Chapitre 2

Généralités sur la structuration de la vidéo

Je reglerai au plus vite
les petits pépins.

L'objectif de ce chapitre est de rappeler la structure hiérarchique d'un document vidéo et de décrire brièvement les méthodes utilisées dans ce travail pour segmenter la vidéo en plans élémentaires et pour suivre automatiquement ou semi-automatiquement des régions mobiles à travers le temps. Nous nous servirons des méthodes développées par l'équipe Vista de l'Inria Rennes suite à une collaboration entre trois partenaires : Alcatel Alsthom Recherche et les deux projets de l'Inria Movi et Vista. Après une courte présentation de ces outils nous évaluons leurs performances sur des exemples réels. Ensuite, nous décrirons le corpus vidéo utilisé dans l'évaluation de nos approches, et avant de conclure on introduit le future standard MPEG-7.

2.1 Structure hiérarchique de la vidéo

Plusieurs catégories des documents vidéo existent dans la réalité : les journaux télévisés, les émissions sportives (foot, tennis, etc.), les dessins animés, les films, etc. Parmi eux il y en a qui possèdent un modèle a priori (e.g. les journaux télévisés) et d'autres qui possèdent une histoire (e.g. les films). L'analyse des émissions de journaux télévisés définit le document comme étant une alternance entre une séquence d'images montrant le présentateur et des séquences d'images associées à des reportages [132]. Les plans sont classifiés selon un modèle spécifique à chaque type de plan : début, présentateurs, publicités, météo et fin.

Si on considère la structuration et la présentation des films cinématographiques, ceux-ci ne disposent d'aucune information a priori sur leur présentation. Leur traitement devra donc être très général et pourra donc servir à tout autre type de documents vidéo.

Un film n'est pas simplement une suite quelconque d'images, il récite une histoire et donc il est possible d'y associer une structure de composition hiérarchique [85] [68]. En général, cette histoire est décomposée en plusieurs épisodes ou *séquences narratives* et chaque épisode est formé de plusieurs *scènes* ou actions. A son tour chaque scène regroupe une suite continue dans le temps de segments d'images ayant un rapport sémantique entre eux. Ces segments sont souvent connus sous le nom des *plans cinématographiques.*

Les différents niveaux physiques et sémantiques qui peuvent être distingués dans la structure d'un film sont :

- **Plan vidéo :** une suite d'images filmée sans interruption temporelle par la même caméra. Il représente l'unité physique de base pour la construction et la manipulation de la vidéo.

- **Image clé :** une image du plan vidéo qui peut représenter visuellement son contenu.

- **Scène vidéo :** est définie comme étant une collection de plans, adjacents dans le temps, ayant un lien sémantique entre eux et qui possèdent une unité d'action et de lieu.

- **Groupe vidéo :** une entité intermédiaire entre les plans physiques et les scènes sémantiques. Dans la cas d'une vidéo segmentée en régions on peut distinguer entre "groupe de plans" et "groupe d'objets".

- **Groupe de scènes :** une suite de scènes non successives où l'unité de lieu est conservée.

- **Séquence narrative :** une suite de scènes successives où l'unité d'action est conservée. Deux séquences narratives sont souvent séparées par un effet de transition de type *fondu au noir.*

La figure 2.1 illustre ces différents niveaux et les liens entre eux. Les niveaux *plan*, *scène* et *séquence* sont considérés comme des niveaux de base pour la production et l'analyse des films [68] [50]. Rui et al. [105] dans leur contribution à MPEG-7 proposent une structure similaire à celle de la figure 2.1. Au niveau "groupe vidéo" ils fabriquent des groupes de plans locaux (dans un intervalle de temps) pour ensuite construire les scènes. Donc, pour eux, c'est un niveau caché à l'utilisateur final de la vidéo structurée. Par contre, dans ce travail, et comme nous le verrons dans les chapitres suivants, les groupes d'objets (plans) sont des entités essentielles pour une vidéo structurée et interactive. Aussi, à partir de deux niveaux "groupe vidéo" et "scènes" on déduit un nouveau niveau "groupe de scènes" qui n'a pas les mêmes caractéristiques que le niveau "séquence narrative".

Un document vidéo structuré comportera à la fois des informations liées à la structure du document, comme la position des plans et les relations qui les unissent pour former les scènes, et des informations relatives au contenu des plans, comme la description des objets en présence et les relations de similarités.

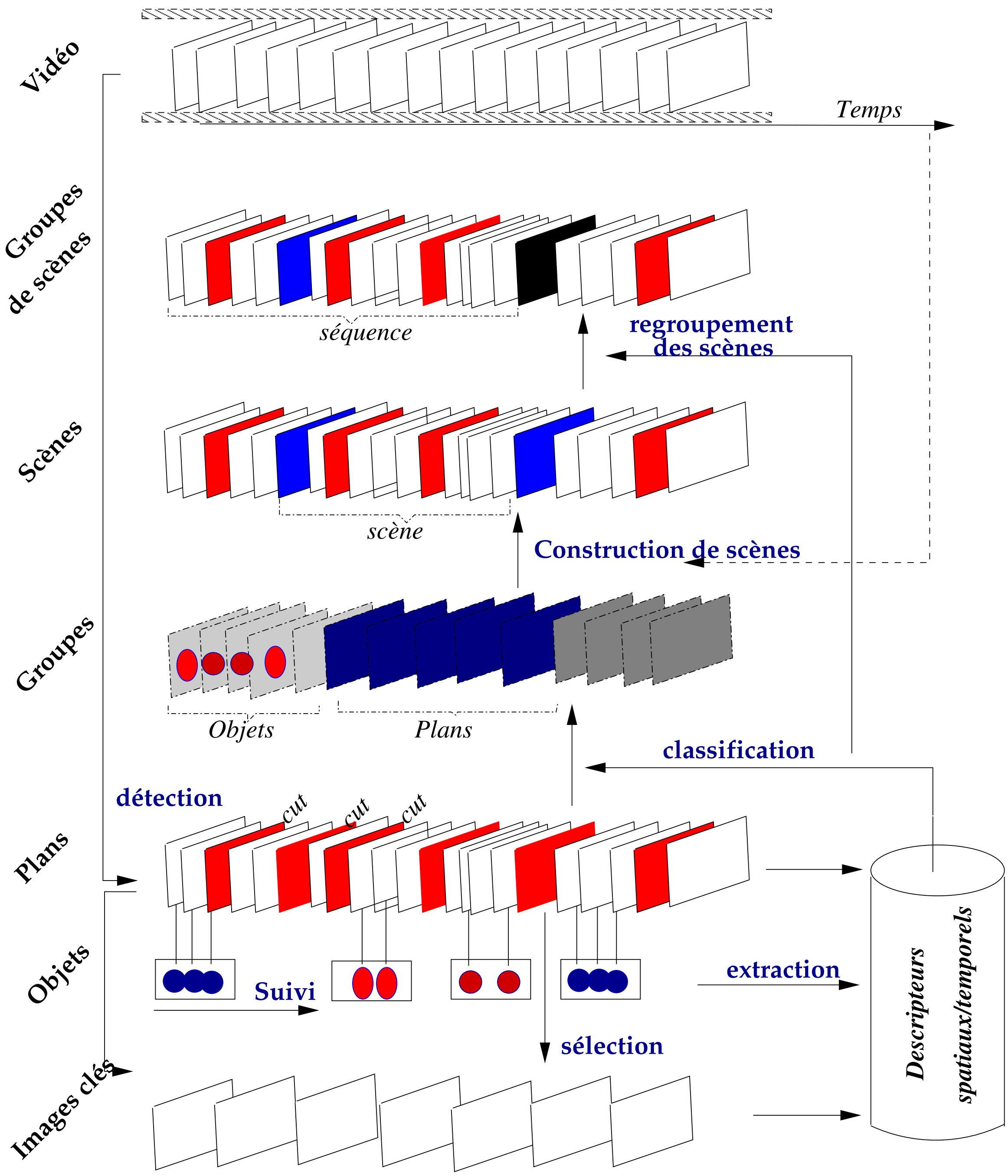

FIG. 2.1: *Une représentation hiérarchique de la structure de la vidéo*

2.2 Vers une structuration automatique

Définition 2.2.1 *Structuration : La structuration est un processus de segmentation spatio-temporelle des documents en segments de différentes unités et de construction des relations entre ces segments. En d'autre termes, c'est une opération d'abstraction dont le but est de trouver les différentes corrélations sémantiques au cours du temps.*

• **Aperçus des problèmes :** L'handicap majeur de l'utilisation de la vidéo est dû d'une part à sa taille importante et à son format purement séquentiel, et d'autre part à la difficulté d'en extraire une information sémantique à la différence du texte.

A l'heure actuelle la production et la diffusion de la vidéo numérique au niveau mondial est gigantesque. Par exemple, l'Inathèque de France[1] accumule environ 17.000 heures par an d'émissions télévisées [65] soit 85.000 heures depuis son ouverture (en Janvier 1995). Aussi, si on prend l'exemple important des bases d'images et vidéos constituées sur les serveurs web du monde entier, le volume du contenu accessible et la rapidité de son évolution sont considérables.

Du fait que la structure d'un film n'est pas fournie a priori, le seul moyen de le consulter est de se servir des outils classiques de navigation et de recherche des lecteurs vidéos : lecture arrière et avant accélérée, à vitesse normale ou ralentie, l'arrêt sur l'image et un ascenseur permet, à l'aide d'un curseur, un positionnement par accès direct aléatoire dans le document.

• **La structuration :** Pour permettre de gérer la taille importante de la vidéo et d'explorer son contenu d'une façon non séquentielle et efficace, la structure intrinsèque de la vidéo doit être identifiée. Comme nous l'avons mentionné dans l'introduction, par analogie avec les livres, la structure hiérarchique peut servir de table de matières pour la vidéo.

La structuration manuelle de la vidéo semble être inefficace : quantité immense de données, coût de la main d'oeuvre, subjectivité et nature monotone et fatigante des travaux. En conséquence, la recherche dans ce domaine est orientée depuis quelques années vers des méthodes automatiques de structuration et d'indexation par le contenu [12] [129] [53]. La figure 2.2 résume les points essentiels de la structure d'un système de vidéothèque adapté par Yeo et Yeung [129].

La suite de ce chapitre se focalise sur les méthodes dont nous nous sommes servi pour segmenter la vidéo en plans et pour localiser et suivre des objets à travers le temps.

2.3 Segmentation temporelle en plans

• **Principe :** La détection des ruptures de plans est basée sur la segmentation ou partitionnement de la bande vidéo en unités adjacentes simples. Un plan, défini par la succession de plusieurs images représentant une action spatiale et temporelle continue, est considéré comme l'unité de base de manipulation de la vidéo (figure 2.3). Le partitionnement de

1. département de l'Institut National de l'Audiovisuel (INA) chargé de la mise en oeuvre du dépôt légal de la radio-télévision française

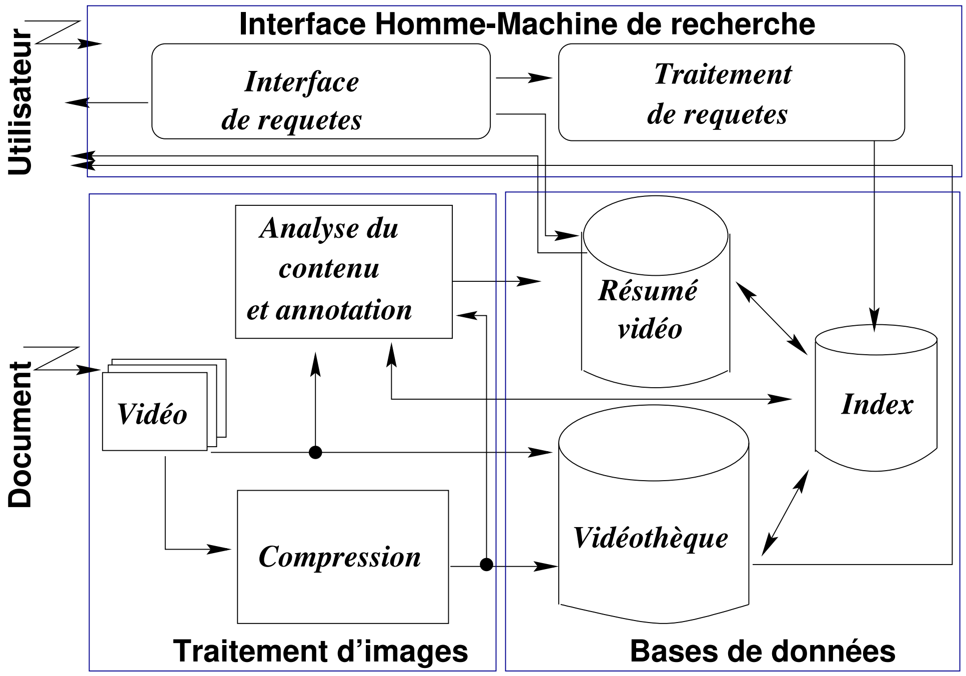

FIG. 2.2: *Schéma de la structure d'un système de vidéothèque.*

la vidéo est peut être vu comme un processus de recherche de rupture ou d'anomalie de continuité d'un certain critère quantitatif. Par exemple, la différence entre deux images consécutives constitue un critère quantitatif simple qui peut être utilisé pour détecter des ruptures brusques de plans.

• **Les méthodes existantes :** Actuellement, le problème de détection des ruptures de plan est bien maîtrisé et plusieurs méthodes efficaces ont été proposées dans la littérature [4] [122] [13] [14]. Ces méthodes peuvent être subdivisées en deux grandes catégories : celles qui cherchent à évaluer une *similarité* entre images successives et d'autres qui cherchent plutôt à évaluer si le contenu à $t+1$ est une suite *cohérente au contenu* de l'image t.

Parmi les méthodes de la première catégorie on peut distinguer entre les méthodes qui analysent la distribution des couleurs en utilisant des histogrammes globaux [89] ou locaux [72], les méthodes basées sur la comparaisons de primitives extraites des images comme le contour [131] et les foyers d'expansion [4] (section 2.3.2), et les méthodes qui analysent directement la bande compressée MPEG-1 [128] [24].

La deuxième catégorie regroupe les méthodes qui examinent la manière dont une information liée au mouvement estimé dans la séquence peut décrire "le degré de continuité" du contenu [14] (section 2.3.1).

Une comparaison quantitative des performances des méthodes existantes est difficile, chacun des travaux utilisant son propre jeu de tests. Une étude comparative récente est menée par [40]. Parmi les méthodes évaluées, les meilleures performances sont obtenues par [128], mais elles restent très moyennes, notamment en ce qui concerne la détection des

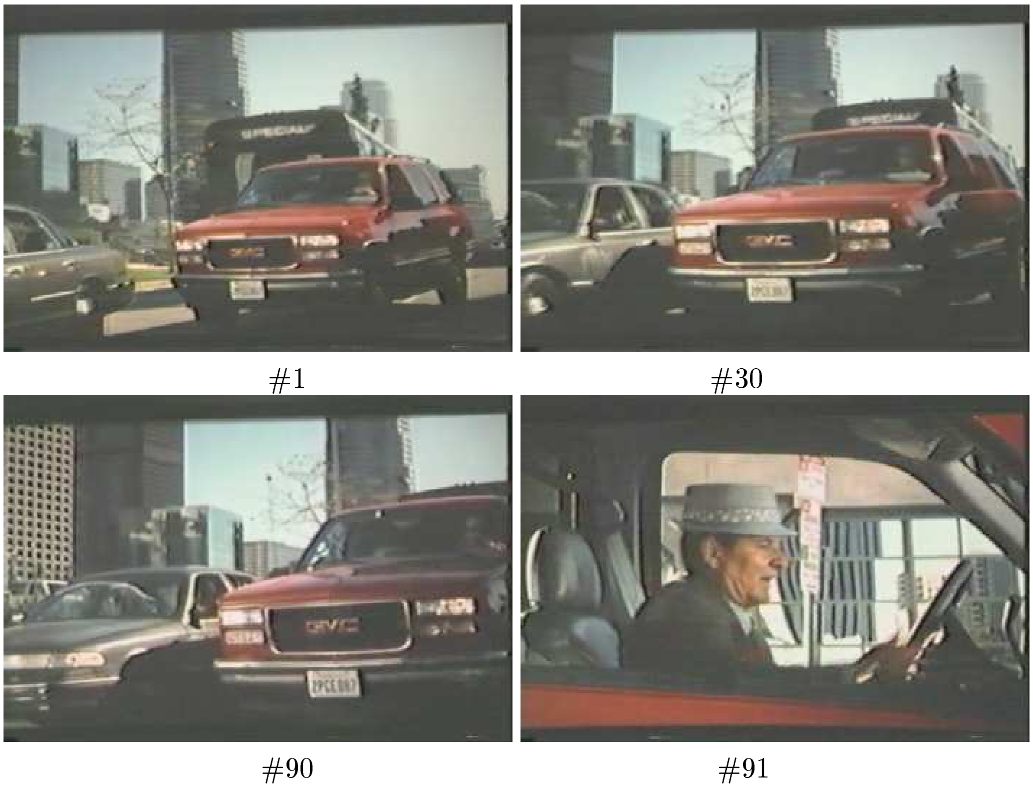

FIG. 2.3: *Exemple d'un changement de plan dans une séquence vidéo.*

transitions progressives (fondu enchaîné, fondu au noir, etc.).

• **Critères de choix d'une méthode :** Le choix d'une méthode de découpage de la vidéo en plans dépend des besoins de l'application en terme de rapidité de calcul et de précision des résultats. Les techniques assez simples mènent à des performances pouvant satisfaire un certain nombre d'applications importantes qui réclament surtout la rapidité de traitement. Le traitement de la représentation compressée MPEG est alors une option pertinente. D'autres applications peuvent être moins exigeantes en temps de calcul, mais nécessitent des résultats complets et précis. Par exemple, la qualité de la fabrication des scènes dépend de la pertinence des résultats de segmentation en plans et de classification de ces plans. Donc un facteur à prendre en compte aussi pour choisir une méthode de découpage en plans est la présence éventuelle d'autres étapes d'analyse dans le système, généralement en aval. En effet, si ces étapes utilisent le résultat du découpage en plans, il faut étudier son impact sur les modules qui en dépendent. D'autre part, il faut évaluer le coût calculatoire du découpage en plans, relativement au coût total de calcul d'un système d'analyse du contenu - et le coût de correction manuelle.

Dans la section suivante nous présentons brièvement la méthode de Bouthemy et al. [14] employée dans ce travail suite à une collaboration déjà mentionnée au début de ce chapitre. Les avantages et les inconvénients de cette méthode sont aussi rappelés. Ensuite, nous présentons la méthode de Ardebilian et al. [4] dont nous nous somme servis dans la segmentation en plans de la séquence “Danse avec les loups” pour un but de construction des scènes [52] [51] (chapitre 7).

2.3.1 Mesure de cohérence temporelle du contenu

Cette méthode proposée par Bouthemy et al. [14] est basée sur l'estimation robuste d'un modèle affine de mouvement dominant. Le modèle affine (équation 2.1) permet de se rendre compte d'une large variété de mouvements $2D$ (translation, rotation, divergence, etc.), dont l'estimation peut se faire d'une manière fiable.

$$\overrightarrow{w}_{\ominus}(p_i) = \begin{pmatrix} a_1 + a_2(x_i - x_g) + a_3(y_i - y_g) \\ a_4 + a_5(x_i - x_g) + a_6(y_i - y_g) \end{pmatrix} \tag{2.1}$$

où $\overrightarrow{w}_{\ominus}(p_i)$ représente le vecteur de déplacement à un point $p_i = (x_i, y_i)$ du point référence (centre de gravité du support). Les paramètres du modèle $\ominus = (a_1, ..., a_6)$ sont estimés par la technique de moindres-carrés pondérés itérés.

Après l'estimation d'un tel modèle pour une image donnée l'ensemble total des pixels dénoté par S est divisé en deux sous ensembles : pixels “conformes” (S_c) ou "non-conformes" au modèle de mouvement. Une comparaison des valeurs de w_i avec un certain seuil permet de juger si un pixel est conforme ou non au modèle. A partir de ça ils définissent une mesure de “cohérence inter-images”

$$\Im_t = \frac{n_c(t)}{n(t)} \tag{2.2}$$

où n et n_c représentent les cardinaux respectifs des ensembles S et S_c.

Cette mesure est ensuite exploitée pour détecter les changements de plans. Pour deux images successives appartenant au même plan, le support d'estimation est de taille importante. Par contre, dans le cas d'une paire d'images aux contenus très différents, le modèle de mouvement estimé ne sera pas capable d'expliquer la transformation entre les deux images, et donc la taille du support d'estimation sera faible. Le test cumulatif de Hinkley est utilisé plutôt qu'un simple seuillage sur les valeurs de $\Im_t$ afin d'améliorer la détection de changement de plan et de distinguer entre différents types de transition entre deux plans (cut, fondu, volet, etc.).

L'avantage de cette technique est qu'elle permet de différencier un changement de contenu dû à un mouvement de la caméra d'un changement induit par un effet de transition progressive (fondu par exemple). Par contre, le modèle affine $2D$ utilisé par cette technique peut être mis en défaut et le support du mouvement estimé est de taille faible, notamment en présence des grandes variations d'illuminations et lorsque le fond de la scène est constitué de plusieurs grands éléments subissant des mouvements différents, ou d'éléments non-rigides.

2.3.2 Méthode d'analyse de la scène 3D

En 1996, à la différence de toutes les méthodes statistiques de détections de plans citées ci-dessus, Ardebilian et al. proposent une méthode qui touche à la structure géométrique de l'image [4].

L'idée de base consiste à approximer l'image par des segments de droites. Cela suppose que l'image est segmentée préalablement en éléments pertinents susceptibles d'êtres approximés par des droites. L'image de contours est calculée premièrement à partir d'une carte de gradient. Ensuite, une approximation communément appelée *line fitting* des points de contours est réalisée à l'aide de la transformation *point-to-line* de Hough [61]. La transformée de Hough permet, en changeant d'espace de paramètres, de détecter des droites et d'en donner une équation. Elle est basée sur la dualité de l'équation d'une droite

$$x \sin\theta - y \cos\theta = \rho \tag{2.3}$$

En effet, une droite peut être définie soit par un couple (ρ, θ) soit par deux points de l'image, l'équation 2.3 permettant le passage d'une représentation à l'autre.

Les indices extraits d'une image sont représentés par les points d'intersections des droites obtenues après une deuxième application de la *transformée de Hough* sur la trace du contour de l'image. Le but de la deuxième transformée de Hough est de réduire au minimum le nombre des droites qui approximent l'image. Une détection de plan est ainsi signalée lorsqu'un changement significatif du nombre de ces indices 3D entre deux images successives est réalisé. Les transitions progressives ne sont cependant pas considérées par cette étude.

Un taux de plus de 95% d'identifications correctes des "cuts" est réclamé par les auteurs de cette méthode. Cependant, il parait bien que la méthode est sensible aux occultations et/ou apparitions des objets (mobiles) dans la scène qui conduisent à des variations majeures dans l'image de contours. Sur l'exemple de la figure 2.3 il est très probable d'avoir détecté un changement de plan avant l'image 90.

2.4 Suivi d'objets dans la vidéo

Le suivi d'objets quelconques dans une séquence vidéo réelle est une tâche très délicate, surtout quand ces objets sont non-rigides, le fond de la scène n'est pas fixe et dans le cas de plusieurs objets mobiles dans la même scène. Lorsqu'il s'agit de suivre des objets particuliers comme le visage, la main, le bras, etc. la tâche est plus simple car le modèle et les contraintes de variations de ces objets sont connus a priori [95] [82] [114]. D'autres critères de type géométrique et/ou statistique peuvent aussi être introduits dans l'identification des ces objets [16] [64].

Le suivi d'une primitive d'un instant t vers un instant $t+1$ est généralement basé sur une prédiction de la position de cette primitive recherchée à l'instant $t+1$. Dans ce mémoire on utilise le terme apparence pour désigner une occurrence de l'objet suivi à un instant t du plan. Dans le système conçu pour la construction de la vidéo interactive [8] [56] (chapitre 8), deux techniques de suivis d'objets *automatique* et *semi-automatique* sont intégrées. Ces techniques ont été proposées par l'équipe Vista [42].

2.4.1 Approche automatique

Nous rappelons ici les grandes lignes de cette approche. Pour plus de détails le lecteur peut s'adresser a la thèse de Gelgon [41] (première partie) ou bien à [42].

Cette approche peut être résumée en deux étapes principales :

1. La première étape consiste à chercher une partition de l'image au sens du mouvement. Dans un premier temps une partition spatiale de l'image est calculée. L'estimation de cette partition est effectuée à l'aide d'une méthode classique de segmentation d'images : ils supposent que la distribution de niveaux de gris des pixels de l'image est un mélange de lois gaussiennes pour lequel l'algorithme EM pourra être appliqué. Ensuite on applique une procédure de fusion des régions adjacentes dans l'image basée sur un critère de minimisation d'énergie entre ces régions. La fonction d'énergie est exprimée en fonction des descripteurs extraits des régions. Dans un deuxième temps, les régions spatiales (de la nouvelle partition ainsi obtenue) sont caractérisées par des modèles de mouvements 2D (affine à 6 paramètres). Cette caractérisation permet de définir une mesure de cohérence de mouvements entre deux régions spatiales. Cette mesure, prenant en compte la *continuité* du champ de mouvement à la frontière des deux régions, évalue la différence entre les deux champs de vecteurs de vitesse étendus au support formé par l'union des deux régions. Une étiquette de mouvement est associée à chaque région.

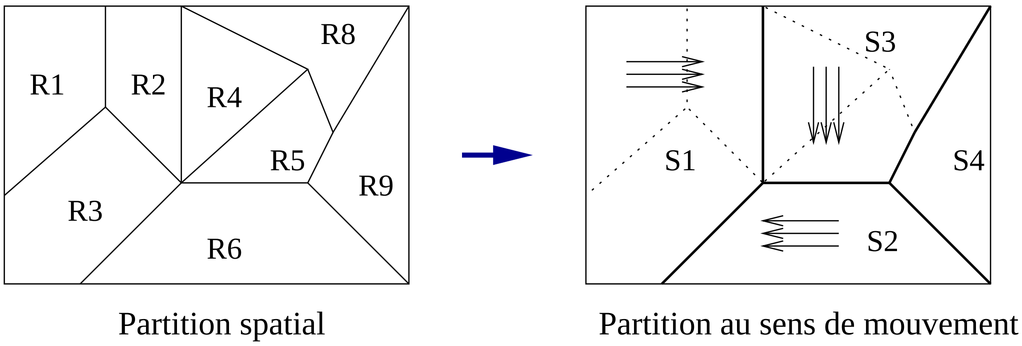

FIG. 2.4: *Partition au sens de mouvement*

2. La deuxième étape est basée sur un principe de propagation des étiquettes de deux phases : prédiction et mise à jour des configurations d'étiquettes définissant deux partitions à l'instant t et $t + 1$. La prédiction de la configuration d'étiquettes construite à l'instant t est utilisée comme configuration initiale d'étiquettes, relative à la segmentation spatiale, à l'instant $t + 1$. La mise à jour d'une configuration d'étiquettes revient à supprimer ou rajouter des régions à l'instant $t + 1$.

L'utilisation du mouvement estimé par l'union des régions spatiales suppose qu'un modèle affine soit à même de la décrire, ce qui n'est pas nécessairement vrai. Ceci peut mettre en défaut une telle approche. D'un autre coté la méthode est basée sur une segmentation de l'image en régions dont le nombre de régions et la qualité de la segmentation sont difficiles à être parfaitement estimés. Une illustration des résultats de cette approche est présentée dans la figure 2.5.

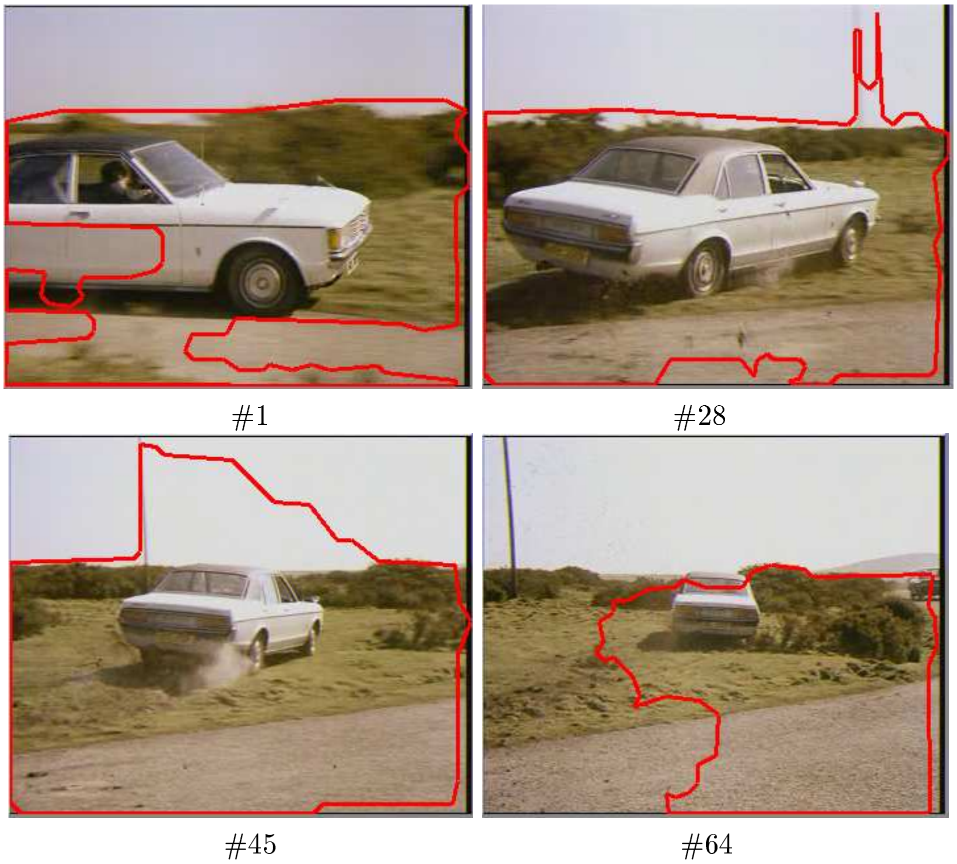

FIG. 2.5: *Résultat de suivi par l'approche automatique de [42] sur la séquence "Ford blanche" de* 66 *images.*

2.4.2 Approche semi-automatique

Lorsque le contenu de la scène est trop complexe pour l'approche automatique et quand un objet d'intérêt pour l'utilisateur n'est pas détecté automatiquement, une approche semi-automatique de suivi d'objets est avantageuse. L'approche possède une étape d'initialisation : la zone d'intérêt est définie manuellement dans une image du plan vidéo (suivi arrière, avant, et/ou dans les deux sens). Ensuite, pour chacun des couples d'images successives, la procédure suivante est alors effectuée. Entre les instants t et $t+1$, un modèle affine $2D$ du mouvement dominant sur cette zone est estimé (voir section 2.3.1). Les sommets de la zone sont projetés dans le sens de mouvement, et le polygone projeté définit la zone à $t+1$. Le nombre de sommets du polygone reste constant dans le temps. En cas de "décrochage du suivi" une alarme est fournie à l'utilisateur et de nouveau la zone d'intérêt est définie.

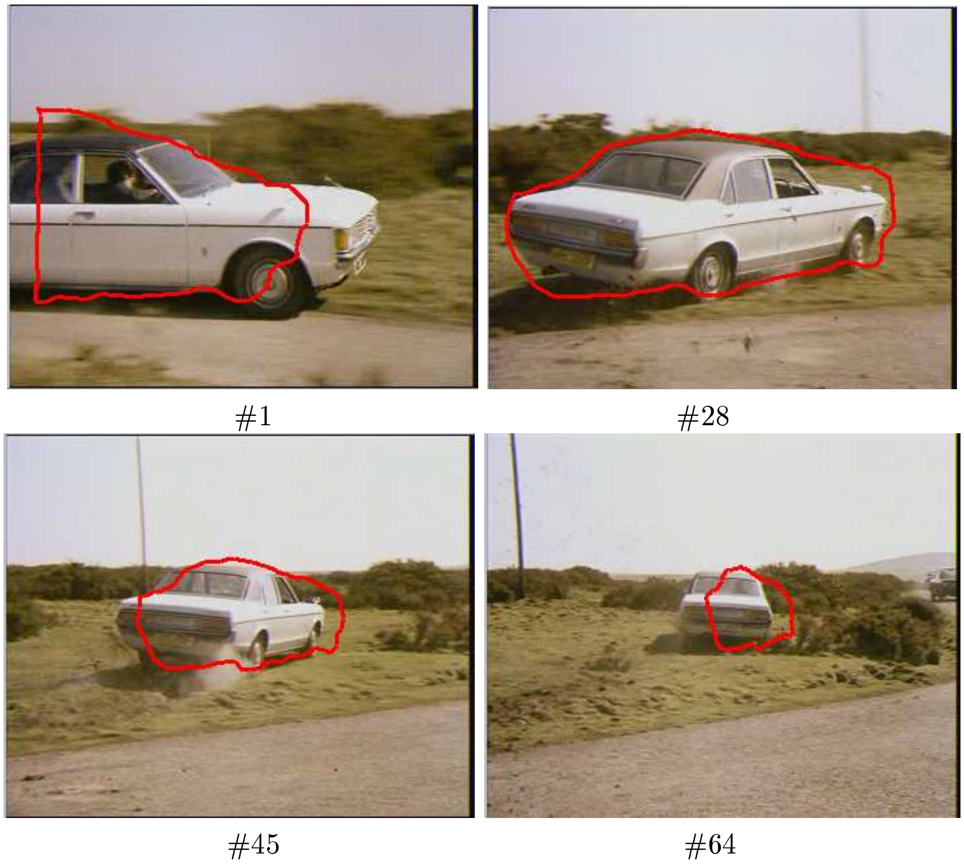

FIG. 2.6: *Résultat de suivi par l'approche semi-automatique de [41] sur la séquence d'images "Ford blanche". Le suivi de la voiture est initialisié manuellement à l'image* 28.

2.5 Corpus d'expérimentation

Les expérimentations que nous avons mené durant cette thèse pour valider les approches proposées dans les chapitres suivants, ont eu lieu sur différentes séquences du film "Chapeau Melon et Bottes de Cuir", *Avengers-1* et *Avengers-2* de la série télévision connue sous le nom *Avengers*. Ce film est fourni par l'Institut Nationale de l'Audiovisuelle (INA) comme un support standard d'évaluation des méthodes d'indexations de la vidéo en France.

Les séquences *Avengers-1* et *Avengers-2* comportent 1391 et 2749 apparences d'objets respectivement. Elles seront détaillées dans les sections 4.4 et 5.4. Ces séquences de test ont des durées assez limitées. Cette limitation est le fait de deux handicaps majeurs :

- Les outils de suivi d'objets que nous venons de présenter ne fournissent pas une segmentation parfaite des objets (voir figures 2.5 et 2.6), donc une correction manuelle du suivi et/ou un suivi manuel des objets dans les plans sont les moyens pour résoudre le problème fondamental étudié dans le contexte de cette thèse : "la variabilité de l'apparence intra-plan des objets suivis". Le coût de cette fastidieuse correction a donc limité notre base expérimentale. Cette correction manuelle reste cependant imprécise pour nous placer dans le cas réaliste d'un futur suivi de bonne qualité, mais qui sera toujours loin d'être parfait.

- Les données visuelles (images, objets segmentés) demandent un espace disque gigantesque (environ 1.04 Giga octets pour une 1 minute de vidéo MPEG) : Les outils de préparation de la vidéo utilisés dans ce travail (segmentation en plans, suivi des objets, caractérisation de bas niveaux des images/objets) exploitent des séquences MPEG décompressés aux formats PPM [2] de haute qualité; les résultats de suivis d'objets sont aussi stockés aux formats PPM; à chaque image segmentée une image de masque lui correspond. Aussi, quelque descripteurs utilisés dans notre système comme les invariants différentiels proposés par Schmid et Mohr [110], sont extraits des images de format PGM (*portable gray map*) de niveaux de gris. Donc, pour une séquence vidéo MPEG de durée d'une minute (équivalent à 1500 images PPM), il faut un espace disque d'environ 1 Giga octets [3], si on considère qu'un objet au moins est localisé dans chacune des images décompressées et sans compter la taille des imagettes des objets segmentés (format PPM et GIF), ni la taille des descripteurs de bas niveaux extraits des objets segmentés. La conception d'une architecture travaillant sur des données compressées est donc à l'ordre du jour, mais hors du cadre de ce travail.

Par contre, ce corpus expérimental est suffisant pour valider nos approches dans des conditions réelles de changements d'images et d'apparences d'objets très variables.

2.6 MPEG-7 : interface pour la description du contenu multimédia

La future norme MPEG-7, annoncée pour septembre 2001, porte essentiellement sur la standardisation des descripteurs associés au contenu des documents audiovisuels [87]. Cette norme associe à un document audiovisuel des descripteurs, à la fois de bas niveau (décrivant les formes, couleurs, ...) et des descripteurs sémantiques pour lesquels l'intervention humaine est indispensable. Par contre, MPEG-7 ne s'intéresse ni à la définition des algorithmes d'extraction des descripteurs, ni à la manière avec laquelle on doit construire un moteur de recherche capable d'exploiter les descripteurs. La figure 2.7 illustre le domaine d'intérêt de MPEG-7.

Plusieurs groupes travaillent aujourd'hui autour de la standardisation de MPEG-7; leurs efforts se focalisent essentiellement sur les trois points suivants :

1. Normalisation d'un ensemble de descripteurs (D) où chaque descripteur sert à représenter un aspect particulier du document audiovisuel (effets sonores, couleur, texture, mouvement, suivis d'objets, etc.).

2. Portable pixel map, image 352x288, de taille 304143 octets.
3. $[(1500 \times 304143) \times 2 + (1500 \times 101391)]/(2 \times 1024) = 1039565.918$ octets.

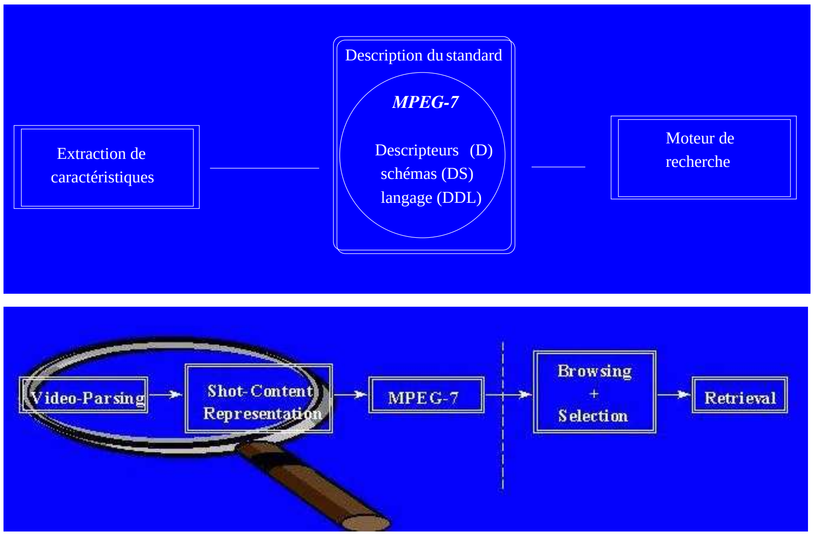

FIG. 2.7: *Le champ d'action du groupe MPEG-7.*

2. Collection de schémas de description (DS) où chaque schéma de description spécifie la structure et la sémantique des rapports entres les éléments qui composent le schéma.

3. Définition d'un langage de description (DDL) qui permet de mettre à jour les schémas et l'ensemble de descripteurs.

Une description MPEG-7 est formée par un schéma (DS) ainsi que par des instances de chaque descripteur et de chaque schéma définis dans le schéma en question (il est possible d'inclure des schémas dans un autre schéma).

MPEG-7 n'est pas encore un standard et donc l'information autour de lui n'est pas stable. Mais cet effort de normalisation semble essentiel, dans la mesure où les informations sont géographiquement très distribuées, et que de nombreux prototypes industriels d'indexation concurrents sont en cours de développement.

2.7 Récapitulatif du chapitre

Dans un premier temps, nous avons rappelé la structure hiérarchique d'un film vidéo, tout en proposant une extension de cette hiérarchie en définissant deux niveaux supplémentaires : *groupes d'objets* à travers la vidéo et *groupes de scènes* non successives dans le temps. Une telle extension pourra être suggérée pour le prochain standard MPEG-7. La construction du niveau *groupes d'objets* est discutée dans les chapitres 4 et 5. Les niveaux *images clés* et *scènes* feront l'objet des chapitres 6 et 7 respectivement.

Dans un second temps, nous avons présenté les outils de base utilisés durant cette thèse pour segmenter la vidéo en plans et pour suivre automatiquement ou semi-automatiquement les objets dans les plans. Les limites de ces outils sont présentées ainsi que leurs influences sur la démarche expérimentale de nos approches proposées.