Résumés automatiques de séquences vidéo

Itheri Yahiaoui, Bernard Merialdo, Benoit Huet
Département de Communication Multimedia
Institut Eurécom,BP 193
F-06904 Sophia-Antipolis, FRANCE
{Itheri.Yahiaoui,Bernard.Merialdo,Benoit.Huet}@eurecom.fr

Résumé: Dans cet article, nous proposons une nouvelle approche pour la construction automatique de résumés de séquences vidéo. En établissant un critère de distance entre un ensemble de plans et la vidéo originale, on définit le meilleur résumé comme étant l'ensemble de plans le plus proche de la vidéo originale, c'est-à-dire présentant des caractéristiques aussi similaires que possible. Nous proposons une caractérisation des plans basée sur le mouvement, puis sur une combinaison avec des informations de couleur, et étudions diverses approches de construction du résumé optimal.

1 Introduction

Avec l'avènement de l'informatique, la croissance de la puissance de calcul et de la capacité de stockage, nous sommes amenés à manipuler quotidiennement un flot croissant d'informations multimédia. Par ailleurs, la vidéo numérique devient de plus en plus utilisée dans de multiples domaines, comme par exemple l'éducation et le divertissement. Cependant les utilisateurs n'ont pas assez de temps pour consulter l'ensemble des informations disponibles. Il est donc indispensable de développer des techniques d'organisation et de recherche de documents multimédia permettant de faciliter l'accès à ces informations.

Dans cet article, nous proposons une construction automatique d'un résumé vidéo présentant des caractéristiques aussi proches que possible de la vidéo originale. Nous étudions, en premier lieu, l'utilisation du mouvement comme critère de comparaison, puis nous le combinons avec des informations de couleur. Nous présentons plusieurs méthodes de sélection de sous-ensembles de scènes représentatives en utilisant différentes mesures.

Cet article est organisé comme suit: la deuxième section présente les différents types des résumés vidéo et les principaux travaux effectués dans ce domaine. La troisième section décrit les méthodes d'évaluation existantes et celle que nous proposons. Dans la section 4, nous développons les méthodes et les espaces que nous avons utilisés pour construire nos résumés. Finalement, nous discutons les résultats obtenus et nous concluons sur les perspectives de ce travail.

2 Résumé Vidéo

La construction d'un résumé est un processus de condensation d'un document (textuel ou multimédia) vers une version courte et ce en préservant le contenu essentiel des informations. Le résumé peut être :
- Indicatif : il contient juste assez d'informations pour permettre de juger la pertinence du document entier ou bien pour donner une indication brève sur le sujet principal.
- Instructif : il peut être utilisé afin de se substituer au document source, il retient les détails importants en réduisant la quantité d'information présentée.
- Appréciatif : il capte le point de vue du réalisateur du document sur le sujet présenté dans ce dernier.

Un résumé vidéo peut être conçu soit sous forme d'un document hypermédia composé d'un ensemble d'images représentatives du contenu de la vidéo, soit comme une séquence audio-visuelle, de durée réduite, construite en prenant des extraits importants de la séquence originale. Nous nous intéressons à cette deuxième approche. Une fois le résumé vidéo construit, il peut servir à:
- juger la pertinence de la vidéo originale,
- enrichir les résultats d'une recherche,
- avoir un point de vue rapide sur une collection de vidéo.

Ce domaine de recherche est très récent. Parmi les principaux travaux réalisés; on peut mentionner :
- Le projet Informedia au CMU [3] qui définit une librairie de vidéo intelligente accessible par l'utilisateur à travers une interface de recherche dans une base de données. Cette équipe de recherche s'est concentrée sur le "Video skimming" qui est une tâche proche de la création de résumés. La vidéo est segmentée en utilisant les histogrammes de couleurs. La sélection des scènes importantes se fait par la détection des visages, les mouvements de caméra, la capture de texte, et le signal audio. Par la suite, des règles de sélection basées sur ces différents critères sont appliquées.
- Le projet MoCa à l'université de Mannheim [4, 8, 9, 11] a porté sur les bandes-annonces de film. La vidéo est segmentée en utilisant les vecteurs de cohérence ainsi que l'audio. Pour la sélection des scènes importantes, on fait appel à la détection et la reconnaissance des visages ainsi qu'une pondération des différents paramètres qui caractérisent un film.

3 Evaluation des résumés

Parmi les applications du traitement du langage naturel, la création de résumés de texte a été amplement étudiée par rapport à celle des autres documents multimédia. Plusieurs systèmes de construction de résumés de texte ont été réalisés et commercialisés ces dernières années. Une évaluation de ces systèmes a été établie au sein du projet TIPSTER [5].

Il faut noter que le problème d'évaluation est très critique, et plusieurs questions se posent concernant les méthodes et les types d'évaluations les plus appropriés. Nous pouvons comparer différents résumés selon diverses manières, par exemple on peut comparer un résumé automatique et un résumé manuel, ou un résumé automatique et un document entier, ou bien deux résumés automatiques générés de différentes manières[5]. En général, les méthodes d'évaluation sont classifiées en deux grandes catégories. La première est l'évaluation extrinsèque, dans laquelle la qualité du résumé est jugée en fonction de la manière dont il affecte la réalisation d'autres tâches déjà préparées par un agent humain. La deuxième est l'évaluation intrinsèque, où des personnes jugent directement la qualité du résumé en l'analysant selon plusieurs critères.

Afin d'éviter l'intervention d'un agent humain et pour avoir un résumé objectif sans assistance, nous proposons une évaluation automatique basée sur des méthodes mathématiques dans des espaces vectoriels et probabilistes. Le meilleur résumé sera le plus proche possible de la vidéo originale selon un ensemble de caractéristiques et une distance que nous définissons et choisissons parmi d'autres.

4 Approche globale

Le but principal de l'analyse du contenu de la vidéo est de représenter les données visuelles de manière à permettre une segmentation et une classification ainsi qu'une recherche efficaces et utiles.

La plupart des techniques de description de vidéo commencent par détecter les transitions entre les plans puis sélectionnent une image représentative de chaque plan. Ensuite, des méthodes d'analyse d'images fixes sont appliquées à ces images représentatives. D'autres techniques consistent à classifier ces images clés pour donner à l'utilisateur une présentation hiérarchique d'une base de vidéo qui sera utilisée pour la recherche.

Dans ce qui suit, nous décrivons les algorithmes et les outils que nous proposons pour produire automatiquement un résumé vidéo. Quelque soit la méthode utilisée pour construire le résumé, la question qui se pose est de savoir comment évaluer d'une manière objective le résultat. Cette évaluation doit pouvoir se faire quelque soit le degré de complexité de la méthode et surtout sans assistance humaine.

4.1 Résumé Optimal

Une vidéo est un ensemble d'images; ces images sont regroupées sous forme de plans, un plan correspond à une seule prise de vue. La concaténation de ces plans constitue la vidéo entière. Dans ce travail, nous considérons le résumé comme étant une concaténation d'un sous-ensemble de plans sélectionnés parmi l'ensemble global qui représente la vidéo originale. Le meilleur résumé est constitué par le

sous-ensemble des plans formant le résumé le plus proche de la video, c'est à dire qui minimisent la distance entre la vidéo originale et le résumé. Nous utilisons une contrainte de durée, par laquelle le résumé doit avoir une taille prédéfinie, soit en terme de nombre de plans, soit en terme de durée.
Soit:

$V = \{P_1, P_2, ...P_k\}$ P_i un plan de V,
$R = \{P_{i1}, P_{i2}, ...P_{im}\}$ un résumé de V,

Soit $d(R, V)$ une distance entre un résumé et la vidéo dans l'espace des caractéristiques considérées.
Le résumé optimal vérifie:

$$\hat{\mathrm{R}} = argmin\ d\,(R, V).$$

où le minimum est calculé sur l'ensemble des résumés de V satisfaisant à la contrainte de durée.

4.2 Caractérisation de la vidéo

Une séquence vidéo n'est pas seulement une collection d'images individuelles, c'est aussi une évolution d'un ensemble de relations spatio-temporelles entre les objets. Afin de prendre en compte ces caractéristiques, nous avons besoin d'une représentation significative de cette évolution.
Nous proposons un ensemble de représentations dérivées de caractéristiques telles que le mouvement et la couleur.

Notons:
$V = \{f_1, f_2, f_3.....f_n\}$ f_i une frame de la vidéo V,
$P_i = \{f_{i1}, f_{i2}, ...f_{im}\}$ f_{ij} une frame du plan P_i,

$C\,(V)$:Vecteur caractéristique de la vidéo,
$C\,(R)$:Vecteur caractéristique du résumé,
$C\,(f_i)$:Vecteur caractéristique d'une frame i,
$C\,(P_j)$:Vecteur caractéristique d'un plan j.

La première caractérisation que nous étudions est basée sur le mouvement. Afin de diminuer le temps de calcul, nous utilisons directement les vecteurs de mouvement des macro-blocs M des frames P dans le flux mpeg, ce qui permet d'éviter d'avoir à reconstituer les images. La quantité de mouvement correspondant à une frame est la norme du vecteur moyen de mouvement. La vidéo est représentée ensuite par l'histogramme des quantités de mouvement Q de l'ensemble des frames et chaque plan est caractérisé par l'histogramme des quantités de mouvements des frames Q(f) le composant.

- C_1(V) = Histogramme(Q(fp_i)) $fp_i \in$V,
- C_1(P) = Histogramme(Q(fp_j)) $fp_j \in$P,

- $Q\,(fp) = \sqrt{\left(\sum_{k=1}^{64} M_{xk}\right)^2 + \left(\sum_{k=1}^{64} M_{yk}\right)^2}$

Notre deuxième représentation consiste à caractériser chaque frame P par l'histogramme des quantités de mouvements de ses macro-blocs puis calculer l'histogramme moyen à partir de ces derniers.

- $Q\,(mb_r) = \sqrt{M_{xr}^2 + M_{yr}^2}$.
- $C_2\,(fp) = Histogramme(Q\,(mb_r))$.
- C_2(P) = Moy($C_2(fp_j)$) $\quad fp_j \in$P,
- C_2(V) = Moy($C_2(fp_i)$) $\quad fp_i \in$V.

Afin d'améliorer notre représentation et de rajouter de l'information, nous combinons le mouvement et la couleur. Les histogrammes de couleurs sont obtenus après la décompression totale du flux Mpeg. Ces histogrammes sont calculés dans l'espace HSV. Les vecteurs que nous utilisons actuellement sont la concaténation des histogrammes de mouvement et de couleurs. Nous normalisons les vecteurs caractéristiques de façon à ce que toutes les valeurs soient comprises dans l'intervalle [0.0 , 1.0]

- $C_3\,(f) = Hist - clr(f)$.
- C_3(V) = Moy($C_2(fp_i)$,$C_3(f_i)$) $\quad fp_i, f_i \in$V.
- C_3(P) = Moy($C_2(fp_j)$,$C_3(f_j)$) $\quad fp_j, f_j \in$P.

4.3 Construction du résumé

L'approche que nous développons se base sur l'idée de l'évaluation automatique et peut être divisée en trois étapes principales:
- Segmentation de la vidéo originale en plans,
- Calcul des caractéristiques des frames et des plans,
- Sélection optimale des plans qui composeront le résumé.

Pour la segmentation en plans et la détection de coupures, nous utilisons les résultats de la segmentation manuelle faite dans le cadre de l'Action Indexation Multimédia qui a fait le sujet d'une étude conjointe entre le CLIPS, le LIP6, l'INA et Eurecom[10]. Différents types de coupures ont été définis, mais cela n'affecte pas les fondements de notre travail.
La caractérisation des frames, des plans et de la vidéo sous forme de vecteurs dans un espace paramétrique est déjà détaillée dans la section précédente.
Après le calcul de ces vecteurs caractéristiques, nous procédons à la sélection optimale des plans constituant le résumé. L'énumération exhaustive de tous les résumés possibles a un coût prohibitif. Aussi nous recherchons une méthode sous-optimale permettant d'avoir rapidement un résumé de bonne qualité. Pour cela nous proposons trois méthodes différentes. Ces méthodes utilisent toutes un principe de sélection progressive. Nous construisons tout d'abord un résumé avec un seul plan, puis

un autre avec deux plans et ainsi de suite jusqu'à la sélection du nombre prédéterminé des plans désirés. Ce nombre reflète la proportion de la taille du résumé par rapport à celle de la vidéo originale:

$V = \{P_1, P_2, ...P_k\}$ P_i un plan de V,
$R_0 = \{\ \}$ R_0 résumé à l'étape 0.
$R_1 = \{P_{1i}\}$ P_{1i} un plan du résumé R_1 à l'étape 1.
$R_2 = \{P_{2i}, P_{2j}\}$ P_{2j} un plan du résumé R_2 à l'étape 2.
.
.
.
$R_l = \{P_{le_1}, P_{le_2}, ...P_{le_l}\}$ P_{le_j} un plan du résumé R_l à l'étape l.

.
.
$R_k = \{P_1, P_2, ...P_k\} = V$ R=V, k nombre de plans de la vidéo.

4.3.1 Sélection directe

Le principe est simple: à une étape (t) de la sélection, on essaie un par un les plans qui n'appartiennent pas au résumé R_t avec l'ensemble des plans déjà sélectionnés, et on conserve la meilleure combinaison.
$V = \{P_1, P_2, ...P_k\}$,
Si
$R_t = \{P_{t1}, P_{t2}, ...P_{tl}\}$,
alors
$R_{t+1} = R_t \cup argmin_{P \in V, P \notin R_t} d\left(C\left(V\right), C\left(R_t \cup P\right)\right).$

4.3.2 Classification K-means

Nous remarquons que dans la sélection directe par Greedy nous ne remettons jamais en cause le choix précédent: à chaque étape t, nous rajoutons aux plans déjà sélectionnés un plan parmi le reste. Pour tenter d'éviter les maxima locaux, nous étudions une autre approche dans laquelle nous remettons en cause les plans sélectionnés à chaque étape. Cette méthode utilise un arbre binaire où chaque noeud est associé à un plan, et où on permet de remplacer un plan du résumé par ses deux fils. Le problème est alors de construire un arbre efficace.
Nous utilisons l'algorithme K-means sur l'ensemble des plans. A chaque itération, les classes sont séparées en 2 sous-classes de manière à construire un arbre binaire non forcément équilibré. A chaque noeuds de cet arbre, on associe le plans le plus proche du centroide de la classe. La sélection du résumé se fait à partir de l'arbre, en partant de la racine aux feuilles, en remplaçant à chaque étape un noeud par ses deux descendants. Nous définisons un résumé à 1,2 ou t plans de la manière suivante:
$R_1 = \{P_{racine}\}$,
$R_2 = \{P_{fils1-racine}, P_{fils2-racine}\}$
Afin de construire un résumé à t+1 plans à partir d'un résumé de t plans, nous

remplaçons tour à tour chaque plan par ses deux fils. Les deux plans fils (P_{fils_1}, P_{fils_2}) qui minimisent le plus la distance du nouveau résumé R_{t+1} par rapport a la vidéo originale V sont gardés à la place de leur parent.
Si
$R_t = \{P_{t1}, P_{t2}, ...P_{tl}\}$,
alors:
$R_{ti} = R_t - \{P_i\} \cup \{P_{fils_{i1}}, P_{fils_{i2}}\}$,
$R_{t+1} = argmin_i d\left(C\left(V\right), C\left(R_{ti}\right)\right)$

4.3.3 Classification ACP-Selection par Greedy

L'arbre K-means n'est pas équilibré. Pour avoir un arbre équilibré et s'assurer que la moyenne des deux vecteurs représentant des deux classes est égale à la moyenne de l'ensemble des vecteurs compris dans ces deux classes, nous utilisons un autre mécanisme de regroupement: l'analyse en composantes principales (A.C.P). Nous rangeons l'ensemble des vecteurs caractéristiques des plans dans une matrice, et nous calculons les vecteurs propres à cette dernière en utilisant l'algorithme de Jacobi. Ensuite nous projetons l'ensemble des vecteurs dans l'espace propre sur l'axe de plus forte variance, puis nous divisons l'ensemble des vecteurs en deux classes de tailles égales. Nous itérons le processus sur chaque classe, de la même manière que le K-means, nous obtenons ainsi un arbre mais binaire équilibré. Le processus de sélection est le même que celui appliqué à l'arbre du Kmeans.

5 Résultats des expériences

Afin d'évaluer les méthodes que nous avons implémentées pour la construction du résumé, nous avons effectué une série d'expériences avec des vidéo Mpeg1 de l'INA (Journaux télévisés, Documentaire et Série de science fiction).

La Figure 1, présente l'évolution de l'erreur entre résumé et vidéo originale, en fonction du nombre de plans utilisés pour construire les résumés. Chaque graphe représente une caractérisation différente (Histogramme(Q(fp_i)), Moy(Hist-mvt(fp_i)), Moy(Hist-mvt(fp_j), Hist-clr(f_j))) et montre les résultats pour chacune des 3 méthodes de sélection (Greedy, Kmeans, ACP).

Nous remarquons rapidement que la qualité du résumé s'améliore en général en fonction de la taille de ce dernier, c'est à dire que le résumé est plus proche de la vidéo originale selon les caractéristiques choisies. Il est intéressant de noter que le choix des plans constituant le résumé est fait d'une manière objective sans aucune intervention d'un agent humain quelconque.

En comparant les 3 méthodes de constructions de résumés que nous avons utilisées (Greedy, Kmeans, ACP) nous constatons que la méthode Greedy donne de meilleurs résultats. Nous expliquons ceci par le fait que la sélection du prochain plan s'effectue parmi l'ensemble des plans non sélectionnés. Et ce, malgré le fait que Greedy ne remette pas en cause ses choix précédents. Par contre les deux méthodes

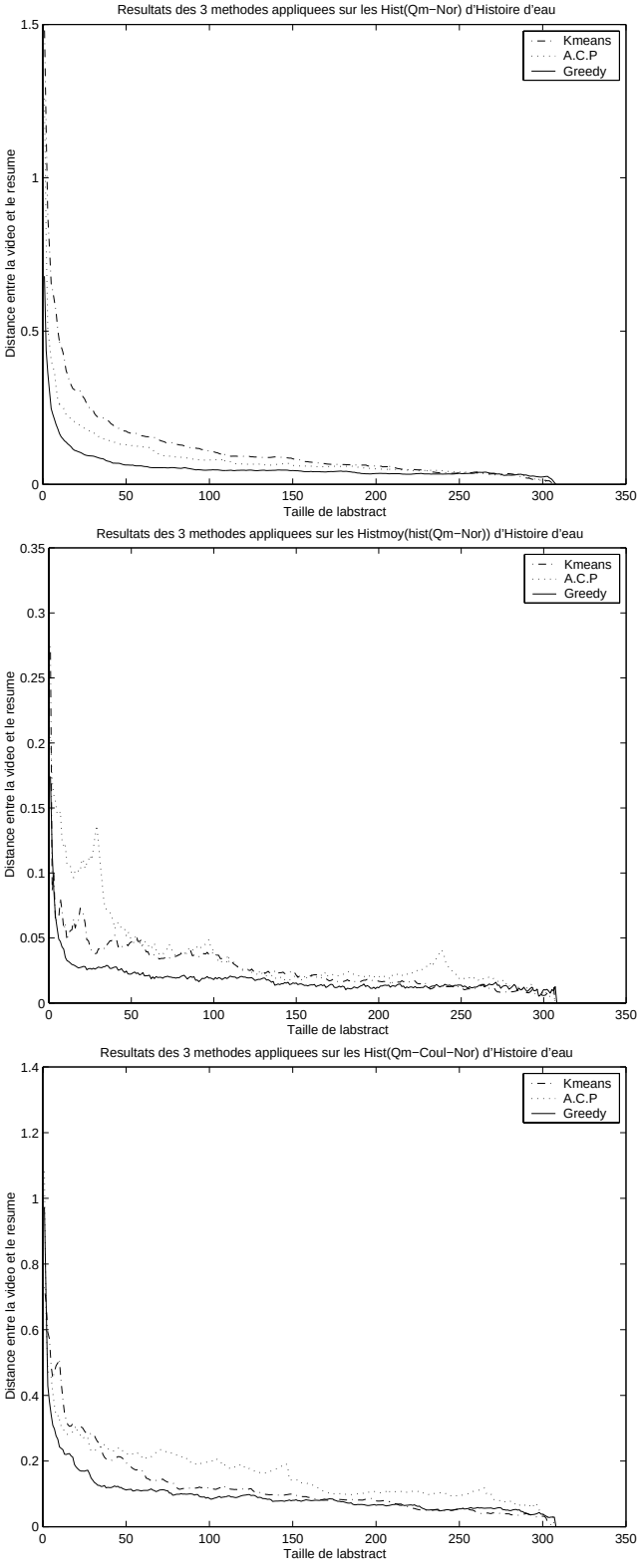

FIG. 1 – *Distance entre la vidéo et le résumé calculée avec les 3 méthodes de sélection appliquées à la vidéo "Histoire d'eau"*

Kmeans et Acp remettent en cause leur choix déjà établi mais ont une plage de sélection restreinte au niveau inferieur de l'arbre.
Nous remarquons aussi que les méthodes utilisant Kmeans et Acp ne donnent pas toujours une courbe decroissante. Nous expliquons ce phénomène par le fait que chaque noeud représente un sous-arbre dont le nombre d'élements diffère en fonction du niveau de l'arbre ainsi que la nature de ce dernier (équilibré ou non); aussi pendant la phase de classification et de construction de l'arbre, on approxime la moyenne d'une classe par un représentant ce qui entraîne une erreur qui peut augmenter d'un niveau à l'autre dans l'arbre.

6 Conclusion

Dans ce papier nous avons présenté plusieurs méthodes de construction automatique de résumé vidéo en variant les manières de caractérisation et de sélection des plans qui constituent le résumé. En premier lieu, nous présentons la vidéo originale ainsi que les unités sémantiques, dans notre cas les plans, qui la constituent dans un espace vectoriel. Puis nous établissons une classification hiérarchique et enfin une sélection qui se base sur l'optimisation de la différence entre le résumé construit et la vidéo originale. Cette approche est inspirée de la méthode d'évaluation automatique que nous voulons appliquer pour éviter toute intervention subjectivite de l'être humain. Parmi nos perspectives, l'enrichissement de la représentation en rajoutant d'autres caractéristiques comme la direction de mouvement, la texture et l'audio, ainsi que d'élargir la base de test.

Remerciements: Cette recherche est soutenue par les membres industriels d'Eurecom: Ascom, Cegetel, France Telecom, Hitachi, IBM France, Motorola,Swisscon, Texas Instruments, et Thomson CSF.

Références

[1] J. Girgensohn, A.; Boreczky. Time-constrained keyframe selection technique. *Multimedia Computing and Systems,IEEE International Conference*, 1:756–761, 1999.

[2] Andrew B. Lippman Giridharan Iyengar. Content-based browsing and editing of unstructured video. New York, USA, 30 July - 2 August 2000. ICME2000.

[3] Michael A. Smith Takeo Kanade. Video skimming and characterization through the combination of image and language understanding. IEEE International Workshop on Content-based. Access of Image and Video Databases(ICCV98-Bombay,India), 1998.

[4] Silvia Pfeiffer Lienhart and Wolfgang Effelsberg. Video abstracting. In *Communications of ACM*, December 1997.

[5] Inderjeet Mani and Mark T. Maybury. *Advences in Automatic Text Summarization*. The MIT Press, 1999.

[6] A.E Maybury, M.T.; Merlino. Multimedia summaries of broadcast news. *Intelligent Information Systems*, IIS '97:442–449, 1997.

[7] R.; Sawhney H.; Wan C. Pope, A.; Kumar. Video abstraction: summarizing video content for retrieval and visualization. *Conference Record of the Thirty-Second Asilomar Conference*, 1:915–919, 1998.

[8] Silvia Pfeiffer Rainer Lienhart and Wolfgang Effelsberg. Scene determination based on video and audio features.

[9] Silvia Pfeiffer Rainer Lienhart and stephan Fischer. Automatic movie abstracting.

[10] Stéphane Marchand-Maillet Rosa Ruiloba, Philippe Joly and Georges Quénot. Towards a standard protocol for the evaluation of video-to-shots segmentation algorithms. Toulouse, France, October 25-27 1999. European Workshop on Content-Based Multimedia Indexing, CBMI'99. http://www-asim.lip6.fr/AIM/.

[11] Gerald Kuhne Silvia Pfeiffer, Rainer Lienhart and Wolfgang Effelsberg. The moca project: Movie content analysis research at the university of mannheim.

[12] J.R. Smith. Videozoom spatio-temporal video browser. *Multimedia,IEEE Transactions*, 1(2):157–171, June 1999.