Chapitre 9

Conclusion générale

Saluer ce qui va pour mieux dire
ce qui ne va pas.

Dans le contexte du prochain standard proclamé MPEG-7, nous avons présenté un travail qui touche différents problèmes liés à l'indexation de la vidéo par le contenu. Nos contributions fondamentales ont porté sur la caractérisation des objets suivis par des modèles statistiques adaptés à leurs changements d'apparences intra-plan, la classification automatique et semi-automatique des objets suivis basée sur les modèles estimés, la représentation optimale d'un objet suivi (apparences-clés), et enfin sur la classification des plans adjacents en des groupes sémantiques (macro-segmentation en scènes). L'objectif de telles tâches automatiques est d'optimiser les coûts d'interaction entre l'utilisateur et la machine lors de la construction et la présentation d'un document vidéo afin de satisfaire une partie des besoins de MPEG-7 ou de permettre une interactivité Homme/Vidéo.

9.1 Bilan de travail

Le problème principal à résoudre pour associer des entités dans une vidéo est la grande variabilité de leurs apparences. Cette variabilité intra-plan est due principalement aux changements d'apparence de l'objet au cours du temps et au manque de robustesse des descripteurs classiques à ces changements de l'image, sans compter les occultations qui ne sont pas réellement modélisables dans ce contexte.

Nous nous sommes penchés d'abord sur ce problème de la variabilité intra-plan d'un objet suivi dans l'espace de descripteurs de couleurs. Ceci conduit à une représentation spatiale dans l'espace de descripteurs qui n'est pas compacte et souvent multimodale. Pour modéliser cet ensemble d'aspects, nous avons retenu un modèle statistique : une distribution représentée par un mélange de lois gaussiennes. L'estimation de cette distribution est vue comme un problème de recherche des classes d'apparences intra-plan d'un objet suivi. Le mélange gaussien multivariée dans le cas automatique possède des propriétés intéressantes : on peut montrer qu'il peut bien approcher des distributions complexes ; de plus il existe maintenant des outils qui permettent de sélectionner automatiquement le modèle gaussien le mieux adapté aux données et le nombre convenable de classes d'apparences intra-plan, structure qui diffère d'un objet suivi à un autre.

Les expérimentations de cette approche montrent que les distributions des objets suivis correspondent bien à des mélanges de loi gaussiennes, et que le choix automatique de la structure du mélange est bien adapté aux données mais par contre ce choix n'est pas toujours optimal.

L'intérêt de cette approche est double : (1) La caractérisation des objets suivis par des modèles de mélange gaussien permet de faire une mise en correspondance efficace entre eux qui tient compte des changements intra-plan d'apparence des objets. (2) L'indexation des objets suivis d'une large séquence vidéo sur la base des paramètres gaussiens est maintenant faisable vu la taille raisonnable de ces descripteurs statistiques et le nombre limité des classes d'apparences retenues pour une séquence vidéo.

Cette approche de modélisation de la variabilité nous a permis d'aborder ensuite le problème de classification des objets suivis. La classification des apparences d'objets suivis à travers la vidéo dans des groupes homogènes créee des liens entre les objets "identiques" du film. Ces liens peuvent être explorés par l'utilisateur final de la vidéo hyperliée.

Dans un premier temps, nous avons proposé une approche de classification semi automatique des objets suivis. L'auteur de la vidéo hyperliée choisit d'une manière interactive la liste des modèles d'objets suivis pour une séquence vidéo. Les mélanges gaussiens estimés pour ces modèles d'objets sont liés ensuite par une loi globale de mélange gaussien. Nous avons proposé d'utiliser le maximum a posteriori (MAP) pour classer toute apparence dans le film dans l'un des modèles d'objets. En intégrant l'aspect temporel une deuxième fois lors du classement des apparences, nous avons proposé d'utiliser une méthode robuste, le vote majoritaire, pour classer toutes les apparences d'un objet suivi dans le même modèle d'objet suivi.

L'étude comparative menée dans ce chapitre a montré une amélioration considérable allant jusqu'au 35% des pourcentages de bon classement des apparences d'objets suivis par l'approche proposée vis-à-vis de la méthode classique de "moyenne temporelle de descripteurs". La méthode de classement robuste donne des résultats meilleurs de l'ordre de 10% que la méthode de MAP. La séquence vidéo Avengers-1 utilisée dans ces expériences contient des modèles d'objets suivis de nombre d'apparences très limités. Cette situation explique la dégradation des résultats lorsque le nombre de composantes gaussiennes dépasse 2. On s'aperçoit dans ce cas que les critères BIC et ICL ont tendance à choisir un nombre de classes d'apparences élevé mais avec des modèles gaussiens linéaires. Les mauvais résultats de classement sont obtenus quand le critère d'entropie NEC est employé.

La performance de l'approche est la meilleure quand la modélisation est effectuée dans l'espace réduit des histogrammes de couleurs RGB. Par contre elle est la plus mauvaise dans l'espace réduit des histogrammes de couleurs normalisés rgb. La normalisation des couleurs conduit théoriquement à une représentation invariante aux changements de luminosité mais elle conduit aussi à une réduction drastique de l'information; il en est de même pour la réduction de l'espace des histogrammes normalisés par une analyse en composantes principales; tout cela explique la dégradation des résultats, en particulier dans ces cas on observe un chevauchement important entre les modèles gaussiens des différents modèles d'objets suivis.

Dans un second temps, nous avons proposé une automatisation de l'approche de classification des objets suivis. La classification ascendante hiérarchique est appliquée sur la matrice de distances entre les objets suivis. La distance proposée entre deux objets suivis est la distance minimale de Kullback (ou de Bhattacharyya) entre leurs classes d'apparences intra-plan. Des résultats satisfaisants de l'ordre de 80% ont été obtenus dans ce cas automatique et sur une base d'objets vidéo très complexe. Il ressort de cette validation expérimentale que les résultats sont les meilleurs lorsque la distance de Kullback entre deux classes d'apparences et la distance de "saut maximum" entre deux groupes d'objets suivis dans la hiérarchie construite sont employées. Une sélection interactive du nombre de classes d'objets ainsi que des outils interactifs pour éditer les résultats de la classification ont été proposés dans un système de vidéothèque hyperliée.

Nous nous intéressons ensuite au problème de l'extraction automatique des apparences-clés d'un objet suivi. L'extraction des apparences-clés permet une visualisation rapide du contenu d'un plan et un appariement de deux objets suivis (deux plans) sur la base des apparences-clés. Pour résoudre ce problème, nous proposons une technique de classification automatique, le mélange gaussien, pour regrouper les apparences similaires dans des classes d'apparences intra-plan. Ensuite, nous effectuons une sélection des apparences médianes de ces classes. Et enfin nous ne gardons que les apparences-clés les plus représentatives qui vérifient deux tests de dissimilarité spatiale et temporelle. L'avantage de cette approche est qu'elle permet une sélection automatique et optimale des apparences les plus représentatives de l'objet suivi. Les expérimentations ont montré l'efficacité et la facilité de la mise en oeuvre d'une telle approche.

Afin de compléter la construction automatique de la structure de haut niveau d'un film vidéo, nous avons abordé à la fin de cette thèse le problème d'identification des scènes. Ce problème est très délicat parce que d'une part il n'existe pas une formulation précise de la rupture entre deux scènes et d'autre part il est lié aux problèmes de caractérisation fiable du contenu des plans et de classification automatique des plans. Nous avons étendu un travail antérieur sur la macro-segmentation en proposant d'apparier deux plans vidéo sur la base de plusieurs descripteurs de bas niveau : le contenu du plan, la distribution globale des couleurs et la distribution spatiale des couleurs extraites 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. Il ressort à ce point un problème difficile de fusion de descripteurs hétérogènes à résoudre. Nous avons proposé un algorithme de fusion des groupes de plans des différents descripteurs plutôt que les descripteurs eux mêmes. Une amélioration nette des résultats de l'approche de base est réalisée sur plusieurs séquences vidéo. Lors de la construction des groupes de plans selon un descripteur donné, un seuil de dissimilarité spatiale et un autre de voisinage temporel sont fixés a priori. Il est normal que les résultats dépendent de ces deux paramètres de seuillage qui ne sont pas facile à choisir dans certains cas.

Une fois que nous avons défini les structures de bas et de haut niveau d'un film vidéo, nous avons développé deux prototypes (VideoPrep et VideoClic) qui facilitent les interactions Homme/Machine et Machine/Vidéo, lors de la construction et l'utilisation d'une vidéo interactive.

9.2 Perspectives de recherche

Les points suivants peuvent être améliorés d'une façon immédiate :

Classification robuste des apparences intra-plan Lors du suivi d'un objet à travers le temps une certaine continuité dans l'apparence de l'objet est généralement conservée. Or quelques-unes de ces apparences sont parfois extrêmes. Celles-ci résultent par exemple des occultations partielles significatives (voir les dernières apparences de la figure 3.2). La représentation de ces apparences dans l'espace de descripteurs correspond à des points atypiques (*outliers*). Lors de l'étape de la modélisation de la variabilité intra-plan, on constate que de tels points tendent à former des classes d'apparences comportant un faible nombre de représentants ou bien des classes d'apparences avec des variances très grandes. L'algorithme EM réestime les centres des classes à chaque itération. Il faut donc s'attendre à ce que cet algorithme se montre assez sensible à la présence de valeurs atypiques. Il serait donc très intéressant de rendre le processus moins sensible aux valeurs extrêmes. Les algorithmes robustes de k-médianes ou de Ransac pourraient être envisageables pour ce problème.

Apprentissage mixte Dans le cas de l'approche de classification supervisée, une relaxation de la désignation d'un objet suivi modèle sur plusieurs plans vidéo (par exemple les vues de la voiture Mercedes du haut -filmé du ciel- et de face) permettrait de résoudre le problème des données limitées, de prendre en compte dans la modélisation statistique de l'apparence intra-plan plusieurs aspects de l'objet suivi, et donc d'augmenter le taux de reconnaissance des apparences requêtes.

Appariement mixte L'identification de deux plans ne peut pas être détectée sur la base d'un seul descripteur comme nous l'avons vu dans le chapitre 7. D'autre part l'étude présentée dans le chapitre 3 montre que l'identification de deux objets sur la base de leurs deux apparences médianes n'est pas la bonne stratégie. En représentant le plan vidéo par un nombre optimal d'images-clés (extension de l'approche de sélection des apparences-clés) nous envisageons d'effectuer un appariement mixte (plusieurs images-clés et plusieurs descripteurs) de deux plans. La difficulté ici est la définition d'une fonction statistique qui tranche entre les deux hypothèses de ressemblance ou dissimilarité.

Les aspects suivants demandent un travail plus important :

Modélisation des régions d'objets suivis La modélisation des objets suivis dans ce travail est faite par des descripteurs globaux. Pour rendre compte des parties qui apparaissent et disparaissent il serait intéressant d'avoir des descripteurs plus locaux. Ces descripteurs peuvent être centrés sur des points d'intérêts (voir section 3.3) ou des régions.

Entre ces descripteurs locaux existent des relations de position comme “voisin et au-dessus”. Il faut dans ce cas modéliser chaque descripteur local et pour cela les algorithmes EM peuvent à nouveau être utilisés, mais une telle approche nécessite le suivi de ces descripteurs, et, pour ne pas perdre d'information, la prise en compte de contraintes de positionnement relatif.

Relevance feedback Revenons sur le cas d'une classification automatique par le CAH appliqué à une matrice de distances entre les classes d'apparences intra-plan des objets. Nous avons observé expérimentalement qu'un retour par l'utilisateur d'une version retouchée “des classes d'objets” à un niveau très fin de la hiérarchie permet probablement de poursuivre automatiquement une construction correcte des classes d'objets à des niveaux plus élevés dans la hiérarchie. Une intégration de l'approche “relevance feedback” dans la construction des groupes d'objets est une voie intéressante. Reste à savoir comment et à quel niveau l'appliquer ?

Vers une modélisation intra-plan généralisée La densité du mélange gaussien est couramment utilisée en statistique pour modéliser des distributions gaussiennes. Mais il est très difficile de vérifier cette hypothèse de normalité dans des espaces de descripteurs à grandes dimensions. On pourrait considérer l'utilisation d'une loi générale comme la loi statistique gamma qui permet de modéliser n'importe quel type de distribution. Mais il faut être capable d'estimer les paramètres du mélange de lois gammas et mettre en oeuvre des critères de sélection de la structure du mélange.

Vers une sélection automatique des descripteurs Le nombre de descripteurs calculables sur les images est très grand. Le problème est de savoir comment déterminer automatiquement quel est le descripteur le plus discriminant pour une image donnée. Les statistiques doivent apporter une réponse, mais il faut garder à l'esprit que le problème est difficile car on ne dispose que de quelques images exemples, de beaucoup de contre-exemples, et que le nombre de descripteurs envisageables est infini.