EXPLOITATION DES RELATIONS TEMPORELLES ENTRE EVENEMENTS PRESENTS DANS LES DOCUMENTS AUDIOVISUELS

Zein Al Abidin IBRAHIM [1], Isabelle FERRANE [1] et Philippe JOLY [1]

[1] Université Paul Sabatier
IRIT, Toulouse
{Ibrahim, Ferrane, Joly}@irit.fr

Résumé : Le but de notre travail est de caractériser des structures intentionnelles dans des documents multimédia et particulièrement dans les vidéos. Pour cela, nous devons d'abord obtenir différentes segmentations du document pour étudier ensuite toutes les relations qui peuvent être observées entre ces segmentations. En ce qui nous concerne, nous ne disposons pas d'informations préalables concernant le type de la vidéo (sport, nouvelles…), la structure, ou le type de structure à rechercher. Notre travail est basé sur différents outils de segmentation fournissant automatiquement des unités temporelles en fonction de caractéristiques spécifiques essentiellement de bas niveau. A partir des segmentations obtenues, considérées deux à deux, nous effectuons une analyse des relations temporelles susceptibles d'exister entre les différents segments. Nous étudions la fréquence des réalisations de ces relations. Afin de donner une interprétation sémantique à ces observations, nous proposons une nouvelle représentation des relations temporelles, que nous appliquerons pour l'exemple aux relations de Allen. Après évocation du problème relatif aux erreurs de segmentation, nous présentons les premiers résultats obtenus après une première phase expérimentale produite sur des programmes TV et notamment des journaux télévisés.

Mots-clés : indexation multimédia, segmentation temporelle, relations de Allen.

1 Introduction

Beaucoup d'outils automatiques d'indexation de contenu audiovisuel produisent des résultats correspondant à des segments ou à des parties temporelles identifiant la présence ou non d'un objet spécifique. La plupart de ces outils se basent sur les résultats de l'extraction de caractéristiques bas niveaux : taux d'activité, couleur dominante, musique ou parole présents sur la bande sonore, textes ou caractères incrustés à l'écran … Leur but est de détecter des événements afin de créer des sommaires ou de déterminer des entrées temporelles dans un flot de données audiovisuelles. Ces premiers résultats sont ensuite utilisés soit pour constituer des index, soit pour servir de base à des étapes d'analyse ultérieures. La principale difficulté que l'on rencontre concerne l'interprétation sémantique que l'on peut attribuer aux index produits. Comme l'information pertinente utilisée durant le

processus d'analyse peut provenir de sources diverses (visuelles, audio, ou même textuelles) présentes dans un même document, on peut espérer que la combinaison de ces informations de bas-niveau puisse faire émerger une interprétation sémantique de plus haut niveau. De manière générale, les outils d'indexation peuvent être classés sur la base des caractéristiques utilisées dans le processus d'analyse. Dans la littérature, les caractéristiques extraites à partir de l'analyse de mouvement sont employées dans (Tovinkere & Qian, 2001) pour détecter un ensemble d'événements qui peuvent être présents dans une vidéo de matchs de football, et dans (Bonzanini et al., 2001) pour extraire un résumé et les moments importants de ce même type de vidéo. D'autres techniques sont basées sur des caractéristiques de couleur et de mouvement comme dans (Xie et al., 2002). Elles sont utilisées pour classifier une vidéo de match de football en deux phases : jeu et arrêt. La couleur et la forme sont employées dans (Avrithis et al., 2000) pour segmenter une vidéo de journaux télévisés en classes sémantiques. Une structuration des vidéos de matchs de basket-ball en classes sémantiques est proposée dans (Zhou et al., 2000) en se basant sur les règles du domaine et en combinant couleur, forme, mouvement et texture. D'autres techniques se basent sur des caractéristiques audio (Rui et al., 2000) et visent à résumer des vidéos de base-ball, ou à classifier les événements acoustiques rencontrés lors de matchs de football en classes sémantiques (Lefevre et al., 2002). Certains utilisent des caractéristiques multimodales (Eickeler & Muller, 1999) pour indexer les vidéos de journaux télévisés en détectant six classes sémantiques (Han et al., 2002) ou pour résumer automatiquement des vidéos de base-ball et des programmes de TV de Formule 1 (Petrovic et al., 2002).

Ces différentes techniques présentent plusieurs limitations notamment, l'utilisation de connaissances a priori sur le type d'analyse à effectuer et sur celui du document à traiter. Ceci suppose que l'on dispose en amont de l'indexation, d'une information précise sur ce que l'on va rechercher, sur le type de vidéo analysé ou sur les règles qui existent dans le domaine auquel appartient la vidéo (règles de jeu utilisées dans un match de football, règles utilisées dans la phase de production du type de document vidéo...). Dans ce cas, la portée de ces techniques est limitée à un contenu spécifique. Elles ne peuvent donc pas être employées telles quelles pour analyser de nouveaux types de contenu ni pour rechercher des événements qui ne sont pas prédéfinis. Des efforts de généralisation de ces techniques de détection d'événements ont été faits. Par exemple, dans (Duan et al., 2002), la généralisation a été faite à l'échelle d' « enregistrements d'événements sportifs » mais ceci reste malgré tout limité à un domaine spécifique (le domaine sportif).

Notre objectif est donc de se baser sur plusieurs systèmes de segmentation fournissant des informations sémantiques de plus ou moins bas niveau relatives à l'évolution du contenu d'un document, de les combiner afin d'observer les relations temporelles qui peuvent exister entre les événements ainsi identifiés. Ces observations permettront de déduire d'autres informations plus précises décrivant la structure temporelle du document. Notre approche est différente des techniques mentionnées ci-dessus car les structures que nous recherchons ne sont pas prédéfinies et aucune information

préalable n'est employée. Nous nous focalisons seulement sur l'étude des segments temporels que peuvent fournir différents systèmes de segmentation appliqués à différents médias (image, audio).

Cet article est organisé comme suit : dans la deuxième section, nous présentons différentes techniques de manipulation des relations temporelles et nous introduisons un mode de représentation graphique de celles-ci. Nous montrons ensuite comment cette représentation peut être employée pour identifier de manière générale des relations significatives entre événements, ce qui nous conduit à aborder les problèmes de quantification et de classification des relations. Enfin, à la fin de cette même section, nous illustrons notre approche en l'appliquant aux relations de Allen. Dans la troisième section, nous donnons quelques résultats des travaux expérimentaux menés jusqu'à présent sur un ensemble de quatre segmentations temporelles d'un journal télévisé. Enfin, dans la quatrième et dernière section, nous concluons et présentons nos travaux futurs dans ce domaine.

2 Information temporelle

2.1 Vue d'ensemble

L'analyse des relations temporelles entre les événements présents dans un même document audiovisuel est une question importante et ce pour différentes raisons. Les résultats d'une telle analyse peuvent être employés pour comparer le contenu d'un document donné à une structure temporelle prédéfinie (avec des HMM hiérarchiques par exemple) afin d'identifier des moments clés spécifiques, ou d'établir automatiquement une représentation temporelle de l'évolution du contenu. La situation actuelle démontre que de tels outils sont toujours construits sur la connaissance a priori de la façon dont les événements sont temporellement reliés entre eux dans les documents audiovisuels. Par exemple, nous pouvons employer le fait que dans un journal télévisé, la présence du présentateur alterne avec les reportages, ou que dans une émission de variété, la performance d'un artiste est suivie d'applaudissements présents sur la bande sonore.
Pour étendre l'analyse temporelle du contenu audio ou visuel d'un document aux relations susceptibles d'exister entre n'importe quel genre d'événements, qu'elles soient flagrantes ou subtiles et imprévisibles, il est nécessaire de disposer d'outils de représentation et de raisonnement temporels.

Hayes a introduit six notions différentes pour représenter des relations temporelles (Hayes, 1995) à savoir : la dimension physique de base, la « time-line », les intervalles de temps, les points temporels, la quantité de temps ou durée et les positions de temps. Ce problème a également été abordé par plusieurs chercheurs. Nous pouvons trouver un état de l'art des différentes approches de représentation et de raisonnement

temporels en se référant à Chittaro et Montanari ((Chittaro & Montanari, 1996) et (Chittaro & Montanari, 2000)), Vila (Vila, 1994), et Pani (Pani, 2001).

Les modèles existants et permettant d'exprimer les relations temporelles peuvent être divisés en deux catégories : les modèles basés sur la notion de point temporel (Vilain & Kautz, 1986) et ceux basés sur la notion d'intervalle (Allen, 1983).

Dans le premier type de modèles, les points sont des unités élémentaires répartis le long de l'axe du temps. Chaque événement est associé à un point temporel. Soient deux événements **e1** et **e2**, trois relations temporelles peuvent être déterminées entre eux. Un événement peut être *avant* (<), *après* (>) ou *simultané* (=) à un deuxième événement. Ces relations sont des relations basées sur la notion de point temporel. Un exemple de représentation de ce type est la « time-line », sur laquelle les objets sont placés sur plusieurs axes de temps. Cette représentation a par la suite été également employée comme représentation basée sur la notion d'intervalle. Nous pouvons trouver le modèle de « time-line » utilisé dans diverses applications telles que HyTime (HyTime, 1992).

Les modèles basés sur la notion d'intervalle considèrent les entités élémentaires comme des intervalles de temps qui peuvent être ordonnés selon différentes relations. Les modèles existants sont principalement basés sur les relations définies par Allen dans (Allen, 1983).

2.2 Représentation et raisonnement temporels

Considérons deux segmentations temporelles d'un même document vidéo, **S1** et **S2**, effectuées pour procéder à l'analyse du document. Cette première étape produit des segments qui peuvent être vus comme des intervalles temporels localisant chacun une partie de la vidéo dans laquelle un événement spécifique a été détecté. Chaque système de segmentation produit une séquence d'intervalles où un seul type d'événement se produit (effets de transition progressifs, présence d'un personnage à l'image, présence de musique sur la bande sonore, etc..). Ainsi, on peut considérer le résultat d'une segmentation donnée, comme un ensemble de segments temporellement disjoints. Les segmentations S1 et S2 comportant respectivement N et M segments successifs seront définies par : $S1 = \{s_{1i}\}_{i\in[1,N]}$ et $S2 = \{s_{2j}\}_{j\in[1,M]}$

Un intervalle temporel est caractérisé par deux points correspondant à ses extrémités. Soit s_{1i} et s_{2j} deux segments issus respectivement de la segmentation S1 et de la segmentation S2. Chacun de ces segments est caractérisé par son point de début (d) et son point de fin (f), soit respectivement $[s_{1id}, s_{1if}]$ et $[s_{2jd}, s_{2jf}]$. Nous pouvons représenter la relation temporelle entre ces segments à l'aide de trois variables, comme proposé dans (Moulin, 1992) :

1) **Lap** = $s_{2jd} - s_{1if}$
2) **DB** = $s_{1id} - s_{2jd}$
3) **DE** = $s_{2jf} - s_{1if}$

Ainsi, une relation temporelle entre deux segments peut être représentée dans un espace à trois dimensions. Un point dans cet espace détermine une relation entre deux intervalles. Pour deux segmentations **S1** et **S2**, les trois paramètres caractérisant chaque couple de segments (s_{1i}, s_{2j}) peuvent être évalués et représentés par un point dans l'espace 3D.
A chaque point de l'espace 3D (i.e. pour chaque relation temporelle potentielle entre deux segments), nous associons un accumulateur qui comptabilise les votes c'est-à-dire qui compte le nombre de fois où la relation associée est observée entre deux segmentations. Ces accumulateurs sont regroupés en une matrice appelée par la suite Matrice des Relations Temporelles (MRT). Cette matrice peut être employée directement pour déterminer les fréquences des relations potentielles ainsi que pour observer les distributions des votes. La mise en évidence de distributions remarquables, permettra d'identifier une règle générale caractérisant le comportement temporel des événements segmentés.
La taille de la MRT est directement liée à la durée du document et peut par conséquent être très grande. Ainsi, le premier problème à surmonter est la quantification de l'espace 3D afin de produire une matrice de taille acceptable. Une fois que la matrice est créée, initialisée et remplie (avec les différents votes), l'étape de quantification peut être exécutée.

2.2.1 Quantification

L'étape de quantification de la matrice dépend de l'échelle des caractéristiques de bas niveau. En ce qui concerne le contenu visuel d'un document, l'extraction de caractéristiques de bas niveau peut associer une valeur à chaque image ou à des fenêtres d'une durée de 1 seconde. Dans le cas de l'analyse audio, les résultats produits ne correspondent pas nécessairement aux mêmes unités de temps, la durée des segments sur lesquels les caractéristiques sont déterminées étant relativement variable. Par conséquent, nous devons dans chaque cas, définir des intervalles basés sur la plus grande échelle temporelle employée pour exprimer les caractéristiques.
D'une façon plus générale, la quantification de cet espace mène aux mêmes problèmes que ceux généralement identifiés dans des méthodes de vote. En particulier, la taille de la matrice doit être limitée. Comme la variation maximale des paramètres est inférieure ou égale à la différence entre le début du premier intervalle et la fin du second nous pouvons a priori identifier les frontières de cet espace. En outre, dans le cas d'un espace de dimension élevée, nous pouvons directement appliquer un processus hiérarchique de discrétisation au lieu de procéder à une quantification brute et ainsi se concentrer progressivement sur les sous parties de l'espace qui reçoivent plus de voix que les autres (Li & Lavin, 1986).

2.2.2 Classification

Une fois que l'étape de vote a été exécutée, c'est-à-dire lorsque tous les couples possibles de segments ont été traités et que les votes ont été effectués, la MRT doit être analysée pour identifier par exemple les relations les plus fréquentes entre les

caractéristiques prises en compte. À la différence d'autres techniques de vote, s'intéresser uniquement à la valeur maximale n'est pas suffisant pour identifier entièrement une relation. En effet, la plupart des relations sémantiques temporelles déterminent des sous espaces de la MRT dans lesquels les votes sont distribués. Ainsi, la première étape de l'analyse de la MRT est de localiser ces zones. Cette localisation peut être réalisée par des méthodes de clustering ou des outils de séparation de classes.

Une autre approche consiste à définir a priori les relations sémantiques à observer, comme par exemple en prenant les relations de Allen. Cette approche consiste alors à identifier les sous-parties disjointes de l'espace de vote qui peuvent être associées aux relations remarquables comme cela est illustré dans l'exemple donné en section 2.2.3.

Ensuite, le nombre d'occurrence de chaque relation **R** observée entre deux caractéristiques est calculé en effectuant la somme des votes contenus dans la sous partie associée.

2.2.3 Exemple

Allen a proposé un ensemble complet de relations temporelles pouvant exister entre deux intervalles. Ainsi, pour deux intervalles donnés, il définit treize possibilités distinctes de relier temporellement ces segments. Dans le tableau 1 les douze premières lignes représentent les six relations directes et leur relation inverse respective et la dernière ligne correspond au cas où les deux segments débutent et se terminent en même temps. Puisqu'un intervalle est défini par deux points (son début et sa fin) on peut ramener un modèle basé sur des relations entre intervalles à un modèle basé sur les points en considérant une relation entre intervalles comme une conjonction de relations entre les points correspondants aux extrémités des segments observés ($[s_{1id}, s_{1if}]$ et $[s_{2jd}, s_{2jf}]$ cf . Tableau 1).
Dans la relation '*before* (<) ainsi que pour son inverse, (>) nous ajoutons une contrainte, nommée 'α', pour limiter la détection de ces deux types de relation à deux intervalles dont la comparaison reste significative et porteuse d'informations pertinentes à propos de la structure du contenu.

Relation	**Symbole et *Inverse***	***Notation Point***	**Exemple**
s_{1i} **before (α)** s_{2j}	**<** **>**	$s_{1id}<s_{1if}<s_{2jd}<s_{2jf}$ **&** $(0<s_{2jd} - s_{1if} \leq \alpha)$	**AAA** $\Leftrightarrow$ **BBBB** $d = s_{2jd} - s_{1if} \leq \alpha$
s_{1i} **meets** s_{2j}	**m** ***mi***	$s_{1id}<s_{1if}=s_{2jd}<s_{2jf}$	**AAAAA** **BBBB**
s_{1i} **overlaps** s_{2j}	**o** ***oi***	$s_{1id}<s_{2jd}<s_{1if}<s_{2jf}$	**AAAAA** **BBBBB**
s_{1i} **starts** s_{2j}	**s** ***si***	$s_{1id}=s_{2jd}<s_{1if}<s_{2jf}$	**AAAA** **BBBBBBBB**
s_{1i} **finishes** s_{2j}	**f** ***fi***	$s_{2jd}<s_{1id}<s_{1if}=s_{2jf}$	**AAA** **BBBBBBB**

Exploitation des relations temporelles

s_{1i} **equals** s_{2j}	= =	$s_{1id}=s_{2jd}<s_{1if}=s_{2jf}$	**AAAAAA** **BBBBBB**
s_{1i} **during** s_{2j}	**d** ***di***	$s_{2jd}<s_{1id}<s_{1if}<s_{2jf}$	**AAAAA** **BBBBBBBBB**

Tableau 1

Si nous représentons chaque relation temporelle entre deux intervalles à l'aide des trois paramètres **Lap**, **DE**, **DB** définis plus haut, et que nous appliquons cette représentation aux relations de Allen, nous constatons que celles-ci définissent des contraintes entre ces paramètres comme nous le montrons dans le Tableau 2.

Par exemple, la relation '*during'* correspond à la définition suivante:

$$\mathbf{s_{2jd} < s_{1id} < s_{1if} < s_{2jf}}$$

D'où on peut déduire que :

$s_{2jd} - s_{1if} < s_{1id} - s_{1if} < s_{1if} - s_{1if} < s_{2jf} - s_{1if}$ => **Lap** < 0 < **DE**.

$s_{2jd} - s_{2jd} < s_{1id} - s_{2jd} < s_{1if} - s_{2jd} < s_{2jf} - s_{2jd}$ => 0 < **DB** < **-Lap**

Pour la relation *'meet inverse'* (mi), nous avons les contraintes suivantes:

$$\mathbf{s_{2jd} < s_{2jf} = s_{1id} < s_{1if}}$$

$s_{2jd} - s_{2jd} < s_{1id} - s_{2jd} = s_{2jf} - s_{2jd} < s_{1if} - s_{2jd}$ => 0 < **DB**.

$s_{2jd} - s_{1if} < s_{2jf} - s_{1if} < s_{1if} - s_{1if}$ => **DE** < 0.

$s_{2jd} - s_{1if} < s_{2jd} - s_{1if} = s_{1id} - s_{1if} < 0$ => **Lap** < 0

DE – DB = $s_{2jf} - s_{1if} - s_{1id} + s_{2jd} = (s_{2jf} - s_{1id}) + (s_{2jd} - s_{1if})$ = **Lap**,
$s_{2jf} - s_{1id} = 0$ parce que s_{2j} 'meet' s_{1i}.

De la même manière, nous pouvons déduire des contraintes similaires pour chacune des autres relations de Allen. Ces contraintes définissent des sous-espaces dans la MRT localisant les votes de chaque relation.

Relation	***Lap***	***DB***	***DE***
<	**0 < Lap ≤ α**	**DB< -Lap**	**DE > Lap**
m	**Lap = 0**	**DB < 0**	**DE > 0**
o	**Lap < 0**	**DB < 0**	**DE > 0**
s	**Lap < 0**	**DB = 0**	**DE > 0**
f	**Lap < 0**	**DB > 0**	**DE = 0**
=	**Lap < 0**	**DB = 0**	**DE = 0**
d	**Lap < 0**	**DB < -Lap**	**DE > 0**
>	**DE-DB < Lap < 0 et 0<DB –DE+Lap ≤ α**	**DB > 0**	**DE < 0**
mi	**Lap =DE-DB**	**DB > 0**	**DE < 0**
oi	**Lap < DE-DB <0**	**DB > 0**	**DE < 0**
si	**Lap < DE**	**DB = 0**	**DE < 0**

fi	**Lap < 0**	**DB < 0**	**DE = 0**
di	**Lap < DE**	**DB < 0**	**DE < 0**

Tableau 2 : Contraintes sur les paramètres caractérisant les relations de Allen

Une quantification a priori de l'espace utilisant les relations de Allen est définie par les règles données dans le Tableau 2.

Pour simplifier la représentation dans l'espace 3D, nous établissons la correspondance suivante : DE=x, DB=y, et Lap = z ;

Considérons la région définie par la relation '*before'*. Les contraintes définies dans le tableau 2 indiquent qu'une relation établie entre deux intervalles s_{1i} et s_{2j} sera la relation '*before'* si et seulement si :

1) $0 < z \leq \alpha$
2) $y < -z$
3) $x > z$

Ainsi, les votes qui correspondent à la relation '*before'* seront cumulés dans la zone délimitée par les contraintes mentionnées ci-dessus et représentée sur la figure 1.

La relation '*meet*' (Fig. 2) correspond à une zone restreinte au plan défini par l'équation: z = 0. Plus précisément, la relation '*meet'* peut être établie entre s_{1i} et s_{2j} si et seulement si on vérifie :

1) $z = 0$
2) $y < 0$
3) $x > 0$

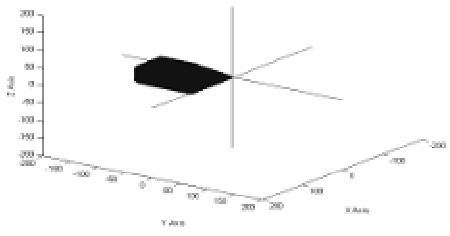

Fig. 1 - La relation *before*

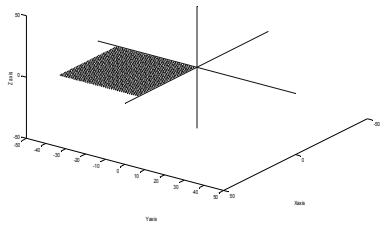

Fig. 2 - La relation *meet*

On peut également considérer que la relation '*equal'* peut être établie entre s_{1i} et s_{2j} ssi on vérifie :

1) $z < 0$
2) $y = 0$
3) $x = 0$

ce qui correspond à la demi-droite représentée dans la figure 3.

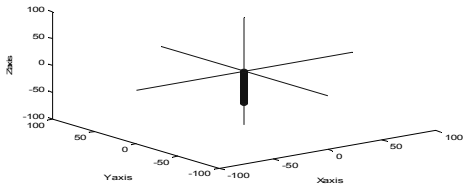

Fig. 3 - La relation *equal*

Suivant le même principe, nous pouvons identifier les régions de la MRT correspondant à toutes les autres relations.

Avec cette représentation, nous pouvons ainsi établir un lien entre une observation donnée entre deux intervalles et une relation. En d'autres termes, les paramètres définissant l'observation d'une relation et vérifiant les équations définissant une zone donnée permettront d'identifier la relation sémantique correspondante.

2.3 Gestion d'erreur

Une segmentation temporelle, si elle est obtenue automatiquement, peut être imprécise. Plusieurs sources d'imprécision existent. Des erreurs peuvent être dues aux performances des outils d'extraction des caractéristiques de bas niveau utilisés pour traiter le flot de données audiovisuelles. Ceci peut conduire à voter pour une relation qui en réalité n'est pas la relation qui devrait être observée mais une relation voisine. L'effet de bord ainsi introduit par l'imprécision de certains outils utilisés pour la segmentation, doit être pris en compte. On ne peut pas être sûr à cent pour cent de la légitimité du vote enregistré. Une approche possible pour prendre en compte cette incertitude consisterait à considérer une distribution floue des votes sur les relations voisines.
Dans le cas des relations de Allen et comme pour la plupart des approches dans ce domaine, nous pouvons employer le principe des relations voisines de Freska (Freska, 1992). Soient A et B, deux événements vérifiant la relation temporelle '*meet'*. En déplaçant ou en déformant légèrement les segments, nous pouvons changer la relation en *'before'* ou *'overlap'*. Par conséquent, les relations *'before'* et *'overlap'* sont considérées comme relations conceptuellement voisines de la relation *'meet'*. Au contraire, la relation *'equal'*, par exemple, n'est pas une voisine conceptuelle de `*meet'* , car elle ne peut pas être obtenue directement à partir *de 'meet'* par

déformation ou translation temporelle des intervalles. Dans notre cas, chaque relation est représentée par une zone et ses voisines sont les zones adjacentes.

Une extension topologique du principe de voisinage défini par Freska à n'importe quelle relation peut ici être formellement définie par :

Soit **Ri** (Xi, Yi, Zi) et **Rj** (Xj, Yj, Zj) deux zones compactes et soit $\mathbf{R_k}$ = Ri ∩ Rj
ou X_k = (Xi ∩ Xj), Y_k = (Yi ∩ Yj), et Z_k = (Zi ∩ Zj).
Ri a **Rj** comme voisine directe si un seul des paramètres de **Rk** est vide.

Par exemple, si Ri = ***meet*** . alors Xi correspond à **DE** > 0, Yi à **DB** < 0, et Zi à **Lap** = 0.

Ri (Xi, Yi, Zi) = **meet (**]0 +∞[**,**]-∞ 0[, {0}**)**

Si Rj = ***overlap*** alors Xj correspond à **DE** > 0, Yj à **DB** < 0, et Zj à **Lap** < 0.
Rj (Xj, Yj, Zj) = **O (**]0 +∞[,]-∞ 0[,]-∞ 0[**)**

Nous avons alors (Xi ∩ Xj) =]0 +∞[, (Yi ∩ Yj) =]-∞ 0[, et (Zi ∩ Zj) = Ø. Comme on a seulement (Zi ∩ Zj) = Ø, alors ***meet*** et ***overlap*** sont des voisines directes.

Ce n'est pas le cas pour les relations ***Meet*** et ***During*** pour lesquelles deux paramètres sont vide : Y_k = Ø, et Z_k = Ø.

Ces liens de voisinages entre relations sont illustrés dans la figure 4 où sont représentées les zones correspondant aux relations 'meet' (gris intermédiaire), 'overlap' (gris clair), et 'start' (gris foncé).

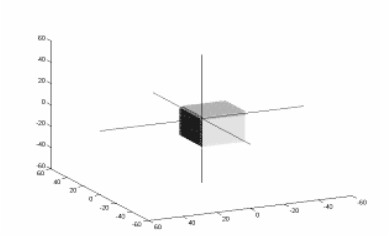

Fig. 4 : liens de voisinage entre relations Meet, Overlap et start

3 Expérimentation

Nous avons calculé plusieurs matrices de type MRT sur des segmentations temporelles effectuées sur des vidéos de journaux télévisés. Cinq segmentations différentes ont été réalisées sur un même document. Elles sont relatives à la présence

ou non du présentateur à l'écran et, du point de vue sonore, à la présence ou non de parole, de silence, de musique et d'applaudissements. Nous avons construit quatre MRT pour observer les relations temporelles entre la segmentation visuelle (présentateur) et chacune des segmentations sonores (parole, silence, musique, applaudissements). A la différence de l'exemple donné dans la section 3, nous n'avons pas de connaissance a priori sur les relations potentiellement observables. Seule l'étude des résultats obtenus par le calcul des MRT doit nous permettre d'observer des régularités et d'en déduire la présence de relations pertinentes entre segments.

En observant la première MRT dont la représentation est données figure 5, nous constatons que les points représentant les relations entre le présentateur et les segments de la parole sont distribués le long de lignes parallèles, toutes inclues dans le plan d'équation $z = ay + b$; pour x arbitraire. Les cercles représentent les points exclus après l'étape de quantification et donc considérés comme non pertinents (i.e. quand $Lap > \alpha$). Cette distribution peut être également observée dans la figure 9.

Après la projection de cette MRT sur l'axe XOZ (Fig.6), nous observons également que la plupart des points se retrouvent sur une même ligne passant par le centre de l'axe XOZ.

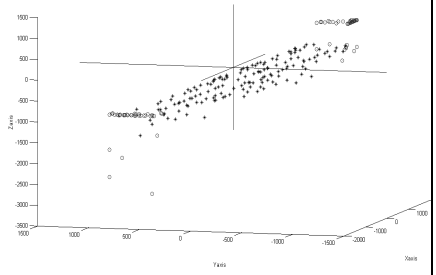

Fig. 5 : Relation entre présentateur et parole

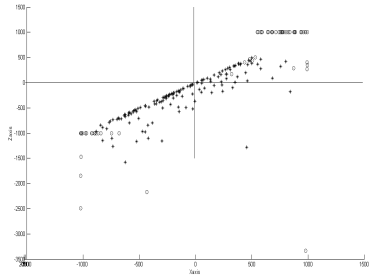

Fig. 6 : Projection sur l'axe XOZ des points de la figure 5

Dans la figure 7 nous avons représenté les relations qui ont pu être observées entre le présentateur et les segments musicaux. Dans la figure 8 sont représentées celles observées entre le présentateur et les applaudissements. Nous pouvons remarquer que des relations entre les événements « présence du présentateur » et « segments musicaux » existent, ce qui n'est quasiment pas le cas pour les relations entre « présentateur » et « applaudissements ». Les applaudissements détectés sont en fait tous des erreurs occasionnelles de l'outil de segmentation concerné. Bien que lorsque le présentateur parle, la présence de musique soit très rare, il arrive que le journal soit

entrecoupé de publicités souvent accompagnées de séquences musicales (jingles par exemple). En ce qui concerne la relation entre « présentateur » et « silence » (fig. 9), le nombre de points augmente et la projection des points sur l'axe XOZ (fig. 10) est également une droite qui passe par le centre. Une transformation de Hough pourrait être employée pour identifier cette droite ou celles qui apparaissent sur la figure 7. Ce type d'analyse fournit des informations plus précises sur le contenu puisque nous avons des relations distribuées suivant des types de zones larges et clairement identifiables (droites, plans, …) qui ne correspondent pas aux régions identifiées par les relations de Allen.

Dans la figure 10, nous pouvons observer que les points projetés sont distribués sur le plan où $x \approx z$. Cela signifie qu'effectivement, DE ≈ LAP, et donc que $s_{2jd} - s_{1if} \approx s_{2jd} - s_{1if}$. Nous en déduisons que $s_{2jf} \approx s_{2jd}$ ce qui signifie que les segments de la seconde caractéristique sont très courts. Celle-ci correspond aux segments de silence présents dans un journal télévisé, qui sont effectivement très courts dans ce type de document.

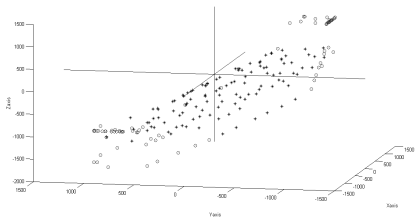

Fig. 7 : Relations entre présentateur et musique

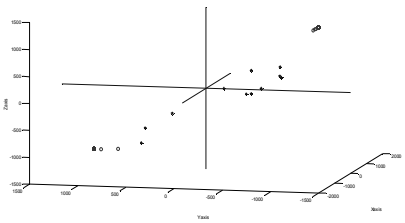

Fig. 8 : Relations entre présentateur et applaudissement

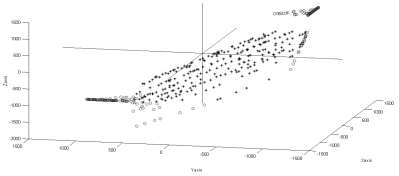

Fig. 9 : Relations entre présentateur et silence

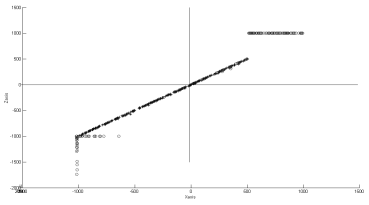

Fig.10 : Projection sur l'axe XOZ des points de la figure 9

4 Conclusion et travaux futurs

Nous avons présentés dans cet article une nouvelle technique pour mettre en évidence les relations temporelles significatives entre des événements issus de la segmentation

produite par des outils automatiques. Les relations temporelles entre deux intervalles appartenant à deux segmentations différentes d'un même document sont représentées par des points dans un espace à trois dimensions. L'espace des observations peut être discrétisé en différentes zones, chacune d'elles pouvant représenter une relation sémantique. Cette discrétisation peut être effectuée en ayant recours à une méthode de classification traditionnelle ou bien en partant de relations temporelles déjà connues comme nous l'avons fait à titre d'exemple en utilisant les relations temporelles de Allen. Nous avons ensuite présenté les premiers résultats d'une expérimentation effectuée sur des documents vidéos, notamment des journaux télévisés, ce qui nous a permis d'observer la distribution des points dans l'espace de représentation.

Plusieurs perspectives peuvent être envisagées pour la poursuite de ce travail. Un de nos objectifs est d'aborder le problème lié aux performances plus ou moins bonnes des outils de segmentation utilisés, ce qui peut influer sur la fiabilité des observations réalisées. Comme mentionné précédemment, les votes enregistrés dans la MRT peuvent être distribués dans le voisinage d'une relation, la notion de voisinage étant définie par exemple suivant le graphe de voisinage de Freska appliqué aux relations de Allen. La détermination des liens de voisinage entre relations dans l'espace de la MRT peut être réalisé en utilisant la distance entre les zones de l'espace à trois dimensions les plus proches. Un arbre de voisinage peut être automatiquement construit reliant chaque sous-espace de la MRT. Les poids associés à un vote peuvent être distribués d'une manière plus précise tout en utilisant la topologie de la MRT au lieu d'employer un arbre de voisinage. Le calcul des poids peuvent dépendre de différents critères : de la position du point dans une zone (c.-à-d. de la distance entre ce point et le centre de la zone), de la distance qui la sépare à d'autres zones, de la taille des zones, du nombre de points inclus…

Nous avons également l'intention d'explorer la conjonction des relations observées d'une manière hiérarchique. Pouvoir manipuler une conjonction d'un grand ensemble de relations devrait nous permettre d'identifier des événements temporels complexes.

Enfin, notre objectif à plus long terme est d'explorer l'utilisation de la MRT comme modèle qui puisse caractériser l'évolution temporelle de différents types de documents.

Références

Allen J. F. (1983). Maintaining Knowledge about Temporal Intervals (Tome 26 (11)). Communication of the ACM. p. 832 – 843.

Avrithis Y., Tsapatsoulis N. & Kollias S.(2000). Broadcast News Parsing Using Visual Cues: A Robust Face Detection Approach. In IEEE International Conference on Multimedia and Expo. New York City, USA.

Bonzanini A., Leonardi R., & Migliorati P. (2001). Exploitation of Temporal Dependencies of Descriptors to Extract Semantic information. International Workshop on Very Low Bitrate Video Coding. Athen, Greece.

Chittaro L. & Montanari A. (1996). Trends in Temporal Representation and Reasoning (Tome 11(3)). The Knowledge Engineering Review. p. 281-288.

Chittaro L. & Montanari A. (2000). Temporal Representation and Reasoning in Artificial Intelligence: Issues and Approaches (Tome 28). Annals of Mathematics and Artificial Intelligence. p. 47-106.

Duan L., Xu M., Xiao-Dong Yu, & Qi Tian (2002). A unified framework for semantic shot classification in sports videos. In Proc. of the tenth ACM international conference on Multimedia. p. 219-220. Juan-les-Pins, France.

Eickeler S. & Muller S. (1999). Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models (Tome 6). In Proc. IEEE ICASSP. P. 2997-3000. Phoenix, USA.

Freska C. (1992). Temporal Reasoning Based on Semi-intervals (Tome 54). Artificial Intelligence. p.199-227.

Han M., Hua W., Xu W., & Gong Y. (2002). An integrated baseball digest system using maximum entropy method. In Proc. ACM Multimedia 2002. p. 347-350. Juan Les Pins, France.

Hayes J. Patrick. (1995). A Catalog of temporal theories. Technical report UIUC-BI-AI- 96-01, University of Illinois.

HyTime (1992) Information Technology, "Hypermedia / Time-based Structuring Language (HyTime)", ISO/IEC 10743.

Li H. & Lavin M. A. (1986). Fast Hough Transform: A Hierarchical Approach (Tome 36). Journal on Graphical Models and Image Processing (CVGIP). p.139-161.

Lefevre S., Maillard B., & Vincent N. (2002). 3 classes segmentation for analysis of football audio sequences. In Proc. ICDSP'2002. Santorin, Greece.

Moulin B. (1992). Conceptual graph approach for the representation of temporal information in discourse (Tome 5 (3)). Knowledge based systems. p 183 –192.

Pani A. K. (2001). Temporal representation and reasoning in artificial intelligence: A review. Mathematical and Computer Modelling. p. 55–80.

Petrovic M., Mihajlovic V., Jonker W., & Djordievic-Kajan S. (2002). Multi-modal extraction of highlights from tv formula 1 programs. In Proc. ICME'2002. Lausanne, Switzerland.

Rui Y., Gupta A., & Acero A. (2002). Automatically extracting highlights for TV baseball programs. In Proc. of the eight ACM international conf. on Mult. p. 105–115. California, USA.

Tovinkere V., Qian R. J. (2001). Detecting Semantic Events in Soccer Games: Toward a Complete Solution. In Proc. ICME'2001. p. 1040-1043. Tokyo, Japan.

Vila L. (1994). A Survey on Temporal Reasoning in Artificial Intelligence (Tome 7(1)). Artificial Intelligence Communications. p. 4 -28.

Vilain M., & Kautz H. A. (1986). Constraint propagation algorithms for temporal reasoning. In AAAI-86. p. 132-144.

Xie L., Chang S-F., Divakaran A., and Sun H.(2002). Structure analysis of soccer video with Hidden Markov Models. In Proc. International Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, USA.

Zhou W., Vellaikal A., & Kuo C.-C. J.(2000). Rule-based Video Classification System for Basketball Video Indexing. In Proc. of the 2000 ACM Mult. Workshops. p. 213-216. Los angelos, USA.