Construction Automatique de Résumés Multi-vidéos

Itheri Yahiaoui - Bernard Merialdo - Benoit Huet
Institut Eurécom
Département Communications Multimédia
BP 193 – 06904 Sophia-Antipolis- France
{Itheri.Yahiaoui,Bernard.Merialdo,Benoit.Huet}@eurecom.fr

Résumé
Dans cet article, nous présentons une approche pour la création automatique de résumés de plusieurs vidéos, comme par exemple des épisodes de séries télévisées. Cette méthodologie est basée sur le principe d'utilisateur simulé afin d'évaluer la qualité du résumé vidéo d'une manière automatique mais inspirée de la perception humaine. Il faut noter que pour les résumés multi-vidéos, il est nécessaire non seulement d'identifier les informations qui sont importantes dans une vidéo, mais aussi celles qui caractérisent cette vidéo par rapport aux autres. Afin de valider notre idée, des résultats expérimentaux sont présentés.

Mots Clef
Résumé, Vidéo, Classification.

1.Introduction
La croissance rapide des documents multimédia, comme par exemple l'énorme flux de vidéos qui se trouvent sur les ordinateurs personnels et autres équipements, nécessite le développement de nombreux outils pour leur manipulation. La création automatique de résumés vidéo est un outil performant qui permet de résumer le contenu général de la vidéo et de présenter les parties les plus pertinentes sous forme d'une séquence audiovisuelle ou d'un ensemble d'images représentatives. Les résumés vidéos permettent d'avoir rapidement une idée sur le contenu de très grandes bases de vidéos, sans nécessiter la visualisation et l'interprétation de l'ensemble des vidéos. Cela permet aussi de juger et d'évaluer la pertinence d'un document multimédia par rapport aux autres.
Plusieurs travaux se sont déjà attaqués au problème de la construction automatique du résumé d'une vidéo. Cependant la plupart des approches actuelles souffrent des limitations suivantes:

- L'évaluation de la qualité du résumé est très difficile, il est donc très délicat d'apporter un jugement sur la performance du résumé résultant. Même si cette dernière est calculée en utilisant un critère et une mesure mathématiques, l'interprétation et la compréhension du sens restent très complexes.
- En général, les travaux actuels se concentrent plus sur la construction de résumé d'une seule vidéo à la fois et ne s'attardent pas sur le problème de résumés multi-vidéos où d'autres contraintes s'imposent et d'autres éléments doivent être pris en compte.

Dans cet article, nous nous intéressons au problème de la construction automatique des résumés de plusieurs vidéos, comme par exemple des épisodes d'une série télévisée. Une application envisagée est par exemple de faciliter le choix pour un utilisateur dont le magnétoscope numérique a effectué automatiquement un certain nombre d'enregistrements. Une solution consiste à faire de façon séparée le résumé de chaque vidéo. Cependant, cette approche ne prend pas en compte les similitudes pouvant apparaître dans les différentes vidéos, et ces résumés pourraient être redondants. Notre approche est basée sur le principe d'utilisateur simulé afin d'évaluer la qualité du résumé vidéo d'une manière automatique qui s'inspire de la perception humaine.

2.Principe de l'Utilisateur Simulé
Nous nous inspirons du principe de l' utilisateur simulé qui se base sur un critère mathématique simulant l'évaluation d'un vrai utilisateur. Tout d'abord, nous définissons une vraie expérience, une tâche que les utilisateurs doivent accomplir et pour laquelle une mesure de performance peut être définie. Puis grâce à certaines hypothèses réalistes, il est possible de prédire le comportement des utilisateurs lors de la réalisation de cette tâche. Ce modèle de comportement est ensuite utilisé par notre algorithme pour construire et évaluer automatiquement des résumés vidéos .
L'application du principe de l'utilisateur simulé au problème des résumés multi-vidéos nous conduit à proposer le scénario suivant afin d'effectuer l'expérience:

- L'ensemble des résumés sont montrés à l'utilisateur,
- Un extrait aléatoire choisi d'une vidéo quelconque lui est ensuite présenté,
- L'utilisateur essaye de deviner de quelle vidéo cette extrait provient.

Le comportement simulé de l'utilisateur est le suivant:

- Si l'extrait contient des images similaires à une ou plusieurs images appartenant à une seule vidéo, il donnera comme réponse la vidéo correspondante (cette réponse n'est pas nécessairement correcte),

- Si l'extrait contient des images qui sont similaires à d'autres images appartenant à différents résumés, la situation est ambiguë et l'utilisateur ne peut pas donner une réponse définitive,
- Si l'extrait ne contient aucune image qui ressemble aux images des résumés, l'utilisateur ne peut pas répondre.

La performance de l'utilisateur dans cette expérience est le pourcentage de ses réponses correctes pour tous les extraits possibles de l'ensemble des vidéos qui lui sont montrés. C'est seulement dans le premier cas décrit précédemment que l'utilisateur peut identifier une vidéo particulière. Mais cette réponse n'est pas forcément correcte, parce qu'une image d'un extrait d'une vidéo peut être similaire à une image d'un résumé d'une autre vidéo.
Cette approche fait deux hypothèses: la première est que l'utilisateur a une mémoire visuelle parfaite, donc il est capable d'identifier immédiatement les images lui étant montrées. La deuxième est que nous utilisons une mesure mathématique de similarité entre images à la place du jugement de l'utilisateur.

3.Approche Globale

Dans ce papier, nous présentons différents algorithmes que nous utilisons pour la construction automatique de résumés multi-vidéos. Le principe de l'utilisateur simulé est ensuite utilisé pour l' évaluation de la qualité des résumés construits. Enfin, nous comparons et discutons les résultats d'évaluation afin de définir l'algorithme le plus approprié pour cette application.
En premier lieu nous décrivons le principe de trois algorithmes basés sur une idée expérimentée ultérieurement (plus de détails sont disponible dans [12]), ensuite nous expliciterons le reste des algorithmes.
Le procédé de construction de résumés multi-vidéos est divisé en cinq étapes. Les trois premières ainsi que la dernière sont communes aux six algorithmes, cependant la quatrième qui effectue la sélection des éléments à inclure dans le résumé est spécifique à chaque algorithme.

Pré-traitement du flux vidéo: Nous éliminons le générique de début et celui de la fin, ceci bien qu'ils comportent des éléments importants de la vidéo, ils ne spécifient pas un épisode donné.

Construction de vecteurs caractéristiques: Cette étape consiste à l'analyse du contenu des vidéos afin de représenter les données visuelles sous forme de vecteurs caractéristiques. Les images sont divisées en neuf régions égales pour lesquelles des histogrammes de couleurs sont calculés. Ensuite les neuf histogrammes sont concaténés pour former le vecteur caractéristique de l'image correspondante. Afin de diminuer le coût de calcul et l'espace mémoire, nous faisons un sous-échantillonnage de la vidéo telle qu'une seule image par seconde est prise en compte.

Classification: Les images sont classifiées par une étape initiale où on crée une nouvelle classe chaque fois que la distance de l'image par rapport aux classes existantes est supérieure à un certain seuil. Ensuite plusieurs itérations du type k-Means sont réalisées afin de raffiner les classes. Cette classification d'images basée sur la comparaison des histogrammes respectifs produit des classes d'images visuellement similaires.

Sélection des segments vidéo: Pour chaque épisode, nous sélectionnons les classes caractéristiques les plus importantes en se basant sur six méthodes différentes qui seront présentées dans la section suivante.

Présentation du résumé : Le résumé global peut être construit et présenté à l'utilisateur sous forme d'une séquence audio-visuelle d'une durée réduite ou d'une grille d'images représentatives où chaque ligne représente un épisode. Le nombre d'images sélectionnées par épisode est défini par l'utilisateur.

4.Différentes méthodes de sélection

Une fois que les images sont classifiées, les vidéos seront décrites par des ensembles d'images représentatives parmi les classes les plus importantes. Les méthodes proposées permettant de calculer le degré d'importance de chaque classe sont définies comme suit :

Méthode 1 : Une mesure de couverture basée sur le critère d'évaluation décrit précédemment est utilisée pour la sélection. On attribue une valeur de couverture à chaque classe. Elle représente le nombre d'extraits d'une durée prédéfinie qui contiennent cette classe. Dans cette méthode la couverture est calculée en utilisant seulement la vidéo courante pour laquelle on sélectionne une classe à rajouter. Une classe doit être sélectionnée une seule fois, elle ne peut pas représenter deux vidéos à la fois dans le même résumé global. Pour respecter cette contrainte, nous utilisons une couverture conditionnelle. Tous les extraits qui contiennent des classes déjà sélectionnées seront négligés.

Méthode 2 : Cette méthode est similaire à la première. La seule différence est que la couverture des classes candidates dans les autres vidéos est prise en compte pendant la sélection. Afin de diminuer les cas ambigus et erronés, nous utilisons un coefficient négatif pour imposer une pénalité sur les classes ayant une large couverture dans les autres vidéos.

Méthode 3 : Afin de comparer la sélection dépendante et indépendante comme étant une expérience de base pour valider l'importance et la spécificité des résumés multi-vidéos, nous construisons les résumés de chaque vidéo. Lorsque nous choisissons de nouvelles classes à inclure dans le résumé d'une vidéo donnée, nous ignorons la présence de ces classes dans les autres résumés. Cependant , une classe peut apparaître deux ou plusieurs fois dans le résumé global constitué de la concaténation des différents résumés correspondant à chaque vidéo.

Méthode 4 : Afin d'éliminer tous les cas ambigus dans notre expérience simulée, nous développons un algorithme basé sur le calcul de la couverture de la même façon que les méthodes précédentes, sauf que les classes candidates ne doivent pas être présentes ni dans les autres

résumés ni dans les extraits qui contiennent des classes déjà sélectionnées dans les résumés des autres vidéos.

Méthode 5 : Basée sur le travail de Uchihashi et Foote [9], qui définissent une mesure pour le calcul de l'importance des plans, nous adaptons cette mesure à notre méthode de construction de résumés multi-vidéos. Les plans sont construits à partir de notre classification en concaténant les images successives appartenant à la même classe. La mesure de l'importance d'un plan est à peine modifiée par rapport au travail original tel que le poids d'une classe Wi, qui est la proportion de plans parmi les plans de la vidéo entière qui appartiennent à la classe i, est calculé ainsi $W_i = S_i \Big/ \sum_{j=1}^{C} S_j$ où C est le nombre de classes composées de toutes les images des épisodes vidéo prises en compte et S_i est la longueur totale de l'ensemble des plans appartenant à la classe i, obtenue par la sommation des longueurs de tous les plans appartenant à la classe. Donc l'importance I du plan j (de la classe k) est $I_j = L_j \log 1/W_k$ où L_j est la longueur du plan j.

Un plan est important s'il est long et ne ressemble pas à la plupart des autres plans. Afin de représenter chaque vidéo par des plans spécifiques et les plus long possible, nous calculons le facteur d'importance pour tous les plans possibles. Ensuite nous sélectionnons dans chaque vidéo les plans les plus importants afin de les inclure dans les résumés correspondants.

4.6 Méthode 6 : L'idée principale de cette méthode est de faire un parallèle avec les méthodologies de construction de résumés texte[6], où la formule TF-IDF a prouvé qu'elle est très intéressante. Pour les résumés texte, cette approche est basée sur les mots qui forment les unités de base, cependant pour les résumés multi-vidéo les unités sont les classes. Alors l'importance I de la classe c est calculée de la manière suivante $I_c = L_c \log n/nc$ où L_c est la longueur (durée totale) de la classe c, n le nombre de vidéos et nc le nombre des vidéos contenant au moins une image de la classe c. Ayant calculé l'importance de chaque classe, nous sélectionnons les plus importantes qui composerons le résumé global. Dans le cas où une classe est présente dans plusieurs vidéos, nous devons déterminer à quel résumé nous l'affectons. Nous faisons ceci en calculant pour chaque vidéo la proportion d'images appartenant à cette classe présentent dans la vidéo, et on choisit la plus probable.

5 Expériences

Dans cette section nous présentons les résultats d'évaluation en utilisant le principe d'utilisateur simulé sur les résumés vidéos construits avec les six algorithmes. Nous avons effectué une série d'expériences avec des vidéos de type Mpeg1. Ces vidéos représentent six épisodes de la série télévisée « Friends ». Nous avons choisi de présenter nos résumés sous la forme d'une grille avec six images par épisode. Ceci est particulièrement adapté à l'affichage sur un écran de télévision ou d'ordinateur.

La figure 1 illustre le résultat de notre premier algorithme de création de résumé. Chaque ligne de cette grille d'image est spécifique à une des six vidéos.

Figure 1

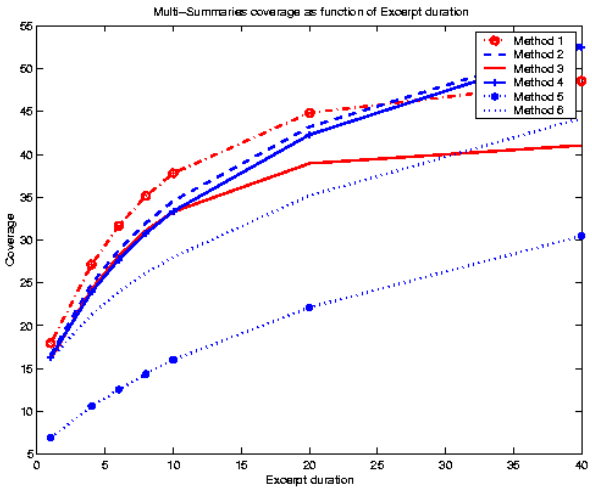

Figure 2

Le graphe dans la figure 2 présente les performances respectives de ces six méthodes quand la durée de l'extrait utilisé lors de l'évaluation varie. Nous notons que les deux premières méthodes qui construisent les résumés en se basant sur un critère mathématique inspiré du critère d'évaluation lui même donnent les meilleurs performances. Nous notons aussi que les résumés multi-épisodes (méthodes 1 et 2) sont plus efficaces que les résumés vidéos simples (méthode 3). Comme prévu la cinquième méthode ne donne pas de bons résultats. Ceci est du au fait que les plans sont sélectionnés selon leur longueur et leur faible nombre d'occurrences. Certainement, les plans rares ont une petite couverture à

travers la vidéo. La méthode 6, inspirée de la formule TF-IDF donne des résultats moyens par rapport aux autres. Nous notons aussi que les résultats de la quatrième méthode sont comparables à ceux de la deuxième, et que les deux donnent la meilleure couverture pour des extraits de longue durée.

6.Robustesse des résumés

Ayant construit des résumés de multi-vidéos en utilisant un certain nombre de méthodes, il est intéressant d'évaluer la performance des résumés pour une durée d'extrait donnée. Les quatre premières méthodes sont dépendantes de la durée d'extrait cependant les deux dernières ne le sont pas. Afin d'étudier la robustesse, les résumés sont construits pour différentes durées d'extrait et ensuite évalués avec plusieurs durées d'extrait. La Figure 3 présente les résultats de cette expérience pour des résumés construits avec la première méthode. Notons que cette méthode suggère que la couverture doit être maximale lorsque la durée d'extrait utilisée pour la construction et l'évaluation est la même. A part le cas des résumés construits avec une durée d'extrait égale à 1 seconde, toutes les méthodes restantes donnent des performances similaires et de bon niveau.

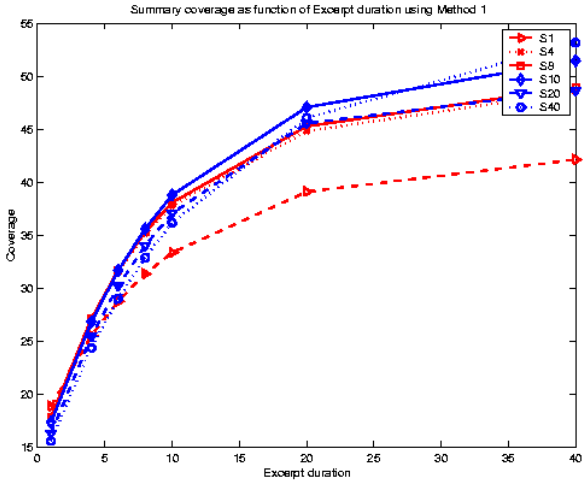

Figure 3

7.Conclusion

Une comparaison de différentes0 approches de construction automatique de résumés multi-vidéos a été présentée. En se basant sur le principe de l'utilisateur simulé, nous évaluons les résultats obtenus par six méthodologies différentes. Nos expériences démontrent que lorsque la construction et l'évaluation sont effectuées avec le même principe, les meilleurs résultats sont réalisés. La méthode proposée donne clairement de meilleurs résultats que la méthode de Uchihashi and Foote [9] et que la méthode inspirée de la formule TD-IDF. Notre évaluation de la robustesse des résumés montre qu'il est possible d'obtenir des résultats raisonnables avec des résumés construits en utilisant une durée d'extrait spécifique.

Bibliographie

[1] Anastasios D. Doulamis, Nikolaos D. Doulamis and Stefanos D. Kollias. Efficient Video Summarization based on a Fuzzy Video Content Representation. IEEE International Symposium on Circuits and Systems, vol. 4, pp. 301-304, 2000.

[2] Benoit Huet, Itheri Yahiaoui, Bernard Merialdo. Multi-Episodes Video Summaries. International Conference on Media Futures, pp. 231- 234, 2001.

[3] Bernard Merialdo. Automatic Indexing of Tv News. Workshop on Image Analysis for Multimedia Integrated Services, pp. 99-104, 1997.

[4] Giridharan Iyengar and Andrew B. Lippman. Videobook: An Experiment in Characterization of Video. IEEE International Conference on Image Processing, vol. 3, pp. 855-858, 1996.

[5] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg. Video Abstracting. In Communications of ACM, vol. 40, no. 12, pp 54-62, December 1997.

[6] I Mani and M. T. Maybury. Advances in Automatic Text Summarization. MIT Press, 1999.

[7] Mark T. Maybury and Andrew E. Merlino. Multimedia Summaries of Broadcast News. IEEE Intelligent Information Systems, pp. 442 -449, 1997.

[8] M.A. Smith and T. Kanade. Video Skimming and Characterization through the Combination of Image and Language Understanding. IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 61-70, 1998.

[9] Shingo Uchihashi and Jonathan Foote. Summarizing Video Using a Shot Importance Measure and a Frame-Packing Algorithm. IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 3041-3044, 1999.

[10] Nuno Vasconcelos and Andrew Lippman. Bayesian Modeling of Video Editing and Structure: Semantic Features for Video Summarisation and Browsing. IEEE International Conference on Image Processing, vol. 3, pp. 153-157, 1998.

[11] Yihong Gong; Xin Liu. Generating Optimal Video Summaries. IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1559-1562, 2000.

[12] I Yahiaoui, B. Merialdo and B. Huet. Generating Summaries of Multi-episodes Video. IEEE ICME 2001.