Utilisation de corpus pour l'Indexation Vidéo de Journaux Télévisés
A Corpus-based Approach to Video Indexing for TV News

Bernard Merialdo
Département Communications Multimédia
Institut Eurecom
BP 193, 06904 Sophia-Antipolis, FRANCE

merialdo@eurecom.fr

Résumé:

L'indexation vidéo est un élément essentiel pour l'accès aux documents multimédia par le contenu. Dans cet article, nous proposons une méthodologie d'indexation basée sur la construction et l'utilisation de corpus de vidéo annotée, à la fois pour l'apprentissage et pour l'évaluation des techniques mises en oeuvre. Nous commentons rapidement les avantages et inconvénients d'une telle approche, puis nous décrivons comment nous l'avons utilisée pour une application d'indexation. Nous décrivons un ensemble d'Outils d'Annotation de Vidéo qui facilite la construction de telles bases de données. Nous proposons également un format unique pour stocker les informations d'annotation. Puis nous présentons comment ces données sont utilisées pour l'analyse automatique d'enregistrements de Journaux Télévisés, suivant une approche proposée par Smoliar and Zhang. Nous proposons des améliorations de certaines étapes, par exemple pour la détection de personnes, nous utilisons un arbre de décision qui est construit automatiquement à partir d'exemples. Enfin, nous montrons comment le résultat de l'analyse peut servir à construire un interface utilisateur qui permet un accès hypermédia au contenu de ces enregistrements.

Abstract:

Video indexing is an important component for content-based access to multimedia documents. In this paper, we propose an approach to video indexing which is based on the construction and usage of annotated video corpora, both for training and for evaluation purposes. We briefly discuss the characteristics and advantages of this approach, then we describe the steps that we have followed to use it for video indexing applications. We describe a set of Video Corpus Annotation Tools that facilitate the construction of such databases. We also propose a standardized format for storing these annotations. Then we describe how we use this approach in the automatic analysis of TV News recordings, following techniques proposed by Smoliar and Zhang. We propose improvements on some steps, for example our person detection step uses a decision tree that is automatically constructed from the manually annotated video corpus. Finally, we propose a user interface which summarizes the results of this analysis and provides an hypermedia access to the video recordings.

1 Introduction

La prolifération des documents multimédia fait qu'il devient très important d'être capable d'organiser efficacement et d'accéder facilement à ces documents. Les applications de recherche et de filtrage de documents multimédia nécessitent de pouvoir indexer ces documents d'après leur contenu. L'indexation automatique est difficile, et si l'indexation de documents textuels est étudiée depuis longtemps, l'indexation d'autres types de données telles que l'audio ou la vidéo est plus récente car elle utilise des techniques de reconnaissance de formes qui sont encore coûteuses et imparfaites.

Pour développer des méthodes fiables d'indexation vidéo, il est nécessaire de pouvoir disposer de volumes importants de vidéo annotée, c'est-à-dire de vidéo associée à des informations précises et fiables sur son contenu, par exemple la position des frontières de plan, la présence et la position de certains objets, les mouvements, etc... Un tel corpus est utile pour l'apprentissage, lorsque certains algorithmes doivent être adaptés (par exemple en ajustant des seuils), ainsi que pour l'évaluation lorsque l'on veut mesurer la performance de certaines méthodes. Dans ces deux cas, la qualité et la quantité des données sont des facteurs importants: pour l'évaluation parce que l'on ne peut garantir un résultat que si il a été obtenu sur un nombre suffisant d'échantillons, et pour l'apprentissage, parce que de nombreuses méthodes utilisent des mesures statistiques qui fournissent des estimations dont la qualité dépend directement de la quantité de données traitée. Pour produire de façon efficace de telles données, il faut disposer d'outils adaptés permettant une production de volume à un coût raisonnable.

Les approches basées sur une analyse quantitative des données se sont révélées très productives dans de nombreux domaines, par exemple en autres: la reconnaissance de parole (voir les bases telles

TIMIT[8]), l'analyse du langage naturel (voir le Penn Treebank [14]), et des applications comme la recherche documentaire (avec la série des bases de données TREC [10]). Dans chacun de ces domaines, l'utilisation de grandes masses de données associées à des informations précises a été très productive. La disponibilité des mêmes données pour toutes les équipes de recherche a engendré des progrès importants dans l'amélioration des algorithmes dans ces domaines.

Nous pensons qu'il est important que des bases de données comparables soient développées pour l'analyse de la vidéo. Dans la suite de cet article, nous présentons d'abord quelques idées sur la construction efficace de ce genre de base, en décrivant un ensemble d'Outils d'Annotation Vidéo. Nous proposons également un format générique pour représenter les différentes informations qui sont rajoutées par ces outils. Enfin, nous décrivons une application d'analyse automatique des Journaux Télévisés où nous utilisons ces informations pour améliorer certaines étapes de l'approche proposée par Smoliar and Zhang. Finalement, nous montrons comment un interface utilisateur peut être construit à partir des résultats de l'analyse.

1.1 Enregistrements de Journaux Télévisés

Nous avons enregistré une série de six journaux télévisés de CNN World News, pendant une semaine de Septembre 1996. Chaque enregistrement dure environ 25 minutes. Nous les avons digitalisés à 12,5 images par seconde, avec une résolution de 384x288 pixels, et compressés avec un compresseur MPEG-1. Cela représente en tout 500 Moctets de données, ce qui permet de les stocker sur un CD-ROM (les paramètres d'enregistrements et de compression ont été sélectionnés à cet effet).

2 Outils d'Annotation Vidéo

L'annotation manuelle de grandes quantités de données peut être une activité très consommatrice de temps. Pour être efficace, il est nécessaire de disposer d'outils adéquats. Ces outils doivent être conçus pour respecter deux critères:

- la vitesse d'annotation: pour traiter de gros volumes, ces outils doivent être rapides,
- la qualité de l'annotation: pour être utiles, les annotations doivent satisfaire un niveau de qualité pré-établi. Idéalement, les annotations devraient être parfaites, mais cela est parfois trop coûteux à garantir, et il faut alors trouver le bon compromis pour avoir le meilleur rapport qualité/coût.

Les outils que nous allons décrire ont été conçus pour les besoins de l'analyse des journaux télévisés, toutefois, nous pensons qu'ils sont suffisamment généraux pour s'adapter à de nombreuses autres applications. Nous avons réalisé des outils pour les opérations suivantes:

- segmentation manuelle: pour séparer une séquence vidéo en segments consécutifs,
- vérification de la segmentation: pour la localisation des séparations à l'image près,
- classification de segments: pour affecter une catégorie à chaque segment (dans notre cas, ces catégories sont PERSONNE, METEO, SPORT...),
- identification de personne: pour reconnaître l'apparition de la même personne dans des segments différents.

2.1 Segmentation de Vidéo

Le premier traitement d'une séquence vidéo est souvent la détection des séparations de plans, pour découper la séquence en plans consécutifs. Des algorithmes efficaces ont été proposés pour détecter automatiquement ces coupures, toutefois ils ne sont pas parfaits, surtout lorsque l'on souhaite traiter les différents effets de transition tels que le fondu-enchaîné ou les volets. Le premier outil que nous proposons a pour but de construire rapidement une segmentation manuelle de la vidéo. Le principe en est simple: la vidéo est jouée à l'écran, et l'opérateur appuie sur un bouton à chaque fois qu'il observe un nouveau plan. Cependant quelques ajouts sont nécessaires pour construire un outil efficace et agréable. L'aspect de l'interface est montré dans la figure ci-dessous:

Nous avons ajouté trois améliorations pour rendre cet outil plus utilisable:

- à cause du temps de réaction de l'opérateur, l'instant où il appuie ne correspond pas exactement au moment de la coupure, mais est légèrement en retard. Nous avons donc ajouté une procédure d'alignement automatique qui ajuste la coupure sur l'endroit où la différence des images consécutives est la plus grande, dans l'intervalle précédent (nous avons trouvé qu'une seconde était une durée adéquate),
- il est possible de rejouer la vidéo en accéléré, ce qui permet de créer la segmentation en

moins de temps que ne dure la vidéo. Bien sûr, en augmentant la vitesse de défilement, on augmente aussi l'incertitude sur la position de la coupure, mais l'ajustement automatique permet généralement un recadrage correct. Dans nos expériences, nous avons trouvé qu'il était possible de rejouer la vidéo en double vitesse, ce qui permet de fournir la segmentation d'une heure de vidéo en une demi-heure de temps.

- l'opérateur peut toujours se tromper, et il faut donc faciliter la prévention, la détection et la correction des erreurs. Trois types d'erreurs de segmentation sont possible: ajout d'une frontière inexistante, oubli d'une frontière, mauvais placement d'une frontière. Nous avons trouvé que les erreurs du troisième type étaient prises en compte par la procédure d'alignement automatique (pour autant que l'erreur ne soit pas trop grande) et qu'il n'y avait pas d'oubli (l'opérateur pouvant rester suffisamment concentré pour ne pas manquer de frontière). Par contre, il arrive que l'opérateur ajoute des frontières en trop à la suite d'un geste malencontreux. Dans ces cas, il s'en rend généralement compte immédiatement. Nous avons donc ajouté à l'interface un bouton UNDO permettant d'effacer la dernière entrée. Pour des raisons de commodité, l'effacement peut également être obtenu en appuyant sur le bouton droit de la souris (ce qui évite le déplacement jusqu'au bouton).

2.2 Vérification de la segmentation

La segmentation manuelle obtenue comme précédemment contient le nombre correct de segments avec des frontières définies de façon approchée (enregistrée manuellement avec un décalage, puis ajustées automatiquement). Pour évaluer la qualité de l'ajustement, nous avons construit un outil de Vérification de la Segmentation qui visualise les images adjacentes à chaque frontière, en donnant à l'opérateur la possibilité de modifier cette frontière en cliquant sur une des images. Nous avons trouvé que l'affichage de 5 images (sur une vidéo échantillonnée à 12,5 images par seconde) était suffisant pour valider une frontière (mais ce nombre pourrait facilement être modifié). Lorsque la transition entre les plans est franche, l'ajustement automatique est facile. Dans les cas d'une transition plus progressive, comme pour les fondus-enchaînés, il est plus délicat de déterminer un emplacement précis pour la frontière. Les figures ci-dessous montrent deux exemples de frontières pour ces deux types de situation: *(certaines images ont été dégradées pour préserver l'anonymat*)

Remarquons que si l'on avait besoin de marquer le type de chaque transition, il serait facile de le faire en rajoutant à cet interface un bouton pour chaque type, comme cela est le cas dans l'outil de classification de segment décrit dans la section suivante.

2.3 Classification de segment

Lors de la classification de segments, on affecte à chaque segment vidéo une étiquette donnant des indications sur son contenu. Pour l'application aux Journaux Télévisés, nous avons défini les classes suivantes:

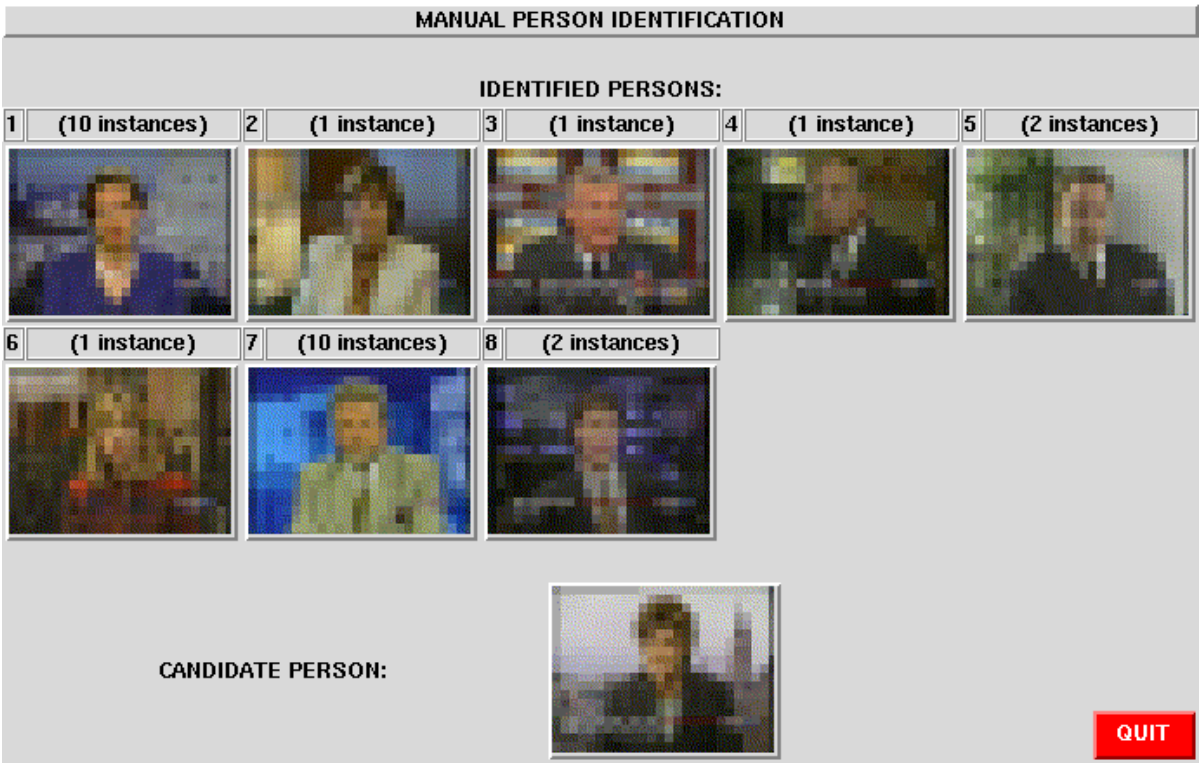
- PERSONNE (PERSON): représente une personne interviewée en plan fixe,
- CARTE (MAP): une carte géographique, ou un graphique, avec pas ou peu d'animation,
- NOIR (BLACK): un suite d'images noires qui délimite une suite de clips publicitaires,
- SPORT: un panneau de résultats sportifs, tels le football ou le base-ball,
- METEO (WEATHER): le bulletin météorologique. Les images sont composées d'un fond

formé d'une carte ou d'une vue satellite, devant lequel un présentateur signale les différents éléments, avec parfois une animation de ces éléments.

Dans tous ces exemples, il suffit d'observer une image représentative pour être capable de déterminer la classe d'un segment. L'outil de Classification de Segment affiche donc les images représentatives des différents segments, au-dessus d'un ensemble de boutons indiquant les classes possibles. L'opérateur appuie pour chaque segment sur le bouton correspondant à la classe correcte. L'interface correspondant est présenté dans la figure suivante:

2.4 *Identification des personnes*

Nous avons également besoin d'identifier les différents segments où apparaît la même personne. Dans ce but, nous avons réalisé un outil d'Identification de Personne qui affiche séquentiellement les images représentatives des segments vidéo qui ont été classifiés comme des plans PERSONNE, ainsi que les images des personnes déjà identifiées. Si la nouvelle image correspond à une nouvelle personne, l'opérateur clique sur l'image, et l'outil l'ajoute dans la liste des personnes identifiées. Si la nouvelle image représente une personne connue, l'opérateur clique sur l'image de cette dernière et rajoute ainsi une instance de la personne. La figure suivante montre l'aspect de l'interface de cet outil alors que huit personnes ont déjà été identifiées:



3 Fichier d'Annotation

Au fur et à mesure que l'on déroule les différentes phases de l'analyse sur une vidéo, on crée de nouvelles informations d'annotation. Il est commode que ces différentes informations soient stockées selon un format homogène, pour que l'on puisse facilement le manipuler, le visualiser, le modifier, que ce soit par des procédures manuelles ou automatiques. Nous avons défini un format de fichier d'annotation utilisant une syntaxe SGML. Chaque information est représentée par une balise SGML dont le nom indique le type d'information et les attributs servent à coder les caractéristiques de cette information. La relation entre les annotations et la vidéo s'effectue au travers des numéros d'image dans la séquence. Cette représentation permet de définir des structures hiérarchiques, car une balise peut inclure d'autres balises, par exemple un élément de macro-segmentation peut contenir plusieurs éléments de micro-segmentation.

Nous utilisons actuellement les balises suivantes:

- SHOT (plan): une séquence d'images représentant un plan,
- PUB: une suite de plans représentant une séquence publicitaire,
- REPORT: un reportage, suite de plans limitée par la présence du présentateur,
- INTERVIEW: un reportage contenant une majorité de plan de personnes (y compris éventuellement le présentateur).

Chaque balise contient des attributs qui servent à représenter les propriétés de chaque élément d'information. Nous utilisons par exemple les attributs suivants:

- START: le numéro de la première image du segment,
- END: le numéro de la dernière image du segment,
- PERSON: indique que le plan représente une personne,
- ID: dans le cas d'une personne, identifie la personne (par rapport aux autres personnes présentes dans la vidéo).

Ci-dessous est présenté un extrait de fichier d'annotation correspondant à un interview:

```
<INTERVIEW>
<SHOT START=7106 END=7568 PERSON ID=0>
<SHOT START=7569 END=7721 PERSON ID=6>
<SHOT START=7722 END=7852>
<SHOT START=7853 END=8027 PERSON ID=6>
<SHOT START=8028 END=8339>
<SHOT START=8340 END=8410>
<SHOT START=8411 END=8454 PERSON ID=6>
<SHOT START=8455 END=8472>
<SHOT START=8473 END=8488 PERSON ID=6>
<SHOT START=8489 END=8502 PERSON ID=0>
<SHOT START=8503 END=8523>
<SHOT START=8524 END=8943 PERSON ID=6>
<SHOT START=8944 END=8984 PERSON ID=0>
</INTERVIEW>
```

4 Indexation automatique

Dans cette partie, nous présentons l'analyse automatique des enregistrements de journaux télévisés. Notre travail suit l'approche proposée par Zhang and Smoliar, avec les différences suivantes:

- nous utilisons un arbre de décision pour l'identification des plans de personne,
- nous utilisons un critère différent pour trouver les apparitions du présentateur,
- les résultats de l'analyse servent à construire un interface hypermédia permettant d'accéder directement à l'intérieur des enregistrements.

L'analyse automatique se compose des étapes suivantes:

- segmentation de la vidéo,
- suppression des publicités,
- classification des segments pour reconnaître les plans contenant des personnes,
- identification du présentateur,
- séparation de l'enregistrements en reportages et interviews,
- construction de l'interface utilisateur,

Nous avons utilisé les Outils d'Annotation Vidéo décrits précédemment pour annoter les deux premiers enregistrements. Les données annotées ont été utilisées pour ajuster ou pour évaluer les différentes étapes du traitement.

4.1 Segmentation de la vidéo

Les frontières des différents plans sont détectées en calculant la distance entre deux images successives et en la comparant à un seuil pré-établi. Nous calculons cette distance à partir des histogrammes d'intensité des images [3].

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

De nombreuses méthodes de détection plus élaborées ont été proposées, en particulier pour mieux traiter les effets spéciaux tels les fondus-enchaînés, les volets, les mouvements de caméra etc... Nous avons trouvé que la méthode des histogrammes fournissait des résultats satisfaisants sur les données considérées. Le seuil de détection de coupure a été déterminé à partir des valeurs des distances trouvées aux positions obtenues par la segmentation manuelle (après ajustement automatique).

4.2 Suppression des publicités

Dans les enregistrements que nous utilisons, les plages de publicité sont introduites par des écrans noirs, correspondant probablement à un changement

de source vidéo au centre de production. La classification manuelle des segments permet d'obtenir des exemples de ces écrans noirs, qui sont moyennés pour construire un modèle. Les segments des nouveaux enregistrements sont alors comparés à ce modèle, ce qui permet d'identifier les coupures publicitaires. Les plans apparaissant entre ces coupures sont alors identifiés comme des parties de publicités et exclus de la suite de l'analyse.

4.3 Identification de personnes

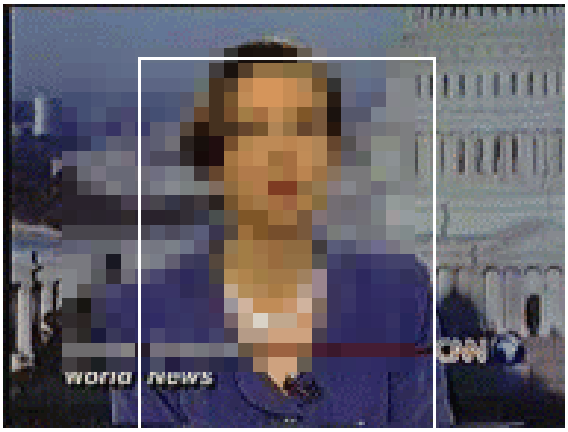
Pour identifier les segments contenant des plans de personnes (en particulier le présentateur), Zhang a proposé une méthode basée sur la constatation que, dans un tel segment, il y a peu de mouvement, et que ce mouvement est situé dans une région centrale de l'écran, là où se trouve la personne, avec un fond qui ne bouge pas. Plus précisément, Zhang définit un modèle d'image contenant deux zones fixes A et B, où A correspond à l'emplacement de la personne, et B ne couvre que du fond. Sur l'image entière, ainsi que sur chacune des régions A et B, on calcule la moyenne et la variance des différences d'image sur tout le segment. Les considérations exprimées précédemment sont alors formalisées par les inégalités:

$$\mu \le t_1 \qquad \sigma^2 \le t_2$$

$$\mu_A \ge t_{1A} \quad \sigma_A{}^2 \ge t_{2A}$$

$$\mu_B \le t_{1B} \quad \sigma_B{}^2 \le t_{2B}$$

où les t_i sont des seuils choisis de façon appropriée.



Deux distances sont utilisées: la différence des histogrammes, et le nombre de pixels différents, si bien que 12 seuils doivent être définis pour appliquer cette méthode.

Trouver les bonnes valeurs de ces seuils est une tâche difficile, c'est pourquoi nous avons choisi de modifier cette procédure en utilisant un arbre de décision qui est construit automatiquement à partir de données d'apprentissage, et qui permet d'identifier le type d'un segment en posant des questions sur les valeurs des différents paramètres mesurés sur ce segment. La construction de l'arbre peut se faire de façon entièrement automatique, et donc la nouvelle procédure ne nécessite aucune intervention manuelle.

La procédure modifiée fonctionne de la façon suivante: à partir de la vidéo annotée, on calcule les valeurs des moyennes et variances définies précédemment pour chaque segment, ainsi que la valeur du type attribué à ce segment. Cela fournit un vecteur pour tous les segments s de l'ensemble d'apprentissage:

$$\left\{ \mu_X(s), \sigma_X{}^2(s) \right\}, t(s)$$

composé de 12 valeurs des paramètres et de la valeur du type. Nous avons seulement considéré les trois types suivants:

- PERSON pour les plans de personne,
- FIXED pour les plans montrant des cartes ou des graphiques, éventuellement animés,
- OTHER pour tous les autres types.

(nous avons introduit les segments FIXED car ils contiennent peu de mouvement et ont donc des valeurs de paramètres souvent proches de celles des segments PERSON).

La procédure de construction de l'arbre de décision réalise une classification descendante [4], où les classes sont divisées récursivement pour minimiser l'entropie de la probabilité du type selon la classe:

$$H(t/c) = -\sum_c p(c) \cdot \sum_t p(t/c) \cdot \log p(t/c)$$

Initialement, tous les segments d'apprentissage sont dans une même classe. On considère alors les questions de la forme:

$$Q_i(x) \;=\; p_i(s) \le x$$

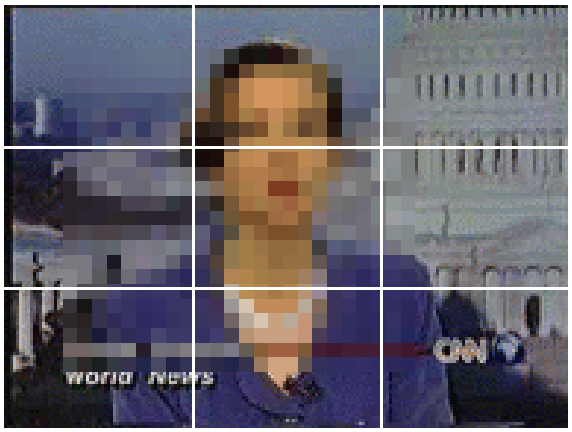
c'est-à-dire que l'on compare la valeur du paramètre i (parmi les 12) avec un seuil x. Remarquons que si nous avons N segments d'apprentissage, il ne peut y avoir plus de N valeurs différentes pour chaque paramètre, et donc que toutes les questions de cette forme ne peuvent engendrer plus de 12xN divisions différentes de la classe. Nous pouvons donc énumérer toutes ces divisions et conserver celle qui minimise l'entropie résultante.

$$Q_0 \;=\; \underset{Q}{\operatorname{argmin}}\, H(t/Q)$$

Les classes sont récursivement divisées jusqu'à ce que la meilleure question trouvée n'améliore plus l'entropie calculée sur un jeu de données de contrôle ("held-out"). Cela signifie alors que la question n'est plus suffisamment générale.

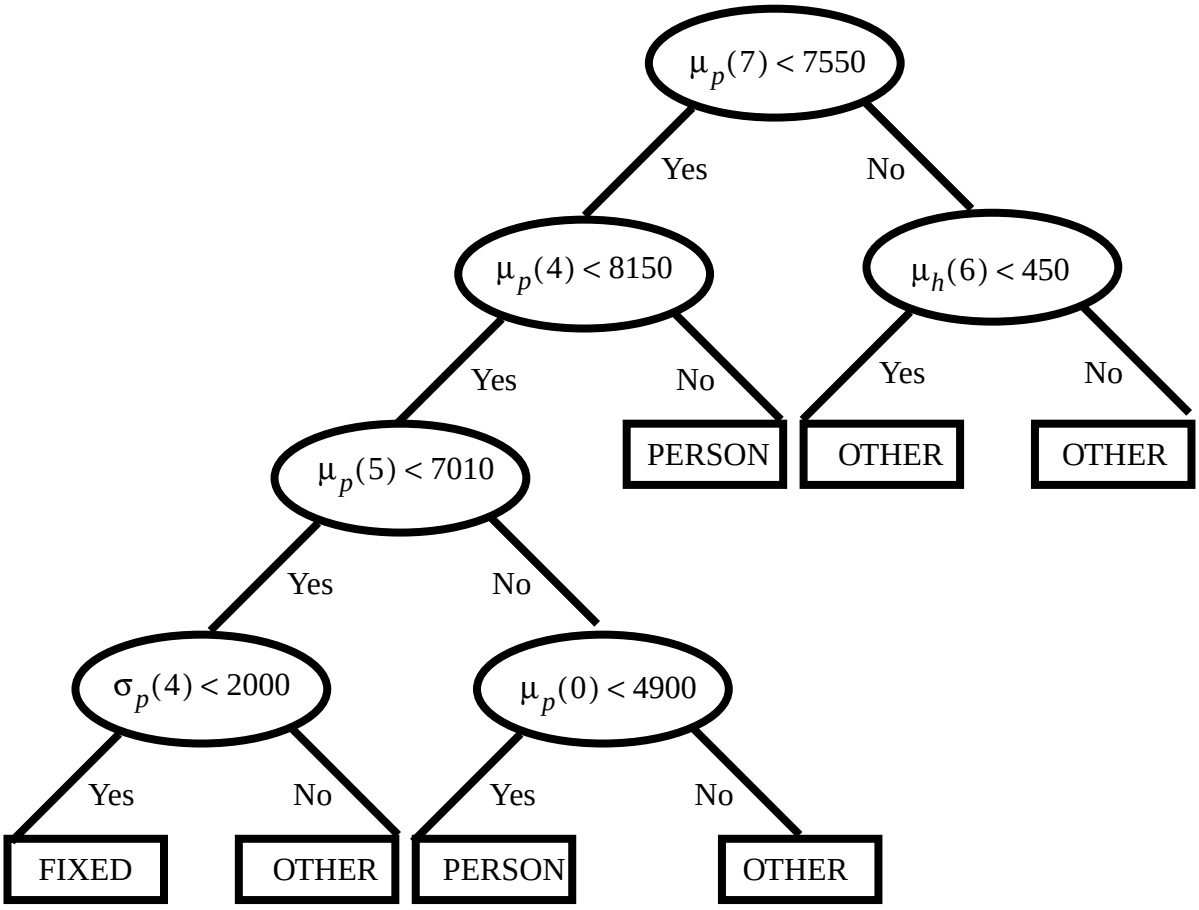
Nous avons appliqué cette procédure en utilisant la

moitié du premier enregistrement annoté manuellement comme données d'apprentissage, la seconde moitié comme données de contrôle, et le second enregistrement pour l'évaluation de la performance. Nous avons également réalisé une expérience où nous avons utilisé un découpage régulier 3x3 de l'image à la place de la classification A-B. La figure suivante montre la position de ces régions.



L'arbre qui a été construit avec la grille 3x3 est explicité dans la figure suivante. Chaque noeud correspond à une question posée sur la valeur d'un paramètre. Les paramètres sont les moyennes et variances pour les distances d'histogramme ou de pixels (respectivement $\mu_p(z)$, $\mu_h(z)$, $\sigma_p(z)$, $\sigma_h(z)$, z étant un numéro de zone entre 0 et 9, avec 0 pour la zone haut-gauche et 2 pour la zone bas-gauche). A chaque feuille de l'arbre, on affecte le type ayant la plus forte probabilité aux nouveaux segments qui vérifient toutes les questions-réponses menant à cette feuille.

(On pourra remarquer que l'arbre obtenu directement par cette méthode peut parfois être simplifié: dans l'exemple les deux feuilles de droite ont toutes deux l'étiquette OTHER car, bien que la question précédente ait amélioré l'entropie, elle n'a pas modifiée l'étiquette la plus fréquente dans les feuilles. Des questions supplémentaires auraient permis de raffiner ces étiquettes, mais elles n'ont pas été jugées suffisamment significatives par la procédure de sélection.)



Les résultats de l'évaluation de la procédure de détection de personne sur le second enregistrement sont présentés dans le tableau suivant (les nombres entre parenthèses indiquent le nombre de segments erronés ayant été classifiés en ce type):

	PERSON	FIXED	OTHER	Précision
référence	43	25	258	
A-B	19 (16)	9 (4)	238 (40)	0.82
3x3	34 (11)	4 (8)	241 (28)	0.86

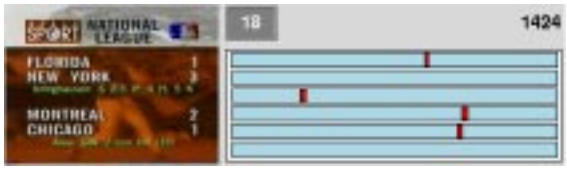
4.4 Identification du présentateur

En fait, Zhang a utilisé plusieurs modèles de découpage en régions pour trouver le présentateur (un ou deux présentateurs, présence ou non d'une icône sur la gauche ou la droite). Il compare aussi les segments entre eux pour voir s'ils contiennent des images similaires. Les segments qui respectent un modèle et ressemblent à d'autres segments (une autre occurrence de la même scène) sont identifiés comme correspondant à un (ou plusieurs) présentateur.

Dans nos enregistrements, il n'y a toujours qu'un présentateur et pas d'icône, de sorte qu'un seul modèle suffit, par contre on rencontre souvent des segments qui se ressemblent mais ne correspondent pas au présentateur. C'est le cas par exemple lors d'un interview où les images de la personne interviewée alternent avec celles du journaliste. Nous avons classifié les segments de tous les enregistrements en utilisant un algorithme de k-moyenne, ce qui permet de visualiser les modes d'apparition de certains segments. Par exemple, la première figure représente les occurrences d'une personne interviewée dans le journal télévisé du quatrième jour.

La figure suivant montre les occurrences du présentateur du premier journal:

La dernière figure montre les occurrences d'un panneau d'affichage de résultats sportifs (on peut remarquer qu'il apparaît dans plusieurs journaux):



Ces observations suggèrent d'utiliser les modes d'apparition des segments pour identifier ceux correspondant au présentateur. Nous avons évalué trois critères pour réaliser cette identification;

- la plus longue durée totale D d'apparition durant un journal,
- le plus grand nombre N d'apparitions,
- D x variance des apparitions.

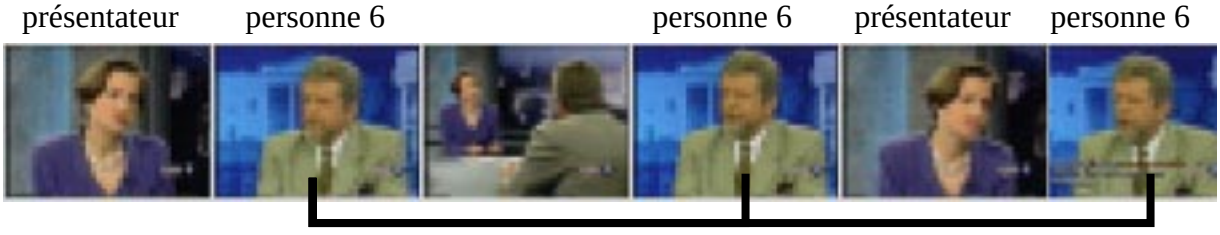
En examinant les six enregistrements, nous avons mesuré les résultats suivants (on calcule la valeur du critère pour tous les segments, et on regarde les segments correspondant à la valeur maximum):

critère	identification du présentateur
durée D	0
nombre N	3
D x variance	5

Lorsque l'on n'applique ces critères aux seuls segments où des personnes ont été détectées, le troisième critère fournit une identification parfaite du présentateur.

4.5 Séparation en reportages

Lorsque le présentateur a été identifié, les segments apparaissant entre deux occurrences du présentateur sont considérées comme des sujets.



Comme le présentateur peut parfois mener un interview, on fusionne dans le même sujet des segments ayant des images semblables, comme indiqué dans la figure précédente.

En fonction du nombre de segments contenant des personnes dans un sujet, on classifie le sujet comme “reportage” ou comme “interview”.

5 Interface utilisateur

Les informations obtenues comme résultat de l’analyse permettent de construire un index hypermédia qui autorise l’accès direct au contenu des enregistrements. La figure suivante montre une représentation de cet interface:

L’interface présente un résumé du contenu des six enregistrements. Chaque ligne d’images représente une journée, et contient les images représentatives des six reportages ou interviews les plus longs ce jour-là. Au-dessus de chaque image, un segment de couleur indique la position et la durée de ce reportage dans le journal. En cliquant sur une image, on peut consulter les détails concernant ce reportage (durée, indice de début et fin...) ou bien déclencher la visualisation de la partie de l’enregistrement correspondante.

D’autres présentations sont également disponibles. Par exemple, le bouton TOPICS permet de visualiser les images représentatives de tous les reportages de la semaine, classés par durée décroissante. Le bouton PEOPLE permet de visualiser les images représentatives de tous les segments qui ont été classifiés comme PERSONNE, classés par durée d’apparition décroissante. Ceci permet à un utilisateur d’avoir rapidement une vue d’ensemble du contenu de tous les enregistrements.

6 Travaux antérieurs

De nombreux systèmes d’annotation de vidéo ou de documents multimédia ont déjà été proposés. Toutefois, ils diffèrent généralement des Outils d’Annotation que nous proposons car ils adressent d’autres besoins que l’annotation en volume. Certains systèmes se concentrent sur les aspects d’utilisabilité [11] [15], pour donner à un utilisateur un moyen commode d’indexer certains événements dans des enregistrements vidéo, puis d’y accéder rapidement. D’autres systèmes s’intéressent plus aux problèmes d’architecture liés à l’implantation de ces mécanismes [7] [12], par exemple pour les questions

de stockage et de synchronisation. De nombreux systèmes utilisent déjà de la vidéo annotée pour l'apprentissage et l'évaluation, mais la question des outils d'annotation et de leur efficacité est rarement abordée.

Les Journaux Télévisés sont une source intéressante de documents multimédia, et de nombreux projets se sont intéressés à différentes approches pour leur traitement automatique [6], que ce soit en utilisant les sous-titres [2] [5], la reconnaissance de parole [9] [13], ou encore l'analyse de la vidéo [16] [1].

La partie d'analyse présentée dans cet article est une amélioration de l'approche proposée par Zhang.

7 Conclusion

Nous avons présenté une approche de l'indexation vidéo basée sur la construction d'un corpus de vidéo annotée, qui est ensuite utilisé pour l'apprentissage et l'évaluation d'algorithmes d'analyse de Journaux Télévisés. Nous pensons qu'une telle approche basée sur les données, qui s'est déjà montrée fructueuse dans différents domaines, est également appropriée pour l'indexation vidéo. Pour la mettre en oeuvre, il faut disposer d'outils performants d'annotation, permettant de traiter rapidement de grands volumes de données. Nous avons proposé quelques exemples de tels outils, et montré comment leurs résultats sont utilisés dans le cadre d'une analyse automatique de Journaux Télévisés.

8 Références

[1] P. Aigrain, H. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications*, 3(3):179–202, November 1996.

[2] D. R. Bacher and C. J. Lindblad. Content-based indexing of captioned video on the viewstation. Technical report, Massachusetts Institute of Technology, TNS, October 1995.

[3] Gerard Benedetti, Benoit Bodin, Franck Lhuisset, Olivier Martineau, and B. Merialdo. A structured video browsing tool. In Leonard J. Bass and Claude Unger, editors, *Engineering for Human-Computer Interaction*, pages 17–26. Chapman & Hall, 1996.

[4] L. Breiman, J. H. Friedman, R.A. Olsen, and C. J. Stone. *Classification and regression trees*. The Wadsworth Statistical Probability series. Wadsworth, 1984.

[5] M. Brown, J. Foote, G. Jones, K. Sparck-Jones, and S. Young. Automatic content-based retrieval of broadcast news. *ACM Multimedia Conference*, November 1995.

[6] C. L. Compton and P. D. Bosco. Internet CNN NEWSROOM: A digital video news magazine and library. In *International Conference on Multimedia Computing and Systems*. IEEE Computer Society, May 1995.

[7] J. Gabbe, A. Ginsberg, and B. Robinson. Towards intelligent recognition of multimedia episodes in real-time applications. In *Proceedings of the ACM Multimedia Conference*, 1994.

[8] J. S. Garofolo. *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[9] P. Gelin and C. Wellekens. Keyword spotting for video soundtrack indexing. In *Proceedings of the IEEE ICASSP96*, 1996.

[10] D. Harman. Overview of the first TREC conference. In *Proceedings of the Sixteenth ACM SIGIR*, pages 36–47, 1993.

[11] B. Harrison and R. Baecker. Designing video annotation and analysis systems. In *Proceedings of Graphics Interface '92*, pages 157–166, May 1992.

[12] R. Hjelsvold, S. Langorgen, R. Midtstraum, and Olav Sandsta. Integrated video archive tools. *ACM Multimedia Conference*, November 1995.

[13] G. Jones, J. Foote, K. Sparck-Jones, and S. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the Nineteenth ACM SIGIR Conference*, pages 30–39, New York, August18–22 1996.

[14] M. Marcus, B. Santorini, and M A. Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational linguistics*, 19:313–330, 1993.

[15] K. Weber and A. Poon. Marquee: A tool for real-time video logging. In *Proceedings of ACM CHI'94*, volume 2, page 203, 1994.

[16] H. Zhang, S. Yeo Tan, S. Smoliar, and G. Yihong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2:256–266, 1995.