

Une Approche Conceptuelle pour la Modélisation et la Structuration Sémantique des Documents Vidéos

Mbarek Charhad^{*}, Mounir Zrigui^{**} et Georges Quénot^{*}

^{*} Laboratoire CLIPS-IMAG, BP 53-38041 Grenoble cedex9

Mbarek.Charhad@imag.fr

Georges.Quenot@imag.fr

^{**} Laboratoire RIADI, Unité Monastir Faculté des Sciences de Monastir, Tunisie

mounir.zrigui@fsm.rnu.tn

Résumé: L'accroissement rapide de la masse de documents vidéos numériques, aussi bien sur Internet que dans les archives et les bases personnelles, nécessite la mise en place d'outils et de méthodes capables de décrire et de modéliser le contenu, afin de faciliter l'accès, la consultation et aussi la navigation dans ces documents. Nous présentons dans cet article, une approche de modélisation conceptuelle du contenu sémantique. Celle-ci se base sur une description par des concepts et des relations conceptuelles. Les concepts sont rassemblés en des classes de sorte qu'on puisse pour chaque classe énumérer la liste des instances correspondante. Des relations sémantiques telle que les relations spatiales sont utilisées pour spécifier des emplacements, des relations temporelles couvrent le critère dynamique de la vidéo et d'autres relations faisant référence à des activités (s'adresser, marcher, rencontrer, etc....) permettent de décrire ce qui se passe.

Les concepts et les relations conceptuelles sont représentés dans un schéma de modélisation en utilisant le formalisme des graphes conceptuels. Ce schéma permet de conserver la structure hiérarchique du document vidéo du fait que les descriptions sont structurées de façon à respecter la chronologie et l'ordre de passage dans le document vidéo.

Nous avons appliqué ce schéma de modélisation sur le corpus TREC2003 (120 heures de journaux télévisés). Notre objectif est d'intégrer ce schéma de structuration sémantique dans le cadre d'un système d'indexation et de recherche vidéo par le contenu afin d'améliorer sa performance en termes de rappel et de précision.

Mots clés : document vidéo, graphe conceptuel, modélisation, contenu sémantique, annotation, base de connaissances, ontologie.

1 INTRODUCTION

Le volume considérable de données vidéos actuellement disponibles, associé à la généralisation de leur utilisation pour de nombreuses applications, représente actuellement l'enjeu principal des études sur le traitement de l'information au niveau représentation, indexation et aussi au niveau manipulation. La masse importante de ces données (que ce soit sur Internet soit dans des bases personnelles) augmente la difficulté de leur accès : d'où la nécessité de développer des nouvelles méthodes plus efficaces pour la classification, l'accès et l'indexation par le contenu.

Plusieurs approches d'analyse et de modélisation de la vidéo ont été mises en place et décrites dans la littérature. Ces approches présentent des inconvénients liés d'une part, à la pauvreté de la description sémantique du document vidéo et d'autre part à

l'absence d'un modèle de description générique qui soit indépendant du type de document vidéo (film, journal télévisé, documentaire, etc.).

Parmi ces approches, nous trouvons celles basées sur l'analyse dite « bas niveau » du contenu vidéo sur les pistes visuelles et/ou auditives. Cette analyse offre généralement des techniques de segmentation fournissant des informations sur les aspects compositionnels et syntaxiques (Celentano & Gaggi, 2002) des contenus. Elle apporte aussi des progrès en matière d'extraction des descripteurs visuels du contenu et facilite par conséquent l'accès aux unités d'informations dans les contenus et leur manipulation.

Des normes telles que Dublin Core, Mpeg7, XML Schéma, etc., ont été mises en place pour représenter de façon formelle et cohérente les descriptions qui fournissent des méta-données (titre, auteur, format, etc.). Elles visent aussi à fusionner les aspects conceptuels (haut niveau) des descriptions des

contenus et les caractéristiques numériques (bas niveau) pour permettre le traitement des informations audio-visuelles à plusieurs niveaux.

Les progrès dans le domaine de l'analyse bas niveau et les normes de description constituent des outils fondamentaux pour une description des contenus de la vidéo. Il reste alors de préciser à quel degré ces différentes approches peuvent satisfaire les besoins attendus par un système de recherche d'information vidéo par le contenu sémantique. Notons que d'une manière générale, les utilisateurs des systèmes de recherche d'information par le contenu préfèrent utiliser des requêtes en langage naturel. En se basant sur des approches comme celles que nous venons de décrire, il est presque impossible de répondre avec précision à ce genre de requêtes.

Pour remédier à ces différents problèmes, nous proposons dans cet article une approche de modélisation vidéo pour la description du contenu sémantique. Cette approche se base sur une description conceptuelle du contenu. Elle permet de réunir les éléments d'informations issus de divers sous-média (image, son, texte). Le critère générique de cette approche se situe surtout dans la manière dont les informations contenues dans le document vidéo seront représentées. Prenons par exemple l'élément d'information « personne ». Une personne peut être vue, (apparaissant à l'écran) ou entendue (faisant un discours); une information textuelle indiquant son identité peut encore s'afficher en bas de l'écran. La combinaison de deux ou plus de ces possibilités de description est aussi envisageable.

Dans cette approche de modélisation, un élément d'information correspond à un concept (information sémantique perceptible par l'homme sur une entité concrète). Il s'agit de modéliser le contenu vidéo selon le point de vue d'un utilisateur. Une telle modélisation facilite l'accès à différentes parties du document et leur manipulation. L'exploitation des relations conceptuelles nous permet de spécifier les requêtes des utilisateurs et de décrire avec précision le contenu de chaque segment vidéo. Pour représenter les différentes facettes de notre approche, nous avons choisi le formalisme des graphes conceptuels. Ce choix est guidé par les exigences liées aux contraintes de modularité, et d'expressivité sémantique. Ces contraintes sont parfaitement prises en compte dans le formalisme des graphes conceptuels.

La suite de cet article est structurée comme suit : dans la section 2, nous présentons quelques travaux de recherche sur la modélisation et la représentation du contenu des documents vidéos. Dans la section 3, nous détaillons la structure de notre approche de modélisation. Nous proposons une extension du schéma de modélisation dans la section 4. La section 5 de cet article porte sur une discussion et l'application de cette approche. Nous évoquons en conclusion quelques perspectives que nous souhaitons développer ultérieurement.

2 TRAVAUX SIMILAIRES

Dans cette section, nous décrivons un panorama d'approches proposées dans la littérature. En premier lieu, nous nous intéressons aux travaux qui portent sur la modélisation par des graphes conceptuels et la représentation symbolique du contenu. En second lieu, nous présentons d'autres approches de modélisation par le contenu des documents vidéos.

Dans la littérature, plusieurs travaux basés sur le formalisme des graphes conceptuels ont conduit à la mise en œuvre de quelques systèmes. On trouve par exemple le système EMIR² (Mechkour, 1995), le système RELIEF (Ounis, 1995), le système CoGiTant (Genest, 2000) et aussi les travaux de Michel Chein sur le formalisme de graphes conceptuels (Chein & Mugnier, 1992). Ces systèmes se distinguent par le contexte d'application et aussi par la manière d'exploiter les graphes conceptuels. En effet, certains systèmes utilisent les graphes conceptuels uniquement pour la partie représentation, d'autres au niveau indexation et interrogation.

Le système EMIR² (Mechkour, 1995) est un modèle général pour la représentation symbolique du contenu sémantique des images fixes. Ce modèle fournit un cadre dans lequel les caractéristiques de l'image peuvent être représentées. Dans ce modèle, une image est décrite par l'ensemble des vues suivantes correspondant au niveau de description logique de l'image :

La vue structurelle : l'image est vue comme un ensemble d'objets. Les objets de l'image pouvant eux-mêmes être composés d'autres objets. La vue structurelle de l'image a la forme d'un graphe orienté ayant pour noeuds les objets de l'image, et pour arcs des relations de composition structurelle. Cette vue permet de décomposer l'image en conservant la structure des éléments présents dans l'image.

La vue spatiale : cette vue est très importante dans le cas de la recherche d'images. Elle permet de décrire la forme des objets et les relations spatiales 2D (dans le plan de l'image) ou 3D (dans la scène représentée dans l'image) qui les lient ensemble. Il s'agit de positionner les objets les uns par rapport aux autres. La vue spatiale d'une image est un graphe dans lequel les noeuds sont les objets spatiaux, et les arcs sont des relations spatiales.

La vue perceptive : cette vue regroupe un certain nombre d'attributs attachés à la perception visuelle des images. Trois attributs sont pris en compte dans EMIR² : la couleur, la texture et la brillance.

La vue symbolique : la vue symbolique associe une description sémantique à l'image et aux objets de l'image. La description regroupe un ensemble de termes pouvant être reliés ensemble par des relations sémantiques.

Le système RELIEF (Ounis, 1995) est un système de recherche d'images basé sur le formalisme des

graphes conceptuels. Ce système utilise l'opération de projection pour comparer un graphe conceptuel décrivant un document à celui de la requête.

Le système CoGitant (Genest, 2000) se base sur une extension du formalisme des graphes conceptuels pour améliorer la précision du système. En effet l'objectif est de réduire au maximum l'imprécision provoquée par l'utilisation de l'opération de projection pour la phase d'interrogation du système. D. Genest dans ces travaux, propose un mécanisme pour que les documents génériques par rapport à la requête soient aussi pertinents.

D'autres approches de modélisation et de structuration ont été également développées. Des méthodes utilisant des méta-informations (titre, taille, format, etc...) traitent la vidéo dans son ensemble. Parmi les outils utilisés, on trouve des ensembles des descripteurs Dublin Core (DC) (Hunter & Armstrong, 1999). La modélisation peut être aussi plus spécifique du fait qu'elle ne s'intéresse pas uniquement aux données génériques (méta-informations). Plusieurs approches décrivent des schémas de modélisation en se basant sur des représentations du contenu à différentes échelles. Ces représentations peuvent être au niveau signal (Lee, Li & Xiong, 1999) en utilisant uniquement des descripteurs numériques (couleur, texture, luminosité, mouvement de caméra, etc.) (Etievant & Lebourgeois & Jolion, 1999) ou elles peuvent contenir des critères symboliques de haut niveau tels que par exemple des objets visuels (description symbolique, Seyrat & Durand & Faudemay, 1998).

La modélisation peut prendre une dimension plus statique. En effet, plusieurs méthodes proposées se basent sur la description du contenu vidéo sous forme d'un SGBD relationnelle ou bien orienté objets (Rowe & Boreczky & Eads, 1994), (Hollfelder & Everts & hiel, 2000).

Quelle que soit l'approche adoptée, l'exploitation de la modélisation se fait en prévoyant une manière d'interroger une base de documents vidéos pour faciliter l'accès et la consultation du contenu. Ceci n'est réalisable que par l'intégration de cette modélisation dans le cadre d'un système d'indexation et recherche qui représente l'interface entre les utilisateurs et la base des documents. Il est donc nécessaire d'évoquer les diverses possibilités d'indexer un document vidéo.

L'indexation des documents vidéos peut être distinguée dans les trois catégories suivantes :

Indexation à niveau haut : cette approche emploie un ensemble de termes (ou concepts) pour annoter des vidéos. Ces termes sont organisés en des catégories ou des classes de haut niveau comme par exemple l'action, le temps, le lieu, personnage, etc.

Indexation de bas niveau : ces techniques permettent d'accéder au contenu des vidéos par des descripteurs de bas niveau tels que la couleur, la

texture, etc. Les algorithmes d'analyse du signal doivent extraire ces descripteurs, les organiser et employer des techniques de recherche par similarité pour la recherche de séquences vidéos.

Indexation spécifique à un domaine : ces techniques emploient la structure de haut niveau de la vidéo pour contraindre l'extraction des descripteurs visuels de bas niveau. Elles sont efficaces seulement dans leur domaine prévu d'application.

3. SCHEMA DE MODÉLISATION DU CONTENU VIDÉO

L'information du contenu vidéo peut être représentée sous plusieurs niveaux : les informations *physiques* constituées des données binaires du contenu qui n'est utilisable que par l'ordinateur, et les informations de description qui permettent de transformer les informations physiques en connaissances exploitables par l'utilisateur, ce qui permet de renforcer « l'interface » entre l'homme et la machine et d'exploiter donc plus facilement le contenu vidéo. Pour fournir plusieurs niveaux d'abstraction en exploitation du contenu vidéo, nous proposons un schéma de modélisation à deux niveaux : *structure*, et *sémantique*.

La modélisation de la structure du contenu permet de décrire l'organisation qui peut représenter l'information du contenu. Cette modélisation est souvent basée sur la structure classique du document vidéo (la séquence, la scène, le plan.). Les descriptions de ce niveau sont calculées automatiquement en utilisant des descripteurs visuels. Ce niveau de modélisation est dépourvu de toute description sémantique.

La modélisation sémantique est une abstraction qui permet de lier des descriptions de bas niveau avec le monde réel. Par exemple, un ensemble d'objets de la vidéo peut correspondre à un personnage, un endroit ou une chose concrète dans le monde réel. Une relation conceptuelle entre eux peut être par exemple « le personnage X est plus âgé que le personnage Y » ou « X et Y sont des amis ». Nous proposons de créer un schéma de modélisation qui permet de décrire la signification des descriptions situées au niveau de la structure. Nous exploitons la notion des concepts et des relations conceptuelles pour représenter les occurrences (éléments d'informations) décrites dans la partie structure.

Notre approche de modélisation se situe au niveau de la description conceptuelle (Charhad & Quénot, 2004). Il s'agit de concevoir un schéma de représentation générique indépendant du contenu vidéo mais qui permet de renseigner les différents éléments d'informations se situant au niveau sémantique (figure 1).

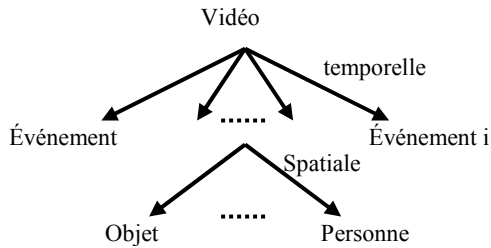


Figure 1 : structure sémantique du contenu vidéo

3.1 Modélisation Spatiale

La modélisation spatiale du contenu vidéo permet d’enrichir les interprétations reliées aux différentes séquences vidéos. S’agissant de décrire les positions spatiales (à droite, à gauche, etc.) des objets les uns par rapport aux autres, cette modélisation dépend du contenu visuel du document. Dans notre modèle, la spécification des relations spatiales ne se limite pas uniquement à décrire les positions spatiales. Nous cherchons plutôt à étendre la description du ‘localité’. Par exemple par modélisation spatiale nous pouvons décrire le lieu de déroulement d’un évènement (pays, ville, région). Ceci permet de mieux interpréter le contenu de la vidéo et d’exploiter d’autres éléments d’information (y compris les informations textuelles et auditives).

Chaque description sera représentée sous forme graphique en utilisant le formalisme des graphes conceptuels. La figure 2 illustre un exemple de description spatiale.

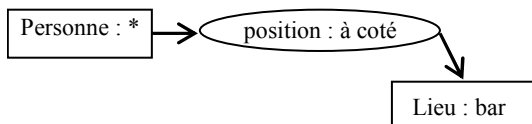


Figure 2: Modélisation spatiale avec le formalisme des graphes conceptuels (« une personne est à côté d’un bar »)

L’ensemble des relations spatiales ainsi que les descriptions sémantiques se référant au concept « localité », est classé dans un treillis de relations (figure 3).

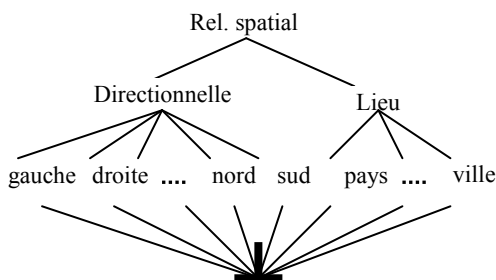


Figure 3: treillis des relations spatiales

L’utilisation des relations directionnelles requiert la mise en place d’une liste d’axiomes. Prenons deux objets obj1 et obj2 situés sur un même plan. Les lignes suivantes montrent un exemple d’axiome utilisable pour spécifier la position entre ces deux objets :

obj1 est au dessus de obj2
si et seulement si :
obj2 est au dessous de obj1.

À chaque relation bidirectionnelle, nous proposons un axiome. L’objectif est d’éviter de tout mettre dans le schéma de modélisation. On se limite à une seule description mais on garde la possibilité d’en inférer d’autres en utilisant ces axiomes.

3.2 Modélisation temporelle

La nature temporelle du document vidéo est une caractéristique spécifique qui peut être exploitée pour modéliser le contenu vidéo. Elle est fondée particulièrement sur la description des évènements. Un évènement est un ensemble d’états et d’actions perceptibles comme ayant une unité par un utilisateur. Il est nécessaire de classer et d’organiser suivant un ordre chronologique l’ensemble de ces évènements.

Dans le cas d’une description conceptuelle avec le formalisme des graphes conceptuels, la notion de temps paraît beaucoup moins importante que les autres informations. Ceci résulte de la manipulation des graphes basés sur la logique du premier ordre (Sowa, 1984).

Nous utilisons l’ensemble des relations temporelles d’Allen (Hjelsvold & Midtstraum, 1994) pour la modélisation temporelle. Ces relations exploitent implicitement la notion du temps. Elles sont calculées à partir de la prise en compte de la sémantique des liens temporels (starts, before, finish, after, meets, etc.).

Nous représentons une interprétation temporelle par deux types de Graphes Conceptuels (GCs). Le premier modélise la représentation sémantique du contenu et le second le diagramme temporel exprimant les contraintes temporelles entre les différentes situations décrites dans la séquence vidéo.

L’ensemble des relations temporelles est classé dans un treillis de relations (figure 4).

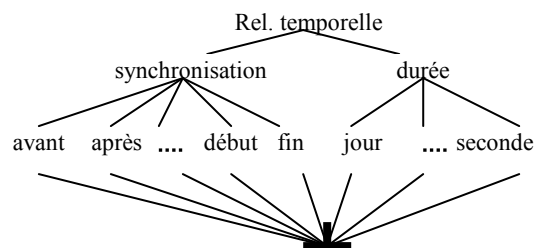


Figure 4: treillis de relations temporelles

Etant donné evt1 et evt2 deux évènements dans une séquence audiovisuelle. Pour déterminer la chronologie entre ces deux évènements, une liste d'axiomes doit être mise en place. Nous présentons dans ce qui suit un exemple :

Si
 evt2 commence avant evt1 et se termine durant evt1
 Alors
 evt1 recouvre evt2
 evt2 recouvre evt1

La description sous forme graphique d'une modélisation temporelle prend la même allure que celle de toute autre modélisation sémantique. Il n'est pas évident de décrire l'aspect dynamique du contenu vidéo par des représentations statiques en utilisant le formalisme des graphes conceptuels. Notre objectif dans la partie modélisation temporelle n'est pas de résoudre ce problème. Nous nous intéressons plutôt à la manière de prendre en compte de l'ensemble des interprétations contenant des descriptions temporelles.

Un exemple : la description temporelle présentée dans la figure suivante illustre un évènement politique qui dure 20s et qui se déroule juste après un évènement sportif.

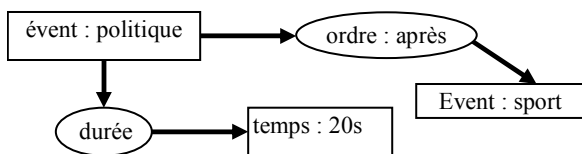


Figure 5 : représentation des relations temporelles avec GCs.

3.3 Annotation

L'annotation consiste à mettre en place une description textuelle ou numérique du contenu vidéo, quelle que soit la partie de document considérée (Assfalg & Bertini & Colombo & Del Bimbo, 2002). La description elle-même peut également prendre le nom d'annotation pour une partie de document. Si l'on parle d'annotation d'un document, alors on fait référence à l'ensemble des annotations de ses différentes parties.

Rowe & Boreczky (L. A. Rowe, J. S. Boreczky, C. A. Eads, 1994) considèrent trois types de caractéristiques : les *données bibliographiques* (titre, résumé, sujet, genre, producteur, acteurs, etc.) ; les *données de structure* (hiérarchie de plans, scènes, séquences) et les *données de contenu* qui peuvent être :

- un ensemble d'images clé du document ;
- des mots-clés extraits de la bande-son ou des sous-titres ;

- des indices d'objets qui indiquent les images d'entrée et de sortie de chaque objet ou personnage significatif.

L'enjeu principal de cette partie est de construire un vocabulaire générique permettant d'indexer les documents vidéos.

Nous avons utilisé l'outil Video-Annex (Lin, Tseng & Smith, 2004) pour annoter une partie du corpus (collection des documents vidéos du TREC 2003). La spécificité de cet outil est de fournir, en plus de la segmentation en plans du document vidéo, une liste des concepts (~130 concepts) qui correspondent dans la plupart des cas aux éléments d'informations contenus dans le document vidéo. Ces concepts sont regroupés en des catégories. Ce qui facilite la tâche de l'utilisateur du système. La figure suivante présente une architecture globale pour la description du contenu sémantique d'un document vidéo.

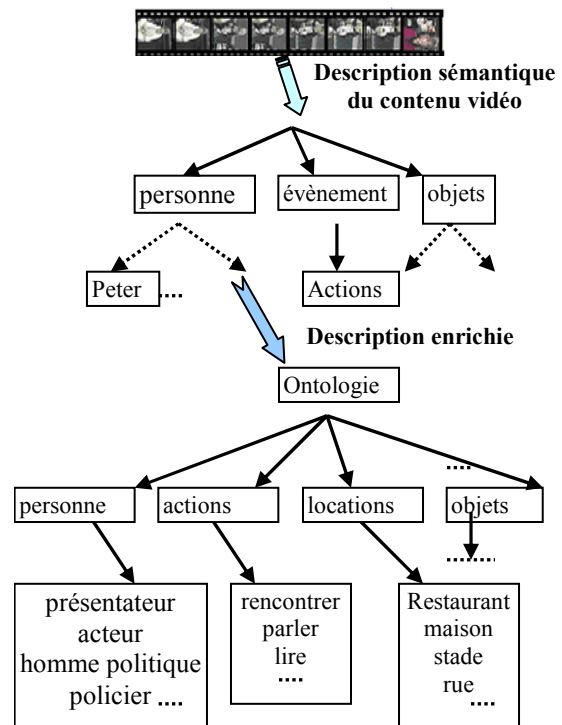


Figure 6: architecture pour la description du contenu sémantique

4. STRUCTURATION DU CONTENU VIDÉO

Définir une structure appropriée à un document vidéo qui permet de prendre en compte à la fois la sémantique et les descriptions bas niveau, constitue une étape essentielle pour le processus d'indexation par le contenu.

La structuration la plus classique se base sur une décomposition linéaire de la vidéo. Cette décomposition se base sur des descripteurs visuels permet de segmenter le document en des unités élémentaires dites 'plans'. Les méthodes utilisées dans

cette étape sont totalement automatisées. Parmi les inconvénients de cette approche de segmentation : l'absence de la notion sémantique lors de la description.

La décomposition linéaire peut prendre une dimension hiérarchique. L'ensemble des plans successifs sera regroupé en scène suivant l'unité de lieu. L'ensemble des scènes et plans, formant la même unité de sujet, sera regroupé en séquence. Nous présentons tout d'abord une définition de chaque unité.

Plan : c'est une série d'images acquise par une seule caméra et qui représente une action continue dans le temps et dans l'espace.

Scène : c'est une suite de plans formant la même unité de lieu.

Séquence : regroupe une suite des scènes et plans ayant la même unité de sujet.

La structure linéaire d'un document vidéo, dite aussi structure physique du fait qu'elle constitue une segmentation physique du document, tient en compte du facteur temporel lors de la décomposition. En effet, parmi les paramètres générés après une opération de segmentation de ce genre, on trouve principalement les intervalles de temps délimitant chaque plan, scène et séquence.

La figure suivante illustre un exemple de structuration physique du document vidéo.

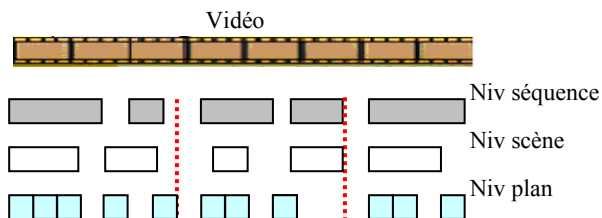


Figure 7 : structure physique d'un document vidéo

La structure linéaire du contenu ne permet pas de fournir une information sur le contenu sémantique. Ce qui rend son usage pour des procédures d'indexation et de recherche par le contenu ambigu.

Pour pouvoir décrire des caractéristiques visuelles du contenu, Il est nécessaire de procéder à une micro segmentation. Cette étape consiste, tout en préservant la procédure de segmentation en plans / scène / séquence, d'approfondir la phase de segmentation. En effet, pour chaque plan, il s'agit d'extraire une image clé et de la partitionner (segmentation spatiale) en des régions. Chaque région correspond à une entité visuelle.

De nombreuses méthodes proposées dans la littérature portent sur la segmentation du contenu audiovisuel. Ces approches se distinguent essentiellement par les techniques utilisées pour partitionner le document vidéo. D'une manière

générale, les paramètres exploités sont des descripteurs bas niveau (segmentation en plan, exploitation du flux audio et segmentation par séparation parole, musique etc...). D'autres approches utilisent les descriptions spatio-temporelles (J. Ma & Knight & Peng, 1997) pour segmenter le contenu vidéo.

4.1 Vers une structuration relationnelle

La représentation des relations au niveau de l'index permet aussi d'éviter certaines ambiguïtés pouvant altérer la précision des réponses fournies par le système. En effet, l'utilisation des mots clés pour indexer un document, génère souvent des ambiguïtés au niveau de la recherche et du choix des termes spécifiques pour que le résultat de la recherche soit le plus pertinent que possible. Une solution possible consiste à utiliser des termes plus expressifs et plus génériques (concepts) interconnectés par des relations sémantiques. Ainsi, par exemple dans le système RIME (Recherche d'Information MEDicale) (Mulhem & Berrut, 1995) par exemple, un formalisme de description conceptuelle a été utilisé comme langage d'indexation. Il s'agit des arborescences sémantiques dont le but est d'explicitier les relations sémantiques entre les divers concepts.

5. REPRÉSENTATION DES CONNAISSANCES ET RECHERCHE D'INFORMATION

La description et la représentation du contenu sémantique de la vidéo sont deux étapes assez importantes et complexes. La diversité du contenu au niveau type d'information (visuelle, auditive et parfois textuelle) provoque une multitude de possibilités pour l'analyse du contenu. En se plaçant au niveau description symbolique, dans laquelle un vidéo est vue comme une agrégation d'actions, d'objets visuels, de personnes, etc., l'intervention de l'opérateur humain s'avère indispensable pour décrire le contenu sémantique, classifier l'information et choisir pour chaque portion vidéo une description pertinente par rapport à son contenu. C'est à ce niveau que la difficulté de ce processus apparaît. En effet, pour une même portion vidéo, nous pouvons associer plusieurs interprétations plus ou moins différentes. Ceci dépend de la personne qui intervient pour interpréter et aussi par le choix du sous-média qui est pris en considération lors de cette phase.

Le problème classique à ce niveau résulte de la contrainte de spécifique / générique qui peut fortement influencer par exemple le résultat d'un système d'indexation et de recherche. Pour résoudre et réduire ces ambiguïtés, notre apport consiste à modéliser les interprétations associées aux différentes portions vidéos sous forme graphique facilement compréhensible par les utilisateurs et aussi traitable par un système d'indexation et de recherche vidéo basé sur le contenu sémantique. Pour cela il est nécessaire de prévoir un formalisme générique

permettant de représenter ce contenu et de tenir compte de l'ensemble de ces contraintes.

Le formalisme choisi doit permettre de représenter l'ensemble des informations de façon "uniformisée", ce qui devrait faciliter les processus d'indexation et de recherche par le contenu. Ainsi, une même interprétation décrite sous plusieurs formes différentes doit aboutir à une même représentation et inversement, cette dernière doit permettre de générer plusieurs interprétations équivalentes dans leur contenu sémantique. Les autres exigences importantes dans le choix du formalisme concernent les contraintes de puissance d'expression et de modularité. Toutes ces considérations ont conduit à orienter notre choix vers le formalisme des graphes conceptuels (GCs) tel que défini par J. F. Sowa (Sowa J.F, 1984). Cette représentation sémantique, sous forme de graphes conceptuels, est alors utilisée à tous les niveaux pour la représentation du contenu d'un document vidéo.

Dans le paragraphe suivant, nous présentons une brève introduction au formalisme des graphes conceptuels.

C'est un formalisme général de représentation des connaissances fondé sur la logique, permettant une meilleure étude (Sowa J F, 1984). C'est un langage de représentation des connaissances expressives et compréhensibles (voir figure 8). Il a été conçu dans l'objectif de développer un système de logique qui représente de façon plus simple et plus commode les structures du langage naturel.

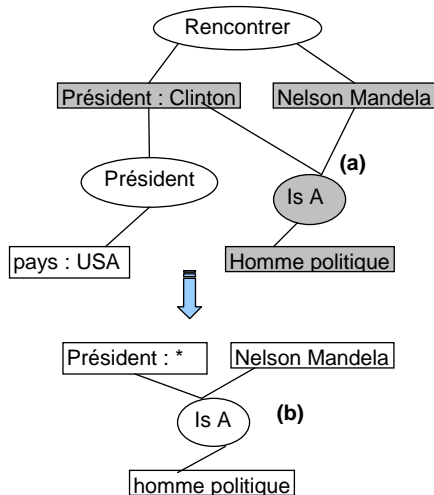


Figure 8: opérations de projection des GCs.

La définition développée par (Chein, 1992) relève un sens plus formel. En effet, Chein définit les graphes conceptuels, comme un graphe bipartite et connexe, mais l'exprime de manière plus formelle sous la forme suivante : $G = (C, R, A, \text{lab.})$

- C est un ensemble non - vide de sommets conceptuels, correspondant aux concepts instanciés du graphe.

- R est un ensemble de sommets relationnels, correspondant aux relations instanciées du graphe.
- A est une relation sur $(C \setminus R) * (C \setminus R)$, correspondant aux arêtes du graphe. Nous représenterons cette relation par un ensemble de couples de la forme (a, b) où $a, b \in (C \setminus R)$.
- lab. est une application qui associe à chaque sommet du graphe une étiquette.

5.1 Description conceptuelle étendue

Dans cette section, nous présentons la démarche suivie pour la classification des informations conceptuelles qui composent notre schéma de modélisation.

Comme nous l'avons mentionné dans la section annotation, le corpus contient essentiellement des journaux télévisés. Par défaut un journal télévisé se caractérise par sa structure particulière (plateau/reportage), ce qui facilite l'interprétation des différentes séquences. Notre approche de modélisation, ne tient pas compte de cette structure. Par contre notre objectif est de représenter le maximum d'informations contenues dans le document. Nous avons défini un certain nombre des classes de concept que nous estimons représentatives par rapport au contenu du document.

Trois catégories de classes de concept (personne, objet, évènement) sont décrites dans ce qui suit.

Les objets visuels composés des entités (des concepts logiques) qui peuvent être reliés à une ou plusieurs régions dans une image clé d'un plan vidéo. Un objet visuel peut être défini aussi comme une collection de régions visuelles, qui ont été groupées ensemble sous quelques critères définis par la connaissance du domaine. Ces objets devraient également satisfaire quelques conditions comme la conformité sémantique, la représentation d'un objet réel pour les utilisateurs.

Les évènements décrivent des interactions et les états des entités visuelles (acteur, objet, localité). Dans des cas d'étude spécifique sur des documents vidéo (sport par exemple). L'extraction des évènements peut être une tâche automatisée. En effet, une grammaire d'évènement qui se compose des règles pour décrire des types d'évènements est possible. Par exemple, dans une séquence vidéo contenant du sport, si un d'objet rond est à l'intérieur d'un objet filet pendant un moment et ceci est suivi de cris et siffles, il peut être identifié comme un évènement de but par exemple.

Les personnages sont des éléments d'informations qui sont reliés aux concepts évènements. L'interprétation d'une séquence vidéo contenant une ou plusieurs personnes, est généralement reliée à un état ou bien une activité. Il est possible de générer plusieurs interprétations plus ou moins différentes du même concept personne contenu dans la même

séquence. En effet, nous pouvons spécifier l'identité, la classe sociale, son activité etc.

La description conceptuelle n'est pas forcément restreint à ces trois catégories de classes de concepts, mais elle peut s'étendre à d'autres informations.

Le tableau suivant présente quelques classes et des instances utilisées dans le schéma de modélisation proposé dans l'approche de modélisation.

Classe	Instance de concept
Personne	<ul style="list-style-type: none"> ▪ identité ▪ fonction
Évènement	<ul style="list-style-type: none"> ▪ politique ▪ sportif ▪ économique ▪ etc....
Emplacement	<ul style="list-style-type: none"> ▪ Extérieur/ intérieur ▪ Nom de pays ▪ Localités
Objets	Entités visuelles
Activité	Réelle / filmique

Vu la diversité du contenu vidéo, il arrive souvent que pour une même séquence vidéo, on puisse associer plusieurs interprétations qui se différencient surtout au niveau abstraction et aussi au niveau source d'information utilisée (audio, image, texte). Ce dernier est un critère de distinction entre les différentes formes de requêtes. En effet, une requête basée sur le flux visuel a la forme suivante : « rechercher les séquences vidéo dans lesquelles une personne X apparaît ». Par contre celle basée sur le flux audio, elle aura la forme : « rechercher les séquences vidéo dans lesquelles une personne X parle ».

Les différentes possibilités d'interpréter la même séquence vidéo peuvent générer des problèmes au niveau application du schéma de modélisation. En d'autres termes si ce schéma est utilisé comme base d'index dans un système d'indexation et de recherche, il est nécessaire de trouver des moyens pour surmonter des problèmes de linguistique tels que la synonymie ou bien l'homonymie. Supposons par exemple que l'utilisateur formule sa requête en utilisant le concept « automobile », le système à ce niveau là ne peut pas forcément reconnaître qu'il peut retourner également les documents indexés par le concept « voiture », les deux termes sont synonymes.

Pour cela nous proposons une extension du schéma de modélisation, en associant une structure ontologique qui en fait représente une sorte de base de connaissances. Cette structure sera exploitée surtout

pour enrichir la description au niveau schéma de modélisation. Elle permet aussi de restructurer les requêtes de sorte qu'on puisse spécifier ou bien généraliser la description formulée. La figure suivante, décrit l'architecture du schéma de modélisation étendue.

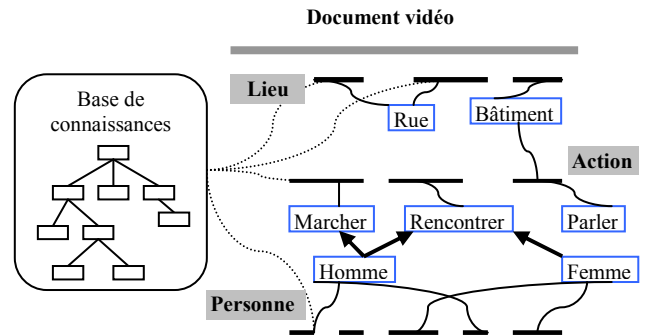


Figure 9: exemple d'un modèle d'annotation de contenu vidéo

5. DISCUSSION ET TRAVAUX FUTURS

L'enjeu principal de la modélisation et la structuration du contenu sémantique des documents vidéo, consiste à faciliter l'accès et la consultation du document en se basant sur une description sémantique. Le schéma de modélisation proposé dans cet article, représente un premier palier à franchir avant d'arriver à un cadre applicatif plus concret pour les utilisateurs. En effet, nous souhaitons exploiter cette modélisation dans le cadre d'un système d'indexation et de recherche par le contenu sémantique des documents vidéos dans le but d'améliorer la performance du système et de rendre le processus de recherche générique, indépendant et orienté utilisateurs.

Le choix de modéliser le niveau conceptuel de la vidéo résulte d'une part, du manque d'outils pour l'analyse du contenu sémantique et les limites des méthodes proposées pour exploiter la richesse sémantique de la vidéo. D'autre part, cette modélisation est en fait inspirée de l'image réelle à partir de laquelle l'utilisateur interprète une séquence vidéo.

Notre schéma de modélisation a été testé en grande partie sur le corpus TREC 2003. Nous avons restructuré l'ensemble des descriptions générées lors de la phase d'annotation. L'idée consiste à associer aux concepts des relations sémantiques en tenant compte des interprétations de chaque séquence vidéo. Plus l'interprétation est développée plus la chance d'avoir des relations dans la description augmente. Nous avons détaillé la modélisation spatio-temporelle dans notre schéma vu l'importance d'informations temporelles et spatiales lors de la description du contenu vidéo aux niveaux perception visuelle et sémantique. La nature temporelle de la vidéo est un aspect qui ne peut pas être négligé dans l'interprétation du contenu vidéo. L'ensemble des relations

temporelles est souvent implicite et désigne la chronologie des événements même si de temps en temps il n'est pas nécessaire d'exploiter la notion du temps proprement dit.

Nous avons proposé une sous-catégorisation dans chaque forme de modélisation. L'objectif est de faire la distinction entre une relation conceptuelle et un élément d'information qui peut inférer une relation conceptuelle. Par exemple dans le cas de la modélisation temporelle, il existe deux formes d'informations temporelles : les relations (before, after, etc.) et les temps sémantiques (durée, intervalle,...).

D'autres formes de modélisation ont été aussi traitées dans notre approche et même si elles ne sont pas détaillées dans cet article, elles sont implicitement représentées dans le schéma de modélisation. Nous évoquons ici, tout ce qui a un lien avec l'extraction des entités nommées (identité de personnages, nom de pays, acronymes, etc.). Ces informations sont sémantiques et infèrent généralement des descriptions sémantiques soit au niveau horizontal aussi bien au niveau spécification / généralisation.

CONCLUSION

Nous avons proposé dans cet article, un schéma pour la modélisation conceptuelle du contenu sémantique des documents vidéos. Ce schéma décrit un cadre unifié pour la spécification des critères sémantiques regroupant des informations issues des différentes modalités (visuelle, auditive et textuelle). La représentation se base sur des listes des concepts (personnes, objets, des événements). Cette approche vise à résoudre plusieurs problèmes tels que :

- ✦ Exploiter des descriptions sémantiques du contenu.
- ✦ Faciliter la manipulation et l'accès aux grandes bases de données vidéos en se basant sur la description niveau symbolique.
- ✦ Regrouper les descriptions issues des différents sous-médias dans un même schéma de modélisation.

Nous avons procédé à une analyse du contenu en s'attachant à décrire des critères objectifs (symboliques). À l'heure actuelle, mis à part quelques cas spécifiques, il est impossible de mettre en place un outil capable d'identifier automatiquement un nombre significatif de descripteurs symboliques. Nous avons proposé un processus semi-automatique pour pouvoir décrire le contenu. Ce processus utilise une base de connaissance qui jouera le rôle de système d'assistance. Cette représentation de connaissances permet aussi d'enrichir le schéma de modélisation, par la correspondance entre les différentes occurrences contenues dans le document vidéo.

Il reste à exploiter ce schéma de modélisation dans un processus d'indexation et de recherche pour

pouvoir évaluer sa précision et son efficacité pour la couverture des besoins en informations de l'utilisateur d'un système de recherche vidéo, basé sur le contenu sémantique.

Références

- (Assfalg & Bertini, & Colombo & Del Bimbo, 2002) Jürgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo: "*Semantic Annotation of Sports Videos*". IEEE MultiMedia 9(2): 52-60 (2002).
- (Celentano, Gaggi, 2002) Augusto Celentano, Ombretta Gaggi: "*Schema modelling for automatic generation of multimedia presentations*", 14th international conference on Software engineering and knowledge engineering, July 15-19, Ischia, Italy. ACM, 2002
- (Charhad & Quénot, 2004) M. Charhad and G. Quénot: "*Semantic Video Content Indexing and Retrieval using Conceptual Graphs*", ICTTA'2004, Damascus, Syria, 19-23 April, 2004.
- (Chein & Mugnier, 1992) Michel Chein, Marie-Laure Mugnier "*Conceptual Graphs : fundamental notions*", Revue d'intelligence artificielle, 1992.
- (Dagtas & Ghafoor, 1999) Serhan Dagtas, Arif Ghafoor: "*Indexing and retrieval of video based on spatial relation sequences*", ACM Multimedia (2) 1999.
- (Etievent & Lebourgeois & Jolion, 1999) Emmanuel Etievent, Frank Lebourgeois, Jean-Michel Jolion, "*Assisted Video Sequences Indexing: Motion Analysis Based on Interest Points*", Journal-ref: Iciap 99, Venezia, 27-29, 1999.
- (Genest, 2000) David Genest : "*Extension du modèle des graphes conceptuels pour la recherche d'informations* ", Thèse de doctorat en informatique. Université de Montpellier II, 159 pages.
- (Hjelsvold & Midtstraum, 1994) Rune Hjelsvold, Roger Midtstraum: "*Modelling and Querying Video Data*" proceedings of the 20 th VLDB Conference Santiago, chile, 1994.
- (Hollfelder & Everts & Thiel, 2000) Silvia Hollfelder, André Everts and Ulrich Thiel: "*Designing for Semantic Access: A Video Browsing System*". Multimedia Tools and Applications 11(3): 281-293 (2000).
- (Hunter & Armstrong, 1999): J. Hunter, L. Armstrong, "*A Comparison of Schemas for Video Metadata Representation* ", WWW8, Toronto, May, 1999
- (J. Ma & Knight & Peng, 1997) Jixin Ma, Brian Knight, and Taixin Peng: "*Representing Temporal Relationships Between Events and Their Effects*", Proceedings of the Fourth International Workshop on Temporal Representation and Reasoning, IEEE Computer Society Press, pp.148-152, 1997.
- (Lee, Li & Xiong, 1999) John Chung-Mong Lee, Quing Li and Wei Xiong: "*Automatic and Dynamic Video Manipulation*", in Furht'99, pp 317-343, 1999.
- (Lin, Tseng & Smith, 2003) Ching-Yung Lin, Belle L Tseng and John R Smith, "*VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept*

- Learning*," IEEE Intl. Conf. on Multimedia & Expo (ICME), Baltimore, July 2003.
- (Mechkour, 1995) Mourad Mechkour, "*EMIR2: An Extended Model for Image Representation and Retrieval*." in DEXA'95, Database and Expert system Applications, London, September, pp395-404, 1995.
- (Mulhem & Berrut, 1995) Philippe Mulhem and Catherine Berrut, The RIME Prototype, in Proceedings of the final Workshop on Multimedia Information Retrieval (MIRO '95), Glasgow, Scotland, UK, 1995.
- (Ounis, 1998) Iadh Ounis and Marius Pasca, "*RELIEF: Combining expressiveness and rapidity into a single system*", in 21st International ACM SIGIR, ACM Press, Melbourne, Australia, August 24-28, pp266-274, 1998.
- (Ronfard & Tien, 2003) R. Ronfard, Tien T: "*A framework for aligning and indexing movies with their script*", IEEE International Conference on Multimedia and Expo July 2003 Baltimore, USA.
- (Rowe & Boreczky & Eads, 1994) Lawrence A. Rowe, John S. Boreczky, Charles A. Eads: "*Indexes for User Access to Large Video Databases*". Storage and Retrieval for Image and Video Databases (SPIE) 1994
- (Seyrat & Durand & Faudemay, 1998) Seyrat Claude, Durand Gwenaël, Faudemay Pascal "*Méthode d'indexation multimédia fondée sur les Objets Visuels*" CORESA'98, Lannion, France, juin 1998
- (Sowa, 1984) Sowa John F, "*Conceptual Structures: Information Processing in Mind and Machines*", Addison-Wesley publishing company, 1984.