Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique

Mbarek Charhad

UNIVERSITE JOSEPH FOURIER

École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique

UFR IMA

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER- Grenoble I

Discipline : Informatique

Mbarek CHARHAD

TITRE

Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique

Date de soutenance : 28 novembre 2005

Composition du Jury :

Président :	***M. Jean Caelen***
Rapporteurs :	***M. Bernard Mérialdo***
	Mme. Sylvie Calabretto
Examinateur :	***Mme. Catherine Berrut***
Directeur de thèse :	***M .Georges Quénot***

CLIPS - Communication Langagière et Interaction Personne Système

CLIPS-IMAG

Remerciements

Je tiens à remercier

M. Jean Caelen directeur de recherche CNRS et directeur du laboratoire CLIPS, qui m'a fait l'honneur de présider ce jury.

Les rapporteurs sur ce travail : M. Bernard Mérialdo, Professeur à l'institut EURECOM Sophia-Antipolis, Mme Sylvie Calabretto, Maître de Conférence HDR, de l'INSA de Lyon, pour avoir accepté d'être rapporteur et pour l'intérêt qu'ils ont manifesté pour ce travail;

Mme Catherine Berrut, professeur de l'Université Joseph Fourier, pour avoir accepté de participer au jury et pour ses conseils, ses aides et son encouragement qui m'ont été très précieux.

M. Georges Quénot, chargé de recherche CNRS, qui a dirigé ce travail : ses remarques attentives, son encadrement de mon travail depuis le D.E.A et pendant ces années de thèse, la gentillesse et la patience qu'il a manifesté à mon égard durant cette thèse;

M. Philippe Mulhem, chargé de recherche CNRS, ses conseils, sa connaissance du domaine et son écoute attentive m'ont été très utiles;

Tous les membres de l'équipe MRIM pour leur accueil dans l'équipe, et pour leurs questions pertinentes lors des réunions, qui ont fait avancer ce travail.

Saïd, Mohammed, et Khaled qui m'ont soutenu comme seuls savent le faire les amis;

Ma famille qui a su manifester son soutien et m'entourer d'affection malgré les kilomètres.

Résumé

Les avancées technologiques dans le domaine du multimédia, associées à la généralisation de leur utilisation dans de nombreuses applications (archivages télévisuelles, vidéosurveillances, etc..), ont rendu possible le stockage des grandes collections de documents vidéo dans des systèmes informatiques. Pour permettre une exploitation efficace de ces collections, il est nécessaire de mettre en place des outils facilitant l'accès à leurs documents et la manipulation de ceux-ci. Une indexation par mots-clés (issus de la transcription de la parole et ou de sous-titre dans le document vidéo) est parfois possible. Cependant, l'utilisation de concepts peut améliorer les résultats de processus d'indexation et de recherche d'information parce qu'elle enlève les ambiguïtés entre les sens des mots-clés dus à la synonymie et l'homonymie. La précision de la description sera encore meilleure si, en plus des concepts non ambigus, des relations entre ces concepts sont indexées.

Les documents vidéo ont un caractère multimédia qui fait que la recherche par le contenu dans ceux-ci présente un certain nombre de spécificités. Par exemple, un concept donné (personne, objet...) peut être interprété de différentes manières : il peut être vu, il peut être entendu ou il peut être mentionné. Des combinaisons de ces cas peuvent également se produire. Naturellement, ces distinctions sont importantes pour l'utilisateur. Des requêtes impliquant le concept C comme par exemple : « rechercher les segments vidéos montrant une image de C » ou comme : « rechercher les segments vidéos dans lesquels on parle de C » sont susceptibles de produire des réponses tout à fait différentes. Dans le premier cas, on rechercherait C dans le contenu visuel tandis que dans le second, on rechercherait dans le contenu audio un segment dans la transcription duquel C est mentionné.

Cette étude s'inscrit dans un contexte de modélisation, indexation et recherche d'information multimédia. Au niveau théorique, notre contribution consiste à la proposition d'un modèle pour la représentation du contenu sémantique des documents vidéo. Ce modèle permet la prise en compte synthétique et intégrée des éléments d'informations issus de chacune des modalités (image, texte, son). L'instanciation de ce modèle est réalisée à l'aide du formalisme des graphes conceptuels. Le choix de ce formalisme est justifié par son expressivité et son adéquation au contexte d'indexation et de recherche d'information par le contenu.

Notre contribution au niveau expérimental consiste à l'implémentation (en partie) du prototype CLOVIS[1]. Nous avons intégré le modèle proposé dans d'un système d'indexation et de recherche vidéo par le contenu pour évaluer ses apports en termes d'efficacité et de précision.

Mots-clés : Recherche d'information multimédia, indexation conceptuel, document vidéo, graphe conceptuel, ontologie.

[1] **C**onceptual **L**ayer **O**rganization for **V**ideo **I**ndexing and **S**earch

Abstract

Advances in multimedia technologies have made possible the storage of huge collections of video documents on computer systems. In order to allow an efficient exploitation of these collections, it is necessary to design tools for content-based access to their documents. As this is the case for text documents, keyword based indexing and retrieval can be used (from speech transcript and/or closed captions for instance). Concept based indexing is an improvement over keyword based indexing because it removes the ambiguities between keyword senses due to synonymy and homonymy. The precision will be even better if, additionally to non-ambiguous concepts, relations between these concepts are indexed.

In the case of video, there are a number of specificities due to its multimedia aspect. For instance, a given concept (person, object ...) can be present in different ways: it can be seen, it can be heard, it can be talked of, and combinations of these representations can also occur. Of course, these distinctions are important for the user. Queries involving a concept C as: “Show me a picture of C” or as “I want to know what C2 has said about C” are likely to give quite different answers. The first one would look for C in the image track while the second would look in the audio track for a segment in which C2 is the speaker and C is mentioned in the speech.

The context of this study is multimedia information modelling, indexing and retrieval. At the theoretical level, our contribution consists in the proposal of a model for the representation of the semantic contents of video documents. This model permits the synthetic and integrated taking into account of data elements from each media (image, text, audio). The instantiation of this model is implemented using the conceptual graph (CG) formalism. The choice of this formalism is justified by its expressivity and its adequacy with content-based information indexing and retrieval.

Our experimental contribution consists in the (partial) implementation of the CLOVIS prototype. We have integrated the proposed model in the video indexing and retrieval system by content in order to evaluate its contributions in terms of effectiveness and precision.

Keywords: Multimedia information retrieval, conceptual indexing, video document, conceptual graph, ontology.

Table des matières

Table des figures

Chapitre VI

-

Introduction & Généralités

Chapitre I

Introduction

I.1 Contexte et problématique

Avec la multiplication des chaînes de télévision et grâce à des capacités de stockage sans cesse grandissantes, des centaines de milliers d'heures de données vidéo numériques sont stockées et archivées. De plus, les avancées technologiques réalisées ces dernières années dans le domaine de l'informatique (espaces de stockage de plus en plus considérables, numérisation des données, etc.) ont permis de simplifier l'utilisation de données vidéo dans d'autres domaines et aussi par le grand public. L'importance que représente un document vidéo est principalement due à sa richesse et expressivité sémantique. L'une des spécificités du document vidéo, sa nature hétérogène, le rend sémantiquement plus expressif mais, en même temps, elle laisse sa structure plus ambiguë. En effet, il est difficile de mettre en évidence une structure unique dans un document vidéo intégrant simultanément de l'image, du texte et du son).

En ce qui concerne la variété du contenu, ces documents peuvent être : des journaux télévisés, des émissions sportives, des films, des documentaires, des émissions de télé-réalité, des enregistrements de vidéosurveillance, etc. Chaque type de document possède sa propre structure qui le distingue.

Avec la croissance constante de masses de données vidéo, outre les problèmes de stockage et d'archivage, les problèmes d'utilisation, de recherche, de navigation et d'extraction se posent. La consultation des documents vidéo doit en effet être facile. Par conséquent un processus d'indexation et de recherche doit être mis en place de manière à accélérer la recherche de l'information souhaitée.

Savoir manipuler de l'information vidéo correspond à un fort besoin dans diverses industries de production, d'archivage ou de distribution de contenu vidéo. De façon très informelle, les systèmes manipulant la vidéo sont caractérisés par le traitement informatique intégré d'information exprimée dans divers média (son, image, texte). Les éléments clés permettant cette intégration sont la puissance et les possibilités croissantes de la nouvelle génération d'ordinateurs personnels ou de serveurs de calcul et de stockage qui apparaissent sur le marché. Ces ordinateurs se distinguent de la génération précédente par leur capacité à manipuler l'ensemble de ces média, leur stockage ainsi que leur échange sous une forme entièrement numérique. Il devient ainsi possible de restituer de l'information aussi bien sous forme audiovisuelle que textuelle et graphique à un ou plusieurs utilisateurs connectés à travers le réseau.

À l'heure actuelle, plusieurs propositions ont été faites pour l'indexation et la recherche par le contenu mais nous remarquons que la plupart d'entre elles mettent en avant un cadre d'étude très restreint et spécifique ou bien proposent une approche basée sur un type spécifique de média (texte, image ou audio). Ceci ne donne pas toujours des résultats satisfaisants et efficaces. Il est donc impératif d'accorder plus d'attention aux aspects de représentation et d'analyse du contenu vidéo d'une manière plus globale. C'est ce que nous voulons réaliser

dans le cadre de nos travaux. Nous nous intéressons aux aspects modélisation et représentation du contenu vidéo.

Un autre problème important pour l'efficacité des processus d'indexation et de recherche est la prise en compte de la variation individuelle dans l'interprétation des documents vidéo. En raison en particulier de la nature visuelle du signal vidéo, les données vidéo sont perçues et interprétées différemment par des personnes différentes. Il est impossible de représenter toutes les interprétations possibles par des mots clés (texte) car il n'est pas possible de les prévoir toutes au moment de l'indexation. En outre, la représentation d'un petit segment de signal vidéo par un grand nombre de mots clés mènera à l'explosion de la base d'indexation. D'autre part, les mots clés ne peuvent pas représenter la nature temporelle des signaux vidéo ni les rapports sémantiques des descriptions du contenu (règles d'inférence et hiérarchie, etc..).

Les représentations de document vidéo diffèrent selon le domaine d'application (archivage, recherche d'information, édition, etc.). Une expression du contenu sémantique n'est pas utile à toutes les applications manipulant de la vidéo. Ainsi une application d'archivage vidéo peut s'intéresser aux pixels de chaque image et aussi leurs positions spatiales pour des fins de compression, alors qu'une opération de montage manipule l'image entière comme unité de base et la considère comme l'unité élémentaire pour la vidéo.

La problématique de la représentation du document vidéo, de son contenu aussi bien sémantique que structurel se retrouve principalement lors des phases d'indexation, d'expression des requêtes et d'interaction avec l'utilisateur. D'une manière générale, on distingue deux possibilités pour représenter la vidéo :

- ✓ *La transmission ou le stockage* : on se place ici dans un contexte de codage d'information. Dans ce cas, il est nécessaire de passer par des étapes de compression qui produit une représentation codée plus compacte et difficile à manipuler. La phase de décompression servira à récupérer le maximum d'information originale du document. d'un point de vue pratique, c'est à dire avec le moins de dégradation possible selon une perception humaine. Plusieurs standards ont été définis. Citons par exemple les normes MJPEG, MPEG, etc. (voir figure ci-dessous).
- ✓ *L'indexation et la recherche* : dans ce cas, il est plutôt question de manipuler la vidéo pour mettre en place une base d'indexation qui constitue une représentation virtuelle du document vidéo servant d'intermédiaire entre le document et les besoins d'information exprimés sous forme de requêtes. Parmi les propositions qui ont été faites dans ce contexte, nous citons la norme MPEG7 et Dublin Core.

La figure I.1 récapitule ce que nous venons de décrire pour les deux possibilités de représenter les données vidéo.

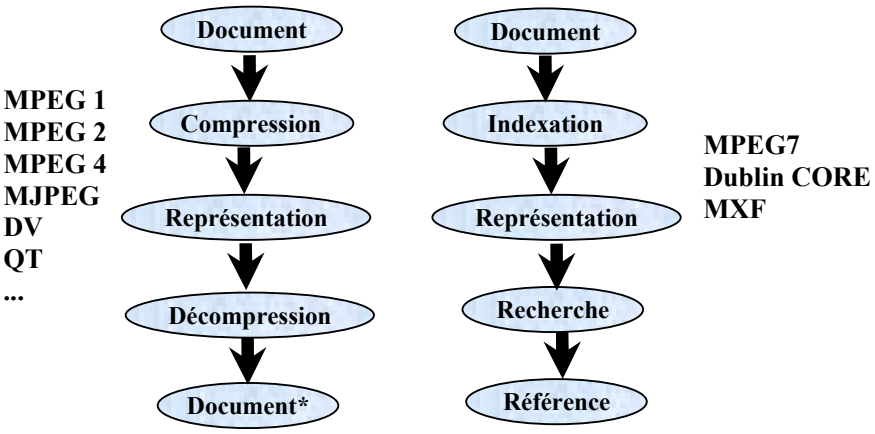


Figure I.1: Représentation des documents vidéo

Le problème de la représentation est donc important d'un point de vue théorique puisqu'il conditionne celui de la réutilisation et de l'accès. Il s'agit de générer une information qui servira de point de repère pour accéder aux segments vidéo et pour identifier les zones d'intérêt dans les documents. Il est important aussi pour ses applications pratiques, du fait de la numérisation de la production et des archives du cinéma et de la télévision, de la généralisation de la vidéo numérique familiale et de la diffusion croissante de vidéos sur internet.

Le besoin de systèmes informatiques pour manipuler les données vidéo est de plus en plus fort. Ceux-ci doivent être capables de gérer les données vidéo d'une manière efficace et précise. Dans un contexte réel, lorsqu'un utilisateur souhaite rechercher des documents vidéo ou des portions de documents vidéo, il est souvent plus pratique pour lui d'utiliser une information sémantique (un événement, un concept spécifique) pour obtenir les réponses les plus pertinentes. Or les systèmes actuels ne satisfont pas vraiment ce besoin du fait que dans la plupart des cas leurs auteurs privilégient un type spécifique de média. Par exemple, ils indexent les vidéos en exploitant uniquement leur aspect visuel. Il est difficile de traiter efficacement l'information sémantique par une telle approche. À l'exception de quelques approches proposées dans le cadre d'études très spécifiques (vidéosurveillance, émissions sportives, etc..), la plupart des études font le choix de d'un processus totalement automatique incapable de prendre en compte le point de vue que peut avoir un simple utilisateur visualisant une séquence vidéo. Il est souvent difficile, en partant d'une analyse bas niveau du contenu vidéo, d'atteindre le niveau sémantique. Partant de cette idée, et considérant aussi que pour un utilisateur d'un système de recherche d'information, ce qui compte le plus c'est la précision et la pertinence des réponses fournies pour sa requête, nous cherchons à mettre en place un modèle de représentation conceptuel du contenu vidéo. Ce modèle devra permettre d'intégrer les différents types de médias (images, texte et audio) et aussi de spécifier la manière dont le document vidéo sera indexé. Ce modèle a pour but de permettre une description fine et précise du contenu et, par exemple, de représenter différemment les cas d'une personne qui parle et d'une personne dont on parle. Ce que nous proposons dans le cadre de cette thèse concerne une modélisation du contenu « haut niveau » du document vidéo.

I.2 Objectifs de la thèse

Ce travail s'inscrit dans un contexte d'indexation et de recherche d'information Multimédia. Dans la perspective d'améliorer l'efficacité des systèmes de recherche d'information vidéo en termes de rappel et de précision, nous proposons une modélisation du contenu sémantique de la vidéo. Cette modélisation permet d'intégrer les descriptions issues de différents types de média et l'aspect temporel de la vidéo. Les interprétations du contenu d'un document vidéo sont souvent sémantiquement très riches et variées. Ceci rend la tâche de l'indexeur plus compliquée lorsqu'il s'agit d'une simple indexation soit par des mots-clés du fait qu'il doit choisir les meilleurs index pour décrire un contenu très riche en information. Soit en exploitant des descripteurs de bas niveau qui ne reflète pas les descriptions sémantiques du contenu vidéo. La même difficulté est aussi envisageable lors de processus de recherche. En effet, l'utilisation de « sac de mots-clés » pour rechercher un segment vidéo ne permet pas d'avoir des réponses pertinentes.

Pour surmonter ces difficultés, nous proposons de modéliser le contenu d'un document vidéo en exploitant la notion de concept et de relation conceptuelle. Cette proposition permettra de faciliter l'indexation et la recherche des documents vidéo par le contenu sémantique.

Nous montrerons, lors de la validation et de l'évaluation, les apports tant sur le plan théorique que sur le plan expérimental de notre proposition.

En résumé nos principaux objectifs sont :

♦ Définir un modèle pour la description du contenu des médias, qui soit approprié à la représentation du contenu des documents vidéo. Ce modèle doit prendre en compte les éléments d'information à différents niveaux de description (signal et sémantique) et les différents médias présents et qui soit adapté pour être intégré dans un système d'indexation et de recherche vidéo;

♦ Définir un modèle conceptuel générique qui met en avant les aspects représentation conceptuelle et interprétation sémantique. En effet, notre principal objectif est d'aboutir à un modèle qui soit à la fois compréhensif, simple à interpréter et exprime clairement les descriptions associées à chaque document vidéo. Pour pouvoir atteindre cet objectif et dans la perspective d'avoir un modèle d'indexation et de recherche par le contenu vidéo efficace et précis, il donc nécessaire de mettre en œuvre des représentations par des concepts et des relations conceptuelles.

Les objectifs présentés ci-dessus se situent clairement dans un contexte applicatif. Par conséquent, pour répondre à ces objectifs, il est question de mettre en œuvre un ou plusieurs modèles selon la spécificité de tâche. Ces modèles seront ensuite validés et évalués par des expérimentations sur une grande collection de documents vidéo afin de prouver leurs apports surtout dans le cadre du processus d'indexation et de recherche vidéo par le contenu. Parallèlement, ce dernier point nous permettra de découvrir les réels problèmes liés notamment à la description de contenu sémantique des documents vidéo.

I.3 Plan de la thèse

La suite de ce rapport est organisée en deux parties : Une première partie contenant un chapitre introductif qui porte sur des généralités du domaine de la recherche d'information et quelques définitions liée au vocabulaire de la vidéo (chapitre 2) ensuite, un état de l'art

(chapitres 3 et 4). La seconde partie décrit à la fois le modèle proposé (chapitre 5) et le prototype développé ainsi que les réalisations qui en découlent (chapitre 6).

Chapitre 2. Document Vidéo : Généralité et Caractéristiques

Ce chapitre comporte une introduction au domaine de la recherche d'information et une présentation des caractéristiques des documents vidéo. L'étude de ces caractéristiques est nécessaire pour se familiariser avec le vocabulaire utilisé pour la vidéo et de mettre en évidence les complexités et les difficultés susceptibles d'être rencontrées lors de l'analyse des documents vidéo.

Chapitre 3. Indexation et Recherche par le Contenu dans les Documents Vidéo : État de l'Art

Ce chapitre présente l'état de l'art des documents vidéo selon trois points de vue : analyse, description et indexation. L'analyse du contenu des médias permet d'identifier les caractéristiques de chaque type de média (image, audio et texte) et de donner une vision des possibilités de représentation du contenu dans ce domaine. Enfin, l'étude de la modélisation des documents vidéo effectuée selon les besoins d'intégration fine entre segments de média montre les limitations des modèles actuels. L'étude de la description vidéo est effectuée dans deux contextes : les standards et les travaux de recherche; elle permet de spécifier un modèle générique en se basant sur les standards, parmi lesquels les normes MPEG-7 ou Dublin Core.

Chapitre 4. Modélisation du Contenu des Documents Vidéo : État de l'Art

Ce chapitre concerne l'état de l'art en modélisation de la vidéo. On distingue trois classes de modélisation : hiérarchique, en strates et à base d'objet. Ce chapitre permet de situer le contexte de notre problématique de recherche en énumérant l'ensemble de travaux existant et leurs limites pour une description générique.

Nous donnons à la fin de ce chapitre la description du formalisme de représentation que nous utiliserons dans notre contribution et nous mettons en avant les apports de d'exploitation des bases de connaissances pour enrichir la description de la vidéo.

Chapitre 5.

Ce chapitre décrit le modèle que nous proposons. Il comprend quatre parties.

Partie I - Modélisation Multifacette et Multimodale : Représentation générique

Cette partie présente le modèle de représentation du contenu vidéo. C'est une proposition d'un schéma générique intégrant plusieurs facettes de description. Dans un premier temps on décrit le modèle de base de cette étude (EMIR²) proposé pour la représentation et la recherche symbolique des images fixes. Nous montrons via l'aspect générique de ce modèle, qu'il est possible de l'étendre pour l'adapter à la représentation du contenu vidéo. Nous proposons de spécifier l'ensemble des facettes dans le modèle EMIR² et aussi les facettes à ajouter pour modéliser le contenu vidéo..

Partie II - Modélisation Multifacette et Multimodale : Représentation spécifique

Cette partie décrit une représentation plus spécifique de notre modèle. Il s'agit de spécifier les caractéristiques spécifiques à chaque média (image, audio) pour la modélisation du contenu vidéo. Dans ce chapitre, on s'intéresse à une description du contenu vidéo qui soit multimodale et multifacette mais qui soit également simple à intégrer. On s'intéresse aussi à l'exploitation des descriptions conceptuelles du contenu.

Partie III - Proposition d'une Ontologie pour Enrichir le Modèle

Cette partie présente une extension de notre modèle par l'utilisation d'ontologies.

Partie IV - Modèle d'Indexation et de Recherche Vidéo par le contenu

Dans cette partie, nous décrivons le modèle d'indexation et de recherche vidéo.

Chapitre 6. Réalisation et Expérimentation

Ce chapitre décrit l'implémentation du prototype CLOVIS[2] qui s'appuie sur le modèle de représentation multifacette et multimodale proposé dans le chapitre 5. Il fournit des outils pour modéliser / indexer le contenu de la vidéo.

Chapitre 7. Conclusion

Dans ce dernier chapitre, nous effectuons un résumé sur l'apport de cette thèse. Nous tirons aussi le bilan des réalisations et nous mentionnons les perspectives de recherche suggérées par ce travail.

[2] **C**onceptual **L**ayer **O**rganization for **V**ideo **I**ndexing and **S**earch

Chapitre II

La Recherche d'Information Vidéo : Généralité et Caractéristiques

L'objectif de ce chapitre est d'expliciter le cadre général d'étude de notre problématique de recherche. Dans la première partie, nous présentons les différents aspects liés au domaine la Recherche d'Information (RI). Dans la deuxième partie, nous passons en revue les caractéristiques spécifiques des documents vidéo et nous donnons quelques définitions.

II.1 La Recherche d'Information

La *Recherche d'Information* (RI) est un champ d'études historiquement organisé autour des documentalistes et des institutions chargées de gérer un grand nombre de documents, principalement textuels (tandis que le champ des bases de données se structurait principalement autour des informaticiens et de la gestion des systèmes d'information de l'entreprise). Les méthodes et les concepts en vigueur dans la Recherche d'Information dépendent fortement de ses origines historiques, et sont plus adaptées aux systèmes d'information documentaires que les méthodes issues de bases de données, car plus centrées sur les besoins des utilisateurs. Par exemple les notions de reformulation de requêtes, de pertinence utilisateur, de *besoin d'information* proviennent de la RI.

II.1.1 Indexation et calcul de correspondance dans les SRI

Dans un Système de Recherche d'Information (SRI) trois dimensions sont prises en compte :

- l'*unité d'indexation* concerne le choix de ce qu'il y a à indexer. On peut ainsi indexer un document de façon individuelle, ou bien un ensemble de documents (par exemple des séries d'images) quand le fond documentaire est trop grand. La prise en compte de la structure du document, sous-jacente encore récemment dans un texte numérisé, mais désormais explicitée, se révèle alors d'importance.
- La *structuration* qui décrit l'organisation des descripteurs du document : cela peut aller de l'indexation à plat, qui consiste à placer des mots-clés les uns à la suite des autres, jusqu'à la mise en place de structures de représentation de connaissances, en passant par de simples rubriques de différents types.
- Les *vocabulaires d'indexation* peuvent être classés en deux grandes catégories qui sont les vocabulaires *libres* (construits *a posteriori*) et les vocabulaires *contrôlés* (des thesaurus par exemple), ces derniers étant construits *a priori* afin de remédier, en apportant de la cohérence d'ensemble, aux problèmes issus d'une part de la polysémie des mots de la langue et d'autre part de l'accroissement anarchique de la taille du vocabulaire d'indexation. Une procédure régulière de mise à jour du thésaurus est généralement prévue.

On pourrait rajouter à ces dimensions une quatrième qui tiendrait au *type de l'analyse* effectuée afin de mettre en place l'indexation : des systèmes manuels aux systèmes

automatiques qui ne gardent que les mots du titre ou du résumé d'un document, il est possible d'ajouter par exemple des termes fondamentaux non présents dans celui-ci (par exemple le domaine scientifique d'un article pointu non cité dans celui-ci). L'indexation automatique, utilise des méthodes mathématiques afin d'extraire les termes statistiquement les plus représentatifs d'un texte, alors supposés les plus pertinents. Le schéma de pondération dit « TF.IDF » est l'un des plus répandus dans le milieu de la RI. TF et IDF signifient respectivement "term frequency" et "inverted document frequency". TF est la fréquence d'apparition d'un terme dans un document. Il mesure l'importance du terme par rapport au document. IDF est la fréquence des documents indexés par un terme. Il mesure l'importance du terme en lui-même.

Une fois la base d'indexation construite, l'interrogation est la fonction principale des SRI. Elle offre à l'utilisateur les moyens d'exprimer son besoin par une requête construite selon un modèle prédéfini. L'étape suivante est alors la *mise en correspondance* entre la requête d'une part, et un document d'autre part. Une requête contient des critères décrivant les caractéristiques souhaitées des documents recherchés et le modèle de requête n'est pas indépendant du modèle de document choisi. La *fonction de correspondance* a pour rôle d'établir une base de comparaison entre les deux.

Passons ici très rapidement en revue quelques modèles classiques de la recherche d'informations. Malgré leur simplicité, ces modèles restent les plus utilisés, en combinaison avec des mécanismes permettant de pallier leur pauvreté d'expression.

Le ***modèle vectoriel*** représente un document D_i par un vecteur de dimension n représentant un ensemble de descripteurs ou mots-clés : $D_i = (d_{i1}\ d_{i2}\ ...\ d_{in})$ où n est le nombre de descripteurs connus et d_{ij} représente le poids affecté au descripteur j dans le document D_i. Une requête est de la même manière exprimée par un vecteur dans l'espace des descripteurs : $R = (r_1 r_2 ... r_n)$. La mesure dans l'espace des descripteurs (supposé euclidien) du cosinus entre un vecteur document et le vecteur requête est une mesure de similarité typique liée à ce modèle.

La puissance du modèle vectoriel réside dans sa simplicité conceptuelle et de mise en œuvre. Documents et requêtes sont exprimés de la même manière, et la mesure de similarité permet de classer simplement les documents retrouvés en fonction de leur pertinence vis-à-vis de la requête. Plusieurs problèmes importants subsistent : l'orthogonalité implicite de l'espace de représentation suppose une indépendance entre les termes, ce qui est une hypothèse très forte.

Le ***modèle booléen*** est basé sur l'utilisation de la logique de Boole pour proposer une représentation des requêtes, et repose sur une représentation classique des documents à base de mots clés. Trois types d'opérateurs (*et*, *ou*, *non*) servent à lier les critères de recherche formant une requête, ce qui permet d'y répondre en appliquant simplement ces opérations logiques sur des ensembles de documents extraits de listes inverses. L'expressivité du modèle est supérieure à celle du modèle vectoriel, puisqu'il permet de retrouver tout sous-ensemble particulier d'une collection de documents. Le modèle booléen peut être pondéré (c'est par exemple le modèle utilisé par Altavista).

Dans les ***modèles logiques***, un document est considéré comme pertinent s'il implique logiquement la requête (ce qui est par exemple trivial dans le modèle booléen). Les modèles logiques fournissent un cadre unificateur pour la recherche d'informations et permettent de prendre en compte toutes sortes de connaissances structurelles sur les documents, les contenus multifacettes, la représentation des connaissances, l'inférence, *etc.* Cette approche associe une grande puissance d'expression à une gestion uniforme des connaissances, mais souffre de

limitations liées à sa complexité théorique et pratique, ainsi qu'aux difficultés qu'il y a à mettre en place des modèles opérationnels liés à des indexations symboliques complexes (graphes conceptuels par exemple).

Le ***modèle probabiliste*** Le modèle probabiliste essaye d'estimer la probabilité qu'un utilisateur a de trouver un document d pertinent. Ce modèle suppose qu'il y a un sous-ensemble R de documents que l'utilisateur veut retrouver parmi ceux disponibles, les autres documents $\overline{R}$ étant considérés comme non pertinents. Un document *d* et une requête *q* sont représentés par un vecteur comme dans le modèle vectoriel, mais les poids sont binaires. Si $p(R/\vec{d})$ est la probabilité que le document *d* soit pertinent pour la requête *q* et si $p(\overline{R}/\vec{d})$ est la probabilité que le document *d* ne soit pas pertinent pour la requête q, alors la similarité entre *d* et *q* est :

$$sim(d,q) = \frac{P(R|\vec{d})}{P(\overline{R}|\vec{d})}.$$

II.1.2 Modèles théorique et opérationnel.

On distingue un modèle théorique (par exemple l'implication logique entre un document et une requête dans le contexte de la logique du premier ordre ; Figure II. 1) et un modèle opérationnel (par exemple la projection de graphes dans le contexte des graphes conceptuels ; Figure II. 2).

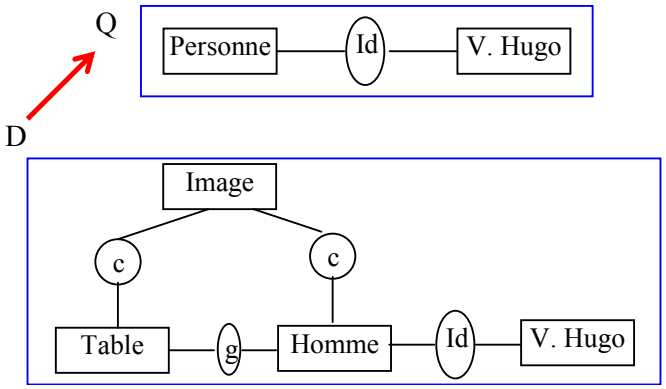


Figure II. 1: Modèle théorique

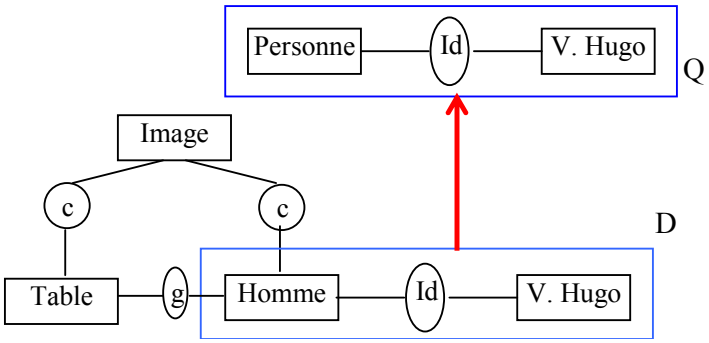


Figure II. 2: Modèle opérationnel

II.1.3 Cycle de recherche d'information.

La forme de représentation de l'indexation, des requêtes et la fonction de correspondance permettent de fournir un certain nombre de documents candidats à l'utilisateur. Un *cycle de recherche d'information* présente les différentes étapes d'une recherche :

- **la formulation** de la requête est l'étape d'expression par l'utilisateur de son besoin d'information dans la forme imposée par le système, cette étape peut être aidée par le système, par exemple si le vocabulaire est contrôlé, il y a lieu de fournir à l'utilisateur un accès au thésaurus ;
- **L'interrogation,** c'est-à-dire l'expression du besoin d'information de l'utilisateur sous la forme d'une requête, la recherche dans le corpus, et la présentation des résultats. Cette phase nécessite un modèle de représentation du besoin de l'utilisateur, appelé modèle de requête, ainsi qu'une fonction de correspondance qui doit évaluer la pertinence des documents par rapport à la requête. La réponse du système est un ensemble de références à des documents qui obtiennent une valeur de correspondance élevée. Cet ensemble est généralement présenté sous la forme d'une liste ordonnée suivant la valeur de correspondance.

  D'autres paramètres peuvent être considérés : le nombre de documents à présenter, la quantité d'information à fournir pour chaque document, le format de présentation utilisé, etc. Éventuellement, le système propose un mécanisme de retour de pertinence ("*relevance feedback*" [Salton 90]) : quand le résultat de la recherche n'est pas satisfaisant, le système reformule automatiquement la requête, en fonction du jugement de pertinence de l'utilisateur sur les documents déjà proposés. Il y a alors un apprentissage par étapes du besoin de l'utilisateur. Cette méthode permet à l'utilisateur de s'abstraire en partie des problèmes de formulation : syntaxe et complexité de la requête. De plus, certains concepts difficiles à exprimer le seront plus facilement "par l'exemple".
- **L'indexation :** dans un document textuel, l'indexation consiste à repérer dans celui-ci certains mots ou expressions particulièrement significatifs (appelés *termes*) dans un contexte donné, et à créer un lien entre ces termes et le texte original. Par exemple, les *pages d'index* d'un livre reprennent (parfois) les termes significatifs apparaissant dans le livre, et les relient aux pages du livre où ces termes apparaissent. Ceci facilite pour le lecteur la localisation des pages ou des sections où l'on mentionne un sujet particulier.

  On peut indexer sur le même principe des non textuels comme les images ou les vidéos. L'indexation peut être manuelle (faite par un humain), automatique (créée par un programme informatique), ou à divers degrés intermédiaires « assistée » ou semi-automatique (par exemple créée par un humain assisté d'un programme proposant des termes). L'indexation manuelle est généralement coûteuse : pour indexer correctement un texte scientifique d'un certain niveau, il faut faire intervenir des personnes qui soient elles-mêmes capables de comprendre le contenu du texte, ce qui impose un coût non négligeable.

  Bien que l'indexation se base sur des techniques relativement établies, il peut y avoir plusieurs indexations différentes d'un même texte, aussi valables les unes que les autres, en fonction de l'usage qui doit en être fait et du public auquel elles s'adressent.

- **Fonction de correspondance :** le système évalue la pertinence (la valeur de correspondance) des documents par rapport à la requête. La fonction de correspondance est un élément clé d'un SRI, car la qualité des résultats dépend de l'aptitude du système à calculer une pertinence des documents la plus proche possible du jugement de l'utilisateur. La figure II.3 récapitule l'ensemble des étapes décrites dans le cycle de recherche d'information.

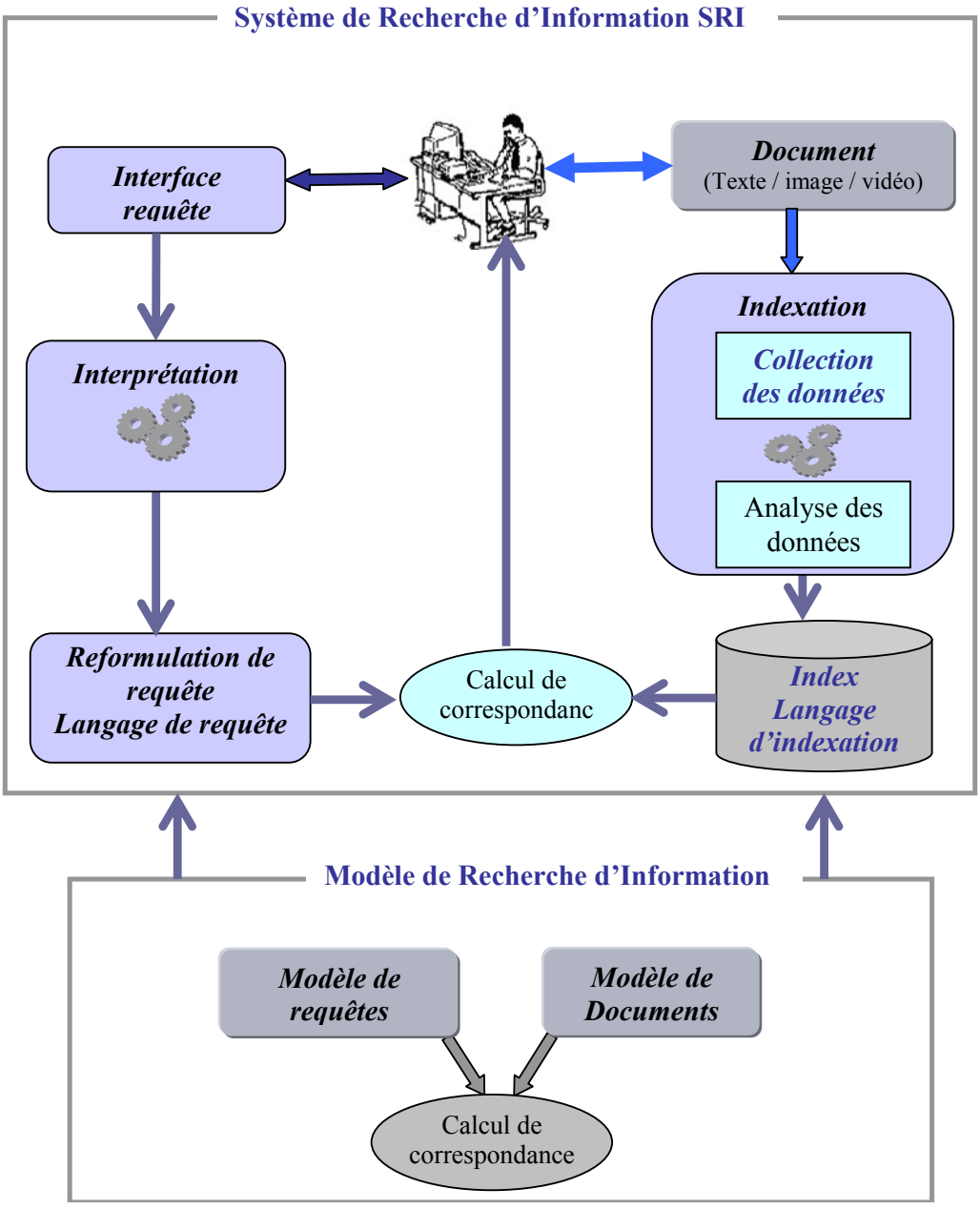


Figure II. 3 : Processus de recherche d'information

II.1.4 Évaluation

C'est dans le domaine des SRI textuels que se sont mis en place les concepts liés à l'*évaluation* des systèmes d'informations.

La ***pertinence*** recouvre des notions différentes selon que l'on se place du point de vue de la machine ou de l'utilisateur. Ainsi, du point de vue du système, la pertinence est la correspondance entre l'énoncé d'un besoin d'information (une requête) et un document, c'est-à-dire le point auquel le document recouvre la matière de l'énoncé du besoin. Le problème pour le concepteur du système est alors d'anticiper tous les besoins auxquels le SRI devra répondre.

Du point de vue de l'utilisateur, la pertinence dépend de l'utilité de chaque document que lui présente le SRI. Ainsi, la pertinence système et la pertinence utilisateur peuvent différer quand un document correspond -- du point de vue du système -- parfaitement à la requête, tandis que l'utilisateur peut n'en avoir que faire (par exemple parce qu'il en connaît déjà parfaitement le contenu). L'utilité d'un document pour l'utilisateur ne peut être mesurée qu'à travers les jugements que celui-ci émet lorsque le SRI lui présente celui-là. Elle dépend naturellement du contexte, *i.e.* de facteurs aussi variés que le but poursuivi par l'utilisateur ou que le contexte socioculturel dans lequel est menée la recherche.

Plusieurs concepts de mesure ont été mis en place afin d'évaluer la pertinence du système de recherche d'information :

- le ***rappel*** mesure la proportion de documents pertinents retrouvés par rapport au nombre total de documents pertinents dans la collection.
- le ***silence*** est la proportion complémentaire du rappel, c'est-à-dire la proportion de documents pertinents que le système n'a pas retrouvés. Un silence important peut par exemple être dû à une mauvaise indexation des documents, ou encore à une recherche trop stricte ou peu adaptée à la requête.
- la ***précision*** mesure la proportion de documents pertinents retrouvés parmi l'ensemble des documents retrouvés.
- le ***bruit*** est la mesure complémentaire de la précision, et donne la proportion de documents non pertinents parmi ceux proposés à l'utilisateur par le système.

Plusieurs remarques doivent venir compléter ces définitions. Ainsi, le calcul de ces mesures nécessite de connaître *a priori* quels sont les documents pertinents pour une requête, ce que seul un utilisateur est capable de dire. De plus, l'accord de deux utilisateurs différents n'est pas obligatoire : une requête que l'un estime satisfaite peut être considérée différemment par l'autre. Également, si la précision reste mesurable à travers les jugements de pertinence de l'utilisateur, le rappel est plus problématique : comment peut-on évaluer quels sont les documents pertinents non fournis par le système ? Il s'agit alors de mettre en place des bases de tests et des jeux de requêtes parfaitement connus afin de pouvoir tester les performances de systèmes différents

II.1.5 Interaction

La gestion de l'interaction de l'utilisateur avec les systèmes de recherche d'information en est devenue une composante naturelle et obligatoire.

Une première interactivité consiste en la possibilité de reformulation conduisant éventuellement à des mécanismes de *bouclage de pertinence* (*relevance feedback* en anglais). Celui-ci est né de la constatation d'une part que l'indexation était en général imparfaite et d'autre part que l'utilisateur avait de grandes difficultés à formuler dès la première tentative la bonne requête (ce qui traduit un décalage entre la fonction de pertinence du SRI et celle de l'utilisateur). L'idée est alors de prendre en compte la pertinence utilisateur pour améliorer les performances du système tout en tenant compte de ses performances passées. La recherche d'information passe alors par une suite d'étapes indépendantes au statut de processus *itératif*, dans lequel se met en place une véritable coopération permettant par un jeu de reformulations de la requête d'aboutir à un résultat satisfaisant pour l'utilisateur. À partir d'une première requête, le système fournit à l'utilisateur un ensemble de documents dont l'utilisateur évalue la pertinence, ce qui conduit à une reformulation automatique de la requête qui tient compte de ce retour de l'utilisateur. Cette nouvelle requête fournit alors un nouvel ensemble de documents, à nouveau évalué, et ainsi de suite jusqu'à satisfaction de l'utilisateur.

On remarquera qu'alors que dans le bouclage de pertinence la reformulation de la requête est automatique, certains systèmes utilisent une autre forme de formulation interactive mais non automatique qui consiste à fournir à l'utilisateur toutes indications utiles à la (re-)formulation (nombre d'occurrence du terme dans la base, utilisation du terme, termes proches dans un thésaurus, *etc.*). A noter également que si certain considèrent que cette technique est utile pour améliorer les résultats d'*une* session de recherche, elle reste peu concluante dans son utilisation pour modifier les représentations des documents dans la base, ce qui correspondrait à une utilisation à long terme, probablement à cause du caractère très subjectif des jugements des utilisateurs.

L'espace documentaire de navigation peut être construit *a priori*, par exemple, un modèle de recherche d'informations extrême est celui qui a cours dans les systèmes hypermédia : les requêtes en sont absentes, tandis que toute la recherche se fait lors de la navigation dans la base de documents (le Web ou un système de fichier sur une machine). Certains documents peuvent également être *construits au besoin*, le plus souvent comme manière d'organiser et de présenter les résultats d'une requête à l'utilisateur, l'accès aux documents numériques étant ainsi direct et immédiat. Il est également possible de considérer la mise en place automatique d'hyperliens dans les documents.

II.1.6 Recherche d'informations et connaissances

Il est possible de regrouper les connaissances prises en compte dans un SRI en trois classes principales : connaissances sur les documents, connaissances sur les concepts du domaine de l'application et connaissances sur les utilisateurs.

Les *connaissances sur les documents* sont en fait les index de ces documents, et *explicitent* des connaissances contenues dans et sur les documents, telles qu'elles ont été interprétées pendant l'indexation.

Les *connaissances sur les concepts du domaine de l'application* concernent le plus souvent le vocabulaire d'indexation et la manière d'organiser celui-ci en indexation structurée. L'organisation des connaissances de description est en effet nécessaire afin de guider la description des documents, que ce soit en phase d'indexation ou de recherche. On mettra dans cette catégorie essentiellement les thésaurus, qui permettent dans le cadre d'un vocabulaire

contrôlé de regrouper les différents termes utilisés ainsi que certaines relations entre ces termes, telles que la *synonymie* et les relations *spécifique/générique.*

L'organisation des connaissances du domaine peut donc varier d'une simple organisation de termes en thésaurus à une véritable organisation en base de connaissances (réseaux sémantique) en vue d'inférences préétablies à la conception.

Les *connaissances sur les utilisateurs du système* concernent tout d'abord ce qu'il est possible de savoir sur les besoins d'information des utilisateurs auxquels le système va avoir pour objectif de répondre. Ensuite, pour chaque utilisateur particulier, des connaissances peuvent être mises en place comme des profils ou des modèles d'utilisateurs que le système pourra alors créer et utiliser afin de répondre au mieux aux requêtes. Les connaissances sur le besoin d'information de l'utilisateur peuvent également être considérées. Par exemple définit différents besoins d'information dans une collection d'images correspondant à différentes stratégies de recherche :

- la *demande précise* : quand l'utilisateur sait parfaitement ce qu'il cherche, voire connaît le document dont il a besoin. Le documentaliste doit alors chercher précisément.
- la *demande exploratoire* : quand l'utilisateur veut se faire une idée d'une collection donnée sans *a priori*. Il s'agit alors de lui proposer des extraits jugés représentatifs de la base.
- la *demande thématique* est destinée à illustrer un thème. Le type de raisonnement alors suivi par l'utilisateur est un raisonnement pas association d'idées stimulé par la visualisation des documents.

II.1.7 Indexation et recherche de documents multimédia

Nous nous plaçons tout d'abord dans le cadre d'un document numérique multimédia faisant au moins appel à une forme d'appropriation non textuelle (image ou son par exemple).

L'indexation doit alors prendre en compte différents médias : texte, image, musique par exemple. Dans la lignée de l'indexation de documents textuels, l'approche la plus standard consiste à décrire un document quelconque avec une notice bibliographique. Celle-ci peut être mise en place pour répondre aux visées de l'institution (par exemple une agence de presse indexera ses images suivant son propre format). Il est également possible d'utiliser les métadonnées standard mise en place par la communauté des documentalistes, par exemple le Dublin Core [Hunter 99] (Titre, Auteur, Sujet, *etc.*) permettant de décrire minimalement tout document trouvé sur le Web.

Il est également possible de considérer le document comme un signal sur lequel sont calculées des caractéristiques, exprimées dans un langage de description. Répondre à une requête posée dans le même langage revient à calculer une *similarité* entre la requête et les index.

Sur une image par exemple, calculer un histogramme de couleur revient à extraire du signal brut un ensemble de composantes couleurs, lesquelles sont considérées comme une description de l'image. Une requête consiste alors en la description d'un histogramme par l'utilisateur, et en la comparaison de celui-ci avec ceux-là, un tri sur les résultats de la fonction de similarité permettant de proposer une suite ordonnée d'images solutions.

Trois remarques sont ici nécessaires. En premier lieu, on ne traite que du signal, c'est à dire que le niveau de sens atteint par les descripteurs est celui d'un *résultat de calcul*. Il est alors nécessaire de connaître l'algorithme d'extraction pour pouvoir les interpréter. Deuxièmement, il devient possible de fournir comme requête un document (de la même forme que celui que l'on cherche, par exemple une image), à charge pour le système d'en extraire les descripteurs pour former une requête. La similarité entre document requête et documents réponses ne résulte toujours alors que d'une similarité calculée, laquelle peut correspondre plus ou moins bien à une similarité au sens de l'utilisateur, dans le cadre d'une tâche donnée. Troisièmement, le fait d'extraire des mots clés d'un texte procède de la même démarche, c'est à dire que le texte est considéré comme signal sur des éléments duquel (les mots) un traitement statistique est réalisé. Même si le niveau symbolique n'est pas atteint, l'adéquation du système fonctionnel de la langue (les lettres, les mots, les textes) à la représentation machine fait qu'il est possible d'obtenir de bons résultats, car la machine manipule les même éléments que l'être humain.

II.2 Le média vidéo

II.2.1 Forme

D'un point de vue physique (ou informatique), un document (ou un flux) vidéo est une combinaison de sous-médias ou « pistes » organisés suivant un axe temporel. Chaque piste est présente sous la forme d'un flux d'éléments et les flux correspondants aux différentes pistes sont synchronisés entre eux. Ces différents flux peuvent contenir des images, du son ou du texte :

- ✓ **Image animée :** tous les documents (ou flux) vidéo contiennent une piste « image ». Les éléments de cette piste sont des images émises à une fréquence fixe (typiquement de 24 à 30 par secondes). Certains documents vidéo peuvent contenir plusieurs pistes image en parallèle. Ce cas est assez rare.
- ✓ **Son :** la plupart des documents (ou flux) vidéo contiennent aussi une (ou plusieurs) piste(s) « audio ». Les éléments de cette piste sont des échantillons émis à une fréquence fixe (typiquement de 16000 à 48000 par secondes). Une piste audio peut encore être composée de plusieurs flux de tels éléments en parallèle (cas de la stéréo, du codage à 6 canaux). Un document (ou flux) vidéo peut contenir plusieurs pistes audio en parallèle (correspondant à plusieurs langues par exemple.
- ✓ **Texte :** certains documents (ou flux) vidéo contiennent aussi une (ou plusieurs) pistes textuelles. Les éléments de cette piste ne sont généralement pas émis à une fréquence fixe mais plutôt par paquets accompagnés des informations permettant de les synchroniser avec les autres flux (temps de début et de fin; dans le cas des pistes image et son, la synchronisation se fait sur la base de l'émission régulière et à une fréquence fixe des éléments).

Les documents (ou les flux) vidéos numérique, comme les pistes qui les composent, sont en pratique toujours représentés par des séries de bits. En pratique aussi, les documents et les flux sont très souvent compressés sous forme numérique. Les méthodes de compressions efficaces induisent presque toujours des pertes mais la différence entre les versions originales et les versions compressées puis décompressées sont souvent peu perceptibles par un humain grâce à l'exploitation des redondances spatiales et temporelles présentes dans les documents

et grâce à l'exploitation des caractéristiques des systèmes perceptifs humains par les algorithmes de compression et de décompression (ce que nous ne percevons pas peut être supprimé ou rendu de manière approximative).

Toujours du point de vue physique, un document (ou flux) vidéo peut avoir une structure temporelle. Des marqueurs inclus dans les flux ou disponibles séparément peuvent délimiter des segments, éventuellement de manière hiérarchique.

Le contenu des documents (ou flux) vidéos est « rendu » à l'utilisateur par le moyen d'un matériel adapté qui convertit les flux d'éléments représentés sous forme numérique vers une modalité physique appropriée (par l'intermédiaire d'écran, de haut-parleurs, …) pour une perception par celui-ci.

Du point de vue de l'humain qui regarde et écoute le document ou le flux, la vidéo apparaît comme une expérience « globale ». Le spectateur n'a pas forcément conscience de la présence des différentes pistes et encore moins des éléments individuels qui les composent. L'expérience globale lui fait percevoir des éléments qui ont du sens pour lui (objets, personnes, événements, …) et des relations entre eux.

II.2.2 Contenu

Les documents vidéo peuvent avoir des contenus extrêmement variés (journaux télévisés, documentaires, films, publicité, vidéosurveillance, etc..). La plupart du temps, ces documents ont une (ou plusieurs) structures internes. Comme les documents eux-mêmes, ces structures peuvent être très variées. Les structures dont nous parlons ici sont conceptuellement différentes des structures physiques dont nous avons parlé dans la section précédente bien qu'il arrive souvent qu'elles coïncident en pratique lorsque les deux types existent. Le type de structure dont il est question ici est relatif au contenu sémantique du document. Il s'agit de structures qui ont un sens pour l'utilisateur. À ce titre, elles peuvent parfois apparaître mal définies, ambiguës ou subjectives.

Ces structures se présentent souvent de manière hiérarchique : un document est décomposé hiérarchiquement en unités plus petites selon un arbre (pas forcément régulier, et notamment parfois de profondeur variable). Une décomposition classique (mais particulière) distingue par exemple les niveaux « vidéo » (document dans son ensemble : suite de séquences), « séquence » (suite de scènes), « scène » (suite de plans), « plans » (suite d'images), « images » (images physiques, éléments de la piste image), « régions » (régions dans une image). La figure II.4 montre une telle décomposition dans le cas d'un journal télévisé.

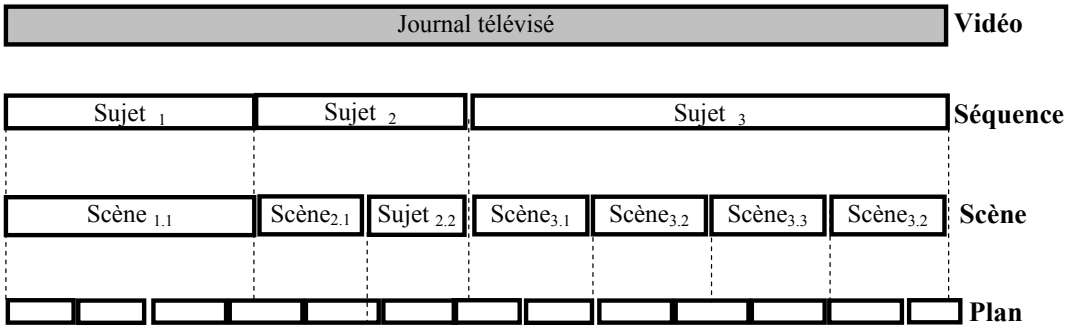


Figure II.4: Structure hiérarchique d'un journal télévisé selon le modèle « vidéo, séquence, scène, plan », les scènes peuvent correspondre à une apparition du présentateur ou à un reportage sur le terrain, les plans correspondent à une prise unique de caméra.

Pour des raisons pratiques et de sens, toutes les unités correspondent à un segment continu correspondant à un intervalle temporel dans lequel on considère toujours les différentes pistes ensemble et alignées excepté dans le cas où, pour les niveaux les plus bas, on « entre » dans un média particulier. Toujours pour des raisons pratiques, les éléments inférieurs de la hiérarchie ne se recouvrent pas entre eux, ils se suivent temporellement les unes les autres et, ensemble, ils recouvrent l'élément de niveau juste supérieur dans l'arbre hiérarchique.

Notons que la décomposition est valable simultanément pour toutes les pistes sauf pour les niveaux les plus bas où elle peut devenir spécifique à un média (le découpage en plans spécifique média image par exemple). Pour le média audio, une ou plusieurs autres décompositions de niveau(x) le(s) plus bas existent aussi. De même pour le sous média texte (télétexte).

II.2.3 Indexation et Recherche

Il existe à présent de nombreuses méthodes d'indexation d'images basées sur des caractéristiques proches de la "représentation signal" de celles-ci (couleur, texture, disposition de points caractéristiques...). Mais il n'existe pas d'algorithmes pour trouver une signification au contenu d'une image ou d'une vidéo, dans le cas général. Dans des cadres restreints, il est possible de définir des critères objectifs permettant de classifier automatiquement ce qui apparaît dans la vidéo, lui donnant ainsi un sens. Pour la surveillance d'autoroutes par exemple, le problème est restreint par le fait que l'on ne traite que des véhicules sur des images prises par une caméra fixe. Il est alors possible de caractériser de façon non ambiguë ces véhicules, qui apparaissent comme des régions de petite taille sur un fond stable.

Dans le cadre de l'indexation de séquences vidéo, la définition a priori de critères objectifs et universels d'interprétation est impossible compte tenu de la diversité des contenus. Pour prendre un exemple, considérons un journal télévisé. Dans ce type de documents se succèdent des plans quasiment statiques, avec des reportages sur le terrain, où les mouvements peuvent être rapides, et les objets d'intérêts de quantité et de nature totalement variables.

Dans ces conditions, une analyse entièrement automatique ne peut fournir d'information sur le contenu sémantique de la vidéo que dans des cas spécifiques prévus à l'avance. Pour une

interprétation générale il faut passer par l'interprétation d'un opérateur humain. L'avantage d'une indexation manuelle est de produire des tables d'index représentant explicitement le contenu sémantique des documents (souvent à l'aide de mots-clefs).

Cependant la qualité des index manuels dépend de la capacité de l'opérateur à décrire l'ensemble d'une scène en quelques descripteurs, dans le temps imparti: ainsi, si chaque scène doit être décrite en peu de temps, la description sera plus superficielle, et donc moins d'information sera disponible pour rechercher les documents répondant à une requête donnée.

Pour mieux comprendre la terminologie liée à la vidéo, nous proposons une comparaison de ses caractéristiques par rapport à d'autres types de documents. Dans [Hampapur 95], différents points pour spécifier un document vidéo ont été signalés.

- ✓ **Volume :** les données vidéo sont trop volumineuses et nécessitent plus d'espace de stockage par rapport à l'image et au texte. En effet, une seconde de vidéo MPEG contient 25 ou 30 images.
- ✓ **Hétérogénéité du contenu :** les informations contenues dans un document vidéo proviennent de sources variées (audio, texte, image). Cette hétérogénéité cause souvent des problèmes au niveau segmentation et analyse du contenu. Pour le cas de documents images ou textuels le problème d'hétérogénéité ne se pose pas.
- ✓ **La nature spatio-temporelle :** les données textuelles sont ni spatiales ni temporelles. Dans les données images, la représentation spatiale est bien définie par contre il n'y a pas la notion du temps dans les images fixes. Dans un document vidéo l'aspect spatio-temporel est une de ses caractéristiques.
- ✓ **L'expressivité sémantique :** il est parfois facile d'interpréter le contenu d'un document en visualisant son contenu mais il est souvent encore mieux de disposer également d'une information sonore. Un document vidéo possède une expressivité sémantique beaucoup plus riche qu'un document textuel ou même qu'un document image.
- ✓ **Durée :** pour représenter le contenu d'une image fixe, la durée de présentation est variable. Dans un document vidéo, une image possède une durée de présentation bien définie qui dépend du type de vidéo (~1/30 seconde pour le cas de vidéo NTSC et 1/25 seconde pour le cas des vidéos PAL / SECAM).
- ✓ **Variabilité de la qualité :** la qualité d'une vidéo dépend de la résolution dans laquelle elle est codée ainsi que du taux de compression utilisé pour son codage. Un même contenu peut être codé avec des résolutions très variables en fonction des capacités de stockage et/ou de transmission. Cela a un impact important sur la qualité des traitements qui peuvent être effectués pour une indexation automatique. De ce point de vue, le cas de la vidéo est différent de celui du texte pour lequel le contenu et son interprétation ne sont pas ou sont peu affectés par le format dans lequel il est encodé.

II.2.4 Définition de flux vidéo (ou flux audiovisuel)

Pour mieux représenter un document vidéo, nous décrivons dans ce qui suit les informations liées à son contenu. Nous essayons en premier lieu de définir la notion du flux vidéo afin de la distinguer de la notion du contenu ces deux termes sont souvent utilisés dans ce rapport.

Par définition, Un **flux** est un déplacement d'éléments dans le temps et dans l'espace. Dans le cas de document vidéo, ces éléments sont les données audio et visuelles.

Un document vidéo (ou document audiovisuel) peut être perçu comme une superposition de flux mais la manière dont il a été composé et la mise ensemble des éléments ne sauraient se réduire à une superposition de flux. Un document audiovisuel possède une structure physique décrivant la mise ensemble des images et la synchronisation du son plus ou moins complexe, tandis que sa structure logique, non explicitée, recouvre toutes les analyses possibles du document. Dans ce dernier cas il s'agit de la notion du contenu de document. Le terme contenu désigne l'ensemble des informations formant une vidéo (image, audio, texte).

(a) Définition de flux visuel

Si on filme une scène, c'est à dire qu'on enregistre à l'aide d'une caméra 25 ou 30 images toutes les secondes, et qu'on fait défiler ces images au même rythme devant un téléspectateur, celui-ci ne sera pas à même de distinguer dans ce qu'il voit une suite d'images. En effet, la persistance des images sur sa rétine fait qu'une image est remplacée par une autre, de telle sorte que les zones qui ne changent pas sont perçues commue un continuum stable, tandis que les mouvements sont lissés par la perception.

(b) Définition de flux audio

À la composante simplement visuelle d'un flux vidéo, on peut superposer un flux sonore. Si celui-ci a été enregistré en même temps que les images par la caméra, il est synchronisé avec le flux vidéo.

II.2.5 Éléments de structure

- **Les images** : les images numériques sont aujourd'hui disponibles en grande quantité, non seulement en milieu professionnel mais également pour des particuliers, grâce à la prolifération des appareils photo numériques. Leur acquisition, archivage, recherche et présentation constituent des tâches nouvelles

Une image numérique peut être représentée comme une matrice de pixels, mais il existe un autre type d'images numériques, les images vectorielles, dont le principe est de représenter, autant que cela est possible de le faire, les données de l'image par des formes géométriques qui vont pouvoir être décrites d'un point de vue formel.

Les images vectorielles sont des représentations d'entités géométriques telles qu'un cercle, un rectangle ou un segment. Ceux-ci sont représentés par des formules mathématiques (un rectangle est défini par deux points, un cercle par un centre et un rayon, une courbe par plusieurs points et une équation). Pour les présenter, il faut dessiner ces objets dans un système de coordonnées et le projeter sur une surface. Cela peut être une matrice de pixels qui sera traité par la suite comme une image matricielle, mais les objets peuvent être directement dessinés sur un papier ou sur certains types d'écrans. L'intérêt des images vectorielles consiste dans :

- leur taille réduite (au lieur de stocker des informations pour chaque point de l'image, il suffit de stocker la description géométrique des objets),

- leur invariabilité face aux modifications géométriques (il suffit de changer le système de coordonnées et redessiner les objets)

- la facilité de sélectionner ou ignorer une partie de l'image (en effet les objets sont manipulables séparément) Une image numérique matricielle est composée d'unités élémentaires (appelées pixels) qui représentent chacune une portion de l'image.

Une image est définie par : le nombre de pixels qui la compose en largeur et en hauteur et l'étendue des teintes de gris ou des couleurs que peut prendre chaque pixel (on parle de dynamique de l'image) [Boudry 2002].

Une image matricielle occupe beaucoup de place en mémoire (en général : nombre de lignes × nombre de colonnes × nombre d'octets codant les couleurs de chaque pixel) Plusieurs algorithmes de codage de l'image ont été développés pour compresser les images matricielles.

La manipulation des images passe par des méthodes et des systèmes de traitement et de gestion d'images. Dans une image on peut identifier *un fond* ou un *arrière plan* et *des objets*. Les objets sont délimités par un *contour*. De plus on peut également définir des *régions* en identifiant leurs frontières. Les régions peuvent être des formes géométriques simples (rectangle, ellipse, ..). Elles peuvent également être définies par les contours de différents objets. Des systèmes analysent les images numériques pour en extraire des signatures, des informations sur leurs textures, sur les couleurs dominantes, sur les contours, sur les objets présents, etc.…, [Hauptmann 02] Ces informations sont codées sous forme numérique, lisible pour un ordinateur mais pas forcément pour un humain. Les documents numériques images, peuvent être enrichis avec des métadonnées qui permettent de stocker des informations non visuelles, telles que la date de création/modification, les éventuelles caractéristiques de prise de vue s'il s'agit d'images prises par des appareils photo numériques, … Pour leur classement, recherche et réutilisation, les systèmes de manipulation des images se servent de ces informations obtenues automatiquement ainsi que des informations et descriptions créées manuellement. En effet, lorsqu'un être humain interprète une image, il la décrit généralement sous une forme textuelle ou symbolique.

Audio : par rapport au texte et aux images fixes numériques le son introduit une autre dimension qui est le temps. En effet par nature (le son est la perception d'une vibration) le son a une durée. Le son intemporel n'existe pas. La perception de l'information véhiculée par les documents sonores nécessite un certain temps.

Pour les documents numériques sonores, la notion de débit devient importante. En effet, pour pouvoir rejouer une musique ou un dialogue il faut disposer d'un certain débit de données. En général ce débit est exprimé en kbit/secondes. Le débit typique d'un morceau enregistré sur CD audio est de 44100*32 = 1411.2 kbit/s. Ce débit peut être compressé pour arriver jusqu'à 128 kbit/s.

Dans un document numérique sonore, des séquences peuvent être identifiées en précisant leurs bornes temporelles sur l'échelle de temps du document. Sur un morceau de musique de 3 minutes, nous pouvons dire par exemple que nous nous intéressons à la séquence qui commence à la 30ème seconde et qui se termine à la 40ème seconde.

Les documents vidéo : un document numérique vidéo contient souvent du son. Ce son est synchronisé avec les images dans le temps. Il peut avoir été enregistré en même temps que les images, ou bien en différé.

Une séquence vidéo est une mise en ordre des éléments d'un document numérique vidéo selon un ensemble de règles. Il s'agit des images numériques disposées chronologiquement avec éventuellement une bande son. Un document numérique vidéo peut être considéré comme une seule ou un ensemble de séquences vidéo. Une séquence vidéo est identifiée par son instant de départ et de fin sur l'échelle de temps du document vidéo numérique. En

général les documents vidéo sont le résultat d'un montage qui consiste à coller l'un après l'autre des *plans* avec d'éventuels effets de transition.

(a) ***Plan vidéo :*** Les plans se combinent à leur tour en des unités sémantiquement plus cohérentes appelées scènes. Dans le langage cinématographique, le découpage désigne la division en plans et scènes. Une scène est souvent considérée comme étant l'unité de référence de la vidéo. En d'autres termes, quand on décrit une vidéo à quelqu'un qui ne l'a pas vue, on donne thème général de chaque scène. D'un point de vue interprétation, rien ne prouve qu'une unité sémantique de la vidéo coïncide nécessairement avec la division en plans. La fin d'une unité sémantique peut ne pas se situer sur une transition de plan. Cependant, les transitions dans le contenu audio semblent plus proches des changements de thème dans le document vidéo. Le découpage de la bande son peut être par exemple basé sur le changement de locuteur.

Un plan est simplement défini dans un cadre de montage vidéo à partir d'une série d'images acquises par une seule caméra. La segmentation de la vidéo en plan peut être faite par un processus automatique qui se base sur la détection de transitions entre les plans. Par contre, pour segmenter une vidéo en des parties selon une description sémantique : unité de lieu (scène) ou unité de sujet (séquences), on doit généralement faire appel à un opérateur humain.

(b) ***Scène :*** une scène est constituée d'un ensemble de plan ayant une même unité de lieu.

Au niveau visuel, une scène vidéo soulève des problèmes tels que par exemple :

- Comment délimiter une scène?
- Sur quelle logique doit-on se baser pour déterminer l'enchaînement des scènes?

(c) ***Séquence :*** Une séquence regroupe diverses plans et scènes. Elle constitue une unité de sujet (par exemple un reportage dans un journal télévisé).

Parmi ces trois unités, le plan représente l'entité de base. Au niveau syntaxique, une telle structuration ressemble à celle d'un document textuel (mot, phrase, paragraphe).

(d) ***Unité audiovisuelle (UAV) :*** Par unité audiovisuelle, on définit une entité abstraite représentant un segment quelconque de document vidéo. Une unité audiovisuelle dépend de type de l'information ainsi que la manière selon laquelle ce type d'information est segmenté. Une unité audiovisuelle sera identifiée par la manière dont le document vidéo est segmenté. Par exemple, si on prend le cas d'une suite d'image on peut considérer qu'un plan représente une unité audiovisuelle. Par contre, pour le contenu audio, la segmentation selon le changement de locuteur peut constituer l'unité de repérage.

(e) ***Élément d'intérêt :*** On entend par élément d'intérêt un concept (visuel ou audio) décrivant de manière pertinente l'unité audiovisuelle. Il apparaît donc que n'importe quel élément du moment où il a été repéré dans le document comme élément d'intérêt donc il y aura autant d'élément d'intérêt de façon de mener une analyse particulière sur le document vidéo.

(f) Les *métadonnées(ou méta-information)* sont des données structurées qui décrivent d'autres renseignements. Dans le cas de la vidéo, les métadonnées décrivent les

informations ressources sur le document. Ces ressources peuvent être intégrées directement dans la description.

Les métadonnées peuvent être définies comme étant des données relatives à d'autres données (données sur des données). Par conséquent, une notice sur le document peut-être considérée comme métadonnée

Utilisée dans le contexte de la recherche d'information, les métadonnées sont perçues comme des informations de fond qui décrivent le contenu et autres propriétés et caractéristiques des données. On distingue trois types de catégories :

i. Métadonnées techniques : elles donnent les informations techniques sur le programme (ex : format d'image en TV, type de support, format d'enregistrement).

ii. Métadonnées administratives : elles donnent les renseignements nécessaires à la réalisation de tâches administratives liées à l'exploitation des contenus (droits d'auteur, droits de diffusion…).

iii. Métadonnées descriptives : elles indiquent le contenu des documents. Elles peuvent être relatives à un document entier ou à des séquences de document. Dans certains cas rares, il peut y avoir une description scène par scène avec le nom des intervenants ou acteurs, le dialogue, le résumé de la scène, des thèmes associés.

(g) ***Image-clé (keyframe) :*** une image-clé est une image qui contient toutes les informations nécessaires à son affichage. C'est une image complète qui va servir de référence pour la reconstruction des images partielles de la séquence. Théoriquement, le choix de l'image représentation d'un plan vidéo se base sur le critère de stabilité. Par conséquent, une image-clé est souvent celle qui correspond à l'image la plus similaire dans le plan.

(h) ***Les mouvements de caméra :*** il est très rare qu'un film soit composé uniquement d'une succession de plans fixes. Dans la très grande majorité des cas, la caméra est mobile et permet ainsi d'accompagner l'action. Ce mouvement de la caméra peut être discret ou ostentatoire. Lorsqu'il est discret, il permet de suivre l'action, d'accompagner un personnage, tout en sachant se faire oublier. C'est le cas d'un panoramique ou d'un travelling lent et fluide que l'on ne remarque pas, à condition qu'il ne soit ni en avance, ni en retard sur l'action. Cette dernière condition résume tout l'art d'un mouvement de caméra réussi. Si la caméra est en avance ou en retard sur l'action, son mouvement devient immanquablement ostentatoire. Il est perçu comme tel par le spectateur et trahit en quelque sorte la présence du réalisateur. Ce genre d'effet peut être voulu et permet de souligner une action particulière. La description des mouvements de caméra est réalisée par une combinaison de mouvements élémentaires illustrés dans la figure II.5. De plus, le zoom (ou changement de focale), qui n'est pas à proprement parler un mouvement de la caméra, est un degré de liberté supplémentaire qui leur est souvent associé.

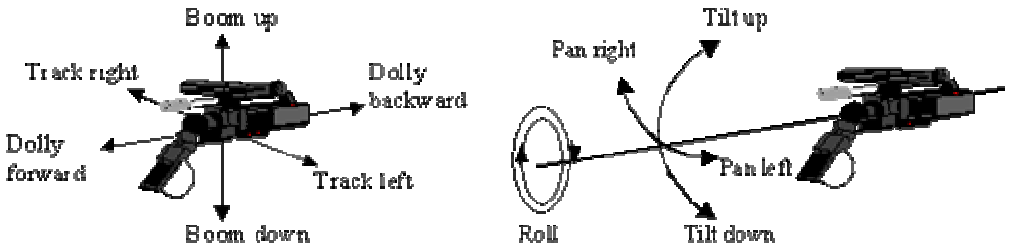


Figure II. 5 : Les mouvements de caméra

Conclusion

Dans ce chapitre, nous avons introduit les concepts de base liés au domaine de la recherche d'information. Nous avons présenté les composantes essentielles d'un système de recherche d'information (SRI). L'objectif de cette partie est de donner une première idée sur le contexte générale de notre problématique de recherche.

Dans un contexte de recherche d'information, l'interaction entre l'homme et le système pour une tâche impliquant la manipulation d'information nécessite que le système fournisse un ensemble de services de recherche à l'utilisateur. Pour le cas des documents vidéo, cette interaction peut être illustrée par des tâches telles que les approches de modélisation du contenu, les techniques d'analyse signal, les travaux sur l'indexation et la recherche d'information.… Certains de ces approches ont été récemment concrétisés par la définition de standards multimédias comme MPEG-7 ou Dublin Core.

Dans une seconde partie de ce chapitre, nous avons présenté un certain nombre de notions que nous pourrons ainsi réutiliser dans la suite du mémoire : documents vidéo, d'indexation, recherche d'information. Nous avons également évoqué quelques spécificités des documents vidéo notamment les aspects de multimodalité et d'hétérogénéité du contenu. L'analyse descriptive des documents vidéo a pour objectif d'une part de se familiariser avec ce nouveau concept et d'autre part de bien connaître et spécifier les vocabulaires.

État de l’art : Modélisation, Indexation et Recherche Vidéo

Chapitre III

Indexation et Recherche par le Contenu dans les Documents Vidéo : État de l'Art

Dans ce chapitre, nous présentons un panorama des travaux liés aux problèmes de l'accès aux documents vidéo et de leur manipulation. Plus précisément, nous nous intéressons aux travaux sur l'indexation et la recherche par le contenu. L'objectif est de faire une idée sur quelques systèmes et approches proposés dans la littérature. Nous introduisons aussi à la fin de ce chapitre quelques outils et schéma de description vidéo.

III.1 Introduction

La recherche d'information vidéo est une nécessité pour un grand nombre d'utilisateurs qui gèrent des informations audio / vidéo dans plusieurs secteurs d'activités. Beaucoup de travaux de recherche ont focalisé leurs efforts sur les aspects technologiques comme l'analyse et l'extraction automatique de l'information audiovisuelle et il existe assez peu de travaux qui prennent en compte les vrais besoins des utilisateurs.

La consultation d'une collection de données vidéo comporte généralement deux phases. La première consiste à spécifier le besoin d'information (le document désiré, un segment du document, etc.). La deuxième concerne l'accès au document ou au segment de document répondant au besoin. Une analyse exhaustive du contenu des documents à chaque nouvelle requête est exclue du fait, d'une part, du gros volume de données que représente une vidéo et, d'autre part, de la structure temporelle qui caractérise une vidéo. Il est donc nécessaire que les documents soient préalablement indexés.

Dans un contexte de recherche d'information vidéo par le contenu, la modélisation constitue une tâche importante et nécessaire à partir de laquelle les index seront formulés et grâce à laquelle le processus de recherche sera plus efficace et précis. Nous détaillerons dans le chapitre suivant (chapitre IV) les différents points liés à la modélisation des documents vidéo en faisant références à l'ensemble des travaux proposés dans l'état de l'art. Nous nous focaliserons dans ce chapitre sur les méthodes d'indexation et de recherche. Une partie de ce chapitre est consacrée à l'analyse du contenu et à la description de la structure de document notamment la structure temporelle.

III.2 Analyse du contenu des documents vidéo

III.2.1 Analyse signal vs analyse sémantique

An niveau analyse du contenu, on peut distinguer deux niveaux de descriptions qui sont liés aux données vidéo :

- Le niveau signal (ou « bas-niveau ») : proche de la représentation numérique des documents, il s'attache à décrire les caractéristiques « physiques » des segments d'une vidéo comme la couleur, la texture et la forme. Les informations

correspondant à ce niveau sont en général de type numérique (tableaux de nombres, histogrammes de couleur par exemple).

- Le niveau sémantique (ou « haut-niveau ») : proche de la façon dont les humains se représentent le contenu des documents, il vise la description des concepts présents et des relations entre eux.. Les informations correspondant à ce niveau sont en général de type symbolique (concepts, relations, graphes).

Certaines applications nécessitent à la fois une description signal et symbolique, se pose alors le problème du fossé sémantique. En effet, dans un contexte d'indexation et de recherche vidéo par le contenu, l'interrogation s'effectue souvent au niveau sémantique. Plusieurs approches ont été proposées pour réduire ce fossé sémantique consistant à intégrer des informations décrivant la structure de document.

Les systèmes existants pour l'analyse du contenu vidéo peuvent être classés en deux catégories : ceux qui exploitent le contenu bas-niveau ([Fablet 00], [Etievent 99], [Kobla 00], [Koubaroulis 97], [Pickering 02]) et ceux qui essayent d'extraire la sémantique à partir des signaux vidéo [Gunsel 96] [Souvannavong 02]. La plupart des systèmes existants sont classés dans la première catégorie. Un des inconvénients de ce type de systèmes provient du fait qu'ils travaillent en général sur un seul des médias présents (image ou audio). Ils n'exploitent donc pas toute la richesse sémantique présente dans la vidéo (image, audio et texte).

Un autre inconvénient de l'analyse de bas niveau est qu'elle ne peut en pratique être exploitée que dans le contexte d'une recherche par l'exemple. Il faut donc déjà disposer d'exemples de ce que l'on cherche pour lancer une requête. De plus, la similarité est basée sur une proximité dans les espaces des descripteurs signal comme la couleur, la texture, le mouvement, etc. Par exemple, dans le cas d'une recherche par similarité visuelle à partir d'une image exemple, une image recherchée peut être proche de la requête dans l'espace de description bas-niveau sans pour autant correspondre à ce que l'utilisateur recherche dans l'espace conceptuel dans lequel il pense. Lorsqu'il utilise des systèmes dans lequel les requêtes se font par images exemples, l'utilisateur ne dispose que d'images qui ne correspondent qu'approximativement à ce qu'il recherche. Dans la plupart des cas aussi, l'utilisateur ignore tout des analyses de bas-niveau qui sont effectuées sur les images, il ne comprend donc pas les rapprochements effectués par le système et il ne sait pas comment utiliser efficacement la fonctionnalité de recherche par similarité visuelle.

Ce problème existe aussi pour les systèmes de segmentation en plans, en histoires ou en scènes, et pour les systèmes d'extraction d'images clés. Ceux-ci travaillent principalement au niveau du signal et ont du mal à extraire des informations réellement utiles et sensées pour les utilisateurs.

III.2.2 Indexation manuelle vs indexation automatique

L'exploitation des documents vidéos dans des nombreux domaines (les médias, la médecine, l'éducation, etc..) rend nécessaire la mise en place d'outils qui soient efficace, précise et rapide. Or ces exigences nous paraissent contradictoires du fait que l'efficacité et la précision sont souvent reliées à l'opérateur par contre, la rapidité (temps d'exécution) est une caractéristique du système.

Dans un contexte d'indexation et de recherche d'information, l'utilisateur a plusieurs choix. Premièrement, un processus d'indexation manuelle. Cette indexation revient à annoter

manuellement le document vidéo. Mais face à l'augmentation incessante de la masse de données vidéo, cette tâche devient de plus en plus impossible à mettre en place. Deuxièmement, une indexation automatique des documents vidéo. Cette indexation malgré qu'elle soit applicable pour des grandes collections de document vidéo (gain de temps), pour beaucoup d'applications pratiques, la qualité de l'indexation est très insuffisante pour avoir des recherches précises et efficaces.

Il existe des solutions d'indexation dites « mixtes » ou « indexation assistée ». Dans ce type d'indexation, l'utilisateur intervient une plusieurs fois durant le processus d'indexation soit pour annoter ou raffiner les résultats d'indexation automatique.

Il n'y a pas de lien a priori ou nécessaire entre le niveau de l'indexation effectuée (signal ou sémantique) et la façon dont elle est effectuée (manuelle ou automatique). En pratique cependant, ce qui peut facilement être fait automatiquement est le plus souvent de bas niveau (calcul de descripteurs sous forme de tableaux de nombres) et ce qui doit être indexé à haut niveau (concepts et relation) ne peut l'être que manuellement (pour des raisons de faisabilité ou de qualité). Toutefois, dans de cas d'études spécifiques, une analyse automatique peut être exploitée pour et la détection et la reconnaissance des quelques concepts ou événements restreints à un genre spécifique de document vidéo (Sport par exemple). Ceci est beaucoup plus difficile voire impossible dans l'état de l'art actuel dans un cadre générique.

III.3 Les bases de données vidéo

Un système de bases de données vidéo est un système qui stocke et gère un ensemble de documents vidéo et qui permet d'accéder à ceux-ci à partir de références ou de requêtes. Quelques travaux de recherche sur les bases de données vidéo ont été présentés dans ([Lozano 00], [Lawrence 94], [Decleir 99]). Les concepts principaux à considérer dans ces systèmes : la structuration de la base, l'indexation des documents, et la formalisation des requêtes.

Les fonctionnalités d'une base de données vidéo sont illustrées au travers de deux exemples :

- ✓ Le premier traite les aspects de stockage et modélisation des données [hjelsvold94]. Les travaux développés dans [Lawrence 94], [Decleir 99] présentent principalement un modèle des données pour la gestion de base de données vidéo (voir figure III.1). Plus spécifiquement, le modèle a été conçu pour satisfaire des besoins de stockage et de structuration des documents vidéo ainsi que pour le partage des données. Des expérimentations ont été réalisées dans ce contexte et testées sur des journaux télévisés où le travail d'archivage représente une tâche de gestion immense à cause de la grande masse de données.
- ✓ Le second décrit une approche orientée objet pour modéliser les données vidéo [Lozano 00]. Cette approche permet de décrire le contenu de façon sous forme de base de données où chaque élément d'information est défini comme étant un objet.

La modélisation orientée objet offre des outils et des méthodes qui facilitent de façon significative la réutilisation de tout ou partie de contenus vidéo. Elle propose aussi des langages de requêtes permettant d'interroger des bases de données vidéo.

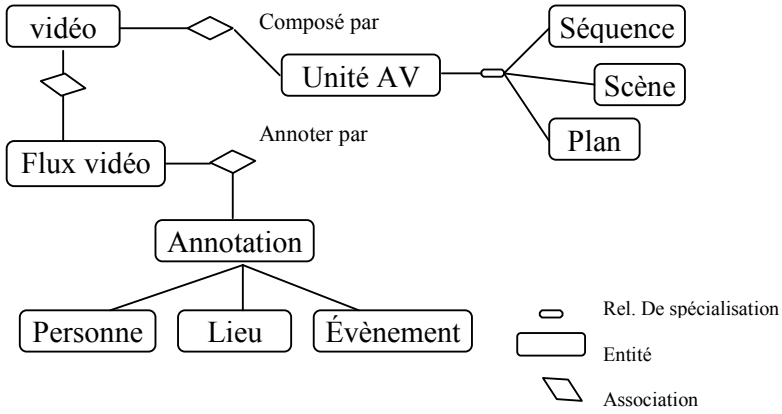


Figure III.1 : Le modèle de Hjelsvold

Dans ces travaux sur la modélisation de données vidéo, Hjelsvold a proposé un modèle de données formé de trois parties : stockage, structure et accès. Ces parties sont assez distinctes quoique reliées par des associations.

III.4 Indexation et recherche des documents vidéo

Certaines caractéristiques telles que la couleur, la forme, la texture sont indexées automatiquement. D'autres informations d'un niveau sémantique ne peuvent être obtenues que via une interaction avec un utilisateur sous formes des annotations (concepts). Dans tous les cas, il est nécessaire que le document soit segmenté pour faciliter son indexation. Un segment vidéo est une partie du document dont le découpage dépend de sous média en question. En effet, pour le sous média visuel ce segment peut correspondre à un plan, une scène ou une séquence.

Lors d'un processus de recherche, l'utilisateur souhaite consulter uniquement le segment qui l'intéresse. Pour conséquent, une représentation du document rend plus facilite l'indexation et améliore le processus de recherche.

Comme nous l'avons cité précédemment, L'indexation des documents vidéo repose sur trois types d'approches : les approches d'indexation manuelle, les approches d'indexation automatique et les Approches mixtes faisant intervenir un opérateur humain. Les approches manuelles consistent à associer à un segment une description conceptuelle, en général lié à une base de connaissances. L'intérêt de l'indexation manuelle est qu'un opérateur humain analyse et interprète le contenu pour le synthétiser et le reformuler. L'indexation automatique repose sur des algorithmes associant automatiquement des descripteurs à des parties de document. Dans le cas des documents textuels, à l'exception des mots vides, comme les conjonctions, pronoms, etc., chaque mot peut servir comme index du document qui le contient.

L'indexation du document vidéo est une tâche difficile et compliquée. En effet, ce type document ne se décompose pas en unités facilement repérables comme le cas pour le document textuel. Il faut donc disposer d'outils capables de segmenter le contenu et de le décrire.

III.4.1 Indexation manuelle ou assistée : annotation

Le processus d'annotation consiste à attribuer des descriptions pour le contenu de chaque séquence vidéo. L'annotation est une tâche souvent considérée comme étant un travail laborieux qui nécessite l'intervention de l'opérateur humain et qui dépend d'un processus totalement manuel. Cependant, l'annotation reste toujours plus sollicitée pour la description du contenu sémantique d'un document vidéo. En effet, elle permet d'analyser le contenu selon un point de vue utilisateur et ceci coïncide donc avec l'image réelle qu'un utilisateur peut retenir en regardant un une séquence vidéo.

Pour l'annotation conceptuelle, il est important de mettre en place une ou plusieurs ontologie(s) pour faciliter l'interprétation du contenu vidéo. En effet, l'utilisateur effectue l'indexation manuelle en recourant à des concepts de l'ontologie. L'ontologie est présentée dans l'outil d'annotation sous forme d'arbre graphique, ce qui permet à l'utilisateur de la parcourir rapidement et de sélectionner à tous les niveaux (hiérarchies de concepts) un concept qui lui semble pertinent pour son indexation.

Dans le cas où l'annotateur a une idée précise de ce qu'il veut mais ne connaît pas précisément le concept correspondant, la visualisation arborescente n'est pas adaptée : l'utilisation d'une terminologie permet d'assister l'utilisateur dans sa recherche. Celui-ci organise, par rapport à un élément d'information, les concepts correspondants selon un ordre de pertinence. Il existe des systèmes spécifiques pour réaliser les annotations qui génèreront l'ensemble des interprétations qu'on peut associer aux segments vidéo. Nous présentons dans ce qui suit une liste non exhaustive des systèmes d'annotation vidéo.

Plusieurs travaux proposés ([Vasconcelos 97], [Arslan 02], [Timothy 94]) dans ce contexte suggèrent l'utilisation d'un lexique bien déterminé permettant d'unifier les descriptions associées au document. De façon générale, le processus d'annotation est une tâche difficile qui nécessité l'attention de l'opérateur humain dans le choix des descriptions associées au document. D'autre part, les annotations dépendent aussi de la manière avec laquelle le document est structuré. En effet, associer une description à un plan vidéo est souvent sémantiquement mois riche qu'une description associée à une scène. Nous détaillerons Dans ce qui suit quelques outils d'annotation vidéo ensuite, nous passerons en revue quelques travaux de recherche sur l'extraction automatique de l'information contenue dans la vidéo.

III.4.1.1 Video-Annex : un outil d'annotation conceptuelle

L'une des spécificités de l'outil Video-Annex [Lin 03] semble être le fait que ces annotations conceptuelles peuvent s'appliquer à la fois sur le document complet, et sur des parties du document (segment vidéo ou image clé d'un plan). L'annotation qui porte sur le document en entier est effectuée à l'aide de possibilité de champs d'annotation libre (voir figure III.2). Ce champ permet aussi d'utiliser d'autres concepts qui ne figurent pas forcément dans la liste.

Cette annotation peut être collaborative ou bien indépendante. En ce qui concerne l'annotation collaborative, son rôle est de permettre le partage des interprétations variées. Ces informations peuvent être génériques (titre, auteur, date, ..) et ne posent pas de problème de principe. Par contre, l'annotation conceptuelle libre est subjective et permet d'élargir le vocabulaire d'annotation,

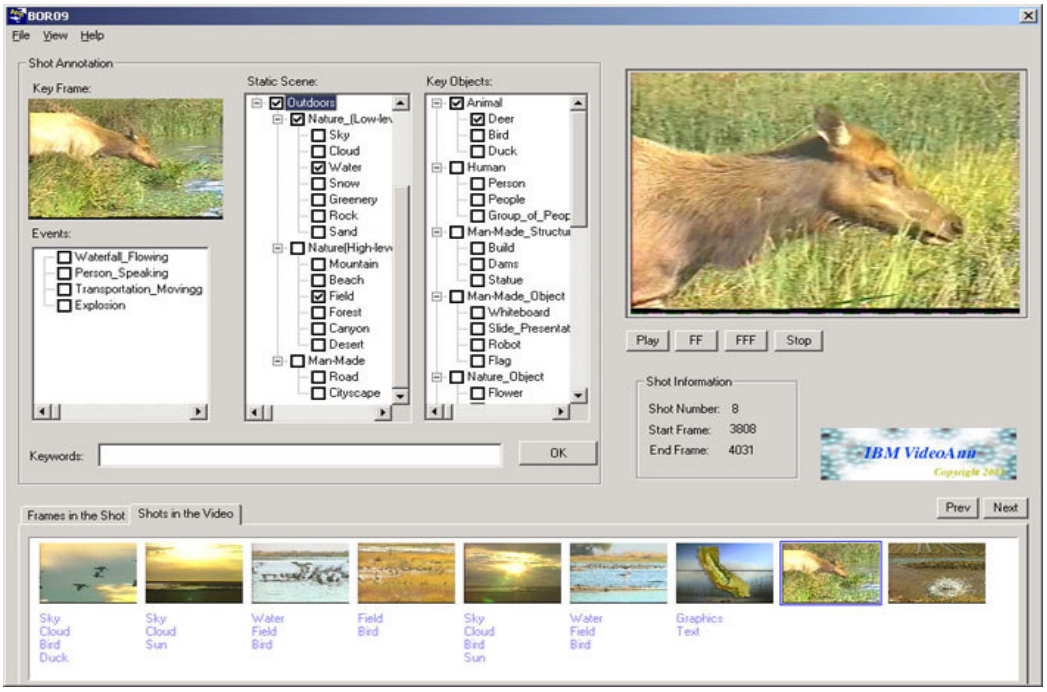


Figure III.2 : Interface de l'outil d'annotation Video-Annex

Cet outil est divisé en trois régions :

(a) Zone pour affichage et visualisation du plan vidéo

(b) Zone d'annotation du plan, dans laquelle apparaît l'image-clé de chaque plan et aussi les concepts sélectionnés pour annoter ce plan.

(c) Zone d'affichage, dans laquelle apparaît l'ensemble et des plans dans la vidéo (onglet2) et aussi les images de chaque plan (onglet1).

Nous avons exploité les annotations générées l'outil Video-Annex sur la collection TRECVID pour la description et la modélisation du contenu visuel que nous allons détailler dans le chapitre V. Notons enfin que cet outil se base sur une structure ontologique pour l'annotation composée de trois catégories de concept (scène, objets, événements) comme indiqué dans la figure ci-dessus.

III.4.1.2 Smart VideoText

Smart VideoText est un système d'annotation vidéo basé sur le formalisme des graphes conceptuels [Sowa 84], [Chein 92] proposé par [kokkoras 02]. Dans ce système, les portions vidéo représentent les nœuds du graphe. Ces portions sont identifiées par des références liées à la structure physique du document (identifiant du plan, numéro de l'image dans le plan, etc.) et aussi par les annotations libres.

L'idée de base du modèle d'annotation Smart VideoText est de relier les descriptions du contenu de document, décrites par des annotations, au flux vidéo. Chaque annotation sera représentée par un segment logique qui est en général une partie du flux vidéo.

III.4.1.3 COALA – Log Creator –EPFL

Le projet COALA (content Oriented Audiovisuel Library) conduit par l'EPFL en Suisse a débouché sur la réalisation d'une plate-forme prototype d'indexation et d'annotations des journaux télévisés de la TSR (Télévision Suisse Romande). Contrairement aux autres outils, il se présente comme une application du Web spécialisé dans l'annotation d'un genre particulier de document vidéo.

Le système se compose d'une interface (Log Creator) de segmentation et d'annotation des journaux télévisés [Fatemi 01]. La segmentation consiste à découper a priori le document selon une structure hiérarchique (voir figure III.3). L'annotation consiste à ajouter une description appropriée à chaque segment.

Figure III.3 : Interface de segmentation du système Log Creator

III.4.1.4 Autres système d'annotations audiovisuelles

Le système d'annotation Anvil [Kipp 01] développé par M. Kipp a été originellement développé pour l'étude de la gestuelle. Le système Anvil permet une annotation suivant des schémas d'annotations prédéfinis par l'utilisateur. La trace d'observation issue de l'annotation est composée d'une marque temporelle de début et de fin d'événement, de l'événement lui-même puis d'un ensemble d'attributs relatifs à cet événement. Suite à une annotation, le logiciel propose des fonctions de regroupement de données suivant les besoins d'analyse de l'utilisateur. On peut noter également que le fichier de traces résultant est au format XML (eXtensible Markup Language) [Bray 98] afin de permettre l'exportation des résultats d'annotation vers d'autres applications.

Le système SignStream [Neidle 01] a été élaboré pour l'étude de données audiovisuelles sur le langage parlé et langage des signes. Le système permet d'annoter les vidéos à partir d'un ensemble d'éléments prédéfinis par le logiciel ou définis par l'utilisateur. De plus, le logiciel

propose un module de création de script c'est-à-dire le scénario correspondant au corpus audiovisuel, ce script permettant par la suite de naviguer dans les médias utilisés. Finalement, le résultat de l'annotation peut être exporté au format XML.

Nous avons fait ici un rapide tour d'horizon des systèmes disponibles pour l'annotation et l'analyse de documents vidéo. Il nous a permis de voir leurs principes de fonctionnement et quelques uns de leurs applications dans le cadre de la recherche. Nous avons pu remarquer que la plupart des travaux se proposent de configurer les annotations autorisées pour permettre ainsi aux chercheurs de bénéficier uniquement d'annotations pertinentes pour la description du contenu vidéo. De plus, ces logiciels présentent, en général, la possibilité de gérer et d'annoter une collection pouvant comporter plusieurs documents vidéo. Une autre caractéristique commune à ces systèmes, c'est la possibilité de traduire les annotations sous forme graphique ou textuelles permettant ainsi aux utilisateurs de disposer d'une vue qui leur convient.

III.4.2 Indexation et recherche automatique et catégorisation par média

Bien que le processus d'indexation et de recherche automatique vise à traiter automatiquement les documents de manière à rendre plus efficace l'accès ultérieur aux documents de ces corpus. Celui-ci reste loin d'être efficace et précis notamment pour le cas des documents vidéo.

D'une manière générale, On distingue deux types de système d'indexation de séquences vidéo. Il existe d'une part *les systèmes dits génériques* [Amato98] qui permettent d'obtenir une classification des différentes séquences vidéo disponibles sans prendre en compte des informations de nature contextuelle. Ces systèmes permettent, par exemple, de classer les différentes séquences vidéo en fonction de la scène (intérieure ou extérieure), de la caméra (statique ou en mouvement), etc.

D'autre part, les *systèmes dits spécifiques* [Hollfelder 00] ne permettent d'indexer qu'un type bien particulier des séquences vidéo, comme par exemple les vidéosurveillances, les vidéos contenant du sport (football, tennis, etc.) Dans ce cas, l'indexation est contextuelle car elle est basée sur une problématique précise et le résultat obtenu répond aux attentes des utilisateurs.

Les systèmes spécifiques, même si leur utilisation est limitée à un type de séquence vidéo, permettent cependant de répondre à de nombreuses demandes de la part des utilisateurs de système de recherche vidéo.

Les systèmes spécifiques d'indexation vidéo dépendent donc fortement du contexte défini au préalable. L'indexation des séquences vidéo peut être vue comme une classification binaire de l'ensemble des images. Soit la suite d'images correspond à l'événement prédéfini, soit elle n'y correspond pas. Le problème revient alors à la détection des événements prédéfinis dans des séquences vidéo, ou des scénarios. L'outil idéal dans le domaine de l'indexation selon une problématique donnée serait un logiciel qui fonctionnerait en deux étapes. Après une phase d'apprentissage des séquences vidéo correspondant ou non à l'événement prédéfini, une étape de reconnaissance permettrait de classer chaque suite d'images en deux classes : celles qui correspondent à l'événement prédéfini et celles qui n'y correspondent pas. Cette étape de reconnaissance serait bien sûr basée sur le processus d'apprentissage effectué dans un premier temps.

III.4.2.1 Indexation et recherche d'image

Le processus d'extraction d'une image-clé dans un plan vidéo passe forcément par l'idée qu'une telle image doit capturer le contenu sémantique d'un plan. Il est donc inutile d'entrer dans des calculs compliqués et longs sur toutes les images du plan vidéo. Malheureusement les techniques existantes ne sont pas assez avancées pour déterminer efficacement l'image-clé d'un plan.

Une vidéo peut être considérée comme une grande base d'images auxquelles pourraient s'appliquer les techniques d'indexation par le contenu, développés pour l'image fixe, comme QBIC [Holt 97], VIR [Bach 96], MARS [Hunag 00] pour ne citer que les plus connus. L'application de ces techniques s'effectue souvent sur l'image-clé de chaque plan vidéo vu part les volumes de données sont trop considérables.

Les techniques d'indexation vidéo basées sur le contenu visuel qui sont actuellement proposées choisissent donc de ne conserver qu'une une seule image-clé par plan (keyframe) [Pickering 02] mais en perdant ainsi toutes autres informations telles que par exemple les mouvements des objets. Par conséquent, ceci limite considérablement la portée de l'analyse par contenu d'un document vidéo dans les différentes applications.

L'aspect structure représente aussi une information à exploiter dans le cadre de la modélisation par le contenu du document.

Les techniques pour la recherche d'image sont principalement basées sur l'analyse perceptive du contenu. Les requêtes types sont en général composées d'une ou des plusieurs images exemples dont le système va se servir pour retourner en réponses des images similaires d'un point vue « perceptif » (couleur, texture, etc.). Parmi les systèmes de recherche de ce type, on trouve le système VisualSeek [Smith 96], le système NETRA [Ma 99] et MARS [Huang 00] pour n'en citer que quelques uns. La plupart des systèmes de recherche d'image sont conçus pour soutenir la technique de requête par image exemple. L'utilisateur peut interroger des bases de données images en utilisant ces systèmes, si la requête peut être exprimée en termes d'un ou plusieurs exemples. Les systèmes permettent à l'utilisateur de définir l'importance ou la pertinence des différents attributs d'images. Les attributs typiques incluent la couleur, texture, disposition et forme des objets, etc.

Généralement, ce type de système permet aux utilisateurs d'attribuer l'importance relative aux images retournées en réponses à une requête. Ceci permettra ensuite de raffiner le processus de recherche et d'améliorer la précision du système.

III.4.2.2 Indexation et recherche d'une séquence d'images

Les techniques de recherche d'une séquence d'images suivent une approche semblable à celle de la recherche d'une image. La plupart des techniques se basent sur la technique de recherche par image exemple. Les exemples tels que WebSeek et VisualSeek [Smith 96], aussi bien que d'autres méthodes [Yeung 95], [Mohan 98], [Pickering 02].

La technique de la recherche de séquence d'images à été adaptée pour la recherche par le contenu visuel dans les vidéos. Cette adaptation permet d'avoir comme élément d'indexation des descriptions de couleur, de taille, de position spatiale des objets visuels, et trajectoire de mouvement spécifique. Tous ces descripteurs de bas niveau peuvent former une base pour un index. Pour les mettre en œuvre, il est nécessaire de disposer d'algorithmes de segmentation

temporelle des vidéos et d'extraction d'images-clés, dans les segments produits [Naphade 98], [Yeo 95], [Meng 95], [Patel 97].

Les plans représentent des unités élémentaires de base de la vidéo. Une transition de plan implique un changement physique cinématographie (coupure de la caméra ou changement de scène). Une frontière de plan ne correspond cependant pas forcément à un changement de scène. Une scène ici est définie comme étant un segment vidéo avec une continuité sémantique (unité d'action et de lieu).

III.4.2.3 Indexation et recherche Audio

Les récents progrès constatés dans le domaine de la reconnaissance de la parole et les technologies relatives de traitement de la parole rend possible l'exploitation de ce type de sous média pour l'indexation et la recherche des documents vidéo. À l'heure actuelle, la plupart des travaux proposés pour l'indexation et la recherche vidéo par le contenu audio ont été évalués sur des applications spécifiques (émissions radiodiffusées, journaux télévisés, etc.).

En se basant sur le contenu audio, il est possible d'appliquer un grand nombre de modèles et d'outils d'analyse de l'audio. Comme par exemple la séparation des sources d'informations dans l'audio (parole, musique, bruit) [Kemp 00], [Pinquier 01], la transcription automatique de la parole [Gauvain 02] qui consiste à passer d'un flux de parole (signal) à une information textuelle plus simple à analyser. D'autres techniques de traitement spécifiques à l'audio telles que par exemple la détection de changement de locuteur [Kown 02], [Nam 97], [Srinivasan 00] permettent de décomposer le document vidéo en des portions plus cohérentes. Chaque segment correspond à la parole d'un seul locuteur dans le document.

D'autres propositions pour la segmentation du contenu audio des documents vidéo ont été proposées dans [Naphade 00], [Akutsu 98], [Liu 98]. Le contenu audio d'un document vidéo offre un potentiel énorme pour la capture de la sémantique mais il faut pour cela concevoir des systèmes assez intelligents pour pouvoir l'interpréter. L'analyse de scènes auditives essaie de saisir de l'information dans la piste audio. Deux classes le plus fréquemment utilisées dans l'analyse de scènes auditives incluent la parole et la musique. Naphade et Huang ont employé les modèles de Markov cachés pour classifier les flux audio. Zhang et Kuo [Zhang 00] ont employé des algorithmes basés sur des heuristiques pour la classification audio.

Les systèmes d'analyse audio peuvent être classés dans deux catégories : ceux basés sur des heuristiques et des règles, et ceux basés sur des modèles d'apprentissage [Naphade 00]. La plupart des systèmes commencent par extraire des segments contenant la parole puis appliquant des techniques existantes de reconnaissance automatique de la parole pour générer des transcriptions sous une forme textuelle simple à manipuler.

III.4.2.4 Indexation et recherche Vidéo

Si pour le cas de différents médias, les techniques pour l'indexation et la recherche ne manquent pas et sont variées [Marchand 00], [Hampapur 99], les techniques multimodales sont encore peu nombreuses. La plupart des techniques de recherche multimédia exploitent l'information temporelle dans le document [Nam 97], [Wang 00]. Le projet Informedia [Wactlar 96] utilise le flux visuel pour la segmentation et le flux audio pour la classification du contenu. De tels systèmes existent également pour des domaines visuels particuliers comme les journaux télévisés [Nakamura 97] et les émissions sportives [Zhang 00], [Kobla

00]. Les structures présentées par Yeung et Liu [Yeung 95] utilisent des modèles de Markov pour l'analyse de la parole, la segmentation en histoire, la détection des événements, etc. Un autre domaine qui a été aussi traité dans plusieurs travaux est la détection et la vérification du locuteur [Rodriguez 04] en combinant des caractéristiques audio et visuelles. Le contenu riche en information pose également des défis difficiles pour l'indexation. En effet, cette expressivité sémantique qui s'étend sur plus qu'un seul type de média nécessite des techniques spécifiques à chaque média afin de les traiter efficacement. C'est une raison pour laquelle peu de techniques peuvent prétendre être vraiment multimodales en termes d'utilisation combinée des flux audio et visuel.

Le domaine de l'indexation et de la recherche par le contenu des documents vidéo a également donnée lieu à plusieurs travaux qui ont été proposés dans le cadre des projets de recherche. Nous présentons quelques projets.

Projet SESAME [3]: ce projet a pour objectif d'étudier des solutions nouvelles à la problématique de l'indexation et de la recherche par le contenu de séquences audiovisuelles. Plus précisément, il vise à mettre au point un système de recherche d'information multimédia évolutif avec des capacités d'apprentissage par acquisition incrémentale de plusieurs types de connaissances (stratégiques, épisodiques, etc.). Il devra être utilisable aussi bien par des utilisateurs occasionnels (étudiants, professionnels de l'audiovisuels) que par des utilisateurs expérimentés.

Projet MUMIS [4] : ce projet a pour but principal de développer et d'intégrer des technologies de bases qui supportent l'indexation conceptuelle automatique de données vidéo et permettent la recherche de contenu dans des archives digitales multimédias. Le projet examine de manière plus précise le rôle que des annotations résultant d'analyses linguistiques poussées, combinées avec des informations spécifiques au domaine d'application (ici le football) peut jouer pour indexer de longues séquences vidéo (une rencontre de football). Ce projet est composé de deux composants : l'un « hors ligne », qui est responsable de la génération automatique d'annotations formelles pour l'indexation conceptuelle du matériel vidéo. Cette indexation se fait sur la base d'information temporelle extraite des multiples documents. L'autre « en ligne » qui est responsable de l'accès en temps réel aux archives multimédias annotées par le premier composant. Ici, nous nous concentrons sur la partie hors ligne.

Projet AGIR [5]: est un projet ambitieux et récent, établi entre plusieurs établissements français tels que l'INA, l'INRIA. Il comporte toute la chaîne de traitement des données multimédias : extractions des caractéristiques médias, langage de description multimédias et applications. L'objectif de ce projet est de développer des technologies et des outils nécessaires pour mettre en oeuvre une architecture pour l'Indexation et la Recherche par le contenu de données multimédia, conforme aux exigences exprimées dans le contexte de la normalisation internationale.

---

3 Système d'Exploitation de Séquences Audiovisuelles et Multimédias enrichies par l'Expérience [http://lisisun1.insa-lyon.fr/projets/descrippr21.htm]

4 Multimedia Indexing and Searching Environment [http:// parlevink.cs.utwente.nl/ projects/mumis/index.html]

[5]Architecture Globale pour l'Indexation et la Recherche [http://www.ina.fr/recherche /projets/finis/agir.fr.html]

Projet DICEMAN [6]: c'est un projet européen qui vise à développer un modèle de référence pour l'indexation, la description et l'échange de contenus audiovisuels en se basant sur la future norme internationale MPEG-7, qui a débuté en avril 1998 pour une durée de deux ans. L'objectif principal de DICEMAN consiste à permettre l'échange de contenu audiovisuel sur Internet, et répond donc à un problème majeur auquel sont confrontés les départements d'archives, leurs clients, et de manière plus générale, l'ensemble des détenteurs de contenus.

Projet INFORMEDIA [7] : ce projet a permit la mise en place de nouvelles approches pour l'indexation automatique, la navigation, la visualisation et la recherche vidéo. Ces approches sont intégrées dans un système pour l'usage dans des environnements tels que l'éducation. Il se base sur une combinaison des caractéristiques audio et visuelles afin de d'analyser le contenu du document. il utilise les connaissances du domaine pour repérer les *événements* signifiants dans la vidéo.

Projet FISHCLAR [8]: c'est un projet qui a abouti au développement d'un système de vidéo numérique qui permet d'enregistrer les journaux télévisés selon les préférences d'un utilisateur, avec des techniques avancées (SMS/WAP/PDA) pour la recherche et les résumés des vidéos enregistrés.

Discussion

Les informations présentées par les différentes techniques sont exploitables dans le cadre d'indexation et de recherche d'information multimédia mais des limites existent surtout lorsqu'il s'agissait d'une indexation par le contenu audiovisuel, en effet les applications spécifiques à l'analyse des images sont basées plus sur l'étude de l'aspect physique de l'image que sur son contenu sémantique. Une telle indexation par le contenu audiovisuel dans ce cas est moins précise parce que les techniques utilisées n'interprètent pas le contenu sémantique. Concernant les applications dédiées à l'étude du flux audio, même si le taux de reconnaissance automatique de la parole est élevé, l'indexation vidéo revient à une indexation textuelle par des mots clés. Cette indexation est subjective surtout lorsqu'il s'agit d'un document audiovisuel où l'interprétation physique et sémantique de signal est utile pour localiser un passage précis.

Les principales limites de l'indexation automatique proviennent de l'exploitation d'algorithmes qui utilisent uniquement l'information contenue dans les documents alors que l'interprétation ne peut valablement être faite qu'en utilisant aussi de l'information contextuelle accessible *hors* des documents. Alors que l'interprétation va de la globalité du contexte à la localité du contenu, l'analyse automatique part de l'analyse les unités locales composant le contenu vers la globalité du document. On n'obtient de manière automatique que des descripteurs qui ne reflètent que le contenu *physique* des documents.

---

[6]Distributed Internet Content Exchange using MPEG7 descriptors and Agent Negotiation [www.cordis.lu/ infowin/acts/analysys/ products/thematic/agents/ch3/**diceman**.htm]

[7] Digital Video Library [http://www.informedia.cs.cmu.edu/]

[8] Fischclar project [http://www.cdvp.dcu.ie/aboutfischlar.html]

III.5 Outils et schéma de descriptions

À l'heure actuelle, les initiatives les plus avancées pour la normalisation des outils et des schémas de description proviennent de Dublin Core (DC) [Hunter 99] et du standard MPEG7 [Paek 99].

L'initiative du DC consiste en l'élaboration de champs prévus pour l'entrée de métadonnées servant à la description de documents électroniques, en vue de permettre leur diffusion en réseau. On pense évidemment à Internet et à l'indexation précise que pourraient effectuer les moteurs de recherche présents sur le Web, grâce aux indications facilement repérables contenues dans les métadonnées. Soulignons ici la responsabilité que peut assumer un créateur, en choisissant lui-même les termes d'indexation du document à diffuser, ce qui allégerait encore la tâche de l'indexeur.

Le standard MPEG7 a été établi précisément pour la description du contenu audiovisuel, Il s'agit de définir un "standard de descripteurs vidéo", afin d'accéder directement à des extraits de documents multimédia.

En utilisant les travaux existants sur l'analyse automatique du flux visuel, on pourra extraire automatiquement certains aspects formels de l'image : couleurs, formes, mouvements de caméra, ainsi que des éléments du contenu. Il s'agit donc de mettre en place un langage favorisant plusieurs niveaux de description selon les usages qui peuvent en être faits. MPEG7 prévoit la possibilité d'une indexation basée sur le contenu, réalisée par des documentalistes et en mode textuel. Il semble donc se dégager des tendances qui se dessinent, une part de plus en plus importante attribuée à l'indexation automatique, et qu'on se dirige vers une division du travail entre l'ordinateur et l'indexeur humain. La machine prendrait en charge l'analyse plan par plan, en raison de sa capacité à assimiler un grand nombre de données en peu de temps, et s'arrêterait au contenu des images; l'indexeur humain se concentrerait sur les produits finis, en raison de sa concision et de son esprit de synthèse, et se chargerait de l'interprétation des images.

III.5.1 MPEG7

Depuis octobre 1998, MPEG (Moving Picture Experts Group) [Paek 99] a démarré un nouveau travail pour offrir des solutions au problème de la description de contenu multimédia. Ce nouvel outil, appelé "*Multimedia Content Description Interface*" (Mpeg-7), vise à créer un standard de description des données multimédias qui répondront aux exigences opérationnelles qu'elles soient de nature temps réel ou non temps réel. MPEG7 se veut suffisamment générique pour répondre aux besoins d'un grand nombre d'applications.

La norme MPEG7 a été établie pour la description du contenu audiovisuel, Il s'agit de définir un "standard de descripteurs multimédias", afin d'accéder directement à des extraits de films et de pouvoir les visualiser. MPEG7 s'intéresse à la description du contenu des différentes scènes. Cette description concerne à la fois les éléments visuels (couleur, texture, position) et les informations d'un niveau d'abstraction supérieur telles que le nom d'une personne présente sur la vidéo et son activité.

Le choix des descripteurs, pour la mise sur pied d'un vocabulaire contrôlé, est nécessaire pour éviter le plus possible les divergences d'interprétations. Une telle mesure augmente les chances qu'un utilisateur tire profit de l'exploitation du système.

III.5.1.1 Objectifs et principes de Mpeg-7

Mpeg-7 a pour objectif de décrire les informations audiovisuelles. Ces descriptions doivent permettre en particulier la recherche et le filtrage de données audiovisuelles. Pour ces descriptions, la norme MPEG propose :

- Des *descripteurs* (Ds) qui présentent les parties distinctives ou des caractéristiques des données qui sont significatives pour quelque chose ou quelqu'un (ex : un histogramme de d'intensité lumineuse, la moyenne des composants fréquents, le texte d'un titre, etc.),
- Des *schémas de description* (DSs) qui comportent en particulier des relations entre descripteurs permettant de spécifier des entités de plus haut niveau, pouvant aller jusqu'à un niveau sémantique,
- Un *langage de définition de description* (DDL) qui doit permettre la création des nouveaux schémas de description et de nouveaux descripteurs. Il doit permettre aussi la modification et l'extension des schémas de description et des descripteurs existants,
- Des formats de codage qui permettent de réaliser des fonctions comme la compression efficace, la correction d'erreur, l'accès direct, etc.
- La prise en compte des contraintes système et performance. La Figure III.4 présente les relations entre Ds, DSs et DDL [Paek 99] Permettant de mieux comprendre les principes de conception ci-dessus et les relations entre eux. Des données audiovisuelles à partir des sources matérielles sont spécifiées sous forme de caractéristiques par le système d'observation ou l'utilisateur. Ces caractéristiques sont regroupées en descripteurs, c'est-à-dire un descripteur représente un ensemble des caractéristiques ou au moins une caractéristique. Les descripteurs sont utilisés pour créer des schémas de description. Un descripteur peut appartenir à plusieurs schémas. Un schéma peut être aussi défini à partir d'autres schémas.

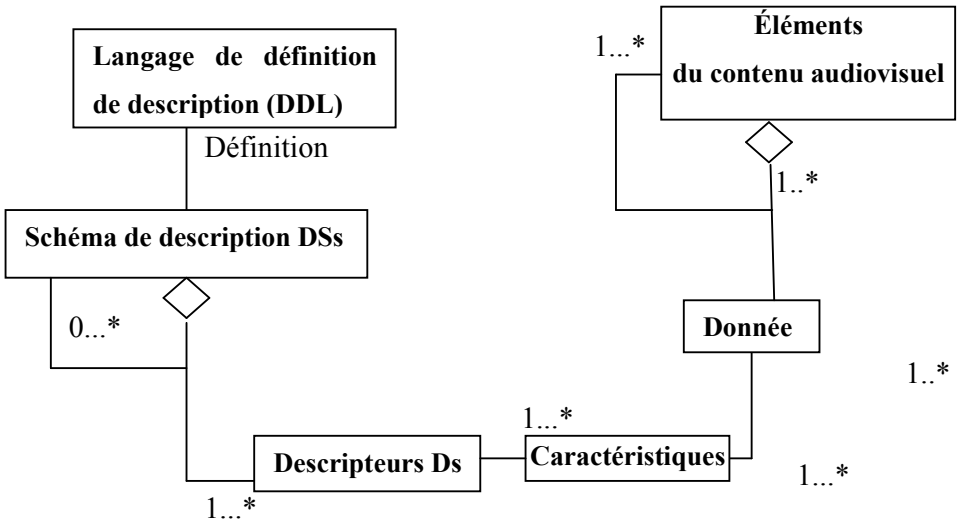


Figure III.4 : Présentation des relations entre Ds et DSs

La figure suivante présente les principes d'une chaîne de traitement utilisant le standard Mpeg7. Cette chaîne inclut une extraction des caractéristiques (analyse), une phase de description, application (un moteur de recherche).

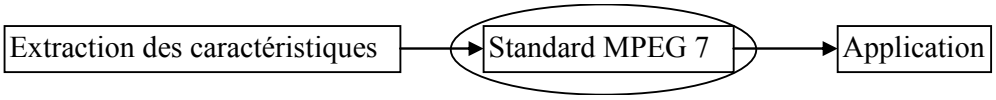


Figure III.5 : Chaîne de traitement vidéo avec Mpeg-7

III.5.1.2 Les parties de MPEG-7

La norme MPEG 7 est composée des parties suivantes :

1. Le MPEG 7 Système : les outils qui sont nécessaires pour préparer les Descriptions MPEG 7 pour le transport efficace et le stockage et permettre la synchronisation entre le contenu en des descriptions. Outils liés à gestion et protection de propriété intellectuelle.
2. Le MPEG 7 Langage de Définition de Description : le langage pour définir des nouveaux Arrangements de Description et peut-être finalement aussi pour de nouveaux Descripteurs. Il permet aussi l'extension et la modification d'Arrangements de Description existants. XML [Bray98] a été choisi pour fournir la base pour le DDL.
3. MPEG7 Audio : Les Descripteurs et les Arrangements de Description traitant seulement les descriptions Audio.
4. MPEG 7 Visuel : les Descripteurs et les Arrangements de Description traitant seulement les descriptions Visuelles.
5. MPEG 7 Arrangements de Description Multimédia : les Descripteurs et les Arrangements de Description traitant avec fonctions génériques et descriptions multimédia.
6. MPEG 7 Logiciel de Référence : Une mise en oeuvre de logiciel des parties appropriées du Norme MPEG 7.
7. Le MPEG 7 Conformité : Des directives et des procédures pour évaluer la conformité de la mise en oeuvre de MPEG 7.

III.5.2 Dublin Core

Dublin Core est un ensemble d'éléments de métadonnées destiné à présenter des ressources du WEB. Il est donc très général et peut s'appliquer en particulier à des données audiovisuelles. Les quinze éléments de Dublin Core ont été étendus de façon à utiliser des sous éléments permettant de créer un schéma de description vidéo. À un niveau élevé ces quinze éléments de Dublin Core peuvent être utilisés pour décrire des informations de nature bibliographique à propos du document (Par exemple, titre, auteur, date, etc.). Une extension de quatre éléments (type, format, relation et traitement) permet de décrire des informations de plus bas niveau.

La figure III.6 (extraite de [Hunter 99]), présente la structure logique, structure des composants et leurs attributs de Dublin Core pour une proposition de schémas de description vidéo.

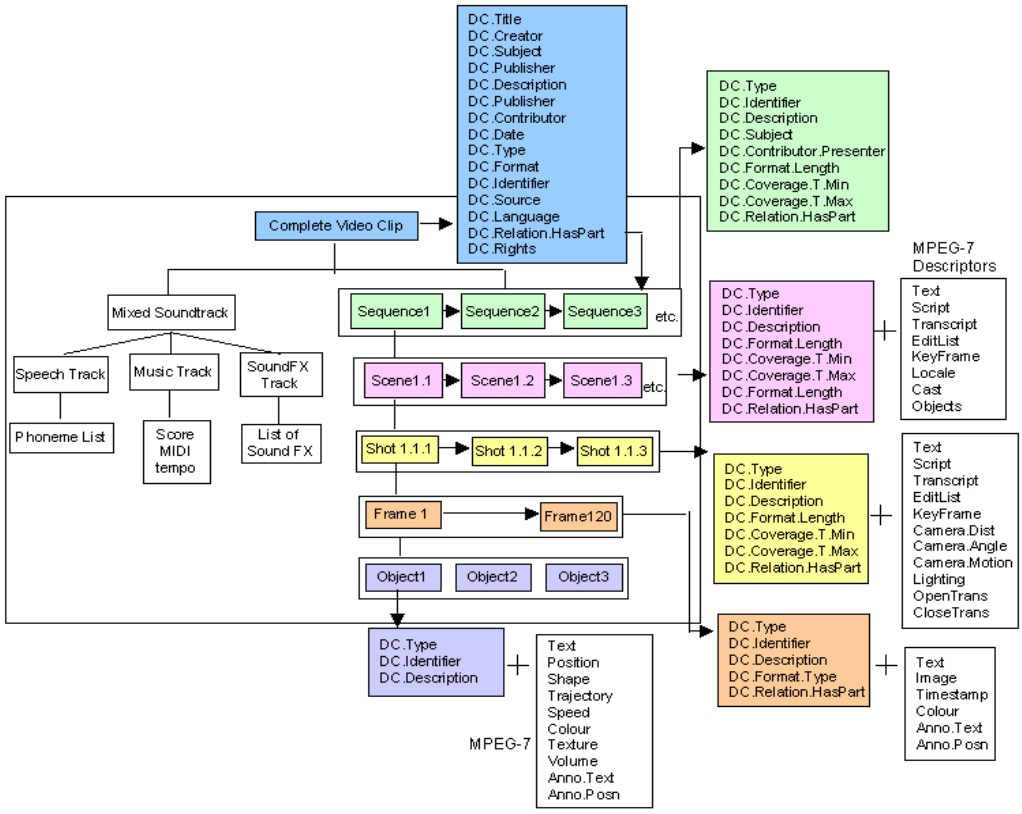


Figure III.6: La structure de hiérarchie et les attributs de la vidéo

Nous décrivons ci-dessous les quatre éléments étendus de Dublin Core pour décrire les vidéos.

1. DC.Type

Des documents d'un film et d'une vidéo peuvent être structurés suivants : *sequences, scenes, shots, frames*. Chaque *sequence* consiste en des *scenes* successives, chaque *scene* consiste en des *shots* successifs, chaque *shot* consiste en des *frames* successives, chaque *frame* peut être divisé en des *régions* représentant des personnages ou objets. La hiérarchie de la structure des documents vidéo est donc :

Par exemple, deux DC.Types, la première spécifie une scène dans une séquence d'un document, et la deuxième spécifie une image dans un plan d'une scène :

DC.Type = "Image.Moving.Film.Documentary.sequence.scene"

DC.Type = "Image.Moving.TV.News.sequence.scene.shot.frame"

2.DC.Format

Cet élément définit le format des données de la ressource. Le format peut être choisi dans une liste qui est actuellement en cours de définition. Par exemple, sont déjà définis :

DC.Format.type = 35mm film, VHS etc.

DC.Format.colour.depth = 256

DC.Format.length = 31 mins.

DC.Format.vidéocodec = MJPEG, MPEG1, MPEG2, AVI, QT, etc.

DC.Format.framerate = 25

DC.Format.sound = Yes/No

3. DC.Relation

Cet élément décrit les relations hiérarchiques de la structure. Il consiste en deux sous-éléments *HasPart* et *IsPartOf* qui sont paramètres par un attribut *Content*. Par exemple, la valeur de *Relation* de la scene3.2 peut être :

Relation.HasPart Content= shot3.2.1, shot3.2.2, shot3.2.3

Relation.IsPartOf Content= sequence3

4. DC.Coverage

Cet élément est utilisé pour décrire la localisation temporelle des *clips, scenes, shots*, etc dans une vidéo. Le format de valeur du temps peut être une durée (*frame number)*, un temps de codage SMPTE ou un temps absolu à partir de début. Par exemple, le moment où une ressource est déclenchée peut être décrit de la façon suivante :

Coverage.t.min scheme=SMPTE content="09:45:23;14"

Coverage.t.max scheme=SMPTE content="09:45:32;1"

De plus, les sous-éléments de Coverage, Coverage.x, Coverage.y, Coverage.z, Coverage.line, Coverage.polygon et Coverage.3D peuvent être utilisés pour décrire des localisations spatiales, des mouvements et des formes pour les objets/personnages.

III.5.3 MXF (Material Exchange Format)

Le format MXF [Ive 04], format de fichier ouvert destiné aux échanges de documents audiovisuels et de leurs données et métadonnées associées, a été conçu et mis en œuvre en vue d'améliorer l'inter-opérabilité entre serveurs, postes de travail et appareils de création de contenu. Ces améliorations devraient se traduire par de meilleurs flux et des méthodes de travail plus efficaces qu'avec les formats de fichiers mixtes et propriétaires d'aujourd'hui. Les principaux intervenants du secteur de la radiodiffusion ont conçu le MXF avec une contribution importante des utilisateurs pour s'assurer qu'il réponde vraiment à leurs besoins. Il se présente sous la forme d'une norme ouverte.

Outre une meilleure inter-opérabilité (une capacité améliorée de travailler avec des fichiers audio et vidéo sur des équipements et des applications différents), l'atout majeur du MXF

réside dans le transport des métadonnées. En le traitant dès le début comme un nouveau format de fichier, ses concepteurs se sont intéressés à la mise en œuvre et à l'utilisation des métadonnées. Cet aspect, non seulement important pour assurer le bon fonctionnement des fichiers MXF, active également de nouveaux et puissants outils pour la gestion du support et l'amélioration des flux de création de contenu en éliminant la réintroduction de métadonnées répétitives.

Utilisation de MXF pour le transfert de métadonnées

L'un des objectifs principaux du MXF est d'assurer un transfert transparent du contenu des programmes et des métadonnées associées. Les métadonnées, ou « données sur les données » comme on les appelle parfois, existent aujourd'hui dans tous les systèmes. Par exemple, le code temporel en est une forme. Le problème réside dans le fait que ces informations sont actuellement perdues lorsque le contenu est transféré d'un système à l'autre à cause des incompatibilités. Les systèmes qui auront adopté le MXF communiqueront en utilisant des métadonnées, de la vidéo et de l'audio. Les métadonnées MXF peuvent acheminer des informations sur :

- ✓ la structure du fichier ;
- ✓ le contenu du corps du fichier (par exemple MPEG ou DV, 525 ou .625, etc.) ;
- ✓ des mots clés ou des titres ;
- ✓ des sous-titres ; - des numéros de référence ;
- ✓ des notes de montage ;
- ✓ le lieu, l'heure, la date et le numéro de la version ;
- ✓ etc.

La figure III.7 présentant un exemple de « *tournage d'un documentaire sur la vie sur la vie des animaux sauvages au Kenya »,* illustre l'utilisation du format MXF pour le transfert des données (métadonnées). Les données GPS (c'est-à-dire les coordonnées géographiques de la caméra) sont ajoutées à chaque plan sous forme d'annotation. Ces métadonnées resteront avec les prises à l'intérieur du fichier MXF pendant la durée du tournage du programme. On peut ensuite, à l'aide d'un processus de production automatique, convertir ces coordonnées GPS en métadonnées supplémentaires lisibles par l'opérateur. Une telle automatisation réduit les tâches ordinaires affectées au personnel et améliore la précision des données stockées.

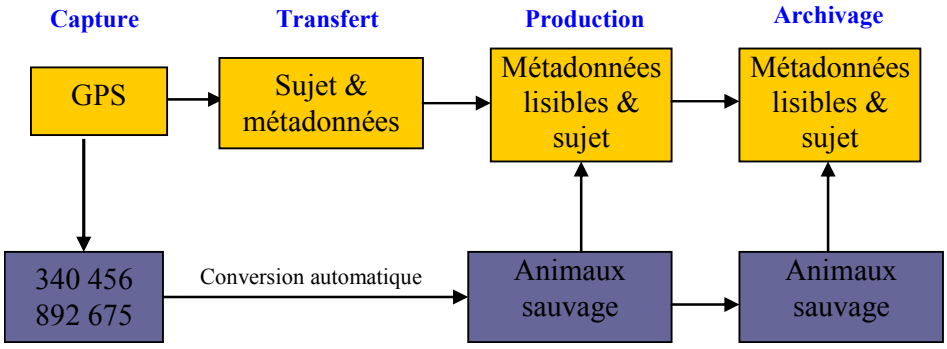


Figure III. 7: Exemple d'utilisation de MXF

Discussion

Différentes propositions de schéma de description d'informations multimédias ont été basées les éléments du Dublin Core ci-dessus pour décrire des documents vidéo.

La conclusion de cette étude est qu'aucun de ces schémas ci-dessus n'est suffisant pour décrire des documents multimédias complexes. Les schémas basés sur Dublin Core ne satisfont qu'un sous-ensemble de besoins de description. Par exemple, il ne peut pas décrire des relations spatiales, temporelles, conceptuelles, etc.

Par comparaison à Mpeg-7, les descriptions proposées de Dublin Core se veulent plus spécifiques et opérationnelles. Bien que ne satisfaisant pas encore à tous les besoins de description d'informations multimédias qui se trouvent dans Mpeg-7 *Requirements*, elles pourraient constituer une première implémentation de Mpeg-7.

Les modèles de description audiovisuelle citée ci dessus ne visent pas tous au même objectif d'où on peut tirer les remarques suivantes :

- Mpeg-7 vise à satisfaire tous les besoins et à être utilisable pour toutes les applications possibles de traitement de l'information audiovisuelle. Pour cela il offre un métalangage de haut niveau.
- Dublin Core propose aussi de couvrir un ensemble de besoins similaire et compatible avec Mpeg-7.

L'ensemble de ces outils et schémas de descriptions ont des limites par rapport à nos besoins. En effet, pour Dublin Core nous pouvons distinguer l'absence des descriptions bas niveau. Les limites les plus importantes sont la synchronisation temporelle et spatiale entre les composants de la vidéo qui est nécessaire à la composition multimédia c'est à dire avec les autres médias (image, texte, son).

Conclusion

Dans ce chapitre, nous avons présenté un panorama des travaux liés aux problèmes de l'accès aux documents vidéo et de leur manipulation. Plus précisément, nous nous somme intéressé aux travaux sur l'indexation et la recherche par le contenu. Nous avons aussi présenté les aspects liés à la structure du contenu vidéo et les outils et schémas de description.

Le problème majeur lié au développement d'un système d'indexation et de recherche des documents vidéo dans un contexte générique consiste à la mise en place des techniques capables de combiner les sous médias pour traiter les documents vidéo. En effet, l'application des études spécifiques à un type de média (image, texte ou audio) telles que proposées dans la littérature est loin d'être efficace pour une application sur un document vidéo. Concernant l'utilisation de l'ensemble des outils et des schémas de description vidéo, ces outils ne permettent pas décrire le contenu et les besoins du document vidéo d'une manière à tenir en compte à la fois de la richesse du contenu (contenu audio, visuel, etc.) et de la diversité des utilisateurs avec leurs besoins et leurs différents points de vue. En effet, les outils et les schémas de descriptions proposées tels que DC ou MPEG7 sont, soit basés sur des descriptions génériques (métadonnées) du contenu vidéo, soit en exploitant des descriptions niveau signal du document vidéo

Une indexation intégralement automatique du contenu vidéo ne peut se réaliser que dans des types de documents vidéo spécifiques (vidéosurveillance, sport, etc.). Pour indexer le contenu

de haut niveau dans un contexte général, l'intervention de l'opérateur humain reste à l'heure actuelle indispensable. Le résultat d'une indexation manuelle du contenu dépend directement du niveau d'expertise de l'opérateur et du temps qui lui alloué pour associer une description à chaque segment vidéo. Généralement, les index se présentent sous une forme textuelle qui peut contenir soit des informations génériques (nom, type de fichier, titre, taille, …) introduites lors de l'acquisition d'un document vidéo, soit des informations sur la structure du document comme par exemple la structure hiérarchique (séquence, scène, plan).

La notion de description sémantique ne coïncide généralement pas avec les méthodes d'analyse des images ou de l'audio. Ceci n'exclut pas l'existence de plusieurs approches « mixtes » qui intègrent les descriptions niveau signal (couleur, texture, mouvement de caméra, etc.…) générée automatiquement avec les descriptions haut niveau issues d'une indexation manuelle ou assistée (concepts) pour l'indexation et la recherche des documents vidéo.

.

Chapitre IV

Modélisation du Contenu des Documents Vidéo : État de l'Art

L'objectif de ce chapitre est de présenter les principaux modèles existants pour la représentation du contenu des documents vidéo ainsi que leurs limites pour décrire efficacement un document vidéo dans un contexte générique.
Ce chapitre est structuré comme suit : nous commençons par présenter quelques notions sur la représentation du contenu. Nous détaillons ensuite quelques approches classiques de modélisation vidéo. La dernière partie du chapitre est consacrée à l'introduction du formalisme des graphes conceptuels. Nous utiliserons ce formalisme comme modèle opérationnel dans le cadre de notre proposition qui sera détaillée dans le chapitre suivant (chapitre V).

IV.1 Introduction

C'est à partir d'une représentation de son contenu qu'il est possible d'exploiter un document vidéo. Une telle représentation est en général réalisée selon in modèle. Pour bien représenter le contenu d'un document vidéo, il est nécessaire d'organiser ce contenu. Un bon modèle de représentation doit permettre cette organisation. Il doit permettre d'intégrer les descriptions issues des différentes modalités (image, audio, texte). Chaque description doit pouvoir servir de base à toute utilisation des documents (indexation, recherche, navigation, etc.).

Un document vidéo est décrit par un ensemble d'éléments constituant un index. La recherche d'un document ou d'une partie de celui-ci repose ensuite sur une recherche dans l'ensemble de ces index et pas directement dans le document initial.

La navigation (autre que séquentielle) repose également sur les index des documents. Dans ce cas, une interface de navigation dépasse les simples capacités de lecture séquentielle du document telle que réalisée par un magnétoscope numérique. Elle permet par exemple une navigation, plan par plan, ou bien une navigation « par concept » (aller directement à la prochaine apparition de la personne X dans le document).

La nécessité de mettre en place des modèles de représentation vidéo résulte de la variété de type des documents vidéo. Cependant, certains types de documents (par exemple journaux télévisés) ont une structure qui leur est propre (sujet, reportages, …). Un modèle trop générique ne sera pas adapté pour la représentation de tous les types de structures spécifiques.

IV.2 Segmentation et description sémantique

IV.2.1 Segmentation temporelle

L'intérêt de la mise en place d'un modèle de représentation consiste à faciliter l'identification et l'annotation des segments (temporels) ayant une unité sémantique. Ces unités sont de granularités variables et résulte d'un découpage temporel qui est éventuellement variables selon la perspective. En effet, des besoins d'informations variés tels

que exprimés par les requêtes sur un concept X : « rechercher les segments vidéos montrant une image de X » et « rechercher les segments vidéos dans lesquels on parle de X », sont susceptibles de produire comme réponses deux unités sémantiques tout à fait différentes selon le média : image (l'unité peut être un plan par exemple) ou audio (l'unité peut être un segment audio).

Comme nous l'avons détaillé dans la sous section II.2.2 du chapitre II, un document vidéo peut être structuré et segmenté de différentes manières selon le média (image, son ou texte) et aussi selon les besoins utilisateurs.

Dans le cas de la piste image, il existe une segmentation « naturelle » qui est la segmentation dite « en plans ». À partir d'un découpage « en plans », il est possible de définir deux types de segmentation supplémentaires :

- ✓ Le premier type est appelé « micro-segmentation ». Elle correspond à une analyse plus fine du contenu d'un plan. Elle consiste principalement à décrire l'intérieur selon les mouvements de caméra dans et / ou par des caractéristiques interprétées du contenu (apparition / disparition d'un objet ou d'une personne).
- ✓ Le deuxième type correspond à l'extraction d'images clés à l'intérieur de chaque plan. Elle constitue une première étape pour la segmentation spatiale de l'image (découpage de l'image en régions, identification des objets) afin de décrire son contenu.

Il existe un autre type de segmentation, plus générique appelé « macro-segmentation ». Ce type de segmentation consiste à partitionner le document en des segments s'insérant (généralement) entre le plan et le document. Ce choix de laisser la notion de « macro-segmentation » vague (absence de définition universelle) est lié à la diversité de types de documents vidéo (journaux télévisés, films émissions sportives, etc.). En effet, le lexique de segmentation peut être varié et parfois spécifique au genre du document en question. On trouve par exemple les notions de scène, séquence, histoire, épisode, mi-temps, thème, etc.

IV.2.2 Exemple classique de segmentation temporelle

La structure la plus classique consiste à décomposer la vidéo en des unités (segments) dont chacune représente un niveau de description. Cette décomposition est similaire à ce qui a été proposé pour la modélisation hiérarchique du contenu vidéo. Tous les éléments de la structuration définis ci-dessous sont repérés par rapport au contenu visuel du document vidéo.

Dans un document vidéo, on distinguera des *séquences* qui correspondent à des unités sémantiques de la vidéo (thème, sujet, etc.). Dans des vidéos de taille importante, plus le contenu est varié plus le nombre de séquences est important. Les séquences peuvent elles-mêmes être décomposées en *scènes*. Ces scènes, sont composées d'un ou plusieurs *plan(s)*. Chaque plan correspond à une prise de vue, avec ou sans mouvement de caméra (plan fixe). Les plans sont séparés par des transitions. La frontière entre deux plans peut être un changement brusque de plan (« *cut* » en anglais), qui est instantané et qui représente 60 à 90% des cas ou bien une transition continue et plus longue (*fondu enchaîné* par exemple).

Dans une vidéo, on peut définir des entités élémentaires (éléments d'information) significatives comme des personnages, des éléments décors, des objets, etc. On peut également définir des objets de granularité plus fine : des parties d'éléments comme le visage du personnage. Ces entités sont regroupées au niveau description du contenu vidéo et sont appelées des entités « *Classe* ».

Chaque apparition d'une classe dans des images consécutives est appelée une *entité visuelle*. Les classes ont souvent des occurrences dans différents plans. Toutes les occurrences apparaissant dans un plan, dans une scène, dans une séquence ou dans la vidéo prise dans son ensemble sont stockées au niveau de la description ce plan, cette scène, cette séquence ou cette vidéo. On peut aussi considérer des micro-segments à l'intérieur des plans dans lesquels les entités visuelles apparaissent ou non. Cette apparition ou non peut même contribuer à définir certains de ces micro-segments.

Dans une vidéo, on peut également définir entités *événement (event)*. Un *événement* est « quelque chose qui arrive ». Il peut être décrit par les classes ou occurrences de classe qui interviennent dans « ce qui arrive » et par des relations qui définissent « ce qui arrive ». Un événement se déroule en général selon un certain intervalle de temps.

Un événement peut être rattaché à un plan, à une scène, à une séquence ou à la vidéo complète (par exemple, le déclenchement d'une explosion, une manifestation, un orage, etc.). On peut aussi rattacher les événements à des micro-segments.

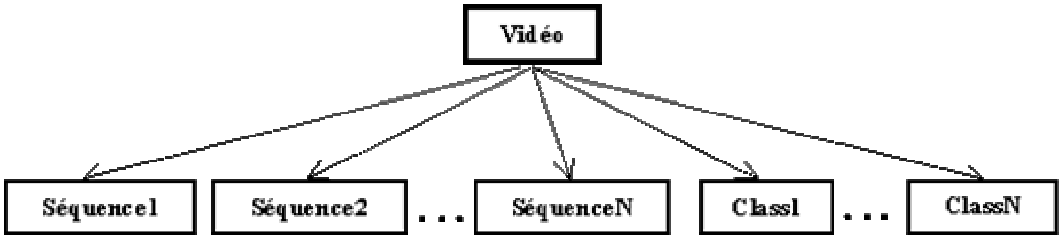


Figure IV. 1 : Structure d'une vidéo en séquences et classes

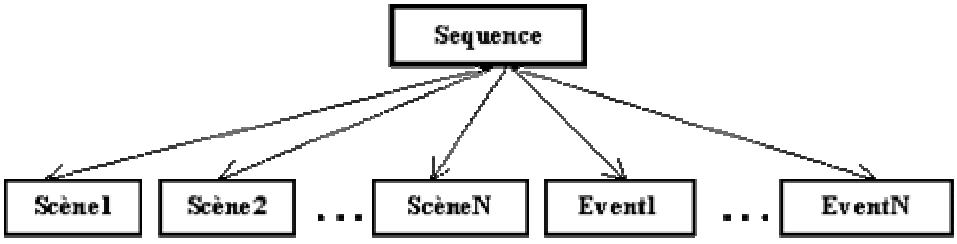


Figure IV. 2 : Structure d'une séquence vidéo en scènes et en événements

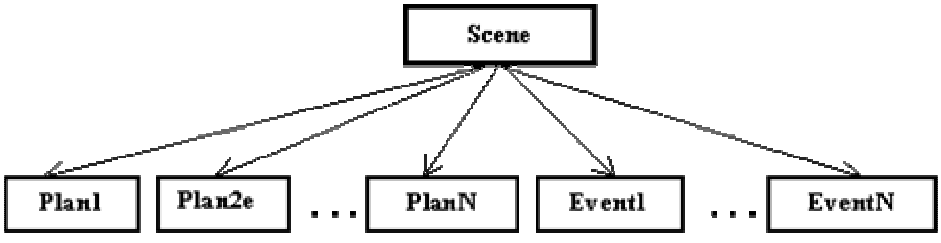


Figure IV. 3 : Structure d'une scène vidéo en plans et événements

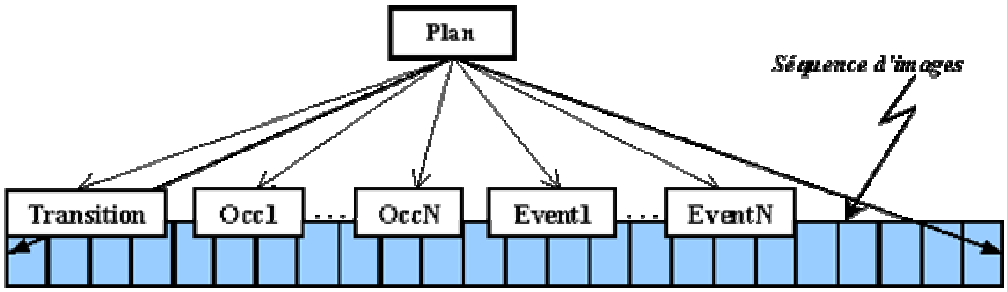


Figure IV. 4 : Structure d'un plan vidéo

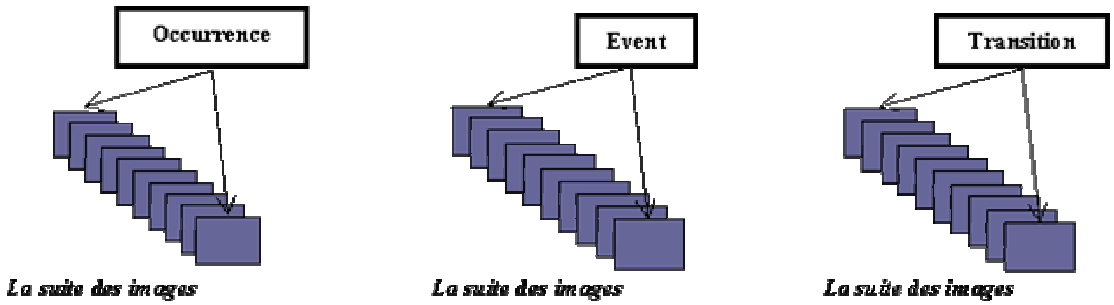


Figure IV. 5: Structure d'une occurrence, d'un événement et d'une transition

IV.2.3 Segmentation Spatiale

Le découpage spatial n'est autre qu'une segmentation du contenu d'une image (segmentation en objets). Cette segmentation plus fine du contenu consiste à partitionner le contenu de l'image en des zones (régions) homogènes (couleur, texture, forme) et/ou correspondant à des objets (ou classes). La figure IV.6 récapitule les deux types de segmentations (temporelle et spatiale).

La segmentation spatiale ne tient compte que du contenu visuel de la vidéo. Elle permet de spécifier les positions entre les objets visuels. Ceci ne permet pas d'assurer une continuité sémantique du contenu vidéo. En effet, en se basant uniquement sur la perception visuelle, une description pourrait être incomplète. Le contenu textuel ou audio dans le document peut par exemple déterminer l'information sémantique (une étiquette pour le cas du texte) de chaque objet identifié lors de la phase de segmentation spatiale.

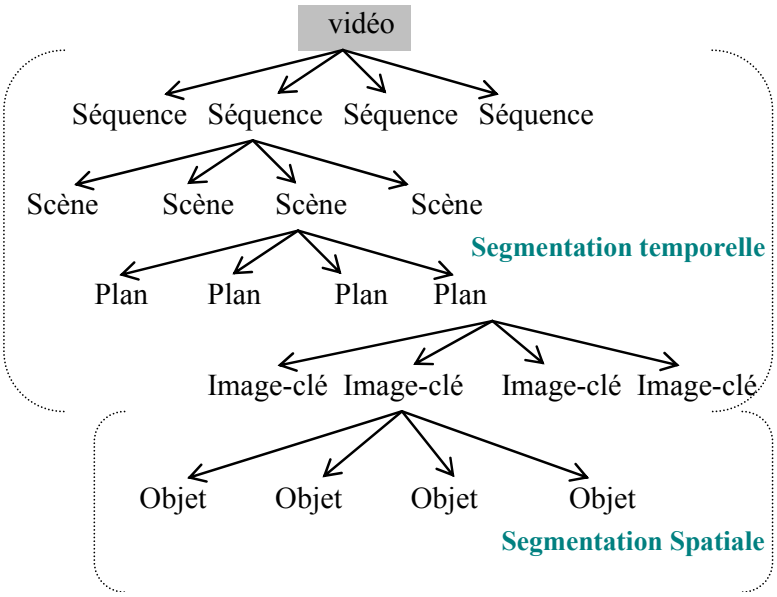


Figure IV. 6 : Exemple de structure spatiale et temporelle d'une vidéo

Il est aussi envisageable de combiner les deux formes de segmentation (spatiale et temporelle). Cette combinaison permet surtout de garder la continuité temporelle du document. Ainsi la segmentation spatiale de deux images proches l'une de l'autre permet de garantir la continuité temporelle. Le lien entre les informations spatiales et temporelles est le plus souvent employé dans un seul sens : on exploite généralement l'information temporelle en premier lieu pour inférer ensuite une description spatiale.

IV.2.4 Description du contenu sémantique de la vidéo

Comme illustré dans la figure IV.7, le contenu sémantique est ce qui a du sens pour un être humain. Il peut en général être décrit par des concepts (classes ou instances) et des relations entre eux. Dans [Lindley 97], Lindley considère que la première tâche de représentation du contenu d'un document vidéo consiste à se donner un ensemble d'unités à même de représenter les flux temporels d'images et de sons, donc de parler (par exemple) d'événements, d'actions et d'objets dans l'espace, de personnages, d'objets impliqués dans des actions, de mise en scène, des positions relatives des objets, de montage, de pensées subjectives à propos du document, etc. Cette approche met en lumière quelques aspects des caractéristiques de haut-niveau :

- ✓ En premier lieu, celles-ci sont le plus souvent des termes, des mots-clé, associés à une partie du document vidéo.
- ✓ En deuxième lieu, les termes sont le plus souvent organisés en catégories correspondant à des niveaux de descriptions. Par exemple, [Amato 98] organisent les descripteurs en objets, lieux, temps, activités et personnes. L'activité est considérée comme centrale pour la recherche, et des liens dans la base de connaissance permettent de mettre en relation des concepts proches, par exemple l'activité *naviguer* sera connectée au lieu *océan* et à la chose *bateau*.

- ✓ En troisième lieu, les différents niveaux de description par des caractéristiques existent : nom d'une personne, lieu, événement, histoire, etc. Par exemple, pour le cas d'un journal télévisé les personnages principaux dans ce type de document sont le présentateur, les reporters et les intervenants. Au niveau interprétation du contenu, un personnage peut être vu, entendu ou mentionné oralement. La combinaison de ces cas est également envisageable. Un aspect plus spécifique concerne le cas des journaux télévisés dans lesquels la fréquence d'apparition du présentateur et des reporters a une importance sémantique.

  Un autre bénéfice attendu concerne la description des actions au sens large. On distingue par exemple les actions filmiques (mouvements de camera et transitions entre plans) et les actions réelles (effectuées par des acteurs).

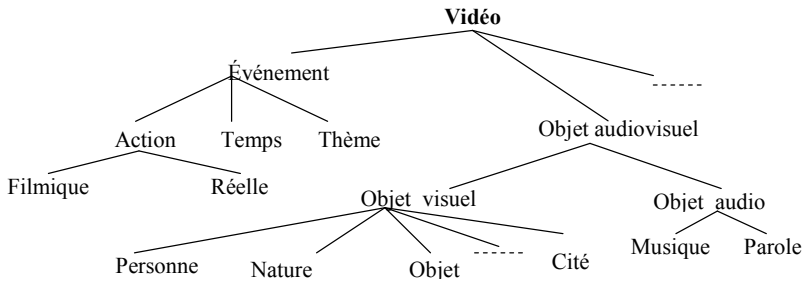


Figure IV. 7 : Description hiérarchique et caractéristiques audiovisuelles

IV.3 Modélisation

Comme pour le cas de document textuel, une vidéo possède une structure spécifique. Cette structure peut être soit homogène du fait qu'elle intègre les différents composants du document ou bien hétérogène lorsqu'il s'agit d'une analyse par média spécifique.

La modélisation d'un document vidéo dans son ensemble, c'est-à-dire la prise en compte de la structure et du contenu, est destinée à accorder une vue synthétique au document de manière de donner à l'utilisateur un aperçu (idée) sur le document. Elle permet de faciliter l'accès au contenu des documents. On peut distinguer trois approches de modélisation des documents vidéo :

- ✓ la modélisation *hiérarchique :* permet d'associer une description sous d'arborescence du contenu du document. Cette forme de modélisation est souvent liée à la segmentation temporelle;
- ✓ la modélisation en *strates* ou « stratification » : associe des annotations aux documents ou aux parties du document vidéo. La stratification est indépendante de la segmentation temporelle du document;
- ✓ la modélisation par *objets* permet de décrire le contenu avec des objets audiovisuel. Elle prend souvent la même structure que la modélisation hiérarchique.

Nous constatons que le rôle de ces différentes approches de modélisation des documents (*en strates*, *hiérarchique* ou *objets)* est de permettre de structurer le contenu d'un document vidéo

pour faciliter son utilisation dans diverses applications telles que la recherche d'information par exemple. Elles ne sont pas exclusives mais complémentaires.

IV.3.1 Modèle hiérarchique

Ce type de modélisation intervient généralement à priori (lors du montage vidéo). Une modélisation hiérarchique donne lieu à une représentation de la vidéo selon une structure arborescente dont la racine est le document vidéo. Une première phase consiste à découper récursivement le document vidéo en des unités plus petites, en général jusqu'au niveau du plan. Une deuxième phase consiste à choisir dans chaque plan une ou plusieurs image(s) clé(s), ces images étant ensuite décomposées en régions ou en objets visuels. La Figure IV. 8 illustre un exemple de modélisation hiérarchique d'une vidéo.

La modélisation hiérarchique de documents vidéo se base généralement sur le plan comme unité élémentaire et sur l'organisation de ces plans en unités de plus haut niveau sémantique, telles que les scènes. Celles-ci peuvent alors être regroupées au sein du document, dans d'autres unités appelées séquences. La structure du document est alors une structure arborescente similaire à une structure documentaire textuelle classique (chapitre, section et paragraphe). La structure arborescente est souvent conçue suivant une hiérarchie document / séquence / scène / plan / image-clé / régions mais d'autres types de hiérarchies sont également possibles. Des caractéristiques de niveau sémantique (classes, événements) ou de niveau signal (descripteurs) peuvent être associées aux différents niveaux de l'arborescence.

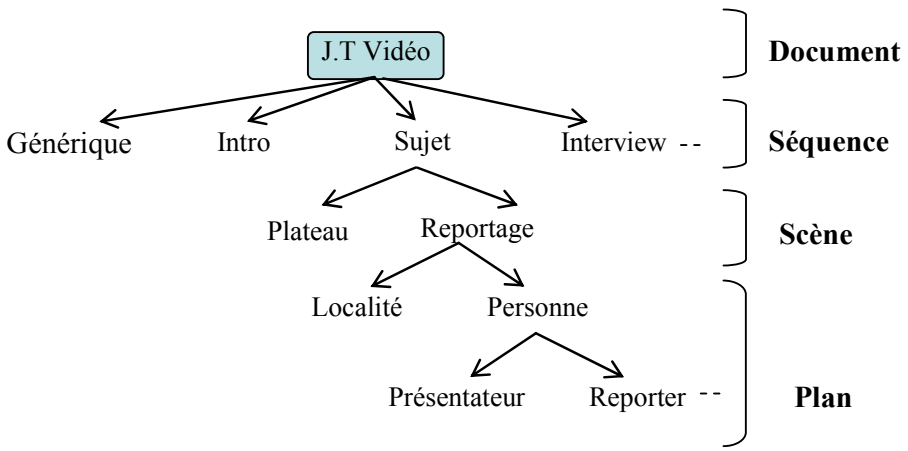


Figure IV. 8 : Modélisation hiérarchique de la vidéo

[Fatemi 99 et al] proposent une structure hiérarchique pour la représentation par le contenu des documents vidéo. Cette structure se base sur une modélisation sous forme de graphes et prend en compte les descriptions temporelles dans le document.

[Corridoni 96 et al] attachent aux différents segments de leur structure hiérarchique filmique un certain nombre de caractéristiques : les caractéristiques générales (auteur, date, titre) sont associées au document dans son ensemble, des descriptions textuelles aux scènes. Les plans sont, quant à eux, caractérisés par des primitives techniques (mouvement de caméra, angle, profondeur de champ).

IV.3.2 La Stratification

La stratification [Prié 99] a été proposée pour décrire le contenu d'une vidéo. Elle consiste à associer a un document vidéo ou à un segment de document des « niveaux d'annotations » (ou strates). La Figure IV. 9 illustre ces différents éléments avec un graphe « Strates-IA » [Prié 99] correspondant à l'annotation d'une interview du cycliste Sandy Casar.

L'avantage d'une telle description est la possibilité de donner une information précise sur une partie de document vidéo tout en offrant une description contextuelle utile pour la compréhension du contenu de cette partie. L'approche de stratification se base sur le principe qu'à toute annotation correspond une définition du segment vidéo annoté. Une description d'un document vidéo dans ce modèle est réalisée par un ensemble de graphes de description éclairés par une base de connaissances.

L'élément de base de Strates-IA proposé par Y. Prié est l'annotation primitive, qui consiste à associer une caractéristique interprétée à un segment de flux vidéo. Ainsi le modèle définit l'*unité audiovisuelle* (UAV) comme une entité abstraite. Une UAV existe dès lors qu'on lui attache un *élément d'annotation* (EA) qui est l'expression d'une annotation. L'élément d'annotation est en *relation d'annotation* (R_a) avec l'UAV. Cet élément est une instance d'un concept. Il est mis en relation avec un élément d'annotation abstrait (EAA). Les différents EAs peuvent être reliés entre eux par une relation générique dénommée relation élémentaire (R_e)

Le contenu doit donc être décrit de la façon la plus globale possible dans le document. Par exemple, un simple descripteur vidéo tel qu'une information textuelle (mot-clé) attachée à au concept personne peut apparaître dans plusieurs plans et on ne doit pas suivre les limites syntaxiques imposées par la segmentation c'est à dire qu'un élément d'annotation peut repérer un segment vidéo dont la dimension dépasse celle fixée par le découpage physique de la vidéo (plan / scène etc.…).

La stratification permet le mélange des considérations différentes sur des morceaux de vidéo identiques. Elle permet de lier les annotations (l'information textuelle utilisée pour la description) et le segment annoté. Un document annoté est alors un ensemble de strates avec leurs annotations. Une structuration des strates entre elles n'existe alors pas a priori. La stratification se caractérise par son indépendance par rapport à toute segmentation, ce qui permet de définir des strates liées à tous les niveaux d'analyses du document que ceux-ci soient consacrés à l'image ou au son.

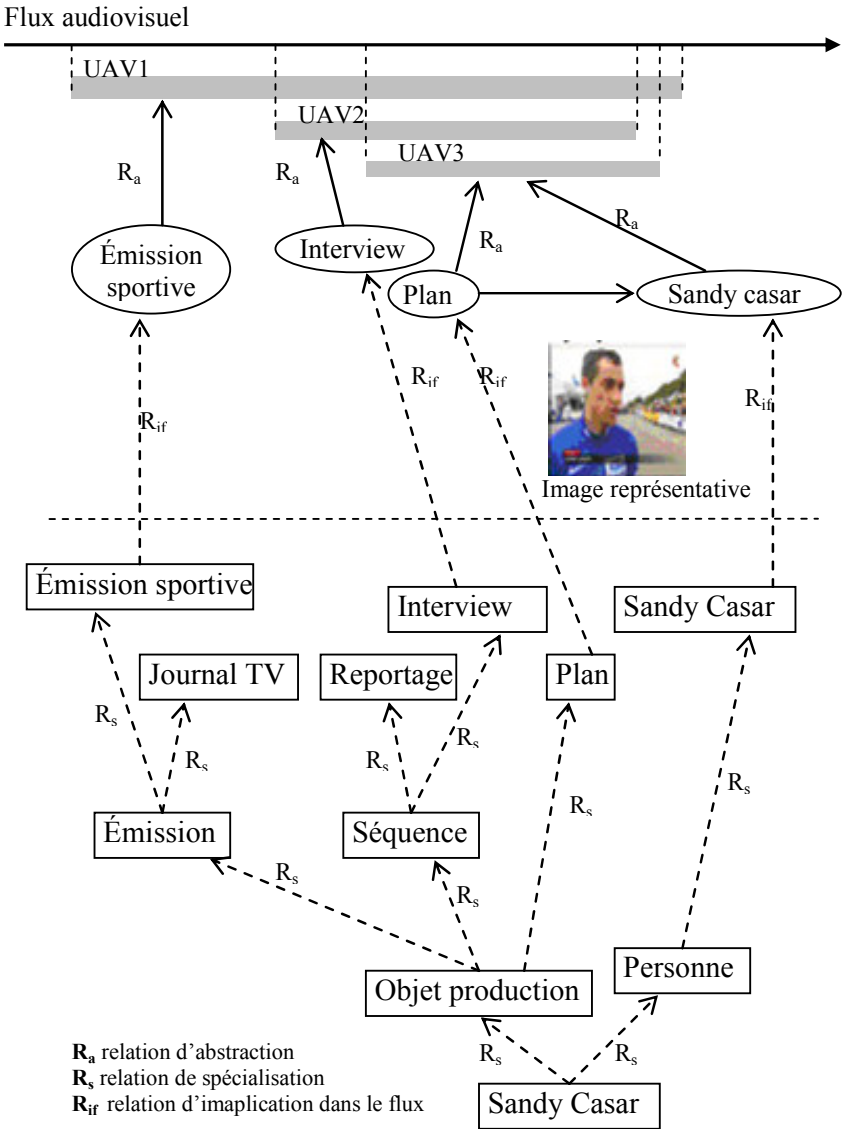


Figure IV. 9 : Structuration des annotations dans le modèle Strates-IA

IV.3.3 La Modélisation à base d'objet

L'élément de base pour décrire un document textuel est le « terme » : mot ou groupe de mots. Chacun des termes apparaissant dans une collection de documents est en général caractérisé par une mesure statistique ou probabiliste de sa fréquence d'apparition dans les documents et dans la collection complète. Dans un document vidéo, la description du contenu se fait à l'aide des objets audiovisuel. Chaque objet audiovisuel est représenté par une description symbolique et des descripteurs (indices) signal.

La modélisation à base d'objets permet surtout de décrire le contenu sémantique du document vidéo. Par exemple, le contenu de la piste image n'est plus considéré comme un tableau de pixels (analyse bas niveau) mais comme un ensemble d'objets (analyse conceptuelle) ayant chacun leurs propres caractéristiques visuelles [Seyrat98]. Cette description a notamment

pour objectif une meilleure structuration de la représentation des vidéos et ainsi permettre par exemple une navigation à différents niveaux de granularité (la vidéo, les scènes, les plans, les objets, etc.). La modélisation du contenu basée sur les objets est nécessaire pour la manipulation du contenu et aussi pour la segmentation spatio-temporelle [Zhong 97] du document.

IV.4 Formalisme de représentation et base de connaissances

Comme nous l'avons mentionné précédemment, l'objectif de la modélisation des documents vidéo est de permettre une exploitation efficace de ceux-ci quelle que soit l'application envisagée. Si, par exemple, cette application est l'indexation et la recherche par le contenu, il est nécessaire que le formalisme de représentation doit permette la définition d'une opération pour le calcul de la pertinence des réponses par rapport à un besoin d'information exprimé sous la forme d'une requête par un utilisateur.

Un formalisme de représentation doit mettre en place une représentation (linéaire ou sous forme des graphes) qui permet de décrire le document (structure et contenu). Un tel formalisme doit être facilement compréhensible par les êtres humains et avoir une capacité de représentation suffisante pour pouvoir être utilisé dans les diverses applications visées. Notre proposition utilisera le formalisme des graphes conceptuels ainsi que des ontologies pour la représentation des connaissances pertinentes dans les contextes visés. Ceux-ci sont présentés et étudiés dans les sections suivantes.

IV.4.1 Les Graphes conceptuels (GCS)

IV.4.1.1 Généralité

Le formalisme des graphes conceptuels est un modèle de représentation de connaissances [Sowa 84]. C'est un ensemble des graphes bipartites, orientés et connexes. Les deux types de nœuds représentent le concept et la relation conceptuelle.

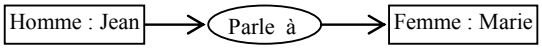
Les nœuds de type concept sont représentés par un couple comprenant un label de type et un référent. Par exemple : [Personne : Bill] désigne un concept de type « personne » et identifié par son nom « Bill ». Les référents peuvent être des constantes correspondant des identifiants uniques ou bien indéfinis désigné par (*) correspondant à un concept générique.

Les nœuds relations sont représentés par les types de relations utilisées. Par exemple, la relation spatiale « à gauche de » peut relier les deux concepts suivants :

[table : *] → (à gauche de) → [personne : V. Hugo]

Une relation est reliée au moins à un concept tandis qu'un concept peut être isolé. Un concept est constitué d'un type de concept (une idée ou représentation de l'esprit qui résume une multiplicité d'objets) et d'un référent. Ce dernier correspond à tout marqueur qui désigne un élément particulier dont le type est celui en question.

La figure suivante montre un exemple de représentation sous forme graphique et sous forme textuelle avec le formalisme des graphes conceptuels.



[Homme : Jean] → (Parle_à) → [Femme : Marie]

Figure IV. 10: Exemple de représentations avec les GCs

Les concepts et les relations conceptuelles sont classés dans des treillis dits *treillis de concepts* et *treillis de relations*.

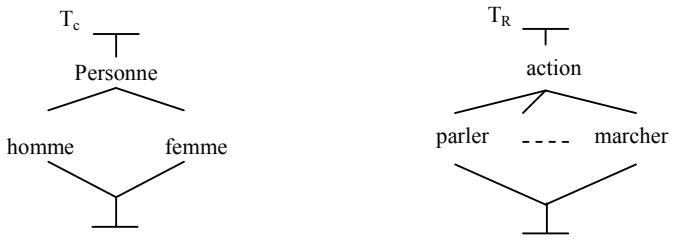


Figure IV. 11 : (a) Exemple de treillis de concepts, (b) exemple de treillis de relations

IV.4.1.2 Notations

Nous rappelons dans cette section, quelques notations de base utilisées dans le formalisme des graphes conceptuels

Un graphe conceptuel g est défini par la donnée d'un triplet (**C**(g), **R**(g), **L**(g)), où :

- **C**(g) est l'ensemble des concepts figurant dans le graphe g;
- **R**(g) est l'ensemble des relations reliant les concepts dans g;
- **L**(g) est un sous-ensemble de **C**(g) × **C**(g) × **R**(g) qui donne pour chaque couple de concepts les relations qui les relient dans le graphe g;

Un concept c est représenté par un couple, c = [type(c) : Référent(c)], où :

- type(c) = t, est le type du concept, t∈ T, T étant l'ensemble des types définis dans le contexte considéré.
- Référent(c) = i est le référent du concept c, i ∈ I, I étant l'ensemble des identifiants représentant les entités du monde réel qu'on veut modéliser par des concepts.

Une relation est représentée par son type, r ∈ R où R désigne l'ensemble des types de relations conceptuelles utilisées.

L'ensemble des concepts et des relations sont partiellement ordonnés par une relation de type spécifique / générique (≤). Les types de concepts et les types de relations sont extraits d'une hiérarchie de types organisée en treillis (nommés respectivement treillis de concepts et treillis de relations). Ces treillis sont limités au sommet par le type universel ⊤ et à la base par le type absurde ⊥. Dans notre définition des graphes conceptuels, le treillis correspond à une structure arborescente.

Les graphes conceptuels peuvent aussi avoir d'autres graphes comme référents. Le graphe inclusif est alors défini comme étant le contexte du ou des graphe(s) inclus. Graphiquement, cela s'exprime en traçant un cadre autour des graphes inclus dans le contexte.

IV.4.1.3 Quelques définitions [Mechkour 95]

Dans [Mechkour 95], un graphe conceptuel canonique et un canon sont définis comme suit :

Définition 1 : ***un graphe conceptuel canonique*** est un graphe qui exprime une combinaison de concepts et de relation conceptuelles. La notion de graphe canonique a été introduite pour limiter la production des graphes conceptuels aux seuls graphes pouvant être interprétés comme ayant un sens dans le contexte de l'application.

Définition 2 : ***un canon*** est une représentation qui permet de définir une base de graphes conceptuels dans un domaine particulier. Il est défini par (T, I, C, B), où :

- ✓ T est le treillis de types de concepts et de relations.
- ✓ I est l'ensemble des instances représentant les entités du monde réel qu'on veut modéliser par les graphes conceptuels.
- ✓ C est la relation de conformité qui lie les éléments de T aux éléments de I
- ✓ B est la base canonique, dans laquelle tous les labels de types sont des éléments de T et tous les référents sont soit *, soit des éléments de I.

Plusieurs règles fondamentales permettent de manipuler de manière cohérente les graphes conceptuels [Sowa 84] :

Règle 1: copie d'un graphe :

Elle permet de copier un graphe dans un ou plusieurs autres graphes.

Règle 2 : Restriction des graphes

La restriction remplace dans un graphe canonique chacun des types par un de ses sous-types ou la valeur de son référent par un plus précis.

Règles 3 : jointure deux graphes :

Nous appelons un concept *commun à deux graphes* un concept *c* figurant dans ces deux graphes. Étant donné un concept *c* commun à deux graphes *a* et *b*, un graphe *d* peut être construit en supprimant *c* dans *a* et en rattachant au concept *c* dans le graphe *b* toutes les relations « pendantes » de *a*.

Règles 4 : Simplification des graphes :

Lorsqu'une relation *r* relie deux concepts *c1* et *c2* plus d'une fois, nous disons qu'il y a des « relations redondantes ». Un graphe peut être simplifié en supprimant les relations redondantes, laissant seulement une occurrence de la relation *r* entre *c1* et *c2*.

Règles 5 : Jointure Maximale :

Un *joint maximal* de deux graphes est obtenu en joignant ces graphes par des sous-graphes compatibles maximaux des deux graphes. Les sous-graphes compatibles sont les plus grands sous-graphes des deux graphes qui ont une restriction en commun.

Règle 6 : Spécialisation :

La restriction et la jointure sont des spécialisations parce qu'elles additionnent les informations des graphes initiaux. Si un graphe *u* peut être retiré d'un graphe *v* par une séquence de restriction et peut être joint avec d'autres graphes conceptuels, alors *u* est appelé une *spécialisation* de v.

Règle 7 : Généralisation :

Si *u* est la spécialisation de *v* alors *v* est la généralisation de *u*. La généralisation est l'inverse de la spécification.

Règle 8 : Projection :

Nous disons qu'il existe une *projection* d'un graphe *b* sur un graphe *a* s'il existe un sous graphe *a'* de a qui est une restriction de *b*. L'opération de projection est l'opération de base du modèle. Une projection d'un graphe *G1* dans un graphe *G2* peut être vue comme la recherche d'une fusion de l'information représentée par *G1* dans *G2*. On appelle alors le graphe *G2* comme étant une spécification de *G1*.

La figure ci-dessous présente un exemple de représentation avec le formalisme des graphes conceptuels.

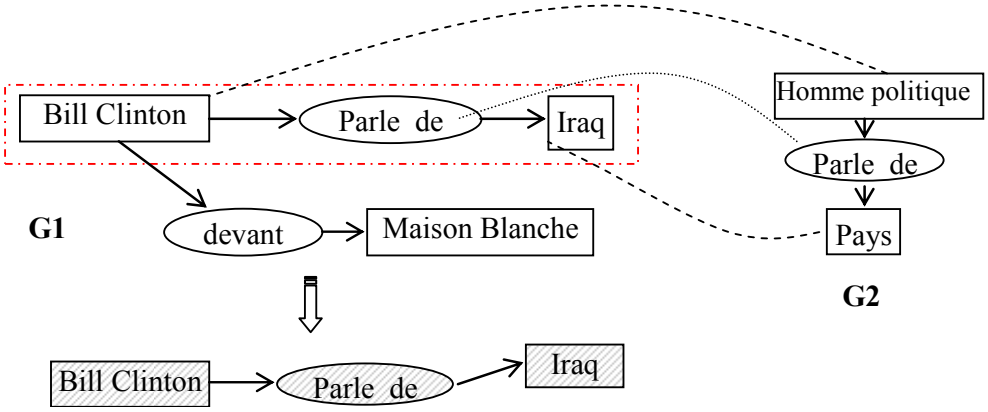


Figure IV. 12: Exemple d'opération de projection de GCs

Deux concepts faisant référence au même objet individuel ou au même ensemble sont dits co-référents. Dans la représentation graphique, le lien de coréférence est exprimé par une ligne pointillée reliant les concepts en question.

La dernière chose que l'on dira à propos des graphes conceptuels est qu'il en existe une forme linéaire qui permet de les mettre facilement en entrée d'un programme informatique. Cette notation linéaire nécessite l'introduction de variables pour mettre en évidence les co-références.

IV.4.1.4 Synthèse

Le formalisme des graphes conceptuels est un formalisme de représentation de connaissance permettant la représentation de relation entre concepts. L'une des particularités de ce formalisme est de permettre la représentation de connaissance sous forme graphique.

Cet aspect graphique nous semble très important car les éléments (concepts et relations) du modèle peuvent être directement et clairement présentés à un utilisateur. Nous pensons en effet que des graphes conceptuels peuvent être facilement crées, compris et manipulés par des personnes qui ne sont pas informaticiens et qui ne connaissent pas l'aspect technique du modèle de représentation de connaissances.

Par « être en relation », on entend l'existence d'un lien sémantique quelconque (par exemple, un même personnage intervient dans deux scènes différentes). Avec cette définition, il est même possible d'étendre la notion de relation aux composants du document audiovisuel. En effet, la structure hiérarchique du document est en elle-même considérée une sorte de description relationnelle (appartenance, inclusion, composé de, etc.…).

La modélisation conceptuelle permet de représenter les choses à l'utilisateur d'une manière différente. Ceci ne veut pas dire que le contenu ainsi décrit soit forcément différent.

IV.4.2 Ontologies

IV.4.2.1 Généralité

Une *ontologie* est un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être :

- des relations sémantiques ;
- des relations de composition et d'héritage (au sens objet).

La structuration des concepts dans une ontologie permet de définir des termes les uns par rapport aux autres, chaque terme étant la représentation textuelle d'un concept.

Les ontologies ont pris récemment une place importante dans le processus d'indexation et de recherche d'information par le contenu ([Zarri 88], [Minghong 99], [Dechilly00]) parce qu'elles aident à donner un sens à l'information. Le terme « ontologie » est emprunté du domaine de la philosophie où il signifie l'étude sur l'existence de l'être en tant qu'être.

Dans le contexte de l'intelligence artificielle, il n'existe pas une définition commune du terme ontologie [Guarino 97], [Gandon 02]. Une des raisons évoquées est que les ontologies se retrouvent dans plusieurs champs d'étude : ingénierie des connaissances, conception de base de données, représentation des connaissances, système à base de connaissance et recherche d'information. La popularité des ontologies provient de ce qu'elles permettent : une compréhension commune et partageable d'un domaine qui peut être communiquée à des humains et des ordinateurs. La définition la plus souvent rencontrée est celle de Gruber : "une ontologie est spécification explicite de la conceptualisation". Une conceptualisation est une vision simplifiée du monde que l'on veut représenter. Guarino, dans [Guarino 97], souligne l'ambiguïté du terme conceptualisation qui doit être pris dans son sens intuitif et propose la définition suivante pour tenir compte du caractère subjectif : "une ontologie est un compte-rendu explicite et partiel de la conceptualisation".

Cependant ce qui importe le plus à propos des ontologies ce ne sont pas leurs définitions mais de savoir à quoi elles servent. Le but, lorsqu'on utilise des ontologies est de permettre aux connaissances d'être partagées et réutilisées. Les ontologies peuvent servir à plusieurs fins, [Mizoguchi96] ont répertorié certains usages.

- ✓ Utilisation en tant que vocabulaire commun pour des communications entre des agents distribués.
- ✓ Utilisation comme schéma conceptuel pour les bases de données.
- ✓ Utilisation comme référence pour des bases de connaissance
- ✓ Utilisation pour répondre à des questions
- ✓ Standardisation (terminologie, concepts)

Il existe plusieurs classifications des ontologies. Tout d'abord on peut catégoriser une ontologie selon le type de langage utilisé pour la construire. Une ontologie est formelle si elle est décrite par un langage artificiel défini de façon formelle par opposition à une ontologie informelle qui utilise des langages naturels. Dans ce mémoire nous ne traiterons que d'ontologies formelles.

Nous pouvons également distinguer le type des ontologies selon leurs objets, ainsi il existe :

- Des ontologies génériques, également appelées « ontologies supérieures », elles contiennent des concepts généraux définies (autant que possible) indépendamment des domaines d'application et qui peuvent être employées dans différents domaines d'application. Le temps, l'espace, les mathématiques, par exemple, sont les exemples des concepts généraux.

- Des ontologies de domaine. Elles sont consacrées à des domaines particuliers et contiennent des concepts spécifiques à ce domaine qui peuvent être employés et réutilisés pour des tâches particulières dans le domaine. Il existe par exemple des ontologies spécifiques pour les domaines des produits chimiques ou de la médecine.

- Des ontologies d'application. Elles contiennent la connaissance consacrée à une tâche particulière, obtenue auprès des experts de l'application. En général elles sont peu réutilisables.

- Des ontologies de représentation ou « méta-ontologies » : Ce sont des ontologies qui définissent les concepts utilisés par les langages de représentation des ontologies.

IV.4.2.2 Ontologies et Recherche d'Information.

Dans cette section, nous étudions l'apport des ontologies pour la recherche d'information (RI). Dans le cadre général que nous considérons est le suivant (*voir figure IV.13*) :

- Un utilisateur exprime son besoin d'information au moyen d'une **requête**.
- Le Système de Recherche d'Information (SRI) met en relations les éléments de la requête de l'utilisateur avec ceux de l'**index** des documents, pour fournir à l'utilisateur un ensemble de documents répondant à ses attentes.

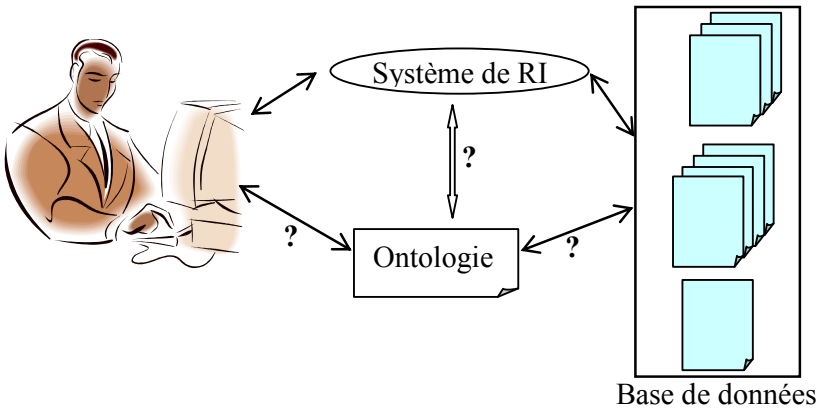


Figure IV. 13: Exemple d'utilisation d'ontologie pour la RI

Ce cadre général couvre l'utilisation d'un système de recherche d'information. Pour ceux qui ont déjà utilisé des moteurs de recherche sur le Web, ils ont certainement été confrontés à des réponses volumineuses contenant des documents n'ayant aucun rapport avec leurs attentes. On comprend dès lors que l'amélioration des performances de ces moteurs de recherche en termes d'efficacité et de précision soit un souci pour les concepteurs de systèmes de recherche d'information. Les ontologies constituent pour cela une technologie jugée prometteuse. Nous allons voir que différentes utilisations des ontologies peuvent être envisagées.

Un des problèmes de l'indexation des documents par des mots, ou groupes de mots, est qu'elle ne prend pas en compte les phénomènes de la langue que sont la synonymie et l'homonymie. Lorsque des mots différents renvoient à un même sens, l'indexeur (personne qui indexe) privilégie en général un seul de ces mots, celui-là même qui apparaît dans le document. Si l'utilisateur emploie dans sa requête un autre mot pour viser le même sens, il n'accède pas aux documents pertinents. La synonymie engendre du silence documentaire (le système a un mauvais rappel). Lorsque des mots identiques renvoient à des sens différents, l'effet inverse se produit. L'emploi du mot par l'utilisateur se traduit par une réponse contenant des documents traitant de l'ensemble des sens possibles… même s'il n'est intéressé que par un seul d'entre eux. L'homonymie engendre du bruit documentaire (le système a une mauvaise précision).

Pour pallier à ces problèmes, une solution consiste à indexer les documents par des concepts, tout en conservant les liens : concept/terme(s). Cette solution se situe dans l'axe des pratiques des documentalistes qui exploitent déjà des thésaurus - des ensembles structurés de termes - pour indexer des documents. Le thésaurus est remplacé par une ontologie, un ensemble structuré de concepts.

Pour montrer l'apport d'une indexation conceptuelle, nous distinguons, d'une part, l'exploitation des liens : concept/terme(s), et d'autre part, l'exploitation des relations sémantiques structurant l'ontologie. Dans ce qui suit, nous considérons que l'utilisateur continuera à employer des mots dans ses requêtes.

Grâce à l'utilisation des liens : concept/terme(s), nous pouvons attendre les deux comportements suivant du système de recherche d'information (SRI) : premièrement, dans le

cas de synonymie, si on suppose que l'utilisateur formule sa requête en employant le mot « politicien », le SRI lui retourne également les documents indexés par le terme « homme politique ». Deuxièmement, dans le cas d'homonymie, si l'utilisateur emploie le mot « pavillon », le SRI demande alors à l'utilisateur si ce mot désigne pour lui une maison, un drapeau etc., puis il lance la recherche des documents indexés par la notion retenue par l'utilisateur.

L'exploitation de la structure de l'ontologie dans le cadre d'un SRI permet la prise en compte des relations existant entre concepts pour la recherche d'information. Par conséquent, on peut attendre du SRI d'autres comportements tels que : si l'utilisateur désire accéder à des documents traitant de l'Europe, en suivant des liens de la relation : spécifique/générique, le SRI peut lui retourner des documents traitant de la France, l'Allemagne, l'Italie, *etc.*, les différents pays composant l'Europe.

En résumé, l'apport d'une ontologie dans le cadre d'une indexation conceptuelle est double :

- Elle permet à l'utilisateur de faire usage d'une autre terminologie (plus précise) que celle présente dans les documents.
- Elle apporte à l'utilisateur une aide pour reformuler ses requêtes sur la base d'une proximité sémantique.

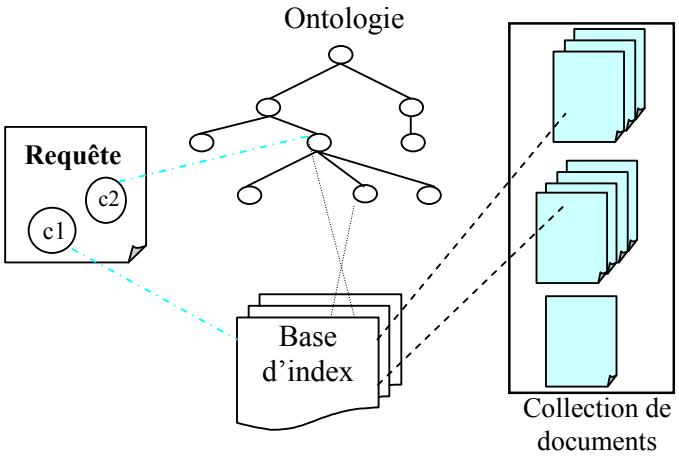


Figure IV. 14 : Utilisation d'ontologie pour le processus d'indexation

L'enjeu d'utiliser une ontologie dans le cadre d'un processus d'indexation et de recherche est de permettre à une large communauté d'utilisateurs de partager la signification de ses éléments, de façon à les utiliser d'une manière cohérente, les utilisateurs étant en premier lieu des êtres humains, et non des machines.

Voyons maintenant comment l'ontologie est exploitée. Lors de processus d'indexation et de recherche, l'ontologie est utilisée pour réaliser des inférences. Nous illustrons ce rôle sur l'exemple suivant :

Requête : [USA_président *x] ; réponse : *x ⊂ {politicien}

Le lien d'inclusion entre les concepts USA_président et politicien permet d'inférer que président est aussi un politicien.

Requête : [USA président *x] ; réponse : *x ⊂{Bill Clinton}

Pour la requête ci-dessus, deux inférences sont possibles. D'une part, comme USA_président a_pour_nom Bill Clinton et que « USA_président » est un politicien. On déduit que Bill Clinton est un politicien. D'autre part, comme Bill Clinton est une Personne qui est un politicien, on peut déduire le concept USA_président utilisé comme requête, désigne une personne.

IV.4.2.3 Synthèse

Toute communauté met en place un vocabulaire commun qui peut être vu comme la base d'un modèle du domaine caractérisant cette communauté. La communication au sein de cette communauté (ou discipline) se fait en conformité avec ce modèle. Le modèle du domaine définit le vocabulaire ainsi que la signification et les relations qui unissent les termes. C*hoisir un mot clé* et *construire une requête* sont des *tâches de recherche de documents* mettant en œuvre de telles relations. On nomme souvent ces modèles « ontologies ». Le terme n'est pas tout à fait exact, mais le mot *ontologie* est bien entré dans les moeurs comme : « ensemble structuré de concepts et de relations reconnus comme existants dans le domaine ».

Une ontologie fournit un vocabulaire, des termes et les relations pour modéliser le domaine. Puisque les ontologies visent la connaissance consensuelle du domaine, leur développement est souvent un processus coopératif faisant participer plusieurs personnes.

Il est extrêmement difficile de structurer un vocabulaire, de créer des ontologies de façon unique. Même dans des domaines bien définis tel que la médecine par exemple, les experts sont rarement d'accord en ce qui concerne les concepts bien établis de leur domaine. En effet, si par exemple deux personnes, dans le cadre d'une tâche donnée, utilisent chacun une ontologie, pour que ces deux personnes puissent communiquer il faut que leurs ontologies se recouvrent suffisamment pour que la compréhension soit assurée. En communiquant, les individus adaptent leur vision, leur interprétation du monde; ils agrandissent ainsi la partie commune de leurs ontologies locales.

IV.5 Conclusion

Nous avons présenté les approches de base pour la modélisation du contenu vidéo. Ces modèles s'appuient sur des structures et des représentations différentes notamment les structures temporelles et spatiales du document. Pour cette raison nous avons aussi détaillé les propositions pour la structuration du contenu d'un document vidéo.

Nous avons introduit brièvement l'existence d'autres modèles de représentations qui visent la description du contenu sémantique. Parmi ces modèles, on trouve ceux qui se basent sur des approches de modélisation objets, la modélisation des actions ou des acteurs.

Les approches de modélisation décrites dans ce chapitre ne permettent de représenter et de décrire le contenu que partiellement. Ce manque d'expressivité de description se traduit par le manque d'exploitation des relations conceptuelles pour associer les différentes descriptions (concepts). Mise à part l'approche de modélisation en strates qui propose quelques relations simples et spécifiques qui permettent de regrouper les annotations sous formes graphique, les approches existantes n'utilisent que de relations de type composition ou appartenance, liées à la structure du document.

.

Proposition & Expérimentations

Chapitre V

Ce chapitre détaille la modélisation formelle de notre proposition. Nous présentons l'ensemble des points liés au modèle proposé. Ce schéma comporte une description multifacette et multimodale du contenu vidéo.
Nous avons choisi de structurer ce chapitre en quatre parties. La première est consacrée à la spécification des aspects génériques liés au modèle notamment les descriptions temporelles et événementielles. La deuxième détaille les descriptions liées aux représentations spécifiques (par média) du document vidéo. Dans la troisième partie, nous présentons la mise en place des ontologies de domaines spécifiques permettant d'enrichir les descriptions formulées dans le modèle proposé. Enfin, la quatrième et dernière partie de ce chapitre porte sur l'utilisation de ce schéma dans un contexte d'indexation et de recherche vidéo par le contenu.

Partie I - Modélisation Multifacette et Multimodale : Représentation générique

V.I.1 Introduction

Dans la littérature, plusieurs approches de modélisation vidéo ont été proposées. Ces approches mettent en œuvre des techniques de modélisation du contenu vidéo. On trouve des propositions telles que VIDEOTEXT [Jiang 97] et VideoGRAPH [Tran 00]. Celles-ci se basent principalement sur l'analyse du contenu visuel par la technique de modélisation en Strates [Prié 99], [Chua99] et aussi sur le principe d'annotation par mots clés pour représenter le contenu sémantique. L'intérêt de la prise en compte des relations sémantiques se situe dans le fait qu'elles peuvent considérablement améliorer l'efficacité de recherche la vidéo fournissant un traitement basé sur la connaissance de la requête. Ceci est dû au fait que les êtres humains utilisent souvent des expressions multiples pour l'interprétation des séquences vidéo similaires. Par exemple, le « sport » et le « base-ball » ne s'assortissent pas syntaxiquement mais s'assortissent conceptuellement. En outre, des associations sémantiques visuelles peuvent être employées pour l'analyse visuelle rapide, flexible et basée sur la connaissance. Les approches basées sur la description du contenu vidéo sont la plupart du temps statiques. Quelques autres approches basées sur la description des caractéristiques de bas niveau de la vidéo telles que le mouvement de caméra [Fablet 00], les transitions entre les plans [Quénot 01] ou les histogrammes de couleur ne sont pas destinées directement à des tâches de modélisation vidéo au niveau sémantique. Cependant, en raison des limitations de la méthodologie utilisée dans ces approches pour la modélisation du contenu sémantique, aucune d'entre elle n'a été proposée pour ordonner les segments vidéo retrouvés.

La plupart des systèmes existant présentent des limites au niveau de l'exploitation des descriptions symboliques pour les différents types de média (image, audio et texte) dans le document vidéo. Pour remédier à celles-ci, nous proposons un modèle de représentation intégré et multimodal décrivant les caractéristiques audio et visuelles. Ce modèle est soutenu par un formalisme de représentation de connaissance.

Dans ce chapitre, nous détaillerons les différents points reliés au modèle que nous avons proposé. Nous rappelons que l'objectif principal consiste en la conception d'un modèle de description des documents vidéo qui soit à la fois multimodal et générique.

Ce travail s'inscrit dans un contexte d'indexation et de recherche vidéo par le contenu sémantique. Il est donc nécessaire qu'avant d'aborder la description du modèle, nous essayions en premier lieu de définir l'ensemble des besoins que notre système de recherche d'information multimédia doit satisfaire. Ces besoins peuvent être regroupés selon deux aspects :

- ✓ Un aspect utilisateur. Celui-ci doit pouvoir spécifier un besoin d'information par l'intermédiaire d'une requête. Ce besoin doit se traduire d'une manière flexible et précise tout en tenant en compte des différents types de média (texte, image et du son), des différents niveaux d'analyse (signal, sémantique) et également du contexte. La modélisation des requêtes constitue une étape commune à toute opération de recherche d'information. Une telle recherche nécessite des méthodes efficaces basées sur la description du contenu. Les requêtes peuvent être formulées en utilisant les différents types de médias (image, texte, audio).
- ✓ Un aspect système qui consiste à traiter efficacement les requêtes et à présenter efficacement les résultats. Ceci nécessite une bonne modélisation des documents ainsi qu'une fonction de correspondance indispensable à tout système de recherche d'informations. Cette dernière mesure la similarité entre requêtes et documents en prenant en compte l'ensemble des index.

La richesse et l'expressivité sémantique d'un document vidéo font apparaître des problèmes liés notamment à la complexité des requêtes utilisées qui résulté de la diversité de besoins d'informations recherchés (audio, visuel, textuel, thématique, etc.).

Prenons l'exemple des requêtes TRECVID [Kraaij 04], nous constatons que dans la même requête plusieurs interprétations sont possibles. « Rechercher les plans vidéos dans lesquels apparaît Bill Clinton qui parle de l'Irak et où au moins une partie du drapeau américain visible »

Dans cet exemple, nous remarquons que pour répondre à ce type de requête, plusieurs médias (ici visuel et audio) doivent être pris en compte. Nous pouvons donc constater dans le cas des vidéos une requête peut prendre plusieurs formes de descriptions telles que par exemple :

Une description objet : ce type de requête est utilisé pour la recherche des objets contenus dans chaque vidéo et satisfaisant les conditions données. Un objet vidéo est tout élément d'information interprété par l'utilisateur (personne, objet, etc.). Un exemple de requêtes contenant une description objet : « rechercher les segments vidéos dont lesquels une personne X apparaît ».

Une description spatiale : ce type de requête est utilisé pour la recherche de vidéo en utilisant les propriétés spatiales entre les objets, ces propriétés peuvent être classées selon trois catégories : les relations directionnelles pour décrire l'espace 2D, les relations topologiques pour décrire les voisinages entre les objets (recouvrement, égalité,..) et les relations directionnelles (sud, nord, etc.).

Une description temporelle : ce type de requête est utilisé pour spécifier l'ordre des occurrences temporelles. Ce type de description peut aussi combiner les requêtes décrit la

structure physique du document. Un exemple : « rechercher les segments vidéos de type plan et qui ont une durée de 3 sec. »

L'étude de ces différents besoins nous amène à la proposition d'un modèle générique qui soit efficace et capable de supporter cette variété de besoins d'information. Pour cela notre modèle sera basé sur une description de contenu et de la structure.

V.I.2 Modèle de base : EMIR²

Nous proposons dans la section suivante une description de l'architecture du modèle EMIR² [Mechkour 95]. Ce modèle combine différentes interprétations (facettes) de l'image pour construire une description complète de son contenu. Chaque interprétation est considérée comme une « facette » (vue) particulière de l'image. L'ensemble des facettes est classé sous deux niveaux de description: le niveau logique et le niveau physique.

- ✓ Au niveau physique, une image est constituée d'une matrice de pixels. Des objets images peuvent être définis par des régions dans cette matrice. L'image et les objets images peuvent être complétés par un certain nombre de signatures ou de descripteurs au niveau signal (couleur, texture, forme, …).
- ✓ Le niveau logique est constitué de l'ensemble des vues décrivant le contenu de l'image; une interprétation complète de l'image se fait par la combinaison des interprétations partielles de chacune de ces vues. La description du contenu de l'image sous forme de graphes conceptuels consiste à relier chaque objet à son interprétation. Ce lien est repéré par un ou plusieurs éléments du graphe se référant à ces objets. Notons que le formalisme choisi pour représenter l'ensemble de ces facettes est le formalisme des graphes conceptuels de Sowa [Sowa 84].

La figure V.1 illustre la description du contenu de l'image en utilisant les interprétations partielles fournies par les différentes vues (facettes).

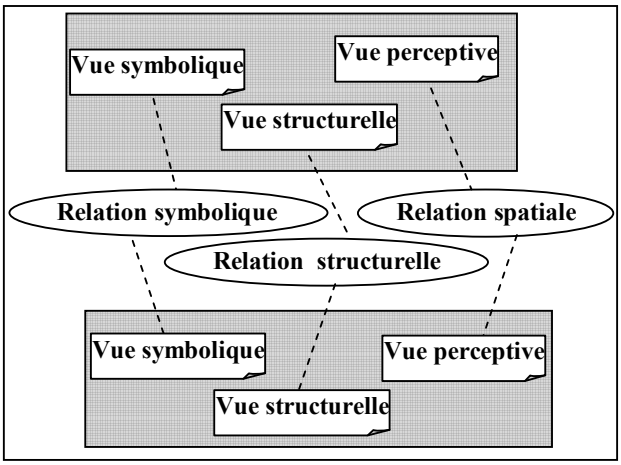


Figure V. 1 : Description d'une image dans le Modèle EMIR²

EMIR^2 suggère trois niveaux de caractérisation :

- ✓ L'image est représentée comme un objet physique ou matrice de pixels.

- ✓ L'image est considérée comme la représentation d'un ensemble d'objets géométriques. Ces objets sont définis par leurs contours et liés entre eux et à l'image par des relations de composition.
- ✓ L'image est caractérisée par un ensemble de descripteurs sémantiques de son contenu. C'est le point de vue symbolique.

L'exemple ci-dessous décrit une représentation du contenu de l'image avec le formalisme des graphes conceptuels.

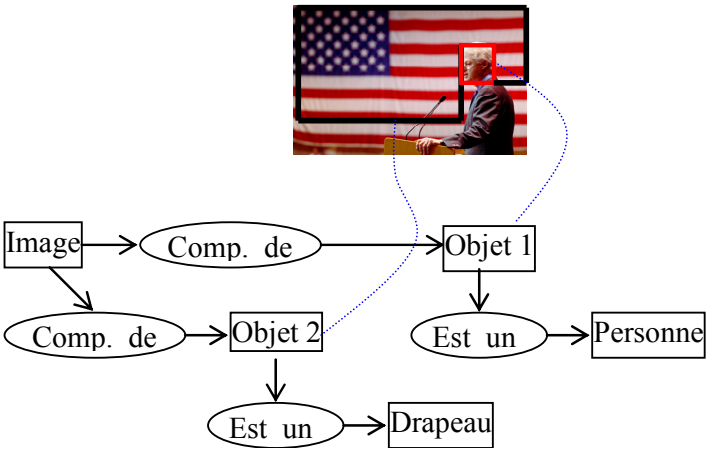


Figure V. :

Figure V. 2 : Exemple de représentation d'images avec le formalisme de GCs

La spécification du modèle de données général permet de représenter une image comme un objet complexe, élaboré à partir d'éléments qui correspondent à des facettes prédéfinies. Cet objet correspond à l'agrégation de différentes facettes dont la pertinence aura été évaluée dans un contexte d'application donné.

V.I.2.1 Spécification de la modélisation multifacettes

V.I.2.1.1 L'objet image

Une image physique est subdivisée en un ensemble de régions représentant son contenu. Une région est par conséquent elle-même un objet physique véhiculant une information indispensable pour la modélisation du contenu global de l'image.

Chaque région est représentée par une abstraction dite « Objet image » dont le référent physique est la région elle-même.

V.I.2.1.2 La notion de facette (ou vue)

À chaque facette, définissant la correspondance avec un type du contenu informatif sur l'ensemble des objets images, est associé un modèle décrivant les objets images, les relations qui les lient et les opérations définies sur ces descriptions.

Les facettes spécifiées sont complémentaires dans un souci de correspondance avec les interprétations qu'elles modélisent et chaque instance du modèle général est une combinaison de facettes pour traduire la richesse des caractéristiques pertinentes des images.

Deux catégories de facettes principales se distinguent : la facette physique qui représente l'entité perçue par l'oeil humain dans sa représentation plane et bidimensionnelle, et la facette logique rassemblant les interprétations de l'image et l'ensemble de ses descripteurs plus sémantiques. Cette dernière se subdivise elle-même en quatre facettes dont la combinaison fournit la caractérisation symbolique de l'image : les facettes structurelles, spatiales, symboliques et perceptives. Toutefois, la facette perceptive est seulement mentionnée, elle n'est pas traitée dans EMIR^2 et fait l'objet de notre étude dans la partie suivante.

V.I.2.2 Les facettes du modèle EMIR²

V.I.2.2.1 La facette perceptive

La facette perceptive regroupe l'ensemble des attributs visuels associés à l'image entière mais aussi aux différents objets image tels que la couleur, la texture, la forme etc. Un des principaux avantages dans l'utilisation des ces attributs réside dans leur caractère objectif est dans la possibilité de les calculer de façon automatique.

Dans EMIR², trois attributs sont pris en compte dans EMIR² : la couleur, la texture et la brillance.

- ✓ La couleur : peut être défini par rapport à un espace de couleur particulier. Il existe plusieurs espaces de couleur (RVB, CMYB, …)
- ✓ La texture : elle est définie comme le graphique qui couvre la surface d'une image ou d'un objet.
- ✓ Brillance : elle mesure la luminosité moyenne de l'image ou d'un objet image.

V.I.2.2.2 La facette structurelle

La facette structurelle représente la décomposition d'une image en objets images. Chaque objet image peut également être décomposé en sous-objets images. La relation de composition associée à cette facette est la relation « contient » qui implique l'inclusion spatiale, les régions correspondant aux objets composants sont incluses dans les limites géométriques de la région décrite par l'objet décomposé.

La facette structurelle est représentée par un graphe conceptuel dont les noeuds sont les objets images et les arcs les instances de la relation de composition.

V.I.2.2.3 La facette spatiale

La facette spatiale décrit les informations géométriques relatives aux objets spatiaux associés aux objets images ainsi que les relations spatiales entre eux. Cette facette permet de caractériser les objets d'une image par leurs formes et leurs positions relatives. Un objet spatial est défini par la donnée d'une forme géométrique (point, segment, polygone) correspondant à son contour.

La facette spatiale est représentée dans les espaces de classiques afin de conférer au modèle une plus grande généralité. Ces espaces sont tout d'abord l'espace euclidien regroupant les

notions de produit scalaire, orthogonalité, angles et normes. Cet espace permet les opérations telles que le calcul du barycentre, de la surface, de la longueur, de la largeur, de la hauteur et du polygone englobant.

L'espace topologique exprime les notions de continuité et de connexion illustrés par les relations Couvre, Touche, Coupe, Dans, Disjoint, Intersection. Cet ensemble de six relations regroupe tous les cas de relation spatiale entre deux objets est invariant par rapport aux transformations géométriques de base.

L'espace vectoriel regroupe les relations de direction fondées sur des notions géographiques (« Nord », « Sud », « Ouest », « Est »).

Enfin l'espace métrique définit les opérations permettant le calcul des caractéristiques géométriques des objets telles que la distance, la distance minimale, la distance normalisée et la distance maximale ainsi que les relations les situant les uns par rapport aux autres. Celles-ci sont au nombre de deux : « Proche » et « Loin » et sont pondérées par une valeur de l'intervalle [0,1] qui mesure le degré d'éloignement des objets spatiaux considérés.

V.I.2.2.4 La facette symbolique

La facette symbolique correspond à la représentation du contenu sémantique d'une image et se définit comme la donnée d'objets symboliques associés aux objets images ainsi que par des relations correspondant à la description de scènes ou d'actions faisant intervenir ces objets.

La facette symbolique s'efforce de prendre en compte les interprétations multiples quant à la sémantique véhiculée par l'image. Elle est fortement contrainte par l'application dans la mesure où l'expression du « sens » est liée à la connaissance du domaine de l'application ainsi que d'un langage d'indexation choisi pour exprimer les relations entre les éléments de connaissance mis en lumière.

Dans EMIR^2, un objet symbolique est décrit par un attribut via un treillis de classes muni d'une relation d'ordre partiel $\leq$. Il est aussi exprimé par un ensemble de couples <propriété, valeur>, le champ propriété correspondant à des propriétés telles que le nom d'un patient et son numéro de dossier dans le contexte des applications médicales.

Les relations symboliques sont des relations binaires associées à des couples de classes de descripteurs R(c1,c2). R décrit donc dans une image une relation entre les objets symboliques o1 et o2 appartenant respectivement aux classes cl_1 et cl_2 si et seulement si $cl_1 \leq c_1$ et $cl_2 \leq c_2$.

V.I.2.2.5 Formalisme de représentation

La modélisation du système de recherche d'images implique la mise en exergue d'un formalisme de représentation des composants du système : requêtes, documents et fonction de correspondance entre requêtes et documents auxquels s'ajoute éventuellement une base de connaissances.

C'est le formalisme des graphes conceptuels [Sowa 84] qui est retenu dans EMIR^2 en raison de son adaptation aisée à l'approche logique de la recherche d'informations notamment dans le cas des documents multimédia [Mechkour 95], [Martinet 04]. Le point de départ est la prise en compte de l'opérateur de projection des graphes conceptuels comme fonction de correspondance.

V.I.2.3 Modélisation EMIR²

Nous développons dans cette partie le modèle d'image ainsi que les modèles correspondant aux facettes de l'image dans le formalisme des graphes conceptuels. Le modèle d'une facette inclut la définition de tous les éléments intervenant dans la composition de la facette pour l'ensemble des images d'une application particulière. Il est considéré ainsi que ses instances comme les composants du canon d'EMIR² et les instances sont elles-mêmes représentées par des graphes canoniques.

Pour chaque type de facette est défini un canon qui est composé du treillis des types de concepts, du treillis des types de relations et d'une base canonique minimale pour contrôler la construction des sous-graphes de la facette.

V.I.2.3.1 Modèle et représentation de la facette structurelle

Le modèle de la facette structurelle est défini comme le couple (Ioi, CONT) avec :

- ✓ Ioi : ensemble des objets images de la facette structurelle.
- ✓ CONT $\subseteq$ Ioi $\times$ Ioi : relation de composition entre les objets images.

La facette structurelle introduit un seul type de concept (ObjetImage) et un seul type de relation conceptuelle (COMP) formant ces treillis de types. Le graphe d'une vue structurelle est construit en liant des concepts de type ObjetImage et dont les référents sont les identifiants d'objets images par la relation conceptuelle de type COMP. Le graphe de base générant par jointure les graphes des facettes structurelles est le suivant :

[ObjetImage]→(COMP)→[ObjetImage]

V.I.2.3.2 Modèle et représentation de la facette spatiale

Le modèle de la facette spatiale est défini par le 6-uplet (Isp, POINT, OS, RSPA, forme, Rsp) avec :

- ✓ Isp : ensemble des identificateurs des objets images de la facette spatiale.
- ✓ POINT = N $\times$ N : ensemble des coordonnées de points.
- ✓ OS : ensemble des objets géométriques de base pour représenter la forme des objets.
- ✓ RSPA = {loin, près, est, ouest, nord, sud, dans, disjoint, touche, couvre, coupe, …} : ensemble des relations spatiales.
- ✓ forme : Isp → P(OS) associe à chaque identifiant des objets spatiaux aux objets géométriques de base.
- ✓ Rsp $\subseteq$ RSPA $\times$ Isp $\times$ Isp : ensemble des relations spatiales qui lient les objets spatiaux dans l'image.

Les types d'objets géométriques représentant les formes possibles pour les objets spatiaux sont représentées par des types de concepts et organisés dans un treillis. Les relations spatiales sont elles représentées par des relations conceptuelles et organisées en treillis. Les objets spatiaux et leur forme sont représentés par des concepts sous-types d'Objet-Spatial dont les référents sont leurs identifiants.

L'instance de la facette spatiale d'une image est alors représentée par un graphe conceptuel composé des concepts correspondants aux objets spatiaux et qui sont liés par les relations spatiales.

V.I.2.3.3 Modèle et représentation de la facette symbolique

La facette symbolique est définie par rapport au modèle sémantique d'une application. Celui-ci regroupe les classes d'objets et leur treillis, la relation de composition entre ces classes, les relations symboliques et les propriétés.

Le modèle sémantique d'une application M est défini par le 7-uplet (IDcl, IDpr, IDrs, VAL_PROP, PROP, RSYMB, domaine) avec :

- ✓ IDcl : ensemble des identificateurs des classes d'objets (relation d'ordre partielle, élément minimal m et élément maximal M). Chaque classe de IDcl est représentée par un type de concept. La relation d'ordre partiel entre les classes d'objets images est représentée dans le treillis des types de concepts.
- ✓ IDpr : ensemble des identificateurs des propriétés représentés par des types de relations conceptuelles dans le treillis des types de relations.
- ✓ IDrs : ensemble des identificateurs des relations symboliques représentés par des types de relations conceptuelles dans le treillis des types de relations.
- ✓ VAL_PROP = Réel U Entier U Chaîne U Booléen : ensemble des définitions de propriétés
- ✓ PROP $\subseteq$ IDpr × IDcl × P(VAL_PROP) : ensemble des définitions de propriétés représentées par le graphe canonique [c]→[p]→[val] où p est la relation conceptuelle correspondant à la propriété p de IDpr.
- ✓ RSYMB $\subseteq$ IDrs × IDcl x IDcl : ensemble des définitions des relations symboliques représentées par le graphe canonique [c1]→[rs]→[c2] où rs est la relation conceptuelle correspondant à la relation rs de IDrs.
- ✓ Domaine : IDpr → P(VAL_PROP) définit pour chaque type de valeurs de propriétés l'ensemble des valeurs possibles

La facette symbolique associe à l'ensemble des objets symboliques leur sémantique définie dans le modèle d'application.

Le modèle de facette symbolique est défini comme le 5-uplet (M, Isy, cl, RI, PI) avec :

- ✓ M : application pour laquelle le contexte symbolique est défini.
- ✓ Isy : ensemble des identificateurs d'objets symboliques.
- ✓ cl : Isy → IDcl associe à chaque objet symbolique sa classe.
- ✓ RI $\subseteq$ IDrs × Isy × Isy : lie les objets symboliques entre eux par les relations de IDrs.
- ✓ PI $\subseteq$ IDpr × Isy × VAL_PROP : lie les objets images à leurs propriétés.

Dans la représentation de la facette symbolique dans le formalisme des graphes conceptuels, le modèle de la facette symbolique est confondu avec celui du modèle de l'application. Il comprend la définition du type de concept représentant les objets symboliques, l'ensemble

des classes et leur relation d'ordre, la définition des propriétés, leurs valeurs et la relation de composition entre les classes d'objets.

Une instance de facette symbolique est représentée par un graphe conceptuel canonique produit par la jointure sur le concept objet symbolique entre les graphes correspondant aux propriétés et ceux correspondant aux objets symboliques liés par des relations symboliques.

V.I.2.3.4 Modèle d'image dans EMIR²

Les représentations index ainsi que les requêtes sont des instances du modèle d'image EMIR^2 qui est donné par l'agrégation de l'ensemble des modèles des facettes et d'un ensemble de relations représentant les liaisons entre les vues :

Le modèle image est défini par le 9-uplet (I_{im}, M_{ph}, M_{st}, M_{pe}, M_{sp}, M_{sy}, L_{sp}, L_{sy}, L_{pe}) avec :

- ✓ I_{im} est l'ensemble des images EMIR^2.
- ✓ M_{ph}, M_{st}, M_{pe}, M_{sp}, M_{sy} sont respectivement les modèles de facette physique, structurelle, perceptive, spatiale et symbolique.
- ✓ L_{sp}, L_{sy}, L_{pe} sont des relations qui associent aux objets images respectivement les objets spatiaux de la facette spatiale, les objets perceptifs de la facette perceptive et les objets symboliques de la facette symbolique.

Le lien entre l'image et ses différentes facettes est représenté par des relations conceptuelles liant le concept représentant l'image ou l'objet image au concept représentant l'objet spatial, perceptif, ou symbolique lui correspondant dans L_{sp}, L_{sy}, L_{pe}. Ces différentes facettes sont dépendantes et forment ensemble une description complète de l'image. La figure V.3 illustre un exemple de description d'image sous forme de graphe conceptuel en utilisant les interprétations des différentes vues :

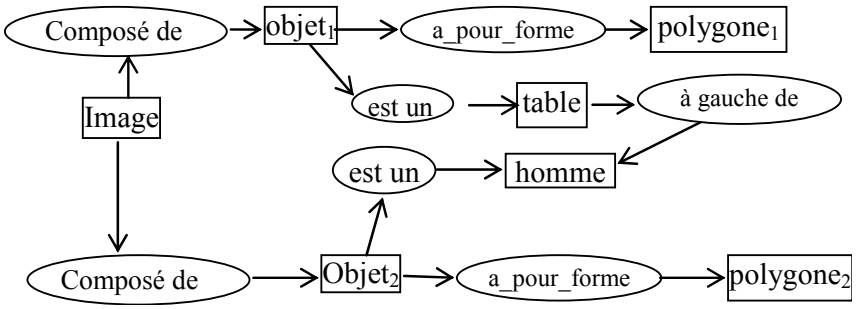


Figure V. 3 : Exemple de représentation formalisme de GCs

V.I.3 Extension du Modèle EMIR²

Un document vidéo présente l'association de plusieurs médias (image, son, texte) dont chacun contient de l'information sous une forme spécifique. Plusieurs nouveaux aspects existent dans la vidéo, tels que, dans la piste image, le mouvement et les changements spatio-temporels des caractéristiques des objets.

La recherche d'un passage précis dans un ensemble de documents vidéo nécessite la prise en compte de ces différents aspects. L'indexation du flux visuel dépend de l'interprétation des

images qu'il contient. On distingue deux niveaux d'interprétation.. Le premier ne fait appel qu'au contenu du segment vidéo considéré. Le second fait intervenir en plus des connaissances complémentaires qui permettent d'enrichir la description issue du premier niveau.

Le contenu d'un plan peut être représenté par une image-clé qui est fixe et dans ce cas on peut appliquer directement le modèle EMIR² pour la description de la partie visuelle de ce plan. Par contre, si l'on considère les images animées du plan comme étant une succession d'images fixes auxquelles est associé l'aspect mouvement de caméra, le modèle EMIR² doit être adapté pour la représentation et la description du document vidéo. Par exemple, chaque image du plan peut être représentée individuellement par un graphe. Les graphes correspondant aux différentes images peuvent ensuite être interconnectés par des relations décrivant le mouvement des objets contenus dans ces images. On peut aussi décrire le contenu d'un plan par un graphe unique au niveau du plan lui-même (sans description image par image) dans lequel chaque objet (spatio-temporel) n'est décrit qu'une seule fois.

Nous proposons une extension du Modèle EMIR² pour l'indexation et la recherche dans une base de documents vidéo [Charhad 04a], [Charhad 04b], [Charhad 05c]. L'extension de ce modèle nécessitera la définition des nouveaux concepts et de nouvelles relations permettant de décrire le contenu du document vidéo.

V.I.3.1 Structure Multifacette

L'approche consiste à considérer chaque séquence vidéo comme une combinaison des diverses interprétations du contenu. Ces interprétations se traduisent par un ensemble d'objets vidéo et de relations qui les lient. Chaque interprétation est alors appelée « vue » ou bien « facette ». Elle-même peut être composée d'autres vues. Les vues sont classées selon le point de vue de l'interpréteur et selon les objets décrits.

Comme le montre la figureV.4, la description symbolique (par concept) du contenu vidéo est une description multimodale (une description spécifique au type de média). En effet, un concept peut être issu de la description du contenu visuel, du contenu audio ou bien du contenu texte.

Afin de simplifier l'analyse de notre proposition, nous allons tout d'abord décrire brièvement chacune des ses composantes. Notons que d'une manière générale, nous avons mis en œuvre des formes de représentation :

- **Générique :** cette forme de représentation regroupe l'ensemble des facettes décrivant les caractéristiques communes dans le document vidéo indépendamment du type de média, telles que par exemple, la nature temporelle de la vidéo. On distingue deux types de facettes :
  - ***Facette temporelle :*** c'est l'ensemble des relations temporelles qui relient les éléments d'informations dans le document vidéo.
  - ***Facette événementielle :*** pour la description des différents événements contenus dans un document vidéo. Un événement est une ou plusieurs actions.
- **Spécifique :** cette forme de représentation permet de décrire le contenu vidéo par média. En effet, il existe plusieurs spécificités de la vidéo qui sont propres à un média

particulier (visuel, audio ou texte). La représentation spécifique contient les facettes suivantes :

- ✓ **facette sémantique :** permet d'associer une description sémantique au contenu visuel, au contenu audio ou au contenu textuel. Cette description est souvent définie par des concepts. Cette facette est composée de trois autres facettes. Une sous-facette visuelle pour décrire le contenu visuel, une sous-facette audio pour la description du contenu audio et une sous-facette texte pour interpréter toutes informations textuelles dans le document vidéo.
- ✓ **Facette signal :** elle permet de décrire les caractéristiques de bas niveau afin de générer des descriptions sémantiques telles par exemple les caractéristiques couleurs dans le contenu visuel. La facette signal regroupe plusieurs sous-facettes, notamment au niveau visuel. Ces sous-facettes sont liées aux caractéristiques couleurs, textures et positions spatiales des objets visuels.

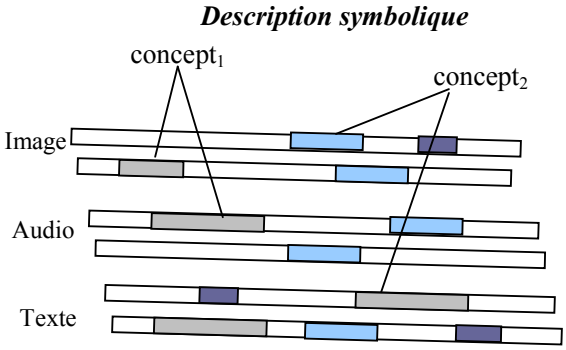


Figure V. 4 : Description conceptuelle multimodale du contenu

La figure V.5 montre l'ensemble de ces facettes. On distingue dans un premier niveau de description quatre catégories de facettes : sémantique, signal, temporelle et événementielle. Chacune de ces facettes est éventuellement composée d'autres facettes ou sous-facettes comme nous l'avons décrit précédemment.

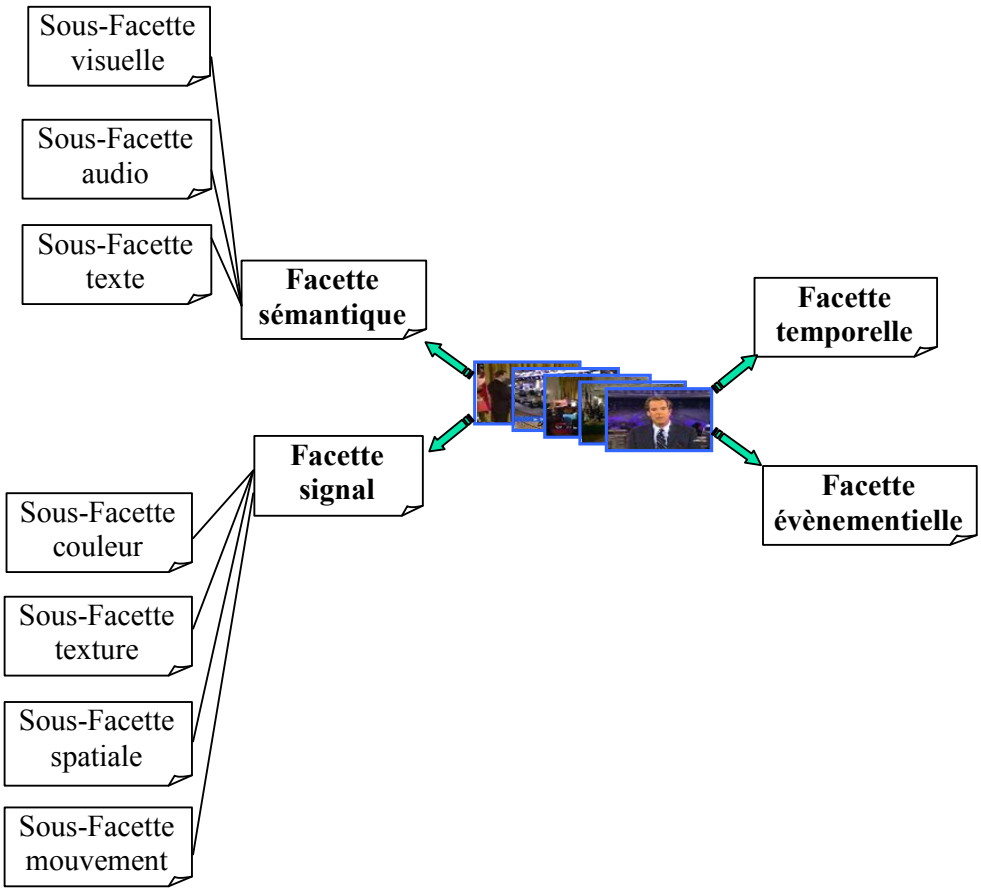


Figure V. 5 : Modélisation multifacette d'un document vidéo

V.I.3.1.1 Facette événementielle

Par définition, un événement est un fait qui survient à un moment donné. Au sens général, il signifie tout ce qui arrive et possède un caractère spécifique. Dans le cas de la vidéo, un événement est quelque chose qui arrive dans un document ou dans un segment de document vidéo.

Dans notre proposition, nous considérons la modélisation du contenu vidéo basée sur la description des événements comme étant une modélisation générique (indépendant d'un média particulier). En effet, généralement même si un événement peut être décrit d'une manière spécifique au type de média (on entend le bruit d'une explosion par exemple), au niveau interprétation la description d'un événement qui est associée au segment vidéo et non au type de média. Un événement peut correspondre à une interaction entre les entités audiovisuelles (personne, objets, etc.).

(a) Modélisation de la facette événementielle

Considérons Ev = {Ve1, Ve2, Ve3, Ve4,…, Ven} l'ensemble des événements. Un événement peut contenir des sous-événements (qui sont aussi des événements).

Un événement peut s'étendre sur un document complet (événement global). Ce type d'événement est généralement décomposable en sous-événements (voir figure V.6). Il en apparaît souvent dans le cas des documents spécifiques tels que les émissions sportives. En effet, un match de football peut constituer un événement et un but marqué dans ce match est alors considéré comme un sous-événement dans ce même document.

Afin de modéliser l'ensemble des événements et des sous-événements dans un document (ou un segment de document) vidéo, nous utilisons une relation de composition « est sous-événement de » qui permet de décomposer un événement en des sous-événements qui peuvent être eux-mêmes décomposables ou élémentaires. Un événement est élémentaire lorsqu'il constitue une description événementielle indécomposable.

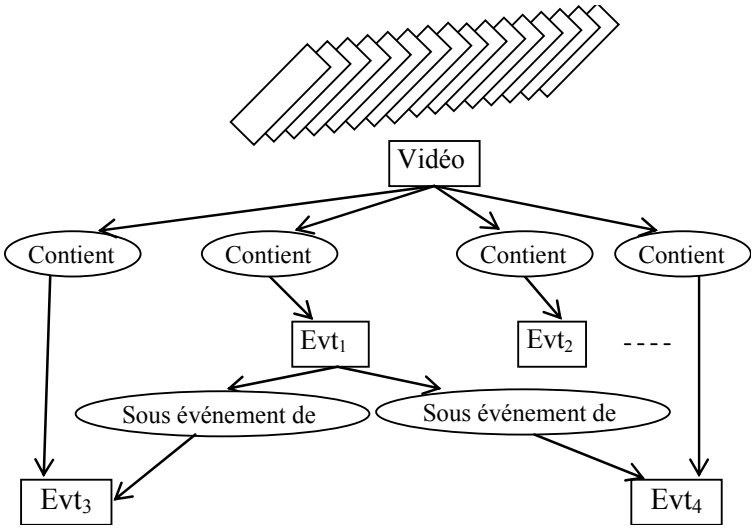


Figure V. 6 : Description événementielle

L'événement Evt_3 correspond à un sous-événement de l'événement Evt_1 dans le contexte du document vidéo. Il peut être une simple action dans le document. Une action peut être « filmique » telle qu'un mouvement de la caméra pour un spécialiste de montage vidéo ou bien elle peut être « réelle » dans la vidéo comme le mouvement des acteurs dans un film par exemple.

Comme le décrit le treillis représenté dans la figure suivante, dans notre proposition pour la modélisation des « événements » dans la vidéo, on se limite à une description des événements représentant les actions réelles dans le document telles que par exemple action d'une personne (exemple : un discours) ou bien action d'un groupe de personne (exemple : une manifestation). Notons que d'une manière générale, la construction de ce type de treillis ainsi que la classification des événements dépend de l'application. D'autres événements un peu plus génériques sont classés selon un sujet particulier dans le document comme par exemple des événements sportifs.

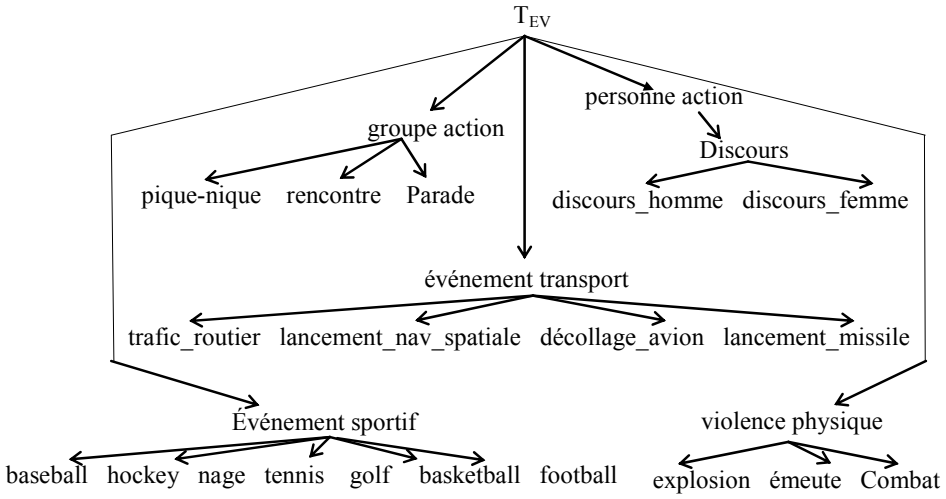


Figure V. 7 : Treillis de concepts « événement »

(b) Spécification conceptuelle

La spécification conceptuelle est la description des différentes actions que l'on peut extraire d'un document vidéo. Chaque description est définie par l'ensemble des faits se produisant dans le passage.

Les descriptions événementielles sont représentées par les concepts « événement » Ve. Chaque représentation est formulée par un ensemble de graphes conceptuels comme pour le cas de la modélisation du contenu visuel.

La génération basique de graphe de représentation est la suivante :

[Ve1]→(sous-événement de)→[Ve2]

La description des événements dans un document vidéo s'attache aussi à la nature temporelle du document. En effet, les événements ont une extension temporelle qui correspond à un intervalle, ce qui permet de définir des relations temporelles entre eux. Les relations temporelles permettent aussi de décrire les concepts (les apparitions / disparitions de ces concepts).

V.I.3.1.2 Facette temporelle

L'information temporelle est une caractéristique spécifique du document vidéo qui permet de synchroniser et d'ordonner l'ensemble des descriptions utilisées pour la modélisation du contenu vidéo. En effet, c'est elle, en donnant au document son caractère dynamique, qui le différencie d'un document classique.

Des nombreux de travaux de recherche [Hijelsvold 94], [Fatemi 99], [Ma 97] ont centré leurs études sur l'exploitation de l'information temporelle dans le document vidéo. Dans cette section, nous n'avons pas l'objectif de re-décrire les approches proposées dans ces travaux. Nous allons plutôt détailler comment nous avons exploité et intégré cette caractéristique temporelle dans notre modèle.

Pour intégrer l'information temporelle dans le modèle, nous proposons de représenter le contenu d'une séquence par deux descriptions conjointes. La première modélise le contenu sémantique pour chaque Unité Audio Visuelle (UAV définit une entité abstraite représentant un segment quelconque de document vidéo) présente dans la séquence. La seconde modélise l'information temporelle entre les différentes UAV décrites dans la séquence.

Les relations d'Allen [Hijelsvold 94] constituent un bon exemple pour décrire le contenu de la vidéo en se basant sur l'information temporelle. Ces relations sont explicitées dans la figure V.8. Sur cette figure, nous illustrons les relations temporelles qui peuvent exister entre deux unités audiovisuelles (UAV). Il existe au total 13 relations d'Allen dont douze relations sont asymétriques. Excepté la relation « égale », les autres peuvent être regroupées deux par deux. En effet, si on a UAV1 chevauche UAV2 alors on peut inférer le que UAV2 est chevauchée par UAV1. C'est pour cette raison dans notre modèle nous ne gardons que sept relations.

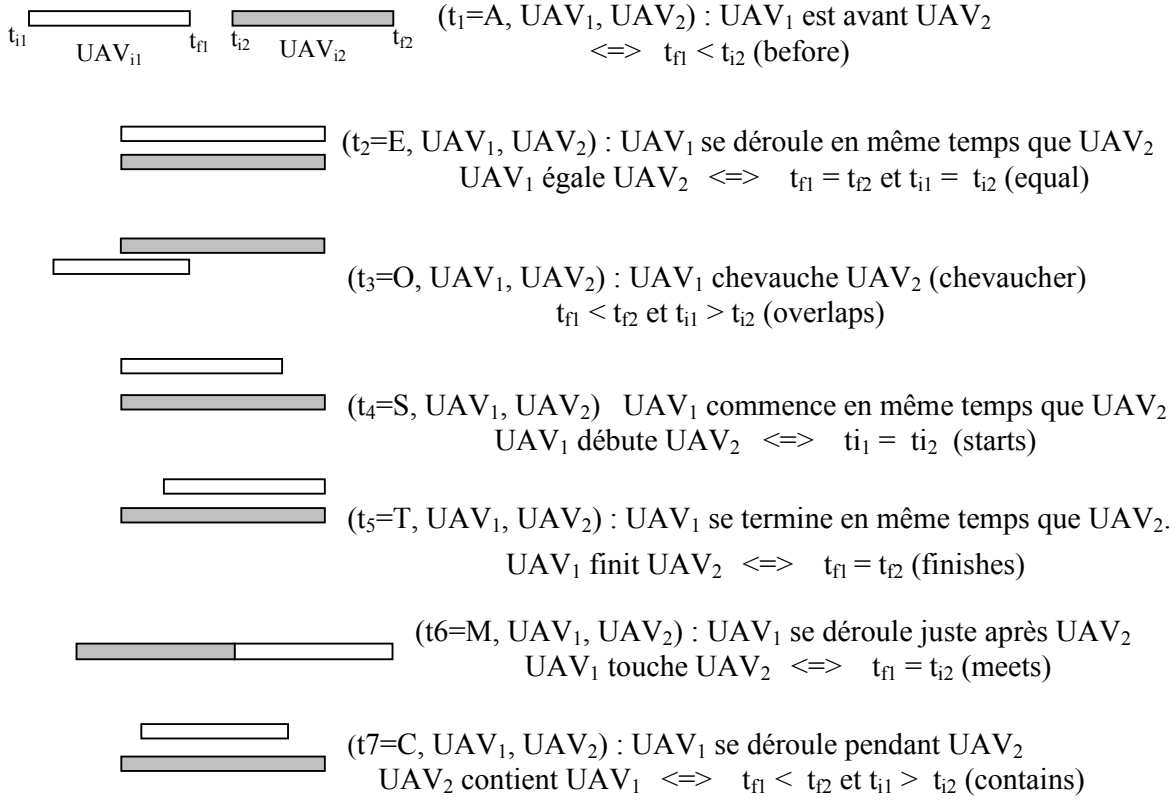


Figure V. 8 : Les relations temporelles d'Allen

L'information temporelle est une caractéristique spécifique à la vidéo. Elle peut être associée dans le flux visuel à l'apparition ou à la disparition des objets visuels par exemple ou bien dans le contenu audio à un changement de thème dans le discours ou à un changement de locuteur.

La segmentation du contenu audio ne correspond pas forcément à un segment issu du contenu visuel d'un point de vue de la durée et des bornes des intervalles. Notons que la segmentation du contenu audio ne correspond pas forcément à celle du contenu visuel.

(a) Modélisation de la facette temporelle

La facette temporelle contient l'ensemble des relations qui relient les entités d'une vidéo dans un ordre temporel bien déterminé.

Afin de modéliser des descriptions temporelles, nous considérons d'abord un ensemble des relations temporelles; sept relations qui sont appropriées pour la description temporelle sont choisies. Ces relations sont décrites entre unités audivisuelles (UAV) (voir figure V.8).

(b) Spécification conceptuelle

Les événements sont reliés deux à deux par des relations temporelles (TeR). Une TeR l'une des sept relations temporelles que nous venons de décrire.

Par exemple, Ve_1 et Ve_2 sont reliés par TeR, est traduit comme UAV_1 se déroule avant UAV_2 (UAV_2 *aprés* UAV_2).

La génération de base pour décrire l'ensemble des graphes attachés à cette facette entre deux unités audiovisuelles est illustré dans la représentation linéaire par:

$[UAV_1] \rightarrow (\ TeR) \rightarrow [UAV_2]$.

Notons que pour cette facette temporelle, notre modélisation est restreinte uniquement à une description par des relations symboliques. Elle ne permet pas de prendre en compte des informations de type « timecode » et durée. Ceci est du principalement à l'utilisation des formalisme des graphes conceptuels qui ne permet pas de représenter des valeurs numériques.

Partie II - Modélisation Multifacette et multimodale- Représentation Spécifique : le Modèle CLOVIS[9]

V.II.1 Document Vidéo : Description Générique

Un élément d'information dans un segment vidéo peut être représenté de différentes manières selon les types de médias (image, audio ou texte). En effet, il peut être vu, entendu, cité dans le discours ou bien apparaître dans les sous-titres qui s'affichent à l'écran (figure V.9). La combinaison de ces cas est aussi envisageable.

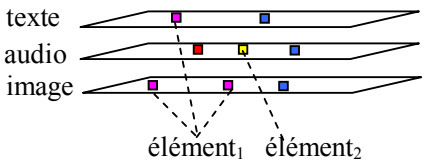


Figure V. 9 : Éléments de description et effet d'ancrage dans un segment vidéo

Du point de vue de la représentation, les éléments d'information dans un document vidéo peuvent être distingués par des relations sémantiques spécifiques à chaque sous média. En effet, pour symboliser un concept visuel on utilise des relations sémantiques de type « apparaît », « disparaît », etc. par contre pour un concept audio, les relations sémantiques sont de type « parle », « parle de », etc.

Les différents éléments d'informations peuvent être symboliques (concepts visuels et concepts audio). Ces éléments sont généralement décrits par des relations conceptuelles telles que celles dont nous avons citées précédemment. Ils peuvent être génériques (auteur, date, format, taille). Ces éléments fournissent des renseignements génériques sur le document vidéo.

V.II.2 Architecture du Modèle CLOVIS

Le modèle permet une description multimodale du contenu vidéo. Nous proposons une modélisation du contenu par la description des caractéristiques audio et visuelles du document vidéo.

Dans cette section nous allons détailler l'architecture globale de notre modèle tel que représenté dans la figure V.10. Il s'agit d'une modélisation multifacette permettant de décrire l'ensemble des caractéristiques audio et visuelles contenues dans le document vidéo. En partant d'une vidéo segmentée en plans, nous associons une description par média (audio et visuel) à chaque plan.

Cette architecture est composée de quatre parties :

---

[9] **C**onceptual **L**ayer **O**rganization for **V**ideo **I**ndexing and **S**earch

Deux premières parties permettent la modélisation des contenus audio et visuel. Pour le contenu audio, nous utilisons la transcription de la parole. Pour le contenu visuel, nous utilisons une image clé représentative de chaque plan. La troisième partie de l'architecture permet la formulation de la requête avec le formalisme des graphes conceptuels. La quatrième partie effectue le calcul de correspondance entre le modèle des documents et le modèle des requêtes.

Dans cette architecture, les notations représentent :

- ✓ Ao : désigne un objet audio. Celui-ci peut être toute information symbolique issue du contenue audio.
- ✓ Io : désigne un objet image. Celui-ci peut être toute information symbolique issue du contenue visuel.
- ✓ SBD (Shot Boundary detection) : segmentation en plans.
- ✓ ASR (Automatic Speech Recongnition) : transcription automatique de la parole.

Le nœud « Bill Clinton » est une instanciation d'une entité abstraite dans le contenu vidéo. Cette entité peut être audio (un objet audio Ao) ou visuelle (objet image Io). On associe les relations « parle » et « apparaît » aux instances des objets. Par contre les objets et leurs instances sont reliés par des relations de spécificité telles que par exemple [Ao1] → (is a) → [Bill Clinton].

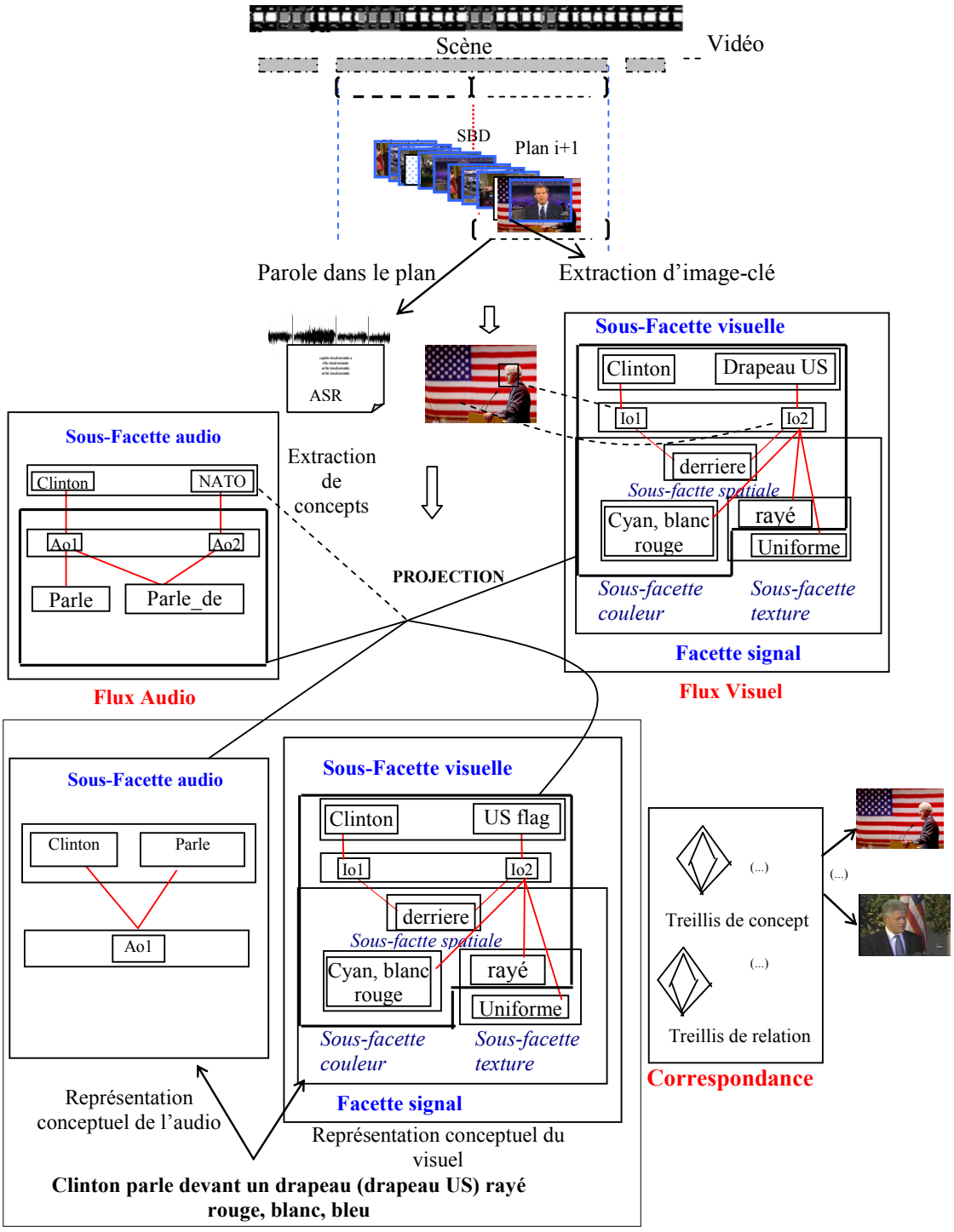


Figure V. 10 : Représentation de l'architecture du modèle CLOVIS

V.II.3 Modélisation du Contenu Visuel

Au niveau de la description du contenu visuel, il existe généralement une relation entre le contenu et les caractéristiques visuelles de bas niveau. En effet, certaines informations sur le contenu peuvent être interprétées en se basant sur la perception visuelle afin de déterminer la couleur d'un objet, sa trajectoire etc.

D'un point de vue pratique, le contenu visuel d'un plan est représenté par une image-clé. Pour une description plus détaillée du contenu visuel, nous proposons de mettre en place un modèle de représentation combinant un ensemble de caractéristiques visuelles. Chaque ensemble des caractéristiques est représenté par une facette (ou vue). Les deux principales catégories de facettes sont les facette signal et symbolique. La facette symbolique, regroupe tous les aspects du contenu sémantique et de son contexte général.

Le modèle se base sur la notion d'objet audiovisuel pour décrire le contenu. Les objets audiovisuels sont des représentations abstraites des éléments d'informations présents dans le document vidéo.

V.II.3.1 Spécification conceptuelle et description du contenu visuel

La description symbolique du contenu visuel s'ajoute aux nombreuses techniques « d'analyse » qui existent déjà et qui mettent en œuvre des critères tels que le mouvement des objets, les histogrammes de couleurs, etc.

Cette description intègre les relations qui existent entre les différents concepts associés pour chaque plan. On peut donc distinguer deux niveaux de description :

- ✓ Le premier niveau s'attache à la description sémantique (décrire l'action ou l'événement associé au segment vidéo),
- ✓ Le deuxième niveau est lié à la description de la structure physique des objets (tels que la couleur, la luminosité et le mouvement apparent des objets).

La figure V.11 montre comment des concepts visuels sont rattachés à un plan vidéo. La description es fait par un ensemble de concepts visuels (cv). Un concept visuel peut être une personne ou n'importe quel objet symbolique qu'un humain peut considérer comme une unité de repérage dans le plan. Ces concepts sont associés à un ou plusieurs intervalles de temps qui correspondent aux moments où ces concepts sont effectivement visibles dans le plan.

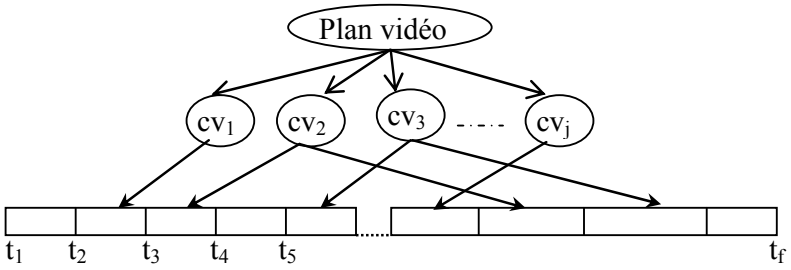


Figure V. 11 : Description symbolique du contenu visuel

Pour une mieux comprendre la tâche de modélisation du contenu visuel, nous proposons dans ce qui suit une structure de représentation générique :

[Flux visuel] → *(Description structurelle)* → *[Séquence d'images]*

[Image] → *(Composé de)* → *[Objet image]*

[Objet image] → *(Description symbolique)* → *[Concept visuel]*

[Objet image] → *(Description spatiale)* → *[Objet image]*

[Concept visuel] → *(Description temporelle)* → *[Concept visuel]*

[Concept visuel] → *(Description symbolique)* → *[Concept visuel]*

La première étape pour décrire le contenu visuel consiste à identifier les passages. Ceci passe par une étape de segmentation où l'ensemble complet du flux est décomposé en un certain nombre de plus petites portions contiguës appelées « *plans* » qui peuvent elles-mêmes être regroupés en « *scènes* ». La deuxième étape est le choix, à partir de chaque plan, d'un passage dans lequel un événement se déroule. Cet événement peut alors être représenté par des caractéristiques de bas niveaux tel que la couleur, la luminosité etc.

V.II.3.1.1 Facette Signal

La facette signal regroupe les informations sur le contenu visuel « de bas niveau » telles que : la couleur, la texture, les positions spatiales, etc. Ces informations sont représentées sous formes de descripteurs numériques. Dans plusieurs cas, il est possible d'exploiter ce type d'information pour inférer une description sémantique par l'agrégation d'un certain nombre de ces critères de bas niveau (par exemple, pour la détection de l'événement « explosion »).

L'aspect signal bien qu'il apparaisse moins important par rapport au contenu symbolique à un spectateur humain, le devient dans le cadre de travaux de modélisation, d'indexation et de recherche de documents vidéo.

L'objectif consiste toujours à automatiser le processus d'indexation tout en tenant compte du point de vue utilisateur lors du processus de recherche. Ces approches sont plus ou moins efficaces lorsqu'il s'agit d'un cadre d'étude spécifique (vidéo à un seul thème : sport par exemple). C'est-à-dire dans le cas où il est facile de prévoir les événements et détecter les entités visuelles contenues dans le document.

EMIR² propose une facette perceptive des images fixes qui décrit uniquement l'aspect physique de l'image tel que la couleur, la texture etc. Dans le cas de la vidéo, la perception visuelle est riche et variée. En effet, dans les images animées on peut décrire par exemple : le mouvement des objets et le suivi de leurs trajectoires, la détection du mouvement de caméra (panoramique, zoom, etc.).

La modélisation de la facette signal (excepté ce qui concerne la sous-facette mouvement) est inspirée des travaux de M. Belkhatir [Belkhatir 05] dans le cadre de ses travaux de thèse. Ces travaux concernant les images fixes sont directement appliqués aux images clés représentant les plans vidéo.

La facette signal décrit le contenu visuel du document vidéo en termes de perception visuelle de l'information vidéo. Elle permet de spécifier les caractéristiques visuelles de bas-niveau de

la vidéo. Formellement, on dénote les éléments de cette facette par des descripteurs images (Ids). Ces descripteurs ne sont pas forcément d'ordre symbolique, mais ils peuvent servir à inférer des descriptions sémantiques.

Dans cette section, nous détaillons les aspects formels liés à la description de facette signal. Nous proposons de décrire les caractéristiques conceptuelles pour la sous-facette couleur, la sous-facette texture, la sous-facette spatiale et la sous-facette mouvement.

- *La sous-facette couleur* pour représenter les caractéristiques couleurs dans le contenu visuel du document. Par exemple, à l'objet image (Io2) sont attribuées les couleurs bleu, blanc et rouge.
- *La sous-facette texture* permet de décrire les propriétés texture dans le contenu visuel. Dans le cas de l'exemple1, Io2 est décrite pas indices textures « rayé » et « uniforme ».
- *La sous-facette spatiale* pour spécifier les relations spatiales (position relative, direction) des objets images.
- *La sous-facette mouvement* pour spécifier les mouvements de la caméra des objets images ainsi que leurs trajectoires.

Pour chaque plan vidéo, la première tâche consiste en le choix de l'image-clé. Nous allons tout d'abord détailler l'extraction et la sélection d'images-clé.

(a) Analyse et extraction de l'image-clé d'un plan vidéo

Le processus d'extraction d'une image-clé ([Naphade 02], [Hampapur 95]) dans un plan vidéo passe forcément par l'idée qu'une telle image doit capturer le contenu sémantique d'un plan. Il est donc inutile d'entrer dans des calculs compliqués et longs sur toutes les images du plan vidéo. En effet, il serait ensuite impossible de conserver et d'utiliser cette information. Malheureusement, les techniques existantes ne sont pas assez avancées pour déterminer efficacement l'image-clé d'un plan. Les algorithmes utilisent souvent des techniques les caractéristiques brutes obtenues lors de l'analyse du contenu visuel tel que la couleur, la texture, le mouvement etc.

Théoriquement, le choix de l'image clé d'un plan vidéo se base sur le critère de stabilité. Par conséquent, une image-clé est souvent celle qui correspond à l'image la plus similaire au plan pris dans son ensemble.

Généralement, le processus principal d'extraction d'images clés est intégré avec les processus de la segmentation en plan. Chaque fois qu'un nouveau plan est identifié, le processus d'extraction d'image-clé est appelé, en utilisant des paramètres déjà calculés pendant la détection de transition entre plans (Shot Boundary Detection SBD) [Quénot 01]. Ces paramètres sont liés aux caractéristiques visuelles telles que la couleur ou les mouvements de caméra. La sélection d'un image-clé peut varier selon la nature et les besoins d'applications. Par exemple, nous avons besoin seulement de quelques images-clé principales (maximum un ou deux) pour un document vidéo décrivant une réunion puisque les mouvements de caméra sont presque statiques.

La figure V.12, montre une architecture pour l'extraction automatique des concepts sémantiques visuels : un réseau « neuronal » à 3 couches avec des possibilités dynamiques de création de noeuds est employé pour les concepts visuels dans chaque image-clé. Des

caractéristiques de couleur et de texture sont calculées pour chaque région de formation comme vecteur d'entrée pour le réseau neuronal. Une fois l'apprentissage du vocabulaire visuel est effectué par le réseau, l'approche ordonne une image-clé quadrillée par mesure d'identification basée sur ces concepts sémantiques visuels. Chaque image-clé sera balayée avec des fenêtres de plusieurs paramètres. Chacun représente un trait visuel caractérisé par un vecteur de description.

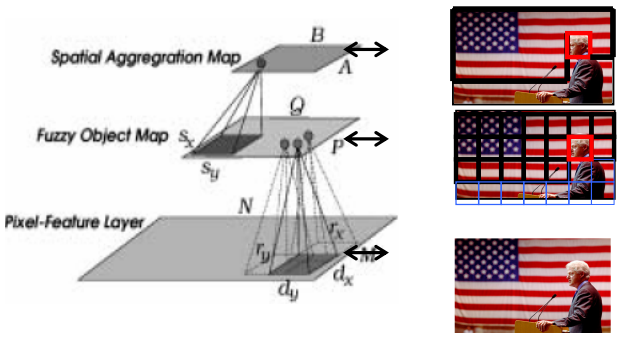


Figure V. 12 : Architecture pour l'extraction du contenu visuel

V.II.3.1.1.1 Sous-facette couleur

(a) Description

Intégrer les descripteurs signal dans un cadre conceptuel (haut-niveau) n'est pas une tâche simple. La première étape consiste à indiquer les données conceptuelles qui correspondent aux descripteurs de bas-niveau. Ceci nécessite la mise en place d'un processus de correspondance entre les noms de couleur et les valeurs numériques correspondant à chacune d'elles. La représentation symbolique de l'information de couleur est guidée par les travaux de recherche sur la catégorisation de couleurs [Berlin 69]. Nous considérerons l'existence d'un système formel Snc [Lammens 94] qui distingue un ensemble de catégories (ou concept) couleur avec un Ccat (Color category). Ces concepts couleur sont désignés par C_i avec $1 \leq i \leq 11$.

Nous utilisons les 11 concepts couleur : « C_1=rouge », « C_2=blanc », « C_3=bleu », « C_4=gris », « C_5=cyan », « C_6=vert », « C_7=jaune », « C_8=violet », « C_9=noir », « C_{10}=peau », « C_{11}=orange ».

(b) Spécification conceptuelle

Dans le modèle proposé par Belkhatir, l'information de couleur relative à un objet image est décrite par une structure de donnée caractérisant qualitativement et/ou quantitativement sa distribution de couleur; celle-ci est appelée « méta-concept ». Différentes telles structures conceptuelles (méta-concepts) sont spécifiées par rapport à différents types de requêtes pouvant être formulées par un utilisateur. Il existe deux types de méta-concepts (chacun ayant des sous-types) : booléen et numérique. L'approche développée permet une interrogation riche par l'intermédiaire de trois opérateurs booléens et six opérateurs de quantification appliqués respectivement à des méta-concepts de type booléen ou numérique (Belkhatir 2005) :

- Un utilisateur a la possibilité d'associer dans une requête des types de concepts de sémantique avec une conjonction booléenne de catégories de couleurs (par exemple : trouver des plans

avec un drapeau rouge *et* blanc *et* bleu), avec une disjonction booléenne de catégories de couleurs (par exemple : trouver des plans de Bill Clinton avec un costume noir *ou* bleu) ou avec une négation de catégories de couleurs (par exemple : trouver des plans avec un ciel *non* gris).

- Un utilisateur a également la possibilité de spécifier des quantifications numériques de catégories de couleurs (par exemple : « *au plus* 25% de bleu », « *au moins* 40% de blanc » ou « *plus de* jaune *que de* rouge »).

Nous ne reprendrons dans notre modèle que la partie relative aux méta-concepts booléens et aux types de requêtes associées (conjonction, disjonction et négation de catégories). Par la suite, « méta-concept » sera synonyme de « méta-concept booléen ». On distingue dans cette approche des méta-concepts « index » caractérisant la distribution de couleur effectivement présente dans un objet image et des méta-concepts « requête » pour traduire les distributions de couleurs (ou des ensembles de distributions de couleurs) spécifiées au sein des requêtes.

Les méta-concepts index pour la couleur représentent la distribution des couleurs des objets image par une conjonction de catégories couleur (C_1 à C_{11}) et sont caractérisés par une structure de vecteur C_{ind} avec un nombre d'éléments égal au nombre de catégories de couleur (11 dans notre cas). Les valeurs $C_{ind}[i]$ avec $i \in [1,11]$ sont des booléens traduisant la présence ou l'absence de la catégorie de couleur correspondante C_i dans l'objet image. Par exemple, un objet image représentant un drapeau américain sera représenté par le vecteur : $<C_1:1, C_2:1, C_3:1, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$. On convient de le représenter aussi par la forme plus claire et compacte : <rouge, blanc, bleu>.

Les méta-concepts requête sont classés en trois groupes selon le type de relation par lequel ils sont impliqués dans une requête : conjonction, disjonction ou négation. Ils sont caractérisés respectivement par les structures de vecteur C_{et} , C_{ou} , et C_{non} avec un nombre d'éléments égal au nombre de catégories de couleur (11). Les valeurs $C_{et}[i]$, $C_{ou}[i]$ et $C_{non}[i]$ avec $i \in [1,11]$ sont des booléens traduisant le fait que la catégorie de couleur correspondante C_i est élément respectivement des conjonctions, disjonctions et/ou négations des catégories de couleur formulées dans la requête. Ces méta-concepts sont notés comme les méta-concepts index avec en indice à la fin « ET », « OU » ou « NON » indiquant le groupe dont ils font partie. Par exemple :

- Le méta-concept $<C_1:1, C_2:1, C_3:1, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>_{ET}$ est celui qui intervient dans la transcription de la requête : « trouver des plans avec un drapeau rouge *et* blanc *et* bleu ». On convient de le représenter aussi par la forme plus claire et compacte : $<\text{rouge, blanc, bleu}>_{ET}$.
- Le méta-concept $<C_1:0, C_2:0, C_3:1, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:1, C_{10}:0, C_{11}:0>_{OU}$ est celui qui intervient dans la transcription de la requête : « trouver des plans de Bill Clinton avec un costume noir *ou* bleu ». De même, on convient de le représenter aussi par la forme plus claire et compacte : $<\text{noir, bleu}>_{OU}$.
- Le méta-concept $<C_1:0, C_2:0, C_3:0, C_4:1, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>_{NON}$ est celui qui intervient dans la transcription de la requête : « trouver des plans avec un ciel *non* gris ». De même, on convient de le représenter aussi par la forme plus claire et compacte : $<\text{gris}>_{NON}$.

Les méta-concepts d'un même groupe peuvent être organisés dans un treillis selon la relation d'ordre partiel « générique > spécifique » selon le fait que si le spécifique est vrai alors le générique l'est aussi. Par exemple :

<rouge, blanc, bleu>$_{ET}$ est un spécifique de <rouge, blanc>$_{ET}$

<noir, bleu>$_{OU}$ est un spécifique de <noir, bleu, vert>$_{OU}$

<gris>$_{NON}$ est un spécifique de <gris, blanc>$_{NON}$.

Les trois treillis peuvent être fusionnés en un treillis unique de méta-concepts requête. De même les méta-concepts index peuvent être organisés en un treillis unique (de type ET). Contrairement au cas des autres types de concepts (sémantiques par exemple), il y a ici une dissymétrie entre la représentation des documents et celle de la requête. Ceci n'est pas un problème dans la mesure où pour chaque paire de concepts associés respectivement à un document et à une requête, il est possible de déterminer (selon la logique booléenne appliquée aux catégories impliquées) si oui ou non ils peuvent être mis en correspondance selon la relation de spécificité intervenant dans l'opération de projection du « graphe requête » dans le « graphe document ». Une telle opération n'est pas triviale mais des solutions existent (Belkhatir, 2005).

Les méta-concepts index sont mis en relation avec les objets images présents dans le document par la relation « has color ». De même, les méta-concepts requête sont mis en relation avec les objets images spécifiés dans la requête par la même relation « has color ».

Une instance de la sous-facette couleur est représentée par un ensemble de graphes conceptuels chacun comprenant un concept de type Io lié par la relation conceptuelle « has color » à un concept couleur. Cette représentation est similaire pour le graphe index et le graphe requête. Elle est généralement décrite par le graphe suivant :

[Io] → (has color) → [C_{xx}]

Si par exemple dans un plan vidéo l'objet Io a une couleur rouge, une représentation dans le formalisme des graphes conceptuels contiendra le sous-graphe :

[Io] → (has color) → [<C_1:1, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$]

ou :

[Io] → (has color) → [<rouge>]

Notons qu'un objet image dans un document est décrit par une conjonction de couleurs. Par exemple un objet image représentant un drapeau américain a une couleur rouge, blanche et bleue. Dans le cas des requêtes, on peur aussi avoir une disjonction de couleur (<col>$_{OU}$). Par exemple : « rechercher des objets images ayant une couleur rouge ou blanche ».

La représentation des cas de conjonction et disjonction des couleurs par des graphes conceptuels est la suivante :

[Io] → (has color) → [<col>$_{ET}$]

[Io] → (has color) → [<col>$_{OU}$]

Avec <col>$_{ET}$ et <col>$_{OU}$ représentant respectivement une combinaison de 11 valeurs booléennes représentant les concepts couleurs. Par exemple :

Une requête spécifiant un Io ayant une couleur rouge, blanche et bleue contient le sous-graphe :

[Io] → (has color) → [<C_1:1, C_2:1, C_3:1, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$]

Une représentation de la requête « rechercher les objets images ayant une couleur rouge ou blanche » contient le sous-graphe suivant :

[Io] → (has color) → [<C_1:1, C_2:1, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{OU}$]

V.II.3.1.1.2 La sous-facette texture

(a) Description

Bien que plusieurs travaux aient été proposés pour l'analyse de la caractéristique texture, peu de propositions ont été faites pour la reconnaissance symbolique de cette caractéristique. Notre représentation symbolique de la texture se base sur les travaux de recherche en nommage et catégorisation de textures proposés par [Bhushan 97]. Nous considérons les concepts suivants de texture comme représentation de chacune de ces catégories : « T_1 = bosselé », « T_2 =craquelé », « T_3 = désordonné », « T_4 =entrelacé », « T_5 = rayé », « T_6 = marbré », « T_7 = rétiforme », « T_8 = sali », « T_9 = tacheté », « T_{10} = uniforme » et « T_{11} = en vague ». Ces concepts texture sont désignés par T_i avec $1 \leq i \leq 11$.

Ces 11 concepts à niveau forment la base de notre cadre de travail pour la caractérisation symbolique de texture.

(b) Spécification Conceptuelle

La forme de la spécification conceptuelle est exactement la même pour la texture que pour la couleur. Les 11 textures de bases jouent le même rôle que les 11 couleurs de base et elles peuvent être combinées de la même façon pour produire des treillis de méta-concepts texture index et de méta-concepts texture requête. La relation « has color » a un équivalent noté « has texture ». Les mêmes notations concises peuvent être utilisées.

Une instance de la sous-facette texture est représentée par un ensemble de graphes conceptuels chacun comprenant un concept de type Io lié par la relation conceptuelle « has texture » à un méta-concept texture. Cette représentation est similaire pour le graphe index et le graphe requête. Elle est généralement décrite par le graphe suivant :

[Io] → (has texture) → [T_{xx}]

Si par exemple l'objet Io a une texture rayée, une représentation dans le formalisme des graphes conceptuels sera comme suite :

[Io] → (has texture) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5: **1**, T_6: 0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$]

ou :

[Io] → (has texture) → [<rayé>]

Notons qu'un objet image dans un document peut être décrit par une conjonction de textures. Par exemple un objet image représentant un drapeau américain a une texture uniforme et rayée. Dans le cas des requêtes, on peut aussi avoir une disjonction de textures (<tex>$_{OU}$). Par exemple, « rechercher les objets images ayant une texture sali ou désordonnée ».

La représentation des cas de conjonction et disjonction des couleurs par des graphes conceptuels est la suivante :

[Io] → (has texture) → [<tex>$_{ET}$]

[Io] → (has texture) → [<tex>$_{OU}$]

Avec <tex>$_{ET}$ et <tex>$_{OU}$ représentant respectivement une combinaison de 11 valeurs booléennes représentant les concepts texture. Par exemple :

Une requête spécifiant un Io ayant une texture uniforme et rayée contient le sous-graphe :

[Io] →(has texture) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5: **1**, T_6: 0, T_7:0, T_8:0, T_9:0, T_{10}:**1**, T_{11}:0>$_{ET}$]

Une représentation de la requête « rechercher les objets images ayant une texture sali ou désordonnée » contient le graphe suivant :

[Io] →(has texture) → [<T_1:0, T_2:0, T_3: **1**, T_4:0, T_5: 0, T_6: 0, T_7:0, T_8:**1**, T_9:0, T_{10}:0, T_{11}:0>$_{OU}$]

V.II.3.1.1.3 La sous-facette spatiale

(a) Description

La facette spatiale permet de spécifier une description spatiale d'un objet par rapport aux autres du point de vue de la perception visuelle. D'autre part, déterminer la position géographique d'un objet vidéo (déterminer le lieu où une telle scène se déroule).

La description spatiale d'une séquence vidéo se traduit selon deux formes. La première au niveau topologique permettant de situer les objets les uns par rapport aux autres. Ceci résulte de la perception visuelle du contenu. Il est souvent qualifié de statique du fait qu'il s'applique aux images fixes. On considère 10 relation qui sont : « à gauche de », « à droite de », « au-dessous de », « au-dessus de », « devant », « intérieur de », « touche », « déconnecté de », « proche de », « loin de ».

Afin de modéliser des données spatiales, nous considérons d'abord un sous-ensemble des relations topologiques. Considérons deux objets images (Io1 et Io2), ces relations sont :

(S_1=C, « devant ») : « Io1 est devant Io2»

(S_2=P, « est intérieur de ») : « Io1 est intérieur de Io2 »

(S_3=T, « touche ») : « Io1 touche Io2 »

(S_4=D, « déconnecté de ») : « Io1 est déconnecté de Io2 »

Des relations directionnelles qui sont invariables par rapport aux transformations géométriques de base :

(S_5, à droite de ») « Io1est à droite de Io2 »

(S_6, à gauche de ») « Io1est à gauche de Io2 »

(S_7, au dessus de ») « Io1est au dessus de Io2 »

(S_8, au dessous de ») « Io1est au dessous de Io2 »

Deux autres relations basées sur les distances entre les objets d'image qui sont :

(S_9, proche de ») « Io1est proche de Io2 »

(S_{10}, loin de ») « Io1est loin de Io2 »

Une instance de la facette spatiale est représentée par un ensemble de GCs, chacun contenant deux types d'objets images liés par les relations spatiales précédemment définies. Un objet image est caractérisé par son centre de gravité io_g, son intérieur, io_i et sa frontière io_b. Pour calculer automatiquement les relations topologiques, deux objets Io1 et Io2 d'image sont caractérisés par des intersections de leurs ensembles d'intérieur et de frontière : io1_i ∩ io2_i, io1_i ∩ io2_b, io1_b ∩ io2_i et io1_b ∩ io2_b. Chaque relation topologique est définie par l'ensemble de ces intersections.

Par exemple : Io1 et Io2 sont déconnectés (S_4, Io1, Io2) si io1_i ∩ io2_i = ∅, io1_i ∩ io2_b = ∅, io1_b ∩ io2_i = ∅ et io1_b ∩ io2_b = ∅. L'intérêt de cette méthode de calcul se fonde sur l'association des relations topologiques à l'ensemble précédent de conditions nécessaires et suffisantes impliquant des attributs des objets spatiaux.

Le calcul des relations directionnelles entre Io1 et Io2 se fonde sur la position relative de leurs centres de la gravité.

(b) Spécification conceptuelle

Les objets images sont représentés par des concepts de type Io. Les relations spatiales sont représentées dans un treillis de relations SpC (voir figue V.13). Chaque pair d'Io est interconnecté par l'une des relations du SpC.

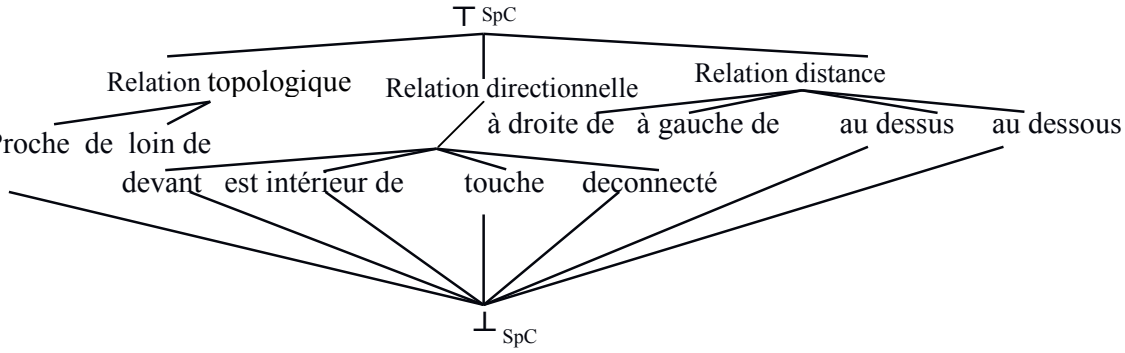


Figure V. 13 : Classification de relations conceptuelles

Une instance de la sous-facette spatiale est représentée par un ensemble de graphes conceptuels chacun comprenant un concept de type Io lié par la relation spatiale. Cette représentation est généralement décrite par le graphe suivant :

[Io] → (SpC) → [Io2]

Considérant Io1 et Io2 deux objets images représentant respectivement les concepts Bill Clinton et drapeau de l'exemple suivant : « les segments vidéo montrant Bill Clinton et au moins un partie du drapeau américain derrière lui ». Une représentation avec le formalisme des graphes conceptuels en exploitant une description spatiale est décrite par le graphe ci-dessous :

[Io1] → (devant) → [Io2]

V.II.3.1.1.4 La sous-facette mouvement

(a) Description

Plusieurs travaux ont été proposés pour le développement des algorithmes et des techniques pour l'estimation des mouvements de caméra et les trajectoires des objets dans le document vidéo. Ces travaux restent à un niveau signal et peu de propositions ont été faites pour la caractérisation des mouvements. Dans le cas de la vidéo, nous distinguons deux types de mouvements. Le premier concerne la description des mouvements de la caméra (considérés au niveau image) et le deuxième décrit les mouvements des objets.

Dans le cas où la caméra n'est pas fixe, un mouvement d'objet peut être « absolu » s'il est estimé par rapport un « fond » ou bien « relatif » s'il est estimé par rapport à une image. Dans le cas inverse (caméra fixe) les notions de « mouvement objet relatif » et « mouvement objet absolu » sont fusionnées.

Un mouvement général peut être décomposé en mouvements élémentaires. Ces mouvements élémentaires peuvent être : une « translation », une « rotation », une « homothétie » ou bien « irrégulier » (non linéaire). Excepté le mouvement « irrégulier » (qui permet uniquement de décrire le mouvement d'objets non rigides), les mouvements élémentaires peuvent être appliqués à la description des mouvements de la caméra et aussi des objets.

Nous distinguons huit mouvements élémentaires qui forment la base de notre cadre pour la caractérisation symbolique de l'aspect « mouvement » dans la vidéo : « M_1 = translation horizontale », « M_2 = translation verticale », « M_3 = translation en profondeur », « M_4 = rotation autour de l'axe horizontal », « M_5 = rotation autour de l'axe vertical », « M_6 = rotation autour de l'axe de profondeur », « M_7 = homothétie » et « M_8 = transformation irrégulière ».

Ces huit mouvements élémentaires sont indépendants les une des autres, ils peuvent être présents ou absents indépendamment les uns des autres et toutes les combinaisons de ceux-ci sont possibles. Les sept premiers correspondent à des degrés de liberté soit de la caméra soit de l'objet (bien que le dernier, « gonflement / rétrécissement », soit exceptionnel dans ce cas). Le huitième n'existe que dans le cas des objets et traduit une déformation de ceux-ci qui ne se réduit pas aux sept premiers mouvements élémentaires (mouvement de foule, vagues, personne qui marche ou se penche...). On peut toutefois le considérer comme indépendant des sept premiers dans la mesure où (en théorie au moins) on peut extraire un mouvement de translation, de rotation ou de grossissement même dans le cas de mouvements irréguliers (déplacement du centre de gravité pour la translation par exemple). On peut toujours trouver une « meilleure translation », une « meilleure rotation » et une « meilleure homothétie » (par moindre carrés par exemple), celles-ci permettent de déterminer l'absence ou la présence (et même l'orientation) des sept premiers mouvements élémentaires. La présence ou l'absence de mouvement résiduel une fois les sept premiers compensés détermine ce qu'il en est pour le huitième.

Les sept premiers concepts « mouvement élémentaires » correspondent à la nomenclature utilisée pour décrire les mouvements de la caméra (voir Figure II.5). Ils correspondent respectivement à : « track », « boom », « dolly», « tilt », « pan », « roll » et « zoom ».

Une description par les concepts mouvements est associée à l'image pour les mouvements de caméra et aux objets image (Io) pour les mouvements des objets. On garde la notation Io qui fait référence à la notion d'objet image dans le plan.

(b) Spécification Conceptuelle

La forme de la spécification conceptuelle est similaire à celle utilisée pour la texture et la couleur. La principale différence concerne le fait qu'une caractéristique élémentaire de mouvement peut être non seulement présente ou absente mais si elle est présente, elle peut être dans une direction ou dans une autre (vers la gauche ou la droite par exemple). Les 8 mouvements de base jouent le même rôle que les 11 couleurs ou les 11 textures de base et elles peuvent être combinées de la même façon pour produire des treillis de méta-concepts texture index et de méta-concepts texture requête. Les relations « has color » et « has texture » ont un équivalent noté « has motion ». Il y a par contre dans le cas du mouvement une extension de l'ensemble des valeurs possibles pour les composantes des vecteurs correspondants aux mouvements élémentaires. En plus du « 0 » et du « 1 » qui ont la même signification (présence ou absence du mouvement élémentaire) nous introduisons les valeurs « −1 » et « +1 » qui spécifient à la fois la présence du mouvement élémentaire et son orientation. Il convient de noter que nous utilisons ici des symboles qui ne s'identifient pas à des nombres entiers. En particulier « +1 » est différent de « 1 ». Des notations concises peuvent également être utilisées. Le tableau V.I indique la correspondance entre les symboles « −1 » et « +1 » et l'orientation correspondante pour les sept premiers mouvements élémentaires.

Type de mouvement	Valeur	signification
Translation horizontale	−1	translation vers la gauche
	+1	translation vers la droite
Translation verticale	−1	translation vers le bas
	+1	translation vers le haut
Translation en profondeur	−1	translation en arrière
	+1	translation en avant
Rotation autour d'un axe horizontal	−1	rotation vers le bas
	+1	rotation vers le haut
Rotation autour d'un axe vertical	−1	rotation vers la gauche
	+1	rotation vers la droite
Rotation autour d'un axe en profondeur	−1	rotation sens inverse d'aiguille d'une montre
	+1	rotation au sens d'aiguille d'une montre
homothétie	−1	réduire
	+1	agrandir

Tableau V.1 : Description des différents types de mouvement

Les méta-concepts mouvement peuvent encore être organisés en treillis car une relation d'ordre partiel entre eux peut également être définie. Celle-ci généralise celle qui est utilisée dans le cas des méta-concepts de couleur ou de texture. En ce qui concerne la relation générique – spécifique, les valeurs de composante « −1 » ou « +1 » induisent une relation de spécificité par rappor à un méta-concept ayant la valeur « 1 » et identique par ailleurs. Ces relations d'ordre partiel permettent encore de déterminer si un méta-concept présent dans une requête peut être projeté sur (s'il généralise) un méta-concept présent dans l'index d'un segment de vidéo.

Une instance de la sous-facette mouvement est représentée par un ensemble de graphes conceptuels chacun comprenant un concept de type Io lié par la relation conceptuelle « has motion » à un méta-concept mouvement. Cette représentation est similaire pour le graphe index et le graphe requête. Elle est généralement décrite par le graphe suivant :

[Io] → (has motion) → [M_{xx}] (pour les objets image)

ou :

[VS] → (has motion) → [M_{xx}] (pour la caméra, VS : Video Segment)

Si par exemple l'objet Io a un mouvement de translation à gauche, une représentation dans le formalisme des graphes conceptuels sera comme suite :

[Io] → (has motion) → [<M_1:−1, M_2:0, M_3: 0, M_4:0, M_5: 0, M_6: 0, M_7:0, M_8:0>]

ou :

[Io] → (has motion) → [<translation à gauche>]

Notons qu'un objet image dans un document peut être décrit par une conjonction de mouvements élémentaires. Par exemple un objet image peut se déplacer à la fois vers la gauche et vers le haut tout en tournant dans le sens des aiguilles d'une montre. Dans le cas des requêtes, on peut aussi avoir une disjonction de mouvements (<mvt>$_{OU}$). Par exemple, « rechercher les objets images se déplaçant vers la droite ou vers le bas ».

La représentation des cas de conjonction et disjonction des couleurs par des graphes conceptuels est la suivante :

[Io] → (has motion) → [<mvt>$_{ET}$]

[Io] → (has motion) → [<mvt>$_{OU}$]

Avec <mvt>$_{ET}$ et <mvt>$_{OU}$ représentant respectivement une combinaison de 8 valeurs booléennes représentant les concepts texture. Par exemple :

Une requête spécifiant un Io ayant un mouvement de translation horizontale (gauche ou droite mais non nulle) et vers le bas contient le sous-graphe :

[Io] → (has motion) → [<M_1:1, M_2:−1, M_3: 0, M_4:0, M_5: 0, M_6: 0, M_7:0, M_8:0>$_{ET}$]

Une représentation de la requête « rechercher un mouvement de caméra panoramique à droite ou zoom out » contient le graphe suivant :

[VS] → (has motion) → [<M_1:+1, M_2:0, M_3: 0, M_4:0, M_5: 0, M_6: 0, M_7:0, M_8:−1>$_{OU}$]

Notons que notre description des objets en mouvement se limite à la description de mouvement apparent des objets « rigides ». Une annotation des objets « déformable » plus précise que « mouvement irrégulier » nécessite une analyse plus fine qui consiste à décomposer l'objet en des sous-objets « non déformables ». Une possibilité d'extension de notre proposition est donc envisageable par la prise en compte des objets « déformables » pour la modélisation de la sous-facette mouvement. À partir de la modélisation des mouvements de ces objets, on peut inférer des descriptions sémantiques associées telles que marcher, se pencher, …

Dans le cas des objets « non rigides » où le mouvement peut être segmenté en des multiples phases durant un même plan, une micro segmentation du plan est nécessaire afin d'extraire des mouvements élémentaires.

Un autre cas important pour la modélisation du mouvement et que nous n'avons pas pris en compte dans notre modèle consiste à décrire le mouvement réel dans une scène entre deux (ou plus) objets par d'autres informations symboliques telles que par exemple « se rapprocher », « s'éloigner », etc.

V.II.3.1.2 Facette Sémantique

La facette sémantique permet de modéliser l'ensemble des descriptions de haut-niveau dans un document vidéo. Nous distinguons trois sous-facettes dont chacune décrit un type de média spécifique (le visuel, l'audio et le texte).

Dans ce qui suit, nous allons détailler la modélisation des sous-facettes visuelle audio. Notons que nous ne détaillerons pas la modélisation de la sous-facette texte pour la raison qu'elle est similaire à celle de l'audio.

V.II.3.1.2.1 Sous-facette visuelle

Cette facette permet de décrire l'information visuelle contenue dans le document vidéo. La sémantique décrite dans cette facette est portée par des objets images (Ios) qui correspondent à des concepts. Soient Io1 et Io2 deux objets images. Dans l'exemple1 Io1 et Io2 correspondent respectivement aux concepts « Bill Clinton » et « drapeau ».

(a) Modèle de la sous- facette visuelle

Si les images fixes sont décrites uniquement par les aspects physiques de l'image tels que la couleur, la texture, etc.; les images animées sont plus riches au niveau sémantique. En effet, la description du mouvement des objets et le suivi de leurs trajectoires ainsi que la description de mouvement de caméra (panoramique, zoom,…) sont des éléments non négligeables pour comprendre le contenu de la vidéo.

L'intégration de l'information au niveau signal et au niveau conceptuel est cruciale puisqu'elle enrichit la structure d'indexation et permet de combiner les aspects signal et sémantique dans le processus de recherche.

Les objets image sont représentés par des concepts Io et des concepts sémantiques visuels qui sont organisés dans un treillis ordonné par de relation de type spécifique/générique Un exemple de la sous-facette visuelle est représentée par un ensemble de graphes conceptuels, chacun contenant un type de Io lié à un VSC (Visual Semantic Concept) par une relation conceptuelle (apparaît). On se base dans ce sur la perception visuelle humaine afin d'interpréter ces informations. Ce type de modélisation permet en fait de répondre à des requêtes spécifiques au contenu visuel.

Le graphe de base pour la génération des représentations liées à cette facette est :

[Io] → (apparaît) → [VSC].

Exemple : « les segments vidéo montrant Bill Clinton et un drapeau », le premier objet Io représente Clinton et le second Io représente un drapeau.

[Bill Clinton] → (apparaît)

[drapeau] →(apparaît)

(b) Extraction de concepts visuels

Les concepts sémantiques visuels (VSC) résultent de l'interprétation du contenu visuel des objets images. Ceux-ci sont extraits pour établir une liste de concepts sémantiques visuels (VSC) qui peuvent être utilisés pour indexer le plan vidéo représenté par une image. Cette liste est enrichie avec des concepts fournis par l'outil Video-Annex [Lin 03]. Cet outil est mis en place pour des annotations semi-automatiques et l'exploitation du contenu visuel sémantique de la vidéo dans le cadre des évaluations TREC vidéo. L'objectif principal est de fournir des annotations collaboratives qui vont servir à une base d'apprentissage pour l'extraction des traits visuels. Nous monterons dans les figures suivantes (V.14 et V.15) un extrait des treillis de concepts proposé dans l'interface de Video-Annex.

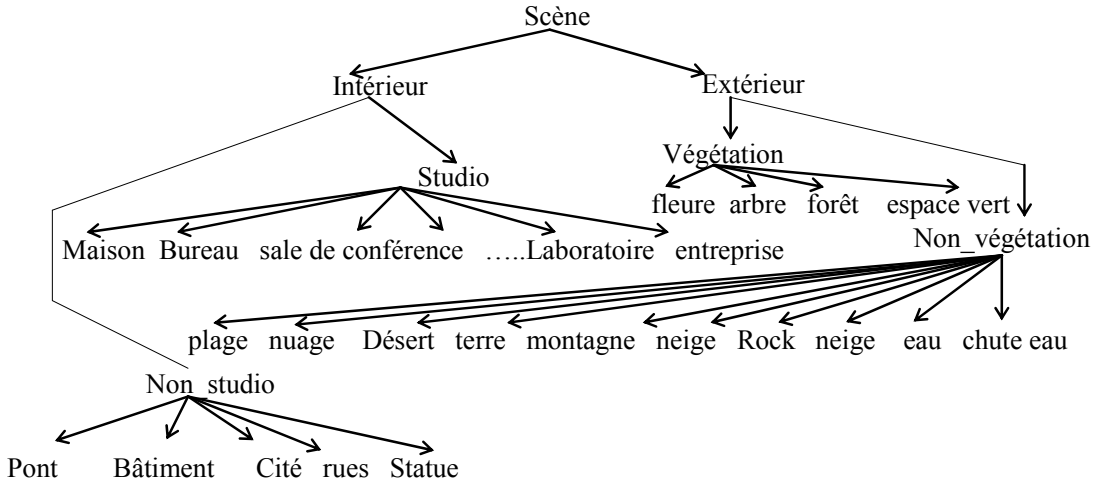


Figure V. 14 : Treillis de concepts visuels (1)

La représentation du contenu audiovisuel (ou contenu vidéo) avec le formalisme des graphes conceptuels est en fait une représentation hiérarchique qui a comme racine le concept abstrait tel que le concept image dans le cas du contenu visuel. Chaque graphe conceptuel est composé d'éléments provenant de deux treillis distincts : le treillis de concepts et le treillis de relations. La figure suivante illustre un ensemble de concepts pour une description générique du contenu visuel d'une portion vidéo :

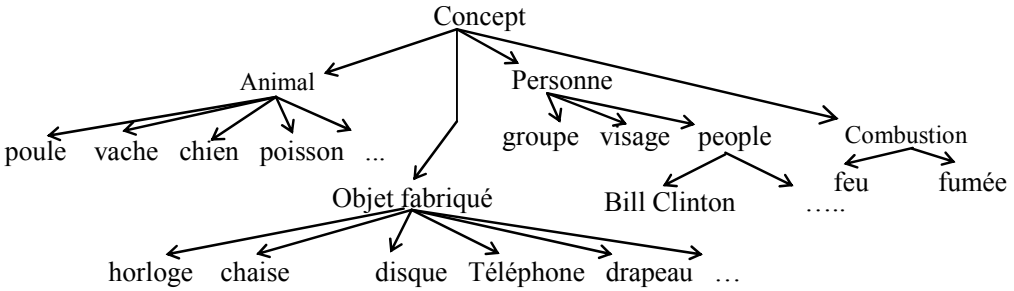


Figure V. 15 : Treillis de concept visuel (2)

(c) Exemple de description

La figure suivante montre un exemple de description du contenu visuel dans le cadre de la modélisation de la facette sémantique visuelle.

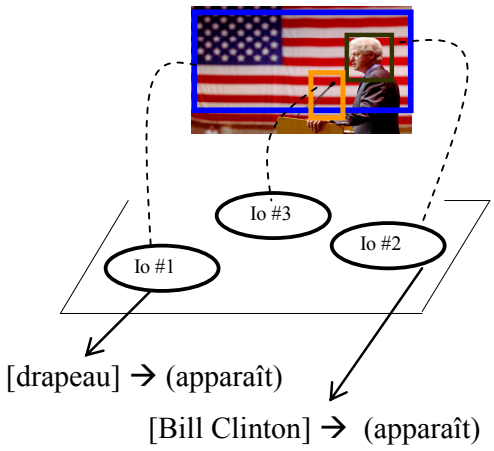


Figure V. 16 : Un exemple de description intégrant des concepts visuels sémantiques

V.II.4 Modélisation du Contenu Audio

L'audio, constitue une composante du document vidéo ayant un contenu riche en information symbolique. Le contenu audio constitue une base pour un grand nombre d'approches proposées dans la littérature. Ces approches se distinguent par la spécificité du cadre applicatif généralement restreint à des applications données.

L'audio possède suffisamment de caractéristiques pour constituer en lui-même le fondement de plusieurs travaux de recherche [Pinquier 01], [Gauvain 02], [kwon 02], [Liu 98], [Harb 02] Du point de vue du contenu sémantique, la parole est plus sollicitée que d'autres types d'informations, tels que par exemple la musique ou le bruit [Pinquier01].

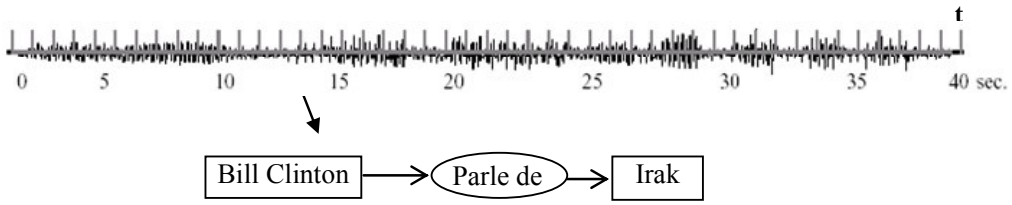
Un aspect important pour l'exploitation de l'audio est la reconnaissance automatique de la parole [Gauvain 02]. Il s'agit d'une transcription textuelle de la parole contenue dans le document qui permet d'avoir énormément d'information sur le contenu. Celle-ci permet d'effectuer des recherches efficaces sur de simples informations textuelles dans des bases de données audiovisuelles. La technique de transcription automatique a été mise en place dans le

cadre de plusieurs travaux de recherche. Les transcriptions viennent généralement avec une segmentation en locuteurs. Chaque segment contient la parole d'un seul locuteur non identifié à ce stade.

V.II.4.1 Spécification conceptuelle et description du contenu audio

En se basant sur la composante parole dans du contenu audio, une autre manière d'indexer le contenu audio qui consiste à déterminer qui parle par la détection et la reconnaissance de l'identité des locuteurs [Rodriguez04] dans le document. Le résultat du processus d'indexation sera de la forme suivante : associer l'identité de locuteur aux segments appropriés locuteur A (seg_1, seg_i, …..), locuteur B (seg_k, seg_m), etc.

La figure V.17 donne un exemple de représentation sous forme de graphe conceptuel du contenu audio d'un segment vidéo dans lequel Bill Clinton fait un discours sur l'Irak.



[Bill Clinton] → (parle de) →[Irak]

Figure V. 17 : Exemple de représentation du contenu audio avec les GCs

La modélisation du contenu audio peut être une tâche plus compliquée que celle de l'image. D'une part, elle dépend uniquement du texte transcrit. D'autre part, on sait qu'aucun processus automatique n'est jamais parfait et donc qu'on aura forcément dans la transcription des erreurs résultant de la reconnaissance automatique. Celles-ci peuvent influencer les résultats de recherche et la pertinence des réponses.

La description du contenu audio a pour objectif de permettre aux utilisateurs de formuler des requêtes sous une forme simple (en langage naturel). Dans ce contexte, l'utilisation de simples mots clés ne permet pas d'obtenir la meilleure précision possible. Nous choisissons donc de mettre en avant une description conceptuelle et relationnelle du contenu.

Un concept qui apparaît dans le graphe fait référence à une description symbolique du contenu. On peut distinguer des cas spécifiques c'est à dire des graphes destinés à représenter un seul type de flux et ceci par l'existence des relations conceptuelles implicites ou explicites.

Dans la figure V.18, nous représentons une représentation abstraite du contenu audio sous forme graphique en utilisant le formalisme des graphes conceptuels.

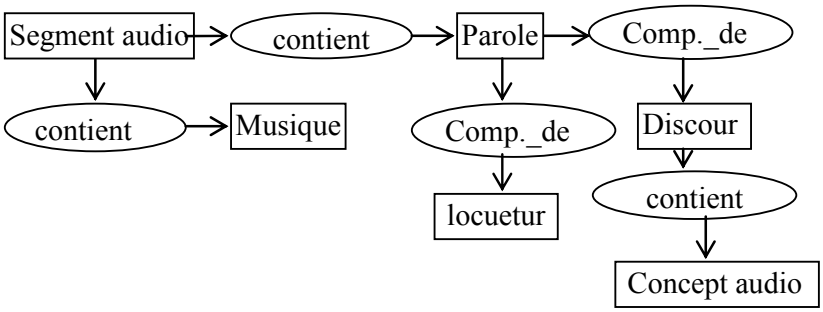


Figure V. 18: Description générique du contenu audio

Une représentation plus concrète avec le formalisme de GCs est illustrée par la figure suivante.

Exemple de description : « *le président américain fait un discours sur l'Irak devant la maison Blanche »*

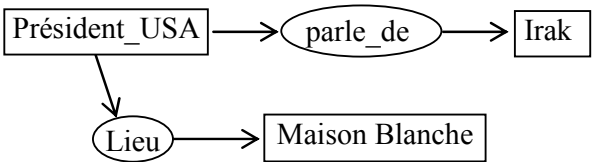


Figure V. 19 : Description du contenu sous forme de GCs

Une description conceptuelle n'est pas forcément la conséquence d'une interprétation par un opérateur humain de ce qu'il voit ou bien de ce qu'il entend. En effet, il existe des descriptions conceptuelles qui résultent de l'analyse du contenu signal pour la détection des actions spécifiques telles que par exemple une action d'applaudissement.

Pour le cas de l'information visuelle, la description fait intervenir différentes interprétations des sous-facettes (couleur, texture, spatiale, etc.). Chacune de ces sous-facettes fournit une description partielle du contenu visuel. Il s'agit en fait d'associer à chaque passage une description qui prend en compte les éléments d'information ainsi que les interactions qui peuvent exister entre eux. Ces interactions peuvent être illustrées par différents types de relations conceptuelles (spatiales, temporelles, etc.). Les relations spatiales dont il est question ici sont celles qui sont mentionnées dans le discours transcrit. Il s'agit par exemple de reconnaître le lieu (pays, ville, …) d'où un intervenant (reporter) parle. Cette description se situe au niveau symbolique et elle est différente de la spécification conceptuelle dans le cas du contenu visuel (sous-facette spatiale).

Pour une mieux comprendre la tâche de modélisation du contenu audio, nous proposons dans ce qui suit une structure de représentation générique :

[Flux audio] → *(Description structurelle)* → *[Transcription]*

[Transcription] → *(Composé de)* → *[Objet audio]*

[Objet audio] → *(Description symbolique)* → *[Concept audio]*

[Objet audio] → *(Description spatiale)* → *[Objet audio]*

[Concept audio] → *(Description temporelle)* → *[Concept audio]*

Notons que la modélisation de la sous-facette audio consiste à décrire uniquement le contenu symbolique de l'audio. Les relations spatiales dont il est question ici sont celles qui sont mentionnées dans le discours transcrit.

V.II.4.2 Segmentation audio et structure du document

L'objectif de la segmentation basée sur le contenu audio est d'obtenir des segments cohérents. Chaque segment possède un contenu spécifique : musique, bruit ou parole. Pour la parole, un autre type de segmentation est possible comme la segmentation du document suivant le tour de parole (changement de locuteur où chaque segment devra contenir la parole prononcé par une seule personne). Dans la littérature, plusieurs travaux sur la segmentation automatique de la parole ont été proposés. Dans [kwon 02], une technique de segmentation du document est basée sur un calcul par distances pondérés des points de changement de locuteur dans le flux audio. L'importance de ce type d'approche apparaît surtout lorsqu'il s'agit de détecter plusieurs locuteurs différents. D'autres techniques plus génériques de segmentation exploitent les caractéristiques du contenu audio pour segmenter le contenu telles que par exemple la détection de passages audio contenant un silence significatif ou bien aussi la détection des jingles qui peuvent aussi servir à inférer des transitions.

Nous nous intéressons à la première catégorie de segmentation (segmentation selon le locuteur) pour la modélisation de la sous-facette audio. Cette catégorie de segmentation est aussi considérée comme une tâche relevant du domaine de traitement automatique de la parole.

Notons aussi que chaque type de document vidéo (émissions, sportives, journal télévisé, etc.) possède une structure qui lui est propre. Dans tous les cas, pour le contenu audio, deux éléments sont nécessaire pour une description sémantique : il s'agit de déterminer « qui parle » et « de quoi on parle ».

Prenons l'exemple des journaux télévisés. Dans ce type de document, le concept locuteur constitue l'acteur principal. Cet élément d'information peut être exploité pour étiqueter les différents segments du journal. D'une manière générale, il existe deux façons pour segmenter le contenu d'un journal télévisé : La première se base sur le découpage thématique du contenu. Un journal sera vu comme étant une suite des thèmes (politique, sport, météo) séparés souvent par des jungles, publicités etc. La deuxième exploite la structure plateau / reportage pour segmenter le contenu du journal. C'est souvent au cours de cette deuxième forme de segmentation qu'il y a transition et changement de locuteur. La figure suivante montre un exemple d'organisation du contenu d'un journal télévisé.

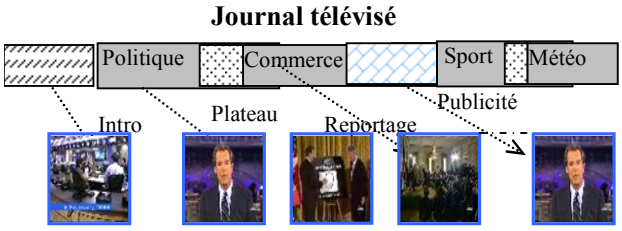


Figure V. 20 : Structure d'un journal télévisé

Le contenu audio permet de décrire toute information : soit de compléter et d'enrichir les descriptions issues du contenu visuel, soit d'apporter une interprétation indépendante du contenu du segment là où le contenu visuel ne le permet pas.

Un objet audio (Ao) peut être n'importe quel élément sémantique du contenu audio, souvent des concepts extraits automatiquement à partir de l'audio tel que par exemple l'identité d'une personne, le lieu, etc. L'information sémantique contient parmi d'autres deux éléments d'informations particuliers. Le premier élément consiste à spécifier qui parle dans chaque segment vidéo (spécifié par la relation parle). Le deuxième élément concerne de l'information dont on parle (spécifié par la relation parle de). L'exemple1 sera décrit par Bill Clinton parle de l'Irak ou bien Bill Clinton parle.

Il existe donc deux relations conceptuelles audio (« parle » et « parle de »). La distinction entre ces deux relations conceptuelles, bien qu'elles puissent exprimer la même signification, permet de tenir compte de la spécificité des requêtes. D'autres relations peuvent être extraites de la transcription (description textuelle d'une action par exemple).

Les concepts et les relations conceptuelles sont organisés dans des treillis partiellement ordonnés par la relation spécifique / générique IS-A (≤). Par exemple, « Irak ≤ pays » signifie qu'Irak est un concept spécifique du concept pays.

V.II.4.3 Modélisation de la sous-facette Audio

Comme nous l'avons déjà mentionné précédemment, les objets audio sont représentés par les concepts Ao. Chaque segment vidéo est représenté par un ensemble de graphes conceptuels comme dans le cas de la modélisation du contenu visuel. Chaque graphe contient des Aos interconnectés pas par des relations conceptuelles (A_{cr} : A*udio* ***c****onceptual* ***r****elation*).

La génération basic de graphe de représentation est la suivante :

[Ao1] → (parle de) → [Ao2] ou bien [Ao1] → (parle).

Si on considère l'exemple suivant: « Clinton parle de l'Irak », cette description peut être formulée par les deux sous graphes tels que :

[Clinton] → (parle)

[Clinton] → (parle de) → [Irak]

Si on considère par exemple deux objets audio (Ao1 et Ao2), ces relations sont (a_1 = Ao1, **parle**) avec Ao1 est dans la classe de concept personne. (a_2 = Ao1, **parle de,** Ao2**)** qui se traduit par « Ao1 parle de Ao2 ». Ao2 peut être n'importe autre concept cité dans le discours.

Nous distinguons trois catégories de concept sémantique audio (Asc) : (Asc_1 = personne, Asc_2 = lieu, Asc_3 = organisation). Ces concepts sont classés dans des treillis comme montré dans la figure V.21 :

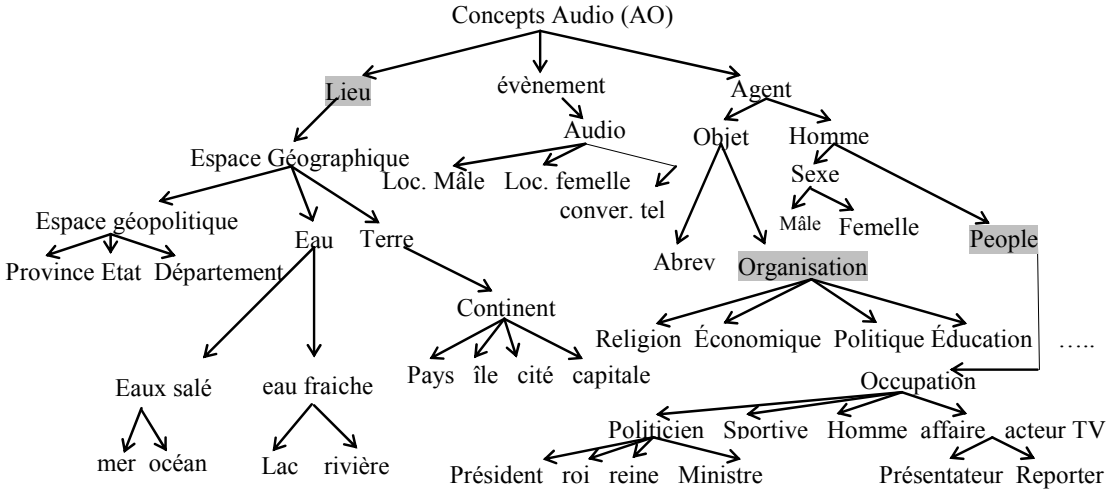


Figure V. 21 : Treillis de concepts audio

V.II.5 Représentation sous forme d'un Graphe unique

La description multifacette et multimodale pour la représentation du contenu vidéo a pour objectif de formuler des descriptions basées sur les caractéristiques audiovisuelles d'un document vidéo. Ces caractéristiques sont généralement dépendantes du type de média. En effet, pour le cas du contenu audio par exemple, la description est basée sur l'information contenue dans le contenu audio (parole, musique,…). La segmentation vidéo dépend du changement des caractéristiques audio (changement du locuteur pour la parole). Dans notre proposition nous avons exploité la parole et particulièrement les transcriptions automatiques.

La modélisation du contenu vidéo avec des concepts et de relations conceptuelles permet de tenir compte de ces différentes spécificités (audio et visuelles). Il en résulte un ensemble de représentations sous forme de sous-graphes pour chaque type média. Cette distinction n'exclut pas le fait d'avoir une représentation unique du contenu. En effet, si on prend le cas de l'exemple suivant : « Bill Clinton fait un discours sur l'Irak » (décrit dans la figure V.22), cette description peut être interprétée de différentes manières. En nous basant sur le contenu audio ou bien par l'intermédiaire de l'information visuelle, nous pouvons par exemple apercevoir une information textuelle à l'écran indiquant qu'il s'agit d'un discours de Bill Clinton sur l'Irak.

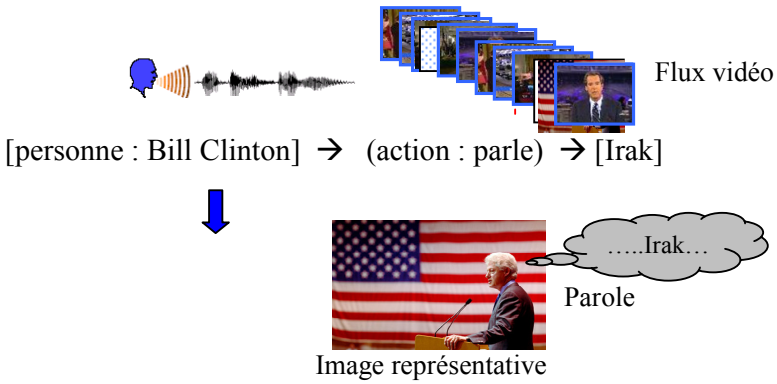


Figure V. 22 : Exemple de description par correspondance Audio / visuel

Deux éléments sont essentiels pour pouvoir regrouper et unifier ces représentations : les interprétations des graphes conceptuels et l'aspect de continuité temporelle dont dépend un document vidéo.

Les représentations avec le formalisme des graphes conceptuels sont illustrées par l'ensemble des concepts et des relations conceptuelles. Ces représentations peuvent apparaître dans des graphes séparés bien qu'il s'agisse parfois d'une description du même concept. Dans l'exemple précèdent, il s'agit du même Concept «Bill Clinton », ce concept est interprété différemment. C'est un concept visuel (on voit son image) ou bien un concept audio (on entend sa voix).

Pour pouvoir intégrer ces descriptions dans un même schéma, nous proposons d'une part, d'associer une information temporelle à chaque description et d'autre part de spécifier les types des relations conceptuelles dans les graphes correspondant à chaque flux (audio ou visuel).

Nous détaillons dans ce qui suit les types de relations utilisés. Ces relations sont classées dans des treillis tels que celui présenté dans la figure V.23 :

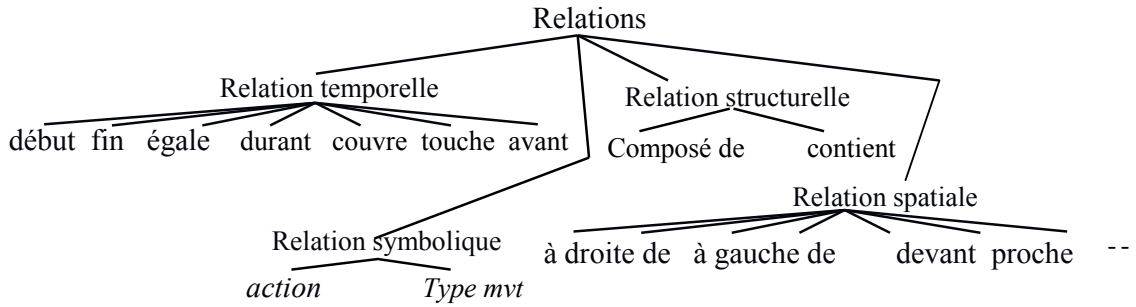


Figure V. 23 : Classification de relations conceptuelles

Les concepts utilisés peuvent être des symboles faisant référence à des objets qui apparaissent dans la vidéo tels que par exemple le suivi d'un objet en mouvement, l'identification de cet objet. On peut distinguer deux catégories de relations :

V.II.5.1 Relation implicite

Dans ce type de relations, un concept fait référence au même objet décrit dans le cas visuel et audio. En effet, la référence à un media peut être retrouvée par une relation conceptuelle implicite. Par exemple, à partir d'une description de type 'un concept *C1 parle de d'un concept C2*) on déduit que C1 est l'acteur (locuteur) et C2 est le sujet (information dans le discours).

La même chose pour le contenu visuel, par exemple, on peut associer une interprétation du contenu en observant juste ce que nous pouvons apercevoir à l'écran (C1 apparaît). Un exemple le concept voiture rouge est représenté par une image-clé dans ce deux cas et comme le montre la figure V.24. Il s'agit élément d'information (C1 = Bill Clinton), qui apparaît et parle.

[Bill Clinton] → (parle)
→ (apparaît)

Figure V. 24 : Exemple de relation implicite

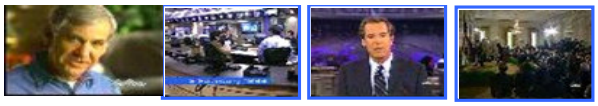
V.II.5.2 Relation explicite

Ce type de relations est utilisé pour regrouper l'ensemble des descriptions plus ou moins variables associées à un segment vidéo. Par exemple, pour un même segment vidéo, la description du contenu audio (une personne qui parle par exemple) ne correspond pas à celle spécifique au contenu visuelle (exemple : une personne qui apparaît à l'écran). C'est le cas d'un journal télévisé où un le présentateur parle d'un sujet et en même temps à l'écran on a l'image d'autres personnes qui sont en lien avec ce que dit le présentateur.

En utilisant uniquement des relations implicites, on obtient deux descriptions plus ou moins distinctes (spécifique à chaque type de média). Pour enrichir ces descriptions, nous proposons d'intégrer dans les représentations un ensemble des relations explicites permettent de relier des concepts situés dans des graphes séparés.

Les relations explicites peuvent être issues d'une interprétation manuelle en utilisant de connaissance externe. Par exemple, décrire le concept qui apparaît à l'écran par des informations qui ne figurent pas dans le contenu. Ces relations peuvent être aussi issues d'une analyse du contenu bas niveau telle que les mouvements de caméra, les trajectoires des objets etc., (relations spatiales par exemple).

Dans la figure V.25, nous donnons un exemple de description avec des relations explicites.



(a) (b) (c) (d)

(a) [homme] → (a pour nom) → [Jacques]

(b) [scène] → (pos. spatiale) → [Interieur] → (is a) → [studio]

(c) [Peter Jenning] → (is a) → [présentateur]

(d) [scène] → (is a) → [reportage] → (a propos de) → [thème : politique]

Figure V. 25 : Exemple de relation explicite

L'ensemble de ces relations permet d'interpréter le contenu en spécifiant un type de média particulier ou bien en intégrant tous les médias dans la même description.

Les connaissances externes sont des informations utiles pour enrichir les descriptions du contenu vidéo. Nous proposons pour l'extension de notre modèle d'associer une ontologie. Nous détaillerons ultérieurement l'apport de cette extension pour la modélisation et la recherche par le contenu. Nous présenterons aussi ce que nous avons réalisé dans ce contexte.

V.II.5.3 Réseaux des graphes

L'information temporelle permet de structurer les différentes descriptions et de les classer dans le même ordre que leur apparition dans la vidéo. Les différentes descriptions du document sont regroupées un schéma constituant une sorte de réseau des graphes. On appelle réseau des graphes, l'ensemble des descriptions sous forme de graphes conceptuels associées aux documents vidéo.

Par rapport à ce que venons de décrire dans la partie I et II de chapitre, La représentation conceptuelle du contenu vidéo sous forme d'un graphe unique obtenu par la combinaison (opération de jointure dans le cas de GCs) entre les différentes facettes reliant les flux visuel et audio. Pour le même exemple : « les vidéos montrant Bill Clinton qui parle de l'Irak et au moins une partie du drapeau américain visible », une description sous forme d'un graphe unique sera comme suit :

JOIN (visual graph)

[Io1] → (vsc) → **[Clinton]**

[Io2] → (vsc) → [drapeau]

[Io2] → (has_color) → [<C_1:1, C_2:1, C_3:1, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0> $_{ET}$]

[Io2] → (has_tex) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5: **1**, T_6: 0, T_7:0, T_8:0, T_9:0, T_{10}:**1**, T_{11}:0> $_{ET}$]

[Io1] → (proche de) → [Io2]

AND (audio graph)

[Ao1] → (is a) → **[Clinton]**

[Ao2] → (is a) → (Irak)

[Ao1] → (parle)

[Ao1] → (parle de) → [Ao2]

BY (concept node)

[Clinton]

Conclusion

Nous avons détaillé dans les deux premières parties de ce chapitre notre proposition pour la modélisation du contenu vidéo. Ce schéma est générique du fait qu'il ne dépend pas d'un genre de vidéo particulier. Il est également assez complet vu ses aspects multifacettes et multimodales.

Le choix de spécifier deux formes de représentation (générique et spécifique) a pour objectif de mettre œuvre l'ensemble des caractéristiques audiovisuelles du document vidéo. Dans la représentation générique, nous modélisons les traits communs tels que les informations temporelles, les descriptions événementielles. Par contre, dans les représentations spécifiques, il est plutôt question d'une modélisation par type de média.

Nous avons ensuite proposé d'intégrer les différentes descriptions afin de mettre en œuvre un modèle qui soit unique et représentatif du contenu.

Dans la partie III de ce chapitre, nous proposons une extension de notre modèle par la projection des différentes descriptions formulées sur des ontologies afin de les enrichir et d'élargir le vocabulaire de description du contenu vidéo. Nous spécifions aussi le cadre applicatif (ses caractéristiques, sa structure).

Partie III - Modélisation Multifacette et Multimodale - Proposition d'une Ontologie pour Enrichir le Modèle

V.III.1 Présentation générale

La description du contenu vidéo est une tâche difficile vu d'une part la variété de son contenu et la masse de données disponible et d'autre part les techniques actuellement disponibles pour extraire des critères objectifs permettant la description. Cette description se fait généralement par l'intermédiaire des « informations textuelles » de ces contenus.

Depuis le début de ce mémoire, le problème que nous posons consiste à modéliser le contenu d'un document vidéo afin de répondre, de manière précise, aux besoins en information d'un utilisateur. Le système de recherche d'information vidéo doit être capable de retrouver des séquences vidéo dont la description ne coïnciderait pas exactement avec les termes employés pour formuler la requête. Pour illustrer ce point, nous présentons l'exemple ci-dessous (illustré dans la figure V.26)

"….u.n. secretary general kofi anan smash into baghdad is a critical opportunity to achieve the outcome that all of us would prefer a peaceful and principled end to this crisis the secretary general is backed by the unambiguous position of the security council saddam hussein must give the weapons inspectors full free unfettered access to all suspected sites anywhere in iraq…"

[Extrait du discours de Bill Clinton sur l'Iraq]
ABC NEW - 1998

Figure V. 26: Exemple d'illustration

Imaginons qu'un utilisateur souhaite « retrouver les segments vidéo montrant un discours de Bill Clinton sur l'Irak ». Une recherche effectuée sur cette description en utilisant des mots clés Bill Clinton, Irak et discours va bien retourner des segments ou au moins l'un de ces trois mots clés existe mais rien ne prouve que dans le ou les segment(s) retrouvé(s) existe bien un segment pertinent par rapport à la de requête. Pour remédier à ce problème, nous avons proposé tout d'abord de mettre en place un modèle qui soit suffisamment riche pour pouvoir décrire le contenu d'un document vidéo. Ce modèle se base sur une description du contenu vidéo par des concepts et des relations conceptuelles en utilisant le formalisme des graphes conceptuels (comme nous l'avons détaillé dans la partie I et II de ce chapitre). Nous avons également proposé d'enrichir ce modèle par des connaissances externes afin d'améliorer le processus d'indexation et de recherche. Ces données externes permettent d'enrichir les descriptions des requêtes utilisateurs et les descriptions utilisées pour indexer les documents. Ces bases de connaissances (souvent on utilise le terme « ontologie ») doivent permettre :

- De fournir un modèle permettant de représenter et d'enrichir la sémantique des index de manière à ce que le système soit capable de faire des raisonnements. Par

exemple, si un segment vidéo est indexé par des concepts « arbitre », « joueur » et « ballon », en utilisant les bases de connaissances, le système doit être capable d'inférer qu'il s'agit d'une scène de sport.

- De naviguer dans la structure des documents retrouvés et d'inférer le contexte.

Pour prendre en compte les besoins de précision, nous proposons de construire une ontologie qui va faciliter l'échange des descriptions et contribuer à améliorer la précision et l'efficacité du système de recherche d'information vidéo.

V.III.2 Construction d'ontologie

En se plaçant au niveau description symbolique, la vidéo peut être 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 segment vidéo une description pertinente par rapport à son contenu. C'est à ce niveau que la difficulté de ce processus se pose. En effet, à un même segment vidéo, nous pouvons associer plusieurs interprétations plus ou moins différentes.

Dans la perspective de décrire le maximum d'informations symboliques contenues dans le document, nous avons défini quelques catégories de concepts que nous considérons comme essentielles pour la représentation du contenu des documents vidéo. Nous décrivons dans ce qui suit trois de ces catégories (personne, objet, organisation) sont décrites dans ce qui suit.

- Les *personnages* sont des éléments d'informations souvent appelés « acteurs » dans un film par exemple. L'interprétation d'un segment 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 le même segment. En effet, nous pouvons spécifier l'identité, le grade, l'activité etc.
- Les *objets visuels* composés des entités (des concepts logiques) qui peuvent être reliées à une ou plusieurs régions dans une image clé d'un plan vidéo. Un objet visuel peut être défini aussi comme une collection de zones de l'image qui ont été groupées ensemble selon des critères définis par la connaissance du domaine. Ces objets devraient également satisfaire quelques conditions comme la conformité sémantique ou la représentation d'un objet réel pour les utilisateurs.
- Les *organisations* décrivent des éléments d'informations symboliques. Par définition, une organisation est un ensemble d'individus regroupés au sein d'une structure régulée dans le but de répondre/d'atteindre à un/des besoin(s)/objectif(s) déterminé(s). Elle peut prendre plusieurs formes qui varient selon les domaines (politique, sportive, média, etc.). Une organisation est souvent désignée par un acronyme. Par exemple l'abréviation ABC « *American Broadcast Channel* » désigne une organisation de type média.

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

La figure V.27 illustre un extrait de l'ontologie que nous avons construit. Notons que lors de la conception nous avons tenu en compte du genre de vidéo (journal télévisé) et de son contenu.

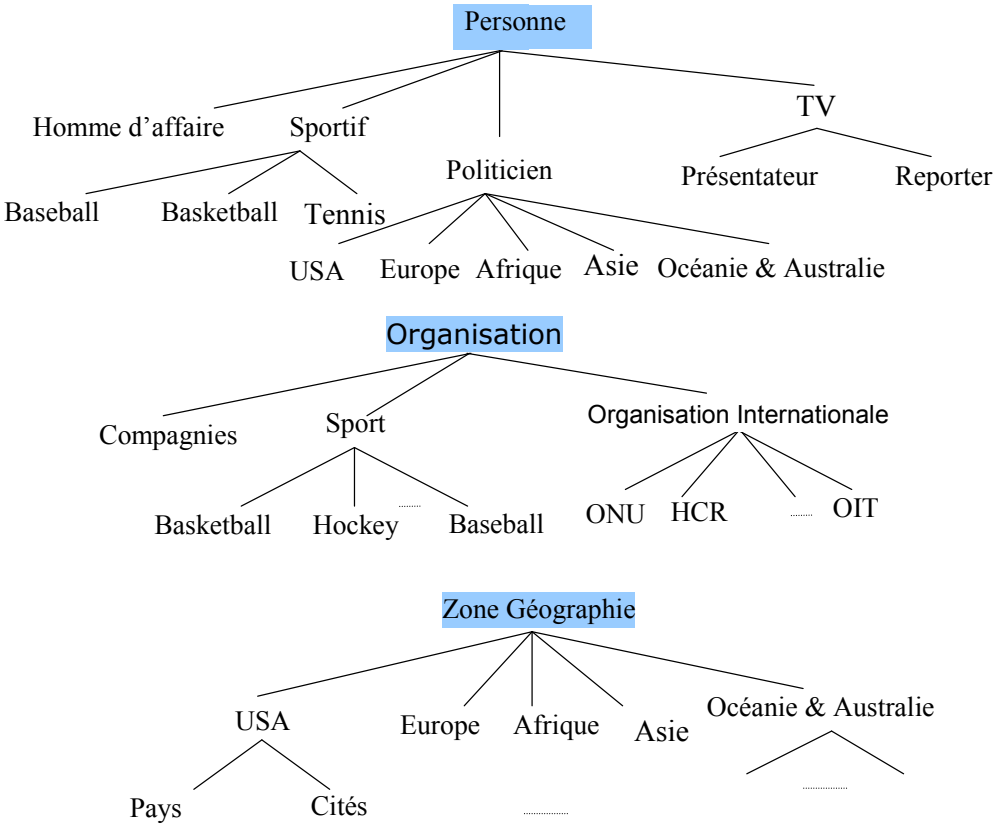


Figure V. 27: Exemple de construction d'ontologie

Discussion

L'utilisation des ontologies reste encore une problématique ouverte du fait de la variété du genre des documents vidéo. Même pour un type spécifique de document vidéo (cas des journaux télévisés par exemple), il n'est pas évident de mettre en place une ontologie que permet de représenter l'ensemble des contenus sémantiques vu la variété thématique dans ce genre de document (politique, sport, etc..).
Dans le cadre de notre travail, nous avons proposé une structure de représentation qui tient compte des éléments d'informations pertinents par rapport aux besoins des utilisateurs dans le document vidéo. Pour le moment, cette structure n'est pas complète mais elle suffit dans le cadre actuel de notre travail pour enrichir les descriptions issues de notre modèle de base.
Nous avons spécifié quelques catégories de concepts à partir desquelles nous avons construit nos ontologies. À l'heure actuelle, ces ontologies se présentent sous forme des hiérarchies de concepts.

Partie IV - Modèle d'Indexation et de Recherche Vidéo par le contenu

V.IV.1 Modèle de documents

Le modèle proposé permet d'intégrer des descriptions qui sont à la fois variées et génériques. Ce schéma est mis en œuvre de manière à répondre aux besoins des utilisateurs d'un système de recherche vidéo. Nous avons détaillé dans la partie modélisation (partie I et II) la sémantique de chaque facette au niveau de représentations générique et spécifique. La définition de la sémantique de chaque facette ou sous-facette est introduite par la spécification conceptuelle de ses éléments d'informations. Cette spécification est ensuite instanciée par une représentation avec le formalisme des graphes conceptuels pour former le modèle de document spécifique chaque facette ou sous-facette. Dans cette section, nous proposons de résumer formellement les représentations de notre modèle pour la description du contenu vidéo.

Les représentations index ainsi que les requêtes sont des instances du modèle de la vidéo qui est donné par l'agrégation de l'ensemble des modèles des facettes et d'un ensemble de relations représentant les liaisons entre les facettes :

Le modèle de la vidéo est défini par le 8-uplet (V_{sh}, M_{tem}, M_{eve}, $M_{sym,}$ M_{sig}, $L_{ss,}$ $L_{ssp,}$ L_{tp}) avec :

- V_{sh} est l'ensemble des segments vidéo.

- M_{tem}, M_{eve}, $M_{sym,}$ M_{sig} sont respectivement les modèles de facette temporelle, événementielle, symbolique et signal. Notons que la facette symbolique et signal sont eux même composées d'autres sous-facettes comme telle que la facette signal qui est composée des sous-facette couleur, texture, spatiale et mouvement.

- $L_{ss,}$ $L_{ssp,}$ L_{tp} sont des relations qui associent respectivement les objets audio et image de la facette symbolique, les objets spatiaux de la facette spatiale, les descriptions temporelle de la facette temporelle.

La figure V.28 présente une architecture générale de notre modèle. Nous pouvons distinguer trois composantes :

- ✓ La première concerne la tâche de modélisation. Nous avons récapitulé notre contribution pour la modélisation multifacette et multimodale basée sur le formalisme des GCs (parties I et II de ce chapitre).
- ✓ La deuxième détaille l'extension du modèle par la proposition d'enrichir le modèle par l'extension du vocabulaire de description en exploitant des données externes sous forme des ontologie ou des hiérarchies de concepts.
- ✓ La troisième et la dernière composante de l'architecture concerne la tâche d'application et l'utilisation de ce modèle pour l'indexation et la recherche par le contenu des documents vidéo.

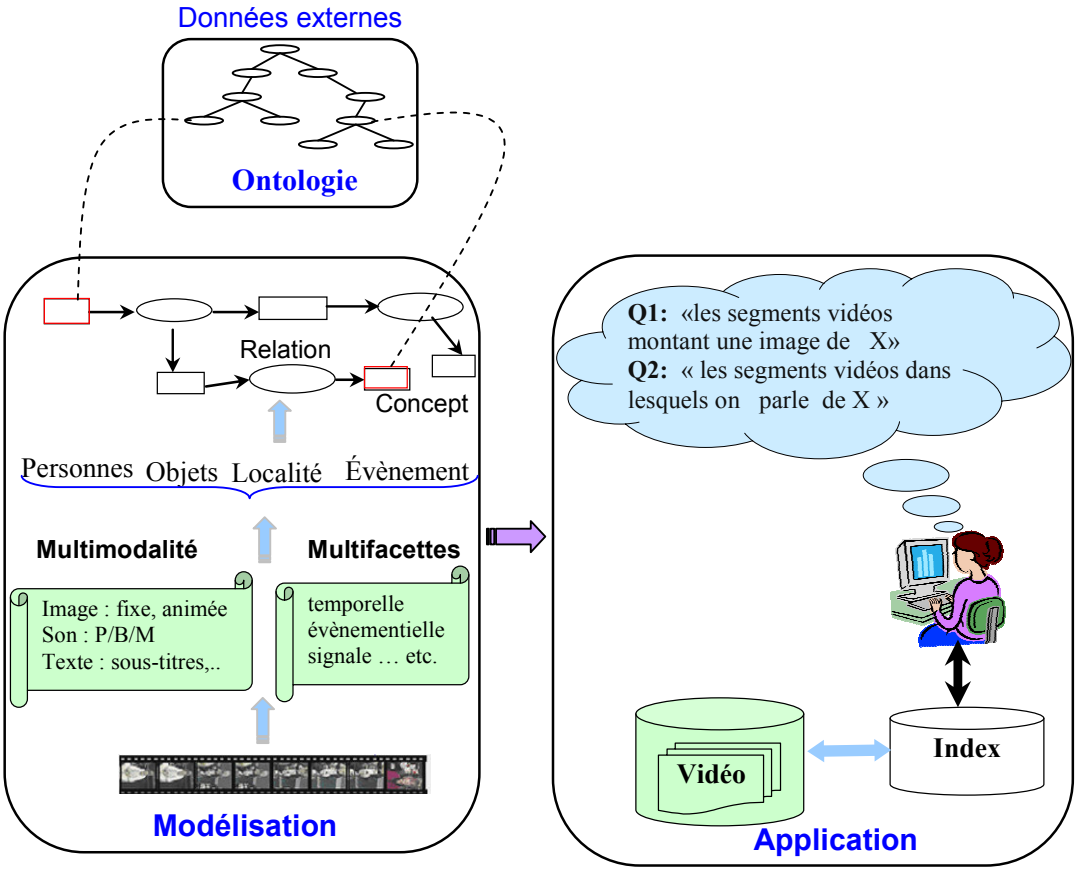


Figure V. 28 : Architecture générale du modèle

V.IV.2 Modèle de requêtes

Notre modèle conceptuel est basé sur des descriptions textuelles unifiées permettant à un utilisateur de d'interroger les différents descriptions visuelles et audio. Ceci optimise évidemment l'interaction d'utilisateur puisque l'utilisateur dirige le processus de recherche en explicitant ses besoins d'information au système. Le processus de recherche d'information est basé sur le formalisme des graphes conceptuels. En effet, chaque requête est également décrite sous la forme d'un graphe conceptuel. La représentation d'une requête utilisateur dans notre modèle est obtenue, comme les représentations d'index, par la combinaison des graphes conceptuels de toutes les facettes visuelles et audio.

Pour rechercher un document ou un segment du document vidéo, les requêtes doivent être riches et comporter des descriptions sur le contenu. Le langage de requête est identique au modèle de représentation du contenu vidéo tel que proposé précédemment. Une requête est alors définie par la combinaison de l'ensemble de facettes et types de média. Par conséquent, nous pouvons distinguer plusieurs catégories de requêtes :

a) Requête événementielle

La requête événementielle porte sur les actions (souvent réelles) dans le document. Par exemple, la requête : « rechercher les segments vidéo montrant un meeting ». Une

représentation avec le formalisme de graphe conceptuel de cette requête est donnée dans la figure suivante :

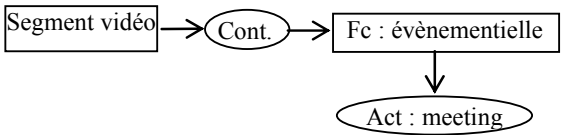


Figure V. 29 : Exemple de requête événementielle

b) Requête temporelle

La requête temporelle intègre des contraintes issues de la facette temporelle. Par exemple, la requête : « rechercher les segments vidéo montrant deux événements qui se terminent en même temps ». Une représentation avec le formalisme de graphe conceptuel de cette requête est donnée dans la figure suivante :

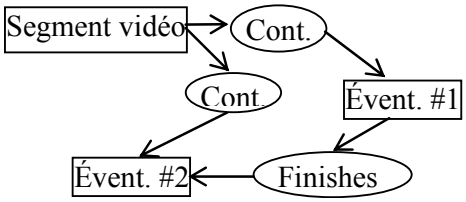


Figure V. 30 : Exemple de requête temporelle

c) Requête sémantique audio

La requête sémantique audio combine les descriptions reliées au contenu audio. Par exemple, la requête : « rechercher les segments vidéo dans lesquels Bill Clinton parle ». Une représentation avec le formalisme de graphe conceptuel de cette requête est donnée dans la figure suivante :

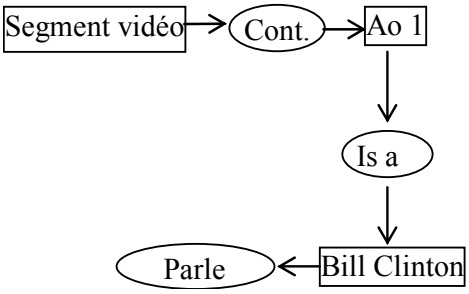


Figure V. 31 : Exemple de requête sémantique audio

d) Requête sémantique visuelle

La requête sémantique visuelle combine les descriptions reliées au contenu visuel. Par exemple, la requête : « rechercher les segments vidéo montrant Bill Clinton ». Une représentation avec le formalisme de graphe conceptuel de cette requête est donnée dans la figure suivante :

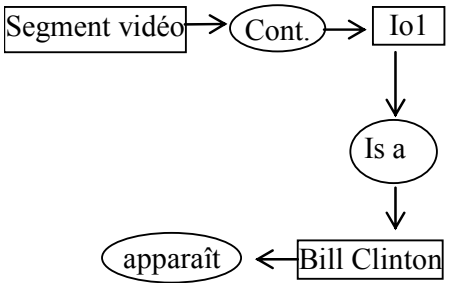


Figure V. 32 : Exemple de requête sémantique visuelle

e) Requête sémantique signal

La requête sémantique signal permet rechercher la vidéo en se basant sur des descriptions bas-niveau. Exemple, la requête : « Rechercher les segments vidéo montrant un drapeau ayant une couleur rouge, blanche et bleue et une texture rayé ».

[Drapeau]→(has_color)→[$<C_1$:**1**, C_2:**1**, C_3:**1**, …, C_9:0$>_{ET}$].

→(has_texture) →[$<T_1$:0, T_2:0, T_3:0, T_4:0, T_5 :**1**, …$>_{ET}$]

f) Requête multimodale

La requête multimodale permet de recherche la vidéo en se basant sur une description issue de plusieurs modalités (image, audio ou texte). Par exemple la requête :

« Rechercher les segments vidéo montrant Bill Clinton qui parle de l'Irak et où au moins une partie du drapeau américain visible »

Une représentation avec le formalisme de graphe conceptuel de cette requête est donnée dans la figure suivante :

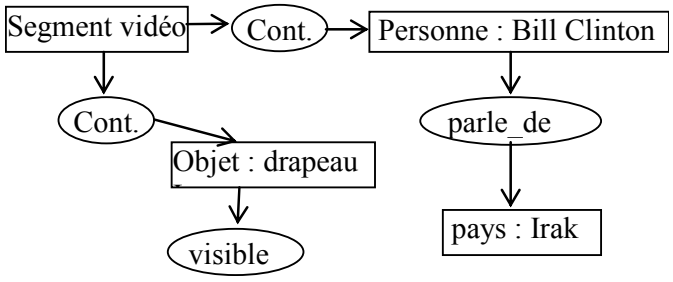


Figure V. 33 : Exemple de requête sémantique multimodale

V.IV.3 Modèle de Correspondance

La correspondance entre un document *d* (*d* pouvant être un document vidéo ou un segment de document vidéo) et une requête *q* est déterminée au moyen de l'opérateur de projection du graphe G_q représentant la requête dans le graphe G_d représentant le document. Il existe une *projection* du graphe G_q sur un graphe G_d s'il existe un sous graphe G'_d de G_d qui est une restriction (spécialisation) de G_q. Étant donnés une requête *q* et un document *d*, il peut exister zéro, une ou plusieurs telles projections. Nous noterons $\Pi(q,d)$ l'ensemble de ces projections.

Lorsque plusieurs documents correspondent à la requête selon une telle correspondance, il est nécessaire de les ordonner. Nous allons pour cela calculer une pertinence pour chacun de ces documents. Celle-ci sera calculée en combinant des mesures dites d'exhaustivité et de spécificité. Ce modèle de correspondance est basé sur une extension du modèle logique de Van Rijsbergen apportée par Nie [Nie 98]. Nous utiliserons une fonction F pour combiner les mesures d'exhaustive et de spécificité :

Pertinence (d, q) = F [E(d → q), S(q → d)]

L'exhaustivité exprime dans quelle mesure le plan satisfait toute la requête. Elle est mesurée par la valeur de E (d→q).

La spécificité mesure l'importance du développement du thème de la requête dans le document. Elle est donnée par la valeur S (q→d).

La fonction F

La fonction F est croissante en fonction de E et S et elle prend ses valeurs dans l'intervalle [0,1]. Cette fonction a les caractéristiques suivantes :

F(a,b) = 0 si a = 0 ou b = 0,

F(a,b) = 1 si a =1 et b = 1,

Nous avons retenu la fonction la plus simple : le produit.

$F(a,b) = a.b$

Dans ce qui suit nous allons présenter une instanciation des mesures d'exhaustivité et de spécificité.

(1) L'Exhaustivité E

Notations

q : requête

d : document

G_q : graphe représentant la requête

G_d : graphe représentant le document

$\Pi(q,d)$: ensemble des projections de G_q sur G_d

m : élément de $\Pi(q,d)$, projections de G_q sur G_d

G'_{dm} : restriction du graphe représentant le document en correspondance avec G_q selon la projection *m*

k : indice pour désigner les nœuds concepts en correspondance dans les graphes G_q et G'_{dm}
K_m : valeur maximum de l'indice k selon la projection m
c_{qmk} : $k^{\text{ième}}$ nœud de G_q selon la projection m
c_{dmk} : $k^{\text{ième}}$ nœud de G'_{dm} selon la projection m
C_{qm} : liste des $c_{qmk} = \{(c_{qmk})$ avec $1 \le k \le K_m\}$
C_{dm} : liste des $c_{dmk} = \{(c_{dmk})$ avec $1 \le k \le K_m\}$

$$E(q,d) = \underset{m \in \prod(q,d)}{MAX} \left[\left(\sum_{1 \le k \le K_m} I(c_{dmk}) \right) + M(C_{qm}, C_{dm}) \right]$$

La fonction I mesure l'importance d'un concept dans le document et la fonction M mesure la correspondance entre les concepts.

La fonction M (correspondance des concepts)

La fonction M est la divergence négative Kullback-Leibler [Zahi 01] entre les probabilités des concepts du document et les concepts de requête (dans le cas où les concepts ne sont pas ambigus $P(c) = 1$).

La forme négative de divergence Kullback-Leibler entre la requête et le document est :

$$M(C_{qm}, C_{dm}) = -KL(C_{qm}, C_{dm}) = \sum_{1 \le k \le K_m} P(c_{qmk}) \log \frac{P(c_{dmk})}{P(c_{qmk})}$$

Pour chaque concept c_d d'un document d, on peut calculer une probabilité $P(c_d)$ mesurant l'importance de ce concept dans le document. Par exemple, pour le contenu audio la valeur de $P(c_d)$ correspond au rapport entre le nombre d'occurrences (n_{cd})du concept c_d en le nombre total de concepts (n_d) dans le document d.

$P(c_d) = n_{cd} / n_d$

$P(c_q)$ est la probabilité d'un concept c_q dans requête q. Dans notre cas, nous supposons que les concepts dans les requêtes sont non ambigus. ($P(c_q) = 1$)

$$M(C_{qm}, C_{dm}) = -KL(C_{qm}, C_{dm}) = \sum_{1 \le k \le K_m} \log P(c_{dmk})$$

(2) La Fonction de Spécificité S

Notre hypothèse est que pour un utilisateur, les documents recherchés sont plus pertinents s'ils sont strictement limités aux concepts du graphe requête G_q. Or un document est souvent décrit par plus des concepts que la requête. Dans le cas de la recherche vidéo, un document

recherché peut correspondre à un document en entier ou à un segment de document (plan, segment audio, etc.).

Si on considère la requête suivante « rechercher les plans vidéos contenant des véhicules », les plans vidéo montrant des voitures et des vélos sont plus appropriés pour sa recherche que des plans contenant des voitures et des bâtiments. En effet, dans les treillis des concepts associés, voitures et vélos sont sémantiquement plus proches du concept véhicule que voitures et bâtiments. Cette notion du rapport entre les concepts sémantiques est évaluée dans le treillis des concepts par les longueurs de chemin entre les concepts des graphes index et du graphe requête.

La valeur de spécificité mesure l'importance des thèmes de requête dans le document vidéo en calculant la somme des longueurs de chemin entre les concepts du graphe requête et les concepts du graphe d'index :

$$S(q,d) = \underset{m \in \prod(q,d)}{SUM} \left[\sum_{1 \le k \le K_m} Sim(c_{qmk}, c_{dmk}) \right]$$

La fonction *Sim* mesure la similarité entre deux concepts. On peur la calculer à partir de la longueur du chemin entre les nœuds concepts dans le treillis de concepts. Par exemple :

$$Sim(c, c') = 1/(1 + D(c, c'))$$

avec D(c,c') distance (nombre d'arcs) entre les concepts c et c' dans le treillis de concepts.

Discussion

L'utilisation du formalisme des graphes conceptuels offre un grand nombre d'avantages. C'est un formalisme de représentation de la connaissance qui est expressif et générique. Ce formalisme offre aussi des avantages majeurs telle que la structure de représentation qui est complètement ouverte et générique et aussi son adapté au contexte de la recherche d'information.

Nous avons présenté dans cette partie les éléments de base pour la définition de la fonction de correspondance. Nous avons détaillé le modèle de requêtes en spécifiant un certain nombre d'exemples des requêtes adaptées au modèle proposé.

Conclusion

Nous avons présenté les différents points liés à la définition d'un modèle générique pour la représentation des documents vidéo adapté à l'indexation et la recherche par le contenu. Pour pouvoir représenter le contenu sémantique de la vidéo, nous avons proposé un modèle basé des descriptions intégrant différentes caractéristiques de la vidéo notamment des informations génériques liées à la structure et d'autre liées à la sémantique. Nous avons choisi les graphes conceptuels comme formalisme de représentation. Ce choix est justifié par plusieurs raisons telles que : le formalisme des graphes conceptuels permet de représenter de manière homogène différentes composantes (documents, requêtes) d'un système de recherche d'information.

Notre contribution s'inspire de l'architecture proposée dans le modèle EMIR². Nous avons décrit le principe de base de ce modèle puis nous en avons proposé une extension pour le rendre valide dans le cas de la description du contenu audiovisuel. L'extension consiste à prévoir d'autres vues logiques ou bien physiques permettant d'associer une interprétation au contenu audiovisuel.

Nous avons proposé un modèle multifacette et multimodal intégrant des éléments d'informations issus de différents types de média. Nous avons choisi donc de structurer la proposition suivant la spécificité du modèle. Nous distinguons deux formes de représentations :

- ✓ *Une représentation spécifique* : elle permet de décrire le contenu vidéo par média. En effet, il existe plusieurs spécificités propres à chaque média (visuel, audio ou texte).
- ✓ *Une représentation générique* : elle regroupe l'ensemble des facettes décrivant les caractéristiques communes dans le document vidéo indépendamment du type de média, telles que, par exemple, la nature temporelle de la vidéo.

Les deux représentations spécifique et générique sont instanciées par le formalisme des graphes conceptuels.

Chapitre VI
Réalisation et Expérimentation

VI.1 Cadre applicatif - cas d'un journal télévisé

Excepté dans quelques cas d'études spécifiques (la vidéosurveillance par exemple), la description du contenu sémantique des documents vidéo est une tâche qui dépasse les capacités des systèmes actuels.

La description du contenu peut être envisagée en prévoyant par exemple une structure ou bien un modèle de représentation basé sur une terminologie bien définie et qui tient en compte de la spécificité du document.

Les journaux télévisés constituent un bon exemple de documents vidéo ayant une organisation physique bien définie. Cette structure est riche tant au niveau sémantique (diversité thématiques) qu'au niveau continuité spatio-temporelle (alternance plateau / reportage et changement de locuteurs). Dans la figure VI.1 nous présentons un exemple de cette structure. Cette structure permet à l'utilisateur une meilleure lisibilité et facilite le suivie et la compréhension du contenu du document. Elle est caractérisée par une alternance (plateau / reportage) et aussi le découpage en sujets.

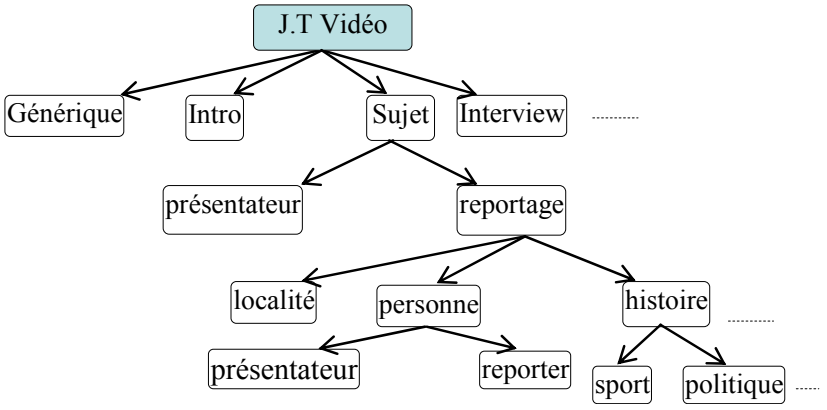


Figure VI. 1: Une structure typique d'un journal télévisé

La structure de ce type de document, bien qu'elle présente une information sur le contenu, ne permet pas de couvrir toutes les informations symboliques (concepts) contenues dans chaque segment vidéo.

Dans un contexte de recherche d'information, pour prendre en compte des besoins d'informations des utilisateurs, il est nécessaire d'exploiter à la fois la structure du document et les éléments d'informations symbolique contenues dans ce document (personnes, sujets, objets, des lieux, etc.). La structure du document paraît aussi importante puisqu'elle définit implicitement des descriptions. Cependant, les éléments d'informations symboliques permettent de répondre aux besoins des utilisateurs qui recherchent souvent des informations liées aux aspects interprétations du contenu et au contexte du document. L'extraction de ces

éléments à partir du contenu du document vidéo nécessite la mise en œuvre d'outils et des techniques appropriées.

VI.2 Corpus : la collection TRECVID

Depuis quelques années, les travaux d'indexation et de la recherche des données vidéo commencent à se structurer autour d'une collection unique dans le cadre des évaluations TRECVID qui réunissent une vingtaine d'équipes de recherche. Un travail commun d'annotation collaborative a été réalisé sur la collection TRECVID 2003. Ce travail a consisté à annoter à l'aide de l'outil Video-Annex [Lin 03] une partie du corpus. Il s'agit de décrire les plans vidéo par des concepts sélectionnés à partir d'une hiérarchie prédéfinie. Cette hiérarchie comporte 133 concepts regroupés en trois catégories : événements, scènes et objets. L'outil Video-Annex permet aussi de spécifier la position du concept dans une image-clé associée à un plan. Ce travail d'annotation a deux retombées majeures. Premièrement, nous disposons d'une grande collection de documents vidéo annotés en utilisant un même vocabulaire d'annotation. Deuxièmement, le lexique commun permet d'unifier les interprétations associées aux documents vidéo de la collection.

TRECVID définit plusieurs tâches telles que la détection de transitions entre les plans (SBD), la segmentation en histoire (SS), la détection de traits visuels (FD) et le processus recherche d'informations(Search). Les « topics » (requêtes) se composent d'une description textuelle courte du besoin de l'information et d'une ou plusieurs images fixes montrant un exemple de réponse pertinente.

Nos travaux ont été réalisés sur les collections TRECVID 2003 & 2004. Les collections TRECVID 2003 et 2004 sont composées de journaux télévisés en anglais. Ces vidéos sont principalement issues de chaînes de télévision Américaines ABC et CNN. Notons que chacune de ces collections contient plusieurs dizaines de milliers plans vidéo. Au total, la collection TRECVID 2003 contient plus de 120 heures de données vidéo. La collection TRECVID2004 contient 80 heures de vidéo en plus des vidéos de la collection TRECVID2003.

L'unité de traitement et d'évaluation est le plan. An niveau visuel, le découpage des vidéos en plan et l'extraction des images-clé ont été réalisés par le système CLIPS [Quénot 01], [Quénot 04].

Le processus d'extraction des images clé est intégré avec les processus de segmentation en plan vidéo. Il se base sur des paramètres calculés au cours du processus de détection de transition entre les plans. Ces paramètres sont reliés aux caractéristiques visuelles telles que la couleur ou les descripteurs de mouvement de la caméra. La sélection des images-clé dépend généralement du contexte d'application. En effet, pour une vidéo décrivant une réunion, il suffit d'une ou deux images-clés pour spécifier le contenu. Ceci est dû aux mouvements de caméra qui, dans ce genre de vidéo, sont limités. Par contre, dans le cas où le document vidéo est un journal télévisé, il est souvent question de multiples prises de vues et par conséquent, les processus d'extraction génèrent un nombre plus élevé de plans vidéo et d'images-clés.

Pour le contenu audio, nous disposons des segmentations en locuteurs et des transcriptions automatiques de la parole générées par le LIMSI.

Afin de d'évaluer l'ensemble des tâches qui seront réalisées durant TRECVID, le corpus est reparti en trois collections : une collection d'apprentissage, une collection de test et une collection de validation.

Notre travail consiste principalement en la proposition d'un modèle de représentation du contenu. Nous avons exploité l'ensemble des données fournies, notamment les transcriptions automatiques de la parole et les annotations du contenu visuel. Nous avons mis en place des outils pour l'extraction des descriptions symboliques du contenu. Au niveau audio, ces outils permettent d'extraire des concepts qui sont classés en plusieurs catégories (personnage, des lieux, etc..). Nous avons aussi mis en place un système pour la détection et la reconnaissance automatique d'identité des locuteurs, appliqué au contenu audio.

VI.3 Extraction des concepts

L'extraction des concepts consiste en une analyse linguistique du contenu de document et en l'identification d'éléments d'informations symboliques selon des listes prédéfinies (identité de personne, ville, cité, pays, organisation, etc.…). S'il est parfois plus facile d'établir un système d'extraction et de reconnaissance automatique de ces informations, il reste encore un grand nombre de cas ambigus et difficile à juger par le système. Par exemple, il est parfois difficile de distinguer dans le texte quand le mot « Washington » est utilisé comme nom de personne ou bien pour designer le nom d'un lieu géographique.

L'extraction et l'identification des « entités nommées » si on veut utiliser un terme plus approprié à cette tâche a fait l'objet de nombreux travaux de recherche [Stevenson 00], [McNamee 02]. Parmi des approches proposées, quelques travaux décrivent l'extraction des entités nommées basée sur des modèle statistiques [Robinson 99] employant HMM (Modèle de Markov caché). Stevenson et Gaizauskas [Stevenson 00] essayent de montrer l'apport de l'utilisation des listes statiques de noms propres comme corpus de reconnaissance dans leur système afin d'identifier les entités nommées. Ils recensent plusieurs méthodes pour filtrer ces listes. Dans le cadre de cette première phase de notre approche de catégorisation, nous nous sommes intéressés à l'extraction et à la reconnaissance des entités nommées. Nous préférons utiliser la notion du concept au lieu d'entité nommée. En effet, un concept est un mot (ou groupe de mots) qui est porteur d'une signification précise par rapport au contenu et qui permet de formuler une description sémantique dans le document. Notre approche est similaire à l'idée proposée par Stevenson et Gaizauskas. Nous avons développé un outil d'extraction et de reconnaissance qui combine des listes prédéfinies pour identifier les noms propres tel que personne, des lieux, des acronymes etc. Nous avons utilisé l'encyclopédie « Wikipedia » comme source d'information pour construire nos listes. Cette encyclopédie possède des dizaines de milliers d'articles sur les divers sujets (politique, sports, sciences, etc.) et dans diverses langues.

Note outil d'extraction de concepts se distingue par rapport à ce qui a été proposé dans la littérature par l'utilisation des patrons linguistiques qui permettent de vérifier l'exactitude des concepts qu'on a pu extraire ainsi que leur appartenance aux classes de concept.

Prenons le même exemple mentionné précédemment : le concept « Washington ». Pour distinguer pour distinguer que ce concept appartient à la catégorie de concept « personne » ou bien à la classe « place géographique », nous utilisons expressions clés citées juste avant où bien ce concept. En effet, si par exemple il est précédé par des termes de type (Mr., George, président, etc.) on peut alors déduire qu'il s'agit bien d'une identité de personne. Par contre, si

ce même concept « Washington » apparaît voisiné par des termes de genre (à, en direct de, etc.), dans ce cas il s'agit bien d'une place géographique.

Pour extraire les acronymes c'est tout simplement la l'aspect syntaxique nous permet d'inférer leur apparition dans le document. Par exemple le terme A.B.C. infère une abréviation d'une expression linguistique (*American Broadcasting Channel*). Dans La figure ci-dessous, nous présentons un exemple de transcription de la parole ainsi que l'affection des catégories aux concepts extraits.

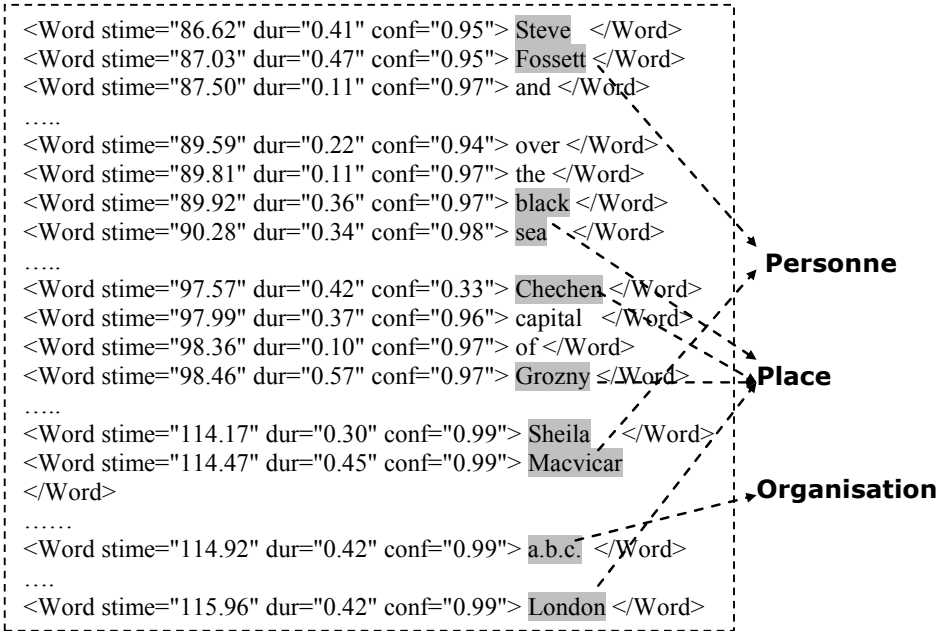


Figure VI. 2: Exemple d'extraction des concepts

Nous avons indiqué trois classes des concepts dans le processus d'extraction : personne, lieu, organisation (ou acronyme). Nous avons extrait de chaque document les concepts qui correspondent à chaque catégorie. Pour extraire ces concepts, nous projetons chaque document à une ontologie spécifique de domaine que nous avons conçu.

Nous détaillons dans ce qui suit le résumé de l'algorithme pour l'extraction automatique:

Pour un segment audio donné

Extraire les Aos par projection des ontologies spécifiques

Vérifier si les Aos sont dans la classe de concept personne

Vérifier si les Aos sont dans la classe de concept place

Vérifier si les Aos sont dans la classe de concept organisation

Si les Aos sont dans la classe de concept personne, spécifier les Aos correspondant aux locuteurs.

VI.4 Détection de l'identité du locuteur par patrons linguistique

Dans cette section nous allons détailler notre proposition d'un outil pour la détection et la reconnaissance d'identité du locuteur [Charhad 05a], [Charhad 05b].

La détection de l'identité du locuteur se fait à partir de la transcription de ce qui est dit et à partir d'une segmentation en locuteurs de la reconnaissance de la parole et/ou par un système de segmentation en locuteurs.

La figure VI.3 montre la structure d'un fichier intégrant les informations issues de la reconnaissance de la parole et de la segmentation en locuteurs. Il est à noter que les résultats de segmentation ne contiennent pas d'information sur l'identité des locuteurs. Aucun modèle audio de locuteur n'est utilisé mais seulement le résultat d'une détection des transitions d'un locuteur.

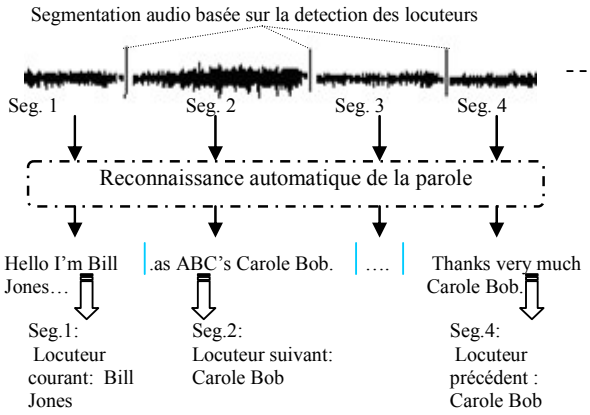


Figure VI.3: Segmentation de l'audio et détection d'identité du locuteur

En partant de cette structure, nous désirons identifier pour chaque segment le nom de locuteur. Nous exploitons pour cela les résultats de la transcription automatique de la parole fournis par le LIMSI[10]. Chaque transcription est segmentée suivant les locuteurs.

VI.5 Les patrons linguistiques

En plus de la structure physique (plateau ou reportage), un journal télévisé possède aussi une structure linguistique lorsqu'il y a un changement du locuteur. En effet, les journalistes utilisent souvent des expressions verbales très marquées dans leurs discours pour par exemple passer la parole d'un locuteur à un autre. Nous exploitons cette structure pour la reconnaissance automatique des identités des locuteurs. L'ensemble des expressions verbales sera considéré comme étant des patrons linguistiques.

L'avantage d'utiliser ce type des patrons c'est qu'ils permettent de distinguer facilement les identités des personnes citées dans le discours de celles des locuteurs. Pour le moment, nous appliquons l'approche de détection sur la transcription automatique de la parole contenue

[10] LISMI : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

dans le journal. Cependant, il reste un grand nombre de cas ambigus qui résultent des erreurs générées lors de la phase de reconnaissance automatique de la parole (ASR). Par exemple, s'il y a une erreur sur l'identité d'une personne ou bien sur un patron linguistique, nos résultats seront eux aussi erronés.

Nous proposons de classer les patrons en trois catégories : la première concerne les patrons pour la détection d'identité du locuteur du segment précédent, la deuxième catégorie contient les patrons qui permettent de détecter l'identité du locuteur dans le segment courant et enfin la troisième catégorie des patrons permet d'identifier l'identité du locuteur dans le segment suivant. Nous détaillerons ces trois catégories dans la section suivante.

VI.II.5.1 Affectation directe

Les patrons linguistiques sont appliqués à la transcription de la parole d'un segment audio issu d'une segmentation selon le changement du locuteur. Ils correspondent à des expressions régulières, en général paramétrables, censées préciser l'identité de la personne qui a parlé, qui parle ou qui va parler respectivement dans les segments précédents, courant et suivant. Ils sont appliqués à chaque segment et lorsqu'ils sont détectés, ils permettent de faire une prédiction de locuteur pour le segment précédent, le même segment ou le segment suivant. Une catégorie de patrons est définie pour chacun de ces trois cas :

- La première catégorie permet de détecter l'identité du locuteur qui est entrain de parler. Par exemple, le locuteur se présente : « …this is c.n.n news i'm [nom] …. » ou bien, lorsqu'il s'agit d'un reportage, généralement à la fin, la personne qui parle mentionne son identité.
- La deuxième catégorie contient les patrons pour détecter l'identité de locuteur qui vient de parler juste avant le locuteur actuel, par exemple « thank you very much [nom] … ».
- La troisième catégorie permet de détecter l'identité de la personne qui va parler (locuteur du segment suivant). Ce passage est souvent exploité lors des transitions plateau / reportage dans le journal. Par exemple à la fin du discours du présentateur on trouve une expression de type « …..[nom] has the latest … ».

À chaque catégorie correspond une liste de patrons linguistiques de détection. L'utilisation des patrons linguistiques permet de distinguer entre l'identité d'une personne mentionnée dans le discours de celle d'un locuteur. Le tableau ci-dessous décrit la liste des patrons de détection que nous avons utilisée dans notre approche. Ces patrons sont en partie spécifiques au corpus sur lequel nous travaillons (TREC vidéo 2003 et 2004, journaux télévisés de CNN et ABC).

Patrons segment précédent (SP)
Thank you very much (name) (name) thanks good morning (name) (name) reporting (name)from a.b.c. sports
Patrons segment suivant (SS)
good morning (name) tonight with (name) here's ABC's / CNN'S (name) with ABC's / CNN'S (name) as ABC's / CNN'S (name) ABC's / CNN's (expr.) correspondent (name) (name) reports With us (name) up to a.b.c.'s (name) we asked a.b.c.'s (name) In (place) (name) From (place) (name) back to (name)
Patterns segment courant (SC)
(name) for ABC news I'm (name) (name) CNN news (place) (name) ABC news (place)

Tableau VI.1: Liste de catégories de patrons

Le patron linguistique « good morning » permet d'indiquer, selon sa position dans le segment, la présence d'une identité du locuteur du segment précédent ou bien d'un locuteur du segment suivant. La position de ce patron dans le segment est donc capitale. En effet, si ce terme apparaît à la fin du segment, il s'agit alors d'un patron linguistique pour détecter un locuteur dans le segment suivant. Par contre, dans le cas où ce patron linguistique apparaîtrait au début du segment, il s'agit alors d'un patron pour détecter un locuteur dans le segment précédent.

Les patrons linguistiques sont tous paramétrés. Ils ont tous au moins un paramètre correspondant à un nom, un prénom ou un couple prénom+nom. Ils peuvent aussi contenir des paramètres correspondant à des lieux géographiques, à des organisations (CNN ou ABC par exemple) ou à des formules comme « good morning », « good evening », « thanks ».

Les patrons linguistiques sont généralement positionnés à la fin du segment audio pour marquer un passage de tour de parole. Par exemple, le patron suivant infère un passage de la parole entre le présentateur du journal et un reporter : « …a.b.c's Sheila Macvicar has the latest ».

Pour la détection et la reconnaissance de l'identité de personne dans chaque segment audio, notre approche se base sur une liste de mots contenant environ 12400 prénoms classés en prénom femme / prénom homme. Par contre, pour la reconnaissance de nom de famille, nous

exploitons une liste de noms communs pour filtrer les termes correspondants à des noms de famille, généralement. Le nom apparaît toujours après les prénoms.

VI.II.5.2 Affectation par propagation

Dans le cas d'un journal télévisé, le locuteur se présente généralement une seule fois lors de sa première intervention. Pour cela, exploiter uniquement des patrons linguistiques pour détecter l'identité des locuteurs ne permet de détecter les locuteurs sur l'ensemble des segments audio. Pour cela, nous avons spécifié en complément de l'affectation directe, un cas d'étude qui consiste à propager les résultats obtenus par application des patrons sur le reste des segments.

Nous exploitons pour cela des informations générées dans l'étape de segmentation et détection de changement du locuteur. Ces informations sur les locuteurs qui sont de type spkr # (où spkr désigne le locuteur et le symbole '#' indique l'indice de locuteur). Rappelons que lors d'un processus de transcription automatique de la parole, quelques informations sont automatiquement générées par le système. Parmi ces informations, on trouve par exemple la référence du locuteur (homme ou femme) et aussi un indice qui va nous permettre d'identifier les locuteurs qui apparaissent plus qu'une seule fois dans le document. La balise « en gris » dans la figure ci-dessous, montre un exemple d'entête de segment audio transcrit. Dans cette balise, l'information *« FS3 »* indique que le locuteur est une femme (*FS : female speech*). Le chiffre trois indique l'indice de locuteur.

```
<SpeechSegment spkr="FS3" stime=".." etime="..">
…
<Word stime=".." dur=".." conf=".."> in </Word>
<Word stime=".." dur=".." conf=".."> the </Word>
<Word stime=".." dur=".." conf=".."> fog </Word>
…

</SpeechSegment>
```

Figure VI.4: Structure du fichier transcription

VI.II.5.3 Évaluations

Les évaluations ont été effectuées sur une partie de la collection TREC vidéo 2003 (quatre journaux télévisés : deux de CNN et deux d'ABC d'une demi-heure environ chacun). Le tableau 2 résume les caractéristiques du corpus effectivement utilisé.

Durée totale	7009.0 s
Durée totale de parole	5249.1 s
Durée totale des journaux	4250.3 s
Parole dans les journaux	3767.1 s
Locuteurs annotés dans les journaux	3677.5 s

Tableau VI.2 - Caractéristiques du corpus

Le tableau ci-dessous (tableau 3) montre les résultats obtenus pour l'indexation de l'identité du locuteur par patrons linguistiques. Les types de patrons « locuteur précédent », « locuteur courant » et « locuteur suivant » permettent de faire une prédiction dans 1.0 %, 6.8 % et 7.0 % des cas (en durée de parole) respectivement, soit une prédiction « directe » dans 14.8 % des cas. La propagation de ces prédictions aux autres segments attribués à un même locuteur au niveau acoustique permet de faire une prédiction dans 52.7 % des cas. Les précisions obtenues sont respectivement de 100 %, 90.2 %, 74.1 %, 83.3 % et 82.4 %. Il est à noter que la propagation a introduit assez peu d'erreurs tout en augmentant de manière très significative la proportion de prédiction. Les erreurs sont dues à une mauvaise détermination de la position des extrémités des segments de parole, à une mauvaise association de segments à un même locuteur ou à une erreur dans la reconnaissance du nom d'une personne. Trois confusions ont été faites par le système de reconnaissance : « Dean Reynolds » pour « Tim Reynolds », « Jim Wooten » pour « Jim Wutton » et « Shimbun Darrow » pour « Siobhan Darrow ». Pour mesurer les effets respectifs des différentes sources d'erreurs, nous avons refait l'évaluation en corrigeant manuellement ces erreurs sur les noms (toutes les autres erreurs de transcriptions étant laissées telles quelles). La précision avant propagation est très nettement améliorée mais la propagation introduit alors plus d'erreurs. Diverses sources d'informations peuvent permettre de filtrer ou de corriger ces erreurs sur les noms (par exemple, les noms proposés ne correspondent pas à ceux de personnes connues).

Prédiction	**Durée prédite**		**Durée correcte**	
Précédent	37.2	1.0 %	37.2	100 %
Courant	250.3	6.8 %	225.9	90.2 %
Suivant	258.2	7.0 %	191.4	74.1 %
Directe	545.8	**14.8** %	454.6	**83.3** %
Propagation	1936.8	**52.7** %	1595.9	**82.4** %

Tableau VI.3 - Résultats pour l'indexation de l'identité des locuteurs. Le pourcentage de durée prédite est relatif à la durée totale ayant pu être annotée manuellement. Le pourcentage de la durée correcte est relatif à la durée prédite.

VI.II.5.4 Interface du système

La figure VI.5 montre l'interface du prototype CLOVIS. La partie « audio layer » permet la détection et la reconnaissance d'identité du locuteur dans un document vidéo. Ce système est basé sur l'analyse de la transcription automatique de la parole L'objectif principal est de déterminer et d'identifier pour chaque segment vidéo, l'identité de la personne qui parle afin de l'exploiter dans le modèle pour la recherche par le contenu sémantique de la vidéo.
Comme nous l'avons mentionné dans la partie modélisation, l'intérêt d'extraire ce type d'information nous facilite la description du contenu audio au niveau conceptuel. En effet le concept locuteur infère deux relations conceptuelles pour la modélisation : les relations « parle » et « parle de ». Notre prototype permet donc de répondre à des requêtes de type : « rechercher les segments vidéo dans lesquels on parle d'un Concept X ». Le concept X peut être soit dans l'une des trois catégories « personne », « lieu » ou « organisation », soit un autre concept extrait de l'audio. Un deuxième type de requête qui consiste à chercher les segments vidéo dans lesquels un Concept X parle. Dans ce dernier cas, le concept X fait partie de la catégorie « personne » (le locuteur tel que le présentateur ou le reporter dans le journal télévisé).
Notons que dans cette interface contient aussi une partie pour la recherche d'information en spécifiant des requêtes au contenu visuel tel que par exemple : « rechercher des segments vidéo montrant une personne à gauche d'un drapeau ». Cette partie n'est pas encore implémentée.

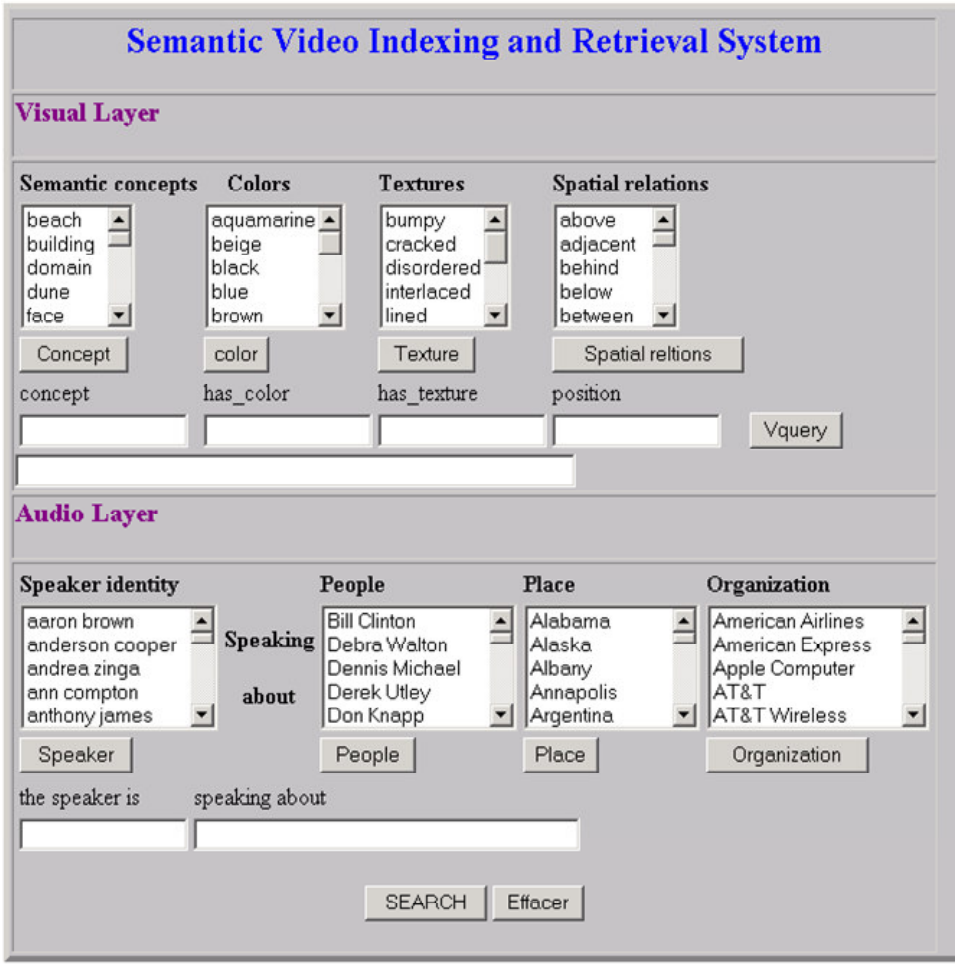


Figure VI.5: Interface –prototype CLOVIS

Pour l'exemple de la requête1 : « *rechercher les segments vidéo dans lesquels Peter jenning parle* », nous présentons ci-dessous les premiers résultats de notre système.

Figure VI.6: Extrait –les premières réponses du système pour la requête1

Discussion

Dans le cas de documents vidéo, la combinaison des éléments d'informations issus de différents sous médias est encore peu abordée dans le contexte de l'indexation et la recherche vidéo par le contenu sémantique. Ceci résulte principalement des difficultés d'intégrer les techniques spécifiques à chaque sous média.

Les expérimentations que nous avons proposées dans ce chapitre sont, en grande partie, appliquées sur la transcription de la parole qui constitue une source d'information symbolique. Notre travail consiste essentiellement en des applications pour l'extraction des concepts. Parmi ces concepts, on trouve le concept « personne ». Une de nos applications concerne la détection et la reconnaissance d'identité de locuteur dans le document vidéo. Ce travail a été poursuivi par d'autres démarches afin d'extraire plus d'information sur le locuteur tel que par exemple détecter le lieu (nom de ville, nom de pays) ou bien détecter le thème (de quoi il parle), etc. Faute de manque de vérité terrain sur ces tâches nous avons évalué cette application que sur une partie de la collection TRECVID.

VI.6 Application à la recherche des « Topics » sur TRECVID 2004

Dans le cadre de l'évaluation sur le corpus TRECVID, la tâche de recherche s'effectue sur des « topic ». Un « topic » est défini comme une description formatée d'un besoin d'information mettant en œuvre de multiples caractéristiques (visuelle, audio, textuelle). La

complexité inhérente à la recherche d'un « topic » est intrinsèquement liée à la difficulté d'interprétation et de mise en œuvre de requêtes soulignant les liens entre les différentes caractérisations exprimées.

Dans le cadre de l'évaluation, nous nous plaçons dans le contexte de la recherche manuelle pour laquelle un utilisateur humain avec une connaissance du système manipulé et de plus familier de son interface est capable d'interpréter les « topics » considérés et de proposer des requêtes. 24 « topics » multimédia proposés par le NIST pour la tâche de recherche expriment un besoin d'information concernant des documents vidéo représentant des personnes, objets, événements, actions… ou encore une combinaison de ces éléments (l'ensemble de ces requêtes sont présentées dans la partie « Annexe » accompagnés de leurs traductions dans le formalisme CLOVIS). Ces « topics » sont mis en œuvre de manière à refléter un ensemble non restreint de requêtes diverses proposées par des utilisateurs réels : ces requêtes sont axées sur la recherche de personnes spécifiques ou de types de personnes, d'objets spécifiques ou bien d'instances de types d'objets, d'activités, de lieux ou bien d'instances de types de lieu ou d'activité. Notons qu'à l'heure actuelle, l'ensemble des résultats que nous avons obtenus n'est pas ordonné selon le modèle de correspondance que nous avons défini dans la section V.IV.3. Ces résultats sont tout simplement triés par ordre chronologique.

Nous comparons les résultats de prototype CLOVIS (appliqué sur les transcriptions) avec les systèmes proposés dans le cadre de l'évaluation TRECVID 2004 opérant une recherche manuelle sur des transcriptions automatiques de la parole.

Le système de NTU (**National Taiwan University)** est basé sur les transcriptions de la parole et les descriptions conceptuelles en utilisant l'ontologie « WordNet » pour le calcul des distances entre les termes. L'approche proposée dans ce système aligne les transcriptions avec la description haut-niveau des traits visuels. Sans utilisation des algorithmes de calcul complexe, cette méthode a prouvé sa performance pour la recherche d'information vidéo.

Le système développé par l'équipe IU (**Indiana University system)** connu sous le nom « viewfinder system », utilise uniquement du texte. Ce système est aussi basé sur les transcriptions de la parole dans la vidéo. Chaque requête est formulée manuellement et est constituée d'une suite de mots supportée par la pondération « tf.idf ». Les requêtes proposées dans le cadre de ce système sont mises en oeuvre par la construction manuelle et la sélection d'exemples visuels.

Dans le cadre des expérimentations TRECVID, nous pouvons remarquer que même si les « topics » sont riches (texte, image, vidéo), au niveau recherche, ce « topic » est généralement traduit sous forme d'un mot ou d'un groupe de mots. Ceci est dû aux processus d'extraction et de caractérisation du contenu audiovisuel de la vidéo.

Afin d'évaluer (en partie) notre proposition, nous avons choisi quelques « topics » du TRECVID pour lesquels il nous semble que l'utilisation de la transcription automatique de la parole suffit pour trouver des réponses pertinentes. Nous présentons dans le tableau VI.4 quelques exemples de ces « topics » avec leurs traductions.

Topic TRECVID 2004	Transcription ClOVIS (sous forme des graphes)
128.US Congressman Henry Hyde's face, whole or part, from any angle	[Henry Hyde] → (parle) ou [Personne] → (parle de) → [Henry Hyde]
133. Saddam Hussein	[Saddam Hussein] → (parle) ou [Personne] → (parle de) → [Saddam Hussein]
134. Boris Yeltsin	[Boris Yeltsin] → (parle) ou [Personne] → (parle de) → [Boris Yeltsin]
135. Sam Donaldson's face. No other people visible with him	[Sam Donaldson] → (parle) ou [Personne] → (parle de) → [Sam Donaldson]
136.Person hitting a golf ball	[Personne] → (parle de) → [P.G.A.]
137. Benjamin Netanyahu	[Benjamin Netanyahu] → (parle) ou [Personne] → (parle de) → [Benjamin Netanyahu]

Tableau VI.4 : Exemple de « topics » TRECVID 2004 et transcription CLOVIS

En plus de la traduction linéaire telle que présentée dans le tableau VI.3, chaque requête est traduite aussi sous forme de représentation graphique suivant la signification du « topic ». Si on prend par exemple le topic 133 (« find shot of Saddam Hussein »), en se basant sur le contenu audio cleui-ci sera décrit de la manière suivante :
[Saddam Hussein] → (parle) « Les plans où Saddam Hussein est le locuteur »
[Personne] → (parle de) → [Saddam Hussein] « Les plans où on parle de Saddam Hussein »

L'application de recherche exploite les résultats de l'extraction automatiques de catégories de concepts (personne, lieu, organisation, etc.) que nous avons détaillée dans les sections précédentes.

Les figures VI.7 et VI.8 montrent respectivement les premiers plans retournés le prototype CLOVIS en réponse aux « topics » (133, 135, 136) et la courbe de rappel précision comparons nos résultats à celle de « IU » (`M_C_1_IuVf1_1`) et de « NTU » (`M_C_2_NTU_NLP_Lab1_2`) pour les même « topics ».

Figure VI.7: les premiers plans retournés par CLOVIS pour les topics 133, 135, 136

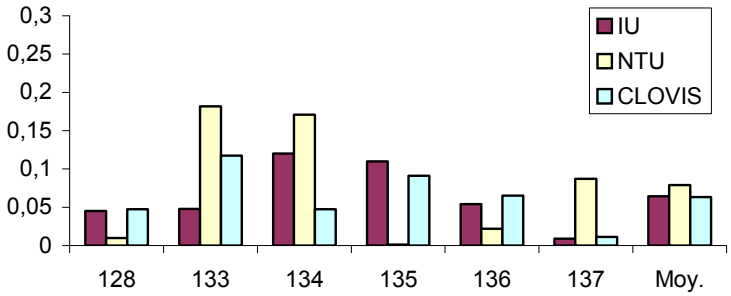


Figure VI. 8 : Courbe de rappel / précision des résultats de CLOVIS sur les « topics » 128, 133, 134, 135,136, 137

En conclusion sur la partie expérimentation, nous pouvons constater que nous sommes encore loin d'une évaluation complète de notre proposition. Ceci résulte principalement de l'absence des vérités terrain pour la recherche par le contenu intégrant la notion de concepts et des relations conceptuelles.

Conclusion et Perspectives

Chapitre VII
Conclusion

Le travail présenté dans ce rapport s'inscrit dans le contexte de la recherche d'information. Il s'agit particulièrement de proposer un modèle et de représentation afin d'améliorer l'indexation et la recherche des documents vidéo par le contenu sémantique.

Nous rappelons nos principaux objectifs que nous avons proposés au début de ce rapport. Il s'agit de :

- ✓ Définir un modèle pour la description du contenu des documents vidéo. Ce modèle doit prendre en compte les éléments d'information à différents niveaux de description (signal et sémantique) et les différents médias présents. Il doit aussi être adapté pour être intégré dans un système d'indexation et de recherche de documents vidéo.
- ✓ Définir un modèle conceptuel générique qui mette en avant les aspects représentation conceptuelle et interprétation sémantique. Pour atteindre cet objectif et dans la perspective d'avoir un modèle d'indexation et de recherche par le contenu vidéo efficace et précis, nous avons choisi de mettre en œuvre des représentations par des concepts et des relations.

Pour mieux atteindre ces objectifs, nous avons tout d'abord détermine l'ensemble des besoins par l'analyse des travaux existants dans le même contexte. Suite à la synthèse de ces travaux, nous avons choisi de nous intéresser prioritairement aux aspects modélisation et représentation du contenu sémantique. Cette modélisation nous a permis de prévoir une nouvelle forme d'indexation du contenu vidéo. Cette forme met en avant l'indexation par concepts et relations conceptuelles. Ceci rend plus facile l'interrogation d'une collection de documents vidéo en se basant sur un langage naturel et sur des descriptions proches de celles qu'un utilisateur exploite pour interpréter le contenu de la vidéo.

Apport théorique

Au niveau formel, notre travail a consisté à définir un modèle pour la représentation par le contenu des documents vidéo intégrant la notion du concept et de relation conceptuelle. Ce modèle est générique du fait qu'il correspond une représentation totalement indépendante de la structure de la vidéo. Dans ce modèle, nous avons distingué deux représentations (générique et spécifique). L'objectif était de tenir compte de la spécificité des descriptions liées à l'aspect multimodal du document.

Le modèle permet de spécifier des descriptions génériques et indépendamment de type de média et du genre de la vidéo (journal télévisé, un documentaire, etc.). Dans ce contexte, nous avons proposé une extension du modèle EMIR² conçu pour la représentation et recherche symbolique des images fixes. L'aspect générique de ce modèle se tient d'une part dans sa capacité d'intégrer une variété des représentations du contenu de l'image et d'autre part dans le procédé opérationnel utilisé pour exploiter ces différentes représentations. En effet, le formalisme des graphes conceptuels dispose d'une grande expressivité et d'une capacité à représenter l'information d'une manière à la fois simple à interpréter par les utilisateurs et adaptée aux processus d'indexation et de recherche d'information.

L'extension du modèle EMIR² consiste à garder les aspects de représentation multifacettes du contenu tels que définis pour le cas des images fixes. La contribution consiste à ajouter des représentations (facettes) spécifiques aux documents vidéo. Nous avons proposé l'ajout de deux facettes suivantes :

- ✓ une facette temporelle pour la description de la nature temporelle dans la vidéo ;
- ✓ une facette événementielle permettant de décrire les événements dans le document vidéo.

Ces deux facettes s'ajoutent à l'ensemble des descriptions définies dans le modèle EMIR² pour former une description plus générique du contenu vidéo.

Le schéma de représentation spécifique est la deuxième partie de notre proposition. Il consiste à mettre en place une représentation spécifique permettant de prendre en compte l'hétérogénéité du contenu vidéo. La modélisation est multifacettes par type de média. Notre proposition est concentrée sur la description du contenu visuel et audio.

En résumé, notre contribution au niveau formel consiste à la proposition d'un modèle et la représentation et la structuration de document vidéo. Cette proposition est instanciée par un modèle opérationnel basé sur une extension formalisme des graphes conceptuels, permettant de définir une représentation du contenu vidéo et une fonction de correspondance pour les comparer. Le formalisme des graphes conceptuels présente de nombreux avantages dans notre contexte d'application. Il permet de représenter tous les composants d'un système de recherche vidéo : requêtes, documents et fonction de correspondance par un même formalisme.

Prototype

Comme nous l'avons mentionné précédemment, Nos expérimentations ont été réalisées sur les collections TRECVID (2003 et 2004). Nous avons développé et testé des outils sur des parties du modèle Notamment sur les résultats des transcriptions automatiques de la parole.

Extraction des concepts : nous avons proposé un outil pour l'extraction des concepts. Cet outil est orienté vers une utilisation dans un contexte spécifique où la liste des concepts extraits automatiquement figure parmi une liste d'entités nommées (nom personne, nom place géographique, nom d'organisation.

Détection et reconnaissance d'identité de locuteur : c'est un outil basé sur l'analyse de transcription automatique pour la détection et la reconnaissance d'identité de locuteur dans un document vidéo. L'objectif principal est de déterminer et d'identifier pour chaque segment vidéo l'identité de la personne qui parle afin de l'exploiter dans le modèle pour la recherche par le contenu sémantique de la vidéo.

Nous avons exploité les résultats de ces outils pour pouvoir modéliser le contenu vidéo. Nous avons donc mis en place une technique qui, à partir des informations sémantiques, permet de générer une description relationnelle (des concepts interconnecté par des relations conceptions). Les relations conceptuelles sont inférées selon le type média. Par exemple, les relations « parle » et « parle de » sont extraites de la transcription et de la segmentation en locuteurs du contenu audio.

Perspectives

Si le processus d'indexation et de recherche d'information par le contenu sémantique demeure un vaste domaine sans solution idéale, les outils et les méthodes que venons de présenter auront participé à un certain avancement vers cette solution. Cependant, nos résultats s'inscrivent dans des perspectives de poursuite de la recherche dans ce domaine.

Modèle

Les perspectives reliées au modèle peuvent être classées en deux catégories selon qu'elles concernent la modélisation au niveau générique ou au niveau spécifique :

En ce qui concerne la modélisation au niveau *spécifiques*, des améliorations peuvent être apportées notamment sur l'intégration des facettes de description issues de différents types de média (image, audio, texte). Cette intégration permettra en effet de d'unifier le schéma de description. Un deuxième point important dans la modélisation au niveau spécifique consiste à la mise en œuvre des propositions formelles pour la modélisation du contenu visuel. Nous n'avons pas encore implémenté les propositions associées au contenu visuel notamment les sous-facettes du contenu signal. Il serait intéressant à court terme des les mettre en œuvre et d'associer ces résultats pour compléter l'architecture du prototype CLOVIS. Rappelons que pour le moment c'est uniquement des résultats, issus de la modélisation du contenu audio, sont intégrés dans ce prototype. Notons que la modélisation du contenu du contenu visuel consiste à exploiter les annotations (description symbolique) pour pouvoir décrire le contenu d'un plan, d'une scène ou d'une séquence vidéo ou bien une description multifacette intégrant les caractéristiques visuelles (couleur, texture, etc.) de l'image clé (description signal). Il s'agit d'une description du contenu de l'image clé de chaque plan vidéo.

Devant un contenu sémantique très riche et variés (plusieurs thèmes) tel que le cas d'un journal télévisé, la description d'information du contenu symbolique notamment pour l'information visuelle nécessite la mise en place des outils capables d'extraire les critères objectifs (concepts visuels) qui permettent d'interpréter le contenu en termes d'événements, d'actions et de concepts visuels, ce qui rendra plus facile la modélisation du document. À l'heure actuelle, l'approche d'annotation est souvent sollicitée pour ce genre des tâches.

En ce qui concerne la modélisation au niveau *générique*, il nous paraît important d'étudier l'extension possible du modèle que nous avons proposé. Cette extension consiste à développer des bases de connaissances (ontologies) afin d'enrichir le modèle proposé.

Prototype

Au niveau expérimental, pour le contenu audio, il est nécessaire d'améliorer les techniques mises en œuvre notamment l'approche d'identification et de reconnaissance d'identité de locuteur. Nous prévoyons aussi d'une part d'intégrer plus de caractéristiques audiovisuelles (visuel et texte dans la vidéo) pour la reconnaissance des locuteurs et d'autre part.

Pour contenu visuel, nous envisageons d'implémenter les propositions dans le modèle et d'intégrer les résultats obtenus dans le prototype.

Bibliographie

Bibliographie

[Ahanger 95] Ahanger G., Benson D., Thomas D. Little C.: "Video Query Formulation", Storage and Retrieval for Image and Video Databases (SPIE) San Diego/La Jolla, CA, USA, 208-291, 1995.

[Akutsu 98] Akutsu M., Hamada A., and Tonomura Y.: “Video handling with music and speech detection,” IEEE Multimedia, vol. 5, no. 3, pp. 17–25, 1998.

[Amato 98] Amato G., Mainetto G., Savino P.: “An Approach to a Content Based Retrieval of Multimedia Data”, Multimedia Tools and Applications, 1998.

[Ardizzone 99] Ardizzone E., Hacid M-H.: "A Semantic Modeling Approach for Video Retrieval by Content", In IEEE International Conference on Multimedia Computing and Systems, (ICMCS'99), 158-162, Florence, Italy, 1999.

[Arslan 02] Arslan U., Donderler M. E., Saykol E., Ulusoy O., Gudukbay U.: "A Semi-Automatic Semantic Annotation Tool for Video Databases." SOFSEM 2002, Workshop on Multimedia Semantics Milovy, Czech Republic: 2002.

[Babaguchi 99] Babaguchi N., Kawai Y., and Kitahash T.: "Event Based Video Indexing by Intermodal Collaboration" Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99) Orlando, USA, 1-9, 1999.

[Bach 96] Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., Shu, C.-F. (1996). “The Virage Image Search Engine: An Open Framework for Image Management”. In: Storage and Retrieval for Still Image and Video Databases IV, 1996.

[Belkhatir 05] Belkhatir M., «Intégration signal/symbole pour l'indexation et la recherche d'images fixes», Thèse Informatique de l’université Joseph Fourier, Grenoble I, septembre 2005.

[Berlin 69] Berlin, B., and Kay, P. Basic color terms: their universality and evolution. Berkeley, CA: University of California Press, 1969.

[Bhushan 97] Nalini Bhushan, A. Ravishankar Rao, Gerald L. Lohse: The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. Cognitive Science 21(2): 219-246 (1997)

[Boudry 02] Boudry, C., En savoir plus sur les images numériques, 2002, http://web.ccr.jussieu.fr/urfist/image_numerique/Image_numerique1.htm

[Bray 98] Bray T., Paoli J, Sperberg M. C., "Extensible Markup Language (XML)" Recommandation du W3C, 10 février 1998.

[Celentano 02] Celentano A., Gaggi O. "Schema modelling for automatic generation of multimedia presentations", 14th international conference on Software engineering and knowledge engineering Ischia, Italy, 593-600, 2002.

[Chang 97] Chang S-F., Chen W., Meng H-J., Sundaram H., Zhong D.: "VideoQ: An Automated Content Based Video Search System Using Visual Cues", 5th ACM conference on Multimedia Los Angeles USA, 313-324, 1997.

[Charhad 04a] Charhad M., Un modèle d'indexation et de recherche de documents vidéos basé sur le formalisme des graphes conceptuels, 22ème Congrès INFORSID'04 (Informatique des Organisations et Systèmes d'Information et de Décision), Biarritz, France, 25-28 Mai, 2004.

[Charhad 04b] Charhad M. and Quénot G., Semantic Video Content Indexing and Retrieval using Conceptual Graphs, in ICTTA, Damascus, Syria, pp19-23, 19-23 Avril, 2004.

[Charhad 05a] Charhad M. and Moraru D. and Ayache S. and Quénot G.: “Speaker Identity Indexing in Audio-Visual Documents”, in Content-Based Multimedia Indexing (CBMI2005), Riga, Latvia, pp: actes sur CD, June 21-23, 2005.

[Charhad 05b] Charhad G., Quénot G. : “Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés”, CORESA’05, 7 - 8 Nov. , to appear, 2005.

[Charhad 05c] Charhad M., Zrigui M., Quénot G., Une Approche Conceptuelle pour la Modélisation et la Structuration Sémantique des Documents Vidéos, in In IEEE Int. Conf. on Sciences of Electronic, Technologies of Information and Telecommunications, (SETIT), Sousse, Tunisia, pp: actes sur CD, 27-31 mars, 2005.

[Chein 92] Chein M., Mugnier M-L.: “Conceptual Graphs: fundamental notions”, Revue d'intelligence artificielle, 1992.

[Chua 99] Chua T-S., Chen L. and Kankanhalli M.: "Stratification Approach to Modelling video", Proceedings of Multi-Media Modeling (MMM’99). Ottawa, Canada, pp 179-192, 1999.

[Corridoni 96] Corridoni J. M., A. Del Bimbo, D. Lucarella, and H. Wenxue: “Multi-perspective navigation of movies”, *Journal of Visual Languages and Computing*, 7:445-466, 1996.

[Dechilly 00] Dechilly T., Bachimont B, "Une ontologie pour éditer des schémas de description audiovisuels, extension pour l'inférence sur les descriptions " Journées francophones d'Ingénierie des Connaissances, Toulouse, mai 2000.

[Decleir 98] Decleir C, Hacid M-S, Kouloumdjian J.: "Modeling and Querying Video Data: A Hybrid Approach." IEEE Workshop of Content-Based Access of Image and Video Databases in conjunction with CVPR Santa Barbara, California: 1998.

[Decleir 99] Decleir C., Hacid M-S., Kouloumdjian J.: "A Database Approach for Modelling and Querying Video data." 15th International Conference on Data Engineering Sydney, Australia: 1999.

[Declerck 01] Declerck Thierry and Peter Wittenburg: "MUMIS - A Multimedia Indexing and Searching Environment", Proceedings of the First International Workshop on Multimedia Annotation, MMA'01 Tokyo, Japan: 2001.

[Derrode 99] Derrode S., Mezhoud R. and Ghorbel F. : "Reconnaissance de formes par invariants complets et convergents - Application à l'indexation de bases d'objets à niveaux de gris", 17ème colloque GRETSI'99 Vannes (France), 1999.

[Dumas 00] M. DumasM., R. Lozano, M.C. Fauvet, H. Martin, P.C. Scholl. "Orthogonally modeling video structuration and annotation: exploting the concept of granularity" AAAI-2000 Workshop on Spatial and Temporal Granularity Ausrin: 2000.

[Etievent 99] Etievent E., Frank Lebourgeois, Jean-Michel Jolion. "Assisted Video Sequences Indexing: Motion Analysis Based on Interest Points", Iciap99, 1999.

[Fablet 00] Fablet R., Bouthemy P. "Statistical motion-based video indexing and retrieval", Conf. on Content-Based Multimedia Information Acces, Paris, France: 602-619, 2000.

[Fatemi 01] Fatemi N. and O. Abou Khaled, COALA: Content Oriented Audiovisual Library Access, Sent for publication, 8th International Conference on Multimedia Modeling (MMM'2001), Amsterdam, Netherlands, November, 2001.

[Faure 98] Faure D. and Nédellec C. and Rouveirol C.: "Acquisition de connaissances sémantiques par des méthodes d'apprentissage: le système ASIUM." 13èmes Journées Francophones sur l'Apprentissage Arras France, 126-137, 1998.

[Gandon 02] Gandon, F.: "Ontology Engineering: A Survey and a Return on Experience", Rapport de Recherche INRIA, 2002.

[Garofolo 99] Garofolo J. S., Voorhees E. M., Anzanne C., and Stanford V. M.: "Spoken document retrieval: 1998 evaluation and investigation of new metrics." In Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio Cambridge, UK, 1-7, 1999.

[Gauvain 02] Gauvain J.L., Lamel L., and Adda G.: "The LIMSI Broadcast News transcription system", in Speech Communication 37, pp. 89-108, 2002.

[Giuseppe 98] Giuseppe A., Mainetto G., Pasquale S.: "An Approach to a Content-Based Retrieval of Multimedia Data", Multimedia Tools Appl 7 (1998): 9-36.

[Guarino 97] Guarino, N. "Understanding, building, and using ontologies: A commentary to Using Explicit Ontologies" in KBS Development, by van Heijst,

Schreiber, and Wielinga". International Journal of Human and Computer Studies 46: 293-310, 1997.

[Gunsel 96] Gunsel B., Ferman A. M., Tekalp M.: "Video Indexing Through Integration of Syntactic and Semantic Features", 3rd IEEE Workshop on Applications of Computer Vision (WACV '96) , 90-95, 1996.

[Hammond 98] Hammond R., Chen L. and Fontaine D.: "An Extensible Spatial-Temporal Model for Semantic Video Segmentation." First International Forum on Multimedia and Image Processing (IFMCP'98) Anchorage, Alaska: 1998.

[Hampapur 95] Hampapur, A., Jain, R., and Weymouth, T., "Digital Video Segmentation", Proc. ACM Multimedia 94, San Francisco, CA, pp. 357-364, 1994.

[Hampapur 99] Hampapur A. "Semantic Video Indexing: Approach and Issue." SIGMOD Record 28: 32-39, 1999.

[Harb 02] Harb H., Chen L.: "Video Scene Description: An Audio Based Approach" HERMES Science Publications:Proceedings of the first Medianet Conference MEDIANET2002 Souss,Tunisia: 2002.

[Hauptmann 97] Hauptmann A. Wactlar H. "Indexing and Search of Multimodal Information." Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-97) Munich, Germany: 1997.

[Hjelsvold 94] Hjelsvold R, Midtstraum R.: "Modelling and Querying Video Data" proceedings of the 20 th VLDB Conference Santiago, chile, 1994.

[Hollfelder 00] Hollfelder S., Everts A. and Thiel U.: "Designing for Semantic Access: A Video Browsing System." Multimedia Tools and Applications 11: 281-293, 2000.

[Holt 97] Holt, B, Weiss, K, Niblack, W, Flickner, M., Petkovic, D.: "The QBIC Project in the Department of Art and Art History at UC Davis". In: Proceedings of the ASIS Annual Meeting. 34. pp. 189-195, 1997.

[Huang 00] Huang T. S. and Naphade M., "MARS (Multimedia Analysis and Retrieval System): A test-bed for video indexing, browsing, searching, filtering and summarization", International Workshop on Multimedia Data Storage, Retrieval, Integration and Application Hong Kong, 2000.

[Hunter 99] Hunter J., Armstrong L., "A Comparison of Schemas for Video Metadata Representation ", WWW8, Toronto, May, 1999.

[Ive 04] John G S Ive: "The Material eXchange Format and the Workflow Revolution", In Professional-MPEG Forum, 2004.

[Iyengar 02] Iyengar G., Nock H., Neti C., Franz M.: "Semantic Indexing of Multimedia using Audio, Text and Visual Cues", Proceedings of ICME2002 Switzerland, 369-372., 2002.

[Jiang 97] Jiang H., Montesi D., Elmagarmid H.: "VideoText database systems", In ICMC and Systems, pp. 334–351, 1997.

[Kemp 00] Kemp T., Schmidt M., Westphal M., Waibel A.: “Strategies for Automatic Segmentation of Audio Data” Proc. of ICASSP-2000, pp. 1423-1426, 2000.

[Kipp 01] Michael Kipp:"Anvil - A Generic Annotation Tool for Multimodal Dialogue", In Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370, Aalborg, September 2001

[Kobla 00] Kobla V., DeMenthon D., and Doermann D., “Identifying sports video using replay, text and camera motion features,” in Proceedings of SPIE Storage and Retrieval for Media Databases, vol. 3972, pp. 332–343, 2000.

[Kobla 98] Kobla V., Doermann D.S. and Faloutsos C.: "Developing High-Level Representations of Video Clips using VideoTrails", In Proceedings of SPIE conference on Storage and Retrieval for Image and Video Databases VI, 81-92, 1998.

[Kokkoras 02] F. Kokkoras F., Jiang H., Vlahavas I.,_, Elmagarmid A.K., Houstis E.N., Aref W.G.: “Smart VideoText: a video data model based on conceptual graphs”, Multimedia Systems, 2002.

[Koubaroulis 97] Koubaroulis D., Matas J., and Kittler J.: "Colour-based image retrieval from video sequences", Proceedings of the Czech Pattern Recognition Workshop University of Brighton, UK, 1-12, 1997.

[Kraaij 04] Kraaij, W. & Smeaton, A. & Over P.: "TRECVID 2004– An Overview".

[Kumar 98] Kumar S., Phanendra **.** B.: "Intelligent multimedia data: data + indices + inference", Multimedia System, 395-407, 1998.

[kwon 02] Kwon S. and Narayanan S.: “Speaker Change Detection Using a New Weighted Distance Measure”, In Proceedings of International Conference Spoken Language Processing. Denver, Colorado, U.S.A., September 16-20, 2002.

[Lammens 94] Lammens J. M. : “A Computational Model of Color Perception and Color Naming,”. Ph.D. dissertation, State University of New York at Buffalo,1994.

[Lawrence 94] Lawrence A. R., Boreczky J. S., Charles A. E.: "Indexes for User Access to Large Video Databases", IEEE Trans. Knowl, Data Eng. Storage and Retrieval for Image and Video Databases (SPIE) Bellingham, Wash, 947-966, 1994.

[Lee 99] Lee J. C-M, Li Q. and Xiong W.: "Automatic and Dynamic Video Manipulation" Handbook of Multimedia Computing, Borko Furht, 317-343, 1999.

[Lin 03] Lin C-Y, Tseng B. L., and Smith J. R.: "VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning," IEEE Intl. Conf. on Multimedia & Expo (ICME), Baltimore, July 2003.

[Lindley 97] Lindley C. A.: "A Multiple-Interpretation Framework for Modelling Video Semantics", ER-97 Workshop on Conceptual Modelling in Multimedia Information Seeking 1997.

[Liu 98] Liu Z., Wang Y., and Chen T., "Audio feature extraction and analysis for scene segmentation and classification," VLSI Signal processing Systems for Signal, Image and Video Technology, vol. 20, pp. 61–79, 1998.

[Lozano 00] Lozano R. E. : "Système de Gestion de Bases de Données Multimédia et Vidéo", Thèse Informatique de l'université Joseph Fourier, Grenoble I, 2000.

[Ma 97] Ma J., Knight B., and Peng T.: "Representing Temporal Relationships between Events and Their Effects", Proceedings of the Fourth International Workshop on Temporal Representation and Reasoning, IEEE Computer Society Press, 148-152, 1997.

[Ma 99] Ma W. and Manjunath B. S., "NETRA: A toolbox for navigating large image databases," Multimedia Systems, vol. 7, no. 3, pp. 184–198, 1999.

[Marchand 00] Marchand M S. "Content-Based Video Retrieval: An Overview." Technical Rep. CUI - University of Geneva, Geneva, Switzerland: 2000.

[Martinet 04] Jean Martinet, Un modèle vectoriel relationnel de recherche d'information adapté aux images, Ph.D. thesis, Université Joseph Fourier, 2004.

[McNamee 02] McNamee P., Mayfield J.: "Entity Extraction without Language-specific Resources", In The 6th Conference on Natural Language Learning, Taipei (2002).

[Mechkour 95] Mechkour M., « EMIR², un modèle étendu de représentation et de correspondance d'images pour la recherche d'informations. Application à un corpus d'images historiques», Thèse Informatique de l'université Joseph Fourier, Grenoble I, novembre 1995.

[Meng 95] Meng J., Juan Y., and Chang S. F., "Scene change detection in a mpeg compressed video sequence," in Proceedings of the SPIE Symposium, vol. 2419, San Jose, CA, Feb. 1995, pp. 1–11.

[Mezoguchi 96] Mizoguchi R., Mihogaoka I., Katherine S. and Mitsuru I.: "Task Ontology Design for Intelligent Educational/Training Systems", Workshop on Architectures and Methods for Designing Cost-Effective and Reusable ITSs, Montreal, 1996.

[Minghong 99] Minghong L., Andreas A., Ansgar B.i, Knut H., and Michael S.: "Ontology for Knowledge Retrieval in Organizational Memories", Workshop on

Learning Software Organizations (LSO) at SEKE'99, Kaiserslautern, Germany. June 1999.

[Mohan 98] Mohan R., "Video sequence matching," in Proceedings of International Conference on Speech, Acoustics and Signal Processing, vol. 6, 1998, pp. 3697–3700.

[Nakamura 97] Nakamura Y. and Kanade T.:"Semantic analysis for video contents extraction-potting by association in news video," in Proceedings of ACM International Multimedia Conference, 1997.

[Nam 97] Nam J., Cetin A., and Tewfik A., "Speaker identification and video analysis for hierarchical video shot classification," in Proceedings of IEEE International Conference on Image Processing, vol. 2, Santa Barbara, CA, pp. 550–555, 1997.

[Naphade 00] Naphade M. and Huang T. S.: "Stochastic modeling of soundtrack for efficient segmentation and indexing of video," in Proceedings of SPIE Storage and Retrieval for Multimedia Databases, vol. 3972, pp. 168–176, 2000.

[Naphade 02] Naphade M.: "Statistical Media Analysis in Video Indexing", Internet Multimedia Management Systems III Boston, 138-150, 2002.

[Naphade 98] Naphade M., Mehrotra R., Ferman A. M., Warnick J., Huang T. S., and Tekalp A. M., "A high performance shot boundary detection algorithm using multiple cues," in Proceedings of IEEE International Conference on Image Processing, vol. 2, Chicago, IL, pp. 884–887, 1998.

[Nastaran 99] Fatemi N. and Philippe Mulhem, "A Conceptual Graph Approach for Video Data Representation and Retrieval", Third International Symposium, IDA-99, Amst, the Netherlands, August 1999.

[Neidle 01] Neidle, C., S. Sclaroff, and V. Athitsos : "SignStream: A Tool for Linguistic and Computer Vision Research", on Visual Gestural Language Data, Behavior Research Methods, Instruments, and Computers ,311-320, 2001.

[Nie 98] Nie, J.: An outline of a General Model for Information retrieval Systems. SIGIR , pp 495-506, Grenoble France, 1998.

[Nishida 99] Nishida, M. and Ariki, Y.,: "Speaker Indexing for News Articles, Debates, and Drama in Broadcasted TV Programs", IEEE International Conference on Multimedia Computing and systems, Volume Two, p.466-471, 1999.

[Paek 99] Paek S., Benitez A. B., and Chang S-F., "Self-Describing Schemes for Interoperable MPEG-7 Multimedia Content Descriptions", Image & Advanced TV Lab, Department of Electrical Engineering Columbia University, 1999.

[Patel 97] Patel N. V. and Sethi I. K., "Video segmentation for video data management," in The Handbook of Multimedia Information Management, W. Grosky, R. Jain, and R. Mehrotra, Eds., Upper Saddle River, NJ: Prentice Hall, PTR, pp. 139–165, 1997.

[Pickering 02] Pickering M., S Rüger, and D Sinclair: "Video retrieval by feature learning in key frames", In Int'l Conf on Image and Video Retrieval (CIVR) London, UK, 309-317, 2002.

[Pinquier 01] Julien Pinquier, Nicolas Chambert, " La première étape d'un système d'Indexation audio (Parole/Musique/Bruit)", quatrième rencontres jeune chercheurs en Parole, Mons(Belgique), septembre 2001.

[Prié 99] PRIÉ Y., "Modélisation de documents audiovisuels en Strates Interconnectées par les Annotations pour l'exploitation contextuelle", thèse Informatique, LISI INSA-Lyon, décembre 1999.

[Qi 00] Qi W., Gu L., Jiang H., Chen X.-R. and Zhang H-J.: "Integrating Visual, Audio And Text Analysis For News Video." 7th IEEE International Conference on Image Processing (ICIP'00) Vancouver, British Columbia, Canada: 2000.

[Quénot 01] Quénot, G. “TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation”, TREC 2001.

[Quénot 04] G. Quénot, D. Mararu, S.Ayache, M. Charhad, L. Besacier, P. Mulhem, M. Guironnet, D. Pellerin, J. Gensel, L. Carminati: CLIPS-LIS-LSR-LABRI Experiments at TRECVID2004: in TRECVID’2004 Workshop, Gaithersburg, MD, USA,, 17-18 November, 2004.

[Régine 02] Régine. A.-O. and Pinquier J. "Reconnaissance et Indexation de documents sonores." Journée AIM Bordeaux, France: 2002.

[Roach 02] Roach M., Mason J., Xu L-Q., and Stentiford F. W. M.: "Recent trends in video analysis: a taxonomy of video classification problems", 6th IASTED Int. Conf. on Internet and Multimedia Systems and Applications Hawaii, 348-353, 2002.

[Robinson 99] Robinson P., Brown, E. Burger J., Chinchor N., Douthat A., Ferro L., Hirschman, L.: Overview: Information extraction from broadcast news. In Proceedings of DARPA Broadcast News Workshop pp 27-30, Herndon, VA (1999).

[Rodriguez 04] Canseco-Rodriguez L., Lamel L., and Gauvain J.L.: “Speaker Diarization from Speech Transcripts”. In International Conference on Speech and Language Processing, pages 1272-1275, Jeju Island, October 2004.

[Rohini 00] Rohini K. Srihari, Zhang Z., Aibing R.: "Intelligent Indexing and Semantic Retrieval of Multimodal Documents", Information Retrieval Volume 2, 245-275, 2000.

[Salton 90] Salton G. and Buckley C.: “Improving retrieval performance by relevance feedback”, Journal of the American Society for Information Science, vol. 41, pp. 288–297, Juin 1990.

[Scheirer97] Scheirer E. and Slaney M., “Construction and evaluation of a robust multifeatures speech/music discriminator,” in Proceedings of IEEE Intl. Conf. on Accoustic, Speech and Signal Processing, vol. 2, Munich, Germany, pp. 1331–1334, 1997.

[Seyrat 98] Seyrat C., Durand G., Faudemay P. "Méthode d'indexation multimédia fondée sur les Objets Visuels" Actes des 4émes Journées d'Echanges Compression et Représentation des Signaux Audiovisuels (CORESA'98), Lannion, France, juin 1998.

[Shearer 96] Shearer K., Bunke H. and Venkatesh S.: "Video Indexing and Similarity Retrieval by Largest Common Subgraph Detection Using Decision Trees" Pattern Recognition volume 34, 19-25, 1996.

[Singla 00] Singla V., Park Y. C., Panchanathan S., Golshani F.: "Video Composition and Retrieval" IEEE International Conference on Multimedia and Expo New York, NY, USA, 1163-1166, 2000.

[Smith 96] Smith J. R and Chang S. F., “Visualseek: A fully automated content-based image query system,” in Proceedings of ACM Multimedia, Boston, MA, Nov. 1996.

[Snoek 01] Snoek C.G.M and Worring M.: "A Review on Multimodal Video Indexing" International Conference on Multimedia and Expo, 21-24, 2002.

[Souvannavong 02] Fabrice Souvannavong, Bernard Mérialdo, Benoit Huet: Semantic Feature Extraction using Mpeg Macro-block Classification. TREC 2002.

[Sowa 84] Sowa, J.F. "Conceptual structures: information processing in mind and machine", Addison-Wesley, 1984.

[Srinivasan 00] Srinivasan S., D. Ponceleon, Amir A, and Petkovic D.: “What is that video anyway?, In search of better browsing”, In Proceedings of IEEE International Conference on Multi-media and Expo, pp 388–392, 2000.

[Stevenson 00]Stevenson M. and Gaizauskas R., “Improving Named Entity Recognition using Annotated Corpora”, LREC Workshop on ``Information Extraction meets Corpus Linguistics", Athens, Greece (2000).

[Timothy 94] Timothy H., Daphne K., Jitendra M., Ogasawara G. H, B., Stuart J. R., and Weber J. "Automatic Symbolic Traffic Scene Analysis Using Belief Networks." in AAAI Workshop on AI in IVHS, Washington: 1994.

[Tran 00] Tran D. A., Hua K. A., and Vu K.: "VideoGraph: A Graphical Object-based Model for Representing and Querying Video Data" In Proceedings of International Conference on Conceptual Modeling, pp. 383-396, Salt Lake city, USA: 2000

[Tran 02] Tran T. T. : “Modélisation et traitement du contenu des médias pour l'édition et la présentation de documents multimédia”, thèse de doctorat, Institut National Polytechnique de Grenoble, mars 2002.

[Tsekeridou 00] Tsekeridou S., Krinidis S., Pitas I.: "Scene Change Detection Based on Audio-Visual Analysis and Interaction", Theoretical Foundations of Computer Vision Dagstuhl, Germany, 214-225, 2000.

[Tusch 00] Tusch R., Kosch H., Böszörményi L. "VIDEX: an integrated generic video indexing approach." ACM Multimedia Los Angeles, USA, 448-451, 2000.

[Vasconcelos 97] Vasconcelos N. and Lippman A.: "Content-based Pre-Indexed Video." Proceedings of International Conference on Image Processing Santa Barbara, California: 1997.

[Verma 03] Verma, R.C., Schmid, C., Mikolajczyk, K. "Face detection and tracking in a video by propagating detection probabilities." IEEE Trans. Pattern Analysis and Machine Intelligence 10, 1215-1228, Verma.

[Wactlar 96] Wactlar H., Kanade T., Smith M., and Stevens S., “Intelligent access to digital video: The informedia project,” IEEE Computer Digital Library Intiaive special issue, 1996.

[Wang 00] Wang Y., Liu Z., and Huang J., “Multimedia content analysis using audio and visual information,” IEEE Signal Processing Magazine, vol. 17, pp. 12–36, 2000.

[Wolf 01] Wolf C, Jolion J.- M, Chassaing F. : "Détection et extraction du texte de la vidéo", 7ème journées d'études et d'échanges, CORESA Dijon: 2001.

[Yang 04] Yang J., Hauptman A., Chen M-Y.: “Finding Person X: Correlating Names with Visual Appearances”, International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004.

[Yeo 95] Yeo B. L. and Liu B., “Rapid scene change detection on compressed video,” IEEE Trans- actions on Circuits and Systems for Video Technology, vol. 5, pp. 533–544, Dec. 1995.

[Yeung 95] Yeung M. M. and Liu B., “Efficient matching and clustering of video shots,” in Proceedings of IEEE International Conference on Image Processing, vol. 1, Washington, D.C., pp. 338–341, 1995.

[Zarri, 88] Zarri G. P.:"Conceptual representation for knowledge bases and “intelligent” information retrieval systems", Proceedings of the 11th annual international ACM SIGIR, May 1988, Grenoble, France.

[Zhai 04] Zhai C. X., John Lafferty J., A study of smoothing methods for language models applied to information retrieval, ACM Transactions on Information Systems, Vol. 2, No. 2, April 2004.

[Zhang 00] Zhang T. and Kuo C., “An integrated approach to multimodal media content analysis,” in Proceedings of SPIE, IS&T Storage and Retrieval for Media Databases, vol. 3972, pp. 506–517, 2000.

[Zhang 93] Zhang H. J., Kankanhalli A., Smoliar S. W.: "Automatic partitioning of full-motion video", Multimedia Systems New York, USA, 10-28, 1993..

[Zhang 94] Zhang H. J., C. Y. Low, and S. Smoliar, “Video parsing using compressed data,” in Proceedings of SPIE Conference on Image and Video Processing II, San Jose, CA, pp. 142–149, 1994.

[Zhong 97] Zhong D. and Shih-Fu Chang S-F.: "Spatio-temporal Video Search Using the Object Based Video Representation", Published in the International Conference on Image Processing, (ICIP'97), October 26-29, in Santa Barbara, CA, 1997.

[Zhuang 98] Zhuang Y., Rui Y., Huang T. S., and Mehrotra S., “Key frame extraction using unsupervised clustering,” in Proceedings of IEEE International Conference on Image Processing, vol. 1, Chicago, IL, pp. 866–870, 1998. http://www.w3.org/TR/1998/REC-xml-19980210.

Annexes

Annexe A

Exemples de Modélisation avec le Formalisme de Graphes Conceptuels Appliqués sur TRECVID 2004

L'ensemble de ces graphes est généré automatiquement à partir des transcriptions audio. Une première étape consiste à extraire les concepts, ensuite on associe ces concepts par des relations spécifiques au contenu audio pour déterminer qui parle et de quoi on parle dans chaque plan.

Video file 19981001_ABCa

Shot N° shot1_3

[Ao1] →(is a)→ [Person:M] → (has_ident) → [unknown]

[Ao2] →(is a)→ [Place: Texas]

[Ao1] → (speaks_about) →[Ao2]

[Ao1] → (speaks)

Shot N° shot1_4

[Ao3] →(is a)→ [Person:M] → (has_ident) → [unknown]

[Ao4] →(is a)→ [Place: Bangalore]

[Ao3] → (speaks_about) →[Ao4]

[Ao3] → (speaks)

Shot N° shot1_5

[Ao5] →(is a)→ [Person:M] → (has_ident) → [ron claiborne]

[Ao6] →(is a)→ [Place: New york]

[Ao5] → (speaks_about) →[Ao6]

[Ao5] → (speaks)

Shot N° shot1_6

[Ao7] →(is a)→ [Person:M] → (has_ident) → [peter jennings]

[Ao8] →(is a)→ [Place: Europe]

[Ao7] → (speaks_about) →[Ao8]

[Ao7] → (speaks)

Shot N° shot1_7

[Ao10] ->(is_a) -> [Person:M] -> (has_ident) -> [peter jennings]

[Ao10] ->(speaks)

[Ao11] ->(is_a) -> [organization:dow jones]

[Ao10] ->(speaks_about) -> [Ao11]

Shot N° shot1_8

[Ao12] ->(is_a) -> [Person:F] -> (has_ident) -> [betsy stark]

[Ao12] ->(speaks)

[Ao13] ->(is_a) -> [acronym:a.b.c.]

[Ao12] ->(speaks_about) -> [Ao13]

[Ao14] ->(is_a) -> [Person:F] -> (has_ident) -> [betsy stark]

Shot N° shot1_9

[Ao14] ->(speaks)

[Ao15] ->(is_a) -> [organization:intel]

[Ao14] ->(speaks_about) -> [Ao15]

[Ao16] ->(is_a) -> [organization:microsoft]

[Ao14] ->(speaks_about) -> [Ao16]

Shot N° shot1_10

[Ao17] ->(is_a) -> [Person:F] -> (has_ident) -> [unknown]

[Ao17] ->(speaks)

[Ao18] ->(is_a) -> [acronym:u.s.]

[Ao17] ->(speaks_about) -> [Ao18]

Annexe B
Les « Topics » TRECVID 2004

Chaque « topic » TRECVID est manuellement transformé en un graphe conceptuel dans le modèle CLOVIS. Ce graphe conceptuel ne traduit pas forcément exactement le « topic » tel qu'il est défini. Des adaptations sont nécessaires parce que les concepts et les relations présentes dans les « topics » ne correspondent pas toujours à ceux et celles du modèle.

Nous avons aussi utilisé des relations approximatives pour avoir des chances de trouver de bonnes réponses lorsque les relations disponibles ou effectivement indexées étaient insuffisantes. Par exemple, pour trouver des groupes de personnes, nous pouvons spécifier l'existence d'une (ou plusieurs personnes) et la présence d'une texture alignée (rayée dans notre classification) pour représenter des personnes côte à côte.

Par ailleurs, suivant les cas et/ou la formulation du « topic », une personne particulière peut être recherchée à partir de l'image ou à partir de l'audio. Par exemple, le contenu du « topic 125 » est destiné à une recherche dans le contenu visuel en exploitant les caractéristiques spatiales et mouvement de objets images. De nombreuses combinaisons restent possibles et nous n'en avons à chaque fois choisie qu'une ou deux. Les graphes sont représentés selon la notation linéaire. Nous présentons également une transcription synthétique en langage naturel.

125. Street scene with multiple pedestrians in motion and multiple vehicles in motion	
Personnes et voitures déconnectées et proches en mouvement	[VS] → (contient) → [Io1] *(avec VS : Video Segment)* [VS] → (contient) → [Io2] [Io1] → (is a) → [personne] [Io2] → (is a) → [voiture] [Io1] → (has object motion) → [<M_1:1, M_2:1, M_3:1, M_4:1, M_5:1, M_6:1, M_7:1, M_8:1>$_{OU}$] [Io2] → (has object motion) → [<M_1:1, M_2:1, M_3:1, M_4:1, M_5:1, M_6:1, M_7:1, M_8:1>$_{OU}$] [Io1] → (proche de) → [Io2] [Io1] → (déconnecté de) → [Io2]

Commentaire

Une interprétation de ce « topic » consiste à associer une description visuelle par deux objets images (Io1 et Io2) référant respectivement aux concepts « personne » et « voiture » en mouvement. Le type de mouvement est non spécifié. Ce qui laisse la possibilité d'avoir toutes les combinaisons possibles de 8 mouvements élémentaires dont une au moins est non nulle. En effet, dans ce cas le mouvement d'un objet image peut être : « translation horizontale (M_1 :1) », « translation verticale (M_2 :1) », « translation en profondeur (M_3 :1)», « rotation autour de l'axe horizontal (M_4 :1) », « rotation autour de l'axe vertical (M_5 :1) », « rotation autour de l'axe de profondeur (M_6 :1) », « homothétie (M_7 :1) » ou bien « transformation non linéaire (M_8 :1) »

Les relations « proche » et « déconnecté » dans la traduction CLOVIS permettent une description spatiale entre les objets images (Io1 et Io2).

126. One or more buildings with flood waters around it/them	
Bâtiment à l'intérieur d'une eau marron et d'aspect sali	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (is a) → [bâtiment] [Io2] → (is a) → [eau] [Io1] → (a l'intérieur de) → [Io2] [Io2] → (has texture) → [<T_1:0, T_2:0, T_3:0, T_4:0, T_5:0, T_6:0, T_7:0, T_8:**1**, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$] [Io2] → (has color) → [<C_1:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:**1**, C_{11}:0>$_{ET}$]
Commentaire	
Les segments vidéo contenant deux objets images (Io1 et Io2) qui correspondent respectivement aux concepts « bâtiment » et « eau ». Une description de l'objet image Io2 dans la sous-facette couleur est représentée par : [Io2] → (has color) → [<C_1:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}: **1**, C_{11}:0>$_{ET}$] signifie que l'objet Io2 a une couleur peau/marron (« rouge (C_1 :0)», « blanc (C_2 :0) », « (C_3 :0) bleu », « gris (C_4 :0) », « cyan (C_5 :0) », « vert (C_6 :0) », « jaune (C_7 :0) », « violet (C_8 :0) », « noir (C_9 :0) », « **marron (C_{10} :1)** », « orange (C_{11} :0) »). Cette description est équivalente à : [Io2] → (has color) → [marron] Une description texture de l'objet image Io2 est représentée par : [Io2] → (has texture) → [sali] « bosselé (T_1 :0) », « craquelé (T_2 :0) », « désordonné (T_3 :0) », « entrelacé (T_4 :0) », « rayé (T_5 :0) », « marbré (T_6 :0) », « rétiforme (T_7 :0) », « **sali (T_8 :1)** », « tacheté (T_9 :0) », « uniforme (T_{10} :0) » et « en vague (T_{11} :0) ».	

127. One or more people and one or more dogs walking together	
Objets images ayant une texture rayée et représentant des personnes et chiens en mouvement de translation ou non linéaire	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (has object motion) → [<M_1:**1**, M_2:0, M_3:**1**, M_4:0, M_5:0, M_6:0, M_7:0, M_8:**1**>$_{OU}$] [Io2] → (has object motion) → [<M_1:**1**, M_2:0, M_3:**1**, M_4:0, M_5:0, M_6:0, M_7:0, M_8:**1**>$_{OU}$] [Io1] → (has texture) → [<T_1:0, T_2:0, T_3:0, T_4:0, T_5:**1**, T_6:0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$] [Io2] → (has texture) → [<T_1:0, T_2:0, T_3:0, T_4:0, T_5:**1**, T_6:0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$] [Io1] → (is a) → [personne] [Io2] → (is a) → [chien]
Commentaire	
Ce « topic » est traduit dans CLOVIS par deux concepts (personne et chien) représenté au niveau visuel par deux objets images ayant une texture rayée (« **rayé (T5 :1)** »). Pour interpréter le concept mouvement (marcher) dans le « topic », nous considérons que ce concept correspond à un mouvement « translation horizontale **(M_1 :1)** », « translation en profondeur **(M_3 :1)** », ou bien un mouvement de « transformation non linéaire **(M_8 :1)** » [Io1] → (has object motion) → [<M_1:**1**, M_2:0, M_3:**1**, M_4:0, M_5:0, M_6:0, M_7:0, M_8:**1**>$_{OU}$] est équivalente à: [Io1] → (has object motion) → [« translation horizontale » ou « translation en profondeur » ou « transformation non linéaire »]	

128. US Congressman Henry Hyde's face, whole or part, from any angle **135. Sam Donaldson's face. No other people visible with him**	
Personne X visible	[VS] → (contient) → [Io1] [Io1] → (is a) → [personne X] [Io1] → (apparaît)
Commentaire	
La traduction des « topics » 128 et 135 est équivalente à rechercher dans le contenu visuel de la vidéo un concept personne X qui est visible. Ce « topic » est décrit l'ensemble des segments vidéo qui contiennent des objets images représenté au niveau symbolique par un concept « personne ». La représentation [Io1] → (apparaît) spécifie l'objet image Io1 est visible dans le segment.	

129. US Capitol dome	
Bâtiment majoritairement blanc avec une texture rayée	[VS] → (contient) → [Io1] [Io1] → (has color) → [$<C_1$:0, C_2:**1**, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0$>_{ET}$] [Io1] → (has texture) → [$<T_1$:0, T_2:0, T_3: 0, T_4:0, T_5:**1**, T_6:0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0$>_{ET}$] [Io1] → (is a) → [bâtiment]

130. Hockey rink with at least one of the nets fully visible from some point of view	
Objet image avec une texture rétiforme au-dessus d'un sol marbré majoritairement blanc	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] →(is a) → [personne] [Io1] → (touche) → [Io2] [Io2] → (has color) → [$<C_1$:0, C_2:**1**, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0$>_{ET}$] [Io2] → (has texture) → [$<T_1$:0, T_2:0, T_3: 0, T_4:0, T_5:0, T_6:0, T_7:**1**, T_8:0, T_9:0, T_{10}:0, T_{11}:0$>_{ET}$]

131. Fingers striking the keys on a keyboard which is at least partially visible	
Personnes touchant un objet image avec une texture rétiforme et une couleur blanche	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] →(is a) → [personne] [Io1] → (touche) → [Io2] [Io2] → (has color) → [<C_1:0, C_2:**1**, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$] [Io2] → (has texture) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5:0, T_6:0, T_7:**1**, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$]

132. People moving a stretcher	
Personne(s) touchant un objet image en mouvement avec objet image au dessus de personne(s)	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (is a) → [personne] [Io1] → (touche) → [Io2] [Io2] → (au dessus de) → [Io1] [Io2] → (has object motion) → [<M_1:**1**, M_2:**1**, M_3:**1**, M_4:**1**, M_5:**1**, M_6:**1**, M_7:**1**, M_8:**1**>$_{OU}$]
Commentaire	
Une interprétation de ce « topic » par une description visuelle liée à deux objets images (Io1 et Io2). L'objet image Io1 est représenté par le concept « personne » L'objet image Io2 est en mouvement. Le type de mouvement ici peut être : « translation horizontale (M_1 :1) », « translation verticale (M_2 :1) », « translation en profondeur (M_3 :1)», « rotation autour de l'axe horizontal (M_4 :1) », « rotation autour de l'axe vertical (M_5 :1) », « rotation autour de l'axe de profondeur (M_6 :1) », « homothétie (M_7 :1) » ou bien « transformation non linéaire (M_8 :1) » Les relations « au dessus de » et « touche » dans la traduction CLOVIS sont utilisées pour associer une description spatiale entre les objets images (Io1 et Io2).	

133. Saddam Hussein **134. Boris Yeltsin** **137. Benjamin Netanyahu**	
Personne X	[VS] →(contient)→ [Ao1] [Ao1] →(is a)→ [personne] [Ao1] → (parle) OU [VS] →(contient)→ [Ao1] [VS] →(contient)→ [Ao2] [Ao1] →(is a)→ [personne] [Ao2] →(is a)→ [personne] [Ao1] → (parle de) →[Ao2]
Commentaire	
les « topics » 133, 134, 137 peuvent être interprétés en utilisant le contenu audio. Cette interprétation correspond soit aux segments vidéo contenant un objet audio (Ao) représenté par le concept « personne X » qui est le locuteur dans ce cas. Soit l'ensemble des segments vidéo contenant deux objets audio (Ao1 et Ao2) avec Ao1 correspond à un locuteur et Ao2 correspond à la personne X recherchée.	
136. Person hitting a golf ball that then goes into the hole	
Personne au-dessus et touchant un sol uniforme et majoritairement vert, lui-même au-dessus d'un objet image blanc	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [VS] → (contient) → [Io3] [Io1] → (is a) → [personne] [Io2] → (is a) → [sol] [Io1] → (au-dessus de) → [Io2] [Io2] → (has color) → [<C_1:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:**1**, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$] [Io2] → (has texture) → [Io1] → (has texture) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5: 0, T_6:0, T_7:0, T_8:0, T_9:0, T_{10}:**1**, T_{11}:0>$_{ET}$] [Io2] → (au dessous) → [Io3] [Io3] → (has color) → [<C_1:0, C_2:**1**, C_3:0, C_4:0, C_5:0, C_6: 0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$]

138. One or people going up or down some visible steps or stairs	
Personnes en mouvement touchant et au-dessus d'escaliers	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [VS] → (contient) → [Io3] [Io1] → (is a) → [personne] , [Io2] → (is a) → [personne] [Io3] → (is a) → [escalier] [Io1] → (touche) → [Io2] → (au-dessus de) → [Io3] [Io1] → (has object motion) → [$<M_1$:0, M_2:**1**, M_3:0, M_4:0, M_5:0, M_6:0, M_7:0, M_8:**1**$>_{OU}$] [Io2] → (has object motion) → [$<M_1$:0, M_2:**1**, M_3:0, M_4:0, M_5:0, M_6:0, M_7:0, M_8:**1**$>_{OU}$]

139. Handheld weapon firing	
Personne touchant un objet image noir et déconnectée du feu	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [VS] → (contient) → [Io3] [Io1] → (is a) → [personne] [Io2] → (is a) → [feu] [Io1] → (deconnecté) → [Io2] [Io3] → (has color) → [$<C_1$:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:**1**, C_{10}:0, C_{11}:0$>_{ET}$]

140.One or more bicycles rolling along 144.One or more horses in motion	
Objet image ayant une texture rayée et représente un objet X en mouvement	[VS] → (contient) → [Io1] [Io1] → (is a) → [Objet X] [Io1] → (has texture) → [$<T_1$:0, T_2:0, T_3:0, T_4:0, T_5:**1**, T_6:0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0$>_{ET}$] [Io1] → (has object motion) → [$<M_1$:**1**, M_2:**1**, M_3:**1**, M_4:**1**, M_5:**1**, M_6:**1**, M_7:**1**, M_8:**1**$>_{OU}$]

141. Find shots of one or more umbrellas	
Objet image noir au-dessus d'une personne	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (is a) → [personne] [Io2] → (au dessus de) → [Io1] [Io2] → (has color) → [<C_1:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:**1,** C_{10}:0, C_{11}:0>$_{ET}$]

142. Tennis player contacting the ball with his or her tennis racket	
Personne touchant un objet image en texture rayée lui même proche d'un objet image majoritairement jaune	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [VS] → (contient) → [Io3] [Io1] → (is a) → [personne] [Io2] → (proche de) → [Io3] [Io2] → (has texture) → [<T_1:0, T_2:0, T_3:0, T_4:0, T_5: **1**, T_6: 0, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$] [Io3] → (has color) → [<C_1:0, C_2:0, C_3:0, C_4:0, C_5:0, C_6:0, C_7:**1**, C_8:0, C_9:0**,** C_{10}:0, C_{11}:0>$_{ET}$]

143. Bill Clinton speaking with at least part of a US flag visible behind him	
Bill Clinton devant un drapeau blanc, bleu et rouge présentant des textures rayée et uniforme	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (is a) → [Bill Clinton] [Io2] → (is a) → [drapeau] [Io2] → (has color) → [<C_1:**1**, C_2:**1**, C_3:**1**, C_4:0, C_5:0, C_6:0, C_7:0, C_8:0, C_9:0, C_{10}:0, C_{11}:0>$_{ET}$] [Io2] → (has texture) → [<T_1:0, T_2:0, T_3: 0, T_4:0, T_5:0, T_6:**1**, T_7:0, T_8:0, T_9:0, T_{10}:0, T_{11}:0>$_{ET}$] [Io1] → (devant) → [Io2]

146. One or more buildings on fire	
Bâtiments proches du feu	[VS] → (contient) → [Io1] [VS] → (contient) → [Io2] [Io1] → (is a) → [bâtiment] [Io2] → (is a) → [feu] [Io1] → (proche de) → [Io2]

147. One or more signs or banners carried by people at a march or protest	
Personnes touchant des objets images au-dessus d'eux-mêmes	[VS] → (contient) → [Io1] [Io1] → (is a) → [bannière] [Io2] → (is a) → [personne] [Io1] → (au dessus de) → [Io2]