Description et indexation automatiques des documents multimédias : du fantasme à la réalité

Patrick Gros

Description et indexation automatique : du fantasme à la réalité

Patrick Gros
IRISA – CNRS Rennes - projet TEXMEX
Patrick.Gros@irisa.fr

L'indexation automatique des documents multimédias a pour but de permettre, par le biais de techniques automatiques ou semi-automatiques, l'exploitation de collections de documents multimédias. L'apparition de ce domaine de recherche, en ce qui concerne les images et les documents audiovisuels, date de la première moitié des années 90, et est donc encore récente. Son émergence a suscité un double mouvement, d'enthousiasme chez les chercheurs qui y ont vu un domaine nouveau d'investigation et qui, selon leur habitude, ont beaucoup promis afin d'attirer des financements pour mener leurs activités, et d'inquiétude chez certains professionnels de la documentation audiovisuelle qui y ont vu une remise en cause de leur métier voire un danger de disparition de leur emploi. Quelques années ayant passé depuis ces débuts, il est intéressant de remettre les choses à plat. Quel est l'objet actuel de l'indexation automatique? Quelles sont ses possibilités, ses applications? Que peut-elle faire ou ne pas faire? Quelles sont les perspectives? C' est à ces questions que nous allons tenter de répondre.

1 Un domaine jeune... au lourd passé

L'expression même d'indexation automatique est sujette à discussion. Elle n'est, tout d'abord, pas largement partagée. Suite au travaux d'IBM sur le système QBIC (Query by Image Content), le domaine s'est développé sous le vocable général d'indexation multimédia « par le contenu », pour le distinguer d'une indexation qualifiée de manuelle. Cette expression est tout à fait malheureuse, laissant entendre qu'il y aurait, d'une part, une indexation basé sur le contenu même et par conséquent objective, face à une indexation basée sur l'interprétation et donc subjective et par conséquent appelée à être reléguée au rayon des pratiques désuètes. On trouve souvent un tel argumentaire dans les thèses du domaine, mais il ne fait que trahir l'ignorance des doctorants, voire de leur encadrants, vis-à-vis du milieu professionnel de la documentation audiovisuelle et de ses méthodes.

La naissance du domaine a, par ailleurs, été marquée par une grande effer-

vescence. Peu d'universités américaines n'ont pas eu leur projet d'indexation automatique d'images. Du coup, il y a eu compétition et surenchère, chacun voulant se démarquer et faire mieux ou, pour le moins, faire une meilleure publicité pour ses propres travaux. Beaucoup se sont rués dans les agences de photo et ont promis d'y remplacer les documentalistes... Puis le soufflet est retombé, car une fois que tout le monde eut essayé de retrouver des images de couchers de soleil en utilisant des histogrammes de couleur, il a bien fallu se rendre à l'évidence que ce n'est pas avec ce genre de technique qu'on allait pouvoir résoudre des problèmes réels et que le chemin était plus ardu que prévu. La place était alors libre pour les quelques équipes qui ont choisi de travailler plus à fond dans le domaine, les autres abandonnant le sujet.

Une autre difficulté vient de l'emploi du terme indexation qui, dans la bouche ou sous la plume de beaucoup, signifie tant la structuration, la description que l'indexation proprement dite des documents, voir la recherche de documents ou la navigation dans des collections. Cette confusion, qui n'est pas sans créer de problèmes de communication avec d'autres domaines, est aussi le signe que les problèmes à résoudre n'étaient pas complètement identifiés encore.

Dans les faits, on peut distinguer plusieurs tâches :

1. la structuration des documents, qui est le plus souvent une segmentation et qui consiste à repérer dans un document des entités d'intérêt ; ainsi cherche-t-on à retrouver les divers plans d'une vidéo, à y isoler les objets en mouvement par exemple ;
2. la description des documents ou des éléments issus de la structuration précédente, qui consiste à calculer un certain nombre de quantités à partir du contenu ; l'idée sous-jacente est bien entendu que l'on pourra par la suite, réduire la manipulation des contenus à celle des descripteurs ainsi calculés ;
3. la sélection, c'est à dire le choix, des descripteurs utiles dans un contexte donné, en fonction de la collection considérée et des requêtes qui devront être traitées ;
4. l'indexation proprement dite, qui va consister à organiser toute cette information pour y accéder de manière efficace et efficiente ;
5. enfin, l'utilisation de ces descripteurs pour y retrouver l'information recherchée et répondre à une requête, ou pour naviguer dans la collection de documents.

D'autres éléments doivent bien entendu être ajoutés à ceux-ci pour obtenir une système complet, avec des interfaces, de l'interactivité...

2 Les images fixes et leurs contextes

Quelques contextes applicatifs. Dans le domaine des images fixes, plusieurs contextes sont intéressants et pourraient bénéficier de techniques auto-

matiques. Côté grand public, la gestion des collections de photos personnelles et familiales est un sujet d'actualité. Avec la diffusion des appareils photo numériques, les photos personnelles sont désormais en format numérique et donc traitables aux moyens de programmes informatiques. Par ailleurs, que ce soit les magnétoscopes, les ordinateurs, peu de foyers ne sont pas équipés d'outils électroniques qui peuvent servir à stocker et manipuler de telles images. Dans ce cas, le problème vient du fait que les photos sont la plupart du temps stockées sans annotation et que toute recherche ou navigation autre que chronologique n'est pas possible ou reste très pénible.

Dans le domaine professionnel, les agences de photos ont été les premières contactées, bien qu'elles ne soient pas les seules à gérer de grandes quantités d'images. L'attention a été fortement focalisée sur le travail des documentalistes qui gèrent les photos et répondent aux requêtes des clients. Il est toutefois très vite apparu que les requêtes étaient de très haut niveau sémantique et que les documentalistes pouvaient s'appuyer sur une bonne connaissance des besoins du client, ou sur une compréhension de sa demande qui n'est pas formelle, et sur leur mémoire du contenu de la collection qu'elles gèrent. Toutes choses qu'un ordinateur a du mal à gérer. Mais il existe d'autres tâches où l'ordinateur peut avoir un apport intéressant. Tout d'abord dans l'aide à l'annotation, dans un fonctionnement semi-automatique, où l'ordinateur peut soit proposer des annotations à un documentaliste qui les valide et complète, soit annoter automatiquement une partie de collection à partir des annotations faites manuellement sur une autre partie. Ce type de fonctionnement peut convenir pour des possesseurs d'archives qui n'ont pas les moyens d'annoter manuellement l'ensemble de leur collection, par exemple les quotidiens régionaux qui récupèrent de grandes quantités d'images de leurs correspondants locaux. Dans ce domaine, le couplage entre appareils photos et systèmes de positionnement par satellites (GPS ou Galileo) apportera une aide appréciable.

La recherche des copies illégales devient incontournable : dès qu'un possesseur d'images veut utiliser internet pour faire connaître son fond, le souci du piratage devient majeur. Mais l'ampleur de la tâche est immense : comment confronter toutes les images que l'on peut trouver sur internet avec celles d' une collection de plusieurs millions d'images ? Clairement, il ne peut y avoir de solution qu'automatisée, au moins pour un premier tri. Il y a là un problème intéressant, qui correspond à un vrai besoin et pour lequel il n'y a pas de solution manuelle possible.

De nombreuses autres applications sont possibles, correspondant à des collections particulières d'images. Il y a, tout d'abord, toutes les applications biométriques. Ainsi le système de vidéo surveillance londonien est-il couplé à un système de reconnaissance de visages qui a permis plusieurs arrestations. Dans le domaine policier toujours, la police judiciaire française utilise un système de reconnaissance pour comparer des images pédophile afin de retrouver celle qui seraient prises dans une même pièce, ou retrou-

ver des indices communs à plusieurs photos. Le domaine médical est aussi fortement demandeur tant pour la formation des médecins que pour l'exploitation des très nombreuses images produites (pour comparer des organes pathologiques, rapprocher des cas atypiques par exemple). La météorologie, les satellites sont de très gros pourvoyeur de données qui ne sont pas exploitées autant qu'elles pourraient.

2.1 Description automatique d'images fixes

Face à ces besoins, de quels outils disposent les chercheurs ? On peut distinguer deux approches, suivant que l'on cherche à décrire l'image dans sa totalité ou seulement des parties de celle-ci.

Des descripteurs globaux. Dans la première approche, on cherche à calculer des descripteurs globaux. En l'absence de toute indication sur l'image, on en est réduit à observer le signal bidimensionnel constituant l'image. Informatiquement, une image se présente comme un table rectangulaire de d'éléments élémentaires appelés pixels, chaque pixel codant sous forme d'un ou plusieurs nombres (généralement trois), l'information d'intensité lumineuse ou de couleur présente en un point. Ce sont donc ces nombres que l'on va utiliser pour décrire les images.

L'information la plus accessible est bien sûr la couleur, puisqu'elle est directement codée au niveau de chaque pixel. Et le plus simple que l'on puisse faire est de déterminer quelle couleur est présente dans l'image et quelle proportion de la surface de l'image elle remplit. C'est ce qu'on appelle un histogramme de couleur. Bien évidemment, un tel procédé de description reste simpliste. Mais il suffit pour rechercher des images de coucher de soleil dans une collection d'images de forêt tropicale... Le procédé peut être amélioré de deux manières : tout d'abord en utilisant plus finement la distribution des couleurs : un pixel jaune entouré de bleu ne donne pas du tout le même effet qu'un pixel jaune entouré de jaune. On peut ainsi pondérer l'importance de chaque pixel dans l'histogramme en fonction de son environnement immédiat. Cela augmente grandement la puissance descriptive de l'histogramme. Par ailleurs, de nombreuses distances existent pour comparer des histogrammes et juger ainsi de la ressemblance entre les images. Le choix de cette distance influe bien entendu sur les résultats, et est un facteur déterminant de la rapidité du système. Les distances les plus intéressantes sont malheureusement très onéreuses à calculer.

Une deuxième approche consiste à faire abstraction de la couleur, mais à s'intéresser uniquement aux distributions locales des intensités lumineuses. On va ainsi chercher à déterminer dans quelle mesure un pixel clair est plutôt entouré de pixels clairs ou foncés. L'idée est de décrire la texture de l'image, c'est à dire le fait qu'une image de nuage ne ressemble pas à une image d'herbes, de marbre ou de feuillage, indépendamment de sa couleur.

Le problème est que la notion de texture n'a pas de définition formelle : de très nombreux descripteurs ont été proposés depuis les co-occurrences d'Haralick en 1973, mais chacun d'eux à un domaine d'usage restreint, sans que ce domaine ne soit lui-même bien défini. Aussi trouve-t-on des descripteurs de texture dans de nombreux outils pour le cas où... mais leur utilité n'est souvent pas prouvée.

Le troisième élément le plus utilisé est la forme. QBIC proposait ainsi une petite palette graphique : l'utilisateur traçait une forme et le système retrouvait les images possédant des formes semblables. Décrire une forme ne pose pas de problème majeur, la transformée de Fourier-Mellin convient tout à fait par exemple. La difficulté est de trouver la forme dans l'image, de la délimiter. En dehors des image ayant un fond uniforme, il n'existe aucune méthode permettant de trouver des formes de manière générale. Cela vient aussi du fait que l'utilisateur humain pense la forme comme étant celle d'un objet, d'une voiture par exemple, et lui associe donc une valeur sémantique. L'ordinateur ne possède pas de telle notion. Pourquoi la voiture serait-elle la forme a extraire, plutôt que l'essuie-glace, qui a au moins une couleur uniforme? Comment avait fait QBIC? Les formes de toutes images avaient été extraites manuellement...

On trouve deux issues possibles. D'une part, on peut renoncer à segmenter l'image et découper l'image selon un quadrillage prédéfini puis essayer de regrouper les carreaux entre eux. C'est simple, mais approximatif. D'un autre côté, de nouveaux algorithmes essayent de ne découper l'image qu'en quelques régions, moins d'une dizaine. Ces régions sont plus grossières, amis semblent bien mieux correspondre aux zones d'intérêt dont on a besoin pour décrire une image. Une fois ces zones délimitées, on peut soit les décrire par leur forme, mais celle-ci reste assez peu fiable, soit par leur couleur ou leur texture. On obtient ainsi une description de l'image en quelques zones qui paraît bien plus puissante que les descriptions globales utilisées jusqu'à présent.

Des descripteurs locaux. Une image peu être recadrée, tournée, agrandie, et on veut pouvoir la reconnaître tout de même. Les descripteurs globaux sont mal adaptés pour faire face à ces transformations. On peut alors chercher à n'utiliser que des descriptions locales. Pour cela, on va chercher des zones d'intérêt dans l'image, puis décrire chacune de ces zones. De manière générale, on peut chercher des points, des courbes ou des régions de l'image.

En ce qui concerne les régions, le problème est étudié de longue date. Les petites régions sont très instables. Celles correspondant à un objet rencontrent le problème évoqué à propos des formes. Reste l'option de ne rechercher que quelques régions dans une image, moins d'une dizaine en général. Cela permet de séparer le centre de l'image de ce qui est soit au premier soit au dernier plan. Mais il n'y a aucune assurance que les régions extraites aient

une quelconque valeur sémantique. Cela dit, on ne sait pas faire mieux ! Le problème est assez similaire pour les courbes : détecter une courbe est difficile. Entre deux images d'une même scène, il suffit d'un changement très mineur d'éclairage ou de point de vue pour que la courbe soit coupée en morceaux. D'autres part, les courbes sont reliées entre elles, et chaque intersection ou jonction pose un problème sans solution immédiate.

L'emploi de points d'intérêt s'est révélée bien plus fructueuse. Tout d'abord, on peut définir mathématiquement ce qu'est un point d'intérêt, par exemple comme étant un point qui ressemble le moins possible à ses voisins (un minimum d'auto-corrélation), ou un point porteur d'un maximum d'information au sens de la théorie de l'information. Ensuite, ces points résistent bien aux transformations que l'on peut faire subir aux images : on dit qu'ils sont répétables. La méthode de Harris modifiée par Schmid s'est révélée une des meilleures sur ce plan. Autour de chacun de ces points, on calcule alors des quantités décrivant le signal, en veillant à ce que ces quantités restent égales si l'on modifie l'image, par exemple, en la faisant tourner. Les deux méthodes les plus répandues et les plus performantes sont les invariants différentiels de Florack que Schmid est la première à avoir utilisé dans un but de description pour la recherche d'images et les invariants SIFT de Lowe.

Ces descripteurs locaux se sont révélés, dans la pratique, extrêmement performants pour des tâches de détection de copie par exemple. Les taux de fausse détection sont très bas. Cela est du au fait qu'ils décrivent de manière très discriminante la texture locale de chaque point. Ils permettent donc de retrouver des objets précis. C'est pourquoi ils sont si adaptés pour la détection de copies.

Vers des descriptions plus sémantiques. Avec toutes ces descriptions, qu'elles soient locales ou globales, on reste au niveau du signal. On aimerait pouvoir traiter les images avec un vocabulaire plus riche sémantiquement. Il n'y a, pour le moment, aucun moyen un tant soit peu générique de trouver des descriptions sémantiques dans les images. Deux voies sont donc explorées.

La première voie consiste à utiliser un apprentissage pour caractériser une classe d'objets ou d'images. Dans un tel processus, on part d'un ensemble d'images ou de régions exemples du concept que l'on veut reconnaître, plus éventuellement d'un ensemble de contre-exemples. On décrit l'ensemble de ces données avec un ou plusieurs descripteurs puis, à l'aide d'un logiciel d'apprentissage, on cherche une limite qui sépare les descripteurs des exemples de ceux des contre-exemples. Lorsque l'on est en présence d'une nouvelle image ou d'une nouvelle région, on calcule simplement son ou ses descripteurs puis on regarde que quel côté de la limite ils se trouvent. S'il est avec les exemples, on déclare avoir reconnu l'objet ou le concept, sinon, on déclare n'avoir rien retrouvé.

Il est évident que la qualité des réponse va largement dépendre de celles

des données utilisées lors du calcul de la limite. Si elles sont précises et surtout très représentatives de celles qui seront à reconnaître dans la suite, il y a quelques chances de succès. Sinon, le système fera beaucoup d'erreurs. La qualité des descripteurs et leur caractère discriminant joue aussi un rôle majeur. Les capacités d'un tel algorithmes dépendent aussi de la collection complète : dans un monde fermé (une collection homogène bien délimitée), on a plus de chance qu'un concept trouve une caractérisation signal simple que dans un monde ouvert, internet par exemple, où on peut trouver les images les plus diverses.

On peut ainsi mettre au point des systèmes de reconnaissance de différents concepts. Pour peu qu'on associe une étiquette à ces ensembles d'images ou de régions que l'on reconnaît, on peut donner l'impression de disposer d'un système de reconnaissance sémantique. En fait, le système ne fait qu'essayer de traduire ce concept en termes de bas niveau, cette traduction étant calculée en fonctions des données fournies lors de l'apprentissage (le calcul de la limite), et sa validité est remise en question dès qu'on l'utilise avec des données très différentes de ces dernières.

Sur le même principe, on peut faire de la propagation de mots-clés. À partir d'images annotées par des mots-clés, on essaye de calculer les caractéristiques signal de bas niveau des images caractérisées par chacun de ces mots-clés, puis par apprentissage, on définit une règle de reconnaissance. On utilise alors ces règles pour « reconnaître ces mots clés » dans de nouvelles images. Lorsqu'un mot clé est reconnu, on annote l'image correspondante. Bien entendu, cet étiquetage ne peut avoir la qualité d'une annotation manuelle.

L'autre voie possible pour obtenir plus de sémantique consiste à traiter des cas particuliers.

Des détecteurs spécialisés. Il existe deux cas d'indices visuels présentant un intérêt tel que de nombreuses études ont été menées spécifiquement pour les détecter et les reconnaître. Ce sont les visages (ou plus généralement les individus) et les lettres.

En matière de visages, il faut bien distinguer le fait de détecter un visage (dire s'il y a un visage dans une image ou non), du fait de localiser ce visage (dire où il est), du fait de pouvoir retrouver plusieurs fois le même visage dans plusieurs images, et enfin de celui de pouvoir reconnaître un visage connu. Chacune de ces capacités fait appel à des techniques qui peuvent être très différentes. La technique de base, celle qui permet la détection ne fonctionne bien que pour des visages vu de face et en plan moyen ou rapproché. Pour la reconnaissance, il faut que l'image ait encore une meilleure qualité.

En matière de texte inclus dans les images, il faut distinguer le texte ajouté, tel celui des sous-titres, de celui qui peut être inclus dans l'image d'origine (le texte des panneaux d'une manifestation). Dans de nombreuses

situations, ce texte apporte une très forte valeur ajoutée sémantique par rapport aux autres descripteurs utilisés, et il pourrait être facilement utilisé dans un système d'aide à l'annotation.

On pourrait imaginer d'autres descripteurs spécifiques. La détection des corps humains sont de bons candidats, d'autant que les applications existent (maisons médicalisées pour détecter la chute des personnes âgées par exemple). Les logos sont aussi des candidats intéressants pour le suivi de l'impact des marques par exemple. Pour les autres, il n'y a par encore de marché suffisant pour motiver (et financer !) des recherches spécifiques.

Combinaisons de descripteurs. Les images sont polysémiques. On peut de plus envisager différents types de description : les mots clés et concepts permettent d'avoir accès à une partie de leur information, un histogramme de couleur donne un autre type d'information, complémentaire du premier. Une difficulté est de faire cohabiter au sein d'un même système toutes ces descriptions, qui sont de type très différent d'un point de vue informatique, puis de pouvoir les faire collaborer. Il faut pour cela un formalisme capable de combiner les différentes facettes de la description d'une image, un langage de requête plus riche, puis un mécanisme de mise en correspondance des requêtes avec les description. Les graphes conceptuels fournissent un outil intéressant pour combiner les descripteurs. Pour les requêtes, le problème est de trouver un langage qui ne soit pas trop abscons, ou une interface qui permette une traduction efficace entre le langage naturel et le langage de requête interne au système. C'est là un sujet de recherche actif.

Avec de telles méthodes, on se rapproche d'une manipulation plus sémantique des images. Du point de vue des traiteurs d'images, on a donc formidablement progressé. En fait, on commence à obtenir uniquement une petite partie de ce que l'on trouve dans un texte : des mots qui désignent des parties de contenu. De tels mots permettent d'utiliser les images dans les moteurs de recherche les plus courant du web : c est effectivement une grande avancée. Mais on va aussi s'affronter aux mêmes limites que celles que rencontrent les textes : ambiguïtés, manque de précision... Dans le cas des images, la linguistique ne pourra pas nous aider.

2.2 Une fois la description effectuée

Décrire ne suffit pas. Un autre paramètre important est la taille des bases à gérer. Tant que l'on reste à des tailles modestes, il est toujours possible pour un ordinateur de passer toutes les images en revue pour résoudre une tâche. Pour des collections de plusieurs millions d'images, cela n'est plus possible si l'on veut garder des temps de réponses assez courts.

Le problème de l'indexation. On fait alors face à un double problème. Premièrement, les données numériques ne peuvent être gérées de manière

efficace par les systèmes de gestion de bases de données (SGBD) habituels. La cause est multiple. D'une part, ces données se présentent sous forme de vecteurs numériques de grande dimension, et on doit utiliser toutes les dimensions à la fois, ce que les SGBD ne savent pas bien faire. D'autre part, on ne fait pas de requêtes exactes, mais approchées : on cherche des images ayant des descripteurs voisins d'un descripteur requête, et non pas strictement égaux. Du coup, on ne cherche pas une valeur, mais les données proches d'une valeur. Dans un espace de grande dimension, les plus proches voisins d'une requête ont toutefois tendance à être aussi éloignés que ses plus lointains voisins, et pour s'assurer qu'un descripteur est bien le plus proche, il faut bien souvent scruter une bonne partie des données, ce qui est très inefficace.

On commence à connaître quelques solutions à ce problème, mais il reste à les évaluer sur de grands ensembles de données. L'idée de base est de confiner la recherche, c'est à dire de réduire le plus vite possible l'ensemble des données où l'on va être forcé de passer en revue. Si l'on veut vraiment aller vite, il faut accepter de n'obtenir qu'un résultat approximatif. Les algorithmes se distinguent alors sur leur manière de gérer cette approximation.

Le deuxième problème posé est celui des descriptions codées sous forme de graphes, comme ceux que l'on peut vouloir utiliser pour combiner différentes descriptions. Comparer des graphes est une opération intrinsèquement compliquée, et on sait pas indexer dans ce cas : on est, sauf exception, obligé de tout passer en revue. Il faut alors utiliser toutes les spécificités des graphes que l'on utilise pour arriver à accélérer les recherches.

Un composant crucial : l'interface. Autre source de problème : l'interface. Un système ne sert à rien s'il ne peut rendre service à un utilisateur. Dans le cas présent, il faut que l'utilisateur puisse exprimer son besoin d'information pour que le système retrouve les images pertinentes dans la collection. On se trouve souvent face à un fossé entre le besoin de l'utilisateur, qu'il peut exprimer en langue naturelle et qui est de très haut niveau sémantique, et les descriptions des images dont on dispose qui sont elles de bas niveau sémantique. C'est le *semantic gap* : comment traduire la recherche d'une image illustrant la mondialisation économique en terme d'images bleue, verte ou rouge ?

La manière la plus courante pour contourner ce problème est l'emploi du paradigme de la recherche par l'exemple : au lieu de formuler la requête textuellement, on fournit au système une image, charge à lui de trouver les images les plus ressemblantes. De nombreuses interfaces permettent de pondérer la prise en compte des différentes facettes de cette images requête, mais il faut bien reconnaître que cette fonctionnalité n'a aucun sens pour la plupart des utilisateurs.

L'utilité de la recherche par l'exemple dépend, bien entendu, des applications : c'est un mode de fonctionnement idéal pour la détection de copies, où

l'image requête est l'image suspecte, et où l'utilisateur ne souhaite pas avoir à décrire l'image pour savoir si elle vient de son stock ou non. Par contre, pour une recherche dans une grande collection d'une agence de photo, cela risque de n'être ni efficace, ni pratique. Divers stratagèmes ont été proposés : le système peut proposer des images issues de la collection pour débuter la recherche, puis demander à l'utilisateur de choisir une deuxième requête dans les résultats d la première recherche et ainsi de suite. Une variante (habituellement appelée *relevance feedback*) consiste à désigner les images les plus intéressantes du résultat fourni plutôt que d'en désigner une seule.

Autre aspect de l'interface, la manière dont elle va présenter les résultats. En ce domaine, l'imagination n'est guère au pouvoir. La plupart des systèmes présentent, quelle que soit la requête, un nombre fixe d'images à l'utilisateur, entre 12 et 20. Ces images sont ordonnées par ordre décroissant de pertinence. Ainsi, ces systèmes retrouvent toujours des images, même si elles n'ont aucun rapport avec la requête. À l'inverse, si de très nombreuses images sont pertinentes, le système n'en présentera qu'une douzaine. Quand à l'ordre selon la pertinence, force est de constater qu'il ne correspond souvent à rien pour l'utilisateur. Quelques propositions plus innovantes ont été faites, mais force est de constater que la communauté de recherche sur les interfaces ne s'est pas encore intéressée au problème.

3 La vidéo et son contexte

Les contextes applicatifs. En matière de vidéo, on retrouve la même diversité d'applications que pour les images fixes : des utilisations grand public avec des vidéos prises à l'aide d'un camescope numérique et l'utilisation domestique de contenu commerciaux, et des utilisations professionnelles, avec les deux grands marchés que sont la télévision et le cinéma, et des secteurs plus spécialisés comme la médecine, la vidéo surveillance...

Une différence notable est que le nombre d'images alors considéré est bien plus grand : une archive de télévision possède des milliards d'images si on compte la vidéo comme autant d'images fixes. Les volumes à traiter sont donc bien plus grands. Autre différence, les enjeux financiers sont bien plus importants, car via la télévision, de nombreux domaines sont concernés : la publicité, le sport, la politique... L'impact sur le public est plus fort. Corollaires, le piratage y est aussi bien plus actif (copies illégales de films), mais les médias sont aussi très surveillés (CSA). Enfin, en d'un point de vue plus technique, faire des recherche par l'exemple n'est guère envisageable !

En matière de cinéma, un des grands enjeux est la lutte contre le piratage des films et DVD, souvent même avant leur sortie en salle ou dans le commerce. La sortie d'un film est un événement. Contrairement à ce qui se passait avec les images fixes, l'enjeu n'est pas tant de reconnaître les copies, mais de savoir comment elles ont été réalisées et quelle filière est en cause.

C'est donc d'un marquage individuel des copies officielles d'un film dont on a besoin. Celui-ci doit être d'une grande robustesse, sachant que certaines copies sont faites par enregistrement au camescope (à 30 Hz) de la projection d'un film dans une salle (à 24 Hz), alors que la personne qui filme n'est pas toujours dans l'axe de projection. Pour le moins, le signal est alors très dégradé !

Autre domaine différent de celui des images fixes, le monitoring de la télévision. Le nombre de chaînes de télévision augmentant, il devient difficile de choisir, trier, analyser tout ce qui est diffusé, que ce soit de manière analogique ou numérique, de manière hertzienne, par le câble, le satellite ou via ADSL. La simple constitution d'un catalogue précis de ce qui est diffusé est un problème ouvert. Évidemment, certaines chaînes disposent en interne de ces catalogues, mais ils ne sont souvent pas publics. Mais d'autres chaînes n'en ont sans doute pas.

Structuration des vidéos. Face à une vidéo, dont la durée peut varier de quelques secondes à plusieurs heures, avant de pouvoir décrire quoi que ce soit, il y a un fort besoin de structuration. Cette structuration peut être une micro-structuration dont l'opération de base est de retrouver les plans de montage de la vidéo. C'est une opération maintenant bien rodée, pour autant que la vidéo ne contienne pas trop de transitions exotiques entre plans. Une fois les plans séparés, il est possible de les étudier plus avant. On peut vouloir retrouver le mouvement de la caméra, ce qui suppose le plus souvent de supposer que le mouvement dominant au sein de la séquence d'image est celui lié au mouvement de la caméra. Cette hypothèse est, par exemple, invalide lors d'un gros plan sur un coureur cycliste. On peut chercher des images clés, soit des images fixes, issues du flux ou reconstruites, qui représentent au mieux l'information contenues dans une séquence. On peut chercher les objets mobiles, c'est à dire ceux n'ayant pas le même mouvement que le reste de la scène filmée.

On peut aussi effectuer une macro-segmentation. Une première manière consiste à regrouper les plans en entités de plus forte granularité : scènes et séquences par exemple. La difficulté est que ces entités sont souvent définies d'une manière difficile à traduire dans un programme informatique. Ainsi le fait que deux plans soient filmés dans un « même lieu », ne permet pas d'en tirer un critère de décision non ambigu. Il existe quelques cas favorables très intéressants comme les journaux télévisés, ou la distinction entre scènes du présentateurs et reportages est assez claire et facile à trouver. Ce n'est certainement pas le cas dans les œuvres de fiction. À l'inverse, on peut partir du flux directement pour rechercher des entités de haut niveau, comme les divers émissions d'un programme de télévision.

Description des vidéos. Une fois le flux segmenté, on peut vouloir décrire tant les diverses entités extraites que les liens spatio-temporels entre ces entités. Pour ce qui concerne l'image, on peut alors utiliser les outils mis au point pour les images fixes. Ces outils peuvent être modifiés pour tenir compte de la redondance d'information présente dans toute séquence d'images. Parmi ces descripteurs, ceux concernant les visages et les textes sont, bien entendu, d'une importance toute particulière dans le cas présent.

L'enjeu principal se situe toutefois dans l'aspect multimédia des documents. La plupart des vidéos comporte une bande sonore, et par le biais de la parole, celle-ci est une source d'information énorme sur le contenu de la vidéo. Il suffit de regarder la télévision sans le son d'une part, sans l'image de l'autre pour s'apercevoir des informations que véhiculent ces deux médias : la structure du flux est fortement portée par l'image, mais le sens est lui fortement porté par la parole et donc le son. À l'analyse de l'image, il est donc indispensable d'ajouter une analyse du son : détection des plages sonores, de la musique de la parole, de bruits particuliers (jingles, sons-clés), segmentation entre les divers locuteurs, reconnaissance des locuteurs, transcription de la parole... La palette est aussi vaste que celle portant sur l'image, mais les difficultés sont propres : à l'image, les divers composants se cachent les uns les autres, sur la bande sonore, ils se superposent. L'information est aussi répartie très différemment : les images ne sont pas très nombreuses (de 24 à 30 par secondes habituellement, mais chacune est porteuse de beaucoup d'information et peut être étudiée pour elle même. À l'inverse, les informations sonores sont très nombreuses, 16 000, 24 000 par seconde, mais chacune n'est en fait porteuse que d'une information infinitésimale (pas plus qu'un pixel individuel d'une image, voire encore moins puisque beaucoup d'information est porté par les fréquences qu'on ne déterminer à partir d'un unique échantillon). Il faut regrouper ces informations pour pouvoir en tirer quelque chose. Mais que regrouper ensemble ? Quand commence les segments sonores intéressants ? Il y a là un vaste et actif champ de recherche.

Une autre difficulté vient du couplage entre les divers médias. Ceux-ci, et cela peut paraître surprenant au premier abord, ne sont pas fortement couplés. Dans une œuvre de fiction, entre deux scènes, la rupture dans la musique et dans l'image peuvent ne pas être exactement au même moment. La voix que l'on entend n'est pas forcément celle associée au visage que l'on voit (voix off par exemple), de même que le nom indiqué dans le sous-titre ne peut correspondre qu'à une des personnes visibles, les sous-titres ne correspondent pas exactement à ce qui est dit, voire même à la traduction de ce qui est dit. Si on prend en compte l'information textuelle disponible, par exemple le guide de programmes correspondant à une chaîne de télévision, on retrouve ce même phénomène de décalage : les horaires indiqués sont inexacts, les inter-programmes ne sont pas détaillés, certaines émissions peuvent être supprimées ou remplacées. Comment analyser tout cela ?

Si l'analyse de la partie image des vidéos ou celle des bandes sonores sont

des domaines déjà bien avancés, profitant d'années de recherche effectuées dans les communautés de traitement du signal et de traitement d'image, cette confrontation multimédia est un domaine tout à fait nouveau, mais tout à fait crucial. La manière de l'aborder est discutée : certains veulent attendre que l'analyse de chaque média soit avancée au maximum pour réduire les problèmes de fusion d'information. D'autres estiment qu'il faut au contraire prendre en compte au plus tôt les divers médias et que le contexte multimédia offre de nouvelles possibilités que l'on ne pourra exploiter si on en reste à des analyses séparées. La question est ouverte.

Plus encore que pour les images fixes, il faut un cadre pour regrouper toutes les informations que l'on peut extraire d'une vidéo. Des modèles issus des probabilités tels les modèles de Markov offrent un cadre intéressant bien adapté à la structuration et à la description des documents temporels. Mais en dehors de ces tâches, ils ne sont pas d'une manipulation aisée, ni pour l'indexation, ni pour l'interrogation de bases de documents, et encore moins pour la formulation de requêtes. Pour organiser toutes ces descriptions, des langages de description ont été proposés. Certains très généraux, tels MPEG'7, d'autres plus ciblés comme TV-Anytime. Basés sur XML, ceux-ci permettent de combiner des informations de toute nature, mais le fait de pouvoir tout combiner ne signifie pas que le résultat de cette combinaison soit vraiment utilisable ! En cette matière, MPEG'7 qui avait des objectifs trop vastes a probablement moins d'avenir que MPEG'21 ou TV-Anytime qui avaient des objectifs mieux définis au départ.

Quelle utilisation des descriptions ? Comment utiliser ces descriptions ? Cela dépend beaucoup des applications. Il existe des applications simples et très ciblées. Par exemple, un documentaliste travaillant sur l'annotation fine d'un flux de télévision passe une partie non négligeable de son temps à faire défiler ce flux à l'aide de fonctions avance et retour rapides pour retrouver les limites de divers plans de montage. Voila une occupation tout à fait fastidieuse qui pourrait bénéficier facilement d'une aide automatique afin de permettre au documentaliste de se consacrer à des tâches d'interprétation et d'annotation des contenus, ou de faire face à la multiplication des documents à annoter.

Pour les particuliers aussi, la recherche d'un enregistrement dans une pile de cassettes VHS est souvent un exercice périlleux à l'issue incertaine. L'arrivée de disques durs dans les magnétoscopes et autres *set-top boxes* offre des possibilités de stockage qui vont rapidement grandir et un système de gestion astucieux de cet espace est absolument indispensable.

4 En conclusion

Comme toute nouvelle technologie, l'indexation automatique des documents multimédias se présente sous deux aspects : possibilité de nouvelles applications, mais mise en cause de manières de faire antérieures. En matière d'emploi, si la première piste ouvre des opportunités, la deuxième suscite légitimement des craintes. Qu'en est-il dans le cas présent ?

L'indexation automatique offre des possibilités tout à fait prometteuses en terme de manipulation de grands volumes de données, en pouvant réaliser des tâches simples comme la détection de copie. Pour de nombreuses autres tâches, les algorithmes sont loin d'offrir des performances, en qualité des résultats, qui permettent d'envisager une documentation complètement automatique. Il y a là des raisons profondes et sérieuses : les langues naturelles ont été un des premiers objets d'études de l'informatique, et leur traitement est très loin d'être parfaitement maîtrisé. La perception d'un document fait appel à tout un univers culturel, social et personnel qui a un fort impact sur cette perception. Et qui est très difficile à appréhender et à manipuler par ordinateur.

Un autre facteur intervient aussi : le volume de documents multimédia générés et stocké augmente à une folle allure. Il est à craindre que la recherche n'avance pas aussi vite en ce moment. Que ce soit sur le web où à la télévision, sans parler de numérisation de fonds plus anciens, c'est plutôt la noyade qui guette que la pénurie de matière. C'est cela, plus que l'émergence de l'indexation automatique qui risque de changer l'environnement d'utilisation de ces documents, que ce soit par des documentalistes, des spécialistes de médias (journalistes, analystes, créateurs, diffuseurs...) que par le grand public. Puissent-ils tous trouver dans la boite à outils de l'indexation automatique les moyens dont ils ont besoin pour faire face et réellement profiter de cet afflux.