Chapitre 7

Construction des scènes pour un film vidéo

Je vais me donner de la marge
côté délais.

Les images successives d'un film prises par la même caméra et sans interruption temporelle sont regroupées dans des *plans vidéo*. A un niveau plus sémantique dans la structure hiérarchique des films (figure 2.1) les plans successifs décrivant une continuité d'action et de lieu peuvent être regroupés dans des *scènes*. Comme nous l'avons déjà mentionné dans l'introduction de ce mémoire, ce niveau permet à l'utilisateur final de la vidéo hyperliée d'explorer un film comme un livre avec une table de matières.

Au contraire de l'identification automatique des ruptures des plans basée sur une simple comparaison des couples des images successives, l'identification des scènes est une tâche très délicate. Ceci est dû d'une part à l'inexistence d'une formulation précise de la rupture entre deux scènes (surtout lorsque l'image est analysée seulement) et d'autre part à la difficulté d'une caractérisation fiable du contenu des plans et de la mise en oeuvre d'une stratégie de classification automatique des plans.

Ce chapitre se focalise sur la construction de scènes pour un film vidéo. D'abord on rappelle les étapes fondamentales de l'approche de macro-segmentation que nous avons proposée lors d'un travail de stage de DEA [51] [53] [52] [25]. Ensuite nous résumons l'extension de cette approche élaborée durant cette thèse [54]. Cette extension porte essentiellement sur la manière de comparer deux plans vidéo sur la base de plusieurs descripteurs.

7.1 Deux étapes pour la macro-segmentation en scènes

La motivation principale de la macro-segmentation est dûe au fait qu'un film vidéo d'une heure par exemple peut contenir plus que 1000 plans, ce qui est difficile à représenter graphiquement et à explorer linéairement par l'utilisateur. Le découpage d'un film vidéo en des unités plus macroscopiques que les plans, mais aussi sémantiques, sera appelé découpage en *scènes*, avec la définition imprécise suivante : une scène est un ensemble de plans adjacents qui comporte une certaine unité de lieu d'acteur, d'action et/ou du son.

La 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, à condition que chaque partie formée soit continue dans le temps – une scène –.

De là, une première étape de classification des plans similaires dans des blocs et une seconde étape de fusionnement des blocs qui se chevauchent dans le temps jusqu'à la vérification de la continuité temporelle, permettent de générer des scènes. Ces deux étapes ont été adoptées dans l'approche que nous avons proposée dans un travail antérieur sur la macro-segmentation. Nous décrivons brièvement ces deux étapes ci-dessous et pour plus de détail le lecteur pourra se reporter aux références [53] [25].

Procédure Macro-segmentation, deux étapes :

1) *Fabrication des blocs* ou de classes de plans. Les plans sont représentés par les images médianes, les histogrammes de couleurs sont extraits de ces images et un algorithme adaptatif de classification est employé pour les regrouper en blocs. A chaque itération de cet algorithme un plan non étiqueté constitue un nouveau bloc ou bien il est classé dans un bloc déjà existant selon une distance de similarité et un seuil fixé a priori. Un algorithme similaire a été employé plus tard par Zhuang [134] pour la sélection des images clés.

2) *Construction de scènes* par fusion de blocs en relation. A partir de la répartition des plans des différents blocs sur l'axe de temps, quatre relations temporelles sont générées entres ces blocs : *Before*, *Meets*, *Overlaps* et *During*. Une relation *Before* décrit une transition entre deux séquences narratives et elle est générée entre deux blocs dont le dernier plan du premier bloc (ordre chronologique) est lié par un effet de transition progressive avec le premier plan du deuxième bloc. Par contre si le raccord entre ces deux plans est de type "cut" alors une relation de type *Meets* est générée. Deux scènes d'une même séquence narrative sont reliées par une relation de type *Meets*. Lorsque deux blocs se chevauchent dans le temps une relation temporelle de type *Overlaps* ou *During* est générée. Deux blocs d'une même scène sont reliées par une relation de type *Overlaps* ou *During*. La figure 7.1 illustre un graphe temporel de blocs pour les 40 premiers plans de la séquence "Dances with Wolves". La construction des scènes se fait donc par fusions successives des blocs reliés par des relations de type *Overlaps* et *During* jusqu'à la vérification d'un critère de continuité temporelle et la détection des relations *Meets* qui déconnecte le graphe temporel de blocs ainsi construit.

Cette approche a donné des résultats satisfaisants sur la séquence "Dances with wolves" comportant environ 10000 images et segmentée en 70 plans par la méthode de [4] décrite dans la section 2.3.2. Les figures B.1 et B.2 (annexe) illustrent les images-médianes de ces plans. Cinq scènes sont identifiées correctement : première scène du plan 1 à plan 27,

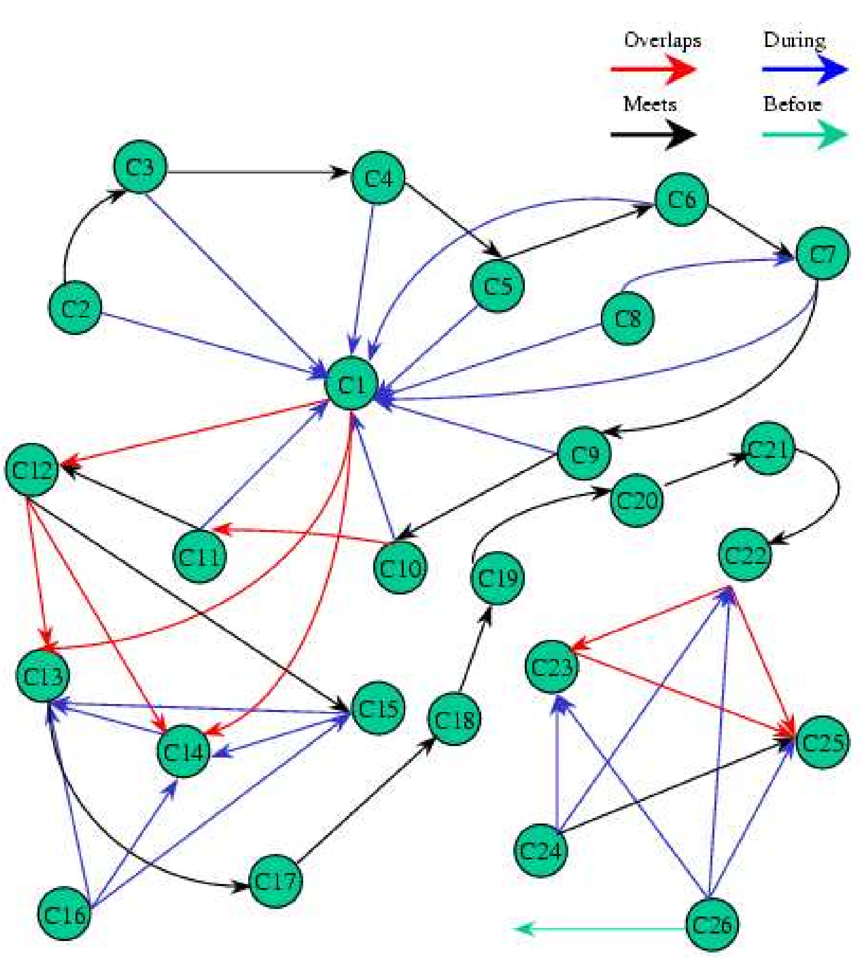


FIG. 7.1: *Exemple d'un graphe temporel de blocs généré sur les* 40 *premiers plans du film "Dances with wolves"; chaque noeud représente un bloc de plans similaires; la similarité est calculée sur la base des histogrammes des images médianes des plans.*

deuxième scène comporte le plan 30 seulement, troisième scène du plan 31 à plan 47, quatrième scène comporte le plan 52 seulement et cinquième scène est composée du plan 53 à plan 67. Les plans restants 28, 29, 31, 32, 48, 50, 51, 68, 69 et 70 sont considérés comme des scènes indépendantes de taille 1. Celles-ci représentent les fausses détections des scènes produites par l'approche.

Dans cette approche seule l'unité de lieu (histogramme global de couleurs) a été considérée dans la fabrication de blocs de plans. Cette restriction peut expliquer le fait d'obtenir des scènes de taille 1. Autrement dit, l'unité de lieu n'est pas toujours capable d'identifier deux plans de la même scène. D'autres indices de comparaisons des plans doivent être pris en considération. Dans ce sens nous discutons l'extension de cette approche dans la section suivante.

Dans le même contexte de travail, Yeung et al. [130] ont adoptés une technique similaire à celle décrite ci-dessus. En se basant sur nos travaux, Mahdi, Chen et Ardebilian [81] [80] ont proposés plusieurs extensions en introduisant la notion du *rythme* des plans successifs et en rajoutant une phase d'identification des plans d'intérieurs et d'extérieurs. Une autre approche de macro-segmentation basée sur des règles déduites des effets de montage a été proposée par Aigrain et Joly [1]. Ils utilisent 10 règles d'identification des ruptures de séquences de scènes dans un film. Mais cette approche paraît très restrictive compte tenu du foisonnement des indices de ruptures de scène qui se trouvent dans différents films. Par exemple cette méthode segmenterait le film "Titanic" (2h50) en trois scènes dont les limites correspondent aux effets de montage réalisés lors des transitions où Rose (Kate Winslett) commence une phase de la narration de son histoire. Enfin, une approche très récente proposée par Sundaram et al. [118] consiste à coupler la bande image et la bande sonore semble être bien adaptée à ce problème d'identification de scènes.

7.2 Résumé de "A mixed classification approach of shots for constructing scene structure for movie films" – IMVIP'2000

Cet article, publié à IMVIP'2000, présente une approche mixte de macro-segmentation de la vidéo en scènes. L'idée ici est d'étudier la similarité entre deux plans vidéo sur la base de plusieurs unités ou descripteurs de bas niveaux. Le problème issu de cette extension de l'approche de base que nous avons présentée dans la section précédente est la fusion de ces descripteurs. Cette phase de fusion des descripteurs devra produire une partition unique et fiable de l'ensemble de tous les plans d'un film vidéo et permet ensuite d'appliquer la deuxième étape de "construction de scènes".

Le contenu visuel d'un plan revêt une masse importante d'informations qui permet normalement de le distinguer ou de l'associer à d'autres plans. Plusieurs autres descripteurs que l'histogramme de couleurs peuvent être utilisés pour caractériser un plan. C'est le cas par exemple des objets identifiables à l'intérieur de chaque plan, de la position relative de ces objets dans chaque plan, des indices indiquant le décor où les images ont été tournées, etc.

Dans notre implémentation, seules l'unité de lieu représentée par l'histogramme et l'autocorrélogramme de couleurs et l'unité de contenu du plan représentée par les objets localisés dans les plans ont été utilisées. L'approche est générale et d'autres descripteurs peuvent être testés. La similarité entre deux plans vidéo sera donc calculée ici sur la base des histogrammes de couleurs, des autocorrélogrammes (voir section 3.2.2.1) et du nombre d'objets similaires contenus dans les plans. Dans ces expérimentations, les objets sont caractérisés par des autocorrélogrammes. Un seuil de similarité a été fixé a priori pour chaque type de distances entres descripteurs. Pour montrer l'importance de cette caractérisation mixte du plan vidéo, considérons l'exemple des plans de la figure 7.2. Chaque plan est représenté par son image médiane. Certaines images contenants d'objets d'intérêts ont été segmentées manuellement. Les plans des couples (7, 9), (33, 50) et (69, 70) appartiennent à une même scène du film "Dances with Wolves" (voir figure B.1). Les deux plans 7 et 9 sont détectés similaires sur la base des histogrammes et des autocorrélogrammes et différents avec le descripteurs de contenu. Le plan 9 ne contient pas des objets d'intérêts. Par contre, les deux plans 33 et 50 sont reliés sur la base de leurs contenus et non pas par les descripteurs globaux. Enfin, les deux derniers plans de la figure 7.2 ont été appariés correctement par les histogrammes de couleurs, mais échec avec les descripteurs de contenu et les autocorrélogrammes.

L'exemple cité plus haut montre que chaque descripteur permet de comparer les plans suivant une caractéristique particulière. Les histogrammes de couleurs tiennent compte de la distribution globales de couleurs, les corrélogrammes tiennent compte des informations locales dans l'image et des environnements de tournage, le contenu des plans permettent de comparer les plans suivant le contenu.

En traitement d'images la comparaison de deux images sur plusieurs points de vue est un problème difficile. Une solution souvent adopté pour ce problème consiste à déduire une distance unique par combinaison des distances des différents attributs. Bisson [11] a proposé une mesure de similarité entre objets dans des modèles à objet qui tiennent compte aussi bien des attributs de chaque objet que des relations qui lient les objets à comparer. Valtchev [123] utilise une distance dérivée de la distance de Mahalanobis pour la construction automatique des taxonomies pour les langages à objets. La distance entre deux images est considérée comme la somme des écarts obtenus sur les attributs et les relations. Jain et al. proposent d'utiliser le barycentre des distances pour la combinaison de la couleur et de la forme lors de la comparaison des images [67].

La méthode de Jain et al. est très simple à mettre en oeuvre mais est-il possible de trouver une pondération entre des entités qui n'ont rien en commun ? Chaque mesure de similarité illustre un aspect particulier de la vidéo, des aspects indépendants les uns des autres. Une esquisse de cette approche a été utilisée dans un premier temps pour la classification des objets dans la vidéo, avec résultats nullement satisfaisants et très sensibles à la variation des coefficients de pondération des distances [35].

Une combinaison linéaire simple des distances entre les plans ne peut donc pas être utilisée pour effectuer une classification automatique des plans. C'est dans cet optique qu'une méthode de construction des blocs par fusion, non pas des distances suivant chaque descripteur, mais des blocs de plans obtenus à partir de chaque descripteur a été proposée. Dans un premier temps nous avons défini une nouvelle distance entre les blocs basée sur l'intersection des ensembles de blocs de plans fabriqués pour chaque descripteur de la vidéo.

plan 69 plan 70

Fig. 7.2: *Similarité entre deux plans caractérisés par trois descripteurs : histogramme de couleurs, autocorrélogramme, et le contenu. Une distance pour chaque type de descripteur est calculée entre les plans (7, 9), (33, 50) et (69, 70). Les plans de chacun de ces couples appartiennent à une même scène du film "Dances with Wolves" (voir figure B.1). Les résultats d'appariements par ces trois descripteurs sont les suivants : les plans* 7 *et* 9 *sont identifiés sur la base des histogrammes et des autocorrélogrammes; les plans* 33 *et* 50 *sont correctement appariés sur la base de descripteurs de contenu seulement; et la similarité des plans* 69 *et* 70 *est détectée seulement sur la base des histogrammes de couleurs.*

Dans un second temps, nous avons proposé un algorithme de fusion de deux étapes : une étape de fusion verticale qui permet la comparaison et la fusion des ensembles construits à partir des descripteurs différents, et une étape de fusion horizontale qui permet un affinage des résultats obtenus à partir de la première étape. En effet, les ensembles de blocs obtenus après la première étape ne sont pas disjoints et l'étape d'affinage les corrige.

A la première étape, le seuil de similarité entre deux plans est choisi empiriquement mais il est strict. Ce seuil strict implique une forte similarité entre les plans regroupés suivant un descripteur donnée. A titre d'exemple, ce seuil est fixé à 0.04 pour les histogrammes de couleurs alors que dans [53] on utilise un seuil de 0.09 sur la séquence "Dances with Wolves". Ce seuil strict donne l'assurance de rejeter les mauvais appariements suivant un descripteur donnée. Les rejets occasionnés par un descripteur donné pourront être corrigés par les autres descripteurs lors de l'opération de fusion.

La construction des blocs de plans à l'aide de cet algorithme de fusion présente l'avantage de tirer le maximum d'informations de chaque descripteur lors de la comparaison de plans. Cette méthode nous a permis de générer des blocs sur des séquences vidéo malgré une forte variabilité dans le type de leur contenus.

L'application de cette approche sur la séquence "Dances with Wolves" a donné des résultats parfaits : les sept scènes de la séquence sont correctement identifiées (voir annexe B). Sur d'autres séquences du film "Avengers", dont leurs contenus sont très complexes, les résultats obtenus sont satisfaisants et nettement meilleurs que lorsqu'un seul descripteurs est utilisé.

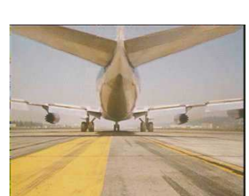
La figure 7.3 illustre les cinq scènes de la séquence Avengers obtenues par l'approche mixte. L'évaluation précise de la qualité de cette segmentation est une tâche délicate car la séparation entre les scènes de la séquence traitée n'est pas bien définie. Une segmentation manuelle de cette séquence pourrait envisager de mettre le dernier plan de la première scène avec la deuxième ou bien de combiner la deuxième scène avec la troisième.

scène composée des plans [1..8] : “l’espion tire sur le prêtre”.

scène composée des plans [9..12] : “l’espion fuit et monte dans sa voiture”.

scène composée des plans [13..15] : “poursuite de deux voitures”.

scène composée des plans [16..23] : “discussion entre Steed, Prêtre et Purdey”.



scène composée de plan 24 : “décollage de l’avion”.

FIG. 7.3: *Les résultats de macrosegmentation de Avengers par l’approche mixte.*