Vers un résumé automatique de séries télévisées basé sur une recherche multimodale d'histoires

Philippe Ercolessi [1], **Christine Sénac** [1], **Hervé Bredin** [2], **Sandrine Mouysset** [1]

1. *Institut de Recherche en Informatique de Toulouse*
   Université Paul Sabatier 118 Route de Narbonne, 31062 Toulouse, France
   ercolessi@irit.fr/senac@irit.fr/mouysset@irit.fr
2. *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur*
   CNRS BP133, 91403 Orsay, France
   bredin@limsi.fr

RÉSUMÉ. *Les séries télévisées récentes multiplient les personnages principaux, développant ainsi des intrigues complexes présentées à travers plusieurs histoires entremêlées. Nous proposons une approche de détection automatique de ces histoires afin de générer un résumé vidéo par extraction de scènes représentatives de ces dernières, et nous présentons un outil de visualisation rapide des histoires et du résumé obtenus. A partir d'une segmentation des épisodes en scènes (présentant une unité de temps, d'action et de contenu sémantique), les histoires s'obtiennent en regroupant les scènes, non nécessairement contiguës, qui présentent une similarité sémantique forte. Les modalités utilisées sont visuelles, audio et textuelles. Nos expérimentations sont menées sur deux séries télévisées de formats différents.*

ABSTRACT. *Modern TV series have complex plots made of several intertwined stories following numerous characters. In this paper, we propose an approach for automatically detecting these stories in order to generate video summaries and we propose a visualization tool to have a quick and easy look at TV series. Based on automatic scene segmentation of each TV series episode (a scene is defined as temporally and spatially continuous and semantically coherent), scenes are clustered into stories, made of (non necessarily adjacent) semantically similar scenes. Visual, audio and text modalities are combined to achieve better scene segmentation and story detection performance. An extraction of salient scenes from stories is performed to create the summary. Experimentations are conducted on two TV series with different formats.*

MOTS-CLÉS : *détection d'histoires, résumé de séries télévisées, classification spectrale, regroupement hiérarchique, multimodalité.*

KEYWORDS: *plot de-interlacing, TV Series summarization, spectral clustering, hierarchical clustering, multimodality.*

DOI:10.3166/DN.15.2.9-34

1. Introduction

La télévision a changé. Depuis la fin des années 1980, on a pu observer une multiplication des émissions et des séries télévisées. Les dix dernières années ont vu apparaître les services de vidéo à la demande et de nouveaux matériels permettant d'enregistrer des milliers d'heures de vidéos. Il est devenu nécessaire de procurer aux utilisateurs des outils efficaces leur permettant de naviguer au sein de ces gigantesques bases de données et de rechercher le contenu dont ils ont besoin le plus rapidement et le plus facilement possible.

Les travaux présentés dans cet article se focalisent sur un type de média bien particulier : les séries télévisées. Nous proposons ici un premier travail sur la génération automatique de résumés pour ce type de vidéos, et cela fait partie d'un plus grand ensemble d'outils de génération de résumés permettant la visualisation rapide et efficace d'une série entière, d'une saison, ou d'épisodes de séries télévisées.

Dans la littérature, il existe deux principaux types de résumés de vidéos : la représentation statique par images clefs (Sujatha, Mudenagudi, 2011) ou l'extraction de séquences dans la vidéo qui forment un résumé de type "bande annonce" ou *skimming* (Jiang *et al.*, 2009). C'est cette dernière forme de résumé que nous proposons dans cet article.

Nous partons du principe qu'un bon résumé doit être capable, en un minimum de temps, de donner le plus d'informations pertinentes possible. La recherche d'informations appropriées au sein de la vidéo est donc un élément nécessaire à la construction d'un bon résumé, et cette recherche passe par la compréhension de la structure de la vidéo. C'est pourquoi dans cet article, nous nous focalisons dans un premier temps sur la recherche des histoires qui constituent une information très importante concernant la narration.

Dans les vieilles séries télévisées comme *Mc Guyver, 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, le nombre de personnages principaux tend à augmenter, et il est devenu courant de voir plusieurs histoires racontées en parallèle dans un même épisode, comme par exemple dans la série *Malcolm*, où la majorité des épisodes sont 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.

Cependant, la structuration automatique d'une vidéo est un domaine de recherche très actif depuis les années 1990. Elle inclut en particulier les méthodes de segmentation telles que la segmentation en plans (Boreczky, Rowe, 1996 ; Hanjalic, 2002) ou en scènes (Yeung *et al.*, 1998 ; Sundaram, Chang, 2002 ; Zhu, Liu, 2009), et plusieurs travaux se concentrent sur la structuration de vidéos issues de la télévision (Abduraman *et al.*, 2011). La segmentation en thèmes ou séquences pour des émissions ayant des structures stables comme les journaux télévisés (Hauptmann, Witbrock, 1998 ; Hsu *et al.*, 2004) a particulièrement été étudiée, ainsi que la structuration d'émissions de sports (Assfalg *et al.*, 2002).

La recherche automatique des histoires pour des épisodes de série télévisée ne peut pas s'appuyer sur le type de méthodes utilisées pour des émissions ayant une structure préformatée. En effet, les épisodes de différentes séries peuvent avoir des formats très différents et des structures narratives très diverses. De plus, les histoires telles que nous les retrouvons dans des épisodes de séries télévisées ne sont pas continues comme les séquences d'un journal, puisque les différentes histoires sont *entrelacées* (évoluant en parallèle et non séquentiellement).

Notre méthode de création de résumé est directement inspirée des méthodes de génération automatique de résumé de texte, basées sur le principe de la catégorisation et de l'extraction de phrases clefs (Zha, 2002 ; Radev *et al.*, 2004 ; Pei-ying, Cun-he, 2009). Nous étudions l'utilisation de méthodes de segmentation et de regroupement des scènes d'un épisode de série télévisée pour rechercher les histoires sur lesquelles nous basons la création du résumé de la vidéo par extraction des scènes représentatives.

2. Présentation générale

L'approche de résumé automatique de documents audiovisuels de type série télévisée que nous proposons s'inspire directement des approches de résumé automatique de texte par extraction. Ce parallèle est mis en évidence dans la figure 1.

Là où un document textuel est constitué d'une suite de mots, un document vidéo peut être décrit comme une suite de plans, qui constituent donc notre unité de base d'un document vidéo.

DÉFINITION 1 (Plan). — *Un plan est une séquence vidéo prise à l'aide d'une caméra sans interruption.*

La question de la détection automatique de changement de plans (ou segmentation en plans) est considérée, dans la plupart des cas, comme un problème résolu (Hanjalic, 2002). Nous supposerons par la suite qu'une segmentation parfaite en plans est d'ores et déjà disponible et ne décrirons pas cette étape.

Plusieurs plans consécutifs peuvent être regroupés pour former des scènes. L'étape suivante vise ainsi à obtenir automatiquement la segmentation d'un document vidéo en scènes.

La méthode proposée provient de nos précédent travaux dans ce domaine (Bredin, 2012 ; Ercolessi *et al.*, 2011). Pourtant, définir la notion de scène est un problème en soi tant le nombre de définitions existantes est proche du nombre de travaux portant sur la question de la segmentation automatique en scènes. Certains considèrent que les scènes n'ont rien à voir avec la sémantique (Sundaram, Chang, 2002) tandis que d'autres affirment le contraire (Zhu, Liu, 2009).

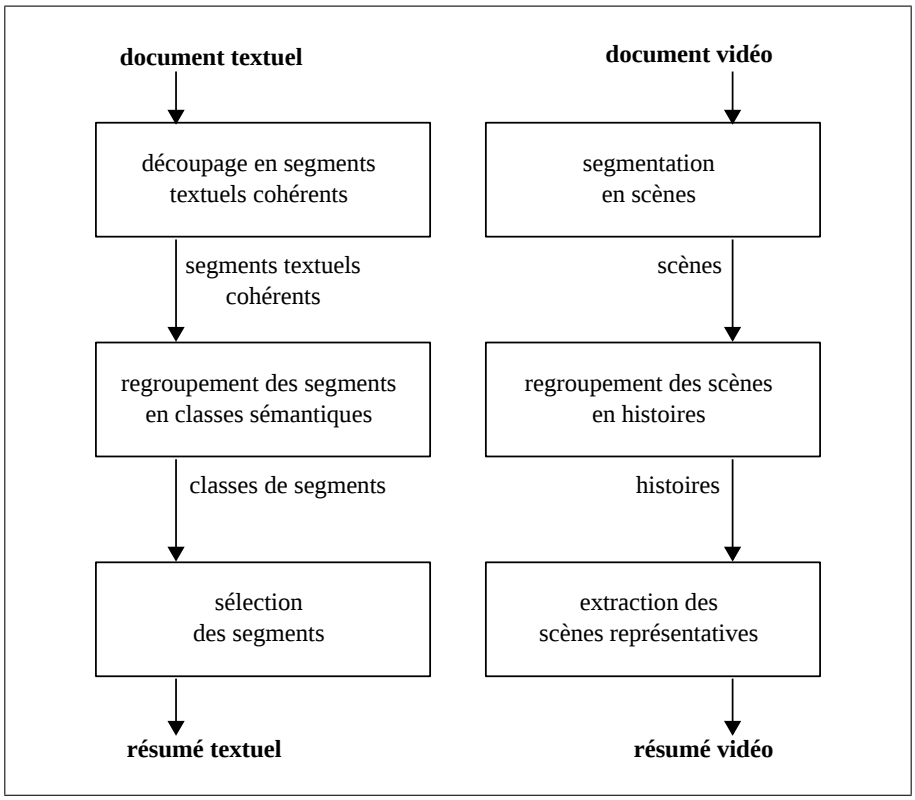

Figure 1. Résumé textuel vs. résumé vidéo

Les scènes dont il est question dans cet article suivent les propriétés suivantes :

DÉFINITION 2 (Scène). — *Une scène est composée d'une suite de plans et possède les caractéristiques suivantes : unité de temps et unité d'action.*

Une scène décrit un événement de manière continue. Ainsi, un flash-back *est considéré comme une nouvelle scène. Il en est de même lorsqu'un changement de décor intervient entre deux plans, indiquant qu'un laps de temps significatif a passé.*

Une scène décrit un seul événement. Même en cas de continuité temporelle, il peut arriver qu'un nouveau personnage ou un événement extérieur vienne perturber le déroulement de l'histoire : dans ce cas, nous considérons qu'une nouvelle scène commence.

La première étape consiste à découper automatiquement le contenu original en phrases pour les documents textuels et en scènes pour les documents vidéos. La section 3 est consacrée à la présentation de notre approche graphique pour la segmentation en scènes. Elle repose sur de multiples indices multimodaux tels que l'informa-

tion de couleurs ou la liste des personnages intervenant dans chaque plan (détectée automatiquement à l'aide de l'information accoustique).

De façon à limiter la redondance dans le résumé final, il est courant de procéder à un regroupement sémantique des phrases dans un document textuel de façon à ne conserver – lors de la phase suivante d'extraction – qu'une seule phrase par groupe sémantique. Afin de poursuivre l'analogie entre résumé textuel et résumé vidéo, nous introduisons la notion d'histoire dans les épisodes de série télévisées. En effet, il n'est pas rare que plusieurs histoires parallèles soient racontées de façon entremêlée au sein d'un même épisode.

DÉFINITION 3 (Histoire). — *Une histoire est un groupe de scènes qui présentent une relation sémantique forte. Cette relation met en exergue une similarité entre les scènes. Cette similarité peut s'exprimer de manière visuelle à travers un décor et/ou des personnages identiques, par exemple. Il peut aussi s'agir d'une synonymie, mettant alors en relief une similarité sémantique entre mots ou expressions. Ces unités de lieu, d'action ou de contexte permettent de déduire un enchaînement entre les scènes d'une même histoire.*

C'est l'objet de la section 4 que de présenter et évaluer notre approche multimodale pour le désentrelacement des histoires contenues dans un épisode. Tout comme le regroupement sémantique des phrases, elle est basée sur plusieurs techniques classiques de regroupement de scènes (ou classification de scènes) décrites à l'aide de différentes modalités.

Le résumé final est alors engendré par l'extraction d'un représentant de chaque groupe de phrases (pour le texte) ou de chaque histoire (pour la vidéo). Cette étape est décrite en détail dans la section 5.

La section 6 est dédiée à l'évaluation des méthodes de segmentation et de désentrelacement des histoires, ainsi qu'à une étude des résumés générés par cette méthode.

Alors qu'il est aisé de présenter un résumé textuel à un utilisateur sous la forme d'un court paragraphe, la présentation d'un résumé vidéo n'est pas si triviale. Nous avons donc mis en place STOVIZ, un démonstrateur en ligne visant à présenter les résumés de plusieurs épisodes de séries télévisées ainsi que la sortie de chacune des étapes intermédiaires. Il est rapidement introduit dans la section 7 et peut être consulté à l'adresse suivante : www.irit.fr/recherches/SAMOVA/ERCOLESSI/StoViz/.

3. Segmentation en scènes

Dans cette section, nous décrivons l'approche STG (*Scene Transition Graph*) permettant de détecter les frontières entre scènes (Yeung *et al.*, 1998) ainsi que son extension appelée approche STG généralisée (Sidiropoulos *et al.*, 2011). Ces deux approches reposent sur le fait que les frontières entre plans sont disponibles. Par ailleurs, la détection de frontières entre scènes revient à résoudre le problème de classification suivant : *"chaque frontière de plan est-elle également une frontière entre deux*

scènes?" Les deux approches, monomodale (description d'un plan par une forme unique de donnée) et multimodale (description du plan par une fusion de plusieurs formes de données), que nous avons développées pour la détection des frontières entre scènes reposent sur la méthode STG généralisée et sur la résolution d'un tel problème de classification.

3.1. Graphe de transition entre scènes (STG)

Soit d_{ij} la distance entre les plans i et j et soit t_{ij} leur distance temporelle. La distance entre deux plans peut être exprimée de manière simple par la distance entre les deux histogrammes de couleur représentant ces deux plans ; cette distance peut également être porteuse d'une information plus sémantique relative, par exemple, à des personnages ou bien à des dialogues.

En supposant que le plan i est antérieur temporellement au plan j, la distance temporelle t_{ij} est simplement définie par la durée entre la fin du plan i et le début du plan j. La distance et la distance temporelle sont ensuite combinées :

$$D_{ij} = \begin{cases} d_{ij} & \text{si } t_{ij} < \Delta_t \\ +\infty & \text{sinon} \end{cases} \quad (1)$$

Un *regroupement agglomératif de type 'complete link'* permet de regrouper les plans suivant cette nouvelle distance D. Un tel regroupement repose sur une phase initiale qui crée autant de groupes qu'il y a d'éléments (des plans dans notre cas). Ensuite les deux groupes les plus proches sont récursivement regroupés jusqu'à ne constituer qu'un seul groupe ou bien jusqu'à ce qu'un critère d'arrêt soit atteint.

La distance d_{ij}^C entre deux groupes C_i et C_j est définie comme étant la distance maximale entre les éléments de chaque groupe :

$$d_{ij}^C = \max_{(i,j)\in\{1,..,N\}X\{1,..,N\}} d_{kl}^e, \ \forall i,j \in \{1,..,N\}, \quad (2)$$

où $|C_i|$ est le nombre d'éléments dans le groupe C_i et d_{kl}^e la distance entre les éléments k de C_i et l de C_j.

Nous choisissons d'arrêter le regroupement lorsque la distance entre les groupes les plus proches dépasse un seuil Δ_d. Globalement, ce processus permet de regrouper les plans similaires dans la mesure où est respectée une contrainte de proximité temporelle dans la vidéo. La figure 2, montre un exemple d'application de cette approche où 11 plans consécutifs sont affectés à 5 groupes différents.

Par conséquent, le graphe de transition entre scènes est construit de façon à affecter un sommet à chaque plan et à relier par un arc chaque paire de sommets correspondant à deux plans consécutifs. Les plans sont ensuite regroupés selon la méthode précédemment évoquée, et les arcs de coupe du graphe (un arc de coupe est un arc dont la suppression entraîne une partition des groupes de sommets en deux sous-ensembles, les groupes de sommets de chaque sous-ensemble étant connectés) sont ensuite détectés

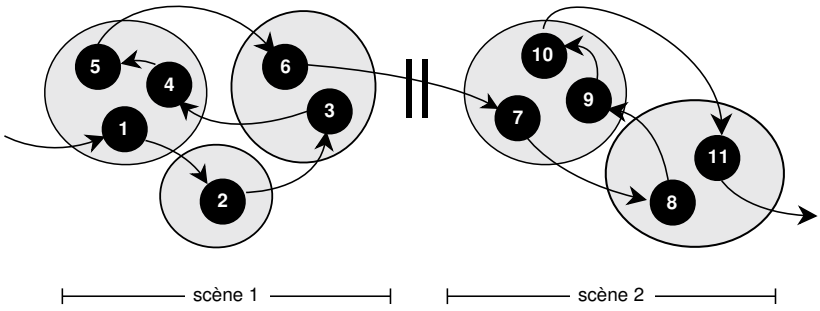

Figure 2. Graphe de transition entre scènes

et les frontières de plan correspondantes sont étiquetées comme des frontières entre scènes. Sur la figure 2, l'arc entre les sommets #6 et #7 est un arc de coupe. Ainsi, la frontière entre les plans #6 et #7 est marquée comme une frontière entre scènes.

3.2. Graphe généralisé de transition entre scènes (GSTG)

Chaque paire de valeurs (Δ_d, Δ_t) conduit à un ensemble différent de frontières entre scènes. Les valeurs optimales (*i.e.* celles produisant le meilleur ensemble de frontières) sont dépendantes de la vidéo. Une manière élégante de s'affranchir partiellement d'un apprentissage a été proposée par Sidiropoulos et al. (2011) grâce à l'introduction du STG généralisé (GSTG). L'idée est de générer un grand ensemble de STG en sélectionnant des valeurs aléatoires pour Δ_d et Δ_t, et de calculer pour chaque frontière de plans le pourcentage de STGs l'ayant détectée comme une frontière entre deux scènes. Sur la figure 3, les frontières de plan ayant un pourcentage p supérieur à un seuil θ sont marquées comme frontières entre scènes (lignes verticales en pointillé).

Sidiropoulos et al. ont trouvé que cette approche donnait de meilleures performances et nos expériences préliminaires ont confirmé cette observation. Non seulement il est plus aisé d'apprendre un seul paramètre (le seuil θ) au lieu de deux (Δ_d et Δ_t), mais également l'approche GSTG donne les meilleures performances (mais bien sûr biaisées) quand les paramètres sont appris directement sur l'ensemble de test.

Plutôt que de sélectionner les valeurs aléatoirement pour Δ_d et Δ_t, notre implémentation génère un ensemble exaustif 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 originale est d'être déterministe et donc de conduire à des résultats reproductibles.

La détermination des STG et GSTG dépend uniquement de la façon dont est calculée la distance entre les plans. Les sections 3.3 et 3.4 montrent les différentes approches que nous avons utilisées pour détecter les scènes en fonction du calcul de la distance entre les plans.

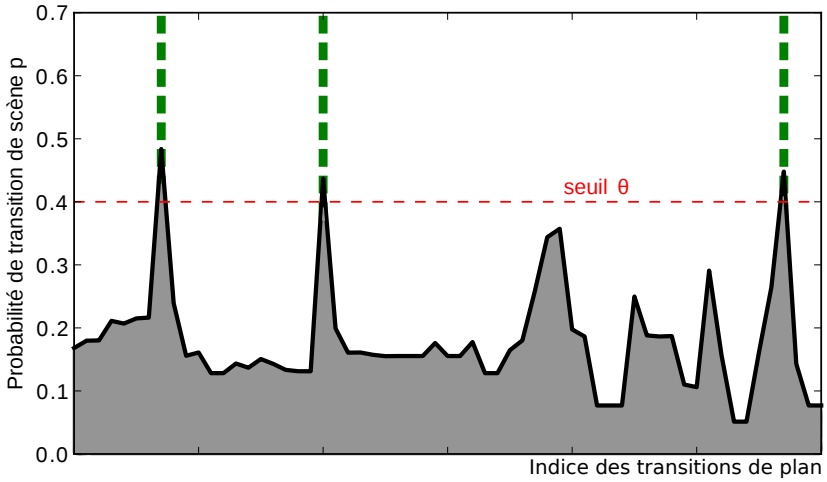

Figure 3. Probabilité de transition entre scènes

3.3. Approches monomodales

Dans cette section, nous décrivons deux nouvelles approches monomodales pour la segmentation de vidéos en scènes. Les deux sont basées sur la méthode GSTG proposée par Sidiropoulos et al. mais diffèrent sur le calcul de la distance entre plans.

Nous avons implémenté un système monomodal de base qui repose sur les histogrammes de couleur HSV. Les histogrammes de couleur (10x10x10 bins) sont extraits chaque seconde et la distance d_{ij}^{HSV} entre deux plans i et j est exprimée par la distance de Manhattan (d_M) minimale entre toutes les paires possibles d'histogrammes issus de ces deux plans.

$$d_{ij}^{\text{HSV}} = \begin{cases} \frac{1}{|H_i|} \sum_{h \in H_i} \min_{g \in H_j} d_M(h,g) & \text{si } |H_i| > |H_j| \\ d_{ji}^{\text{HSV}} & \text{sinon} \end{cases}$$
$$\forall i,j \in \{1,..,N\} \qquad (3)$$

Cependant, on ne peut pas espérer détecter des transitions entre scènes en utilisant uniquement des descripteurs bas-niveau. Aussi, comme résumé sur la figure 4, parallèlement aux histogrammes de couleur HSV de base, nous proposons d'utiliser des descripteurs extraits de la bande son et apportant des informations de niveau sémantique plus élevé.

3.3.1. Segmentation et regroupement en locuteurs (SD)

La segmentation et regroupement en locuteurs (ou speaker diarization) est le processus automatique qui consiste à découper la bande son en segments homogènes au

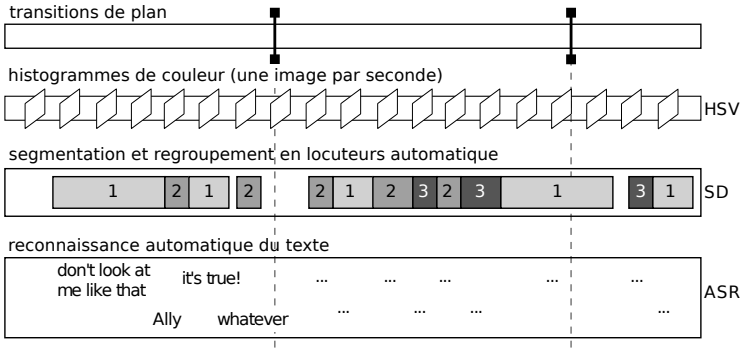

Figure 4. Ensemble des modalités utilisées

sens acoustique et à les regrouper de façon à ce qu'un groupe corresponde aux segments de parole prononcée par un unique locuteur. Dans l'idéal, une fonction bijective est appliquée entre l'ensemble des locuteurs du document sonore et les groupes de segments. La ligne **SD** sur la figure 4 montre un exemple de sortie d'un tel système : les tours de parole sont détectés et étiquetés avec une identité unique de locuteur (1, 2 ou 3). Sur cet exemple, le nombre de locuteurs durant un plan peut varier de zéro à trois.

De façon à calculer une distance unique d_{ij} entre chaque paire de plans (i, j), nous proposons d'utiliser la méthode TF-IDF (de l'anglais Term Frequency-Inverse Document Frequency), empruntée à la communauté de fouille de documents textuels.

Chaque plan s est décrit par un vecteur $X(s)$ de dimension D_{SD} où D_{SD} est le nombre total de locuteurs dans la video et $\text{X}_\lambda(s) = \text{TF}_\lambda(s) \times \text{IDF}_\lambda$ pour $\lambda \in \{1 \dots D_{\text{SD}}\}$:

– Le terme IDF est défini par $\text{IDF}_\lambda = \log(N/N_\lambda)$ où N est le nombre de plans dans la vidéo et N_λ est le nombre de plans durant lesquels le locuteur λ parle.

– Le terme TF est défini par $\text{TF}_s^\lambda = L_\lambda(s)/L(s)$ où $L(s)$ est la durée du plan s et $L_\lambda(s)$ est le temps cumulé de parole du locuteur λ dans le plan s.

La distance d_{ij}^{SD} entre les plans i et j basée sur le résultat du SD est définie comme la distance cosinus entre leurs vecteurs TF-IDF respectifs.

3.3.2. Reconnaissance automatique de la parole (ASR)

Afin d'apporter davantage d'informations sémantiques, nous proposons également d'utiliser la sortie d'un système de reconnaissance automatique de la parole (ASR) (Gauvain *et al.*, 2002) comme modalité complémentaire. A partir de la sortie de l'ASR, le programme TreeTagger (Schmid, 1994) permet d'extraire les lemmes pour les mots reconnus. Chaque plan s est ensuite décrit par un vecteur TF-IDF de dimension D_{ASR}, où D_{ASR} est le nombre total de lemmes uniques reconnus par le sytème ASR :

– Le terme IDF est défini par $\text{IDF}_\ell = \log(N/M_\ell)$ où N est le nombre de plans dans la video et M_ℓ est le nombre de plans contenant au minimum une occurence du $\ell^{ième}$ lemme.

– Le terme TF est défini par $\text{TF}_s^\ell = W_d(s)/W(s)$ où $W_\ell(s)$ du $\ell^{ième}$ lemme dans le plan s et $W(s)$ est le nombre de mots reconnus dans le plan s.

La distance d_{ij}^{ASR}, basée sur la sortie de l'ASR, entre les plans i et j est définie par la distance cosinus entre leurs vecteurs TF-IDF respectifs.

3.4. Fusion multimodale

Dans cette section, nous combinons les approches monomodales décrites ci-dessus afin de produire un système multimodal de détection de frontières entre scènes. La figure 5 montre trois approches possibles de combinaison (précoce, intermédiaire et tardive) appliquées à la méthode basée sur l'utilisation d'un GSTG. Nous ne présentons ici que la fusion intermédiaire qui a conduit aux meilleurs résultats.

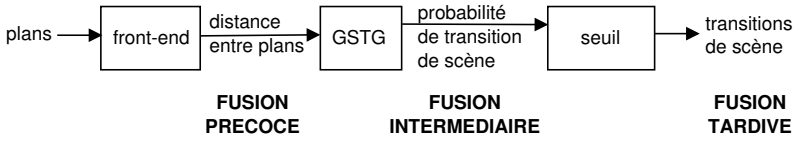

Figure 5. Fusion précoce vs. intermédiaire vs. tardive

La fusion intermédiaire consiste à effectuer une combinaison linéaire des probabilités p des frontières de scènes générées par les GSTGs monomodaux. Pour chaque frontière de plan, sa probabilité d'être une frontière de scène est définie par :

$$p = w_{\text{HSV}} \cdot p_{\text{HSV}} + w_{\text{SD}} \cdot p_{\text{SD}} + w_{\text{ASR}} \cdot p_{\text{ASR}} \tag{4}$$

avec la contrainte sur les poids $w_{\text{HSV}} + w_{\text{SD}} + w_{\text{ASR}} = 1$. Un paradigme de validation croisée est utilisé afin d'estimer le meilleur poids de chaque modalité. Cette approche particulière est celle proposée par Sidiropoulos et al. – même s'ils combinent des modalités différentes des notres.

4. Détection des histoires

Nous venons de voir qu'une scène est une suite de plans consécutifs décrivant un événement unique. Afin de décrire au mieux ces événements qui composent les scènes, nous avons caractérisé ces dernières au moyen des trois modalités précédemment décrites : histogrammes de couleur (qui peuvent rendre compte du contexte visuel), sortie d'un système de segmentation et regroupement en locuteurs (qui fournit une connaissance sur la répartition des locuteurs) et sortie d'un système de reconnaissance automatique de la parole (dont on extrait les lemmes les plus significatifs). La cohérence sémantique des scènes ainsi obtenues automatiquement est évaluée dans la section 6.

Comme illustré sur la figure 6, le désentrelacement des histoires consiste à séparer chaque histoire en regroupant les scènes, non nécessairement adjacentes, qui les composent. Ainsi, le désentrelacement des histoires peut être vu comme un problème de regroupement de scènes.

Le regroupement, que l'on appelle aussi classification, s'effectue généralement en deux étapes. Tout d'abord une mesure de distance doit être calculée pour chaque paire d'entités à regrouper (des scènes dans notre cas). Ensuite, en se basant sur la matrice de distances entre entités, l'algorithme constitue des groupes d'entités homogènes. Le choix de la mesure de distance est critique et doit être effectué en concordance avec les sorties souhaitées du regroupement. Ce choix est décrit dans la section 4.1. De même, il existe une multitude de méthodes de regroupement et nous utilisons deux de ces méthodes dans la section 4.2.

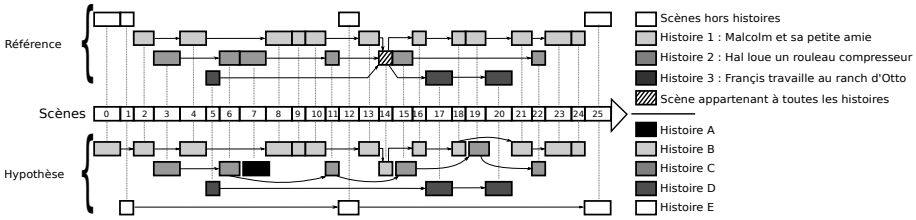

Figure 6. Exemple de désentrelacement des histoires pour un épisode de série télévisée. Cet exemple montre une segmentation en scènes et une représentation du désentrelacement automatique (hypothèse) et manuel (référence) des histoires

4.1. Calcul de distances entre scènes

Inspirés par les unités classiques d'action, lieu et temps du théâtre, nous proposons d'utiliser trois types de distances entre scènes basées sur les trois différentes modalités déjà utilisées pour la constitution des scènes.

4.1.1. Histogrammes de couleur (HSV)

Deux scènes qui se déroulent dans le même lieu ont une plus grande probabilité d'appartenir à la même histoire que si elles prennent place dans des lieux totalement différents. Partant de cette hypothèse, la première mesure de distance est donc basée sur le contenu visuel de chaque scène (à travers l'exploitation des histogrammes de couleur HSV). En effet, bien qu'ils soient considérés comme des informations de bas niveau, on peut attendre des histogrammes de couleur qu'ils apportent des informations relatives aux lieux mais également aux personnages (par la couleur de leurs vêtements par exemple).

4.1.2. Segmentation et regroupement en locuteurs (SD)

De même, il est très probable qu'une histoire suive un personnage particulier (ou un petit groupe de personnages). Ainsi, notre seconde mesure de distance repose sur

l'utilisation de la sortie d'un système de segmentation et regroupement en locuteurs (ou *Speaker Diarization* SD) : un personnage qui parle durant une histoire implique généralement qu'il fait partie de cette histoire. Aussi, de façon à calculer une distance unique d_{ij}^{SD} pour chaque couple de scènes (i, j), nous proposons de comptabiliser le nombre de locuteurs communs pour chaque couple.

Les paires de scènes p_{ij} sont triées par ordre décroissant de locuteurs communs entre les deux scènes i et j. Les paires de scènes ayant obtenu des valeurs identiques sont triées par ordre croissant du nombre total de locuteurs parlant dans p_{ij}. On note R_{ij} le rang de p_{ij} dans ce tri.

La distance d_{ij}^{SD} entre les scènes i et j est définie par :

$$d_{ij}^{\text{SD}} = \frac{R_{ij}}{P_{tot}},\ \forall i, j \in \{1, .., N\}, \tag{5}$$

où P_{tot} est le nombre total de paires de scènes dans l'épisode.

Ainsi, plus deux scènes ont de personnages en commun, plus leur distance sera faible.

4.1.3. Reconnaissance automatique de la parole (ASR)

La troisième modalité a pour but d'essayer de combler le *gap sémantique* en tenant compte des sujets de discussion entre personnages. Les sujets de discussions sont caractérisés par l'ensemble des lemmes prononcés durant la scène et dont l'importance est pondérée par le calcul du TF-IDF de chacun des lemmes.

Cette modalité repose sur la sortie du système d'ASR. L'explication donnée dans la section 3.3.2 pour l'obtention des termes TF et IDF s'applique ici en remplaçant le mot *plan* par *scène*.

La distance basée sur l'ASR d_{ij}^{ASR} entre deux scènes i et j est définie comme la distance cosinus entre leurs vecteurs TF-IDF respectifs.

$$d_{ij}^{\text{ASR}} = d^{cos}(\text{TF-IDF}_i, \text{TF-IDF}_j), \tag{6}$$

où d^{cos} est la distance cosinus entre deux vecteurs.

4.1.4. Fusion multimodale (FUS)

Nous proposons maintenant de fusionner les trois modalités présentées précédemment. Pour cela, nous définissons une matrice dont les éléments sont constitués des distances issues d'HSV, ASR et SD : chaque dimension correspond à une modalité différente. Ensuite un noyau gaussien est appliqué à cette distance de fusion multimodale pour définir la mesure de similarité entre deux scènes. L'utilisation d'un tel noyau gaussien permet de définir des classes sans avoir de connaissance a priori sur les distributions des distances. En d'autres termes, les données, difficilement séparables dans l'espace des données, sont projetées dans un espace de plus grande dimension dans

lequel les données seront séparables. La distance associée à la fusion, notée d^{FUS}, est définie comme suit :

Soit $V \in \mathbb{R}^{N\times N\times 3}$ une matrice composée des distances définies respectivement par les équations (3), (5) et (6) telles que :

$$V_{ij} = [d_{ij}^{\text{HSV}}, d_{ij}^{\text{SD}}, d_{ij}^{\text{ASR}}],\ \forall i,j \in \{1,..,N\}. \tag{7}$$

La distance de la fusion d^{FUS} est définie par la mesure gaussienne suivante :

$$d_{ij}^{FUS} = e^{-\frac{\|V_{ij}\|_2^2}{2\sigma^2}},\ \forall i,j \in \{1,..,N\}, \tag{8}$$

avec $\sigma = \max_{i,j\in\{1,..,N\}} \|V_{ij}\|_2$ et $\|.\|_2$ la norme euclidienne.

4.2. Méthodes de regroupement

Dans le but d'obtenir nos histoires, nous utilisons deux méthodes de classification pour rechercher les groupes de scènes qui nous intéressent : une classification agglomérative et une classification spectrale. Les méthodes de classification permettent de classer les scènes en groupes de telle sorte qu'un groupe de scènes corresponde à une histoire.

4.2.1. Classification agglomérative de type average-link

Le processus de classification s'effectue séquentiellement en agglomérant à chaque étape les deux groupes les plus proches. Ici, la distance d_{ij}^C entre deux groupes C_i et C_j est définie comme étant la distance moyenne entre les éléments de ces deux groupes :

$$d_{ij}^C = \frac{1}{|C_i||C_j|}\sum_{k=1}^{|C_i|}\sum_{l=1}^{|C_j|} d_{kl}^e,\ \forall i,j \in \{1,..,N\}, \tag{9}$$

où $|C_i|$ est le nombre d'éléments dans le groupe C_i et d_{kl}^e la distance entre les éléments k de C_i et l de C_j.

Le regroupement se termine lorsque la dérivée première de la distance entre les deux groupes les plus proches atteint son maximum. Ainsi, le nombre optimal de groupes est déterminé automatiquement.

4.2.2. Classification spectrale

La classification spectrale consiste à créer, à partir des éléments spectraux d'une matrice d'affinité gaussienne, un espace de dimension réduite dans lequel les données transformées seront linéairement séparables et pourront donc être plus facilement regroupées en classes (Ng *et al.*, 2001 ; Shi, Malik, 2000). Ci-dessous, nous présentons la méthode de classification spectrale adaptée à la fusion multimodale et nous proposons une heuristique pour déterminer automatiquement le nombre de classes, noté k et ainsi, fournir une classification spectrale totalement non supervisée.

Principe. Soit l'ensemble de données V défini par l'équation (7) et illustré par le graphe (a) de la figure 7. La matrice d'affinité gaussienne représente la distance de fusion multimodale d_{ij}^{FUS} définie par l'équation (8) entre toutes les scènes i et j, pour $i, j \in \{1, .., N\}$.

L'algorithme utilisé ici projette les données originales dans un espace défini par les k plus grands vecteurs propres d'une matrice d'affinité gaussienne normalisée (graphe (b) de la figure 7). Dans cet espace, la méthode des *K-means* est appliquée sur ces nouvelles données (Maila, Shi, 2001) afin de les regrouper en k classes.

Une relation d'équivalence permet de passer directement de la partition dans l'espace spectral à celle dans l'espace de données initiales.

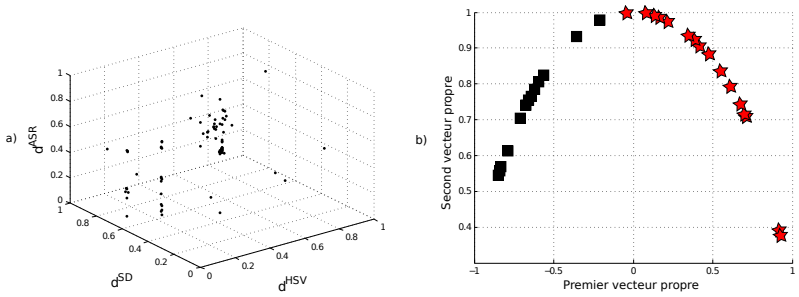

Figure 7. Méthode de classification spectrale avec FUS
(a) Matrice V définie dans l'équation (7)
(b) Regroupement dans la projection spectrale pour $k = 2$

Nombre de classes k. Le nombre k de classes (ou groupes) est obtenu à partir de la matrice d'affinité gaussienne réordonnancée par classe (Mouysset *et al.*, 2011). En effet, dans le cas idéal, la matrice affinité bien que dense, possède une structure numérique proche de celle d'une matrice bloc-diagonale, chaque bloc correspondant à une classe. Les blocs diagonaux de cette matrice représentent l'affinité intraclasse alors que les blocs hors diagonaux représentent l'affinité interclasses. On choisit le nombre de groupes k de façon à minimiser l'affinité intergroupes et à maximiser l'affinité intragroupe. Pour cela, on calcule le ratio moyen en norme de Frobenius entre tous les blocs hors diagonaux et les blocs diagonaux de la matrice affinité réordonnancée par classe. Un ratio proche de 0 montre que la matrice affinité a une structure bloc-diagonale proche du cas idéal définissant ainsi une bonne partition des données. On réalise donc une classification spectrale pour différentes valeurs de k et le regroupement final correspondra à la valeur de k pour laquelle le ratio est minimal.

4.3. Episodes dirigés par les personnages et sélection de la méthode de classification

Les épisodes de séries télévisées peuvent avoir différents formats. Dans cette section, nous distinguons deux types d'épisodes en fonction de la manière dont sont orga-

nisées les histoires qui les composent : les épisodes dirigés par les personnages (notés EdP) qui sont généralement constitués de 3 ou 4 histoires centrées sur des communautés indépendantes de personnages, et les autres épisodes (notés $\overline{\texttt{EdP}}$) pour lesquels les personnages ne suffisent pas à définir les histoires.

Nous proposons de classer automatiquement chaque épisode en deux catégories : EdP ou $\overline{\texttt{EdP}}$. Notre approche est inspirée par les graphes sociaux de personnages (RoleNet) (Weng *et al.*, 2009).

Cette approche vise à appliquer des méthodes de regroupement de scènes différentes en fonction du type d'épisodes détecté. Les EdP étant composés d'histoires fortement dépendantes des personnages qui y apparaissent, un regroupement average-link avec la seule modalité SD sera appliqué pour la détection des histoires. Pour les $\overline{\texttt{EdP}}$, la fusion des modalités avec classification spectrale sera préférée pour rechercher les histoires, puisque les personnages seuls ne permettent pas de les définir.

4.3.1. Graphe social des personnages (RoleNet)

Un RoleNet permet d'établir les relations entre personnages de l'épisode. La figure 8 montre la construction d'un tel graphe. Chaque personnage (ou locuteur) est associé à un nœud du graphe. Une arête entre deux nœuds signifie que les deux personnages apparaissent dans au moins une scène commune, chaque arête étant valuée par le nombre de scènes en commun. Le graphe non orienté résultant représente donc les interactions entre personnages intrascènes d'un épisode.

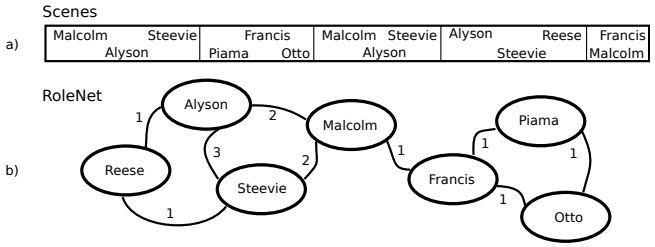

Figure 8. Graphe social des personnages
(a) Liste des personnages pour chaque scène
(b) Graphe social des personnages résultant

4.3.2. Détection des communautés de personnages

Nous proposons d'utiliser une méthode de l'état de l'art basée sur l'approche de *Louvain* (Blondel *et al.*, 2008) et dont le rôle est de détecter les communautés au sein d'un graphe.

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

$$\mathcal{Q} = \frac{1}{\sum_{i,j} A_{ij}} \sum_{i,j}^{N_c} \left[A_{ij} - \frac{\sum_k A_{ik} \sum_k A_{kj}}{\sum_{i,j} A_{ij}} \right] \delta_{ij} \qquad (10)$$

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

$\mathcal{Q}$ 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).

4.3.3. Détection des EdP

Plus la valeur de $\mathcal{Q}$ est élevée, plus les communautés de personnages sont disjointes et plus l'épisode peut être qualifié d'épisode dirigé par les personnages.

Aussi, nous proposons d'utiliser cette valeur pour détecter automatiquement les épisodes dirigés par les personnages. Les épisodes dont la modularité est supérieure à un seuil sont dits `EdP`, les autres étant répertoriés comme $\overline{\texttt{EdP}}$.

Le seuil optimal est automatiquement déterminé par un processus de cross validation 1 contre N.

5. Extraction des scènes représentatives pour la génération d'un résumé vidéo

Il s'agit maintenant de s'appuyer sur les histoires précédemment identifiées afin de générer le résumé.

Nous proposons une méthode inspirée de la génération de résumés de textes par extraction de phrases clefs (Radev *et al.*, 2004). 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.

C'est Luhn (1958) qui a ouvert la voie aux systèmes statistiques de résumé par extraction en construisant une liste de termes importants des documents et en se fondant sur leur fréquence. Seuls sont sélectionnés les termes dont la fréquence appartient à un intervalle prédéfini. Plus une phrase présente des mots appartenant à cette liste, plus elle est pertinente. Radev *et al.* ont profité des avancées dans le domaine des statistiques textuelles en intégrant le tf-idf (Salton, Buckley, 1988) à la méthode de Luhn. La liste des termes importants, que Radev appelle *centroïde*, est composée des n termes avec les plus grands tf-idf. Les phrases sont ensuite classées selon leur similarité au centroïde.

Nous procédons par analogie en considérant qu'une scène correspond à une phrase. Les scènes étant regroupées sous forme d'histoires, nous sélectionnons la scène la plus centrale de chaque histoire comme détaillé ci-dessous.

Soit une histoire H composée d'un ensemble de scènes telle que $H = \{S_0, S_1, ..., S_N\}$. La scène centrale de H, notée $\widehat{S}$, est une scène dont la distance moyenne à toutes les autres scènes de H est la plus faible, et cette distance est calculée sur les mêmes modalités que celles présentées à la section 4.

La distance utilisée dépend du type d'épisode. Pour un EdP la distance est calculée sur la modalité SD (locuteurs) alors qu'elle est calculée sur la fusion de toutes les modalités pour un $\overline{\text{EdP}}$.

$$\widehat{S} = \arg\min_{S_i} \frac{1}{N} \sum_{j=1}^{N} d(S_i, S_j) \tag{11}$$

Le résumé est ensuite composé de ces scènes centrales qui sont concaténées pour former le résumé en respectant l'ordre chronologique d'apparition dans la vidéo initiale.

6. Expérimentations et résultats

Nous avons évalué notre système sur deux séries télévisées : huit épisodes de la série *Ally McBeal* et 7 épisodes de la série *Malcolm*, contenant respectivement 5.5 heures de vidéo, 5 564 plans et 306 scènes et 20 histoires pour *Ally McBeal* et 2.5 heures de vidéo, 196 scènes et 24 histoires pour *Malcolm*.

6.1. Segmentation en scènes

Pour évaluer la méthode de segmentation en scènes, nous avons utilisé les huit épisodes de la première saison d'*Ally McBeal*. Afin de ne pas corrompre le système par des entrées éventuellement erronées, nous avons retouché manuellement les frontières des plans trouvés automatiquement. De plus, les scènes de la référence ont été annotées manuellement en vue de l'évaluation.

Les histogrammes de couleur HSV ont été extraits toutes les secondes en utilisant la librairie OpenCV (Bradski, 2000). La reconnaissance automatique des locuteurs et de la parole ont été générées automatiquement en utilisant le système de reconnaissance de la parole et des locuteurs du LIMSI (Gauvain *et al.*, 2002).

Nous considérons le problème de segmentation comme un problème de détection de frontière, qui sera validé par un calcul de précision et de rappel et de leur combinaison en F-mesure. Une transition détectée est correcte si elle a exactement la même position qu'une transition de la vérité terrain (et incorrecte dans le cas contraire). Aucune tolérance temporelle n'est permise.

Comme seulement 8 épisodes sont annotés, le protocole d'évaluation suit le principe de la validation croisée (leave-one-out cross validation).

Le tableau 1 résume les performances de nos approches. Pour les méthodes monomodales, le système de référence (basé uniquement sur les histogrammes de couleur HSV) obtient largement le meilleur résultat avec une F-mesure de $0,487$.

Les approches basées sur les modalités SD et ASR tendent à détecter beaucoup trop de transitions de scènes (resp. 1100+ et 1700+ scènes détectées alors que le

Tableau 1. Performance des approches mono et multimodales

	Précision	Rappel	F-mesure
HSV	0,45	0,57	0,49
SD	0,16	0,56	0,24
ASR	0,10	0,57	0,18
p(HSV) + p(SD) + p(ASR)	0,49	0,62	**0,54**

nombre correct de transitions de 305) et engendrent des taux de précision très bas (resp. $0,16$ et $0,10$). Ceci peut être expliqué par la façon dont les distances des modalités SD et ASR sont calculées. En effet, dans le cas de l'ASR où un vecteur TF-IDF est associé à chaque plan, deux plans partagent rarement plus de deux mots en commun, la distance sera toujours proche de 1 (c'est-à-dire quasiment maximale). C'est pourquoi le processus de regroupement agglomératif s'arrête très tôt (en fonction du seuil Δ_d), ce qui génère un graphe de transition entre scènes composé de beaucoup de groupes ne contenant qu'un seul plan.

Cependant, bien que les approches basées sur les modalités SD et ASR affichent les plus mauvais scores, leur fusion avec HSV permet d'améliorer la performance du système comme le montre la dernière ligne du tableau 1.

6.2. Détection des histoires

Aux sept épisodes annotés d'Ally McBeal déjà utilisés pour la segmentation en scènes, nous en ajoutons sept nouveaux tirés de la série *Malcolm*. Les annotations concernent la segmentation en scènes et en histoires : chaque histoire est définie par la liste des scènes qui la composent (une scène pouvant appartenir à plusieurs histoires).

Les deux collections sont des comédies américaines, mais offrent des formats très différents : les épisodes d'*Ally McBeal* durent 40 minutes contre 20 pour ceux de *Malcolm*. Et il y a en général plus d'histoires par épisode pour ce dernier et elles sont plus facilement identifiables que celles que l'on trouve dans les épisodes d'*Ally McBeal*.

6.2.1. Métriques

Nous introduisons une métrique d'évaluation empruntée à la communauté travaillant sur la segmentation et regroupement en locuteurs : le *taux d'erreur de segmentation et regroupement en locuteurs* (DER pour *Diarization Error Rate*).

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. Cette liste est appelée *l'hypothèse*. Elle est évaluée en la comparant à une liste manuellement annotée (*la référence*). Sachant que les identifiants de l'hypothèse et de la référence ne sont pas forcément les mêmes, l'évaluation recherche la meilleure correspondance entre les segments de l'hypothèse et ceux de la référence de manière à

ce que le chevauchement total entre les locuteurs de référence et ceux qui leur correspondent dans l'hypothèse soit maximisé. Le DER est défini comme la somme de trois erreurs :

$$DER = \frac{\text{Fausses alarmes} + \text{Non-détections} + \text{Erreurs de locuteurs}}{\text{Durée de l'épisode}} \quad (12)$$

Fausses alarmes : il s'agit de la durée totale des segments pour lesquels de la parole est détectée dans l'hypothèse alors qu'il n'y en a pas dans la référence.

Non-détections : il s'agit de la durée totale des segments pour lesquels de la parole est présente dans la référence mais qui n'est pas détectée dans l'hypopthèse.

Erreurs de locuteurs : il s'agit de la durée totale des segments pour lesquels l'identifiant du locuteur dans la référence n'est pas celui qui correspond avec l'identifiant de l'hypothèse.

Le DER peut directement être utilisé pour le problème de désentrelacement des histoires en utilisant l'analogie suivante : chaque histoire correspond à un locuteur et chaque scène à un segment de parole. Dans notre cas, la distinction entre les segments de parole et de non-parole n'est pas évidente, puisque les segments de scènes couvrent la durée totale de l'épisode. Cette propriété fait que nous obtenons systématiquement des taux de fausses alarmes et de non-détections nuls. Cependant, quelques scènes particulières (sketchs isolés, scènes de transition) ne font parties d'aucune histoire. Elles peuvent alors correspondre à des segments de non-parole dans le calcul du DER.

Contrairement à la F-mesure, le DER mesure l'erreur d'une méthode et non son exactitude. Ainsi, un regroupement parfait obtiendra un DER de 0 alors que le pire regroupement obtiendra un DER proche de 1.

Cette métrique a l'avantage de prendre en compte la durée cumulée des erreurs. Ainsi, nous n'évaluons pas le nombre de scènes mal regroupées, mais plutôt la durée totale de vidéo correspondante.

6.2.2. Résultats

Le tableau 2 résume les résultats obtenus en utilisant toutes les combinaisons de méthodes de regroupement et de modalités utilisées.

Tableau 2. Regroupement de type average-link vs. classification spectrale appliqués aux différentes modalités. ***Aléa.*** *: Evaluation moyenne de 100 regroupements de scènes obtenus à partir de distances aléatoires entre les scènes.* ***Meill.*** *: Meilleure valeur atteignable avec nos méthodes (métrique d'évaluation : DER)*

	Average-link				**Classification spectrale**				**Aléa.**	**Meill.**
	HSV	ASR	SD	FUS	HSV	ASR	SD	FUS	**baselines**	
Moyenne	0,57	0,59	0,37	0,67	0,61	0,56	0,44	0,40	0,55	0,16
Mét. globale	**0,33**									

La colonne *Aléa* est obtenue en utilisant un regroupement average-link basé sur un calcul aléatoire des distances entre les scènes. Les valeurs aléatoires sont la moyenne

des résultats obtenus par 100 regroupements aléatoires auxquels nous avons retranché 3 fois l'écart-type de ces 100 résultats. Ainsi, toutes les valeurs qui sont plus petites que l'aléatoire sont meilleures que 95 % des 100 regroupements aléatoires.

La colonne *Meilleur* montre le meilleur score que nous pouvons obtenir avec nos approches. En effet, aucune des méthodes que nous avons utilisées ici ne permet qu'une scène fasse partie de plus d'une histoire. Ainsi, dans le meilleur des cas, si au moins une scène appartient à deux histoires (ou plus), le DER ne pourra jamais atteindre 0. Cette colonne est là pour illustrer la meilleure valeur de DER que nous pouvons obtenir en tenant compte de cette propriété.

En considérant les distances monomodales, le tableau 2 montre que seule la modalité SD donne des résultats meilleurs que l'aléatoire. La modalité HSV correspond à un descripteur très bas niveau qui n'offre que peu d'information sémantique, et la modalité ASR contient de très nombreuses erreurs provenant du système de reconnaissance de la parole utilisé, ce qui conduit aux mauvais résultats observés pour les expériences menées avec ces deux modalités.

Le tableau 3 montre l'intérêt de la sélection automatique de la méthode de regroupement en fonction de la structure des épisodes. Pour les `EdP`, le choix d'un regroupement average-link avec la modalité SD apporte une amélioration de 13 % par rapport à la classification spectrale avec la fusion des modalités. Pour les $\overline{\texttt{EdP}}$ en revanche, la classification spectrale donne une performance de 5 % meilleure que le regroupement average-link. Globalement, le tableau 2 montre que la sélection automatique de la méthode (ligne *Met. globale*) apporte une amélioration de 4 % par rapport au meilleur système (regroupement average-link associé à la modalité SD), et est inférieure seulement de 16 % par rapport au meilleur résultat pouvant être atteint.

Tableau 3. Episodes dirigés par les personnages (`EdP`) vs. autres épisodes ($\overline{\texttt{EdP}}$) (Evaluation = DER)

EdP		**$\overline{\texttt{EdP}}$**	
average-link (SD)	**spectral (FUS)**	**average-link (SD)**	**spectral (FUS)**
$0,20$	$0,33$	$0,5$	$0,45$

6.3. Extraction des scènes représentatives

Évaluer un résumé (de texte ou de vidéo) était jusqu'à il y a peu, une tâche exclusivement exécutée par l'humain et opérée de deux manières différentes. Dans la première, qui est *intrinsèque*, des tiers jugent la qualité du processus de construction du résumé en se basant directement sur l'analyse du résumé lui-même ; dans la seconde, *extrinsèque*, la qualité du résumé est déterminée en se basant sur la manière dont celui-ci influence l'accomplissement de certaines tâches comme l'interrogation de certaines personnes en leur demandant de répondre en se basant uniquement sur la lecture ou le visionnage des résumés construits.

Mais une telle évaluation, bien que significative, est extrêmement coûteuse en temps et très difficile à mettre en œuvre. Même si l'humain est capable de distinguer un *bon* résumé d'un *mauvais*, le résumé *idéal* n'existe pas, ce qui rend très difficile de définir une mesure de qualité pouvant être calculée automatiquement.

D'abord pour le résumé de texte (Lin, 2004 ; Das, Martins, 2007), puis pour le résumé de vidéo (Li, Mérialdo, 2010), des tentatives d'évaluation automatiques, basées sur des résumés références créés par des humains, sont apparues récemment.

La mesure ROUGE (Lin, 2004) (*Recall-Oriented Understudy for Gisting Evaluation*) détermine à quel point un résumé de texte généré par un système couvre le contenu présent dans un ou plusieurs résumés modèles (produits par des humains) que l'on appelle les références.

La mesure VERT (Li, Mérialdo, 2010) (*Video Evaluation by Relevant Threshold*) s'inspire directement de la ROUGE qui est une mesure orientée rappel. Nous détaillons cette mesure ci-dessous.

La mesure VERT part de l'hypothèse que le résumé à évaluer peut être soit une concaténation d'instants vidéo soit un ensemble d'images clefs (keyframes). Par mesure de simplicité, les explications suivantes se basent sur des images clefs. A chaque image clef est affecté un poids $W_s(f)$ dépendant du rang de l'image clef f dans le résumé. La mesure VERT compare cet ensemble ordonné d'images clefs avec plusieurs ensembles d'images sélectionnées par des humains. La métrique est ainsi définie :

$$VERT(C) = \frac{\sum_{S \in \{RésumésRéférence\}} \sum_{gram_n \in S} W_C(gram_n)}{\sum_{S \in \{RésumésRéférence\}} \sum_{gram_n \in S} W_S(gram_n)} \qquad (13)$$

Avec C le résumé vidéo candidat à évaluer, $gram_n$ un groupe de n images clefs, $W_S(gram_n)$ le poids du groupe $gram_n$ pour un résumé reference S, et $W_C(gram_n)$ le poids du groupe $gram_n$ pour le résumé C. Dans le numérateur de la formule, la somme sur les $W_C(gram_n)$ est effectuée seulement sur les $gram_n$ qui sont présents dans le résumé référence S. VERT calcule un pourcentage de $gram_n$ issus des résumés références et également présents dans le résumé candidat.

Les séries télévisées présentent généralement un résumé au début ou à la fin de certains épisodes, soit pour rappeler ce qu'il s'est passé dans l'épisode précédent, soit pour inciter le spectateur à suivre l'épisode suivant. Ce type de résumé, présent dans certaines séries populaires, est composé de très courtes séquences qui s'enchaînent suivant un rythme très élevé, et qui explique succinctement la ou les trames principales de l'épisode. Les séquences qui le composent ne suivent pas forcement l'ordre chronologique de leurs apparitions lors de l'épisode, et la partie audio utilisée pour une séquence vidéo du résumé n'est pas forcément la partie qui était utilisée dans l'épisode avec cette même séquence vidéo, permettant de multiplier le nombre d'informations fournies au spectateur tout en minimisant la durée du résumé.

Idéalement, nous voudrions atteindre la qualité de ce type de résumés avec notre système de génération automatique. Pour l'instant, notre système de génération de résumé permet de retrouver les différentes trames scénaristiques présentes dans un épisode, de sélectionner une unique scène de chaque histoire détectée et de les concaténer pour obtenir le résumé final.

A l'heure actuelle, nous avons mis en place un protocole de création de résumé par les humains. Pour chaque épisode, nous leur demandons de placer par ordre de représentativité des séquences de chaque histoire. Cependant, cette création de résumé est très coûteuse en temps, et nous avons récolté pour l'instant trop peu de résumés afin de mener à bien une évaluation avec la métrique VERT.

Donc en parallèle, nous avons développé STOVIZ, un système de visualisation des résumés et histoires générés automatiquement (voir section 7) qui permet de faire une évaluation de type intrinsèque. Cette visualisation nous permet d'avoir une idée des personnages présents dans la vidéo, des différentes intrigues qui s'y déroulent et du thème global de l'épisode dans la majorité des cas, tout en ne conservant, en moyenne, que 13 % de la durée de la vidéo d'origine.

Cependant, les scènes retenues sont peu nombreuses, souvent trop longues et offrent un rythme trop lent à la vidéo, ce qui empêche de capter l'attention du spectateur. Nous pensons cependant que la sélection des histoires peut nous permettre de fortement améliorer la qualité des résumés pour le type de média sur lequel nous travaillons.

Aussi, au lieu d'extraire la scène la plus représentative de l'histoire, nous pensons qu'il serait préférable de sélectionner un ensemble d'extraits de scènes pertinents au sens informatif. En effet, les scènes sélectionnées sont souvent très longues, or l'importance d'un sujet n'est pas nécessairement liée à son temps d'exposition. Pour cela, il est nécessaire de pousser la compréhension du contenu de la vidéo qui ne pourra se faire qu'avec l'amélioration des modalités. Une idée serait de se focaliser sur les zones de dialogues et de forte activité visuelle, afin d'en extraire les séquences saillantes.

7. STOVIZ : outil de visualisation

Afin d'étudier la pertinence des résumés que nous générons et des histoires détectées, nous avons mis en place un système de visualisation appelé "Stories Visualization System" ou STOVIZ. Il s'agit d'un outil permettant la visualisation d'une vidéo associée à une frise permettant d'observer de manière rapide et intuitive les différentes histoires détectées, de les comparer avec les histoires annotées manuellement (permettant par la même occasion d'observer la pertinence des annotations) et d'observer le résumé ainsi généré.

La figure 9 montre les fonctionnalités de l'outil STOVIZ, accessible à l'adresse "http://www.irit.fr/recherches/SAMOVA/ERCOLESSI/StoViz/".

STOVIZ permet de visualiser une vidéo en offrant plusieurs possibilités d'interaction avec l'utilisateur pour observer la structure de cette vidéo. En bas de la fenêtre

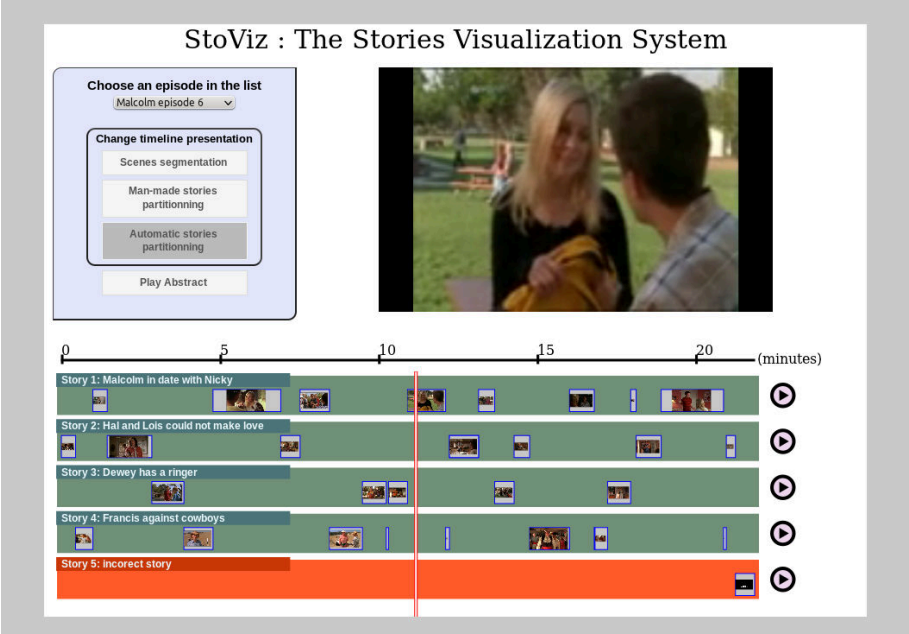

Figure 9. Outil de visualisation des histoires et des résumés

se trouve la *frise* qui permet de visualiser les différentes scènes (rectangles gris clair) ainsi que les différentes histoires détectées (rectangles englobants foncés). Le contenu de la frise peut être modifié à l'aide des boutons *Timeline, Groundtruth et Hypothesis*. *Timeline* permet l'affichage des scènes sur une seule ligne ne donnant pas plus d'information que le découpage en scènes de la vidéo. Le bouton *Hypothesis* permet de réorganiser les scènes de manière à ce qu'elles se positionnent sur différentes lignes, chaque ligne correspondant à une histoire issue de notre méthode de détection automatique. Le bouton *Groundtruth* permet de réorganiser les scènes en histoires telles qu'elles ont été annotées.

Il est possible de visualiser chaque histoire séparément en cliquant sur le rectangle qui lui correspond, et de visualiser le résumé généré automatiquement en cliquant sur le bouton *Play Abstract*.

La sélection de l'épisode se fait à partir de la liste déroulante en haut à droite de la fenêtre.

8. Conclusion

Dans ce papier, nous proposons un premier travail de génération automatique de résumés pour des épisodes de séries télévisées. Il s'agit d'une méthode directement héritée des systèmes classiques de génération automatique de résumés de texte en couplant une méthode de segmentation en scènes à une méthode de regroupement

sémantique en histoires de ces scènes non nécessairement contiguës. Le résumé est alors généré en sélectionnant les scènes représentatives de chaque histoire.

Les séries télévisées n'ayant pas une structure prédéfinie comme par exemple les journaux télévisés, nous ne pouvons pas nous appuyer sur une quelconque modélisation pour extraire les histoires. Les méthodes de segmentation et de regroupement reposent toutes sur l'utilisation de distances multimodales (audio, video, texte) entre des séquences de vidéo (plans, scènes).

Les résultats de la segmentation en scènes et de la détection automatique des histoires sont très prometteurs. La méthode de segmentation proposée permet de détecter correctement plus de la moitié des transitions entre les scènes et notre méthode de regroupement des scènes en histoires permet de se rapprocher de celui réalisé par des êtres humains pour de nombreux épisodes. Cependant, pour certains épisodes de type $\overline{\texttt{EdP}}$, le regroupement en histoires n'est pas conforme à nos attentes car nous nous heurtons au *gap sémantique* que les modalités utilisées ne suffisent pas à franchir. En effet, les similarités sémantiques que nous recherchons sont parfois différentes de celles obtenues à partir des caractéristiques bas niveau liées aux signaux audio et vidéo. De plus, pour l'instant, la méthode de regroupement en histoires ne permet pas d'associer une scène à deux histoires différentes : or, la plupart des épisodes contiennent de telles scènes charnières comme décrit sur la figure 6 (voir scène 14).

Toutefois, ces deux premières étapes de segmentation en scènes et regroupement en histoires nous semblent essentielles à la bonne structuration des épisodes de séries télévisées et donc à la génération du résumé.

Afin d'évaluer la pertinence des histoires détectées et du résumé généré, nous avons développé l'outil de visualisation STOVIZ.

Les résumés générés nous 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 dans la majorité des cas. Cependant, afin de leur donner davantage de rythme, nous travaillons à la recherche de zones d'activité et de dialogue à l'intérieur des scènes en vue de l'extraction de séquences plus courtes, le but ultime étant d'obtenir un résumé s'approchant des résumés que l'on peut trouver au début de certains épisodes et rappelant les intrigues passées.

Bibliographie

Abduraman A. E., Berrani S.-A., Mérialdo B. (2011). *TV program Structuring Techniques : A Review*. Book chapter in TV Content Analysis: Techniques and Applications, October 2011.

Assfalg J., Bertini M., Del Bimbo A., Nunziati W., Pala P. (2002). Soccer Highlights Detection and Recognition using HMMs. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference*, vol. 1, p. 825-828.

Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E. (2008). Fast Unfolding of Community Hierarchies in Large Networks. *Computing Research Repository*.

Boreczky J. S., Rowe L. A. (1996). Comparison of Video Shot Boundary Detection Techniques. In *Storage and Retrieval for Still Image and Video Databases IV*. Los Angeles, California.

Bradski G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Bredin H. (2012). Segmentation of TV Shows into Scenes using Speaker Diarization and Speech Recognition. In *ICASSP 2012, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Kyoto, Japan.

Das D., Martins A. F. T. (2007). *A Survey on Automatic Text Summarization.* Literature Survey for the Language and Statistics II course at Carnegie Mellon University.

Ercolessi P., Bredin H., Sénac C., Joly P. (2011). Segmenting TV Series into Scenes Using Speaker Diarization. In *WIAMIS*.

Gauvain J., Lamel L., Adda G. (2002). The LIMSI Broadcast News Transcription System. *Speech Communication*, vol. 37, n° 1-2, p. 89-109.

Hanjalic A. (2002). Shot-boundary Detection: Unraveled and Resolved? In *Circuits and Systems for Video Technology, IEEE Transactions*, vol. 12, p. 90-105.

Hauptmann A., Witbrock M. (1998). Story Segmentation and Detection of Commercials in Broadcast News Video. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum*, p. 168-179.

Hsu W., Kennedy L., Huang C.-W., Chang S.-F., Lin C.-Y., Iyengar G. (2004). News Video Story Segmentation using Fusion of Multi-level Multi-modal Features in TRECVID 2003. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference*, vol. 3, p. iii - 645-8.

Jiang R., Sadka A., Crookes D. (2009). Advances in Video Summarization and Skimming. In M. Grgic, K. Delac, M. Ghanbari (Eds.), *Recent Advances in Multimedia Signal Processing and Communications*, vol. 231, p. 27-50. Springer Berlin / Heidelberg.

Li Y., Mérialdo B. (2010). VERT: Automatic Evaluation of Video Summaries. In *ACM Multimedia*, p. 851-854.

Lin C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, p. 74-81. Barcelona, Spain, Association for Computational Linguistics.

Luhn H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.*, vol. 2, n° 2, p. 159-165.

Maila M., Shi J. (2001). A random walks view of spectral segmentation. In *"AI and STATISTICS (AISTATS) 2001"*.

Mouysset S., Noailles J., Ruiz D., Guivarch R. (2011). On a strategy for spectral clustering with parallel computation. *High Performance Computing for Computational Science–VECPAR 2010*, p. 408-420.

Newman M. E. J. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, n° 23, p. 8577-8582.

Ng A., Jordan M., Weiss Y. (2001). On Spectral Clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, p. 849–856. MIT Press.

Pei-ying Z., Cun-he L. (2009). Automatic Text Summarization Based on Sentences Clustering and Extraction. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference*, p. 167-170.

Radev D. R., Jing H., Styś M., Tam D. (2004). Centroid-based Summarization of Multiple Documents. In *Information Processing Management*, vol. 40, p. 919-938. Tarrytown, NY, USA, Pergamon Press, Inc.

Salton G., Buckley C. (1988). Term-weighting Approaches in Automatic Text Retrieval. In *Information Processing and Management*, p. 513-523.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49.

Shi J., Malik J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 22, n° 8, p. 888-905.

Sidiropoulos P., Mezaris V., Kompatsiaris I., Meinedo H., Bugalho M., Trancoso I. (2011). Temporal Video Segmentation to Scenes using High-level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sujatha C., Mudenagudi U. (2011). A Study on Keyframe Extraction Methods for Video Summary. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference*, p. 73-77.

Sundaram H., Chang S.-F. (2002). Computable Scenes and Structures in Films. , vol. 4, n° 4, p. 482 - 491.

Weng C.-Y., Chu W.-T., Wu J.-L. (2009). Rolenet: Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions*, vol. 11, n° 2, p. 256-271.

Yeung M., Yeo B., Liu B. (1998). Segmentation of Video by Clustering and Graph Analysis. *Computer Vision and Image Understanding*, vol. 71, p. 94-109.

Zha H. (2002). Generic Summarization and Keyphrase Extraction using Mutual Reinforcement Principle and Sentence Clustering. In *SIGIR*, p. 113-120.

Zhu S., Liu Y. (2009). Video Scene Segmentation and Semantic Representation using a Novel Scheme. In *Multimedia Tools Appl.*, vol. 42, p. 183-205. Hingham, MA, USA, Kluwer Academic Publishers.