INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

$N^o$ _ _ _ _ _ _ _ _ _ _

# THESE

pour obtenir le grade de

**DOCTEUR DE L'INPG**

**Spécialité : IMAGERIE, VISION ET ROBOTIQUE**

préparée au laboratoire GRAVIR-IMAG et INRIA Rhône-Alpes
dans le cadre de l'école Doctorale
Mathématiques, sciences et technologie de l'information

présentée et soutenue publiquement

par

**Riad HAMMOUD**

le 27 février 2001

# Construction et Présentation des Vidéos Interactives

Directeur de thèse : Roger MOHR

Jury

| | |
|---|---|
| M. Augustin LUX | , President |
| Mme. Françoise PRETEUX | , Rapporteur |
| M. Theirry PUN | , Rapporteur |
| M. Liming CHEN | , Rapporteur |
| M. Roger MOHR | , Examinateur |
| Mme. Cordelia SCHMID | , Examinateur |

à mes parents, mes soeurs et frères.

à Aba Saleh et Aba Hadi.

à Safi Dager.

# Remerciements

Trois ans déjà ! Me voilà disant merci à ceux qui ont rendu cette aventure possible et tous ceux qui l'ont ornée d'échanges ou d'amitiés.

Je voudrais donc tout d'abord remercier Roger Mohr, mon directeur de thèse, pour m'avoir accueilli dans l'équipe MOVI. Je lui suis très reconnaissant d'avoir poursuivi mon encadrement lors de son séjour à Melbourne et durant sa direction du centre de recherche européen de XEROX à grenoble. Sa confiance est au commencement de tout.

Je tiens à remercier le centre de recherche d'ALCATEL d'avoir financer cette étude et aussi le chef du projet MOVI, Radu Horaud, pour sa proposition de financer les trois derniers mois (février, mars et avril) de mon séjour à l'INRIA RHÔNE ALPES. Également, je remercie l'équipe de Patrick Bouthemy pour sa collaboration, et aussi l'Institut National de l'Audiovisuel de nous avoir donné la permission d'utiliser leurs données vidéos.

Je remercie vivement les membres de mon jury pour l'honneur qu'ils m'ont fait. Merci à M. Augustin Lux d'avoir accepté de présider mon jury. Merci à MME. Françoise Prêteux et M. Thierry Pun, pour les remarques constructives qui ont permis d'améliorer ce mémoire. Merci à tous ceux qui ont rendu la lecture de ce mémoire et de mes publications plus agréable. Je voudrais en particulier cité Matthieu Personnaz, Roger Mohr, Augustin Lux, Bill Trrigs et Ragini Choudhury.

Je suis très reconnaissant aux gens qui ont consacré du temps pour faire un échange scientifique et technique. Je voudrais en particulier cité Christophe Biernacki, Partick Gros et Pascal Bertolino. Également, je salut tous les stagiaires qui ont travaillé sous ma direction : Donacien Guifokou, MME Leila Cammoun, Mathieu Rudant, Khassoum Wone, et Guillaum Durant.

Ma reconnaissance va encore vers tous les membres du projet MOVI, pour leur esprit d'équipe et leur cordialité, mais aussi plus largement aux membres des équipes IS2, PRIMA et OPERA.

Enfin, il me faut saluer tous les amis qui m'ont entouré pour leurs services et leurs encouragements jusqu'à ce dernier moment.

# Sommaire

# Table des figures

*Chapitre 1*

# Introduction

*Un projet d'entreprise ne peut tenir debout que sur trois pieds : un concept solide, un créateur déterminé et une préparation rigoureuse.*

## 1.1 Contexte et motivation

Depuis l'apparition des standards de compression vidéo MPEG-1 et MPEG-2, l'évolution rapide des performances des processeurs et des technologies de stockage comme le CDROM et le DVD, et la forte présence du Web et la facilité d'y accéder via le réseau Internet, la vidéo numérique dispose aujourd'hui d'une technologie qui la rend accessible comme les autres données.

Les applications de la vidéo sont en constante augmentation : les services de la "vidéo à la demande", l'enseignement à distance, la vidéo surveillance, les bases de données vidéo de films et de journaux télévisés, la "vidéo en continu" sur le Web (*streaming*), "TV Any-Time", etc. De nouvelles fonctionnalités essentiellement liées à l'**interactivité** séduisent de plus en plus les grandes entreprises dans ce domaine d'applications. On peut ainsi dès à présent visionner sur le Web des vidéo panoramiques et obtenir automatiquement des informations supplémentaires à certains moments-clés d'un reportage dite de "**vidéo enrichie**".

### 1.1.1 Structures et hyperliens

Le document vidéo a une structure qui n'apparaît pas explicitement, à la différence d'un ouvrage avec ses chapitres et ses paragraphes. Permettre une manipulation avancée d'une vidéo demandera donc de mettre en évidence ces éléments.

Les premières entités à considérer sont l'équivalent des mots dans un texte: les objets élémentaires dans un film. Il nous faudra donc extraire de tels objets qui se confondent avec le fond; le chapitre 2 détaille cet aspect. L'équivalent de ce que pourrait être un paragraphe serait le plan vidéo, et des structures de plus haut niveau existent comme les scènes et les séquences.

Une fois cette structure créée, des hyperliens sur les entités existantes jouent le même rôle que dans un document écrit; les liens peuvent typiquement être classés dans deux catégories :

- **Lien interne :** il lie des entités dans la vidéo, typiquement lorsqu'une relation chronologique et/ou sémantique existe entre deux ou plusieurs entités d'un document vidéo. Le même objet dans le plan suivant est un exemple (voir figure 1.1).

- **Lien externe :** il s'agit d'un lien classique permettant par exemple d'attacher une page web à un objet dans un plan. Un exemple d'application publicitaire serait ainsi d'attacher un descriptif technique à un véhicule visible dans une vidéo.

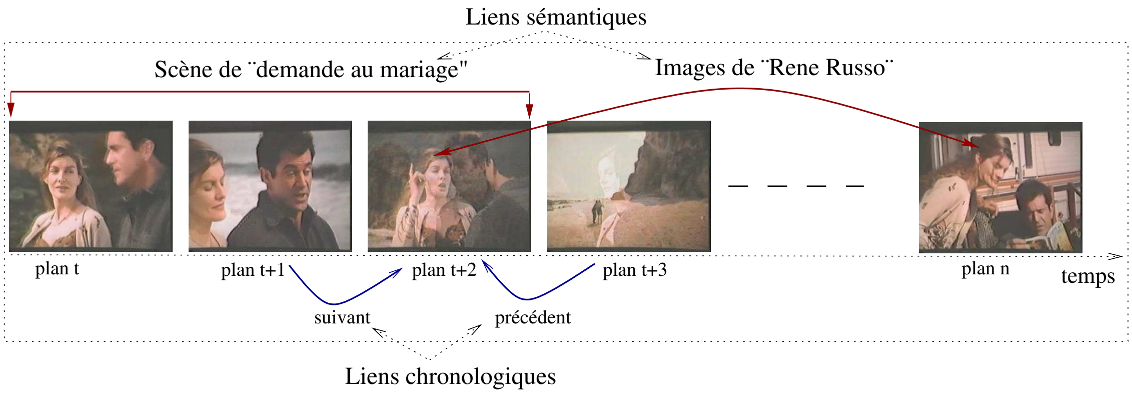


FIG. 1.1: *Exemple des hyperliens intra-film.*

### 1.1.2 Interactivité

Avec un magnétoscope classique, on peut visualiser un film vidéo, rebobiner pour observer un segment plusieurs fois, avec une vitesse rapide ou lente, sauter des segments, ou bien faire une pause. Ces capacités de contrôle de la vidéo rendent la vidéo interactive pour certains utilisateurs. Mais l'interaction ici est sévèrement limitée. Pour être vraiment interactive la vidéo doit offrir des niveaux d'interactions plus liés à son contenu; par exemple on peut imaginer l'option suivante : "*Indiquez l'acteur qui vous intéresse dans le film pour avoir plus d'information sur lui* ou bien "*Un clic sur un objet permettra de visualiser la scène suivante/précédente où il est présent*". Un exemple d'une vidéo interactive est illustré dans la figure 1.2.

Les propositions sur l'interactivité ont commencés à apparaître très récemment au début de l'année 1998. Plusieurs sociétés telles Intervu, Langages virtuels et Artsvidéo Interactive, ont proposé d'insérer de façon transparente aux vidéos une couche d'interactivité compatible avec les trois principaux formats de diffusion sur le Réseau (Realplayer, Quicktime et Windows Media Player). En 1999, deux nouveaux formats de vidéo enrichie

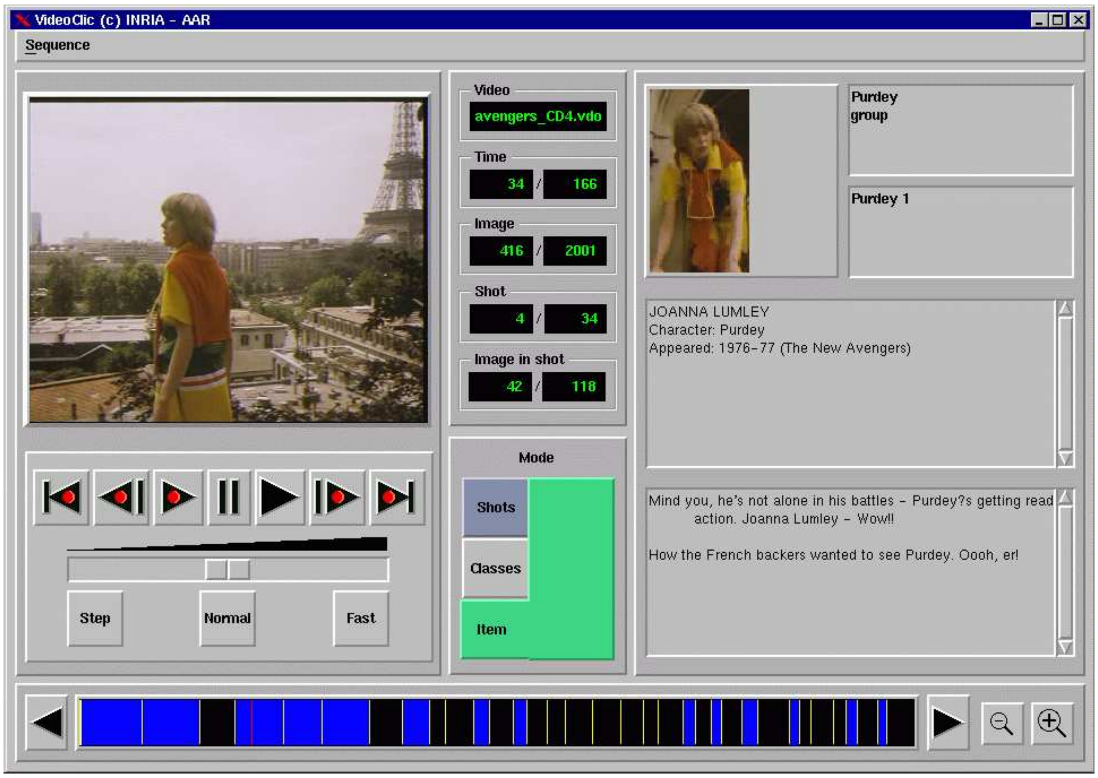


FIG. 1.2: *Application d'une vidéo interactive : l'utilisateur demande plus d'informations sur l'actrice "Purdey" dans le film "Avengers"; par un simple clic sur une de ces apparitions dans la vidéo (coin haut-gauche), des informations diverses (histoire, rôle, ...) s'affichent à coté de la vidéo et des outils de navigation avancés sont activés qui lui permettent par exemple de se positionner directement sur la scène ou l'apparence suivante/précédente de Purdey.*

sont apparus, le SMIL (*Synchronised Multimedia Interface Language*) standard des logiciels Realplayer et Quicktime, et le format ASF (*Advanced Streaming Format*), utilisé par le logiciel Windows Media Player.

L'interaction à laquelle ce travail se limite est liée à la navigation utilisant la structure de vidéo et les hyperliens mentionnés ci-dessus. Cette interaction permet non seulement une navigation en suivant les hyperliens, mais offre des extensions naturelles comme des recherches à granularité variable : parcours accéléré de scènes en scènes à la recherche d'une action particulière où apparaît un acteur donné; recherche plus détaillée d'un plan particulier à l'intérieur d'une scène.

Quelques éléments de cette spécification de l'interactivité sont offerts par des produits commerciaux en vente aujourd'hui, comme par exemple "Mvshots" et "Movideo 2 Studio" conçus par la société française Artsvidéo Interactive. Le produit Mvshots possède une fenêtre de navigation dans la structure plan-séquences, et Movideo offre en plus une interaction au niveau des objets qui ont manuellement été segmentés et rattachés à des liens internes et externes.

## 1.2 Les problèmes soulevés

L'interaction entre l'homme et la vidéo se fait nécessairement au travers de la machine. Il en découle que les problèmes évoqués plus haut interviennent à deux niveaux : au niveau Homme/Machine et au niveau Machine/Vidéo. Au niveau Machine/Vidéo, l'incompréhension est totale : la machine n'est pas capable de *comprendre la vidéo* et donc a priori de la manipuler. Un traitement, globalement désigné sous le terme d'**indexation**, doit être mis en oeuvre pour structurer un document vidéo brut et en faciliter la manipulation. Au niveau Homme/Machine, c'est la représentation qui pose le plus de problèmes. L'homme est capable de *comprendre la vidéo* mais il n'arrive pas à communiquer avec une machine qui a du mal à le comprendre et à lui montrer facilement ce qui l'intéresse. Les requêtes portent généralement sur un concept sémantique entièrement absent chez la machine. Par exemple, la machine n'est pas capable d'interpréter une requête simple du genre : "Trouvez moi la scène où Mr. Bean est sur la plage".

Au niveau conceptuel de cet handicap entre l'homme et la machine se rajoute une difficulté technique dûe à la nature à la fois temporelle et spatiale de la vidéo. D'abord, la vidéo représente un volume énorme d'information (1 heure de vidéo en format compressé correspond à 575 Mega octets environ). Ensuite, les entités présentes dans cette vidéo n'ont pas de caractérisation fiable et ne sont pas extractibles : un personnage se confond plus ou moins dans la scène, et ses aspects varient considérablement durant le film. On se trouve donc dans une situation très différente du texte où un mot se localise facilement; il a un ensemble de sens très limité et très peu de variations orthographiques.

Face à cette double difficulté, nous listons ci-dessous les problèmes à résoudre pour atteindre nos objectifs :

- **Extraction des objets :** L'intérêt qu'un spectateur donne au contenu visuel d'une vidéo se focalise essentiellement sur les objets comme personnages, voitures, objets commerciaux, etc. L'extraction de ces objets, qu'elle soit automatique ou semi-automatique, représente une tâche très complexe. Ceci est dû principalement au fait que les modèles de ces objets ne sont pas connus a priori, au contraire de l'extraction des visages ou des doigts par exemple.

- **Indexation d'objets vidéo :** L'indexation de toutes les images ou objets d'une vidéo n'est pas faisable à cause du nombre énorme des images à indexer (à un film de 1h30 correspondent 129600 images). Les dimensions des descripteurs à extraire de ces images sont généralement assez grandes et celles-ci rendent inefficaces les structures de données souvent utilisées pour gérer ces descripteurs et rechercher des requêtes.

- **Appariement des objets :** Il s'agit ici d'identifier les différentes apparences du même objet. Ces apparences intra- et inter-plans vidéo sont très variables et rendent leur identification très délicate. Le changement d'apparence est dû à la nature de la vidéo : les images sont acquises dans des conditions naturelles et le montage des scènes exige de filmer les objets (acteurs, voitures, ...) sous une grande variété de prises de vues, de changements d'échelles et d'éclairages, etc.

- **Classification des objets vidéo :** La classification est une technique qui regroupe les apparences d'objets en différentes classes d'équivalence. Une telle technique est pratiquement plus difficile que l'appariement. Car, d'une part la classification hérite les problèmes de l'appariement qui lui représente un processus "interne", et d'autre part en classification automatique il s'agit de déterminer le nombre de groupes d'objets à fabriquer, ce qui n'est pas du tout facile sur des données simulées, voir des objets d'apparences variables.

- **Segmentation élémentaire de la vidéo :** Il s'agit ici de regrouper les images successives en des unités qui correspondent à des prises de vues de la caméra souvent connues sous le nom de "plans vidéo". Notons que toutes les méthodes de traitement du contenu de la vidéo (suivis d'objets dans les plans, similitudes entre plans, etc.) se basent sur ces unités et donc la fiabilité de leurs résultats dépend de la précision de la détection des plans.

- **Sélection des images- apparences-clés :** Pour permettre une visualisation rapide du contenu d'un plan vidéo et afin de simplifier la présentation d'un très grand nombre de plans, la sélection des images ou apparences les plus représentatives du plan ou objet suivi semble être un point important à résoudre. Souvent le plan vidéo est résumé par l'image médiane. Ce choix n'est pas toujours valide surtout quand le plan n'est pas fixe et son contenu varie (objets mobiles).

- **Macro-segmentation de la vidéo en scènes :** Lorsqu'il s'agit d'un film vidéo d'une heure contenant environ par exemple 1000 plans, il devient très difficile de les représenter graphiquement et à les explorer linéairement par l'utilisateur. En effet, un découpage plus macroscopique que les plans vidéo en des unités appelées *scènes* est réalisé. Cette technique de macro-segmentation peut-être vue comme un processus qui consiste à créer une partition de l'ensemble de tous les plans d'un film vidéo. Cela veut dire que deux problèmes se présentent face à un processus de macro-segmentation : le problème de la classification automatique des plans et le problème de la définition de la sémantique lors de la classification des plans.

- **Interaction entre utilisateur et vidéo hyperliée :** Lors de la création d'une vidéo hyperliée l'intervention de l'"'auteur de la vidéo hyperliée" peut-être indispensable, par exemple aux endroits où les méthodes automatiques d'indexations échouent. A ce niveau, il s'agit de définir précisément ces interactions et de fournir des outils efficaces de présentations et d'éditions des résultats d'indexation. Aussi, lors de la présentation d'une vidéo hyperliée à un utilisateur final, il s'agit ici de fournir des interfaces graphiques simples et intuitives, qui reflètent la puissance et la souplesse d'utilisation d'une vidéo hyperliée.

## 1.3 Les contributions de cette thèse

Face aux problèmes cités au-dessus les contributions principales de cette thèse peuvent être résumées par les points suivants :

- **Modélisation de la variabilité intra-plan :** Nous proposons de modéliser la variabilité intra-plan d'un objet suivi d'un plan vidéo, dans l'espace de descripteurs, par une fonction statistique de *densité de mélange gaussien*, qui **capture** les différentes apparences intra-plan de l'objet. L'idée ici est de classifier les apparences intra-plan semblables en des groupes. Ensuite, chaque groupe d'apparences sera représenté par ses *paramètres statistiques*, ce qui **réduit** considérablement le nombre des descripteurs à indexer par objet suivi (par plan vidéo).

- **Identification d'objets à travers la vidéo par une classification supervisée :** Dans un premier temps, l'auteur de la vidéo hyperliée choisi d'une façon interactive les "modèles d'objets suivis". Dans un second temps, une modélisation de la variabilité de ces objets est effectuée. En adoptant le mélange gaussien dans ce cas de discrimination, deux techniques sont employées pour classer les requêtes (toutes les apparences d'objets dans le film) : le **classement individuel** par maximum a posteriori (MAP) et le **classement robuste** par vote majoritaire.

- **Identification d'objets à travers la vidéo par une classification automatique :** Pour automatiser la classification des objets suivis, on calcule des distances adaptées à nos données statistiques représentant les objets suivis : *les densités de mélanges gaussiens*. La distance entre deux objets suivis est calculée par le minimum de la distance de Kullback (ou de Bhattacharyya ) entre les composantes gaussiennes des deux densités de mélanges correspondant. Sur la matrice de distances ainsi obtenue, une classification ascendante hiérarchique est effectuée. Une méthode interactive a été proposée également pour choisir le nombre de classes d'objets suivis. De même des outils graphiques et interactifs sont mis à la disposition de l'auteur de la vidéo interactive pour corriger éventuellement de mauvaises classifications d'objets ainsi obtenues.

- **Sélection des images représentatives d'un plan :** Face au problème de sélection des images- apparences-clés, et en s'inspirant des travaux précédents sur la modélisation de la variabilité intra-plan des objets suivis, il était facile pour nous de proposer une technique performante à ce problème. Après l'identification des classes d'apparences intra-plan, chaque classe d'apparences est d'abord représentée par l'apparence médiane, et ensuite deux tests de vérification temporelle et spatiale permettent de ne garder que les images- apparences-clés les plus représentatives. Ce choix des images-clés est automatique.

- **Macro-segmentation d'une vidéo en scènes :** Pour la macro-segmentation d'un film vidéo, nous avons mené d'abord une étude expérimentale sur la comparaison de deux plans vidéo sur la base de trois descripteurs de bas niveaux : le contenu du plan, la distribution globale de la couleurs et la distribution spatiale de la couleurs de l'image médiane du plan. La distance proposée pour mesurer la similarité entre deux plans sur la base du descripteur du contenu est le nombre des objets similaires de deux plans. Cette étude a démontré l'importance d'une telle stratégie de caractérisation mixte du plan vidéo. De là, nous avons étendu une approche existante de macro-segmentation [53] [25]. Cette extension est résumée par une étape de fusion des classes de plans obtenues par les différents descripteurs. Une amélioration nette des résultats de l'approche de base est réalisée sur plusieurs séquences vidéo.

- **Implémentation de ces outils dans un prototype :** En analysant les performances et les limites des méthodes utilisées lors de la construction de la vidéo hyperliée, nous avons défini les interactions entre l'indexeur, la vidéo et la machine. Ensuite on les a implémentées dans un système de fabrication de la vidéo hyperliée appelé **VideoPrep**. Les résultats obtenus après cette étape sont utilisés par un autre système de présentation de la vidéo hyperliée nommé **VideoClic**.

Des démonstrations pour l'ensemble des approches proposées dans cette thèse sont disponibles sur le web à l'adresse suivante :

`http://www.inrialpes.fr/movi/pub/Demos/DemoRiad`.

## 1.4 Organisation de ce mémoire

**Chapitre 2.** Il rappelle la structure hiérarchique d'un film vidéo, ainsi que les méthodes de segmentation en plans et de suivis d'objets dont nous nous sommes servis durant ce travail.

**Chapitre 3.** Les problèmes d'indexation et de variabilité abordés plus haut, ainsi que l'approche retenue pour palier à ces deux problèmes, sont discutés en détail dans ce chapitre. Les données expérimentales (descripteurs de couleurs) et leur pré-traitement sont décrites ici.

**Chapitre 4.** Il décrit l'approche semi-automatique de classification des objets basée sur les modèles de variabilité estimés dans le chapitre précédent. Une validation expérimentale de cette approche sur la séquence vidéo "**Avengers-1**", ainsi qu'une comparaison de performance avec une méthode classique d'appariements sont également présentées.

**Chapitre 5.** L'objet de ce chapitre est de décrire en détail l'approche de classification automatique des objets suivis dans l'espace de paramètres gaussiens. Les expérimentations sont menées sur la séquence vidéo "**Avengers-2**".

**Chapitre 6.** Il détaille l'algorithme de sélection des apparences-clés d'un objet suivi.

**Chapitre 7.** Ce chapitre rappelle les deux étapes de base pour la macro-segmentation en scènes, ensuite il présente l'extension proposée durant cette thèse. Des résultats sur la séquence "**Dances with Wolves**" sont montrés pour l'approche de base et son extension.

**Chapitre 8.** Les niveaux d'interactions entre Utilisateur/Vidéo, Utilisateur/Machine et Machine/Vidéo ainsi que les deux systèmes **VideoPrep** et **VideoClic** sont l'objet de ce chapitre.

**Chapitre 9.** Cet ultime chapitre présente le bilan du travail effectué durant ces trois ans de thèse. Quelques perspectives de recherches et d'applications terminent ce mémoire.