CLASSIFICATION AUTOMATIQUE DE SEGMENTS VIDÉO

Bernard Merialdo
Institut Eurecom
B.P. 193
06904 Sophia-Antipolis
merialdo@eurecom.fr

1. INTRODUCTION

La disponibilité croissante de documents multimédia sous forme digitale pose le délicat problème de l'accès à ce gigantesque volume d'information. Si la recherche documentaire a apporté depuis longtemps des mécanismes pour retrouver de l'information textuelle, les documents audio-visuels posent des difficultés particulières. En effet, pour définir automatiquement le contenu d'un document audio-visuel, il faut faire appel à des technologies de reconnaissance de formes, soit pour le son, soit pour l'image, dont la complexité est souvent très grande, et dont les performances souffrent de nombreuses limitations.

Dans cet article, nous nous intéressons à l'analyse et à l'indexation automatique de journaux télévisés. Le but est de reconnaître la structure de journal de façon à identifier les différents éléments qui le constituent (présentateur, reportages, interviews, publicités...). A partir de ces éléments, il est ensuite facile de construire, par exemple, un interface utilisateur permettant un accès hypermédia aux différentes parties du journal. Nous décrivons différentes phases du traitement d'analyse et nous détaillons en particulier les algorithmes qui sont mis en oeuvre pour reconnaître des segments vidéo comme étant des plans fixes de personnes interviewées. Cette classification passe par la construction automatique d'un arbre de décision analysant le mouvement dans le segment vidéo.

2. ANALYSE DES JOURNAUX TÉLÉVISÉS

Les journaux télévisés constituent un champ important de l'indexation multimédia. D'une part, ils sont produits et consommés en quantité abondante, ce qui en fait un matériau facilement disponible. D'autre part, bien que présentant une grande diversité, ils sont souvent construit selon des modèles relativement simples, ce qui fournit des indications précieuses pour faciliter leur analyse. Enfin, eu égard à la diversité des thèmes traités et aux intérêts variés du grand nombre d'utilisateurs, la personnalisation de leur contenu en fonction du destinataire constitue une application potentiellement très utile au plus grand nombre.

De nombreux travaux se sont déjà intéressés au problème de l'indexation de journaux télévisés, selon des modalités très variées. Si l'on peut disposer directement des éléments constituant le programme, il devient facile d'assembler ces éléments selon un format prédéfini [Compton 95], ce qui permet de se passer de la phase d'indexation. Lorsqu'on

souhaite traiter des journaux existants, l'approche la plus simple est l'utilisation des sous-titres à l'usage des mal-entendants [Bacher 95][Brown 95] (lorsque ceux-ci sont disponibles), puisqu'elle permet d'utiliser toutes les technologies d'indexation et de recherche basées sur le texte. Les différents mots de la transcription servent alors de pointeurs hypermédia vers la partie du journal correspondante.

Une approche plus sophistiquée est d'utiliser des techniques de reconnaissance de parole pour identifier automatiquement des mots-clés ("word spotting") dans la bande-son du journal [Gelin 96][Jones 96]. Bien que la reconnaissance de parole présente encore des difficultés à résoudre (l'indépendance du locuteur, la taille du vocabulaire, la précision de reconnaissance...) des progrès spectaculaires ont été accomplis ces dernières années.

Une autre perspective est d'analyser la partie vidéo en utilisant des techniques de traitement d'image pour reconnaître la structure et les différentes parties du journal. Dans ce cas, on utilise des paramètres tels que la texture, les couleurs, l'intensité, les formes, la détection de coupures, la reconnaissance de mouvements de caméra, ... pour obtenir de l'information sur le contenu vidéo du journal. Une revue des approches existantes se trouve dans [Aigrain et al, 96]. Notons que l'on peut aussi analyser directement les sous-titres apparaissant sur l'image (la plupart du temps les noms des intervenants) [Lienhart 96]. [Zhang et al, 95] ont proposé une approche d'analyse des journaux télévisés basés sur l'utilisation d'un modèle de présentateur. Notre étude est largement inspirée de leurs travaux.

Enfin, au delà des difficultés propres à l'indexation, d'autres travaux s'intéressent à un niveau plus conceptuel du traitement, en particulier pour les applications de filtrage [Abiza 96].

3. VIDÉO ANNOTÉE

3.1 BASE D'ENREGISTREMENTS

Pour nos besoins, nous avons réalisé des enregistrements de 6 journaux télévisés consécutifs (CNN World News au cours du mois de Septembre 1996) représentant au total environ 3 heures de vidéo. Ces enregistrements ont été digitalisés à l'aide d'une carte SunVideo au format 340x280, et codés en MPEG-1 à 12.5 images/sec. Avec ce codage, les 6 enregistrements (audio et vidéo) tiennent sur un seul CD-ROM.

3.2 FICHIERS D'ANNOTATION

L'analyse de l'enregistrement d'un journal s'effectue en plusieurs étapes mettant en jeu des processus différents. Chaque processus rajoute de l'information à la structure connue du journal. Pour stocker ces informations de façon cohérente, nous avons défini un format de "fichier d'annotation" en nous inspirant de la norme SGML. Chaque information est représentée par une balise dont le nom indique le type de l'information et les attributs contiennent les valeurs des divers paramètres qui caractérisent cette information. Les informations sont reliées à l'enregistrement au travers de time-codes (dans notre cas des numéros d'image dans la vidéo). Cette représentation autorise des structures hiérarchiques, puisque la portée d'une balise peut recouvrir d'autres balises. Nous pouvons ainsi mélanger micro et macro-segmentation dans le même formalisme.

Nous utilisons actuellement les balises suivantes:

- SHOT: un segment vidéo entre deux coupures de plan,
- PUB: un ensemble de segments représentant une série de clips publicitaires,
- REPORT: une suite de SHOTS entre deux occurrences du présentateur,
- INTERVIEW: un reportage contenant principalement des plans de personne.

Chaque balise contient des attributs, dont voici quelques exemples:

- START: numéro de la première image,
- END: numéro de la dernière image,
- PERSON: indique que le segment représente un plan fixe d'une personne,
- ID: dans le cas d'un segment indiquant une personne, donne un numéro identifiant cette personne (et permettant de la reconnaître dans d'autres segments).

Le contenu d'un fichier d'annotation est donc similaire à l'exemple ci-après:

```
<INTERVIEW>
<SHOT START=7106 END=7569 PERSON ID=0>
<SHOT START=7569 END=7722 PERSON ID=6>
<SHOT START=7722 END=7853>
...
<SHOT START=8524 END=8944 PERSON ID=6>
<SHOT START=8944 END=8985 PERSON ID=0>
</INTERVIEW>
```

4. TRAITEMENT DE LA VIDÉO

4.1 SEGMENTATION

La première étape du traitement de la vidéo consiste en une segmentation en utilisant un algorithme de détection de coupures basé sur la variation de l'histogramme d'intensité [Benedetti 95]. Pour chaque image, on calcule l'histogramme de l'intensité des pixels (selon les valeurs de 0 à 255). On compare ensuite les histogrammes de deux images consécutives selon la norme L1:

$$d(H_1, H_2) = \sum_{i=0}^{255} \left| h_1(i) - h_2(i) \right|$$

Une coupure est détectée lorsque cette distance est supérieure à un seuil. Un exemple de graphe représentant les variations de cette distance se trouve dans la figure 1.

La détection des coupures permet de découper la vidéo en segments consécutifs représentant chacun un plan. L'analyse du contenu de ces segments va ensuite permettre de retrouver la structure du journal.

4.2 COMPARAISON D'IMAGES

La deuxième phase du traitement consiste à comparer les segments vidéo pour reconnaître si certains se ressemblent. C'est en particulier le cas pour les segments contenant le présentateur, mais des segments similaires se retrouvent également au cours des reportages ou des interviews. Pour chaque segment, une image représentative est sélectionnée (on choisit pour cela une image au milieu du segment). Les différentes images obtenues sont ensuite comparées en utilisant la distance d'histogrammes

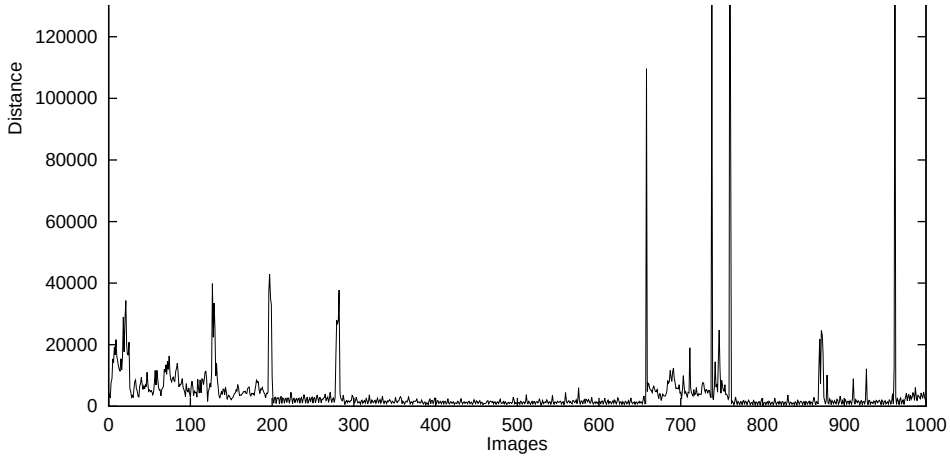


Figure 1: distance entre deux images consécutives

précédemment décrite. On utilise un algorithme de classification où une nouvelle image est intégrée à une classe existante si la distance au centre de classe est inférieure à un seuil, et forme une nouvelle classe sinon. Cette classification peut s'effectuer à l'intérieur de l'enregistrement d'un seul journal, ou bien en prenant en compte les enregistrements de tous les journaux. La figure 2 montre trois exemples d'une telle classification..

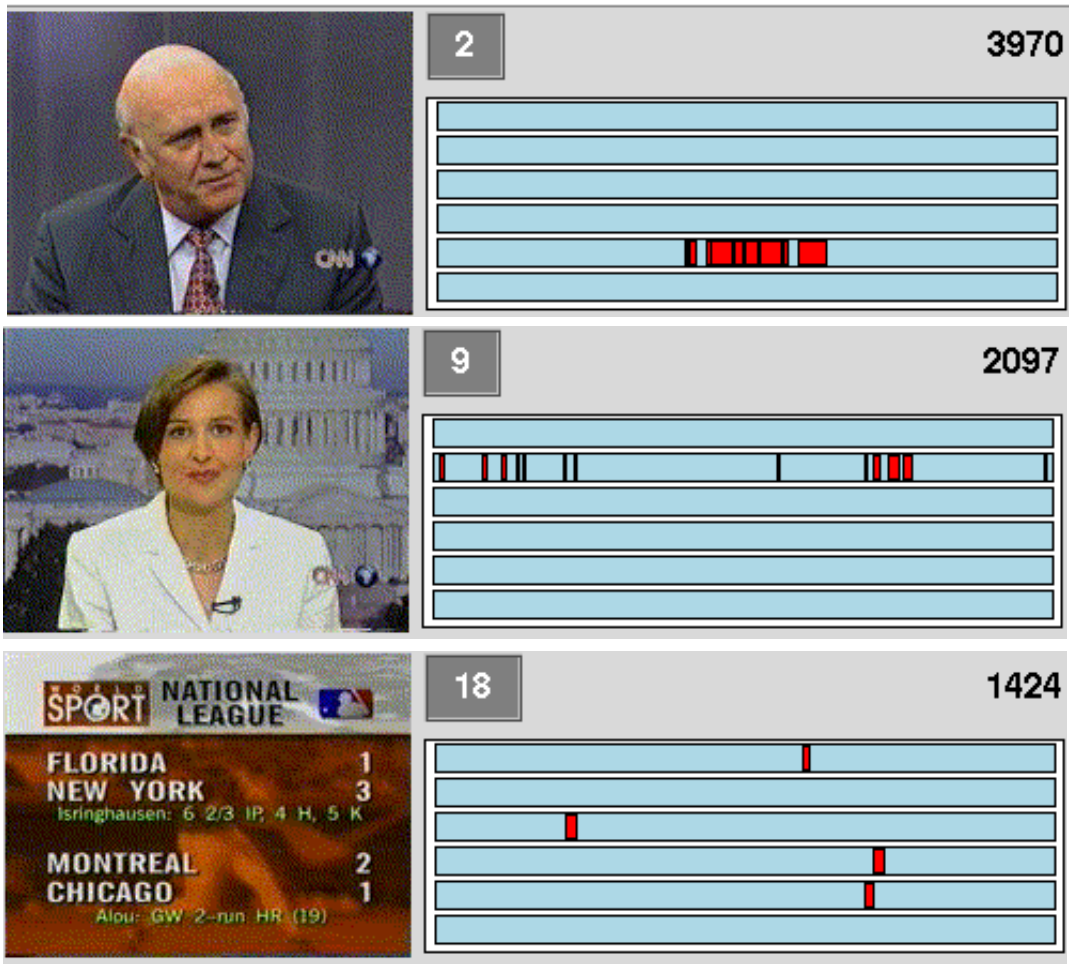


Figure 2: classification des images

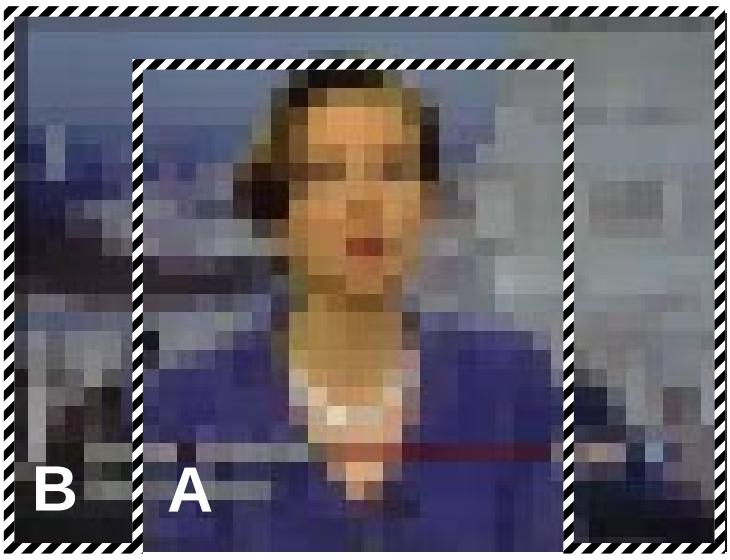
La première image représente une personne apparaissant dans un interview. On remarque que l'on trouve de multiples occurrences de cette image, mais localisées le même jour dans un intervalle de temps réduit. La seconde image est celle d'une présentatrice, dont les occurrences se répartissent tout au long du journal. La troisième image est un panneau fixe indiquant des résultats sportifs. Cette image se retrouve sur plusieurs jours de la semaine.

5. DÉTECTION DE PERSONNES

Un élément important dans un journal télévisé est de reconnaître les personnes qui sont apparues dans le journal. Reconnaître le présentateur permet de trouver le découpage du journal en reportages, et reconnaître les personnes permet de détecter les interviews.

Lors d'un journal, les plans fixes sur une personne présentent des caractéristiques particulières. Le fond est en général fixe, ou bouge très peu, il n'y a pas de mouvement de caméra, le cadrage est toujours similaire.

Pour déterminer qu'un segment vidéo correspond au plan fixe d'une personne, [Zhang et al, 96] proposent d'utiliser une mesure à base de moyenne et variance des différences entre certaines zones d'images consécutives. Ces zones sont choisies selon un ou plusieurs modèles pour définir grossièrement les zones de fond et les zones où apparaît la personne.



L'approche de Zhang consiste à mesurer les variations entre deux images successives:

- pour toute l'image, pour la zone A, et pour la zone B,
- en calculant la moyenne et la variance pour tout le segment considéré (soient les valeurs μ, σ, μ_A, σ_A, μ_B, σ_B),
- en utilisant comme mesure la différence des histogrammes d'intensité ou le nombre de pixels qui changent.

Un segment est classifié comme "potentiellement personne" lorsque l'image totale varie peu, la zone A varie un peu, et la zone B ne varie presque pas, ce qui se traduit par les inégalités (où les t_i sont des seuils à définir):

1) $\mu \leq t_1 \qquad \sigma^2 \leq t_2$

2) $\mu_A \geq t_{1A}$ $\sigma_A{}^2 \geq t_{2A}$

3) $\mu_B \leq t_{1B}$ $\sigma_B{}^2 \leq t_{2B}$

En tenant compte des deux mesures utilisées, cela fait un ensemble de 12 seuils à déterminer pour appliquer cette méthode. D'autre part, nous avons remarqué que dans nos enregistrements figurent également des plans fixes (par exemple des tableaux de résultats) qui vérifient très facilement les inégalités 1) et 3), et donc qui sont souvent susceptibles d'être confondus avec des plans de personne.

Pour éviter une approche du type "trial and error" pour choisir les meilleures valeurs de ces seuils, nous avons utilisé une méthode de classification plus générale basée sur la construction d'un arbre de décision [Breiman et al, 84] qui permet de généraliser les tests précédents et d'obtenir automatiquement les seuils en observant des exemples de données d'apprentissage. Cette construction s'effectue de la manière suivante:

- en annotant manuellement une partie des vidéos, on obtient un ensemble de segments de chaque type avec les valeurs des paramètres correspondants (moyenne et variance pour les différentes zones et mesures). On suppose que ces données sont initialement situées à la racine de l'arbre.
- l'étape de base de la construction de l'arbre consiste à partager ces données de façon "optimale" en posant une question sur les valeurs des paramètres. Nous considérons les questions de la forme *"valeur < seuil"* (pour chaque paramètre, il y a donc autant de questions que de valeurs différentes). Le critère d'optimalité est la minimisation de l'entropie du type du segment connaissant la réponse à la question H(type du segment / noeud(paramètres)). Les questions sont énumérées et évaluées exhaustivement et la meilleure est conservée pour ce noeud de l'arbre. Les données sont alors partagées entre les deux feuilles correspondant à la réponse à la meilleure question.
- ce processus d'éclatement des données d'un noeud continue jusqu'à ce qu'un critère de fin soit satisfait. Ce peut être qu'il n'y a plus d'amélioration de l'entropie par découpage, ou que cette amélioration est trop faible, ou encore, que l'amélioration apportée par cette question sur un autre jeu de données est trop faible. Lorsque plus aucun noeud ne peut être éclaté, la construction de l'arbre est terminée.

5.1 RÉSULTATS

Nous donnons ici les résultats de classification obtenus par différentes méthodes. Deux enregistrements de journaux annotés manuellement sont utilisés: un pour construire l'arbre de décision, l'autre pour évaluer la qualité du résultat. Nous considérons deux types de découpage de l'image en zone: A-B précédemment décrit, et 3x3 correspondant à un quadrillage régulier en 9 parties. Nous considérons deux façons de construire l'arbre: "arbre complet" dans lequel les noeuds sont divisés tant que l'entropie sur les mêmes données diminue, "arbre limité" où la moitié des segments d'apprentissage sert à trouver la meilleure question et l'autre moitié est utilisée pour le critère d'arrêt du découpage.

Le tableau ci-dessous résume les résultats obtenus (pour chaque type, on indique le nombre de segments auxquels on a attribué ce type de façon correcte et entre parenthèse

de façon erronée):.

		PERSONNE	FIXE	AUTRE	Précision
	référence	43	25	258	
Arbre complet	A-B	24 (21)	7 (8)	231 (35)	0.80
	3x3	30 (11)	6 (12)	236 (31)	0.83
Arbre limité	A-B	19 (16)	9 (4)	238 (40)	0.82
	3x3	34 (11)	4 (8)	241 (28)	0.86

On peut remarquer que la limitation de l'arbre par des données extérieures permet d'éviter la sur-adéquation (over-fitting) et d'améliorer la robustesse de la classification. Bien sûr une classification sur des paramètres aussi simples ne peut pas fournir des résultats parfaits, mais elle est fort utile si ces résultats sont combinés avec d'autres méthodes et peut, par exemple, servir de préfiltre à des méthodes de reconnaissance plus coûteuses.

6. CONCLUSION

Nous avons présenté quelques étapes de classification de segments vidéo pour analyser le contenu d'un journal télévisé. Après un découpage en plans, d'une part on regroupe les segments en comparant leurs images représentatives, d'autre part on détecte les plans fixes de personnes en calculant la variation de l'image dans différentes zones. L'utilisation de données d'apprentissage annotées manuellement avec des procédures automatiques pour trouver les critères de décision constitue une approche très générique des problèmes de classification. Elle s'applique naturellement au cas de la classification de segments vidéo, et ainsi permet d'éviter une mise au point manuelle des seuils qui s'avère souvent fastidieuse et incertaine.

7. REMERCIEMENTS

Laurent Doucet, Hector Espinoza, Stephane Heulin et Laurence Thiery ont largement contribué à la réalisation des différents traitements décrits dans cet article.

8. REFERENCES

[[Abiza 96] Y. Abiza, A. Leger and M. Crehange, *"Conceptual Modelling for Information Filtering in Broadcast Interactive Video Applications"*, Multimedia Modeling Conference, Toulouse, November 1996, pp 35-50.

[Aigrain et al, 96] Philippe Aigrain, HongJiang Zhang and Dragutin Petkoviç *"Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review"*, Multimedia Tools and Applications, vol 3, no 3, November 1996, pp: 179-202.

[Bacher 95] D. R. Bacher and C. J. Lindblad, *"Content-based Indexing of Captioned Video on the ViewStation,"* MIT TNS Laboratory, Technical Note, October 1995.

[Benedetti 95] Gerard Benedetti, Benoit Bodin, Franck Lhuisset, Olivier Martineau and B. Merialdo, *"A structured Video Browsing Tool"*, in Engineering for Human-

Computer Interaction, edited by Leonard J. Bass and Claude Unger, Chapman & Hall, 1996, pp: 17-26.

[Breiman et al, 84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *"Classification and Regression Trees",* Wadsworth, Belmont, 1984.

[Brown et al, 95] Martin Brown, Jonathan Foote, Gareth Jones, Karen Sparck-Jones and Steve Young, *"Automatic Content-Based Retrieval of Broadcast News"*, ACM Multimedia Conference, November 1995.

[Compton 95] C. L. Compton and P. D. Bosco, *"Internet CNN NEWSROOM: A Digital Video News Magazine and Library"*, International Conference on Multimedia Computing and Systems", IEEE, May 1995.

[Gelin 96] Philippe Gelin and Christian Wellekens, *"Keyword Spotting Enhancement for Video Soundtrack Indexing"*, Proc. ICSLP '96, Philadelphia, PA, October 1996, vol 2, pp: 527-530.

[Jones 96] G.J.F. Jones, J.T. Foote, K. Sparck-Jones and S.J. Young, *"Retrieving spoken documents by combining multiple index sources"*, Proc. SIGIR'96, Zurich, August 1996, pp: 30-39.

[Lienhart 96] Rainer Lienhart and Frank Stuber, *"Automatic text recognition in digital videos*", in Image and Video Processing IV 1996, Proc. SPIE 2666-20, 1996.

[Zhang et al, 95] HongJiang Zhang, Shuang Yeo Tan, Stephen Smoliar and Gong Yihong, *"Automatic Parsing and indexing of news video"*, Multimedia Systems, ACDM-Springer, vol 2, 1995, pp: 256-266.

EURECOM

M. Gerard EUDE
Direction Scientifique
CNET
38-40 Rue du Général Leclerc
92794 Issy les Moulineaux cedex

Sophia-Antipolis, le 10 Mars 1997

Monsieur,

Veuillez trouver ci-joint le texte de ma présentation pour les journées CORESA 97, ainsi que l'autorisation d'impression. Lors de ma présentation, je souhaiterais projeter des images d'écran depuis mon ordinateur portable (j'ai déjà envoyé un email à ce sujet), j'aimerais donc disposer d'un projecteur type Barco ou équivalent.

Avec mes remerciements,

Bernard Merialdo
Dépt Communications Multimédia
Institut EURECOM