte en un long STG avec beaucoup de groupes composé d'un seul plan (et donc de beaucoup d'arcs de coupure). Ce phénomène est illustré dans par la Figure 4.21, qui montre le fonctionnement d'un STG dans le cas où peu de plans sont regroupés (arrêt précoce du regroupement).

L'approche monomodale basée sur l'utilisation des tours de parole des locuteurs détectés automatiquement (SD) donne de moins bons résultats que l'approche utilisant uni-

quement l'information visuelle (HSV). L'utilisation de celle-ci pour la segmentation avec l'approche GSTG donne une F-Mesure de 0.57 à $\tau = 0$, ce qui est meilleur de 3% en valeur absolue comparé à la meilleure approche présentée précédemment. De plus, la fusion des modalités SD et HSV permet d'améliorer la F-Mesure par rapport à une segmentation basée uniquement sur la couleur en réduisant la sur-segmentation (884 transitions détectées contre 982). Ainsi, l'utilisation conjointe des modalités SD et HSV permet d'augmenter la précision (+3%) à $\tau = 0$, tout en conservant un rappel similaire.

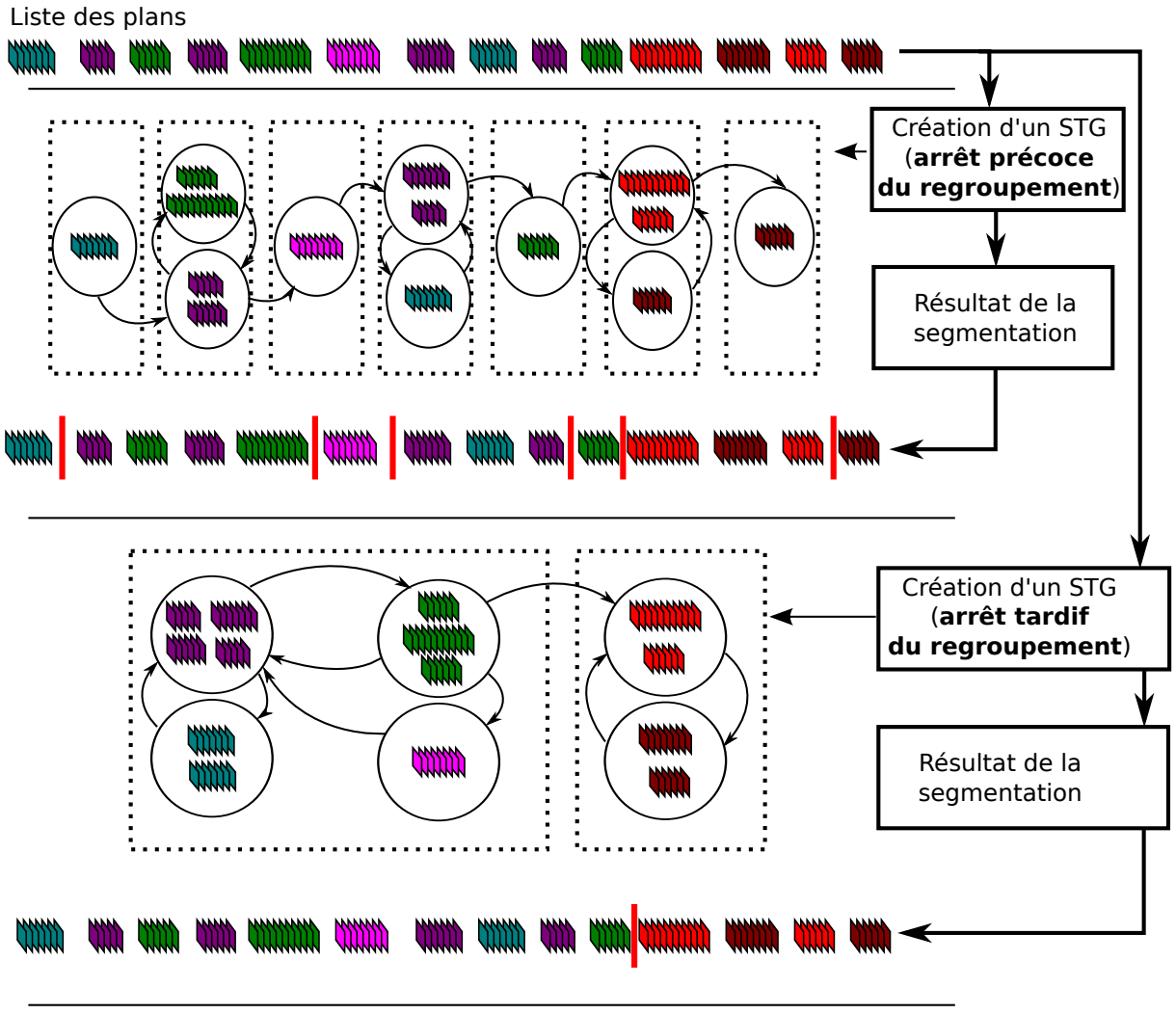

FIGURE *4.21 – Construction d'un STG dans le cas où peu de plans sont regroupés (arrêt précoce du regroupement) et dans le cas où beaucoup de plans sont regroupés (arrêt tardif du regroupement). Dans le premier cas, le STG possède beaucoup d'arcs de coupures résultant en un grand nombre de frontières de scènes détectées. Dans le deuxième cas, beaucoup de plans sont regroupés, et le STG ne possède qu'un seul arc de coupure résultant en une seule frontière de scènes détectée.*

Poids des modalités HSV et SD

Lors de l'apprentissage des paramètres du *GSTG*, un « poids » est donné à chaque modalité (SD et HSV) utilisée pour le calcul des distances entre les plans de façon à optimiser la F-Mesure. La Figure 4.22 illustre les poids attribués à chaque modalité pour les différentes séries télévisées. La modalité SD a une importance de 25% pour les séries *Ally McBeal* et *Malcolm*, et 10% pour *le Trône de Fer*.

Ces résultats confirment les observations faites jusqu'à maintenant, à savoir que pour la série *le Trône de Fer*, les scènes consécutives ont un environnement visuel très différent, impliquant un apport très important de la modalité HSV pour la segmentation en scènes, ce qui induit une faible pondération pour la modalité SD.

Pour les séries *Malcolm* et *Ally McBeal*, l'environnement visuel varie beaucoup moins entre les scènes. La modalité SD devient par là même plus intéressante pour la segmentation.

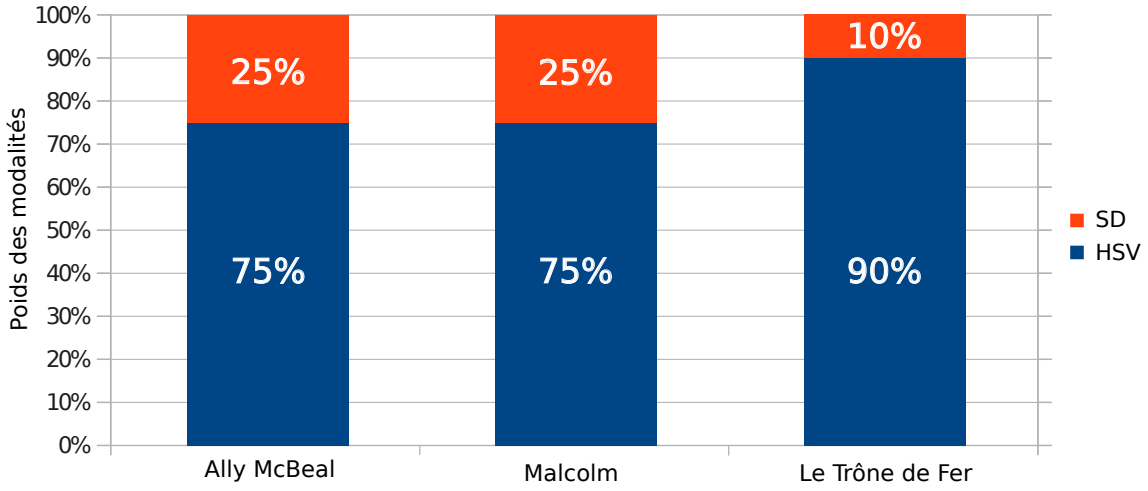

FIGURE *4.22 – Poids associés aux modalités* HSV *(couleur) et* SD *(locuteurs) lors de l'apprentissage du* $GSTG$.

4.5 Conclusion

La Figure 4.23 permet de comparer la F-Mesure obtenue par les différentes approches de segmentation pour les trois séries du corpus, en n'autorisant aucune tolérance sur la position des frontières entre scènes lors de l'évaluation.

La fusion par alignement des frontières provenant d'une segmentation MFG et d'une segmentation de [Yeung 1998] ne permet une amélioration de la segmentation que pour la série *Ally McBeal*. Cependant, pour toutes les séries, la meilleure F-Mesure est obtenue par la fusion des modalités HSV et SD avec une méthode GSTG.

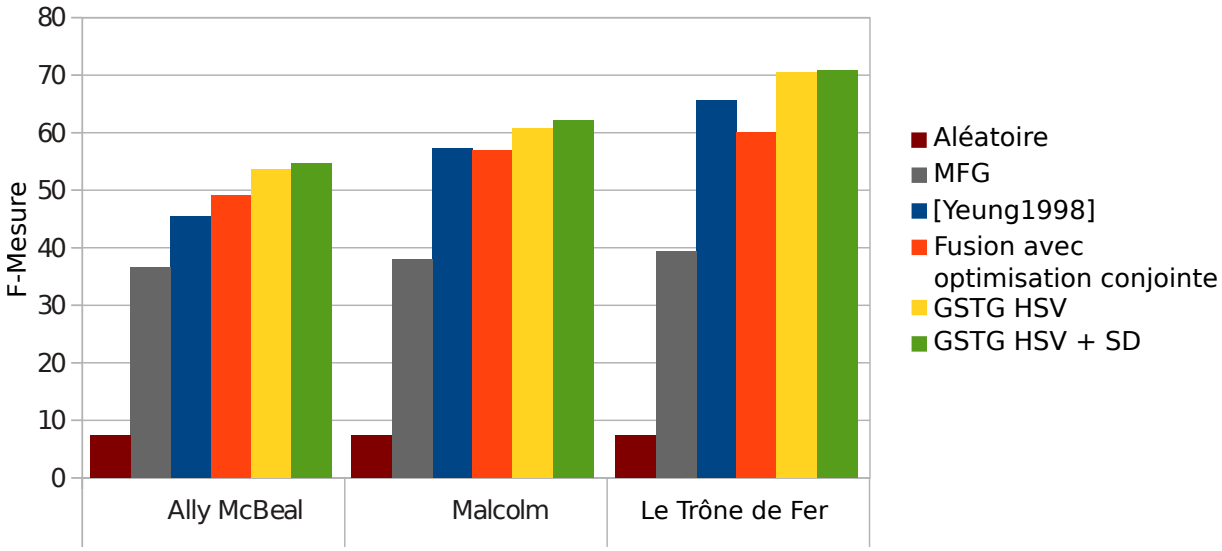

FIGURE *4.23 – F-mesure moyenne par série et méthode de segmentation en scènes.*

Chapitre 5

Regroupement des scènes en histoires

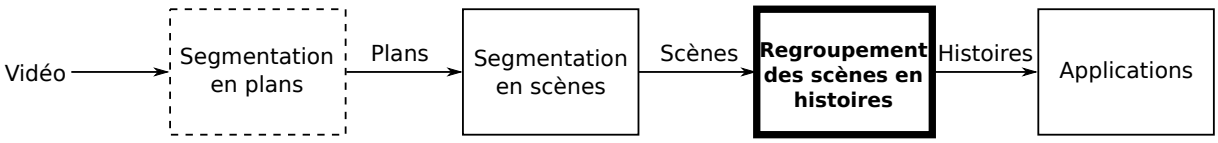

FIGURE *5.1 – Différents niveaux de structuration d'une vidéo : focus sur le regroupement des scènes en histoires.*

Comme illustré dans la Figure 5.1, le niveau de structuration d'un épisode de série télévisée étudié dans ce chapitre concerne le regroupement des scènes en histoires. La définition d'une histoire détaillée au Chapitre 1 (page 11) est la suivante :

Définition d'une histoire
- Une histoire est un ensemble d'évènements qui partagent une même idée maîtresse.
- Une scène décrivant un évènement unique, une histoire est donc l'ensemble de toutes les scènes qui partagent la même idée maîtresse.

La Figure 5.2 illustre le résultat attendu d'un système de regroupement des scènes en histoires pour un épisode de la série *Malcolm* : les histoires sont constituées en regroupant les scènes, non nécessairement adjacentes, qui les composent. Ainsi retrouver les histoires dans un épisode de série télévisée peut être vu comme une tâche de regroupement des scènes.

Tâche de regroupement des scènes en histoire
À partir de l'ensemble des scènes d'un épisode, la tâche de regroupement des scènes en histoires consiste à regrouper les scènes partageant la même idée maîtresse.

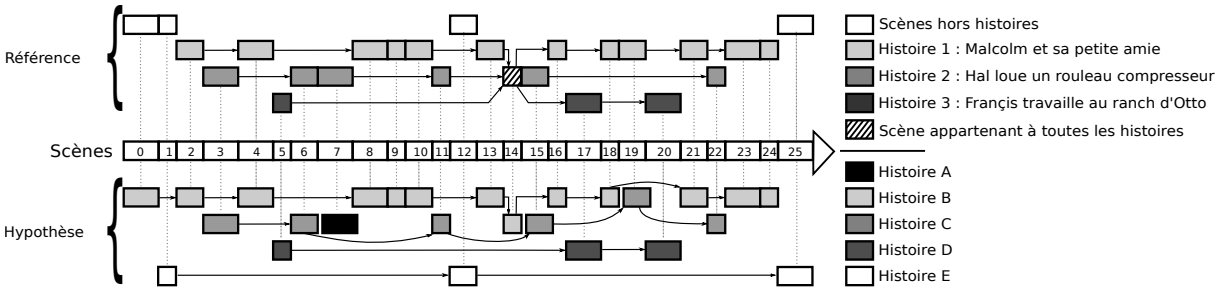

FIGURE *5.2 – Exemple de regroupement des scènes en histoires pour un épisode de série télévisée. Chaque couleur indique les scènes appartenant à une même histoire. La référence représente le résultat d'un regroupement manuel des scènes. L'hypothèse est un exemple du résultat que l'on peut obtenir par une approche de regroupement automatique.*

Le regroupement des scènes en histoires s'effectue généralement en deux étapes illustrées par la Figure 5.3. La première étape consiste à calculer une matrice d'affinité entre les scènes, déterminée à partir d'une mesure de distance pour chaque paire de scènes. Ensuite, un algorithme de regroupement permet de constituer les groupes de scènes homogènes à partir de cette matrice d'affinité.

Le choix de la mesure de distance est critique et doit être effectué en concordance avec les sorties souhaitées du regroupement. Ce choix est étudié dans le Chapitre 3 page 75, où les observations 2, 3 et 5 considèrent que les descripteurs basés sur les modalités HSV (histogrammes de couleur), SD (sortie d'un système de segmentation et regroupement en locuteurs) et ASR (sortie d'un système de transcription automatique de la parole) apportent des informations pertinentes pour le regroupement des scènes en histoires.

Bien qu'il existe de nombreuses métriques permettant d'évaluer l'efficacité des méthodes de regroupement, elles ne sont pas forcément adaptées à l'évaluation d'une tâche spécifique. Nous commençons par discuter de la problématique de l'évaluation de la tâche de regroupement des scènes en histoires dans la Section 5.1.

La Section 5.2 présente le protocole expérimental utilisé pour valider nos expériences.

Il existe une multitude de méthodes de regroupement dont certaines sont décrites dans la Section 2.2 page 44. L'utilisation de ces méthodes et des différents calculs de distances entre scènes pour le regroupement des scènes en histoires est étudiée dans la Section 5.3. La Section 5.4 étudie une approche de fusion des informations provenant des 3 modalités pour améliorer le regroupement.

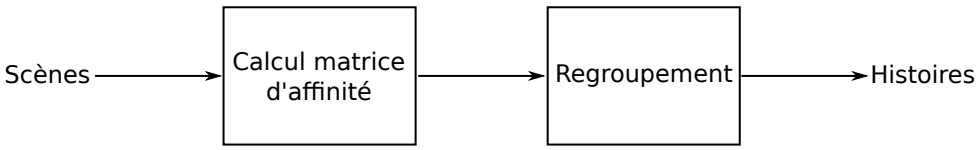

FIGURE *5.3 – Étapes du regroupement des scènes en histoires*

Enfin, la Section 5.5 propose une méthode de sélection automatique de la meilleure approche de regroupement à appliquer à chaque épisode.

5.1 Métriques d'évaluation

La tâche de regroupement des scènes en histoire est, comme son nom l'indique, une tâche de regroupement de données. A notre connaissance, il n'existe pas de travaux discutant de la problématique de l'évaluation d'un regroupement des scènes en histoires. Beaucoup de méthodes et de métriques pour mesurer la qualité de groupes de données sont disponibles dans la littérature scientifique. Cependant, bien qu'il soit possible d'utiliser des méthodes d'évaluation de regroupement génériques, elles ne sont pas forcément adaptées à la tâche que nous souhaitons évaluer.

Dans cette section, nous analysons trois métriques d'évaluation de la qualité du regroupement des scènes histoires : la *F-Mesure*, l'*Adjusted Rand Index* (ARI) et le *Diarization Error Rate* (DER). Ces métriques sont toutes basées sur la comparaison d'un regroupement de référence et d'une hypothèse. Dans notre cas, la référence a été manuellement annotée en associant à chaque scène les identifiants des histoires dont elle fait partie (le corpus et les annotations sont décrits dans la Section 5.2).

5.1.1 F-Mesure

D'après Manning [Manning 2008], il est possible d'évaluer la qualité d'un regroupement en analysant toutes les paires (i, j) des objets regroupés et en répondant au problème de classification suivant : est-ce que les objets i et j font partie du même groupe ?

Dans notre cas, les objets à regrouper sont les scènes de la vidéo. Soit $\mathcal{S}$ l'ensemble des scènes de la vidéo. La tâche de regroupement des scènes en histoires est une tâche de classification $\mathbb{Q}$ des paires de scènes (s, s') telle que

$$\begin{aligned} \mathbb{Q} : \mathcal{S} \times \mathcal{S} &\rightarrow \{0, 1\} \\ (s, s') &\rightarrow \begin{cases} 1 & \text{si } s \text{ et } s' \text{ sont dans la même histoire} \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{5.1}$$

Il est ainsi possible d'évaluer cette tâche de classification par un calcul de précision (P), rappel (R) et F-mesure ($\mathrm{F_{PR}}$) (décrits Section 2.4.3 page 71) en comparant une classification automatique $\mathbb{Q}$ des paires de scènes avec une classification manuelle $\mathbb{M}$ de référence :

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \\ \mathrm{R} &= \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \\ \mathrm{F_{PR}} &= 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \end{aligned}$$

Le comptage des *Vrais Positifs* (VP), *Faux Positifs* (FP), *Vrais Négatifs* (VN) et *Faux Négatifs* (FN) est effectué suivant les formules suivantes :

$$\begin{aligned} \text{VP} &= \text{card}\{(s,s') \in \mathcal{S} \times \mathcal{S} / \mathbb{Q}((s,s')) = 1 \wedge \mathbb{M}((s,s')) = 1\} \\ \text{FP} &= \text{card}\{(s,s') \in \mathcal{S} \times \mathcal{S} / \mathbb{Q}((s,s')) = 1 \wedge \mathbb{M}((s,s')) = 0\} \\ \text{VN} &= \text{card}\{(s,s') \in \mathcal{S} \times \mathcal{S} / \mathbb{Q}((s,s')) = 0 \wedge \mathbb{M}((s,s')) = 0\} \\ \text{FN} &= \text{card}\{(s,s') \in \mathcal{S} \times \mathcal{S} / \mathbb{Q}((s,s')) = 0 \wedge \mathbb{M}((s,s')) = 1\} \end{aligned} \tag{5.2}$$

Limites

La F-Mesure a quelques limites. Par exemple, dans le cas où toutes les scènes sont incorrectement regroupées en une unique histoire, le rappel est égal à 1 (aucun faux négatif, FN=0), pour une précision non nulle. Une F-Mesure élevée peut donc être atteinte par un système se contentant de regrouper toutes les scènes ensemble.

5.1.2 (Adjusted) Rand Index

Le *Rand Index* (RI) est une métrique générique d'évaluation de regroupement proposée par Rand [Rand 1971]. Comme la F-Mesure, elle est basée sur le calcul des valeurs de VP, VN, FP et FN :

$$\text{RI} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{5.3}$$

C'est une mesure de performance qui correspond au pourcentage de paires de scènes pour lesquelles la référence et l'hypothèse sont d'accord. L'avantage du RI est que, contrairement à la F-Mesure, il prend en compte le nombre de VN, et donc un système regroupant toutes les scènes en un unique groupe obtient un *RI* faible.

D'après Hubert et Arabie [Hubert 1985], le *Rand Index* varie entre 0.5 et 1 dans des cas concrets. Ils proposent de réaliser un ajustement du *Rand Index* : l'*Adjusted Rand Index* (ARI). Cette ajustement utilise un calcul de l'*index attendu* qui représente l'espérance mathématique du Rand Index que peut atteindre un regroupement aléatoire des scènes comparé à la référence.

$$\text{ARI} = \frac{\text{RI} - \text{index attendu}}{1 - \text{index attendu}} \tag{5.4}$$

Le calcul de l'ARI est détaillé dans la Section 2.4.3 (page 71).

5.1.3 Diarization Error Rate

La troisième métrique d'évaluation est empruntée à la tâche de *segmentation et regroupement en locuteurs* : le *Diarization Error Rate* (DER) ou taux d'erreur de la segmentation et du regroupement en locuteurs.

La sortie d'un système de segmentation et regroupement en locuteurs consiste en une liste de segments de parole décrits par un début, une fin et l'identifiant du locuteur associé, et appelés « tours de parole ». Une première étape d'un tel système consiste souvent à réaliser une segmentation parole/non parole dans le but de distinguer les segments de parole des segments ne contenant pas de parole. Le DER est une mesure d'erreur qui permet de déterminer à quel point le regroupement des tours de parole des locuteurs de l'hypothèse est différent du regroupement des tours de parole manuellement annotés.

Le calcul du DER est basé sur une recherche de la correspondance optimale entre un locuteur $\ell \in \mathcal{R}$ de la référence et un locuteur de l'hypothèse $\mathcal{H}$. Nous faisons l'hypothèse non restrictive que $|\mathcal{H}| = |\mathcal{R}|$ en ajoutant des locuteurs vides dans l'un ou l'autre des ensembles pour respecter cette égalité. Soit k la matrice de co-occurrence de dimension $|\mathcal{H}| \times |\mathcal{R}|$ telle que $k(\ell, \ell')$ est égale à la durée totale de co-occurence des locuteurs $\ell \in \mathcal{H}$ et $\ell' \in \mathcal{R}$. La correspondance optimale m^* est définie par l'Equation 5.5.

$$m^* = \underset{m \in \mathbb{R}^{\mathcal{H}}}{\operatorname{argmax}} \sum_{\ell \in \mathcal{H}} k(\ell, m(\ell)) \tag{5.5}$$

où m est une fonction bijective qui à tout locuteur de l'hypothèse associe un locuteur de la référence.

$$m : \mathcal{H} \to \mathcal{R} \tag{5.6}$$

Le DER tient compte de trois types d'erreur. Son calcul est illustré dans la Figure 5.4.

$$DER = \frac{\text{Fausses alarmes} + \text{Non-détections} + \text{Erreurs de locuteurs}}{\text{Durée de l'épisode}} \tag{5.7}$$

Fausses alarmes : durée totale des segments de parole de l'hypothèse correspondant à des segments de non parole dans la référence (exemple : segment E1 de la Figure 5.4).

Non-détections : durée totale des segments de parole de la référence correspondant à des segments de non parole dans l'hypothèse (exemple : segment E2 de la Figure 5.4).

Erreurs de locuteurs : durée totale des segments de parole pour lesquels l'identifiant du locuteur ℓ' de la référence ne correspond pas à l'identifiant du locuteur ℓ dans l'hypothèse : $\ell' \neq m^*(\ell)$ (exemple : segment E3 de la Figure 5.4).

Le DER est une mesure d'erreur. Ainsi, un regroupement parfait obtiend un DER de 0. Cette métrique d'évaluation peut directement être transposée au problème de regroupement des scènes en histoires en utilisant les analogies suivantes :

Transposition du DER à la tâche de regroupement des scènes en histoires

Locuteur	$\longleftrightarrow$	Histoire
Tour de parole	$\longleftrightarrow$	Scène
Tour de parole superposée	$\longleftrightarrow$	Scène appartenant à plusieurs histoires
Non-détection	$\longleftrightarrow$	(ne s'applique pas)

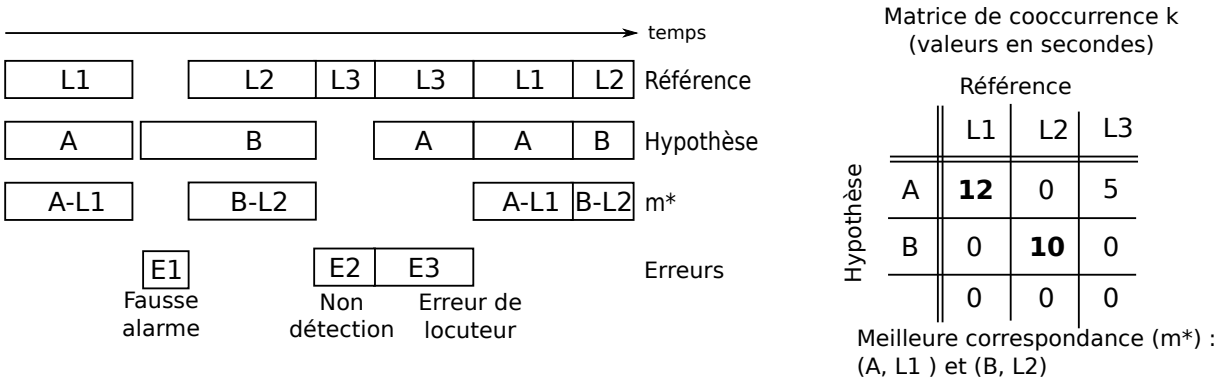

FIGURE *5.4 – Calcul du* Diarization Error Rate

Dans notre cas, puisque toutes les scènes d'une vidéo sont prises en compte par le système de regroupement des scènes en histoires (couvrant ainsi la durée totale de l'épisode), aucun segment de vidéo ne peut être considéré comme une *non-détection*.

5.1.4 Comparaison des métriques

Les trois métriques présentées ici sont généralement d'accord sur le fait qu'un regroupement est meilleur qu'un autre. Ainsi, si h et h' sont deux regroupements et $\mathcal{L}$ et $\mathcal{L}'$ deux métriques d'évaluation, alors $\mathcal{L}(h) > \mathcal{L}(h') \Rightarrow \mathcal{L}'(h) > \mathcal{L}'(h')$. Cependant, leur comportement peut être très différent dans des situations particulières pour l'évaluation du regroupement des scènes en histoires.

Cas des scènes appartenant à plusieurs histoires

La Figure 5.5 montre un épisode composé de deux histoires partageant une scène commune (#7). La première histoire (en blanc) contient 9 scènes et la seconde (en gris foncé) est composée de 4 scènes.

Deux hypothèses (h et h') sont proposées, illustrant le cas où la scène #7 n'est associée qu'à une seule histoire. Dans les deux cas, l'erreur est la même (une scène appartenant à deux histoires n'est associée qu'à l'une d'entre elles). Une métrique idéale doit donner la même valeur aux deux hypothèses ($\mathcal{L}(h) = \mathcal{L}(h')$). C'est le cas pour le DER, mais la F-Mesure et l'ARI retournent des valeurs différentes.

Prise en compte de la durée des scènes

La Figure 5.6 montre un autre comportement intéressant du DER. Le DER ne se contente pas d'évaluer le résultat du regroupement, mais il prend en compte la durée des erreurs. C'est à dire qu'une longue scène a un poids plus important qu'une scène courte.

Prenons un exemple concret pour illustrer ce problème. La Figure 5.6 propose deux hypothèses pour lesquelles une unique scène a été associée à une histoire incorrecte (correspondant à la scène rayée dans la Figure 5.6). Dans l'hypothèse h, la durée de cette

scène est beaucoup plus longue que dans l'hypothèse h'. En leur donnant le même poids, il est possible qu'un regroupement obtienne un bon score alors qu'une grande partie de l'épisode (en durée) est mal regroupée. Ainsi, la métrique DER montre une autre propriété intéressante pour la tâche de regroupement des scènes en histoires puisqu'elle évalue la durée d'un épisode correctement regroupée et non le nombre de scènes.

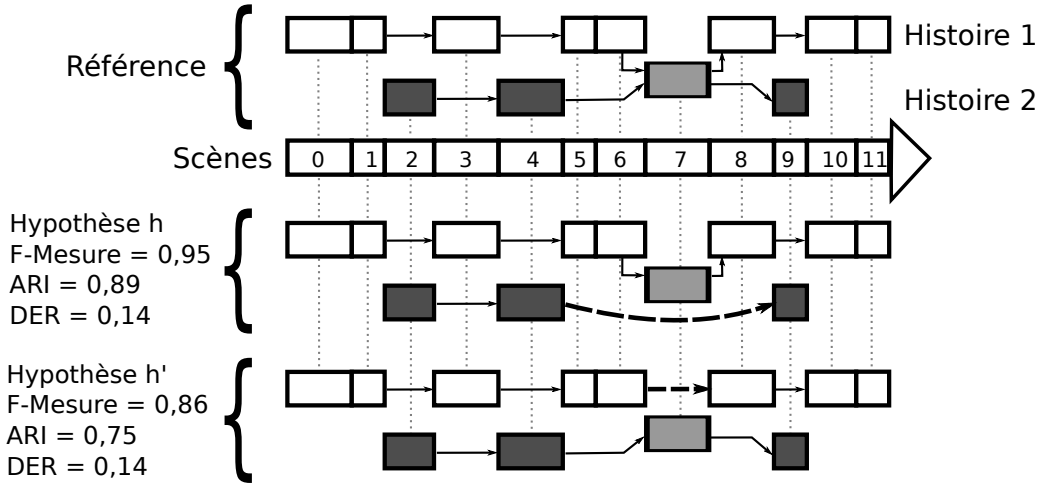

FIGURE *5.5 – Comportement des métriques d'évaluation avec des scènes appartenant à plusieurs histoires*

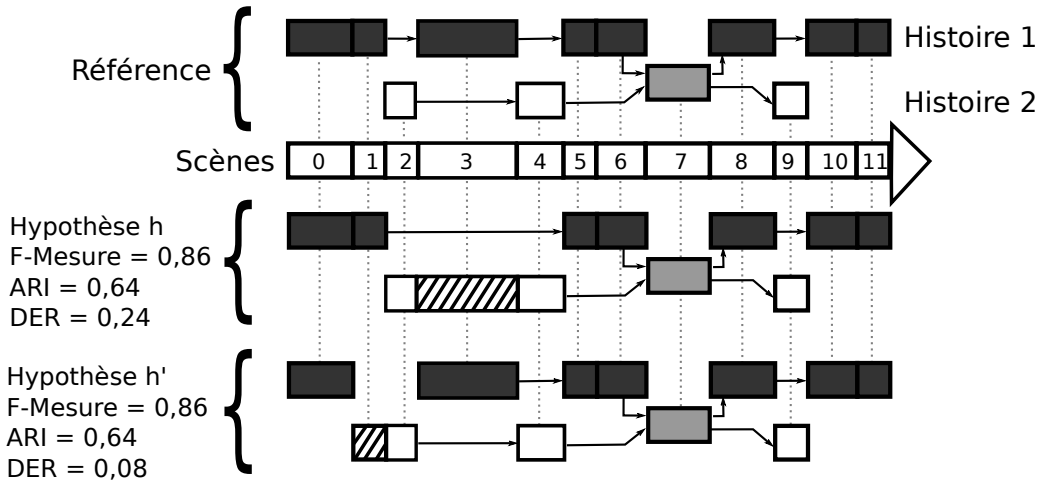

FIGURE *5.6 – Comportement des métriques d'évaluation en fonction de la durée des scènes mal regroupées.*

5.1.5 Conclusion

Le Tableau 5.1 résume les avantages et les inconvénients des trois métriques d'évaluation pour la tâche de regroupement des scènes en histoires. La métrique DER est la seule qui permet de prendre en compte la durée des scènes et le cas des scènes appartenant à plusieurs histoires. Par conséquent, toutes les évaluations discutées dans ce chapitre pour le regroupement des scènes en histoires utilisent cette métrique.

	F-Mesure	**ARI**	**DER**
Prise en compte la durée des scènes	-	-	+
Comportement idéal dans le cas de scènes appartenant à plusieurs histoires	-	-	+

TABLE *5.1 – Avantages et limites des métriques d'évaluation pour le regroupement des scènes en histoires.*

5.2 Protocole expérimental

5.2.1 Corpus

Pour évaluer les algorithmes de regroupement des scènes en histoires, nous avons annoté les histoires de 22 épisodes de séries télévisées : 7 épisodes de la première saison de la série *Ally McBeal*, 7 épisodes de la saison 4 de *Malcolm* et 8 épisodes de la première saison de la série *Le trône de fer*.

Ces 3 séries ont été sélectionnées pour leur style très différent (nombre varié de scènes, d'histoires ou de personnages résumés dans le Tableau 5.2).

- La série *Ally McBeal* propose des épisodes de 40 minutes racontant peu d'histoires en parallèle (2,7 en moyenne), et dont la plupart sont centrés sur le personnage d'*Ally*.
- Les épisodes de la série *Malcolm* sont plus courts, et ont un rythme plus élevé : scènes plus courtes (45 secondes en moyenne contre 57 secondes pour *Ally McBeal* et 100 secondes pour *le Trône de Fer*) et un nombre élevé d'histoires (3,7 par épisode). De plus, les histoires racontées impliquent généralement des personnages très divers.
- La série *le Trône de Fer*, quant à elle, propose des épisodes racontant beaucoup d'histoires différentes (en moyenne 4,2 par épisodes), qui ont souvent comme particularité de se dérouler dans des lieux très différents (château, désert glacé, désert aride, forêt).

Les nombres de scènes et d'histoires correspondent aux scènes et histoires annotées suivant les définitions du Chapitre 1 (page 11). Les tours de parole des locuteurs ont été

obtenus automatiquement en utilisant un système de segmentation et regroupement en locuteurs [Barras 2006]. Dans le but d'évaluer la pertinence des méthodes de regroupement des scènes en histoires sans le biais imposé par les erreurs du système automatique de segmentation et regroupement en locuteurs, les locuteurs présents dans chaque scène ont été manuellement annotés pour l'ensemble du corpus. Ces annotations sont différentes de celles utilisées dans le chapitre précédent : seule la liste des locuteurs présents dans chaque scène a été manuellement annotée, et non les tours de parole des locuteurs.

			Annotation manuelle		
Série	**Nombre d'épisodes**	**Durée totale** (par épisode)	**Nombre de scènes total** (par épisode)	**Nombre d'histoires** (par épisode)	**Nombre de personnages par épisode**
Ally McBeal	7	5h (≈41min)	304 (43 scènes)	19 (2,7 histoires)	15
Malcolm	7	2,5h (≈22min)	196 (28 scènes)	24 (3,4 histoires)	19
Le Trône de Fer	8	7h (≈50min)	244 (30 scènes)	34 (4,2 histoires)	34

TABLE *5.2 – Description du corpus utilisé.*

5.2.2 Scènes ne faisant partie d'aucune histoire

Quelques scènes particulières ne font partie d'aucune histoire (les scènes blanches dans la Figure 5.2). Ces scènes sont les génériques de début et de fin, des sketchs isolés (par exemple, chaque épisode de la série *Malcolm* propose un sketch avant le générique), ou des scènes de transition qui n'ont aucune relation entre elles ni avec les histoires. Puisque ces scènes ne sont pas annotées en histoires, elles peuvent être prises en compte de trois façons différentes lors de l'évaluation :

- nous pouvons considérer qu'elles font toutes partie d'une même histoire (ce qui favorise un regroupement qui tend à « sur-regrouper » les scènes) ;
- nous pouvons considérer que chaque scène est une histoire indépendante (ce qui favorise un regroupement qui tend à « sous-regrouper » les scènes) ;
- nous pouvons ne pas les comptabiliser lors de l'évaluation.

Comme elles ne sont pas annotées comme des histoires, et pour ne pas favoriser un comportement particulier des approches de regroupement étudiées, ces scènes n'interviennent pas dans le calcul de l'évaluation. Cependant, ces scènes ne sont pas supprimées de l'ensemble des scènes à regrouper. Elles doivent donc être prises en compte lors des étapes de regroupement.

5.2.3 Regroupement aléatoire

Nous utilisons un système de regroupement *aléatoire* comme système de référence. Le résultat d'un regroupement aléatoire peut fortement varier en fonction du nombre d'histoires, du nombre de scènes ou de la durée des scènes. C'est pourquoi une valeur de référence est calculée pour chaque série télévisée comme étant le DER moyen du regroupement aléatoire des épisodes de la série.

Un regroupement aléatoire est obtenu en utilisant un regroupement *average-link* (décrit dans la Section 5.3) basé sur des matrices d'affinité entre les scènes aléatoires. La valeur finale du DER est obtenue par un calcul similaire à celui présenté dans le chapitre précédent (page 102). La valeur aléatoire est le DER moyen obtenu par 1000 regroupements aléatoires auquel a été retranché 3 fois l'écart-type de ces 1000 résultats. Ainsi, une évaluation inférieure à cette valeur est meilleure que $99,9\%$ de tous les regroupements aléatoires effectués.

5.2.4 Oracle

La valeur *oracle* montre le meilleur score que peuvent obtenir les approches de regroupement proposées. En effet, aucune des méthodes de regroupement décrites dans ce chapitre n'est en mesure d'associer une scène à plus d'une histoire. Or il est courant qu'une scène décrive plusieurs histoires : c'est le cas pour 50 des 774 scènes annotées du corpus d'évaluation. Ainsi, dans le meilleur des cas, si au moins une scène appartient à deux histoires (ou plus), le DER est supérieur à 0. L'oracle permet d'illustrer la meilleure valeur du DER que peuvent atteindre les approches de regroupement en tenant compte de cette propriété.

5.3 Approches monomodales

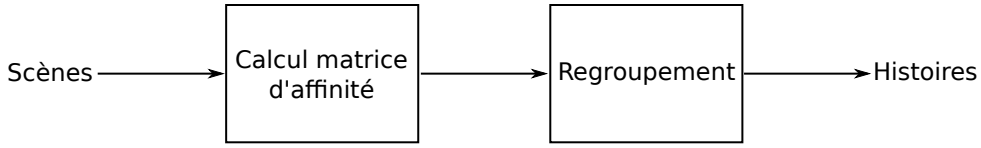

FIGURE 5.7 – *Étapes du regroupement des scènes en histoires*

5.3.1 Méthodes utilisées

Calcul de la matrice d'affinité

La Figure 5.7 montre les deux étapes nécessaires au regroupement des scènes en histoires. La première étape consiste à calculer une matrice d'affinité $\mathcal{A}$ des scènes, telle que $\mathcal{A}_{ij}$ soit la distance entre les scènes s_i et s_j. Les observations 2, 3 et 5 énoncées au Chapitre 3 (page 75) permettent de conclure que les modalités HSV (histogrammes de couleur), SD (sortie d'un système automatique de segmentation et regroupement en locuteurs) et ASR (sortie d'un système de transcription automatique de la parole) apportent des informations pertinentes pour le regroupement des scènes en histoires.

Soit N le nombre de scènes dans une vidéo, le calcul de la matrice d'affinité peut être réalisé suivant plusieurs mesures de distances basées sur ces trois modalités :

- Modalité HSV : $\forall i, j \in [\![1, N]\!]^2, \mathcal{A}_{ij}^{\text{HSV}} = d^{\text{HSV}}(s_i, s_j)$ avec d^{HSV} la distance entre les histogrammes de couleur décrivant les scènes (équation 3.1, page 77).
- Modalité ASR : $\forall i, j \in [\![1, N]\!]^2, \mathcal{A}_{ij}^{\text{ASR}} = d^{\text{ASR}}(s_i, s_j)$ avec d^{ASR} la distance entre les vecteurs TF $\cdot$ IDF des mots reconnus par un système de transcription automatique de la parole dans chaque scène (équation 3.2, page 82).
- Modalité SD : $\forall i, j \in [\![1, N]\!]^2, \mathcal{A}_{ij}^{\text{SD}} = d^{\text{SD}}(s_i, s_j)$ avec d^{SD} la distance basée sur le nombre de locuteurs commun entre les scènes. Les locuteurs présents dans chaque scène sont déterminés par un système automatique de segmentation et regroupement en locuteurs (équation 3.6, page 88).

Dans le but d'évaluer les méthodes de regroupement sans tenir compte des erreurs de la sortie d'un système de segmentation et regroupement en locuteurs, une quatrième matrice d'affinité est calculée :

- Modalité SD manuelle (SDM) : $\forall i, j \in [\![1, N]\!]^2, \mathcal{A}_{ij}^{\text{SDM}} = d^{\text{SDM}}(s_i, s_j)$ avec d^{SDM} la distance basée sur le nombre de locuteurs commun entre les scènes (équation 3.6, page 88), où les locuteurs présents dans chaque scène sont manuellement annotés.

Méthodes de regroupement

Nous testons trois approches pour le regroupement des scènes en histoires. Ces méthodes sont détaillées dans la Section 2.2 de l'état de l'art (page 44) :

- **Regroupement hiérarchique agglomératif** : il consiste à regrouper séquentiellement les deux groupes les plus proches jusqu'à ce qu'un critère d'arrêt soit atteint. Trois types de regroupement agglomératif sont explorés : *complete-link*, *single-link* et *average-link*.
  Le nombre de groupes est déterminé automatiquement. L'arrêt du regroupement correspond à l'étape pour laquelle la dérivée première de la distance entre les deux groupes les plus proches atteint son maximum, tel qu'illustré par la Figure 5.8.

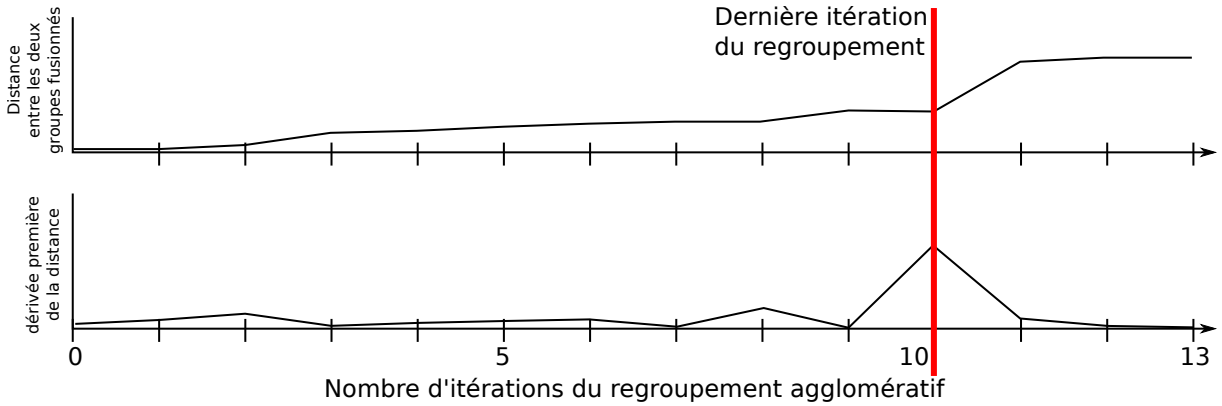

FIGURE *5.8 – Critère d'arrêt du regroupement agglomératif.*

- **Regroupement par détection de communautés dans un graphe** (utilisation de la méthode de Louvain) : les scènes sont modélisées sous forme de graphe (où chaque scène est un nœud du graphe). L'algorithme recherche les communautés de scènes par maximisation d'une mesure de modularité $\mathcal{Q}$, avec

$$\mathcal{Q} = \frac{1}{\sum_{i,j} \mathcal{A}_{ij}} \sum_{i,j} \left[\mathcal{A}_{ij} - \frac{\sum_k \mathcal{A}_{ik} \sum_k \mathcal{A}_{kj}}{\sum_{i,j} \mathcal{A}_{ij}} \right] \delta_{ij} \tag{5.8}$$

  où $\delta_{ij} = 1$ si les nœuds (scènes) i et j appartiennent à la même communauté et 0 sinon.
- **Regroupement spectral** : à partir de la matrice d'affinité $\mathcal{A}$, il consiste à projeter les scènes dans un espace de dimension K dans lequel les scènes sont regroupées par l'algorithme des K-Moyennes. Cet algorithme nécessite de connaître le nombre K de groupes (histoires) *a priori*. Pour automatiser la détermination du nombre d'histoires, plusieurs regroupements sont effectués en faisant varier la valeur de K.

La valeur finale de K est déterminée de manière à ce que le regroupement sélectionné maximise la similarité des scènes intra-groupes et minimise la similarité des scènes inter-groupes (le protocole complet est détaillé dans l'état de l'art, Section 2.2.5, page 50).

Dans cette section, nous analysons le comportement des différentes combinaisons entre mesures de distance et méthodes de regroupement.

5.3.2 Résultats et discussion

		Série		
Méthode de regroupement	**Modalité**	**Ally McBeal**	**Malcolm**	**Le Trône de Fer**
Complete-link	HSV	**0.49**	**0.52**	**0.33**
	ASR	0.71	0.61	0.64
	SD	0.63	0.58	0.63
	SDM	0.58	0.30	0.41
Single-link	HSV	0.56	0.55	0.43
	ASR	0.54	0.55	0.65
	SD	0.56	0.55	0.63
	SDM	0.52	0.42	0.41
Average-link	HSV	0.54	0.54	0.45
	ASR	0.52	0.54	0.57
	SD	0.55	0.63	0.67
	SDM	0.46	0.22	0.30
Louvain	HSV	0.57	0.53	0.45
	ASR	0.54	0.57	0.57
	SD	0.53	0.58	0.62
	SDM	0.39	0.30	0.42
Spectral	HSV	0.54	0.59	0.48
	ASR	0.67	0.61	0.55
	SD	0.62	0.56	0.63
	SDM	0.45	0.39	0.43
Aléatoire		0.64	0.54	0.55
Oracle		0.18	0.14	0.03

TABLE *5.3 – Taux d'erreur (DER) obtenu pour chaque série télévisée en fonction de toutes les combinaisons modalité/méthode de regroupement. Les cases non grisées correspondent aux regroupements meilleurs que l'aléatoire. Les valeurs en « gras » représentent le meilleur DER obtenu pour chaque série en considérant des distances obtenues à partir de descripteurs automatiques. Les valeurs « encadrées » montrent le meilleur DER obtenu pour chaque série en autorisant une annotation manuelle des locuteurs.*

La Tableau 5.3 résume tous les résultats obtenus à partir de chaque combinaison de modalité et de méthode de regroupement. Les valeurs du tableau sont le DER moyen des épisodes de chaque série télévisée.

Approches automatiques

Pour commencer, nous ne considérons que les modalités dont les distances sont basées sur des descripteurs extraits automatiquement (lignes HSV, ASR, et SD du Tableau 5.3). La meilleure approche de regroupement des scènes en histoires est un regroupement *complete-link* associé à la modalité HSV pour les trois séries du corpus. Cette modalité donne de meilleurs résultats pour les séries *Ally McBeal* (DER = 0.49 contre 0.64 pour le regroupement aléatoire) et *le Trône de Fer* (DER = 0.33 contre 0.55 pour l'aléatoire) que pour la série *Malcolm* (DER = 0.52 contre aléatoire = 0.54). Ceci peut s'expliquer par la nature de ces séries. En effet, il y a beaucoup de différences dans la manière dont *Ally McBeal*, *Malcolm* et *le Trône de Fer* sont construites :

- Dans la série *Malcolm*, chaque histoire est généralement centrée sur des évènements qui arrivent à un certain groupe de personnages, avec peu de personnages communs aux différentes histoires. Cependant, toutes les histoires ont généralement un lieu d'origine commun : la maison où vivent Malcolm ses frères et ses parents. Ainsi, des histoires différentes sont composées de scènes ayant un environnement visuel similaire, et il n'est pas étonnant que la modalité HSV donne de moins bons résultats pour cette série que pour les autres.
- Les histoires de la série *Ally McBeal* ont aussi généralement un lieu d'origine commun : le bureau d'Ally. Cependant, les environnements visuels sont plus variés et les lieux relatifs aux histoires sont mieux définis. Par exemple, une histoire se déroule en partie au tribunal tandis qu'une autre, plus proche de la vie personnelle des protagonistes, se déroule dans des lieux publics. Ceci explique la meilleure performance de la modalité HSV pour cette série que pour la série *Malcolm* par exemple. La série *Ally McBeal* est aussi la seule pour laquelle les modalités ASR et SD donnent des résultats meilleurs que le regroupement aléatoire.
- Dans la série *le Trône de Fer*, les histoires se déroulent généralement dans des lieux très différents proposant des environnements visuel variés (désert aride, désert glacé, champ, ville). C'est pourquoi la modalité HSV donne de bons résultats pour le regroupement des scènes en histoires.

Excepté pour la série *Ally McBeal*, l'utilisation des sorties d'un système de transcription automatique de la parole (modalité ASR) ou d'un système automatique de segmentation et regroupement en locuteurs (modalité SD) ne permet pas d'obtenir un regroupement meilleur qu'un regroupement aléatoire. C'est pourquoi nous avons testé les méthodes de regroupement à partir d'une annotation manuelle des locuteurs (modalité SDM).

Utilisation d'une annotation manuelle des locuteurs

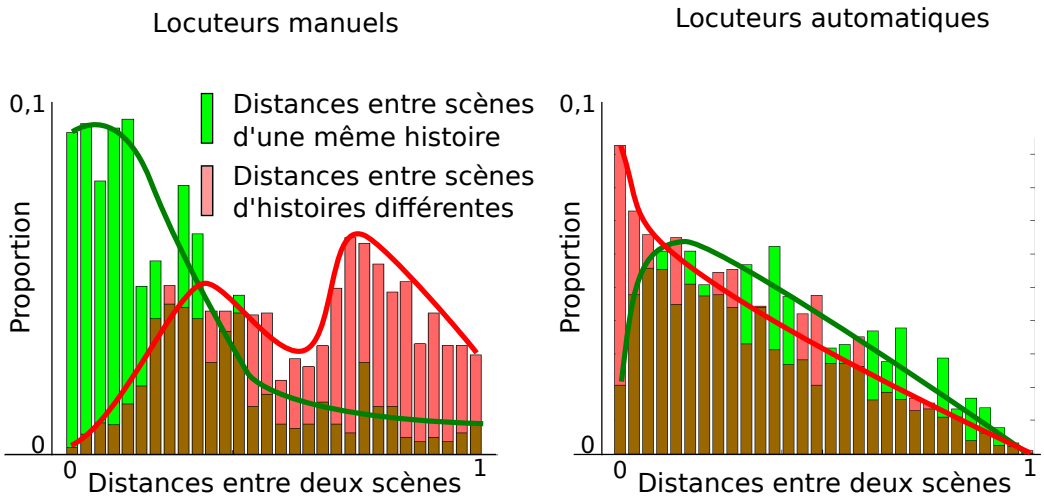

FIGURE *5.9 – Répartition des distances entre scènes pour l'ensemble des épisodes de la série* Malcolm.

La modalité SDM est la plus performante pour le regroupement des scènes en histoires pour les trois séries du corpus. Ainsi, la qualité des informations produites par le système SD a une forte influence sur la performance du regroupement des scènes en histoires. Les approches basées sur la modalité SDM sont parfois plus de deux fois plus performantes que celles basées sur les tours de parole détectés automatiquement. Les erreurs du système de segmentation et regroupement en locuteur sont discutées dans la Section 3.3.2 (page 90).

La Figure 5.9 permet d'expliquer pourquoi cette différence est si importante. Les histogrammes présentés illustrent la distribution des distances d^{SDM} et d^{SD} entre scènes d'une même histoire et celle des distances entre scènes d'histoires différentes. L'histogramme de gauche montre la distribution des distances obtenues en utilisant une annotation manuelle des locuteurs, tandis que l'histogramme de droite a été obtenu en utilisant un système automatique de segmentation et regroupement en locuteurs (locuteurs automatiques).

Considérons les distributions des distances mesurées à partir de locuteurs manuellement annotés. La distribution des distances entre scènes d'une même histoire recouvre partiellement celle des distances entre scènes d'histoires différentes. On remarque cependant que les distances très faibles entre deux scènes nous assurent qu'elles font partie d'une même histoire. À l'inverse, les distances très grandes entre deux scènes indiquent qu'elles font partie de deux histoires différentes.

Cependant, cette observation n'existe pas lorsque l'on considère la distribution des distances obtenues à partir des locuteurs automatiques. La distribution des distances entre scènes d'une même histoire et celle des distances entre scènes d'histoires différentes sont confondues. Il est donc impossible d'obtenir un regroupement correct des scènes en histoires en utilisant les tours de parole des locuteurs détectés automatiquement.

Dans le but de valider les approches de regroupement des scènes en histoires, les expériences qui seront menées à partir de maintenant seront toutes basées sur une annotation manuelle de la présence des locuteurs dans les scènes.

Sélection manuelle *vs.* automatique du nombre d'histoires

Méthode de regroupement	**Modalité**	**Ally McBeal**	**Malcolm**	**Le Trône de Fer**
Complete-link	HSV	0.54 (0.49)	0.53 (0.52)	0.44 (0.33)
	ASR	**0.54** (0.71)	**0.45** (0.61)	**0.61** (0.64)
	SDM	**0.45** (0.58)	**0.29** (0.30)	**0.31** (0.41)
Single-link	HSV	**0.55** (0.56)	0.55 (0.55)	0.48 (0.43)
	ASR	**0.54** (0.54)	**0.54** (0.55)	**0.59** (0.65)
	SDM	0.55 (0.52)	0.49 (0.42)	**0.32** (0.41)
Average-link	HSV	0.55 (0.54)	**0.51** (0.54)	**0.44** (0.45)
	ASR	0.54 (0.52)	**0.53** (0.54)	0.57 (0.57)
	SDM	**0.46** (0.46)	**0.22** (0.22)	**0.29** (0.30)
Spectral	HSV	0.55 (0.54)	**0.51** (0.59)	**0.47** (0.48)
	ASR	**0.53** (0.67)	**0.57** (0.61)	0.55 (0.55)
	SDM	**0.41** (0.45)	**0.31** (0.39)	**0.34** (0.43)
Aléatoire		0.64	0.54	0.55
Oracle		0.18	0.14	0.03

TABLE *5.4 – Taux d'erreur (DER) obtenu pour chaque série télévisée en fonction de toutes les combinaisons modalité/méthode de regroupement. Les résultats sont obtenus en connaissant* a priori *le nombre d'histoires présentes dans chaque épisode. Entre parenthèses est indiqué le résultat obtenu si le choix du nombre d'histoires est laissé à l'appréciation de la méthode de regroupement. Les cases non grisées correspondent aux regroupements meilleurs que l'aléatoire. Les valeurs en « gras » indiquent les approches pour lesquelles la connaissance du nombre d'histoires permet d'améliorer le regroupement des scènes en histoires.*

Une source d'erreurs du regroupement des scènes en histoires concerne la sélection automatique du nombre d'histoires par les algorithmes de regroupement. Le Tableau 5.4 montre les résultats obtenus si le nombre d'histoires est connu *a priori*. Ainsi, l'arrêt d'un regroupement agglomératif intervient lorsque le nombre de groupes est égal au nombre d'histoires. Le regroupement spectral prend directement le nombre d'histoires en entrée de l'algorithme. L'implémentation de l'algorithme de Louvain que nous avons à notre disposition ne permettant pas de déterminer le nombre final de groupes, il n'est pas présent dans ce tableau.

Indiquer manuellement le nombre d'histoires présentes dans une vidéo en entrée d'un *regroupement spectral* permet d'améliorer les résultats : on observe une amélioration allant de 4% à 9% en utilisant la modalité SD pour les trois séries du corpus. Cela s'explique en étudiant le comportement de la sélection automatique du nombre d'histoires par le

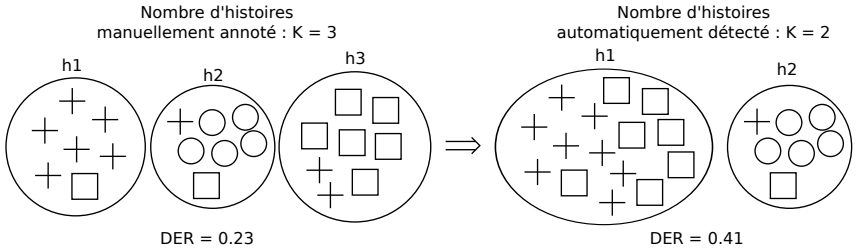

FIGURE *5.10 – Exemple de regroupement pour deux valeurs de K. Chaque entité (+, O et □) est une scène. La forme associée est l'identifiant de l'histoire annotée manuellement. Cet exemple illustre le fait que la connaissance* a priori *du nombre d'histoire permet d'obtenir un meilleur regroupement des scènes en histoires.*

regroupement spectral. En effet, on observe généralement un nombre d'histoires automatiquement détecté bien inférieur au nombre d'histoires manuellement annotées. Ainsi beaucoup de scènes sont incorrectement regroupées dû à une mauvaise sélection de ce nombre comme illustré par la Figure 5.10.

On observe souvent le comportement inverse dans de nombreux cas du *regroupement agglomératif*, et plus particulièrement pour un regroupement *complete-link* basé sur la modalité HSV. Dans ce cas, la connaissance du nombre d'histoires dégrade les résultats (jusqu'à un DER supérieur de 11% pour la série *Le Trône de Fer*). Cette approche de regroupement tend à détecter un nombre d'histoires plus grand que le nombre d'histoires manuellement annotées. Les premières étapes du regroupement permettent généralement de regrouper correctement des scènes appartenant à une même histoire. Cependant, fixer le nombre d'histoires manuellement force le système à regrouper des scènes pour lesquelles il n'est pas en mesure de décider si elles doivent être regroupées ou non. C'est pourquoi, comme illustré par la Figure 5.11, il est parfois préférable d'arrêter le regroupement agglomératif avant d'atteindre le nombre d'histoire annotées.

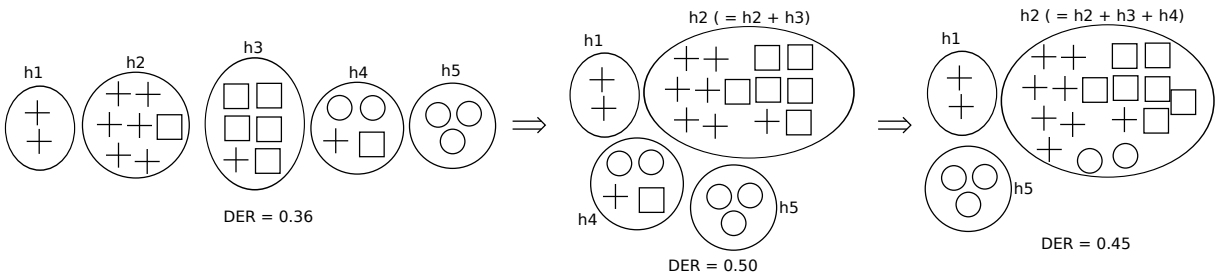

FIGURE *5.11 – Exemple de regroupement agglomératif. Chaque entité (+, O et □) est une scène. La forme associée est l'identifiant de l'histoire annotée manuellement. Cet exemple illustre le fait qu'il est parfois préférable d'arrêter le regroupement avant d'atteindre le nombre réel d'histoires.*

Détail du comportement du regroupement agglomératif

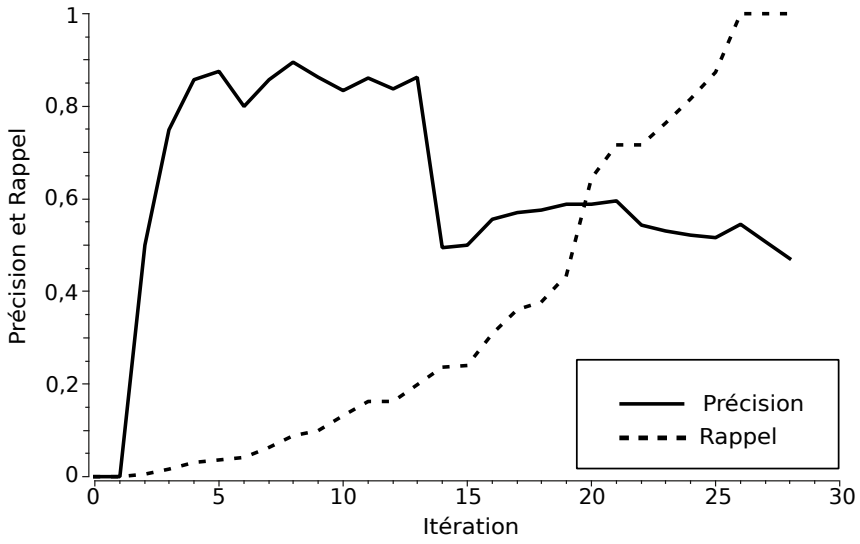

FIGURE *5.12 – Evolution de la précision et du rappel après chaque itération d'un clustering agglomératif.*

La Figure 5.12 montre l'évolution de la précision et du rappel après chaque itération d'un regroupement agglomératif de type *average-link* appliqué à la modalité HSV. La courbe montre que, dans ce cas particulier, la précision est très élevée jusqu'à la $14^{ème}$ itération : peu de regroupements erronés de scènes sont effectués. Cependant, deux groupes correspondant à deux histoires différentes sont agglomérés à la $14^{ème}$ itération, ce qui résulte en une très forte diminution de la précision. Ceci montre les limites des approches agglomératives proposées : les itérations suivantes ne sont pas en mesure de corriger cette mauvaise décision (puisque le regroupement n'est qu'agglomératif), et un seul descripteur ne peut pas toujours porter toute l'information sémantique nécessaire pour résoudre la tâche de regroupement des scènes en histoires.

5.4 Regroupement multimodal

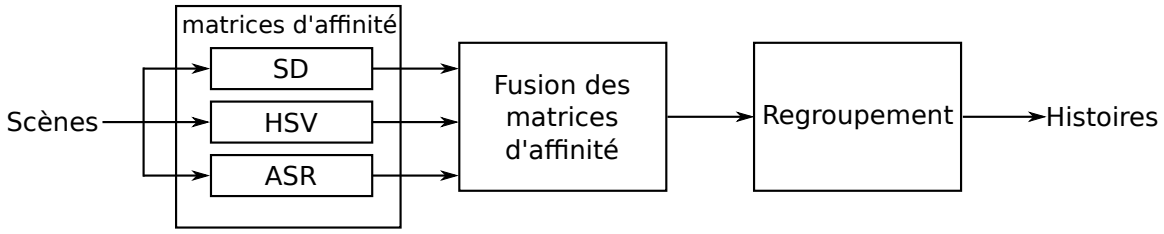

FIGURE *5.13 – Étapes du regroupement multimodal des scènes en histoires.*

Puisqu'une seule modalité ne permet pas d'obtenir un bon regroupement des scènes en histoires, nous proposons une approche de fusion des trois modalités (HSV, SD et ASR) comme illustré par la Figure 5.13. Après un calcul des matrices d'affinité obtenues à partir de chaque modalité, une matrice d'affinité multimodale est calculée et utilisée en entrée d'un algorithme de regroupement.

Dans la section précédente, nous avons vu que les erreurs du système de segmentation et regroupement en locuteurs utilisé impliquent qu'une distance basée sur la modalité SD n'est pas pertinente pour le regroupement des scènes en histoires. Pour valider les approches de fusion multimodale, une annotation manuelle des locuteurs est utilisée pour le calcul de la distance d^{SD} pour toutes les expériences qui seront menées dans la suite de ce chapitre.

5.4.1 Méthode développée

L'approche de fusion étudiée propose de fusionner les trois modalités présentées précédemment pour obtenir une matrice d'affinité unique notée $\mathcal{A}^{\text{FUS}}$. La fusion proposée utilise les trois matrices d'affinité utilisées pour les trois modalités HSV, SD et ASR telle que

$$\forall (i,j) \in [\![1,N]\!]^2, \quad \mathcal{A}_{ij}^{\text{FUS}} = 1 - \left(e^{-\frac{(\mathcal{A}_{ij}^{\text{HSV}})^2 + (\mathcal{A}_{ij}^{\text{SD}})^2 + (\mathcal{A}_{ij}^{\text{ASR}})^2}{2\sigma^2}} \right) \tag{5.9}$$

avec $\sigma^2 = \max_{(i,j)\in[\![1,N]\!]^2} ((\mathcal{A}_{ij}^{\text{HSV}})^2 + (\mathcal{A}_{ij}^{\text{SD}})^2 + (\mathcal{A}_{ij}^{\text{ASR}})^2)$.

Ce type de calcul de matrice d'affinité est très utilisé dans le cadre du regroupement spectral.

Méthode de regroupement	**Modalité**	**Ally McBeal**	**Malcolm**	**Le Trône de Fer**	**Tous**
Average-link	SDM	0.46	**0.22**	**0.30**	**0.34**
	SDM+HSV	0.63	0.65	0.64	0.63
	SDM+HSV+ASR	0.77	0.64	0.65	0.68
Spectral	SDM	0.45	0.39	0.43	0.43
	SDM+HSV	**0.43**	0.33	0.42	0.38
	SDM+HSV+ASR	0.45	0.32	0.36	0.38
Aléatoire		0.64	0.54	0.55	0.57
Oracle		0.18	0.14	0.03	0.11

TABLE *5.5 – Comparaison des approches de fusion pour le regroupement des scènes en histoires (DER). Les cases non grisées correspondent aux regroupements meilleurs que l'aléatoire.*

5.4.2 Résultats

Le Tableau 5.5 résume les résultats obtenus pour chaque série télévisée et pour l'ensemble du corpus avec différentes approches de fusion. Nous comparons la performance des méthodes de regroupement *average-link* (obtenant le meilleur score de regroupement monomodal) avec un regroupement spectral (pour lequel l'approche de fusion a été développée). Dans les deux cas, la matrice d'affinité $\mathcal{A}^{\text{FUS}}$ est utilisée en entrée de l'approche de regroupement.

Les performances de ces algorithmes sont indiquées en utilisant uniquement la modalité SDM, une combinaison des modalités SDM et HSV (SDM+HSV) ou la combinaison des trois modalités (SDM+HSV+ASR).

5.4.3 Discussion

Le Tableau 5.5 confirme les résultats présentés précédemment : la modalité SD donne les meilleurs résultats quand elle est utilisée avec un regroupement *average-link* pour presque toutes les séries. La fusion des distances ne permet pas d'améliorer un regroupement *average-link*. Cependant, elle montre des résultats prometteurs pour le regroupement spectral. En effet, ce dernier obtient les meilleurs résultats en utilisant une matrice d'affinité multimodale plutôt qu'une matrice basée sur la modalité SD uniquement. De plus, la fusion des modalités SD et HSV donne le meilleur résultat pour la série *Ally McBeal*.

Intérêt de la fusion multimodale pour un regroupement spectral

La Figure 5.14 montre l'amélioration apportée par la fusion des trois modalités en utilisant un regroupement spectral sur un épisode de la série *Ally McBeal*. Elle représente le résultat des K-Moyennes obtenu dans l'espace spectral en utilisant la matrice d'affinité $\mathcal{A}^{\text{SD}}$ (en 5.14**a** à gauche de la figure) ou la matrice d'affinité $\mathcal{A}^{\text{FUS}}$ (en 5.14**b** à droite de la

figure). Les symboles « △, ◊ » indiquent les scènes qui doivent être regroupées dans une même histoire, les « ◯ » représentent les scènes qui n'appartiennent à aucune histoire, et les « + » sont les scènes qui appartiennent à plusieurs histoires. Les cercles décrivent les trois groupes obtenus par l'algorithme des K-Moyennes. Dans les deux cas, trois groupes ont été formés par l'approche de regroupement spectral, pour deux histoires annotées dans l'épisode donné en exemple.

Il y a peu de différences dans ces deux exemples, mais on remarque que les trois groupes sont mieux séparés lorsque l'on utilise la matrice d'affinité $\mathcal{A}^{\text{FUS}}$ (5.14**b**) qu'en utilisant la matrice d'affinité $\mathcal{A}^{\text{SD}}$ (5.14**a**). Ainsi le regroupement K-Moyennes effectué à la fin du regroupement spectral est plus simple à réaliser et il permet d'éviter des erreurs comme en 5.14**a**, où deux scènes ◊ sont regroupées à deux scènes ◯ alors qu'elles sont correctement regroupées en 5.14**b**.

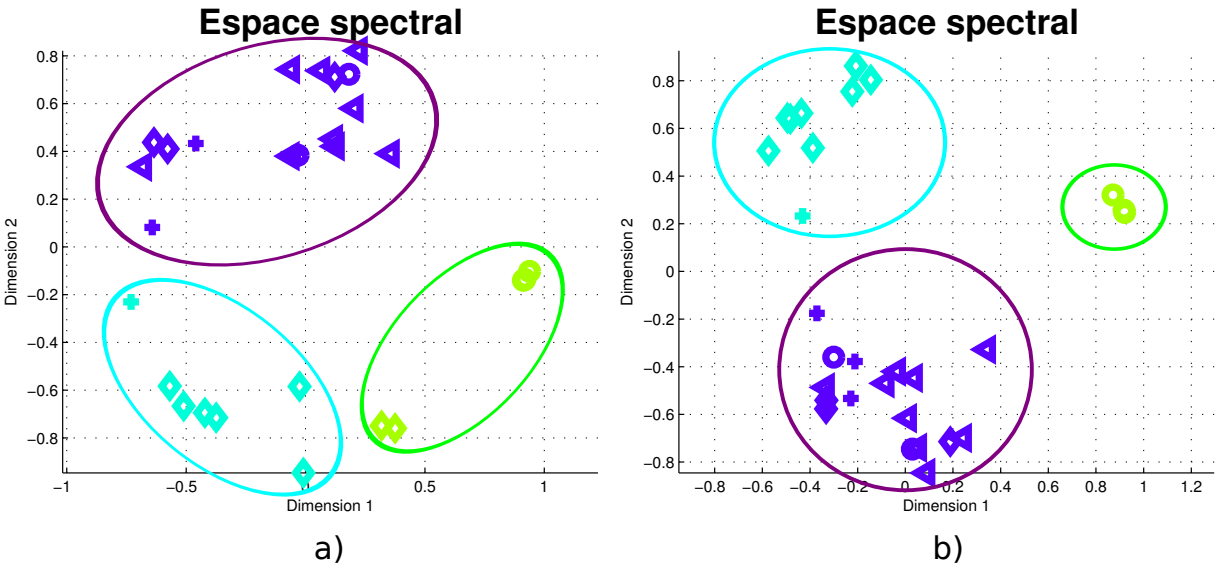

FIGURE *5.14 – Résultat d'un regroupement spectral. Les scènes sont tracées dans l'espace spectral et les cercles indiquent le résultat des K-Moyennes dans cet espace. a) résultat du regroupement spectral en utilisant la modalité SD. b) résultat du regroupement spectral en utilisant la fusion multimodale. Chaque symbole (□, △, ◊) indiquent les scènes qui doivent être regroupées dans une même histoire, les « ◯ » représentent les scènes qui n'appartiennent à aucune histoire, et les « + » sont les scènes qui appartiennent à plusieurs histoires.*

Différences entre les épisodes

En examinant chaque épisode, on remarque que la combinaison d'un regroupement *average-link* avec la modalité SD donne un excellent regroupement des scènes en histoires pour 11 des 22 épisodes annotés (pour un DER moyen de 0.26 contre 0.50 pour les autres épisodes).

Ces épisodes contiennent en général 3 ou 4 histoires indépendantes centrées sur des communautés de personnages disjointes. Nous les appelons des *Épisodes Dirigés par les*

Épisodes	Average-link + SD	Spectral + FUS
EdP	**0.26**	0.35
$\overline{\texttt{EdP}}$	0.50	**0.43**

TABLE *5.6 – Comparaison des EdP et des $\overline{\texttt{EdP}}$ (DER).*

Personnages (EdP). Ce type d'épisode n'est pas spécifique à un genre de série particulier. On les retrouve dans les trois séries de notre corpus. Les autres épisodes proposent des histoires qui sont centrées sur un sujet particulier plutôt que sur un groupe de personnages : ils sont notés $\overline{\texttt{EdP}}$.

Le Tableau 5.6 montre les résultats moyens obtenus pour ces épisodes (EdP) ou les autres épisodes ($\overline{\texttt{EdP}}$) avec un regroupement *average-link* utilisant la modalité SD seule, et un regroupement spectral utilisant la fusion des trois modalités. Pour les EdP, le regroupement *average-link* associé à la modalité SD donne les meilleurs résultats (0.26 contre 0.35 pour un regroupement spectral basé sur la matrice d'affinité $\mathcal{A}^{\text{FUS}}$). Au contraire, le regroupement spectral obtient un DER de 7% meilleur qu'un regroupement *average-link* pour les $\overline{\texttt{EdP}}$.

Dans la section suivante nous étudions une méthode permettant de déterminer automatiquement la meilleure méthode de regroupement à appliquer pour chaque épisode, entre un regroupement *average-link* associé à la modalité SD seule et un regroupement *spectral* associé à la fusion des modalités.

5.5 Sélection automatique de la meilleure méthode de regroupement

Dans les sections précédentes, nous avons investigué l'utilisation de différentes approches de regroupement. Dans la Section 5.3, nous avons testé chaque combinaison de modalité (HSV, SD, ASR) et d'approche de regroupement (agglomératif, Louvain, spectral). Dans la Section 5.4 nous proposons une approche de fusion de ces trois modalités.

Dans cette section, nous proposons une méthode hiérarchique pour le regroupement des scènes en histoires, illustrée par la Figure 5.15, qui consiste à sélectionner la meilleure approche de regroupement à appliquer pour chaque épisode. Cette sélection propose de choisir quelle est la meilleure méthode de regroupement à appliquer à chaque épisode entre les deux approches suivantes :

- un regroupement *average-link* basé sur la matrice d'affinité $\mathcal{A}^{\text{SD}}$ (ALSD) : en moyenne, cette approche donne les meilleurs résultats dans les sections précédentes ;
- un regroupement *spectral* basé sur la matrice d'affinité $\mathcal{A}^{FUS}$ (SCFUS) : cette approche donne les meilleurs résultats pour certains épisodes particuliers.

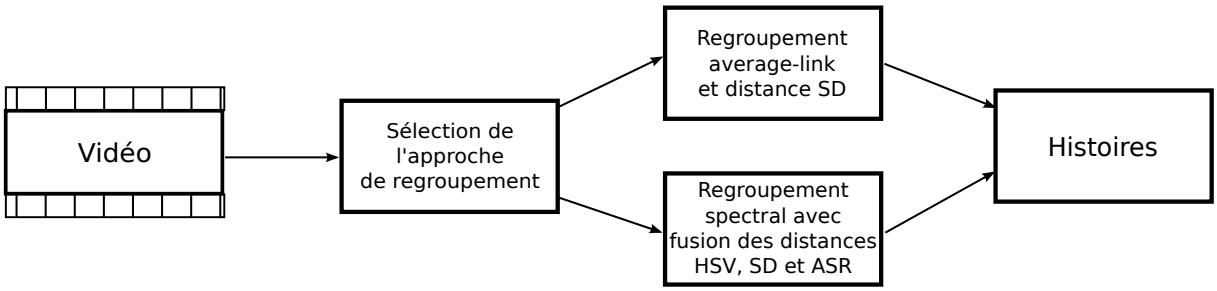

FIGURE *5.15 – Méthode de regroupement des scènes en histoires par sélection de la méthode de regroupement*

5.5.1 Approche de sélection automatique

Notre idée est de modéliser la relation entre les personnages de façon à déterminer un critère permettant la distinction entre ces épisodes. L'approche de regroupement `ALSD` est basée uniquement sur la connaissance des locuteurs présents dans chaque scène pour regrouper les scènes en histoires. Ainsi, si chaque histoire racontée dans un épisode décrit les aventures d'un groupe de personnages qui ne sont que peu impliqués dans les autres histoires, alors l'approche de regroupement `ALSD` permet un regroupement pertinent des scènes. Sinon, si chaque histoire voit intervenir beaucoup de personnages en commun, la connaissance des personnages présents dans chaque scène n'est pas suffisante pour obtenir un bon regroupement des scènes. L'approche `SCFUS` basée sur la fusion des modalités SD, ASR et HSV est alors plus pertinente.

Différences entre les épisodes : graphe social des personnages

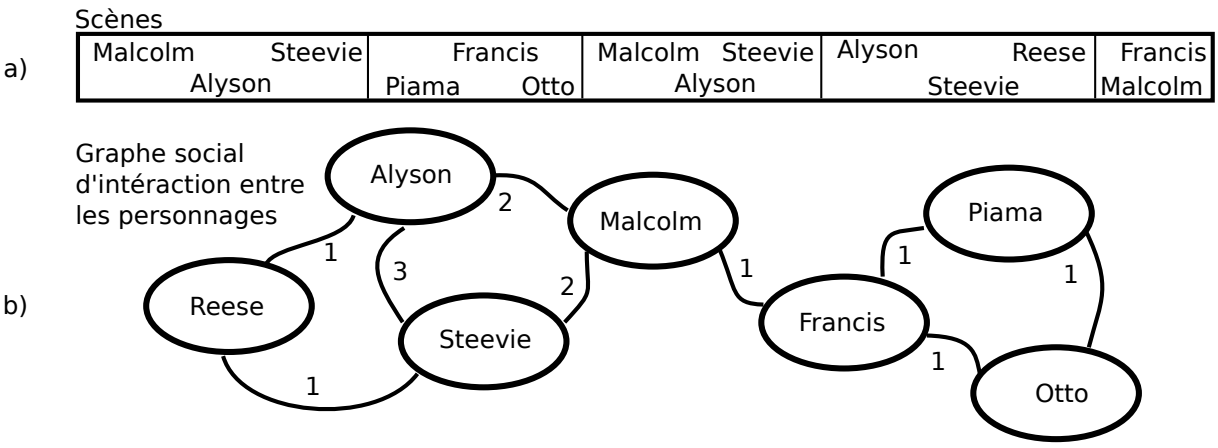

FIGURE *5.16 – (a) Liste des personnages dans chaque scène. (b) Graphe social d'interaction entre les personnages.*

Pour modéliser la relation entre les personnages d'une vidéo, notre approche est inspirée par les graphes sociaux d'interaction des personnages introduits par Weng *et al.* [Weng 2009]. La Figure 5.16 montre la construction d'un tel graphe. Chaque personnage (ou locuteur dans notre cas) est associé à un nœud du graphe. Un arc entre deux nœuds signifie que les personnages correspondants apparaissent dans au moins une scène commune. Ces arcs sont pondérés par le nombre de scènes communes aux deux personnages. Il en résulte un graphe représentant l'interaction sociale entre les personnages de l'épisode.

La Figure 5.17 montre deux graphes d'interaction entre les personnages. Ils sont obtenus à partir de deux épisodes : le premier est un épisode pour lequel l'approche `ALSD` est la plus performante, et le meilleur regroupement est obtenu avec l'approche `SCFUS` pour le deuxième.

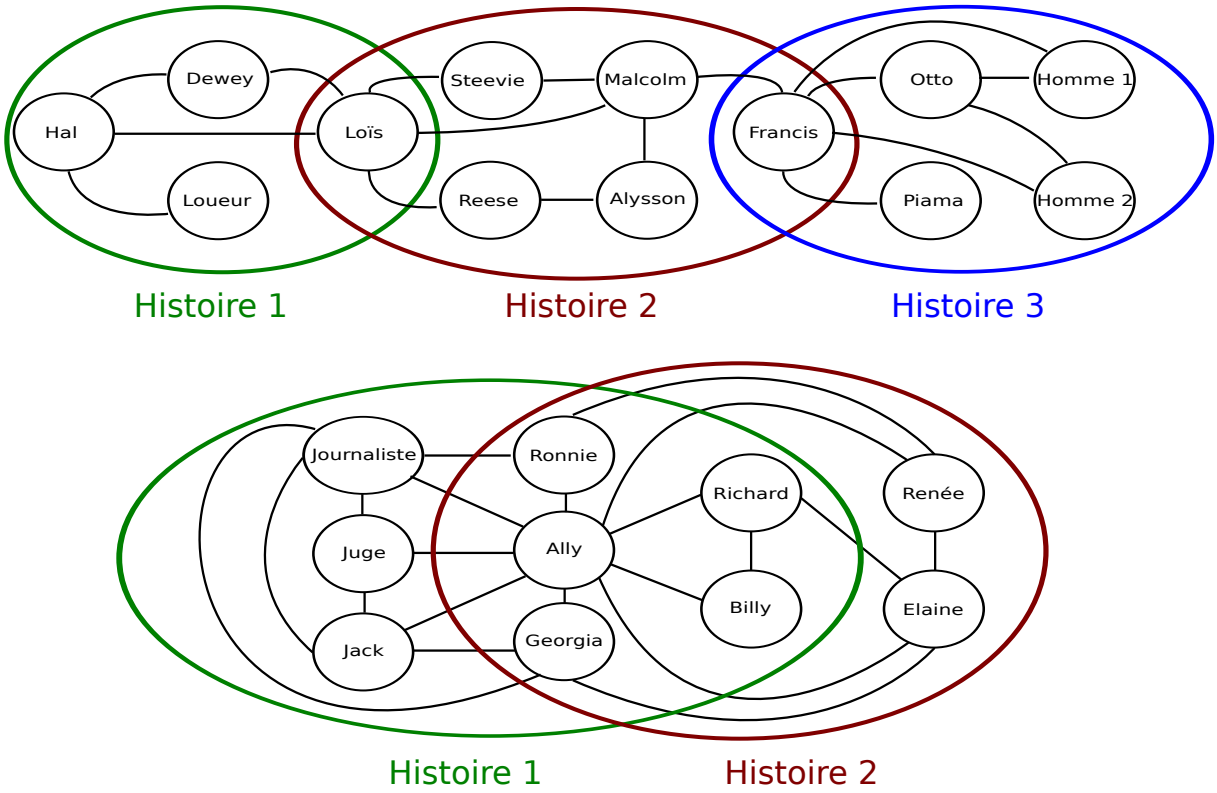

FIGURE *5.17 – Graphes d'interaction entre les personnages. Le premier décrit la relation entre les personnages d'un épisode de la série* Malcolm *pour lequel la meilleure approche de regroupement est l'approche* `ALSD`*. Le deuxième montre la relation entre les personnages d'un épisode de la série* Ally McBeal *pour lequel la meilleure approche de regroupement est l'approche* `SCFUS`*. Les arcs indiquent une relation forte entre les personnages (présence simultanée dans une scène). Il y a peu de liens entre les personnages intervenant dans des histoires différentes pour l'épisode de* Malcolm*. Au contraire, il y a peu de distinction entre les personnages impliqués dans les différentes histoires pour l'épisode d'*Ally McBeal*.*

Les cercles représentant les histoires englobent tous les personnages impliqués dans chaque histoire manuellement annotée. Pour le premier épisode, il existe peu d'interactions entre les personnages intervenant dans deux histoires différentes. Au contraire, dans le deuxième, les deux histoires de cet épisode englobent la majorité des personnages et il y a une forte interaction entre tous les personnages. C'est cette distinction que nous avons choisi d'utiliser pour permettre la sélection de la méthode de regroupement.

Critère de sélection

À partir des graphes d'interaction des personnages, nous proposons d'utiliser l'algorithme de Louvain [Blondel 2008] décrit dans l'état de l'art, Section 2.2 page 44, pour rechercher les communautés de personnages. C'est une méthode de regroupement heuristique basée sur la maximisation d'une quantité appelée modularité et notée $\mathcal{Q}$. $\mathcal{Q}$ peut être vue comme une mesure de la qualité des communautés détectées. Elle est grande lorsque les personnages ont globalement des liens forts avec les personnages de leur communauté et des liens faibles avec les personnages des autres communautés [Newman 2006]. Nous proposons d'utiliser cette valeur $\mathcal{Q}$ pour détecter automatiquement la meilleure méthode de regroupement à appliquer :

- une grande modularité signifie que les communautés de personnages sont fortement séparées. Ainsi, nous pouvons supposer que chaque communauté est impliquée dans une histoire qui lui est particulière, et l'approche de regroupement `ALSD` est la plus pertinente.
- une faible modularité signifie qu'il y a peu de distinctions entre les communautés détectées. Ainsi, une description des scènes basée uniquement sur les personnages n'est pas suffisante pour le regroupement des scènes en histoires, et l'approche `SCFUS` doit être préférée.

La Figure 5.18 montre la modularité de chaque épisode du corpus en fonction de la valeur DER obtenue par un regroupement `SCFUS` (DER_{SP}) et un regroupement `ALSD` (DER_{AL}). Comme la métrique DER est un mesure d'erreur, $log(\text{DER}_{SP}/\text{DER}_{AL}) > 0$ si le regroupement `ALSD` donne le meilleur résultat. À l'inverse, $log(\text{DER}_{SP}/\text{DER}_{AL}) < 0$ si le meilleur résultat est obtenu par un regroupement `SCFUS`. Pour une modularité $\mathcal{Q}$ inférieure à une valeur ϕ, l'approche de regroupement la plus pertinente est toujours `SCFUS`. Lorsque $\mathcal{Q} > \phi$, l'approche de regroupement la plus pertinente est très souvent `ALSD`. La Figure 5.18 prouve donc que la modularité est un critère efficace pour la sélection automatique de la meilleure approche de regroupement. Ainsi, pour un épisode e, la méthode de regroupement utilisée est :

– l'approche `ALSD` si $\mathcal{Q} > \phi$;
– l'approche `SCFUS` sinon.

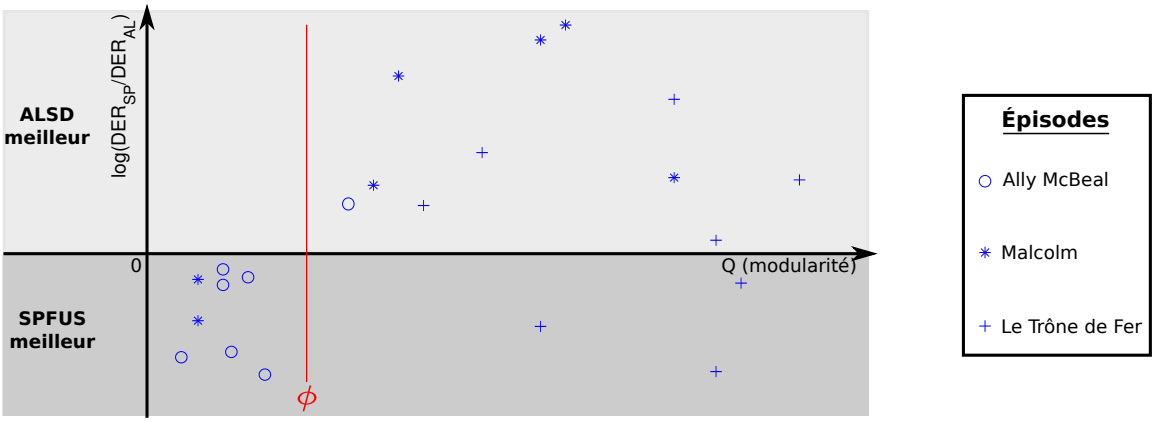

FIGURE 5.18 – *Modularité des épisodes en fonction de la valeur DER obtenue par un regroupement* `SCFUS` *(DER_{SP}) et un regroupement* `ALSD` *(DER_{AL}). $log(DER_{SP}/DER_{AL}) > 0$ si le regroupement* `ALSD` *donne le meilleur résultat. Pour une modularité $\mathcal{Q} < \phi$, l'approche de regroupement la plus pertinente est toujours* `SCFUS`*. Lorsque $\mathcal{Q} > \phi$, l'approche de regroupement la plus pertinente est très souvent* `ALSD`*. $\mathcal{Q}$ peut donc être utilisé comme critère de sélection de la meilleure approche de regroupement à appliquer.*

Optimisation de ϕ

Nous cherchons à déterminer automatiquement la valeur optimale ϕ de manière à maximiser le nombre d'épisodes pour lesquels la meilleure approche de regroupement est correctement sélectionnée. Puisque nous n'avons que 22 épisodes annotés en histoires, la quantité de données est insuffisante pour constituer un ensemble d'apprentissage et un ensemble de test. C'est pourquoi le protocole d'optimisation du paramètre ϕ suit le principe de la validation croisée (*leave-one-out cross validation*). Ainsi, le paramètre $\phi(e)$ pour un épisode $e \in \mathcal{E}$ est déterminé tel que :

$$\phi(e) = \underset{\phi \in \mathbb{R}^+}{\operatorname{argmax}}\ card\{e' \in \mathcal{E} \backslash e \ \ / \ \mathcal{Q} > \phi \wedge DER_{SP}(e') > DER_{AL}(e')\} \tag{5.10}$$

5.5.2 Résultats

Le Tableau 5.7 résume les résultats obtenus pour toutes les collections du corpus d'évaluation. La ligne $e^{\texttt{ALSD}}$ correspond aux résultats moyens des épisodes pour lesquels le regroupement sélectionné est un regroupement `ALSD`. La ligne $e^{\texttt{SCFUS}}$ correspond aux épisodes pour lesquels le regroupement `SCFUS` a été sélectionné.

Pour les épisodes $e^{\texttt{ALSD}}$, le regroupement *average-link* obtient des résultats meilleurs de 9% comparé au regroupement spectral, alors que ce dernier donne un résultat de 7% meilleur que le regroupement *average-link* pour les épisodes $e^{\texttt{SCFUS}}$. Ainsi, la méthode de regroupement des scènes en histoires basée sur une sélection de la meilleure approche

de regroupement permet d'améliorer les résultats de 2% sur l'ensemble de la collection comparé au meilleur système précédent (regroupement *average-link* avec la modalité SD).

Épisodes	Average-link + SD	Spectral + FUS	Sélection automatique	Aléa.	Oracle
Ally McBeal	0.46	**0.45**	**0.43**	0.64	0.18
Malcolm	**0.22**	0.33	**0.21**	0.54	0.14
Le Trône de Fer	**0.30**	0.36	**0.30**	0.55	0.03
Tous	**0.34**	0.38	**0.32**	0.57	0.11
$e^{\texttt{ALSD}}$	**0.26**	0.35	**0.26**	0.51	0.05
$e^{\texttt{SCFUS}}$	0.50	**0.43**	**0.43**	0.61	0.19

TABLE *5.7 – Comparaison de différents ensembles d'épisodes pour les approches de regroupement ALSD, SCFUS et l'approche par sélection automatique de la méthode de regroupement.*

5.5.3 Discussion

Résultat de la classification des épisodes

La Figure 5.19 permet d'observer le résultat de la sélection automatique de la meilleure approche de regroupement. Chaque point représente un épisode qui est tracé sur la figure en fonction du résultat du DER obtenu en utilisant les approches de regroupement `ALSD` et `SCFUS`. La forme associée à chaque point indique quelle méthode a été sélectionnée : un « + » si la méthode retenue pour l'épisode est l'approche `ALSD`, et un « O » si l'approche retenue est `SCFUS`.

La méthode de classification permet de choisir la meilleure méthode de regroupement à appliquer pour 19 des 22 épisodes du corpus. Au total, l'approche `ALSD` a été automatiquement sélectionné pour 14 épisodes et l'approche `SCFUS` pour les 8 restants.

Le Tableau 5.8 montre la répartition du nombre d'épisodes en fonction de l'approche de regroupement des scènes retenue pour chaque série du corpus. L'approche de regroupement `ALSD` n'a été retenue que pour 14% des épisodes de la série *Ally McBeal* (1 épisode sur 7). Le format de cette série permet d'expliquer ce phénomène, puisque beaucoup de personnages participent a plusieurs histoires dans chaque épisode. Ainsi, la modularité du graphe de relation des personnages a tendance à être très faible pour la majorité des épisodes de la série *Ally McBeal*, ce qui explique pourquoi l'approche `ALSD` n'a été retenue que pour un seul épisode.

Pour la série *Malcolm*, l'approche de regroupement `ALSD` a été retenue pour 72% des épisodes du corpus (5 épisodes sur 7). Dans cette série, la majorité des épisodes est composée d'histoires suivant des personnages différents, ce qui explique pourquoi le graphe de relation des personnages obtient une forte modularité pour la majorité des épisodes de cette série.

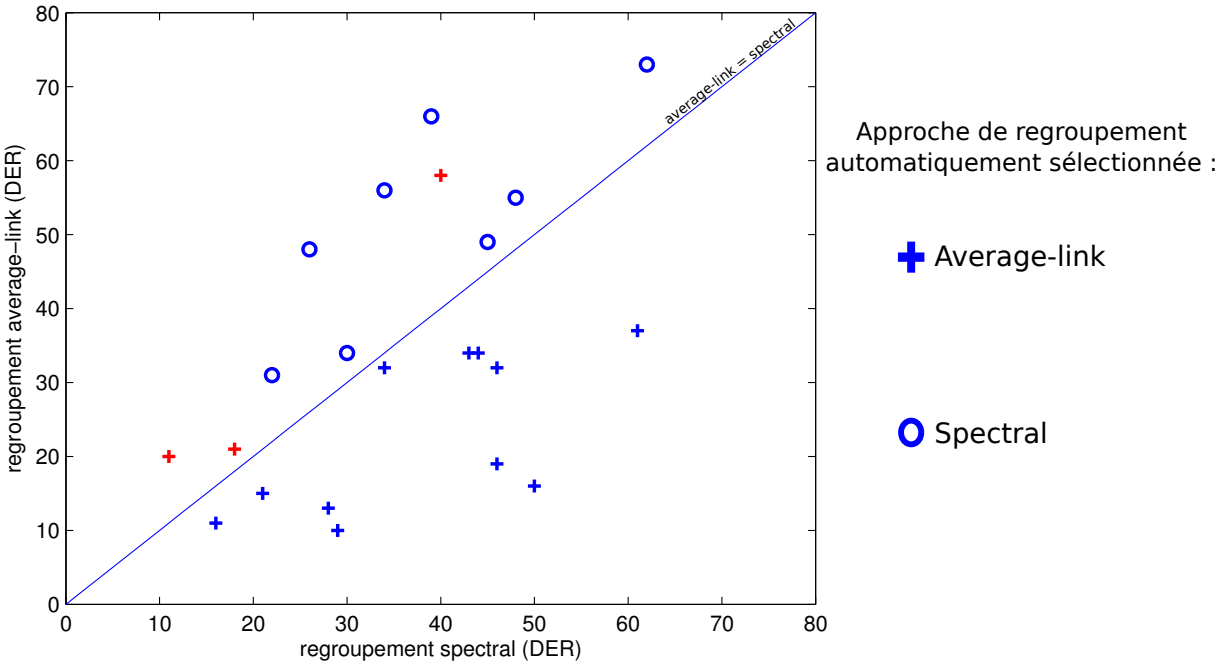

FIGURE *5.19 – DER obtenu par un regroupement average-link ou spectral pour tous les épisodes du corpus. Les « + » représentent les épisodes pour lesquels l'approche* average-link *a été sélectionnée. Les « O » correspondent aux épisodes pour lesquels le regroupement* spectral *a été automatiquement sélectionné. Dans l'idéal, les « + » doivent être tracés sous la ligne oblique (regroupement* average-link $<$ spectral*), et les « O » au dessus (regroupement* spectral $>$ average-link*). En rouge les épisodes mal classés.*

Pour tous les épisodes de la série *le Trône de Fer* l'approche de regroupement retenue est l'approche `ALSD`. Cette série propose des épisodes composés de beaucoup de personnages (en moyenne 35 par épisodes). Or, comme la majorité des différentes histoires se déroule dans des lieux différents, il y a peu d'interactions entre les personnages principaux et aucune entre les nombreux personnages moins importants qui sont présents dans les différentes histoires.

Cependant, la table de vérité pour cette série dans le Tableau 5.8 montre que la meilleure méthode de regroupement a été mal sélectionnée pour 37% des épisodes (3 épisodes sur 8). L'approche de regroupement `ALSD` est sélectionnée alors que l'approche `SCFUS` donne de meilleurs résultats. Pour deux de ces épisodes, à la fois l'approche `ALSD` et `SCFUS` donnent de bons résultats. Elles obtiennent respectivement un DER de 0.21 et 0.18 pour le premier et 0.20 et 0.11 pour le deuxième. Bien que l'approche `SCFUS` soit plus performante, la sélection de l'approche `ALSD` reste pertinente.

Le troisième épisode montre une particularité qui permet d'expliquer pourquoi la modularité n'est pas en mesure de choisir la meilleure approche de regroupement. Trois histoires différentes de cet épisode ont un grand nombre de personnages communs. L'ap-

Ally McBeal

Classification optimale \ Classification automatique	ALSD	SCFUS
ALSD	14%	0%
SCFUS	0%	86%

Malcolm

Classification optimale \ Classification automatique	ALSD	SCFUS
ALSD	72%	0%
SCFUS	0%	28%

Le Trône de Fer

Classification optimale \ Classification automatique	ALSD	SCFUS
ALSD	63%	0%
SCFUS	37%	0%

Tous les épisodes

Classification optimale \ Classification automatique	ALSD	SCFUS
ALSD	50%	0%
SCFUS	14%	36%

TABLE *5.8 – Répartition des épisodes en fonction de l'approche de regroupement des scènes retenue pour chaque série du corpus.*

proche de regroupement ALSD n'est donc pas en mesure de fournir un regroupement des scènes en histoires correct. Cependant, une histoire de l'épisode voit évoluer une communauté de personnages ne rencontrant jamais les personnages des autres histoires. C'est pourquoi la modularité est très forte (les communautés de personnages sont très séparées) et l'approche sélectionnée est l'approche ALSD.

5.6 Conclusion

À notre connaissance, la tâche de regroupement des scènes en histoires pour des épisodes de séries télévisées n'a jamais été étudiée à ce jour. Nos contributions concernent plusieurs points :

- Nous avons exploré l'utilisation de plusieurs approches de regroupements des scènes et des mesures d'affinité entre les scènes à partir de 3 modalités : la couleur (HSV), la présence de locuteurs dans les scènes (SD), les mots prononcés par les personnages (ASR).
- Nous proposons une approche de fusion de ces trois modalités.
- Les approches proposées ayant un comportement très différent en fonction des épisodes, nous proposons une méthode de sélection automatique de la meilleure mé-

thode de regroupement à appliquer à chaque épisode. Nous montrons que cette approche de sélection permet d'améliorer globalement les résultats du regroupement des scènes en histoires de 2%, et qu'elle permet de sélectionner la meilleure méthode de regroupement pour 19 des 22 épisodes du corpus de test.

Pour améliorer les résultats du regroupement des scènes en histoires, plusieurs pistes peuvent être suivies. Les approches de regroupement utilisées dans ce chapitre ne peuvent associer une scène à plusieurs histoires. Faire en sorte que deux histoires convergent à un moment d'un épisode puis divergent (comme illustré dans la Figure 5.15) est une pratique commune dans la construction des épisodes de séries télévisées modernes. Notre corpus d'évaluation contient 774 scènes dont 50 appartiennent à au moins deux histoires. Le DER moyen de 0.11 indiqué par la ligne *oracle* dans le Tableau 5.5 le montre bien : c'est le taux d'erreur le plus bas que l'on peut atteindre avec les méthodes de regroupement proposées. Il est donc nécessaire de détecter automatiquement la présence de ces scènes et d'être en mesure de les associer à plus d'une histoire.

Une autre piste concerne la détermination automatique du nombre d'histoires. Nous avons vu qu'en indiquant manuellement le nombre d'histoires présentes dans un épisode, les résultats d'un regroupement spectral sont améliorés de 4%. Une amélioration de la méthode de sélection du nombre optimal d'histoires doit permettre d'améliorer le regroupement des scènes en histoires.

Chapitre 6

Applications

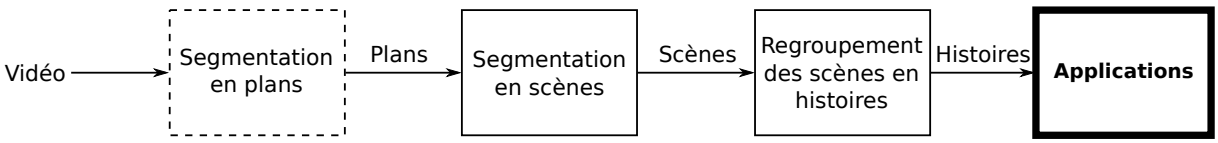

FIGURE *6.1 – Utilisation de la structuration d'une vidéo : applications.*

Comme illustré dans la Figure 6.1, ce chapitre présente les applications utilisant le système de structuration d'épisodes de séries télévisées présenté dans les chapitres précédents. La structuration proposée s'articule autour de deux niveaux : segmentation en scènes d'une part et regroupement des scènes en histoires d'autre part. L'extraction automatique de cette structure peut avoir des applications très variées :

- elle peut servir à enrichir les outils d'indexation de bases de données de vidéos, en permettant une nouvelle catégorisation des séries (par exemple, série développant plusieurs arcs narratifs ou séries linéaires) ;
- elle peut être utilisée dans un système de navigation de vidéo pour permettre un visionnage non linéaire de l'épisode en se concentrant sur certaines histoires particulières ;
- utilisée en entrée d'un système de génération automatique de résumés de vidéos, elle permet de mettre en valeur chaque histoire racontée dans l'épisode.

Nos contributions portent essentiellement sur le deuxième point. Nous avons développé une interface web de navigation de vidéos, StoViz, pour illustrer les résultats de notre segmentation en scènes et du regroupement des scènes en histoires.

StoViz a été développé pour deux applications distinctes :

- **Une application de visualisation et de manipulation de la vidéo**. Elle permet d'avoir un rapide aperçu de la structure narrative d'un épisode, en offrant la possibilité de suivre une histoire particulière indépendamment du reste de l'épisode.
- **Une application d'analyse des erreurs** du système de segmentation en scènes et du système de regroupement des scènes en histoires. Elle permet d'accompagner le chercheur en permettant une lecture pertinente des erreurs des systèmes développés. Cette application peut elle-même être divisée en deux modes distincts.
  - Un mode **segmentation en scènes** qui permet d'évaluer et d'analyser le résultat d'un système automatique de segmentation en scènes.
  - Un mode **regroupement des scènes en histoires** qui permet d'évaluer et d'analyser le résultat d'un système de regroupement des scènes en histoires.

Ce chapitre présente les différentes applications de StoViz. Après une présentation d'outils de navigation de vidéo existants (Section 6.1), une description de l'interface de StoViz est proposée (Section 6.2). La Section 6.3 décrit l'application de visualisation et de manipulation de la vidéo. Ensuite, la section suivante discute de l'application d'évaluation des systèmes de structuration développés. Enfin, la Section 6.5 présente la technologie utilisée pour développer StoViz.

StoViz est une interface web fonctionnant sur les navigateurs récents tels que Firefox, Chrome ou Safari. Elle est disponible en ligne à l'adresse www.irit.fr/recherches/SAMOVA/StoViz/.

6.1 Outils de navigation de vidéo

Il existe de nombreuses solutions logicielles qui permettent de manipuler des documents vidéo. Nous pouvons citer de manière non exhaustive vlc [VideoLan], Windows Media Player [Microsoft], MPlayer [MPlayer], QuickTime [Apple] ou Real Player [RealNetworks]. Tous ces logiciels permettent de naviguer dans la vidéo, mais ils ne donnent aucune information sur la structure de la vidéo, ni aucune aide pour retrouver un évènement particulier qui s'y déroule.

L'outil proposé dans ce chapitre est comparable à Joke-o-mat [Friedland 2009], illustré dans la Figure 6.2, dans le sens où ce sont deux outils permettant de mettre en avant des éléments structurels d'une vidéo. Joke-o-mat est basé sur un outil de détection des « chutes » dans des vidéos comiques (appelées *punchlines*). La « chute » est une réplique comique et percutante constituant la fin d'un moment drôle, d'un dialogue ou d'un évènement.

Joke-o-mat permet de mettre en avant ces évènements particuliers dans un épisode, en autorisant de naviguer dans l'épisode « chute » par « chute » ou scène par scène, et permettant de filtrer les « chutes » ou les scènes en fonction de deux données à sélectionner : personnages impliqués dans la séquence et mots prononcés durant la séquence.

La principale différence entre Joke-o-mat et StoViz vient du fait que Joke-o-mat permet de visualiser des segments de vidéo isolés (des « chutes » ou des scènes) tandis que StoViz permet une visualisation de la structure globale de l'épisode et permet une visualisation de sous-ensembles composés de séquences non contiguës de vidéo (les histoires).

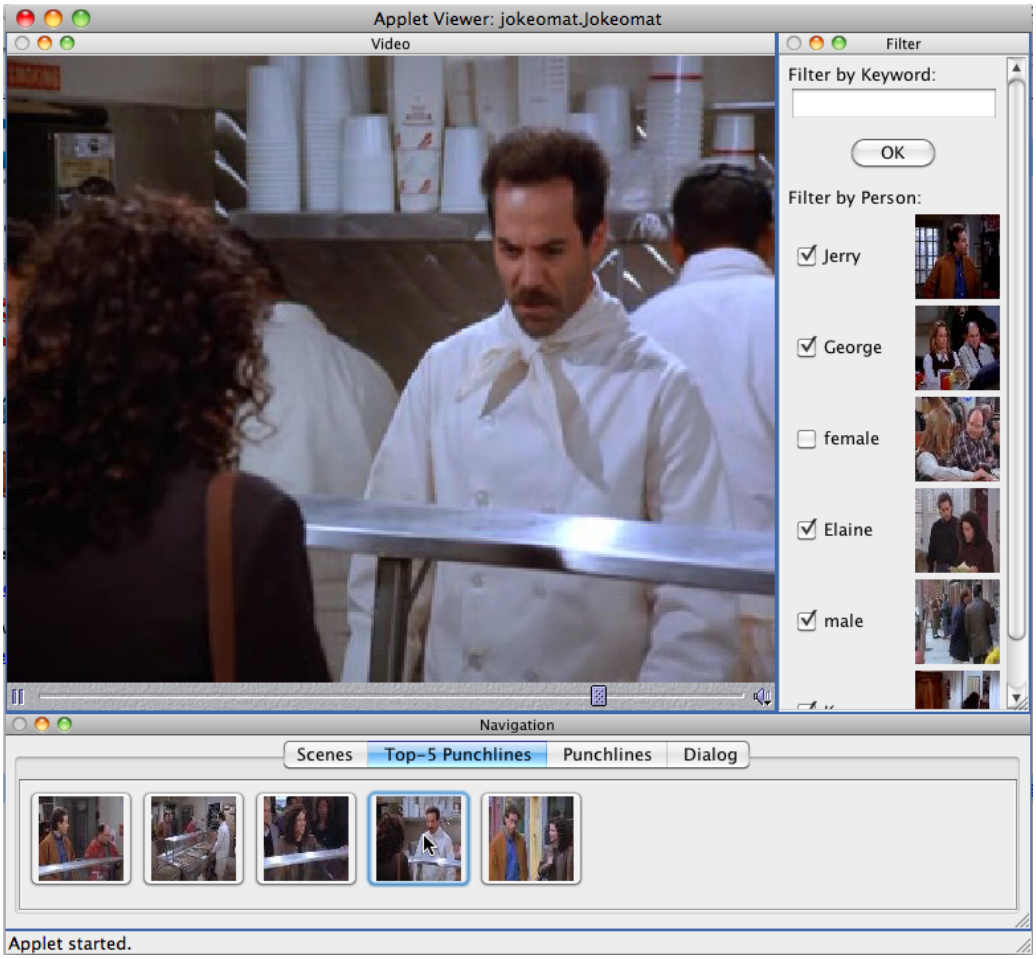

Figure 6.2 – *Interface de l'outil de navigation de vidéo* Joke-o-mat. *Cet outil permet de naviguer scène par scènes, « chute » par « chute »* (punchline par punchline) *ou dialogue par dialogue. La zone décrite sous la vidéo permet de sélectionner quel type d'évènement l'utilisateur souhaite observer (scène, « chute » ou dialogue). Une image représentative de chaque évènement est proposée, et un « clic gauche » permet de naviguer directement au début de l'évènement. Le panneau à droite de la vidéo permet de filtrer les évènements décrits dans le panneau du bas. Filtrer par personne retire du panneau du bas les évènements où apparaissent les personnes « décochées ». Filtrer par « mot-clef »* (keyword) *permet de sélectionner les évènements pour lesquels le mot-clef est prononcé.*

6.2 Aperçu de l'interface

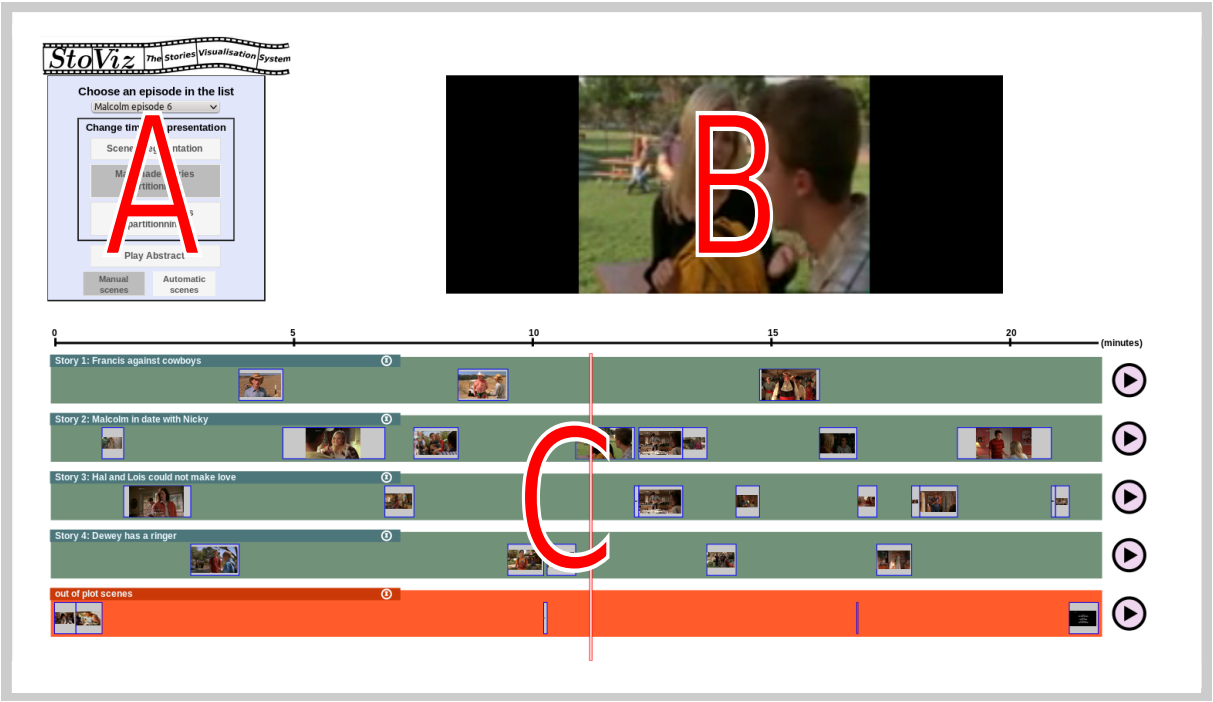

FIGURE *6.3 – Interface web de* STOVIZ.

STOVIZ est une interface de visualisation des résultats des méthodes de segmentation en scènes et de regroupement des scènes en histoires présentées aux Chapitres 4 et 5. Aucun calcul n'est effectué directement par STOVIZ. Ce logiciel a uniquement pour vocation de permettre une visualisation rapide et efficace de la structure d'un épisode de série télévisée et des résultats de nos méthodes de structuration. L'interface du navigateur de vidéo est composée de trois zones illustrées dans la Figure 6.3 :

- La zone **A** contient le *panneau de contrôle*. Il permet à l'utilisateur de sélectionner l'épisode désiré et de modifier le contenu de la zone **C** en sélectionnant le type d'histoire à afficher (détectées automatiquement par le système ou annotées manuellement) ou le type de scènes (manuelles ou automatiques).
- La zone **B** correspond au lecteur vidéo, qui permet de naviguer dans la vidéo, de lancer ou arrêter la lecture ou modifier le volume sonore.
- La structure de l'épisode est affichée dans la zone **C**. Cette zone affiche des lignes que nous appelons *Frises temporelles*. Elles correspondent à un sous-ensemble de la vidéo dont la signification varie en fonction des ordres dictés par le panneau de contrôle. Ainsi, une frise peut correspondre à une histoire, un résumé ou l'ensemble de la vidéo.

Zone d'affichage des frises temporelles

Cette zone affiche toutes les frises relatives à l'état du système. Chaque frise est composée d'un ou plusieurs rectangles, chacun correspondant à une scène de la vidéo. La Figure 6.4 illustre la façon dont une frise est affichée. Elle est composée d'un ensemble de scènes de la vidéo, chacune étant décrite par une image clef qui peut être agrandie en passant le curseur de la souris au-dessus de celle-ci, et d'un rectangle coloré dont la largeur est proportionnelle à la durée de la scène.

Chaque frise possède un court texte au niveau de son coin supérieur gauche décrivant son contenu. Ce contenu dépend des choix effectués sur le panneau de contrôle. Un bouton situé à droite de la frise permet de jouer son contenu indépendamment du reste de la vidéo.

FIGURE 6.4 – *Frises temporelles.*

6.3 Visualisation de la structure d'un épisode

Dans son état initial ou en cliquant sur le bouton *scenes segmentation*, le système affiche une unique frise composée de toutes les scènes de l'épisode, comme illustré par la Figure 6.5.

En cliquant sur un des deux boutons nommés *manual stories* ou *automatic stories*, les scènes se déplacent de manière à former plusieurs frises, chacune décrivant une histoire comme illustré par la Figure 6.6. Cet affichage permet de visualiser chaque histoire séparément, et d'observer la répartition des scènes dans les différentes histoires de l'épisode. La dernière frise, affichée en orange, contient toutes les scènes qui n'ont pas été annotées en histoires.

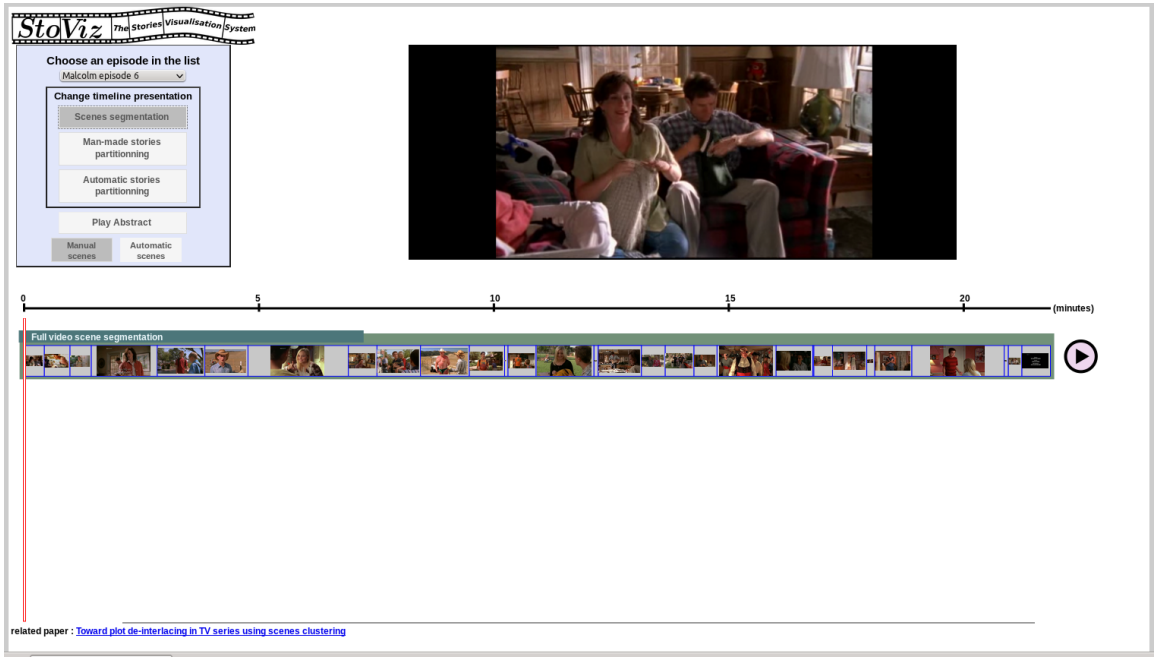

FIGURE *6.5 – Affichage de la segmentation en scènes d'un épisode. Une seule frise affiche l'intégralité des scènes de l'épisode.*

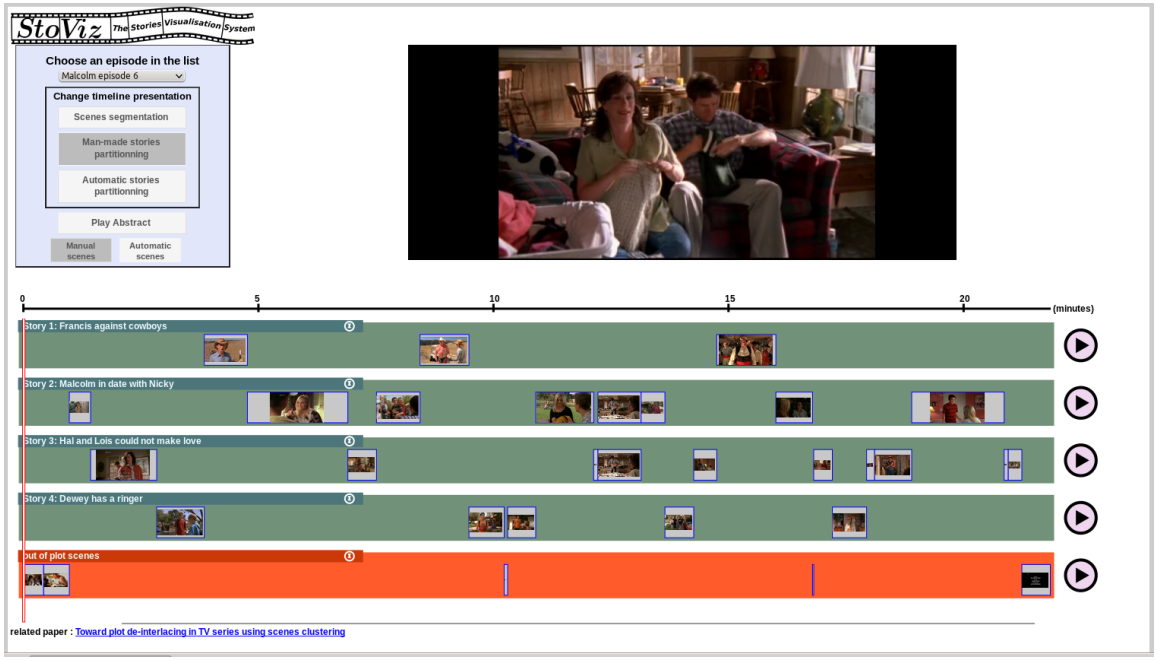

FIGURE *6.6 – Affichage des histoires d'un épisode. Chaque frise montre les scènes appartenant à une histoire de l'épisode.* STOVIZ *permet de jouer chaque histoire indépendamment du reste de l'épisode.*

Résumé de l'épisode

Bien que cela ne fasse pas partie des contributions de la thèse, la structuration automatique d'un épisode de série télévisée proposée dans les chapitres précédents a pour vocation d'améliorer des systèmes d'indexation et de création automatique de résumés de vidéos. Ainsi, le panneau de contrôle de StoViz permet à l'utilisateur de choisir d'observer un extrait de l'épisode sous forme de résumé vidéo.

Le bouton *Abstract* permet d'afficher une frise contenant une suite de scènes représentatives de la vidéo et fournissant à l'utilisateur un résumé compact de l'épisode, comme illustré Figure 6.7. La sélection des scènes à inclure dans le résumé d'un épisode de série télévisée est discutée en perspective de cette thèse.

FIGURE *6.7 – Lecture d'un résumé de l'épisode.*

6.4 Évaluation des systèmes de structuration

Une évaluation par le calcul d'un score (par exemple F-Mesure) ou d'un taux d'erreurs (DER) donne une indication globale sur la performance d'un système mais elle ne permet pas de repérer facilement les causes des erreurs produites par les systèmes automatiques de structuration. Ainsi, StoViz facilite l'évaluation visuelle des systèmes de structuration développés. Cette section est divisée en deux sous-sections : la première présente le mode d'évaluation du système de segmentation en scènes, la deuxième discute du mode d'évaluation du regroupement des scènes en histoires.

6.4.1 Évaluation de la segmentation en scènes

Le panneau de contrôle permet de sélectionner un affichage des scènes annotées manuellement ou des scènes détectées automatiquement par les systèmes présentés au Chapitre 4. Choisir la segmentation automatique des scènes permet d'afficher deux frises comme illustré dans la Figure 6.8 : la première montre la segmentation automatique des scènes et l'autre décrit une segmentation de référence (manuellement annotée).

Pour permettre une évaluation visuelle de la performance du système de segmentation, chaque frontière de scène est décrite par une barre colorée. Sur la frise représentant la segmentation automatique, les *Vrais Positifs* sont indiqués par une barre verte et les *Faux*

Positifs par une barre rouge. Les *Faux Négatifs* sont indiqués en jaune sur la segmentation de référence.

Cette représentation visuelle des frontières correctement détectées et des erreurs de segmentation permet de naviguer rapidement dans la vidéo par un simple « clic gauche » de la souris au niveau de la frontière à observer. Elle est ainsi complémentaire d'un score de précision, rappel et F-Mesure (indiqué dans un cadre en haut à gauche des frises) pour comprendre les erreurs de segmentation et ainsi corriger le comportement de la méthode de segmentation.

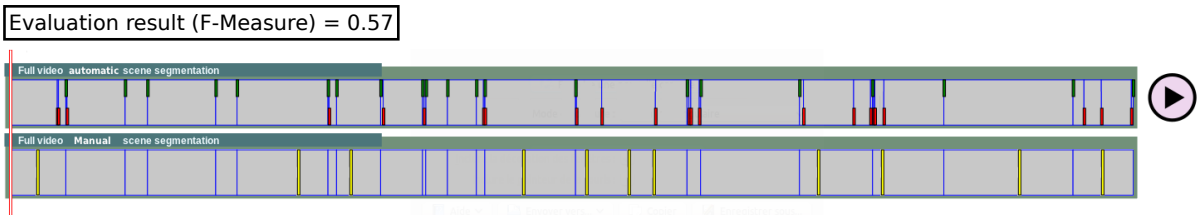

FIGURE *6.8 – Affichage des scènes détectées automatiquement.*

6.4.2 Évaluation des histoires

Deux boutons du panneau de contrôle permettent de modifier l'affichage des frises de manière à visualiser les histoires. *Manual stories* permet d'observer les histoires annotées manuellement et *Automatic stories* permet d'observer les histoires détectées automatiquement par les méthodes présentées au Chapitre 5.

Dans le cas des annotations manuelles, les histoires sont décrites par un titre et une courte description des actions qui s'y déroulent. Ceci permet de comprendre les choix des annotateurs et donc pourquoi une scène se retrouve dans une histoire particulière.

StoViz permet de comparer ces histoires annotées manuellement (référence) avec des histoires détectées automatiquement par notre système (hypothèse). Dans le cas d'un affichage de l'hypothèse, StoViz modifie la couleur du rectangle représentatif des scènes selon qu'elles sont bien ou mal positionnées dans les histoires.

Scène correctement/incorrectement positionnée

Pour savoir si une scène est associée à « la bonne histoire », StoViz commence par rechercher quelle histoire de l'hypothèse correspond au mieux à une histoire de la référence. La correspondance entre les histoires de l'hypothèse et de la référence est effectuée suivant le même principe que celle utilisée pour le calcul du DER expliqué dans la Section 5.1.3 (page 128).

Ainsi, soit $\mathcal{H}$ l'ensemble des histoires de l'hypothèse, et $m^*(H)$ une fonction retournant l'histoire $H' \in \mathcal{R}$ de la référence qui correspond au mieux à l'histoire $H \in \mathcal{H}$ de l'hypothèse. Sachant qu'une histoire est un ensemble de scènes, et qu'une scène peut être

associée à plusieurs histoires, il est possible de déterminer trois états différents sur la position des scènes, qui sont représentés dans StoViz par trois couleurs différentes :

- **gris** : la scène s est correctement positionnée si chaque histoire dans laquelle elle intervient dans l'hypothèse correspond à une histoire où elle intervient dans la référence : $\{H \in \mathcal{H}/s \in H \wedge s \in m^*(H)\} = \{H \in \mathcal{H}/s \in H\}$
- **rouge** : la scène s est incorrectement positionnée si aucune histoire dans laquelle elle intervient dans l'hypothèse ne correspond à une histoire où elle intervient dans la référence : $\{H \in \mathcal{H}/s \in H \wedge s \in m^*(H)\} = \emptyset$
- **jaune** : dans tous les autres cas, cette couleur est utilisée pour décrire une scène qui est partiellement correctement positionnée. Ce cas correspond à une scène pour laquelle seule une partie des histoires de l'hypothèse dans lesquelles elle intervient correspondent à une histoire où elle intervient dans la référence (par exemple, une scène associée à deux histoires dans la référence qui n'est associée qu'à une histoire dans l'hypothèse).

Analyse de l'affichage du regroupement automatique des scènes en histoires

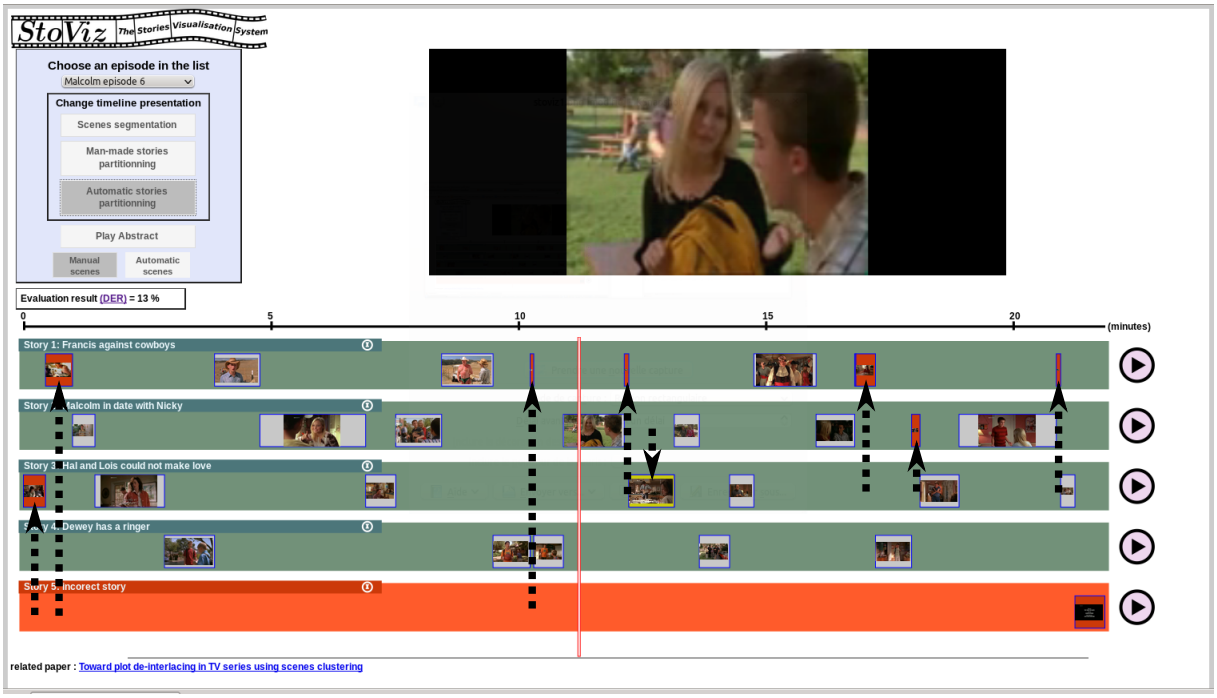

FIGURE 6.9 – *Affichage des histoires d'un épisode détectées automatiquement. Chaque frise montre les scènes appartenant à une histoire de l'épisode.* StoViz *permet de jouer chaque histoire indépendamment du reste de l'épisode.*

La Figure 6.9 montre le résultat d'un regroupement automatique des scènes en histoires pour un épisode de la série *Malcolm*. Le code couleur associé à chaque scène permet de se rendre compte immédiatement si une scène a été correctement associée à une histoire.

De plus, puisque les couleurs ne sont parfois pas suffisantes pour mettre en évidence les erreurs, le changement d'affichage entre un regroupement automatique et un regroupement manuel entraîne une animation où seules les scènes mal positionnées se déplacent. Ce principe est illustré sur la Figure 6.9 par les flèches qui indiquent le déplacement des scènes en passant de l'affichage du regroupement manuel vers le regroupement automatique. La valeur de la métrique est affichée dans un rectangle en dessous du panneau de contrôle. Dans cet exemple, le regroupement obtient un DER de 13% (erreurs).

Dans l'exemple de la Figure 6.9, cinq histoires ont été détectées pour quatre annotées. Les histoires de l'hypothèse correspondant à des histoires de la référence sont affichées dans les frises vertes. Les histoires de l'hypothèse n'ayant aucune correspondance dans la référence sont affichées dans des frises orange. Cette distinction permet ainsi une discrimination visuelle immédiate entre les histoires de l'hypothèse proches des histoires de la référence et les autres.

Cas d'usage pratique

StoViz permet de repérer d'où proviennent les erreurs du regroupement des scènes en histoires grâce aux mécanismes évoqués précédemment. Ainsi, la comparaison visuelle d'un regroupement manuel et d'un regroupement automatique des scènes en histoires permet d'évaluer la qualité d'un regroupement automatique plus efficacement que le calcul de la métrique DER pour les raisons suivantes :

- StoViz permet de comparer immédiatement le nombre d'histoires détectées automatiquement et le nombre d'histoires annotées manuellement.
- StoViz permet de repérer les scènes correctement ou incorrectement associées à une histoire par un code couleur et une animation de ces scènes.
- StoViz facilite l'étude des scènes correctement ou incorrectement regroupées par un « clic gauche » de la souris sur ces scènes.

Dans l'exemple de la Figure 6.9, huit scènes sont indiquées comme étant associées à une mauvaise histoire. StoViz permet de remarquer que cinq d'entre elles sont incorrectement associées à l'histoire 1. L'observation de ces scènes permet de s'apercevoir que ce sont des scènes ne comportant aucun tour de parole. StoViz permet donc de se rendre compte rapidement que l'une des sources d'erreurs de l'approche utilisée ici vient de la gestion des scènes dans lesquelles aucun locuteur ne prend la parole.

Bien qu'il ne permette pas de corriger directement le système de regroupement, StoViz permet de retrouver quels types d'erreurs sont produits par les approches de regroupements étudiées. Ainsi, StoViz permet de cibler les expérimentations qui doivent être effectuées afin de corriger les systèmes de structuration développés.

6.5 Technologies utilisées

Les technologies utilisées pour StoViz ont été choisies pour leur portabilité (StoViz est utilisable sur la plupart des systèmes d'exploitation) et leur pérennité (tous les outils utilisés par StoViz sont inclus dans le code source). StoViz est une plateforme web de visualisation de vidéos développée en HTML5 et Javascript. Le lecteur vidéo utilisé est le lecteur Projekktor [Projekktor]. Il s'agit d'un lecteur vidéo gratuit et open-source développé en HTML5. Le dessin des frises et les animations des scènes sont réalisées à l'aide de la librairie Javascript D3js [D3js] (Data-Driven Documents). Il s'agit d'une librairie permettant de visualiser facilement des données.

Une technologie basée sur HTML5 et Javascript permet d'utiliser StoViz sur des systèmes d'exploitation aussi variés que MacOS, Linux et Windows, ainsi que sur tablette tactile fonctionnant avec un système iOS. StoViz fonctionne sur toutes les plateformes web récentes, comme Firefox, Chrome ou Safari.

L'utilisation de Javascript fait de StoViz un outil fluide et dynamique. Une fois les données transférées sur la machine client (vidéo, images et données de structuration), tous les calculs sont effectués « côté client » en Javascript, ne nécessitant pas de transfert d'information entre le client et le serveur. De plus, l'utilisation d'un lecteur vidéo autorisant le streaming permet d'optimiser le transfert des données initial.

Le code source est disponible à l'adresse suivante : https://bitbucket.org/philerco/stoviz.

6.6 Conclusion

Nous avons développé StoViz pour deux applications distinctes :

- visualiser la structure globale d'un épisode en permettant un aperçu des scènes et des histoires ;
- permettre une évaluation visuelle complémentaire au calcul d'un score pour des méthodes de segmentation en scènes et de regroupement des scènes en histoires.

L'application de visualisation de StoViz peut être étendue. Par exemple, StoViz pourrait être utilisé sur une tablette tactile comme une télécommande. Ainsi, la télécommande permettrait de visualiser la structure de l'épisode visionné, et donc d'améliorer l'efficacité de la recherche d'information et la navigation dans cet épisode.

Pour l'instant, StoViz a principalement prouvé son efficacité pour nous accompagner dans l'analyse des erreurs des systèmes de structuration développés. Il permet d'évaluer visuellement les résultats de méthodes de segmentations en scènes et de regroupement des scènes en histoires. De plus, l'interface est déjà prête pour permettre la visualisation du résultat de systèmes de génération automatique de résumés vidéo.

Conclusion et perspectives

Le travail mené au cours de cette thèse a permis d'explorer plusieurs méthodes pour la segmentation et la structuration d'épisodes de séries télévisées. En particulier, nous avons mesuré l'apport d'une description de segments audiovisuels par les locuteurs pour la segmentation en scènes, et nous avons proposé une nouvelle approche de structuration pour des épisodes de séries télévisées : le regroupement automatique des scènes en histoires.

Dans un premier temps, nous avons proposé un **schéma hiérarchique pour la structure narrative d'un épisode de série télévisée**. Il en ressort différents niveaux de structuration qui sont, dans l'ordre croissant de granularité, le plan, la scène et l'histoire. Ces concepts peuvent avoir une interprétation différente en fonction du domaine étudié (par exemple, une « scène » a une signification différente dans un film ou une pièce de théatre). Nous avons limité nos recherches à des documents audiovisuels ayant subi une étape de « post-production », et plus particulièrement à des épisodes de séries télévisées. Une définition objective de ces différents concepts pour les documents étudiés est ainsi proposée.

Une de nos contributions concerne la **segmentation en scènes des épisodes de séries télévisées**. Nous avons proposé une étude sur l'utilisation des sorties d'un système de segmentation et regroupement en locuteurs pour décrire des segments de vidéo en vue de retrouver les frontières entre les scènes. **Nous avons montré que l'information fournie par les tours de parole des locuteurs est pertinente pour cette tâche de segmentation**. Cependant, elle reste moins performante qu'une approche de segmentation issue de l'état de l'art et basée uniquement sur une analyse de la couleur [Yeung 1998]. Nous avons donc proposé deux approches de fusion des segmentations basées sur la couleur et sur les tours de parole des locuteurs.

La première consiste à **aligner les frontières obtenues à partir de la segmentation basée sur les tours de parole avec celles obtenues à partir de la méthode [Yeung 1998]**. Nous montrons que l'approche de fusion obtient un score F-Mesure meilleur de 3% par rapport à la méthode [Yeung 1998] si l'on utilise des tours de parole manuellement annotés. Cependant, les erreurs introduites par des tours de paroles automatiquement détectés ne permettent pas d'obtenir une meilleure segmentation que l'approche [Yeung 1998].

Nous avons donc utilisé une deuxième approche de fusion inspirée des travaux de Sidiropoulos *et al.* [Sidiropoulos 2011]. Il s'agit d'une **approche consistant à fusionner un très grand nombre de segmentations** obtenues en faisant varier les paramètres de l'approche [Yeung 1998]. Nous montrons que l'utilisation conjointe des tours de parole des locuteurs automatiquement détectés et de la couleur donne un score meilleur de 1 à 3% qu'une approche basée uniquement sur la couleur pour chaque série du corpus de test.

Les épisodes de séries télévisées modernes racontent souvent plusieurs histoires (ou lignes d'action) en parallèle. C'est pourquoi nous proposons une nouvelle **approche de structuration consistant à extraire ces histoires**. Notre approche utilise le résultat d'une segmentation en scènes d'un épisode pour regrouper les scènes appartenant à une même histoire. À notre connaissance, la problématique du regroupement des scènes en histoires telle que nous l'abordons n'a pas été étudiée précédemment. Nos contributions proposent donc un travail exploratoire sur la faisabilité et l'utilité d'une telle approche de structuration.

Nous avons étudié **cinq méthodes de regroupement et trois approches de description des scènes basées sur trois modalités** :

- **la couleur** (histogrammes de couleur) ;
- **les personnages** (tours de parole des locuteurs) ;
- **le texte** (dialogues reconnus par un système de transcription automatique de la parole).

Nous avons montré que la modalité la plus pertinente pour le regroupement des scènes en histoires est celle utilisant les tours de parole des locuteurs. En effet, dans de nombreux épisodes du corpus étudié, les personnages intervenant dans une histoire n'interviennent que rarement dans les autres histoires. Dans ce cas, deux scènes partageant un grand nombre de personnages font généralement partie d'une même histoire.

Nous avons proposé une **approche de fusion des trois modalités** en définissant une mesure de similarité multimodale unique entre les scènes. Bien que le regroupement basé uniquement sur les tours de parole des locuteurs reste globalement meilleur, nous montrons que notre approche de fusion, associée à un regroupement spectral, donne les meilleurs résultats pour certains épisodes particuliers. Ces épisodes ont pour particularité que la majorité des personnages sont impliqués dans toutes les histoires de l'épisode. C'est pourquoi un regroupement basé uniquement sur les tours de paroles des locuteurs n'est pas en mesure de retrouver les histoires annotées, et l'approche multimodale est plus performante.

Pour résoudre ce problème de variabilité de la meilleure approche de regroupement entre les épisodes, une autre de nos contributions consiste à **détecter automatiquement la meilleure approche à appliquer à chaque épisode**. Nous avons montré que l'approche de regroupement basée uniquement sur les tours de parole est la meilleure dans le cas où une communauté de personnages est spécifique à chaque histoire. Dans les autres

cas, l'approche multimodale est plus performante. Nous utilisons un critère permettant de définir à quel point les communautés de personnages sont séparées. Ce critère est alors utilisé pour déterminer quelle méthode de regroupement utiliser (regroupement basé uniquement sur les tours de parole si les communautés sont séparées, approche multimodale sinon). Nous montrons que cette méthode permet de sélectionner la meilleure approche de regroupement pour 19 des 22 épisodes du corpus.

Une autre contribution concerne les applications dérivées de cette nouvelle approche de structuration. Nous avons développé StoViz qui permet de **visualiser un épisode de série télévisée tout en proposant un aperçu visuel de la structure de l'épisode**. Ainsi, l'utilisateur peut naviguer plus efficacement dans l'épisode, ou suivre une histoire particulière. De plus, StoViz propose plusieurs **outils d'évaluation des résultats de la segmentation en scènes et du regroupement des scènes en histoires** utiles au chercheur.

Une dernière contribution du travail effectué dans cette thèse concerne les annotations des scènes et des histoires pour les 22 épisodes du corpus. La plupart des travaux sur la structuration d'épisodes de séries télévisées (comme la segmentation en scènes) se basent sur des annotations personnelles impossibles à reproduire. Pour permettre une comparaison de nos travaux et une diffusion du corpus annoté, **les annotations des scènes et des histoires sont accessible en ligne à l'adresse** `http://herve.niderb.fr/data/ally_mcbeal.html`.

Perspectives à court terme

Les perspectives de recherche à court terme consistent en quelques propositions d'amélioration des approches de regroupement des scènes en histoires. Cette amélioration passe par un perfectionnement des approches de description des scènes et de la méthode de regroupement employée.

La caractérisation des personnages : Nous avons vu que le système nous permettant de détecter les personnages (locuteurs) [Barras 2006] introduit beaucoup d'erreurs en entrée de nos algorithmes de structuration. En effet, nous avons mesuré sur quatre épisodes près de 70% d'erreurs (DER) en moyenne. Ce système a été développé pour des documents de type journaux télévisés ou radio. Ainsi, ses performances sur des épisodes de séries télévisées sont détériorées à cause de différents facteurs : interaction très rapide entre les personnages, parole spontanée (bien que jouée) induisant des tours de parole très courts ou parole/musique superposée. Comme proposé par El-Khoury [El Khoury 2010], des améliorations sont cependant possibles en utilisant des systèmes de détection de visages pour détecter les personnages et en combinant la segmentation et le regroupement en locuteurs à cette détection visuelle des personnages.

La caractérisation des lieux : Dans le Chapitre 3, nous faisons l'hypothèse que la connaissance du lieu où se déroule la scène permet de retrouver les transitions entre scènes et donne une information pertinente pour le regroupement des scènes en histoires. Nous avons utilisé des histogrammes de couleur pour décrire les lieux. Il est intéressant d'étudier d'autres méthodes pour caractériser les lieux et, à la manière d'un système de segmentation et regroupement en locuteurs, être capable de déterminer à quels moments de la vidéo le lieu change (segmentation en lieux) et si deux segments de vidéo prennent place dans un même lieu (regroupement en lieux).

La caractérisation des dialogues : De la même manière que pour le système de segmentation et regroupement en locuteurs, le système utilisé de transcription automatique de la parole [Gauvain 2002] a été développé pour des documents différents des épisodes de séries télévisées. Ainsi, nous observons beaucoup d'erreurs qui se répercutent sur nos algorithmes de structuration. Améliorer le système de transcription utilisé est donc une première perspective concernant la transcription de la parole. L'utilisation que nous faisons des dialogues reconnus est inspirée des approches de classification de textes, mais aucune analyse sémantique des phrases n'est utilisée. Une autre perspective intéressante est de travailler sur une analyse syntaxique et sémantique des phrases reconnues pour décrire les scènes par le sens des phrases et non pas uniquement par les mots reconnus.

Les méthodes de regroupement utilisées : Nous avons vu qu'un des problèmes des approches de regroupements hiérarchiques agglomératifs que nous avons utilisées est qu'il est impossible de remettre en question le regroupement : deux scènes regroupées à une étape du regroupement ne peuvent plus être dissociées. Nous pensons qu'un regroupement avec optimisation globale d'un critère tel que proposé par l'approche de Louvain est plus pertinent que les approches agglomératives étudiées.

Suppression automatique des scènes n'appartenant à aucune histoire : Les scènes n'appartenant à aucune histoire (génériques de début et de fin, sketchs isolés) viennent perturber le regroupement des scènes. Une perspective à court terme de nos travaux concerne la suppression automatique de ces scènes. Nous avons exploré cette idée pour la détection du générique de début de chaque épisode. Le générique est généralement identique pour tous les épisodes d'une même série et d'une même saison. En utilisant les images ou le son, il est possible de comparer les épisodes deux à deux pour retrouver la position d'une longue séquence très similaire. La Figure 1 montre le résultat obtenu en comparant deux épisodes de la série *Le Trône de Fer*. La matrice présentée est une matrice de similarité entre des descripteurs audio (MFCC) extraits chaque seconde de la bande-son des épisodes. Le générique est retrouvé dans cette matrice pour les deux épisodes en recherchant la plus longue séquence audio similaire aux deux épisodes. Nos expériences préliminaires ont montré qu'un tel système permet de retrouver la position du générique pour 17 des 22 épisodes annotés.

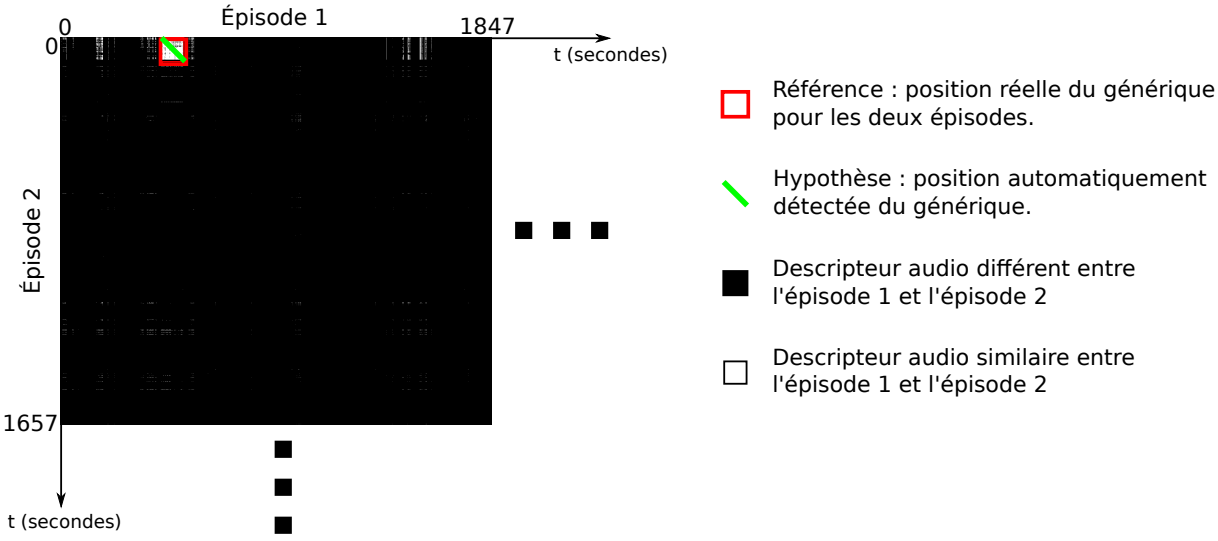

FIGURE *1 – Recherche du générique de début des épisodes de séries télévisées à partir de la bande son en comparant les épisodes deux à deux. Exemple avec deux épisodes de la série* Le Trône de Fer.

Perspectives à long terme

À long terme, les perspectives s'orientent vers une évolution du type de données étudiées et les applications possibles de la structuration proposée.

Recherche des arcs narratifs pour toute une collection : Les travaux sur la recherche des histoires ne concernent que des histoires racontées en parallèle dans un même épisode de série télévisée. Or, dans la majorité des séries modernes, il est courant d'observer des histoires se déroulant sur plusieurs épisodes. Ces histoires sont appelées *arcs narratifs*. Dans la première saison de la série *Ally McBeal*, par exemple, un arc narratif principal est développé autour des relations amoureuses d'Ally. Il se développe durant la majorité des épisodes de la série. Une extension pertinente au travail sur le regroupement des scènes en histoires est de rechercher les histoires dans une collection d'épisodes pouvant posséder des arcs narratifs (comme tous les épisodes d'une même saison d'une série télévisée).

Génération automatique de résumés vidéo : Nous pensons qu'une des applications les plus pertinentes pour la structuration des épisodes de séries télévisées en scènes et histoires concerne la génération automatique de résumés de vidéo. Dans [Ercolessi 2012], nous proposons une approche de génération de résumés vidéo directement inspirée des méthodes de génération automatique de résumé de texte [Radev 2004, Wang 2009].

Plutôt que de comprendre le texte et de générer un résumé fidèle, le principe consiste à identifier les phrases pertinentes qui seraient susceptibles de porter les thématiques principales du document. Ainsi, l'approche fonctionne en trois étapes : découpage du texte

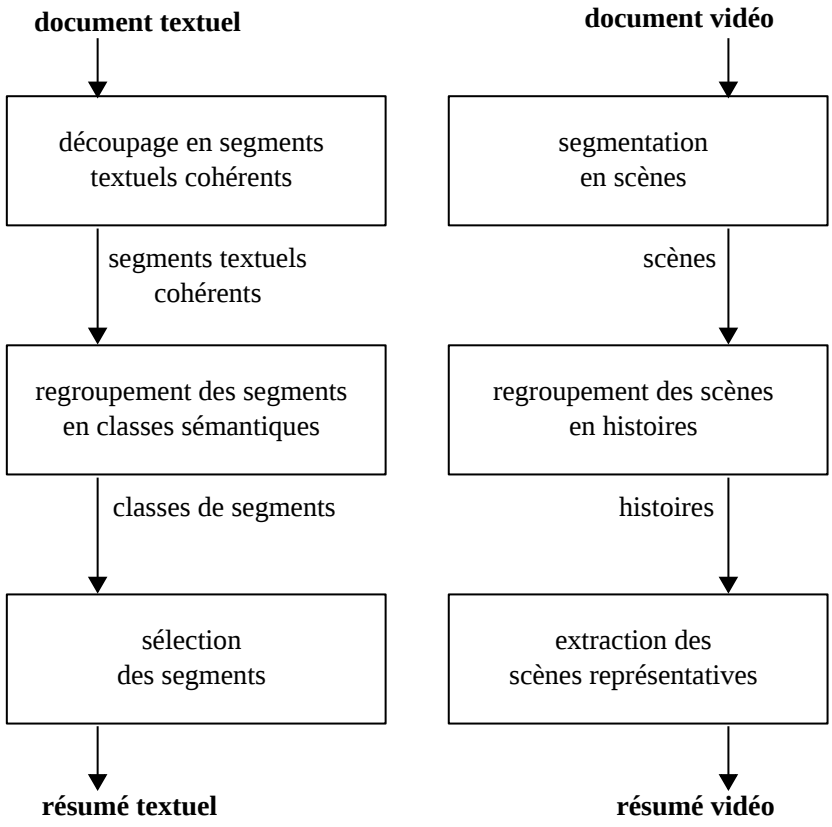

FIGURE *2 – Résumé textuel vs. résumé vidéo*

en segments cohérents (ou en phrases), regroupement des segments en classes sémantiques (thématiques), et enfin, sélection des segments pertinents pour chaque classe.

Nous procédons par analogie pour la génération d'un résumé de vidéo, résumée dans la Figure 2, en considérant qu'une scène correspond à une phrase. Les scènes sont ensuite regroupées sous forme d'histoires (correspondant au regroupement des segments en classes sémantiques). Nous cherchons ensuite à extraire une scène représentative par histoire (sélection des segments), comme étant la scène la plus centrale de chaque histoire (scène dont la distance moyenne à toutes les autres scènes de l'histoire est la plus faible).

Le résumé est ensuite composé de ces scènes centrales qui sont concaténées pour former le résumé final en respectant l'ordre chronologique d'apparition dans la vidéo initiale. Les résumés générés par cette approche permettent d'avoir une idée des personnages présents dans la vidéo, des différentes intrigues qui s'y déroulent, et du thème global de l'épisode.

Actuellement, l'approche proposée est une transposition directe des méthodes de résumé textuel. Cependant, il serait intéressant, plutôt que de s'appuyer sur un nombre restreint de scènes entières, d'utiliser des extraits saillants de ces scènes afin d'offrir un rythme adéquat au résumé. En effet, le but ultime est d'obtenir un résumé s'approchant de ceux que l'on peut trouver au début de certains épisodes et rappelant les intrigues passées.

Annexe A

Publications

Numéro spécial de revue

Philippe Ercolessi, Christine Sénac, Hervé Bredin, Sandrine Mouysset. *Vers un résumé automatique de séries télévisées basé sur une recherche multimodale d'histoires.* Numéro spécial de Document Numérique sur le Résumé Automatique de Documents, pages 41-66, 2012.

Conférences et workshops internationnaux

Philippe Ercolessi, Christine Senac, Philippe Joly, Hervé Bredin. *Segmenting TV Series into Scenes using Speaker Diarization.* International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), pages 13-15. Delft - Pays-bas, Avril 2011.

Philippe Ercolessi, Christine Senac, Hervé Bredin. *Toward Plot De-interlacing in TV Series using Scenes Clustering.* IEEE International Workshop on Content-Based Multimedia Indexing (CBMI 2012), pages 1-6. Annecy - France, 2012.

Philippe Ercolessi, Christine Senac, Hervé Bredin, Sandrine Mouysset. *Hierarchical framework for plot de-interlacing of TV series based on speakers, dialogues and images.* ACM Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA 2012), pages 3-8. Nara - Japon, 2012.

Démos avec actes publiés

Philippe Ercolessi, Christine Senac, Hervé Bredin. *StoViz : Story Visualization of TV Series* ACM Multimedia (ACMMM 2012), pages 1329-1330. Nara - Japon, 2012
http://irit.fr/recherches/SAMOVA/StoViz/

Conférences sans actes publiés

Philippe Ercolessi, Christine Senac, Hervé Bredin, Philippe Joly. *Video Collection Summarization by Semantic Graph Comparison.* Dans : Workshop on Visual Information Processing (EUVIP). Paris - France, 2012

Philippe Ercolessi, Christine Senac, Hervé Bredin, Philippe Joly. *Summarizing Video Collection using Semantic Graph.* Dans : Workshop IRIT/Kyushu Image et Multimedia. Toulouse - France, 2011

Bibliographie

[Abduraman 2011] Alina Elma Abduraman, Sid-Ahmed Berrani & Bernard Mérialdo. TV program structuring techniques : A review. Book chapter in TV Content Analysis : Techniques and Applications, 2011.

[Adams 2002] Brett Adams, Chitra Dorai & Svetha Venkatesh. *Toward Automatic Extraction of Expressive Elements from Motion Pictures : Tempo.* IEEE Transactions on Multimedia, vol. 4, no. 4, pages 472–481, 2002.

[Aigrain 1997] Philippe Aigrain, Philippe Joly & Véronique Longueville. *Medium Knowledge-Based Macro-Segmentation of Video into Sequences.* Intelligent multimedia information retrieval, vol. 25, pages 74–84, 1997.

[Apple] Apple. *QuickTime.* http://www.apple.com/fr/quicktime/.

[Arandjelovic 2005] Ognjen Arandjelovic & Andrew Zisserman. *Automatic Face Recognition for Film Character Retrieval in Feature-Length Films.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pages 860–867, 2005.

[Barras 2006] Claude Barras, Xuan Zhu, Sylvain Meignier & Jean-Luc Gauvain. *Multi-Stage Speaker Diarization of Broadcast News.* IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pages 1505–1512, 2006.

[Bay 2006] Herbert Bay, Tinne Tuytelaars & Luc Van Gool. *Surf : Speeded Up Robust Features.* European Conference on Computer Vision (ECCV), pages 404–417, 2006.

[Belongie 2002] Serge Belongie, Jitendra Malik & Jan Puzicha. *Shape Matching and Object Recognition Using Shape Contexts.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pages 509–522, 2002.

[Benini 2005] Sergio Benini & Riccardo Leonardi. *Identifying Video Content Consistency by Vector Quantization.* Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2005.

[Benini 2006] Sergio Benini, Aldo Bianchetti, Riccardo Leonardi & Pierangelo Migliorati. *Hierarchical Summarization of Videos by Tree-Structured Vector Quantization.* IEEE International Conference on Multimedia and Expo (ICME), pages 969–972, 2006.

[Benini 2007] Sergio Benini, Pierangelo Migliorati & Riccardo Leonardi. *A Statistical Framework for Video Skimming Based on Logical Story Units and Motion Activity.* IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pages 152 –156, 2007.

[Benini 2008] Sergio Benini, Pierangelo. Migliorati & Riccardo. Leonardi. *Retrieval of Video Story Units by Markov Entropy Rate.* IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pages 41–45, 2008.

[Bicego 2006] Manuele Bicego, Andrea Lagorio, Enrico Grosso & Massimo Tistarelli. *On the Use of SIFT Features for Face Authentication.* Computer Vision and Pattern Recognition Workshop (CVPRW), pages 35–35, 2006.

[Blondel 2008] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre. *Fast Unfolding of Communities in Large Networks.* Journal of Statistical Mechanics : Theory and Experiment, no. 10, 2008.

[Bonnabel 2000] Anne-Marie Bonnabel & Marie-Lucile Milhaud. À la Découverte du Théâtre. Ellipses Édition, 2000.

[Bozonnet 2010] Simon Bozonnet, Nicholas WD Evans & Corinne Fredouille. *The liaeurecom RT'09 Speaker Diarization System : Enhancements in Speaker Modelling and Cluster Purification.* IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pages 4958–4961, 2010.

[Bradski 2000] Gary Bradski. *The OpenCV Library.* Dr. Dobb's Journal of Software Tools, 2000.

[Bredin 2012] Hervé Bredin. *Segmentation of TV Shows into Scenes Using Speaker Diarization and Speech Recognition.* International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2377–2380, 2012.

[Chaisorn 2003] Lekha Chaisorn, Tat-Seng Chua & Chin-Hui Lee. *A Multi-modal Approach to Story Segmentation for News Video.* World Wide Web, vol. 6, no. 2, pages 187–208, 2003.

[Chang 2007] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui & Jiebo Luo. *Large-scale Multimodal Semantic Concept Detection for Consumer Video.* ACM International Workshop on Multimedia Information Retrieval, pages 255–264, 2007.

[Chartrand 1999] Suzanne-G. Chartrand, Denis Aubin, Raymond Blain & Claude Simard. Grammaire pédagogique du français d'aujourd'hui. Graficor, 1999.

[Cheng 2008] Wengang Cheng & Jun Lu. *Video Scene Oversegmentation Reduction by Tempo Analysis.* International Conference on Natural Computation (ICNC), vol. 4, pages 296–300, 2008.

[Christel 1999] Michael G Christel, Andreas M Olligschlaeger & Chang Huang. *Interactive Maps for a Digital Video Library.* IEEE International Conference on Multimedia Computing and Systems, vol. 1, pages 381–387, 1999.

[Chua 2004] Tat-Seng Chua, Shih-Fu Chang, Lekha Chaisorn & Winston Hsu. *Story Boundary Detection in Large Broadcast News Video Archives : Techniques, Experience and Trends.* ACM International Conference on Multimedia, pages 656–659, 2004.

[Colonna 2010] Vincent Colonna. L'art des séries télé. Payot, 2010.

[Cour 2008] Timothee Cour, Chris Jordan, Eleni Miltsakaki & Ben Taskar. *Movie/Script : Alignment and Parsing of Video and Text Transcription.* European Conference on Computer Vision (ECCV), pages 158–171, 2008.

[Csurka 2004] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski & Cédric Bray. *Visual Categorization with Bags of Keypoints.* Workshop on Statistical Learning in Computer Vision, pages 1–22, 2004.

[D3js] D3js. http://d3js.org/.

[Darrell 1998] Trevor Darrell, G Gordon, Michael Harville & John Woodfill. *Integrated Person Tracking using Stereo, Color, and Pattern Detection.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pages 601–608, 1998.

[Del Fabro 2013] Manfred Del Fabro & Laszlo Böszörmenyi. *State-of-the-art and Future Challenges in Video Scene Detection : a Survey.* Multimedia Systems, pages 1–28, 2013.

[Deselaers 2008] Thomas Deselaers, Daniel Keysers & Hermann Ney. *Features for Image Retrieval : an Experimental Comparison.* Information Retrieval, vol. 11, no. 2, pages 77–107, 2008.

[Dorkó 2006] Gyuri Dorkó. *Selection of Discriminative Regions and Local Descriptors for Generic Object Class Recognition.* Thèse de doctorat, Institut National Polytechnique de Grenoble, 2006.

[Duygulu 2004] Pinar Duygulu, Jia-Yu Pan & David A. Forsyth. *Towards Auto-documentary : Tracking the Evolution of News Stories.* ACM International Conference on Multimedia, pages 820–827, 2004.

[El Khoury 2010] Elie El Khoury. *Unsupervised Video Indexing based on Audiovisual Characterization of Persons.* Thèse de doctorat, Université de Toulouse, 2010.

[Ercolessi 2011] Philippe Ercolessi, Christine Senac, Hervé Bredin & Philippe Joly. *Segmenting TV Series into Scenes Using Speaker Diarization.* Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pages 13–15, 2011.

[Ercolessi 2012] Philippe Ercolessi, Christine Senac, Hervé Bredin & Sandrine Mouysset, editeurs. Vers un Résumé Automatique de Séries Télévisées Basé sur une Recherche Multimodale d'Histoires. Document numérique. Hermès, http ://www.editions-hermes.fr/, 2012.

[Everingham 2006] Mark Everingham, Josef Sivic & Andrew Zisserman. *Hello ! My name is... Buffy - Automatic Naming of Characters in TV Video.* British Machine Vision Conference (BMVC), pages 889–908, 2006.

[Fiscus 2004] Jonathan G. Fiscus, John S. Garofolo, Audrey N. Le, Alvin F. Martin, David S. Palett, Mark A. Przybocki & Gregory A.. Sanders. *Results of the Fall 2004 STT and MDE Evaluation.* Fall 2004 Rich Transcription Workshop (RT-04), 2004.

[Freeman 1991] William T. Freeman & Edward H. Adelson. *The Design and Use of Steerable Filters.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 13, no. 9, pages 891–906, 1991.

[Freytag 1863] Gustav Freytag. The technique of the drama. S. Hirzel, 1863.

[Friedland 2009] Gerald Friedland, Luke Gottlieb & Adam Janin. *Joke-o-mat : Browsing Sitcoms Punchline by Punchline.* ACM International Conference on Multimedia, pages 1115–1116, 2009.

[Ganchev 2005] Todor Ganchev, Nikos Fakotakis & George Kokkinakis. *Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task.* Proceedings of the SPECOM, pages 191–194, 2005.

[Gauvain 2002] Jean-Luc Gauvain, Lori Lamel & Gilles Adda. *The LIMSI Broadcast News Transcription System.* Speech Communication, vol. 37, no. 1-2, pages 89–109, 2002.

[Goyal 2009] Anuj Goyal, P. Punitha, Frank Hopfgartner & Joemon M. Jose. *Split and Merge Based Story Segmentation in News Videos.* European Conference on IR Research on Advances in Information Retrieval, pages 766–770, 2009.

[Gros 2012] Patrick Gros. *Recent Advances and Challenges in TV Structuring.* Signal Processing Conference (EUSIPCO), pages 2382 –2386, 2012.

[Hanjalic 1999] Alan Hanjalic, Reginald L. Lagendijk & Jan Biemond. *Automatically Segmenting Movies into Logical Story Units.* International Conference on Visual Information and Information Systems, pages 229–236, 1999.

[Haralick 1973] Robert M Haralick, Karthikeyan Shanmugam & Its' Hak Dinstein. *Textural Features for Image Classification.* IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-3, no. 6, pages 610–621, 1973.

[Huang 2008] Jing Huang, Etienne Marcheret, Karthik Visweswariah & Gerasimos Potamianos. *The IBM RT07 Evaluation Systems for Speaker Diarization on Lecture Meetings.* Multimodal Technologies for Perception of Humans, pages 497–508, 2008.

[Hubert 1985] Lawrence Hubert & Phipps Arabie. *Comparing Partitions.* Journal of classification, vol. 2, no. 1, pages 193–218, 1985.

[Ide 2005] Ichiro Ide, Tomoyoshi Kinoshita, Hiroshi Mo, Norio Katayama & Shin'ichi Satoh. *TrackThem : Exploring a Large-scale News Video Archive by Tracking Human Relations.* Asia Information Retrieval Technology, pages 510–515, 2005.

[Jeannin 2001] Sylvie Jeannin & Ajay Divakaran. *MPEG-7 visual motion descriptors.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pages 720–724, 2001.

[Jiang 2010] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subhabrata Bhattacharya, Dan Ellis, Mubarak Shah & Shih-Fu Chang. *Columbia-UCF TRECVID2010 Multimedia Event Detection : Combining Multiple Modalities, Contextual Concepts, and Temporal Matching.* NIST TRECVID Workshop, 2010.

[Karypis 1998] George Karypis & Vipin Kumar. *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs.* SIAM Journal on scientific Computing, vol. 20, no. 1, pages 359–392, 1998.

[Katz 1994] Ephraim Katz. The film encyclopedia. Harper perennial. HarperCollins Publishers, 1994.

[Kender 1998] John R Kender & Boon-Lock Yeo. *Video Scene Segmentation via Continuous Video Coherence.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pages 367 –373, 1998.

[Keysers 2007] Daniel Keysers, Thomas Deselaers, Christian Gollan & Hermann Ney. *Deformation Models for Image Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 29, no. 8, pages 1422–1435, 2007.

[Khan 2004] Shehroz S Khan & Amir Ahmad. *Cluster Center Initialization Algorithm for K-means Clustering.* Pattern recognition letters, vol. 25, no. 11, pages 1293–1302, 2004.

[Koenderink 1987] Jan J Koenderink & Andrea J van Doorn. *Representation of Local Geometry in the Visual System.* Biological Cybernetics, vol. 55, no. 6, pages 367–375, 1987.

[Kwon 2000] Yong-Moo Kwon, Chang-Jun Song & Ig-Jae Kim. *A New Approach for High Level Video Structuring.* IEEE International Conference on Multimedia and Expo (ICME), vol. 2, pages 773–776, 2000.

[Lar 2010] Petit larousse illustre 2011. Larousse, 2010.

[Lazebnik 2003] Svetlana Lazebnik, Cordelia Schmid & Jean Ponce. *Sparse Texture Representation Using Affine-Invariant Neighborhoods.* International Conference on Computer Vision & Pattern Recognition (CVPR), vol. 2, pages 319–324, 2003.

[le Chevalier D'Arcy 1765] M. le Chevalier D'Arcy. Sur la Durée de la Sensation de La Vue. Mémoires de l'Académie des Sciences de Paris, 1765.

[Li 2010] Zhengming Li, Lijie Xue & Fei Tan. *Face Detection in Complex Background Based on Skin Color Features and Improved AdaBoost Algorithms*. IEEE International Conference on Progress in Informatics and Computing (PIC), vol. 2, pages 723–727, 2010.

[Liu 1998] Zhu Liu, Yao Wang & Tsuhan Chen. *Audio Feature Extraction and Analysis for Scene Segmentation and Classification*. Journal of VLSI signal processing systems for signal, image and video technology, vol. 20, pages 61–79, 1998.

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004.

[Lu 2006] Lie Lu, Rui Cai & A. Hanjalic. *Audio Elements Based Auditory Scene Segmentation*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, pages V–V, 2006.

[Ma 2009] Chengyuan Ma, Byungki Byun, Ilseo Kim & Chin-Hui Lee. *A Detection-Based Approach to Broadcast News Video Story Segmentation*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1957–1960, 2009.

[Mahalanobis 1936] Prasanta Chandra Mahalanobis. *On the Generalized Distance in Statistics*. Proceedings of the National Institute of Sciences of India, vol. 2, no. 1, pages 49–55, 1936.

[Manning 2008] Christopher D Manning, Prabhakar Raghavan & Hinrich Schütze. Introduction to information retrieval, volume 1. Cambridge University Press Cambridge, 2008.

[Martin 1977] Marcel Martin. Le Langage Cinématographique. Les Éditeurs français réunis, 1977.

[Mermelstein 1976] Paul Mermelstein. *Distance Measures for Speech Recognition–Psychological and Instrumental*. Joint Workshop on Pattern Recognition and Artificial Intelligence, 1976.

[Microsoft] Microsoft. *Windows Media Player*. http://windows.microsoft.com/fr-fr/windows/windows-media-player.

[Mikolajczyk 2005] Krystian Mikolajczyk & Cordelia Schmid. *A Performance Evaluation of Local Descriptors*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 27, no. 10, pages 1615–1630, 2005.

[Misiti 2007] Michel Misiti, Yves Misiti, Georges Oppenheim & Jean-Michel Poggi. Wavelets and their Applications. Wiley Online Library, 2007.

[Misra 2010] Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha & JoemonM. Jose. *TV News Story Segmentation Based on Semantic Coherence and Content Similarity.* Advances in Multimedia Modeling, vol. 5916, pages 347–357, 2010.

[Mitrović 2010] Dalibor Mitrović, Stefan Hartlieb, Matthias Zeppelzauer & Maia Zaharieva. *Scene Segmentation in Artistic Archive Documentaries.* International Conference on HCI in work and learning, life and leisure : workgroup human-computer interaction and usability engineering, pages 400–410, 2010.

[Moravec 1981] Hans Moravec. *Rover Visual Obstacle Avoidance.* International Joint Conference on Artificial Intelligence (ICAI), pages 785–790, 1981.

[Mouysset 2011] Sandrine Mouysset, Joseph Noailles, Daniel Ruiz & Ronan Guivarch. *On A Strategy for Spectral Clustering with Parallel Computation.* High Performance Computing for Computational Science (VECPAR), pages 408–420, 2011.

[MPlayer] MPlayer. http://www.mplayerhq.hu.

[Müller 2007] Meinard Müller. Information Retrieval for Music and Motion. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[Newman 2006] Mark EJ Newman. *Modularity and Community Structure in Networks.* Proceedings of the National Academy of Sciences of the United States of America, vol. 103, no. 23, pages 8577–8582, 2006.

[Ngo 2003] Chong-Wah Ngo, Yu-Fei Ma & Hong-Jiang Zhang. *Automatic Video Summarization by Graph Modeling.* IEEE International Conference on Computer Vision - Volume 2, pages 104–109, 2003.

[Ngo 2009] Chong-Wah Ngo, Yu-Gang Jiang, Xiao-Yong Wei, Wanlei Zhao, Yang Liu, Jun Wang, Shiai Zhu & Shih-Fu Chang. *VIREO/DVMM at TRECVID 2009 : High-Level Feature Extraction, Automatic Video Search, and Content-Based Copy Detection.* NIST TRECVID Workshop, pages 415–432, 2009.

[Niblack 1993] Carlton W. Niblack, Ron Barber, Will Equitz, Myron D. Flickner, Eduardo H. Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos & Gabriel Taubin. *QBIC project : Querying Images by Content, Using Color, Texture, and Shape.* IS&T/SPIE's Symposium on Electronic Imaging : Science and Technology, pages 173–187, 1993.

[Novak 1992] C.L. Novak & S.A. Shafer. *Anatomy of a Color Histogram.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pages 599–605, 1992.

[Odobez 2003] Jean-Marc Odobez, Daniel Gatica-Perez & Mael Guillemot. *Spectral Structuring of Home Videos.* International Conference on Image and Video Retrieval (CIVR), pages 310–320, 2003.

[Over 2011] Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F Smeaton, Wessel Kraaij, Georges Quénot*et al.* *An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics*. TRECVID 2011-TREC Video Retrieval Evaluation Online, 2011.

[Penet 2012] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier & Patrick Gros. *Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012.

[Peng 2008] Jiang Peng & Qin Xiao Lin. *Automatic Classification Video for Person Indexing*. Congress on Image and Signal Processing (CISP), pages 475–479, 2008.

[Petersohn 2009] Christian Petersohn. *Temporal Video Structuring for Preservation and Annotation of Video Content*. IEEE International Conference on Image Processing (ICIP), pages 93–96, 2009.

[Projekktor] Projekktor. http://www.projekktor.com/.

[Radev 2004] Dragomir R. Radev, Hongyan Jing, Ma Styś & Daniel Tam. *Centroid-based Summarization of Multiple Documents*. Information Processing Management, vol. 40, no. 6, pages 919–938, 2004.

[Rand 1971] William M Rand. *Objective Criteria for the Evaluation of Clustering Methods*. Journal of the American Statistical association, vol. 66, no. 336, pages 846–850, 1971.

[Rasheed 2005] Zeeshan Rasheed & Mubarak Shah. *Detection and Representation of Scenes in Videos*. IEEE Transactions on Multimedia, vol. 7, no. 6, pages 1097 – 1105, 2005.

[RealNetworks] RealNetworks. *RealPlayer*. http://fr.real.com/.

[Rui 1999] Yong Rui, Thomas S Huang & Sharad Mehrotra. *Constructing Table-of-Content for Videos*. Multimedia systems, vol. 7, no. 5, pages 359–368, 1999.

[Salton 1988] Gerard Salton & Christopher Buckley. *Term-weighting Approaches in Automatic Text Retrieval*. Information processing & management, vol. 24, no. 5, pages 513–523, 1988.

[Schmid 1994] Helmut Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. International Conference on New Methods in Language Processing, pages 44–49, 1994.

[Sharff 1982] Stefan Sharff. The Elements of Cinema : Toward a Theory of Synthetics Impact. Columbia University Press, 1982.

[Shi 1994] Jianbo Shi & Carlo Tomasi. *Good features to track*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pages 593–600, 1994.

[Shi 2000] Jianbo Shi & Jitendra Malik. *Normalized Cuts and Image Segmentation.* IEEE Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pages 888–905, 2000.

[Sidiropoulos 2009] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo & Isabel Trancoso. *Multi-modal scene segmentation using scene transition graphs.* ACM International Conference on Multimedia, pages 665–668, 2009.

[Sidiropoulos 2011] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho & Isabel Trancoso. *Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 8, pages 1163–1177, 2011.

[Sidiropoulos 2012] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris & J. Kittler. *Differential Edit Distance : A Metric for Scene Segmentation Evaluation.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 6, pages 904 –914, 2012.

[Smeaton 2003] Alan F. Smeaton & Wessel Kraaij. *Trecvid 2003 - an overview.* TRECVID 2003 - Text REtrieval Conference TRECVID Workshop, 2003.

[Smith 1995] S. M. Smith & J. M. Brady. *SUSAN - A New Approach to Low Level Image Processing.* International Journal of Computer Vision, vol. 23, pages 45–78, 1995.

[Smith 1996] John R. Smith & Shih-Fu Chang. *Automated Binary Texture Feature Sets for Image Retrieval.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pages 2239–2242 vol. 4, 1996.

[Stevens 1937] Stanley Smith Stevens, John Volkmann & Edwin B Newman. *A Scale for the Measurement of the Psychological Magnitude of Pitch.* The Journal of the Acoustical Society of America, vol. 8, pages 185–190, 1937.

[Stoerring 2004] Moritz Stoerring. *Computer Vision and Human Skin Colour.* Thèse de doctorat, Aalborg university, Danemark, 2004.

[Stricker 1995] Markus A. Stricker & Markus Orengo. *Similarity of Color Images.* IS&T/SPIE's Symposium on Electronic Imaging : Science & Technology, vol. 2420, pages 381–392, 1995.

[Störring 1999] Moritz Störring, Hans J. Andersen, , Erik Granum & Erik Granum. *Skin Colour Detection Under Changing Lighting Conditions.* 7th Symposium on Intelligent Robotics Systems, pages 187–195, 1999.

[Sundaram 2002] H. Sundaram & Shih-Fu Chang. *Computable Scenes and Structures in Films.* IEEE Transactions on Multimedia, vol. 4, no. 4, pages 482 – 491, 2002.

[Swain 1991] Michael J. Swain & Dana H. Ballard. *Color Indexing*. International Journal of Computer Vision, vol. 7, no. 1, pages 11–32, 1991.

[Tamura 1978] Hideyuki Tamura, Shunji Mori & Takashi Yamawaki. *Textural Features Corresponding to Visual Perception*. IEEE Transactions on Systems, Man and Cybernetics, vol. 8, no. 6, pages 460–473, 1978.

[Tavanapong 2004] Wallapak Tavanapong & Junyu Zhou. *Shot Clustering Techniques for Story Browsing*. IEEE Transactions on Multimedia, vol. 6, no. 4, pages 517 – 527, 2004.

[Truong 2002] Ba Tu Truong, Svetha Venkatesh & Chitra Dorai. *Scene Extraction in Motion Pictures*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 1, pages 5–15, 2002.

[Ulges 2012] Adrian Ulges, Christian Schulze, Damian Borth & Armin Stahl. *Pornography Detection in Video Benefits (a lot) from a Multi-modal Approach*. ACM International workshop on Audio and multimedia methods for large-scale video analysis (AMVA), pages 21–26, 2012.

[Vadrevu 2011] Srinivas Vadrevu, Choon Hui Teo, Suju Rajan, Kunal Punera, Byron Dom, Alexander J. Smola, Yi Chang & Zhaohui Zheng. *Scalable Clustering of News Search Results*. ACM International Conference on Web search and data mining, pages 675–684, 2011.

[Vallet 2011] Félicien Vallet. *Structuration Automatique de Talk Shows Télévisés*. Thèse de doctorat, Télécom ParisTech, 2011.

[Vendrig 2002] Jeroen Vendrig & Marcel Worring. *Systematic Evaluation of Logical Story Unit Segmentation*. IEEE Transactions on Multimedia, vol. 4, no. 4, pages 492 – 499, 2002.

[VideoLan] VideoLan. *VLC Media player*. http://www.videolan.org/vlc/.

[Viola 2004] Paul Viola & Michael Jones. *Robust real-time face detection*. International Journal of Computer Vision, vol. 57, pages 137–154, 2004.

[Wactlar 1996] Howard D. Wactlar, Takeo Kanade, Michael A. Smith & Scott M. Stevens. *Intelligent Access to Digital Video : Informedia Project*. IEEE Computer, vol. 29, no. 5, pages 46–52, 1996.

[Wang 2009] Meng Wang & Hong-Jiang Zhang. *Video Content Structuring*. Scholarpedia, vol. 4, no. 8, page 9431, 2009.

[Wang 2010] Xiang-Yang Wang, Jun-Feng Wu & Hong-Ying Yang. *Robust Image Retrieval Based on Color Histogram of Local Feature Regions*. Multimedia Tools & Applications, vol. 49, no. 2, pages 323–345, 2010.

[Weng 2008] Ming-Fang Weng & Yung-Yu Chuang. *Multi-cue fusion for semantic video indexing*. ACM International Conference on Multimedia, pages 71–80, 2008.

[Weng 2009] Chung-Yi Weng, Wei-Ta Chu & Ja-Ling Wu. *RoleNet : Movie Analysis from the Perspective of Social Networks*. IEEE Transactions on Multimedia, vol. 11, no. 2, pages 256–271, 2009.

[Wengang 2003] Cheng Wengang & Xu De. *A Novel Approach of Generating Video Scene Structure.* TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region, vol. 1, pages 350–353 Vol.1, 2003.

[Xie 2002] Lexing Xie, Peng Xu, Shih fu Chang A, Ajay Divakaran & Huifang Sun B. *Structure Analysis of Soccer Video with Hidden Markov Models.* Pattern Recognition Letters, pages 767–775, 2002.

[Yang 2007] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann & Chong-Wah Ngo. *Evaluating bag-of-visual-words representations in scene classification.* International workshop on multimedia information retrieval, pages 197–206, 2007.

[Yeung 1998] Minerva Yeung, Boon-Lock Yeo & Bede Liu. *Segmentation of Video by Clustering and Graph Analysis.* Computer Vision and Image Understanding, vol. 71, no. 1, pages 94–109, 1998.

[Zha 2012] Zheng-Jun Zha, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong & T.-S. Chua. *Interactive Video Indexing With Statistical Active Learning.* IEEE Transactions on Multimedia, vol. 14, no. 1, pages 17–27, 2012.

[Zhai 2006] Yun Zhai & Mubarak Shah. *Video Scene Segmentation Using Markov Chain Monte Carlo.* IEEE Transactions on Multimedia, vol. 8, no. 4, pages 686–697, 2006.

[Zhao 2001] Li Zhao, Shi-Qiang Yang & Bo Feng. *Video Scene Detection Using Slide Windows Method Based on Temporal Constrain Shot Similarity.* IEEE International Conference on Multimedia and Expo (ICME), pages 1171–1174, 2001.

[Zhao 2007] Yanjun Zhao, Tao Wang, Peng Wang, Wei Hu, Yangzhou Du, Yimin Zhang & Guangyou Xu. *Scene Segmentation and Categorization Using NCuts.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1 –7, 2007.

Résumé

Nos contributions portent sur l'extraction de la structure narrative d'épisodes de séries télévisées à deux niveaux hiérarchiques. Le premier niveau de structuration consiste à retrouver les transitions entre les scènes à partir d'une analyse de la couleur des images et des locuteurs présents dans les scènes. Nous montrons que l'analyse des locuteurs permet d'améliorer le résultat d'une segmentation en scènes basée sur la couleur.

Il est courant de voir plusieurs histoires (ou lignes d'actions) racontées en parallèle dans un même épisode de série télévisée. Ainsi, le deuxième niveau de structuration consiste à regrouper les scènes en histoires. Nous cherchons à désentrelacer les histoires pour pouvoir, par exemple, visualiser les différentes lignes d'actions indépendamment.

La principale difficulté consiste à déterminer les descripteurs les plus pertinents permettant de regrouper les scènes appartenant à une même histoire. A ce niveau, nous étudions également l'utilisation de descripteurs provenant des trois modalités différentes précédemment exposées. Nous proposons en outre des méthodes permettant de fusionner les informations provenant de ces trois modalités.

Pour répondre à la variabilité de la structure narrative des épisodes de séries télévisées, nous proposons une méthode qui s'adapte à chaque épisode. Elle permet de choisir automatiquement la méthode de regroupement la plus pertinente parmi les différentes méthodes proposées.

Enfin, nous avons développé StoViz, un outil de visualisation de la structure d'un épisode de série télévisée (scènes et histoires). Il permet de faciliter la navigation au sein d'un épisode, en montrant les différentes histoires racontées en parallèle dans l'épisode. Il permet également la lecture des épisodes histoire par histoire, et la visualisation d'un court résumé de l'épisode en donnant un aperçu de chaque histoire qui y est racontée.

Mots clefs : structure narrative, segmentation en scènes, regroupement des scènes en histoires, désentrelacement des histoires, multimodalité, séries télévisées, sélection de méthode de regroupement, couleur, locuteurs, reconnaissance automatique de la parole.

Abstract

Our contributions concern the extraction of the structure of TV series episodes at two hierarchical levels. The first level of structuring is to find the scene transitions based on the analysis of the color information and the speakers involved in the scenes. We show that the analysis of the speakers improves the result of a color-based segmentation into scenes.

It is common to see several stories (or lines of action) told in parallel in a single TV series episode. Thus, the second level of structure is to cluster scenes into stories. We seek to deinterlace the stories in order to visualize the different lines of action independently.

The main difficulty is to determine the most relevant descriptors for grouping scenes belonging to the same story. We explore the use of descriptors from the three different modalities described above. We also propose methods to combine these three modalities.

To address the variability of the narrative structure of TV series episodes, we propose a method that adapts to each episode. It can automatically select the most relevant clustering method among the various methods we propose.

Finally, we developed StoViz, a tool for visualizing the structure of a TV series episode (scenes and stories). It allows an easy browsing of each episode, revealing the different stories told in parallel. It also allows playback of episodes story by story, and visualizing a summary of the episode by providing a short overview of each story.

Keywords : narrative structure, segmentation into scenes, scenes clustering into stories, plot deinterlacing, TV series, color, speakers, automatic speech recognition.