Navigation par le contenu dans les vidéos avec des Tables Graphiques des Matières

Hervé Goëau[1,2], Jérôme Thièvre[1], Marie-Luce Viaud[1], Denis Pellerin[2]

[1]Institut National de l'Audiovisuel (INA)
4, avenue de l'Europe, 94366 Bry-sur-Marne Cedex, France

[2]Grenoble Images Parole Signal Automatique (GIPSA-lab)(ex LIS)
46, avenue Félix Viallet, 38031 Grenoble, France

hgoeau@ina.fr, jthievre@ina.fr
mlviaud@ina.fr, denis.pellerin@lis.ingp.fr

Résumé − Nous présentons une visualisation interactive pour parcourir les vidéos composites tel que les journaux télévisés, les magazines ou certaines émissions sportives. Nous proposons une visualisation, qui, une fois connectée avec un lecteur multimédia standard, offre une solution de parcours rapide de ce type de vidéo. Le système s'appuie sur une vue globale révélant la structure temporelle de la vidéo, permettant ainsi un accès non linéaire au contenu en suggérant des images pertinentes. Notre application est issue d'une plateforme graphique d'expérimentation dédiée au calcul de similarité et au regroupement de contenu visuel autour d'une interface 2D pour un retour utilisateur. L'outil présenté ici est une des premières réalisations et promet un certain potentiel.

Abstract − We present an interactive visualization for browsing structured TV programs such as news, magazines or sports. This visualization, connected with a classical media player, offers a very handy video browser. This system allows a global overview by showing the temporal structure and by giving some semantic information. The drawn structure enables a non linear video access by suggesting relevant frames. This tool is created from a graphic framework designed for computing similarities on visual content, and displaying the associated proximities in a 2D map with graph representation. It is one of its first applications and shows interesting capabilities.

1 Introduction

Une des missions principales de l'Institut National de l'Audiovisuel est de collecter les programmes audiovisuels français et de renforcer leurs accessibilités. La segmentation et l'annotation manuelle d'émission constitue une part importante de l'activité des documentalistes et peut être très coûteuse en temps. Nous nous sommes intéressés dans ce papier au parcours de vidéo structurées. Nous entendons par « structuré » une alternance de type plateau-reportage (journaux télévisés, magazines) ou une répétition d'un contenu visuel similaire comme par exemple la succession d'épreuves d'une discipline sportive par différents athlètes. Le but est d'aider les documentalistes à se repérer rapidement pour effectuer leurs tâches de documentation ou de consultation. Après un bref état de l'art exposant notre positionnement, nous décrirons la plateforme d'expérimentation mise en place pour analyser et décrire des vidéos dans une perspective de visualisation. Ce système permet d'élaborer une application de navigation par le contenu interactive appelée Table Graphique des Matières (TGM) qui, une fois connecté avec un lecteur standard multimédia, offre une solution efficace pour survoler le contenu. La méthode proposée est illustrée par une visualisation typique sur un journal télévisé.

2 Etat de l'art et positionnement

Si les lecteurs multimédia intègrent tous aujourd'hui une barre de lecture et les touches de contrôle héritées des magnétoscopes, ils ne permettent pas une navigation précise : le retour visuel est saccadé et désagréable, l'utilisateur a tendance à se perdre dans le contenu notamment pour de longues vidéos. Pour palier à ce problème d'échelle, il est possible de modifier le pas de défilement des images avec un mouvement vertical de la souris [1]. Cette navigation locale peut être complétée par un aperçu global du contenu visuel, la solution la plus basique étant de faire apparaitre dans une frise des images extraites régulièrement (à la manière des logiciels de montage vidéo). Pour mieux couvrir le contenu sémantique, il est préférable de sélectionner des images clés représentatives de la « nouveauté » visuelle, en s'appuyant sur des courbes d'activité de caractéristiques bas-niveaux ou en se basant sur un modèle d'attention visuel [2]. La juxtaposition d'imagettes donne ainsi un résumé vidéo rappelant un « storyboard » épuré. La mise en forme des imagettes peut être valorisée par de l'appariement d'images, ou « mosaicking », pour condenser un plan séquence possédant un mouvement de caméra ample en une seule image globale. Pour des longs plans statiques, les événements peuvent être détectés et superposés en une seule image [3]. Dans un autre registre, l'étude de la structure narrative de la vidéo permet de détecter des instants clés et de proposer

des points d'entrées pertinents du point de vue sémantique pour une navigation non linéaire. Il est ainsi possible d'aligner le contenu audiovisuel avec une forme narrative par une structuration multimodale par modèles de Markov cachés comme dans [4] où les différentes phases d'un match de tennis sont identifiées et mises en forme dans une table hiérarchique. D'une manière plus générique, il est possible d'exploiter des mesures de similarité afin de passer d'une unité technique (le plan) vers une unité sémantique (la scène). Des algorithmes de regroupement de plans sont alors appliqués sur des caractéristiques bas-niveau. Ces clusters peuvent être validés par un utilisateur à travers une visualisation 2D s'appuyant sur des techniques de réduction de dimensionnalité (MultiDimensional Scaling ou MDS) [5]. En appliquant des contraintes temporelles sur ces regroupements, il est possible de formaliser l'enchainement des plans et de proposer un parcours de lecture à travers un graphe de transition [6]. Nous reprenons dans nos travaux l'idée de faire émerger à partir d'une analyse bas-niveau la structure d'un programme vidéo au sens du montage. Notre proposition contourne le problème délicat de la segmentation en plans, et consiste plutôt à travailler sur une frise d'images extraites à intervalles réguliers que nous restructuront en mêlant approches MDS et technique de visualisation de graphes.

3 La plateforme d'expérimentation

La plateforme d'expérimentation (figure 1) est conçue pour exploiter et mettre en valeur des données calculées sur du contenu visuel. Une cellule de calcul extrait des descripteurs standards en couleur, texture, mouvement, puis les compare pour établir des matrices de similarités. Ces données sont transformées en une représentation sous forme de graphe où chaque noeud symbolise une image clé et où les arêtes, multivaluées, contiennent des distances déterminées à partir des mesures de similarité. Un graphe peut ainsi être chargé dans l'interface graphique et mettre en perspective les données. L'idée centrale est de placer l'utilisateur expert au coeur du système. Son interaction va permettre de déterminer un profil type correspondant à une tâche applicable à un corpus homogène de vidéos : journaux télévisés (JT) d'une même période, même émission, même type de captation de sport, etc. La présence de l'expert se justifie par le fait qu'il paraît très difficile vis-à-vis de l'état de l'art de penser une technologie totalement automatique et suffisamment générique pour répondre à la diversité des vidéos présentes à l'INA. Son action intervient sur des opérations de placements de graphe, de réduction de dimensionnalité, de filtrage, de combinaison de distances et de regroupements d'images.

4 Table Graphique des Matières - TGM

L'idée est d'organiser une succession d'imagettes extraites à intervalles réguliers de manière à proposer une visualisation présentant une mise en forme de la structure

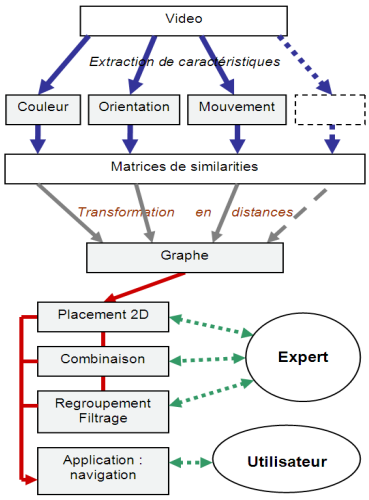

FIG. 1 – La plateforme d'expérimentation.

vidéo, au sens du montage. La manipulation effectuée revient à tordre un fil, ou plutôt un chapelet d'images successives, de manière à faire apparaitre une forme visuelle 2D en adéquation avec la structure du document. La méthode commence un regroupement d'images correspondant à la « colonne vertébrale » du document : ce sont dans la plupart des cas les images du plateau ou du présentateur, ou encore un habillage graphique. Il s'agit donc de les distinguer des inserts tel que les reportages, extraits de films, publicités ...L'expert visualise les imagettes dans le plan et teste différentes distances à travers le retour donné par un algorithme de réduction de dimensionnalité. Le MDS proposé s'appuie sur un algorithme simulant un modèle de force [7]. Chaque noeud/image représente une masse et la distance donnée par une arête induit des forces d'attraction entre chaque couple d'images. La minimisation de l'énergie du modèle converge vers un placement reflétant la proximité existant dans l'espace des caractéristiques. Les figures 2 et 3 illustrent la formation d'un TGM pour un JT. Dans cet exemple, la couleur du décor du plateau est très discriminante vis-à-vis des reportages. L'expert a donc intérêt à appliquer dans le MDS une distance sur des descripteurs couleur. La première figure donne un aperçu sur le jeu d'imagettes extraites toutes les 10 secondes. La distance euclidienne entre des histogrammes couleur avec 64 classes dans l'espace LUV (signature compacte et rapide à calculer) fait émerger deux clusters « présentateur » avec ou sans encart. Ces regroupements sont identifiés par un algorithme agglomératif hiérarchique [8] favorisant l'apparition de clusters de forme arbitraire.

Pour créer la vue finale, 2 contraintes temporelles filtrent les arêtes : ne sont gardés que les liens dans l'ordre chronologique à l'intérieur du ou des clusters désignés comme faisant partie de la colone vertebrale de la vidéo d'une part, puis sur l'ensemble des arêtes d'autre part. Il en résulte un graphe épuré et constitué d'un axe central, la colonne vertébrale de la vidéo, avec des ramifications en forme de boucles. Pour finir, un dernier placement topologique met en forme le graphe avec des contraintes pour permettre une lecture du temps sur l'axe horizontal de la gauche vers la droite et développer les boucles de part et d'autre de la

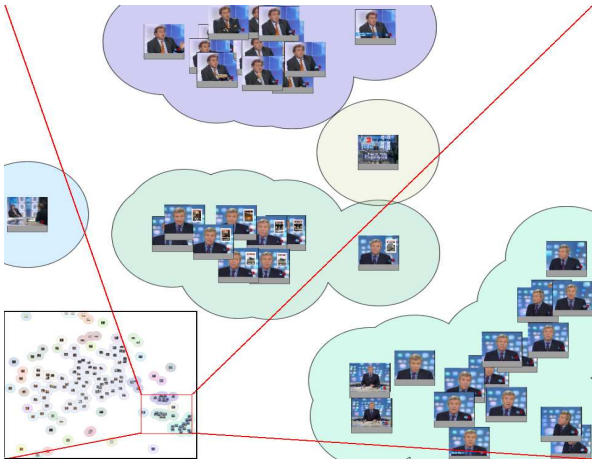

FIG. 2 – Résultat du MDS appliqué selon la distance euclidienne entre histogrammes 64 LUV sur des images extraites toutes les 10 secondes d'un journal télévisé d'une durée de 28 minutes environs.

FIG. 3 – TGM sur le même JT fig.2. La vue révèle la structure narrative de la vidéo : chaque boucle représente un reportage ou un plan d'une interview plateau. La longueur des boucles nous renseigne sur la durée des reportages.

colonne vertébrale 3. La figure 4 indique le sens de lecture d'une TGM. La frise temporelle ainsi réorganisée reflète la structure narrative et donne un aperçu global du contenu. L'interface zoomable et cliquable permet de choisir une zone d'intérêt, et d'accéder localement et non linéairement au contenu. Dans cet exemple (3 chaque boucle est un reportage ou un plan interview, et nous pouvons compter rapidement leur nombre. Nous pouvons évaluer par la longueur des boucles la durée des reportages et repérer très rapidement quelles actualités dominent ce jour de diffusion. Par exemple la $4^{\text{ième}}$ boucle, nettement plus longue nous indique qu'un événement particulier domine l'actualité. La couleur nous donne également quelques renseignements : l'apparence bleutée autour de la $10^{\text{ième}}$ boucle désigne des plans interview/plateaux. Les couleurs chaudes en fin de JT connotent un ton plus léger des thèmes abordés.

Le choix du descripteur est important : par exemple,

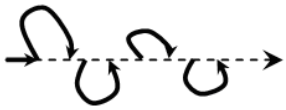

FIG. 4 – La lecture temporelle d'une TGM : l'axe central représente la colonne vertébrale et se lit de la gauche vers la droite. Une boucle représente un insert au montage.

dans ce précédent JT, l'expert aura intérêt à privilégier la distance couleur. Mais pour des émissions sportives, le choix de la caractéristique mouvement sera plus approprié (illustration dans la figure 5).

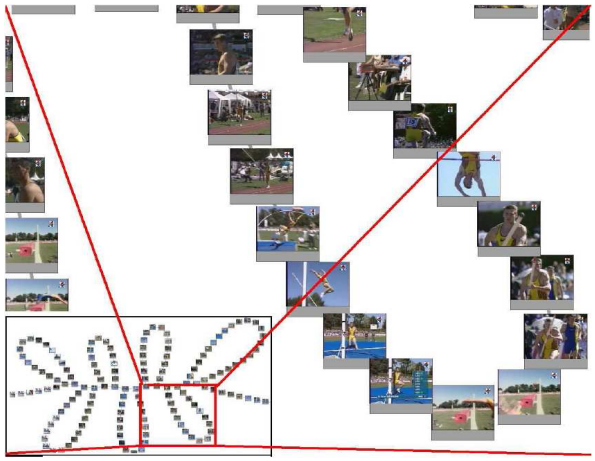

FIG. 5 – Zoom sur une boucle d'un TGM analysant une suite de 7 épreuves sportives. Le rapprochement des séquences est donnée à partir d'une inter corrélation entre signaux représentants la translation horizontale issue d'une estimation de mouvement affine.

5 Evaluation

Les outils de navigation sont difficiles car cela nécessite de bien recenser tous les biais relatifs à la présence d'un utilisateur. Nous avons effectué une précampagne d'évaluation des TGM sur 2 JT (figures 3 et 6) sur un panel de 12 utilisateurs non professionnels de la documentation. Nous avons opté pour un approche comparative avec une visualisation plus compacte et basique sous forme de grille (figure 7). L'évaluation se décompose en 3 phases : des tâches chronométrées de recherche d'image (RI) par le contenu, un questionnaire où une note est attribuée sur les aspects graphiques et de compréhensions des vues, et enfin une interview sur les impressions des utilisateurs. Chaque nouvel utilisateur teste les 2 documents, une seule vue par JT, et à chaque individu, l'ordre de passage est inversé.

Les résultats affirment que, pour les tâches de RI, 50% des utilisateurs sont plus rapides avec les TGM contre 25% plus lent et 25% indifférents. En moyenne, le temps de recherche est de 12 secondes contre 15 pour les vues en grille. Le questionnaire révèle des différences notables, confirmé par le test de Wilcoxon ($p<0.1$), sur l'appréciation des utilisateurs. Un point important révélé par le question-

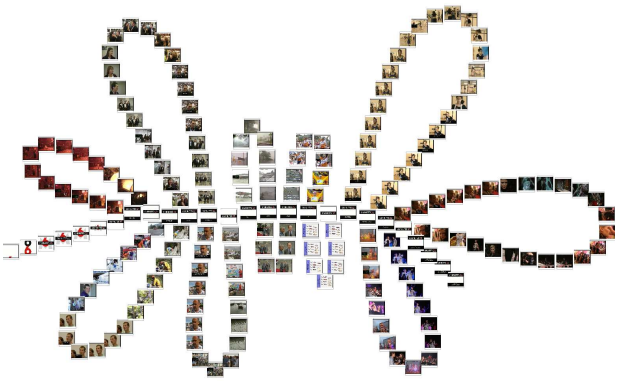

FIG. 6 – TGM d'un JT format court : une distance somme pondérée normalisée de distances entre histogrammes orientation et couleur permet de faire émerger un cluster plateau contenant uniquement des images d'habillage vidéo.

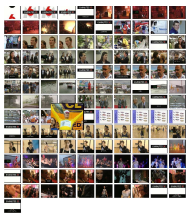

FIG. 7 – Visualisation basique du même JT fig.6

naire est notamment la capacité à revenir et localiser un contenu déjà apperçu dans les TGM au cours d'une précédente tâche.

6 Conclusion

Les TGM sont une première application de notre plateforme d'expérimentation et possèdent un réel intérêt auprès des documentalistes de l'INA pour le travail d'annotation manuel. Elles peuvent être vues comme une signature visuelle de la forme narrative d'un genre vidéo, et permettent des comparaisons rapides entre documents d'un même corpus.

Si la méthode nécessite l'intervention d'un utilisateur expert pour l'analyse de contenu du document, une méthode moins supervisée dans le cas spécifique des JT est en cours d'étude.

Les premières évaluations confirment l'appréciation par les utilisateurs des TM pour appréhender le contenu de la vidéo. La structuration visuelle permet notamment de se repérer plus facilement sur des zones déjà explorée lorsque l'on revient sur le document. Nous avons également intégré dans notre système une visualisation alternative suggérée par quelques utlisateurs intermédiaire entre les TGM et les grilles basiques 8. Toutes ces visualisations tentent de répondre à cette problématique : comment exploiter au mieux l'encombrement d'une surface 2D fixée pour ordonner dans l'espace une suite d'images d'une vidéo tout en

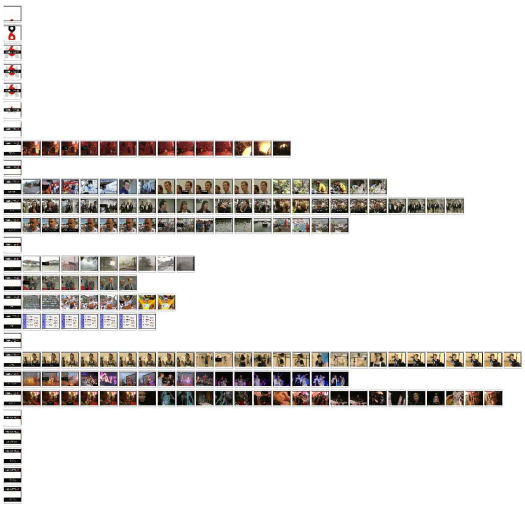

FIG. 8 – Visualisation alternative du TGM du JT fig.6

révélant la structure narrative ? Une étude plus approndie peut être envisagée sur ces problèmes de mise en page.

Références

[1] W. Hurst and P. Jarvers. *Interactive, Dynamic Video Browsing With The Zoomslider Interface.* IEEE ICME, 2005.

[2] M. Guironnet, N. Guyader, D. Pellerin, P. Ladret. *Static and dynamic feature-based visual attention model : comparison with human judgement.* EUSIPCO, 2005.

[3] C. Pal and N. Jojic. *Interactive Montages of Sprites for Indexing and Summarizing Security Video.* IEEE CVPR, 2005.

[4] E. Kijak and G. Gravier and L. Oisel and P. Gros. *Audiovisual integration for tennis broadcast structuring.* Multimedia Tools and Applications, 2006.

[5] M. Campanella and R. Leonardi and P. Migliorati. *Future Viewer : an efficient Framework for navigating and classifiying audio-visual document.* IEEE ICME, 2005.

[6] M. M. Yeung and B. L. Yeo. *Time-constrained clustering for segmentation of video into story units.* IEEE ICPR, 1996.

[7] T. M. J. Fruchterman and E. M. Rheingold. *Graph drawing by forcedirected placement* Software - Practice and Experience, 1991.

[8] S. Guha and R. Rastogi and K. Shim. *CURE : an efficient clustering algorithm for large databases* ACM SIGMOD, 1998.