SESAME : modèle conceptuel de description de documents audiovisuels et assistants d'annotation de contenu *

Yannick Prié[1] Jean-Michel Jolion[2] Alain Mille[3]

[1] LISI INSA-Lyon / Yannick.Prie@insa-lyon.fr
[2] RFV INSA-Lyon / jolion@rfv.insa-lyon.fr
[3] LISA CPE-Lyon / am@cpe.fr

1 INTRODUCTION

Sesame (Système d'Exploration de Séquences Audiovisuelles et Multimédias enrichi par l'Expérience) est un projet rassemblant plusieurs équipes de différents laboratoires (LISI et RFV de l'Insa-Lyon, LIP de l'Ens-Lyon et LISA de Cpe-Lyon) autour de la problématique générale de l'accès à des séquences audiovisuelles (SAV) par le contenu pour différentes tâches comme la recherche, l'analyse, la manipulation, le montage, *etc*. Le groupe de recherche constitué autour de SESAME s'est fixé comme objectif de prendre en compte l'ensemble du problème comme un tout indissociable : modèle de description de morceaux d'audiovisuel, outils d'assistance, modèle de base de données audiovisuelles, architecture parallèle pour les serveurs, *etc*. Cette approche globale facilite la prise en compte mutuelle des contraintes des différents aspects du problème. L'ambition de maquetter par morceaux les différents éléments de SESAME puis d'intégrer les résultats sur un démonstrateur (sous la responsabilité du CISM Lyon1-INSA) est raisonnable compte tenu des partenariats installés avec l'INA d'une part et FRANCE 3 d'autre part.
La communication se focalisera sur quelques aspects particulièrement critiques pour la construction globale de SESAME : un modèle de description de morceaux audiovisuels, des mécanismes d'annotation automatiques ou semi-automatiques permettant de rendre cette description accessible, avant d'aborder la notion plus générale d'assistant de description de documents audiovisuels.

2 LE MODELE DES STRATES-IA

Nous présentons ici rapidement le modèle des *Strates Interconnectées par les Annotations* (pour plus de détails, voir [7]). Nous considérons que quelles que soient les tâches liées à l'utilisation d'un système de recherche d'information audiovisuelle, *indexation* bien sûr,

*Ce travail est en partie financé par France Télécom (CNET/CCETT), contrat de recherche $n°$ 96ME17.

mais aussi *recherche*, *analyse* ou encore *montage*, il s'agit en fait de différents raffinements d'une tâche fondamentale de *description* du contenu de morceaux de documents audiovisuels (tels qu'on les indexe, tels qu'ils existent, ou tels qu'on les arrange en nouveaux morceaux). Pour les tâches d'indexation-recherche en particulier, l'utilisateur doit décrire de manière plus ou moins complète une séquence audiovisuelle, ce qui exige un support de description générique suffisamment riche. Cette modélisation orientée utilisateur est également un prérequis absolu à toute tentative de proposition d'assistance à la tâche de description (qu'elle soit pour l'indexation ou pour la recherche). Considérant de plus certaines spécificités inhérentes au médium audiovisuel (temporalité, construction par le montage, importance du contexte), il est nécessaire de disposer d'un modèle aux larges possibilités d'expressivité et de structuration, dont la complexité de description puisse être variable, qui prenne en compte la dynamique temporelle et permette de considérer des relations contextuelles.

Le principe fondamental de la description de documents audiovisuels est l'*annotation*, qui consiste à attacher à un *morceau* de document une description à base de *caractéristiques* plus ou moins organisées. Les caractéristiques peuvent aller du plus haut niveau conceptuel (noms de personnages, moments narratifs) au plus bas (histogrammes de couleurs, résultats de calculs) auquel cas on privilégiera le terme de *primitives*. Nous regroupons en *dimensions d'analyse* les différentes caractéristiques qui peuvent être repérées suivant une même analyse, que celle-ci soit automatique ou non. Par exemple des dimensions d'analyse automatiques permettront de repérer des plans, ou de distinguer scène de jour ou de nuit, ou encore voix, musique et bruitages ; une dimension d'analyse semi-automatique de nommer des objets importants à l'écran ; et une totalement manuelle de repérer des ambiances (joyeux, glauque, *etc*.).

Attacher une annotation qui représente une caractéristique à un morceau de document audiovisuel peut se faire de deux manières : soit on annote, on décrit des morceaux (un document entier, un plan) déjà déterminés et nommés, c'est l'approche de *segmentation* [10] [2], soit on considère que c'est la présence de l'annotation elle-même qui désigne et découpe le morceau : c'est l'approche de *stratification* [3]. Nous nous plaçons dans le cadre de cette dernière approche, en considérant que *tout* ce qu'il est possible de dire et de calculer sur un document audiovisuel peut-être considéré comme une annotation d'une strate. Le processus d'annotation fonde alors le découpage d'un document audiovisuel en strates, une strate accèdant à l'existence parce qu'au moins un *Elément d'Annotation* (EA) — qui est le représentant informatique d'une caractéristique, par exemple $\langle Plan \rangle$, $\langle Mouvement\_camera \rangle$ — l'annote. Nous appelons *Unité Audiovisuelle* (UAV) le représentant d'un morceau de document, caractérisé par son document fichier-source et sa position temporelle dans ce document (t_1 et t_2). Un ou plusieurs EA peuvent être en *relation d'annotation* R_a avec une UAV. Un EA a un nom, mais possède également des attributs qui viennent le préciser. Par exemple, un EA $\langle Plan \rangle$ peut avoir pour attribut une image caractéristique qui le résume. Un EA $\langle Dialogue \rangle$ une transcription des paroles échangées dans l'UAV qu'il annote, *etc*. Ces principes sont illustrés figure 1.

Afin d'enrichir la description, les éléments d'annotation peuvent être connectés entre eux *à la manière des annotations* à l'aide d'EA de relation, ceci étant valable aussi bien à l'intérieur d'une UAV qu'entre EA annotant des UAV différentes. On pourra par exemple établir une liaison entre un EA $\langle Plan \rangle$ et un EA $\langle Document \rangle$ à l'aide d'un

EA $\langle Contenu\_dans \rangle$ et de deux relations dites élémentaires R_e. De la même manière il sera possible de relier un EA $\langle Voiture \rangle$ à un EA $\langle Forme \rangle$ dont les attributs donneront les caractéristiques de positionnement à l'image, et ce à l'aide de l'EA $\langle Etiquette \rangle$, voir figure 1.

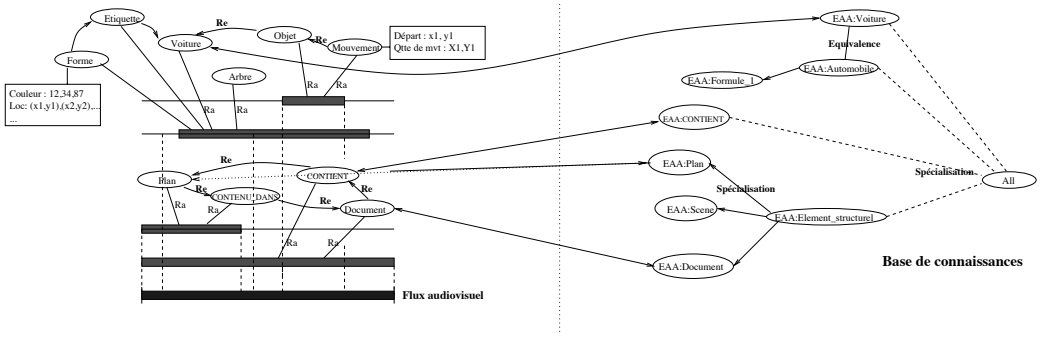

FIG. 1 – *A gauche : unités audiovisuelles annotées, attributs d'EA et liens entre éléments d'annotation. A droite, base de connaissances.*

L'ensemble des EA utilisables fait partie d'un vocabulaire contrôlé à l'aide d'une Base de Connaissances organisée en réseau sémantique d'*Eléments d'Annotation Abstraits* (EAA) dont les EA sont les instances et de relations de concept entre ces EAA. L'EAA $\langle EAA : Plan \rangle$ est ainsi l'abstraction de l'EA $\langle Plan \rangle$, et est par exemple une spécialisation d'un EAA $\langle EAA : Element\_structurel \rangle$, *cf.* figure 1. La figure 2 présente l'ensemble des bases du système et le processus d'annotation.

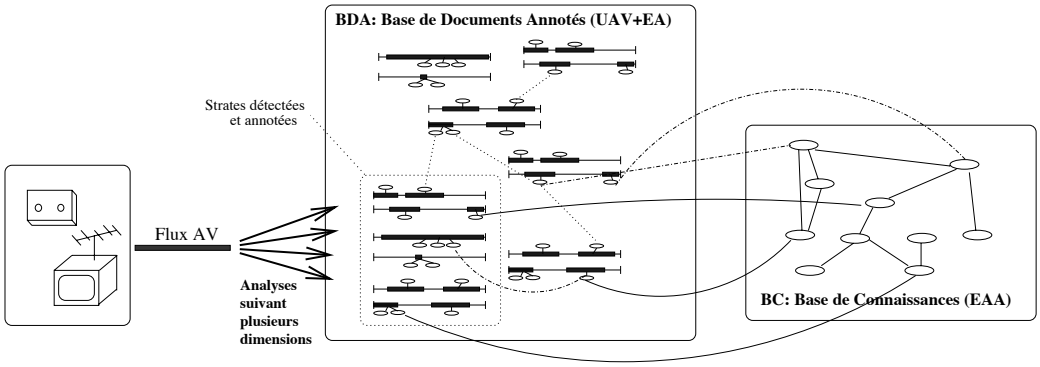

FIG. 2 – *Annotation de documents : un flux est analysé et décrit, puis se fond dans la* Base des documents annotés. *Les EA utilisés sont liés à la base de connaissances*

L'approche de description en Strates Interconnectées par les Annotations permet : la description d'un document audiovisuel avec un degré de granularité quelconque ; l'intégration dans un même schéma de toutes les caractéristiques possibles et l'étude du passage des primitives calculées au caractéristiques de niveau conceptuel (*cf.* [1]) ; enfin la mise en place d'une notion de contexte audiovisuel. En effet, on peut définir des liens contextuels suivants, d'une part, les relations temporelles entre strates, mais aussi considérer un

contexte *conceptuel* basé sur les liens explicites entre éléments d'annotation. Par exemple pour une UAV annotée par l'EA $\langle Zoom\_avant \rangle$ en relation avec un EA $\langle Voiture \rangle$ (figurant le focus du zoom) annotant une autre UAV, on pourra considérer que la seconde UAV est dans un contexte de la première et utiliser cette relation.

3 L'ANALYSE VIDEO COMME SUPPORT D'ANNOTATION POUR CERTAINES DIMENSIONS D'ANALYSE

Les EA couvrent une palette de niveaux d'abstraction très étendue : depuis les caractéristiques statistiques des images jusqu'aux concepts les plus généraux. Les EA de niveau d'abstraction élevé (correspondant à l'interprétation d'un opérateur) sont l'apanage de l'homme par définition, les EA les plus concrets peuvent et doivent faire l'objet d'une élaboration la plus automatique possible. L'analyse des flux vidéo permet de générer directement des EA mais fournit également à l'opérateur qui essaie de décrire un morceau de vidéo une assistance efficace et interactive.
Les dimensions d'analyse que nous avons mises en évidence à ce stade du projet se répartissent en trois classes selon la nature des EA retournés et leur degré d'automaticité :

- soit des informations liées à la nature de la SAV (information proche du signal, description structurelle) dont on peut envisager l'automaticité : détection de plans et d'effets de montage, analyse des mouvements de la caméra, informations statistiques (colorimétrie globale, indices d'information et de texture)...
- soit des informations liées au contenu sémantique de la SAV (information plus conceptuelles, donc plus proche du script) dont il faut prévoir uniquement une semi-automaticité voire une exécution à la demande : détection et analyse des mouvements (hors caméra), analyse des informations textuelles incrustées dans la vidéo (génériques, sous-titres, noms des présentateurs), analyse et suivi de personnages et de visages.
- soit des informations liées à l'accès à la SAV (anticipation sur la phase de recherche, spécificité thématique) en mode (semi) automatique : construction d'une vignette représentative d'un plan, extraction de points d'intérêt pour les requêtes par l'exemple.

Nous allons maintenant détailler quelques-unes de ces dimensions d'analyse.

La détection des «cuts». Un cut, du point de vue de l'analyse d'images, est une rupture dans les caractéristiques principales du signal. L'objectif est de fournir rapidement une localisation de ces cuts. Nous utilisons une méthode tout à fait classique qui combine, pour deux images consécutives, deux mesures : une différence entre les histogrammes de luminance et une mesure d'énergie sur la différence entre les chrominances (en effet, un histogramme de chrominance est peu informatif). Cet algorithme a la particularité de ne pas utiliser toute l'image mais uniquement une version réduite en taille et information, déduite du codage MPEG (nous ne décodons que la composante DC ce qui conduit à une imagette environ 8 fois plus petite que l'image initiale). Elle est donc bien adaptée au codage actuel — compressé — des flux vidéos numériques et valide en cela une contrainte

FIG. 3 – *Visage avec points d'intérêt en superposition sous forme de croix*

imposée par nos partenaires. Cette méthode a un comportement stable même si elle laisse apparaître des différences selon la nature des images (I, B ou P).

La détection des mouvements. Nous avons proposé d'analyser une séquence spatio-temporelle dans son intégralité en représentant celle-ci, non plus comme une simple succession d'images, mais comme un signal tridimensionnel. La transformée de Fourier 3D permet de faire cette analyse globale et d'obtenir une information très riche. Ceci est rendu possible par la nature *a posteriori* des traitements qui autorise l'emploi de méthodes non causales comme la transformée de Fourier. De par le caractère global de cette dernière, les résultats sont plus stables et plus performants que ceux obtenus par les approches classiques reposant sur la mesure de flux optique [5].
Nous avons ainsi mis en évidence des moyens simples de repérage des éléments du mouvement de la caméra (translation, rotation). Grâce à cette estimation du mouvement induit par la prise de vue, on peut tout à fait extraire les objets en mouvement propre et par exemple faire une séparation entre l'avant- et l'arrière-plan [6]. Cette analyse participe donc à l'enrichissement de la description d'une UAV.

Elaboration de points d'intérêt. Le volume de données correspondant à un flux vidéo est souvent un obstacle important pour tous les traitements et caractérisations relevant du domaine du traitement d'images. C'est pourquoi une piste prometteuse de recherche préconise de résumer une image (*cf.* figure 3) à un ensemble de *points d'intérêt*, chaque point étant associé à un vecteur de caractéristiques (le plus souvent des invariants différentiels [9]). Nous avons conçu un nouveau détecteur qui s'appuie sur une mesure multi-échelles d'une énergie de contraste. Le contraste étant une notion plus pauvre que le classique contour, ce détecteur est plus riche dans le panel des points extraits et nous avons montré sa meilleure résistance au bruit de codage (par rapport au détecteur de Plessey et celui du projet SUSAN) induit par l'emploi de techniques de compression telles que Jpeg [4].
Il s'agit pour le moment d'une approche plus image que vidéo et particulièrement bien adaptée à la recherche par exemples. Cependant, dans le cadre d'une vidéo, ce détecteur peut également permettre de construire une image résumé d'une séquence. En effet, dans une séquence donnée, l'image résumée sera celle contenant le plus de points d'intérêt.

4 L'ANALYSE VIDEO COMME ASSISTANT PRIVILÉGIÉ POUR LA DESCRIPTION DE DOCUMENTS AUDIOVISUELS

La description de contenu constitue la base de la plupart des tâches associées à la gestion et l'exploitation de fonds documentaires audiovisuels. Sans vouloir être exhaustif, les tâches suivantes illustrent bien ce constat :

- *l'indexation primaire (ou initiale)* consiste à décrire le document en STRATES-IA correspondant aux objets d'intérêt repérés par l'utilisateur documentaliste.
- *la recherche* consiste à décrire les STRATES-IA que l'on aimerait retrouver comme descripteurs des morceaux audiovisuels souhaités. C'est une sorte «d'exemple» de ce que l'on aimerait retrouver qui est donc ainsi élaboré.
- *l'analyse* consiste à rechercher les morceaux audiovisuels qui «résonnent» avec les STRATES-IA du document analysé. De manière imbriquée, on se retrouve dans le même type de tâche que précédemment.
- *la navigation* consiste à glisser d'une STRATE à une autre (que ce soit intra-documentaire ou inter-documentaire). Il s'agit encore de décrire le morceau que l'on souhaite atteindre.
- *monter* ou *construire* un document audiovisuel peut se concevoir comme une description «à l'avance» des morceaux que l'on souhaite associer. Chacune de ces descriptions peut alors être utilisée (seule ou en relation avec les autres) pour chercher les morceaux qui pourraient convenir.

Les tâches ainsi énoncées sont génériques. Elles seront bien sûr combinées pour le besoin spécifique de chaque type d'utilisateur (documentaliste, archiviste, journaliste, publiciste, *etc*.). Trois niveaux de complexité peuvent être identifiés :

- au niveau de la tâche globale de l'utilisateur, pour réaliser cette tâche selon des règles ou selon un cadre défini ;
- au niveau de la description en tant que telle, pour maîtriser l'extension et la portée des graphes associés ;
- au niveau des annotations elle-mêmes, très nombreuses et dont la sémantique doit être maîtrisée.

La figure 4 montre que ces niveaux d'assistance correspondent à des connaissances différentes : le modèle de tâche global est très spécialisé et lié à l'application mise en place sur le poste de travail ; la base de connaissances audiovisuelles, respectant le métamodèle des STRATES-IA supporte les opérations de mises en contexte qui guident les recherches dans le graphe des UAV ; enfin les connaissances opérationnelles pour l'extraction d'EA du flux sont encapsulées dans les méthodes des assistants d'analyse vidéo.
Ce découpage n'est pas étanche, et en particulier les modèles de plus haut niveau possèdent des connaissances sur ceux des niveaux inférieurs. En particulier, les EA résultat d'une analyse par l'assistant vidéo «connaissent» les paramètres exploités par l'assistant pour l'élaborer (il s'agit d'attributs de l'objet EA).

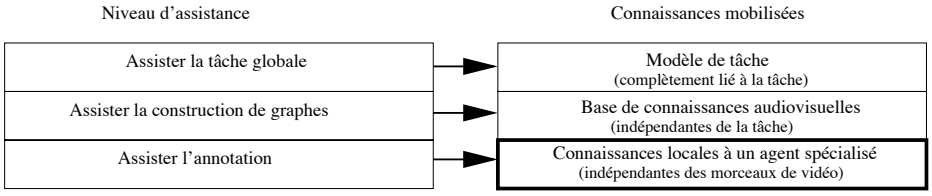

FIG. 4 – *Niveau d'assistances et type de connaissances associées*

La présentation d'un scénario d'utilisation donne une idée d'exploitation concrète d'un tel dispositif. L'audiovisuel est un support d'abord «visuel», et la description d'un morceau de document à partir d'un exemple est une démarche naturelle pour un utilisateur :

- l'utilisateur visionne un morceau de document audiovisuel qui est à sa disposition comme «contenant» un exemple de ce qu'il cherche.
- le morceau sélectionné est annoté comme pour l'indexer, essentiellement en exploitant l'*assistant vidéo* pour élaborer les EA «visuels» qui y sont détectés.
- ce graphe constitue une requête soumise à un *assistant exploitant de la base de connaissances audiovisuelles* pour (par exemple) exécuter un contrôle de pertinence en testant l'exemple sur une base locale retournant des UAV similaires. L'utilisateur a ainsi une estimation immédiate de la pertinence de son exemple sur la base de la visualisation des UAV considérées comme similaires dans la base locale (*cf.* [8]).
- après un éventuel affinement de la description de son exemple (cycle sur les deux premières étapes), l'*assistant d'élaboration de contextes* aide l'utilisateur à déterminer le «degré de voisinage» qu'il souhaite pour les UAV qui seront sélectionnées sur le serveur.
- l'utilisateur peut alors utiliser d'autres outils pour naviguer à partir des éléments de réponse fournis par le serveur.

Le caractère avant tout visuel et interactif de la démarche de l'utilisateur est ainsi complété efficacement par un dispositif de description adapté à la recherche de morceaux audiovisuels sur des serveurs distants gérant des volumes qui peuvent être gigantesques.

5 CONCLUSION

Les travaux de l'équipe SESAME portent leurs efforts essentiellement sur les assistants vidéo et sur les assistants à la tâche de description intégrant ainsi la caractéristique principale du support audiovisuel. L'exploitation et l'intégration des résultats de la recherche en analyse d'image et vidéo sont en effet nécessaires pour permettre de rendre compte facilement du contenu «mesurable» d'un document audiovisuel. Nécessaires, mais non suffisants, c'est leur mise en contexte par le support des STRATES-IA qui complète leur pouvoir d'expression intrinsèque pour les rendre immédiatement utiles et efficaces dans les tâches d'exploitation d'une base de documents audiovisuels.
D'autres sources d'annotation semi-automatique ou automatique sont exploitables, aux premiers rangs desquelles il faut distinguer les sources sonores, textuelles associées au

flux, ou documentaires classiques (les documents de production). Le même principe d'intégration de ces sources par des assistants à la tâche d'annotation laisse espérer la mise à disposition d'outils performants pour prendre en compte toute la complexité de l'audiovisuel dans des tâches utilisateurs qui réclament de la rapidité.
Ainsi, le modèle des STRATES-IA permet d'unifier le principe d'annotation et fournit une représentation facilitant la réutilisation de l'expérience des utilisateurs dans leurs tâches de description comme dans la maîtrise des assistants d'annotation semi-automatique. Capitaliser cette expérience est un enjeu important pour enrichir les connaissances exploitées par les assistants de description.

Références

[1] Chang (S.F.), Chen (W.), Meng (H.J.), Sundaram (H.) et Zhong (D.). – Videoq: An automated content based video search system using visual cues. *In : ACM Multimedia 97*. – Seattle, Nov. 1997.

[2] Corridoni (J. M.), Del Bimbo (A.), Lucarella (D.) et Wenxue (H.). – Multi-perspective navigation of movies. *Journal of Visual Languages and Computing*, vol. 7, 1996, pp. 445–466.

[3] Davis (M.). – Media streams: An iconic visual language for video annotation. *In : Proceedings of the 1993 IEEE Symposium on Visual Languages*. pp. 196–203. – Bergen, Norway, août 1993.

[4] Jolion (J.M.). – *Multiresolution Contrast Based Detection of Interest Points*. – Rapport technique nRR.98.02, RFV-INSA, 1998.

[5] Lebourgeois (F.) et Jolion (J.M.). – Application de la transformée de fourier tridimensionnelle à l'analyse de l'information spatio-temporelle dans les séquences vidéo. *In : CORESA 98*. – Lannion, Juin 1998.

[6] Lebourgeois (F.), Jolion (J.M.) et Awart (P.). – Toward a video description for indexation. *In : 14th Int. Conf. on Pattern Recognotion*. – Sydney, Australie, Aout 1998.

[7] Prié (Y.), Mille (A.) et Pinon (J.M.). – Un approche de modélisation de documents audiovisuels en strates interconnectées par les annotations. *In : Ingénierie des Connaissances 98*. – Nancy, Mai 1998.

[8] Rissland (E.) et Daniels (J.). – The synergistic application of CBR to IR. *Artificial Intelligence Review*, vol. 10, 1996, pp. 441–475.

[9] Schmid (C.) et Mohr (R.). – Local grayvalue invariants for image retrieval. *IEEE trans. on Patt. Anal. Mach. Intell.*, vol. 19, n5, 1997, pp. 530–535.

[10] Yeo (B.L.) et Yeung (M.M.). – Retrieving and visualizing video. *Communications of the ACM*, vol. 40, n12, Dec. 1997, pp. 43–52.