

Une approche de modélisation de documents audiovisuels en strates interconnectées par les annotations

An Approach for Modelling Audiovisual Documents with Annotations Interconnected Strata*

Yannick Prié¹

Alain Mille²

Jean-Marie Pinon¹

¹ INSA-Lyon

² CPE-Lyon

Laboratoire d'Ingénierie des Systèmes d'Information — LISI
Bât 404. INSA, 20 avenue Albert Einstein F69621 VILLEURBANNE
Yannick.Prie@insa-lyon.fr / am@cpe.fr / pinon@lisi.insa-lyon.fr

Résumé

Nous donnons tout d'abord quelques spécificités fondamentales de l'audiovisuel. Nous présentons ensuite une méthode d'analyse des modèles pour la recherche d'information audiovisuelle et les deux principales approches structurelles utilisées. Nous proposons une approche originale de modélisation audiovisuelle en strates interconnectées par les annotations. Un document s'analyse suivant des dimensions d'analyse permettant de repérer des objets d'intérêt qui peuvent être de tous types (structurels, conceptuels, ...). L'annotation est alors structurée à partir d'éléments d'annotation (EA) qui sont des termes représentant aussi bien des objets d'intérêt que des relations, une base de connaissances permettant de contrôler leur utilisation. Nous montrons ensuite comment il est possible de construire à partir de relations entre EA (et donc entre strates) des vues contextuelles de celles-ci, c'est-à-dire de les annoter à l'aide d'EA annotant au départ d'autres strates. Nous discutons pour finir quelques perspectives ouvertes par cette nouvelle approche.

Mots-Clef

Audiovisuel, approche de modélisation, strates, unités audiovisuelles, éléments d'annotation, annotation contextuelle

Abstract

In this paper we first present some fundamental aspects of audiovisual. We then present a method for analysing models for audiovisual information retrieval and the two major structural approaches used in related works. We propose a novel approach for modelling audiovisual documents with annotations interconnected strata. An audiovisual document can be analyzed through some analysis dimensions allowing the detection of objects of interest of all types (structural, conceptual,...). Then the annotation is structured upon annotation elements (AE), which are terms representing both objects of interest and relations, a knowledge base being used to control the necessary AE. We then show how we can build from relations between AE, hence between strata, contextual views of the latter, that is how to annotate strata using AE formerly annotating other strata. We finally discuss some perspectives opened by this new approach.

Keywords

Audiovisual, modelling approach, strata, audiovisual units, annotation elements, contextual annotation.

1 Introduction

Rendus possibles tout à la fois par le développement des capacités de traitement des machines, des moyens de stockage et des débits des réseaux informatiques, les systèmes de recherche d'information audiovisuelle se développent maintenant très rapidement. Destinés principalement au stockage, à l'indexation et à la re-

*Ce travail s'inscrit dans le cadre du projet Sesame financé par le CNET, consultation thématique 96-ME-17. This work is part of the Sesame project granted by the French Center for Telecommunication Research under contract 96-ME-17.

cherche de documents audiovisuels, ils seront de plus en plus utilisés par les organismes de préservation du patrimoine audiovisuel (INAtèque), les chaînes de télévision ou les fournisseurs de contenus audiovisuels aux professionnels voire aux particuliers (INA, autres). Ces systèmes de recherche d'information se caractérisent par la possibilité de mener des recherches par requêtes, par l'exemple, ou bien en naviguant dans la base des documents audiovisuels, et par l'accès direct au flux audiovisuel numérique (compressé en MPEG1 ou 2).

Un nécessaire lien se fait alors jour entre le modèle de *représentation* des documents audiovisuels choisi par les concepteurs, le choix de la *présentation* de celui-ci à l'utilisateur, et enfin l'accès au flux audiovisuel lui-même. Aussi bien les tâches d'*indexation* (choix et mise en place des descripteurs et de la manière de les organiser qui permettront de retrouver plus tard les documents rentrés dans la base), que de *recherche* (à l'aide de ces descripteurs) et d'*assistance* à l'utilisateur, points fondamentaux dans les systèmes de recherche d'information modernes dépendent du choix de représentation des documents. La représentation nous semble à cet égard centrale et dépasse le cadre du traitement d'image dont sont issus la majorité des travaux en indexation audiovisuelle [1]. Le besoin de descriptions de haut-niveau d'abstraction se fait désormais ressentir y compris au niveau des comités de normalisation tels que MPEG4 et surtout MPEG7 [7]. Le problème de l'indexation d'un document audiovisuel est en effet de plus en plus lié au problème de l'*annotation*, c'est-à-dire au fait d'attacher une description à une partie, à un morceau d'un document. Sur la manière de choisir les morceaux, ainsi que leurs arrangements et liens, nous verrons que deux écoles s'affrontent, la *segmentation* et la *stratification*, les morceaux étant dans la première fixés avant la description, tandis que leurs limites sont fixées au besoin dans la seconde. Notre travail s'inscrira dans le cadre d'une approche par stratification. Nous nous appuierons également sur une structuration des strates et du vocabulaire d'annotation empruntées à la représentation de connaissances.

Le cadre empirique dans lequel sera utilisée notre approche est celle du projet SESAME (Système d'Exploration de Séquences Audiovisuelles et Multimédia enrichies par l'Expérience). L'ensemble du projet [3] a pour vocation de couvrir la problématique de la recherche de séquences audiovisuelles par le contenu dans la plupart de ses aspects¹

1. A- Etude de la modélisation des documents audiovisuels en vue d'indexation pour la recherche par le contenu, en particulier pour servir de support de représentation pour exploiter

L'ambition du projet est de permettre la mise à disposition d'un corpus de documents audiovisuels sur la base de leur contenu. Les indexeurs seront par exemple des documentalistes de chaînes de télévision ou d'institutions telle l'INA, tandis que le public des utilisateurs visé se composera de professionnels de la télévision. Le projet SESAME est en partenariat avec l'INA et France3 Rhône-Alpes.

Dans cet article, nous présenterons tout d'abord rapidement quelques-unes des spécificités fondamentales de l'audiovisuel, des documents audiovisuels par rapport aux autres types de documents, avant de nous intéresser aux différentes approches de modélisation couramment utilisées. Nous décrirons ensuite les bases de notre approche et un mécanisme d'inférence contextuelle qui lui est propre. Nous concluerons l'article par une discussion sur les perspectives ouvertes par ce travail inaugural.

2 Documents audiovisuels

À l'instar de l'écrit, l'audiovisuel est un média *séquentiel*, c'est-à-dire qu'il ne peut être appréhendé en un instant, et qu'un *ordre* et une *vitesse* de lecture lui sont inhérentes. De sa composante visuelle et de la *temporalité* attachée — de la cadence de défilement des images — naît le *mouvement* du cadre et des objets représentés. Renforcée par la bande audio, l'*illusion de réalité* est l'effet principal et primaire obtenu dans un *plan* (suite d'images filmées entre deux interruptions de la caméra).

Si l'on quitte le niveau du plan, une deuxième linéarité séquentielle, d'ordre supérieur, vient se superposer à la séquentialité immédiate présentée plus haut : par le biais du *montage*, il devient possible de *raconter*. On peut alors mettre en place une histoire en utilisant des arrangements de plans en scènes (suite de plans consécutifs constituant un tout sous le rapport d'une action dramatique déterminée : par exemple une poursuite ou un dialogue); un document audiovisuel étant alors constitué d'un ensemble de scènes. La structure du document se projette sur l'axe linéaire du temps en ce que nous convenons d'appeler une deuxième séquen-

l'expérience: Equipe «Indexation secondaire», A Mille (Lisa-Cpe-Lyon), JM Pinon, Y Prié (Lisi-Insa-Lyon). B- Etude des caractéristiques extractibles «automatiquement» des flux audiovisuels, et en particulier des flux vidéo: Equipe «Image», E Etievent, JM Jolion, F Lebourgeois (RFV-Insa-Lyon). C- Etude des modèles de données capables de représenter efficacement les différentes dimensions de l'audiovisuel, en particulier les dimensions temporelles: Equipe «Indexation Primaire», C Decleir, MS Hacid, J Kouloumdjian (Lisi-Insa-Lyon). D- Etude d'implantation efficace des bases de documents audiovisuels, capable de servir les recherches («temps réel») en parallèle sur des sites distribués: Equipe «Parallélisme», L Brunie, A Mostefaoui (LIP-ENS-Lyon).

tialité, qui ne concerne plus seulement des images qui se succèdent, mais bien d'autres unités (plans, scènes, documents).

L'audiovisuel étant également un médium *multimodal*, puisqu'on superpose et donc qu'on perçoit les modalités audio et vidéo, l'effet global résulte d'un montage global suivant celles-ci. Ajoutons qu'on peut considérer l'audiovisuel suivant une composante *multimédia* dès l'instant où l'on peut s'intéresser aux différents média de manière indépendante.

Pour terminer cette courte présentation de notre objet d'étude, insistons sur le rôle fondamental joué par le *contexte* dans l'audiovisuel, qui cumule les facteurs contextuels de l'image et du texte, mais également ceux liés aux deux séquentialités (grossièrement, intra- et inter-plans) déjà évoquées.

3 Modèles pour la recherche d'information audiovisuelle

Cette partie traite des systèmes de recherche d'information audiovisuelle et plus particulièrement des approches de modélisation qui leur sont sous-jacentes, dont nous proposons une grille d'analyse. Nous évoquons ensuite les deux approches structurelles proposées dans la littérature, avant de conclure sur les nécessités à notre sens fondamentales à prendre en compte dans toute modélisation audiovisuelle pour la recherche d'information.

3.1 Méthode d'analyse des différentes approches de modélisation

Le concept fondamental en ce qui concerne les médias séquentiels est celui de l'*annotation*, qui consiste à attacher une annotation, une description à une partie du document que l'on traite. Celle-ci représentera par exemple un paragraphe ou quelques mots dans un texte, et pour l'audiovisuel un *morceau* de document se définira par deux instants t_1 et t_2 .

La *granularité* de l'annotation est un premier critère de classification des modèles de représentation. En effet du plus petit intervalle de temps — on peut se limiter à une image, par exemple 1/25ème de seconde pour rester au niveau de la bande vidéo ou bien aller jusqu'au 1/9000ème de seconde, qui est la limite de précision des time-code —, au plus grand (un document entier, par exemple un film de 2 h), tout est possible. Le *découpage* (au sens large du terme) du document en morceaux peut être régulier ou non, et éventuellement multi-couches (on annote le document entier *et* chacun de ses plans).

La complexité de l'annotation elle-même est le critère d'analyse suivant. Une annotation peut-être *simple* ou *atomique* (par exemple un terme, [2]), ou bien *com-*

plexe, c'est-à-dire mettant en jeu une structure d'annotation (par exemple un réseau sémantique [8] ou une phrase iconique [6]).

Les annotations représentent au niveau du flux des *caractéristiques* (*features*) qui peuvent être de toutes sortes et que l'on peut classer en fonction des analyses qui ont permis de les détecter. Ainsi nous proposons d'appeler *primitives* celles des caractéristiques qu'il est possible d'extraire automatiquement du flux audiovisuel, que ce soit au niveau physique (taux de couleurs, texture, luminosité) ou après une certaine abstraction (contours, formes, leurs variations dans le temps, leurs mouvements). Les caractéristiques non primitives sont extraites semi-automatiquement ou manuellement du flux (personnages, objets, actions, sentiments ressentis) et/ou des documents qui y sont liés (titre, acteurs, commentaires).

S'il est possible de considérer un document audiovisuel réduit à un seul morceau qui est le document dans son ensemble, que l'on annote alors globalement par quelques mots-clés (ce qui nous ramène aux systèmes de recherche d'information classiques où une fiche décrit un document quelconque), il convient dès que l'on annote un document audiovisuel par morceaux de considérer la représentation de celui-ci de façon globale, en lien avec les annotations des morceaux dont il est composé. Des considérations sur la *structure* du document audiovisuel sont alors nécessaires, sur les liens entre les différents morceaux que l'on a annotés.

La représentation globale peut alors être *implicite* si le seul lien entre morceaux découle de la continuité du temps (du fait de la succession stricte des plans), ou bien *explicite* si une structure globale entre morceaux est mise en place, laquelle sera la plupart du temps hiérarchique (par exemple, tel morceau-document contient tels morceaux-scènes, qui eux-mêmes sont composés d'ensembles de morceaux-plans). Des liens entre annotations peuvent découler de cette structure (par exemple une scène s'enrichit des annotations des plans dont elle est constituée [2], ou un morceau profite des annotations du morceau qui le contient [10]).

Les modèles de représentation sous-jacents aux systèmes de recherche d'information peuvent être à notre sens systématiquement décomposés suivant les quelques critères que nous proposons : granularité, choix des caractéristiques, complexité de l'annotation, structuration de la représentation.

3.2 Approches structurelles d'annotation

Deux approches principales d'annotation se dégagent dans la littérature, à savoir la *segmentation* et la *stra-*

tification.

Segmenter un document audiovisuel consiste à le découper en segments, ou morceaux audiovisuels de types fixés et bien choisis (plans, scènes, ...) que l'on va ensuite annoter. Une structure est également affectée au document sous la forme d'une arborescence dont les feuilles sont des plans (descendre au niveau de l'image est considéré comme inutile). Cette approche (cf. figure 1) et toutes les variantes qu'elle autorise au niveau de la granularité et de la structure (on peut se limiter à un ensemble de plans consécutifs, ou bien mettre en place une hiérarchie à plusieurs niveaux) est de très loin la plus utilisée de par le monde.

Le modèle le plus avancé à notre connaissance dans cette direction est celui de Corridoni et *al.* [4] dans lequel est développée une véritable grammaire filmique : un film est composé d'unités sémantiques (scènes), et syntaxiques (plans), une scène étant composée de plans délimités par des «ponctuations» (effets d'édition). A chaque unité structurale est associée une annotation structurée.

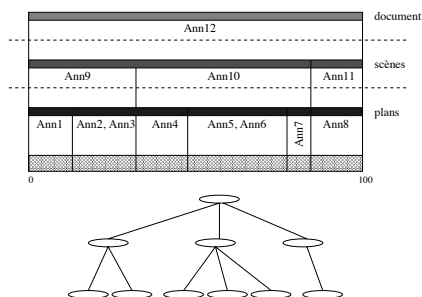


FIG. 1 – Approche de segmentation : décrire des morceaux audiovisuels typés éventuellement organisés entre eux en hiérarchie structurale

L'approche de *stratification* a été développée au MediaLab du MIT par Davenport, Aguiere-Smith [5], puis Davis [6]. Elle s'oppose à la segmentation classique notamment sur les points suivants : en dehors d'eux-mêmes, plans ou scènes ne représentent rien et ne tiennent pas compte du montage, tandis qu'il y a perte de la linéarité du média temporel qu'est l'audiovisuel. Au contraire, on se situe ici dans une dynamique de flux, une *strate* est une sous-partie quelconque du document audiovisuel, à laquelle est attachée une annotation. Le découpage «utile» pour l'utilisateur — qu'on peut rapprocher de la segmentation évoquée plus haut, à la différence qu'ici les morceaux ne sont pas typés — du document se fait alors *après-coup*, en fonction des requêtes, par intersection de strates (cf. figure 2).

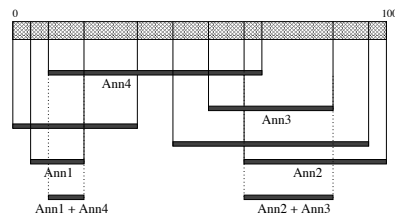


FIG. 2 – Approche de stratification : une annotation de morceaux et un découpage utile a posteriori

L'approche de stratification permet de mieux prendre en compte le caractère fondamental de l'audiovisuel qu'est la temporalité, puisque les bornes d'une strate sont librement posées. Elle néglige néanmoins certaines données structurales (plans, scènes, ...) pourtant couramment admises aussi bien au niveau de la production que de l'étude de la réception de document audiovisuels, elles-mêmes prises en compte dans une approche de segmentation un tant soit peu structurée telle que [4].

Le problème essentiel nous semble résider dans la définition des morceaux temporellement situés et annotés : dans le cas de la segmentation on fixe des morceaux que l'on annote ensuite, tandis qu'en ce qui concerne la stratification, ce sont les annotations temporellement situées qui fixent des strates, «unités de base» d'un second découpage, d'autres unités se dégageant *a posteriori* lorsque l'on considère des combinaisons d'annotations. On se trouve ainsi devant deux démarches différentes, un découpage *a priori* ou *a posteriori*. Le statut d'annotation recouvre en fait dans ces deux cas deux réalités qui diffèrent quelque peu.

En effet, dans le cas de la segmentation, l'annotation est une description d'un morceau de document structurellement fixé, qui est premier. On dira alors que tel plan ou telle séquence est décrit par telle et telle caractéristique, l'ensemble formant l'annotation, ce qui nous rapproche du paradigme classique de description de document. Dans le cas de la stratification au contraire, l'annotation prend un statut d'unité première, la structure *a posteriori* émergente (*i.e.* après requête, quelle que soit la nature de cette dernière) se dotant en même temps qu'elle émerge d'une description-annotation qu'on pourrait dire seconde, ce qui nous éloigne du paradigme évoqué plus haut. Il nous semble que c'est à ce niveau que réside la différence fondamentale entre segmentation et stratification.

3.3 Conclusion

Quelques conclusions sur la modélisation audiovisuelle s'imposent : en premier lieu, la *dynamique temporelle*

liée au flux doit pouvoir être prise en compte, ce qui ne peut se faire qu'à l'aide des *strates* d'une part et d'une *atomicité* des annotations d'autre part. En deuxième lieu, une structuration de l'annotation est nécessaire afin d'augmenter le pouvoir d'expression de la description. La décomposition structurelle hiérarchique d'un document audiovisuel doit également être considérée (*cf.* segmentation). Le niveau de complexité de la représentation, à ces deux niveaux doit se caractériser par sa variabilité. Enfin, il apparaît important de pouvoir tenir compte des relations contextuelles internes à un document audiovisuel.

4 L'approche STRATES-IA

Nous présentons ici notre approche de modélisation audiovisuelle en Strates Interconnectées par les Annotations (STRATES-IA).

Le principe général en est simple : on considère tout ce qui peut être relevé ou dit sur le document audiovisuel considéré — primitive image ou son, caractéristique de haut-niveau conceptuel, notation structurelle — comme caractéristique, comme annotation temporellement située dans le flux, comme terme lié à une strate. Dans un deuxième temps, on met en place des relations inter- et intra-strates entre annotations, à la manière des annotations, afin de pouvoir complexifier la représentation, par exemple en exprimant les relations structurelles d'un document audiovisuel. Enfin, on considère toute annotation (et ses relations) comme *support de relations contextuelles* propres à permettre et à guider une annotation contextuelle qui vient compléter l'annotation originelle. Nous précisons dans la suite les concepts et le vocabulaire liés à cette approche.

4.1 Définitions

Objets d'intérêt audiovisuels, dimensions d'analyse. Nous appelons *objets d'intérêt audiovisuels* les objets — au sens général du terme — qui peuvent être repérés lors de l'appréhension (visionnage / audition) d'un flux audiovisuel. On peut considérer cette notion comme le pendant cognitif et conceptuel des caractéristiques audiovisuelles, qui en sont alors les «réifications informatiques» sous forme symbolique. Un objet d'intérêt peut se référer à des primitives allant du très bas niveau d'abstraction (un taux de couleur) au très haut (une action). Il en existe bien entendu autant que l'on peut mener d'analyses (manuelles ou automatiques) sur un flux. On regroupe en *dimensions d'analyse* les analyses qui permettent de repérer des objets d'intérêt de même niveau d'intérêt. On pourra par exemple considérer la dimension d'analyse liée au repérage des plans ou une dimension

d'analyse liée à celui des visages. Mais aussi des dimensions d'analyse liées à la recherche d'unités structurelles de tout niveau (plan, scènes, documents), ou encore à l'analyse de la présence d'un personnage public sont également possibles : les dimensions ne sont pas figées *a priori* et ne permettent que de regrouper des méthodes et/ou des types de caractéristiques de façon pertinente par rapport aux objectifs de l'analyse en cours.

Flux et unités audiovisuels. Un flux audiovisuel² FAV (ou *séquence audiovisuelle*, SAV) est un fichier informatique contenant des données audio et vidéo, commençant à l'instant 0, et se terminant à un autre instant.

On définit une *unité audiovisuelle* (par la suite écrite UAV) comme une entité abstraite représentant un morceau quelconque du flux audiovisuel. Celle-ci se compose d'un flux audiovisuel (nom du fichier), ainsi que de deux instants t_1 et t_2 référencés par rapport à l'instant 0 du flux, et permettant de situer sans ambiguïté le morceau de document audiovisuel représenté par l'UAV. Une UAV prend naissance à partir du moment où son existence devient d'intérêt dans l'analyse du flux, c'est-à-dire à partir du moment où elle a été repérée comme telle, à savoir comme zone ou segment temporel de flux lié à un objet d'intérêt audiovisuel. Le morceau, via son représentant UAV, accède à l'existence dès lors qu'il a été spécifié, que lui a été attaché une description, que celle-ci soit d'ordre structurel («ceci est un morceau de document que l'on appelle plan») ou autre («ceci est le morceau de document dans lequel apparaît X»).

Toute UAV doit par définition être *annotée*, c'est-à-dire associée à au minimum une caractéristique correspondant à l'objet d'intérêt détecté. On appelle *annotation primitive* cette annotation première et définitoire de l'UAV (figure 3). Une UAV est une entité «biface» au sens où elle permet de faire le lien entre le morceau de flux audiovisuel et la caractéristique qui lui est associée en même temps qu'elle le définit.

Éléments d'annotation. Nous définissons un *élément d'annotation* (par la suite noté EA) comme un *terme* (considéré de la façon la plus large pour l'instant) en *relation d'annotation* R_a avec l'UAV. Les EA sont les expressions de caractéristiques. Celles-ci peuvent représenter des objets d'intérêt qui ont été détectés (et avoir par là même participé de la mise en place des UAV) de n'importe quelle façon (du plus automatique au plus idiosyncrasique). L'analyse utilisée peut s'intéresser à un quelconque des objets d'intérêt d'un

2. A ne pas confondre avec la notion de flux utilisée dans le cadre des serveurs de vidéo à la demande (VoD).

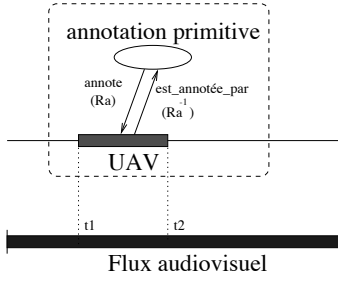


FIG. 3 – Une unité audiovisuelle et l’annotation primitive qui la fonde

flux : analyse de couleurs, de formes, de mouvements, de ruptures de plan, des types de vues, des mouvements de caméra, de la bande-audio, d’objets, d’actions, du type de document, des auteurs, des sensations produites, *etc.*

On aura ainsi par exemple les EA $\langle Plan \rangle$, $\langle Senna \rangle$, $\langle Objet_rond \rangle$, $\langle Histogramme_bleu \rangle$, $\langle Contenu_dans \rangle$ ou encore $\langle Glauque \rangle$. Le principe fondamental ici se trouve dans l’atomicité de l’EA : il est *a priori* possible de considérer tout EA comme tel et de façon atomique, quelle que soit la caractéristique qu’il exprime, et ce à une exception près, qui tient aux caractéristiques à caractère numérique (par exemple un taux de couleur, auquel cas la présence d’un EA indiquera que la valeur numérique est disponible, tandis que cette valeur sera considérée comme attribut de l’EA). D’autres attributs d’EA sont également possibles, par exemple du texte ou un lien sur une image qui lui serait liée.

En plus de l’EA primitif qui fonde une UAV, on se donne la possibilité de mettre en relation d’annotation R_e avec celle-ci autant d’éléments d’annotation supplémentaires que nécessaire. Ceux-ci peuvent avoir été mis en place selon plusieurs mécanismes : en premier lieu la *structuration de l’annotation* par la mise en relation d’annotations jusqu’ici isolées (voir 4.2), ainsi que le *regroupement d’annotations* de même extension temporelle.

Dans ce dernier cas (*cf.* figure 4), on convient simplement de rajouter à l’annotation d’une UAV, par exemple l’EA $\langle Document \rangle$ pour une UAV correspondant à tout le document que l’on considère, d’autres EA correspondant à des UAV de même extension temporelle, par exemple concernant l’auteur ou le producteur du document.

4.2 Relations entre annotations

Selon notre approche il est indispensable d’autoriser la mise en place d’une structuration de l’annotation : pour pouvoir exprimer des annotations plus complexes

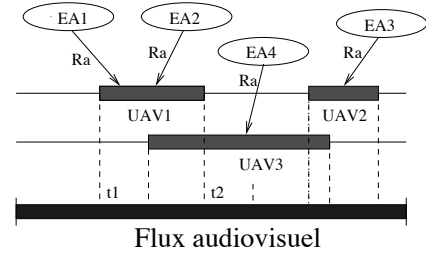


FIG. 4 – Unités audiovisuelles et éléments d’annotation

que les annotations atomiques (les éléments d’annotation, les termes) dont on dispose déjà, il nous faut pouvoir mettre en relation des éléments d’annotation entre-eux, exprimer des relations entre EA. Si certaines de ces relations peuvent à la limite être considérées comme implicites, par simple juxtaposition d’éléments d’annotation (deux EA correspondant à deux personnalités annotant la même UAV signifient sans doute que ces deux personnages sont ensemble sur l’image), d’autres se doivent d’être explicitées, par exemple, «ce plan est contenu dans cette scène», «ce personnage accomplit telle action», «cet objet soleil correspond à ce contour rond et jaune», «ce plan est réutilisé dans ce document», *etc.*

Afin de garder le principe de la stratification, de l’annotation atomique temporellement située, ces relations doivent être exprimées à la manière des annotations, c’est-à-dire à l’aide d’annotations-relations en relation avec les EA qu’elles relient, toutes les relations entre EA se résumant à la *relation élémentaire* R_e . Les véritables relations (celles ayant une signification) sont *nommées* sous la forme d’éléments d’annotation canoniques. Par exemple pour exprimer le fait que «Jean marche», à partir de deux UAV annotées respectivement par $\langle Jean \rangle$ et $\langle Marcher \rangle$ ³, on utilisera les éléments d’annotation canoniques (aux noms inspirés de ceux utilisés par Sowa [9]) $\langle AGNT \rangle$ et $\langle STAT \rangle$ qui annoteront les deux UAV, tandis que la relation élémentaire R_e sera la seule véritable relation entre EA (*cf.* figure 5).

Le fait d’utiliser pour toute relation deux EA canoniques permet de prendre en compte les deux sens de la relation, à savoir que «Jean marche» et que «l’action de marcher est effectuée par Jean». Pour une relation du type «ce plan est contenu dans cette scène» on pourrait par exemple utiliser les EA canoniques $\langle CONTENU_DANS \rangle$ et $\langle CONTIENT \rangle$. On notera que le type de relation-annotation que nous

3. Détectées suivant par exemple les dimensions d’analyse correspondant aux «personnages» et aux «activités».

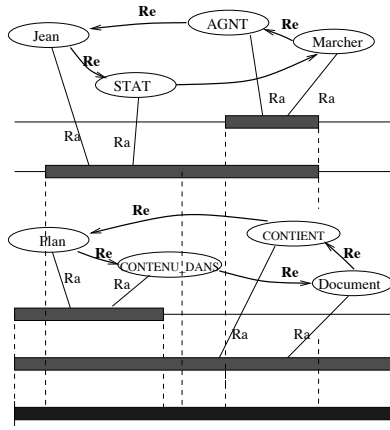


FIG. 5 – Structuration de l'annotation

proposons fonctionne aussi bien entre EA annotant la même UAV qu'entre EA annotant deux UAV différentes. Remarquons d'ores-et-déjà que les relations inter-strates nous permettent de considérer des UAV en relation avec d'autres UAV, dont elles font en quelque sorte partie du contexte.

Finalement, annoter un flux revient à étudier ce dernier *suivant* autant de dimensions d'analyse que nécessaire, afin de mettre en place pour chaque dimension un certain nombre d'UAV ainsi que les éléments d'annotation qui leur sont liés. Certaines relations — toujours sous la forme d'EA, et avec la seule relation élémentaire R_e sont également mises en place. Ceci est illustré par la figure 6.

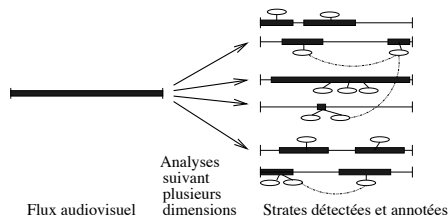


FIG. 6 – Annotation de flux et dimensions d'analyse

4.3 Base de connaissances associée à l'annotation

Les éléments d'annotation se comportent de façon atomique au niveau de l'annotation. Ils seront support d'inférences contextuelles menées sur le réseau d'annotations que constitue un flux audiovisuel ayant été annoté, par exemple par un documentaliste. Afin de faciliter et de contrôler le travail de ce dernier en vue de la recherche et des inférences réalisables, il est né-

cessaire d'envisager ces termes comme issus d'un vocabulaire contrôlé au sein d'une base de connaissance. Nous considérons qu'un élément d'annotation résulte de l'*inscription dans le flux* (par l'intermédiaire des UAV) d'*éléments d'annotations abstraits* (EAA), eux-mêmes décrits dans une base de connaissances BC. Un EAA est en relation d'inscription dans le flux R_{if} avec les EA qui lui sont liés tandis que ceux-ci sont en relation de *décontextualisation* R_d avec l'EAA dont ils sont issus. On aura par exemple $\langle \text{Chirac} \rangle R_d \langle \text{EAA} : \text{Chirac} \rangle$ et $\langle \text{EAA} : \text{Chirac} \rangle R_{if} \langle \text{Chirac} \rangle$, ou encore $\langle \text{EAA} : \text{Maire} \rangle R_{if} \langle \text{Maire} \rangle$ (figure 7).

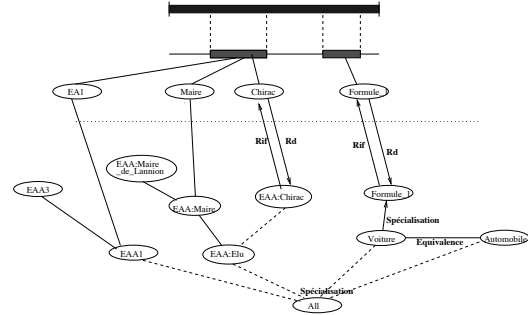


FIG. 7 – Éléments d'annotation abstraits et base de connaissances

Les EAA sont organisés dans la base de connaissances à l'aide de relations, qui sont des relations de concepts (par exemple spécialisation, synonymie, équivalence, etc.). Ces relations permettent d'organiser la base de connaissances en un réseau qu'on ne limitera pas *a priori* à une hiérarchie, quoiqu'il soit bien entendu possible d'en extraire un treillis lié à une relation, pour peu que celle-ci soit un ordre partiel. La base de connaissances joue ici le rôle que joue un thésaurus en Recherche d'Information classique, mais doit également permettre d'organiser toute la connaissance sur les EA et sur leurs attributs possibles (possibilités de relations, domaines de valeurs pour les représentants d'attributs numériques, etc.).

Toute annotation d'un document audiovisuel, c'est-à-dire l'analyse de celui-ci suivant un certain nombre de dimensions finit par se résumer en unités audiovisuelles et en éléments d'annotation (voir figure 6). Cet ensemble d'UAV et d'EA n'est *a priori* pas indépendant du système dont il fait partie : il trouve sa place, il se fonde dans la *base des documents annotés* (BDA — ensemble des UAV et des EA de la base) par les relations que certaines de ses UAV entretiennent avec d'autres UAV de la base par l'intermédiaire des EA. Le flux-fichier n'apparaît plus comme tel, mais s'inscrit comme ensemble d'UAV annotées dans l'ensemble

complet de la base. Il s'inscrit également dans le système plus global composé de la BDA et de la base de connaissances du fait des liens de chacun de ses EA avec les EAA dont ils sont les inscriptions. Ces principes sont illustrés figure 8.

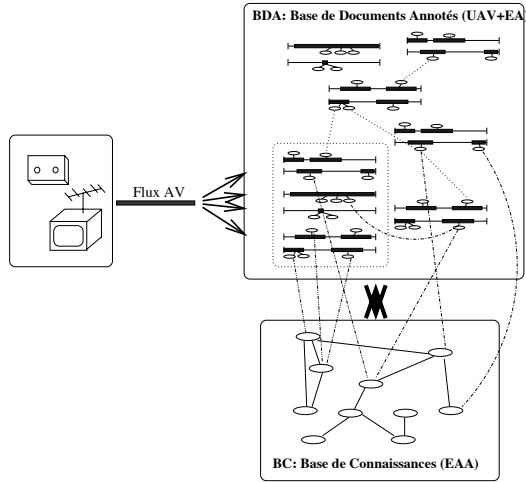


FIG. 8 – Annotation et Bases du système

4.4 Contexte d'une UAV

Nous avons insisté plus haut sur l'importance du contexte dans l'audiovisuel, aussi bien au niveau intra-documentaire — par exemple contexte interne à un plan (tel objet sur lequel la caméra zoome se situe dans telle pièce d'abord évoquée, telle musique accompagne tel personnage ou telle action) ou bien contexte de montage (tel plan ne se comprend que parce que tel autre l'a précédé) — qu'au niveau inter-documentaire (par exemple tel document contient un extrait d'un tel autre, est diffusé dans telle émission).

Si dans notre approche chaque document se fonde dans la BDA en UAV annotées, celles-ci ne sont pas isolées mais doivent être considérées comme insérées dans un réseau d'annotations. Ainsi chaque UAV est en relation temporelle *explicitée* (*i.e.* sous forme d'annotation) avec ses voisines, donc s'insère dans le flux audiovisuel temporel. Elle peut aussi être en relation structurelle explicite avec ses voisines (par exemple entre deux UAV repérées suivant la dimension d'analyse structurelle et annotées par les EA $e_1 = \langle \text{Scene} \rangle$ et $e_2 = \langle \text{Plan} \rangle$ et les EA canoniques de relation $e_3 = \langle \text{CONTIENT} \rangle$ et $e_4 = \langle \text{CONTENU_DANS} \rangle$, avec $e_1 R_e e_3 R_e e_4 R_e e_2$). Enfin, et c'est le cas le plus général, elle est en relation *conceptuelle* avec d'autres UAV (par exemple si un EAA lié à un personnage est plusieurs fois inscrit dans le flux, ou bien si on a une

relation explicitée entre un EA $\langle \text{Voiture} \rangle$ et un EA $\langle \text{Zoom_avant} \rangle$ spécifiant que «le zoom se termine sur la voiture»).

Nous considérons donc que toutes les relations de contexte que nous avons évoquées peuvent être comprises comme des relations entre éléments d'annotation annotant des UAV détectées suivant des dimensions d'analyse pertinentes. En d'autres termes, nous considérons non seulement les *éléments d'annotation* comme annotations simples (ce qu'ils sont pas définition), comme *éléments d'annotations* plus complexes (dans leur organisation en réseau), mais également — et du fait de cette organisation — comme *supports* de mise en relation contextuelle.

Nous définissons dans notre approche le *contexte interne à la base des documents annotés* d'une UAV comme un certain nombre d'UAV qui peuvent lui être liées en considérant des *relations contextuelles internes* $e_1 R_e \dots R_e e_2$ entre EA les annotant respectivement. Un contexte dit *externe à la BDA*, composé d'éléments d'annotation abstraits est également défini en considérant des *relations contextuelles externes*, permettant de lier des EA à des EAA.

4.5 Construction de vues

La possibilité de considérer le contexte d'une UAV nous permet de construire des *vues contextuelles* de celles-ci. De telle vues sont composées d'*éléments d'annotation contextuels*. Ces éléments d'annotation peuvent provenir du contexte interne à la BDA — auquel cas ils viennent annoter une autre UAV que celle qu'ils annotaient à l'origine — ou bien du contexte externe à la BDA, auquel cas ce sont les EAA qui le composent qui s'inscrivent dans le flux pour annoter l'UAV dont on souhaite la vue.

La *construction* d'une vue d'une UAV u se fait en deux étapes. Dans un premier temps, il s'agit de construire le contexte de l'UAV u (*cf.* figure 9, avec R_{syn} relation de synonymie, R_{spec} relation de spécialisation, R_d relation de décontextualisation, $R_{gen} = R_{spec}^{-1}$).

L'annotation contextuelle potentielle interne à la BDA est l'ensemble des EA annotant les UAV annotées par les EA liés par des relations contextuelles internes à l'UAV considérée. Dans la figure 9, on a ainsi les annotations contextuelles potentielles $EA4, EA6, EA10$ qui découlent de la relation $(EA3, R_e, EA4)$ et $EA8, EA9$ qui découlent de la relation $(EA5, R_e, EA1, R_e, EA7, R_d, EAA7, R_{syn}, EAA8, R_{if}, EA8)$ (les relations ainsi que leurs extrémités sont en traits doubles).

L'annotation contextuelle potentielle externe à la BDA est l'ensemble des EAA liés par des relations contextuelles externes à l'UAV considérée. Ainsi dans la figure 9 on a les annotations contextuelles potentielles

notation lié à une unité audiovisuelle. On structure ensuite l'annotation en mettant en place à la manière des annotations des relations entre ces annotations, et donc entre ces strates. Le vocabulaire des termes-EA utilisés est contrôlé à l'aide d'une base de connaissances organisée en réseau sémantique.

L'intérêt de cette «déconstruction» du document est d'autoriser tout lien contextuel explicite à l'intérieur du document, y compris entre UAV ne se superposant pas (et autres que celles découlant de la simple répétition d'un terme-annotation à différents moments), ainsi que vers d'autres documents de la base de documents annotés. Il est alors possible de considérer pour une UAV quelconque un contexte basé sur les relations que ses éléments d'annotation entretiennent avec d'autres EA de la base, y compris en utilisant des relations passant par la base de connaissances : les annotations sont support de relations contextuelles conceptuelles, les relations d'ordre structurel ou temporel en faisant partie. On peut ensuite construire à partir de ce contexte une vue d'une UAV, qui est alors annotée contextuellement, l'ensemble de ces opérations étant défini par un filtre de vues. Ceci nous permet d'exprimer de façon formelle des contextes audiovisuels et de les manipuler. Ajoutons qu'aucune granularité n'étant fixée *a priori*, il est possible de considérer un grand nombre de niveaux de précision différents dans l'annotation, et ce y compris au sein d'un même document. De même, tout type d'annotation est autorisé, et donc tout type d'analyse.

L'approche des STRATES-IA permet une description très souple des documents audiovisuels et surtout fournit les éléments de base afin de définir des stratégies d'indexation ou de recherche s'appuyant sur des besoins et des usages très divers. Ces tâches sont toutes liées à une *description* d'un document audiovisuel : *indexation* bien sûr (suivant des méta-modèles de description liés par exemple aux types de documents), mais aussi *recherche* (on décrit ce que l'on cherche), *analyse* (découverte de nouveaux liens, de nouvelles descriptions), *montage* (appuyé sur une analyse, éventuellement réorganisation d'un graphe d'annotation). La tâche de *navigation* s'appuie sur les liens contextuels d'annotation.

Afin de préparer une validation de l'approche, nous nous attachons désormais à la mise au point de premiers méta-modèles de description (vocabulaires et Bases de Connaissances, méthodologies de description) avec des documentalistes de l'INA et de France 3. Parallèlement, une interface homme-machine correspondant à notre modèle est en cours de mise au point, afin d'une part de permettre une annotation plus aisée et donc la conception des méta-modèles, et d'autre part

de pouvoir tester les utilisations liées aux tâches évoquées plus haut. Plus théoriquement, nous étudions comment un formalisme lié à la théorie des graphes permettra d'exprimer les possibilités de l'approche, ainsi que les mécanismes d'inférence contextuelle à l'œuvre dans les tâches de description de l'utilisateur. Nous menons à ce sujet une réflexion sur les mécanismes d'apprentissage basé sur l'expérience qui permettront d'aider ce dernier. Enfin, l'étude de la projection de notre modèle sur le modèle logique de la base de données des documents audiovisuels est un travail engagé au sein du projet SESAME en connexion avec les travaux des spécialistes des bases de données.

Références

- [1] Aigrain (P.), Petkovic (D.) et Zhang (H.J.). – Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications*, 1996.
- [2] Chua (T.-S.) et Ruan (L.-Q.). – A video retrieval and sequencing system. *ACM Transactions on Information Systems*, vol. 13, n4, octobre 1995, pp. 372–407.
- [3] Collectif. – *Projet SESAME*. – Rapport SESAME1, CPE-Lyon, ENS-Lyon, INSA-Lyon, Juillet 1997.
- [4] 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.
- [5] Davenport (G.), Aguierré Smith (T.) et Pincever (N.). – Cinematic primitives for multimedia. *IEEE Computer Graphics and Applications*, Jul. 1991, pp. 67–74.
- [6] 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.
- [7] Pereira (Fernando). – Mpeg-7 : A standard for content-based audiovisual description. In : *2nd Int. Conf. on Visual Information Systems*, pp. 1–4. – San Diego, Dec. 1997.
- [8] Simmonot (B.). – *Modélisation multi-agent d'un système de recherche d'information multimédia à forte composante vidéo*. – Thèse de doctorat, Université Henri Poincaré — Nancy I, Jan. 1996.
- [9] Sowa (J.-F.). – *Conceptual Structures — Information Processing in Mind and Machine*. – Addison-Wesley, 1984.

- [10] Weiss (R.), Duda (A.) et Gifford (D.K.). – Composition and search with a video algebra. *IEEE Multimedia*, vol. 2, n1, 1995, pp. 12–25.