UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À

L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE DE LA MAITRISE

EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES

PAR

YOUSSOUFOU SIDIBE

UN SYSTÈME POUR L'ANNOTATION SEMI-AUTOMATIQUE DES VIDÉOS ET APPLICATION À L'INDEXATION

AOÛT 2009

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Mohammed-Lamine Kherfi, directeur du mémoire

Département de mathématiques et d'informatique de l'UQTR

M. Fathallah Nouboud, juré

Département de mathématiques et d'informatique de l'UQTR

M. Alain Chalifour, juré

Département de mathématiques et d'informatique de l'UQTR

SOMMAIRE

Le nombre de vidéos disponibles dans les ordinateurs personnels et sur Internet est devenu impressionnant. Il est donc devenu indispensable de développer des outils qui permettent d'organiser ce type de données afin d'en faciliter l'accès et de rendre leur localisation plus rapide et plus efficace. C'est justement à ce problème d'indexation et d'annotation des vidéos personnelles que nous nous sommes attaqués dans ce mémoire.

La plupart des outils déjà existants sont basés soit uniquement sur le contenu des vidéos soit uniquement sur leur description textuelle. Or, ni les uns ni les autres n'arrivent à bien répondre aux besoins de l'utilisateur. En effet, les attributs visuels encodent bien le contenu des vidéos mais ne réussissent pas à représenter leur sémantique. Les attributs textuels quant à eux encodent la sémantique mais pas le contenu. Nous avons donc décidé de développer un outil qui combine les deux types d'attributs.

Afin de pouvoir indexer les vidéos par le texte, ces vidéos doivent disposer d'une description textuelle. Dans la réalité cependant, ceci est rarement le cas. Certains chercheurs se sont intéressés à cette problématique en développant des outils pour l'annotation manuelle. Cependant, ceci peut être très lent et fastidieux. Un groupe de chercheurs a tenté de développer des mécanismes d'annotation automatique. Bien que l'annotation automatique soit beaucoup plus rapide que l'annotation manuelle, les techniques existantes sont encore loin d'être au point. Une troisième piste qui semble être prometteuse consiste en l'annotation semi-automatique. C'est cette piste que nous avons empruntée. Ainsi, nous avons développé un outil qui permet à l'utilisateur d'annoter un nombre très restreint de vidéos, puis nous utilisons un mécanisme qui propage les mots-clés de façon automatique à toutes les vidéos similaires.

Finalement, nous avons exploité un troisième type d'attributs dans l'indexation des vidéos. Ce sont les attributs physiques que nous extrayons de façon automatique à partir des vidéos. Ces attributs sont entre autres le nom du fichier vidéo, sa date de modification, sa taille, la durée de la séquence, le nombre total d'images, etc.

Nos expériences rapportées dans le dernier chapitre ont donné des résultats très concluants surtout lorsque la recherche combine les caractéristiques de bas niveau et celles de haut niveau.

REMERCIEMENTS

Tout d'abord, je remercie le Seigneur qui, par sa grâce, m'a permis de réaliser ces travaux à terme.

Je remercie également tous les membres de ma famille et tous mes amis qui m'ont soutenu et encouragé tout au long de mes études.

Je ne saurai trouver les termes adéquats, pour remercier mon directeur de mémoire Mohammed-Lamine KHERFI, pour l'encadrement qu'il m'a apporté durant toute la durée de ma maîtrise. Sa disponibilité, ses encouragements et son amitié sont les facteurs clés de la réalisation de ce travail.

Je tiens aussi à remercier les membres du jury qui ont bien voulu accepter d'évaluer ce mémoire.

Mes remerciements vont également à l'endroit de tous mes professeurs du département de mathématiques et d'informatique de l'Université du Québec à Trois-Rivières.

Un merci particulier à Chantal GUIMOND, secrétaire du département mathématiques et informatique, pour sa gentillesse.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS ET DES SIGLES

UQTR	Université du Québec à Trois-Rivières
BD	Base de Données
S	Seconde (Unité de temps)
Mo	Méga Octet (Unité de mesure)
MPEG	Moving Picture Experts Group (Standard ISO)
GMM	Gaussian Mixture Model (Mixture de Gaussiennes)

INTRODUCTION GÉNÉRALE

Avec l'évolution technologique, on assiste à une explosion des données multimédia (notamment les images et les séquences audiovisuelles). Cette explosion est due au faible coût des appareils de capture et des appareils de stockage (ordinateurs, disques durs de grandes capacités, appareils photos numériques, webcams, cellulaires, etc.). Face à l'abondance de ces données multimédia, il devient très difficile à un utilisateur de retrouver rapidement une information souhaitée, d'où la nécessité de développer des outils et des techniques efficaces pour la recherche multimédia.

Plusieurs moteurs de recherche et d'indexation ont vu le jour ces dernières années. Ces outils sont exploités dans plusieurs domaines [4] et permettent la recherche par le contenu (images et son) ou tout simplement la recherche par le texte (mots-clés).

Cependant, le domaine de l'indexation de la vidéo ayant vu le jour il y'a quelques années seulement, les systèmes conçus ne sont pas encore assez performants. Ces systèmes exploitent le texte et le contenu. Cela donne la possibilité d'effectuer des recherches qui peuvent être basées sur le texte rattaché à la vidéo (contenu sémantique) ou basées sur le contenu visuel en utilisant ce qu'on appelle les caractéristiques de "bas niveau" (couleur, forme, texture).

Le problème qu'on rencontre au niveau de la recherche par le texte, est soit qu'il n'y ait pas de texte permettant de décrire l'objet recherché, soit que le texte existant est subjectif, ce que nous verrons au chapitre 1.

Au niveau de la recherche par le contenu, un problème non moins important, est que c'est seulement les caractéristiques dites de "bas niveau" comme la couleur, la texture ou la forme, qui sont utilisées pour effectuer une requête. Pour formuler une requête, l'utilisateur doit entrer une vidéo requête, et la comparaison avec les vidéos de la base de données se fait à l'aide d'une mesure de similarité. Cette méthode de recherche basée

seulement sur le contenu s'avère insuffisante à cause du fossé sémantique. Le défi pour ces méthodes de recherche basées seulement sur le contenu purement visuel, est d'assigner une sémantique aux vidéos. L'une des méthodes possibles, pour assigner une sémantique à une vidéo, est l'annotation [1, 2, 3, 4, 6, 20].

C'est dans ce cadre, et pour répondre aux besoins grandissant des utilisateurs, que s'inscrit notre recherche. Nous avons réalisé un outil qui permet l'annotation semi-automatique des vidéos et son application à l'indexation.

L'annotation des données multimédia est une tâche qui consiste à assigner, à chaque document multimédia, ou à des parties de ce document, un mot-clé ou une liste de mots-clés permettant de décrire son contenu sémantique.

Dans le premier chapitre, nous essayerons de donner un aperçu général sur le concept de l'indexation de la vidéo. Nous verrons les motivations qui ont conduit à faire de l'indexation de la vidéo, les méthodes pour faire l'indexation de la vidéo et le besoin de faire de l'annotation.

Dans le deuxième chapitre, que nous intitulons « État de l'art », nous verrons en quoi consiste l'annotation proprement dite, les méthodes d'annotations existante, et pour finir, nous verrons quelques travaux apparentés au nôtre.

Nous présentons, dans le troisième chapitre, l'outil que nous avons développé, et expliquerons son fonctionnement. Nous verrons la manière dont on a procédé pour effectuer l'annotation de notre base de vidéos et l'application à l'indexation.

Dans le quatrième chapitre, nous présentons les expériences effectuées, l'analyse des résultats obtenus et nous en tirons quelques conclusions.

Nous terminons notre mémoire avec une conclusion générale.

CHAPITRE 1

GÉNÉRALITÉS SUR L'INDEXATION DE LA VIDÉO

1.1 Motivations : Pourquoi indexe-t-on les vidéos ?

Aujourd'hui, nous assistons à une explosion de la quantité des données multimédia. Comment se retrouver avec toutes ces données amassées, le plus souvent à partir du réseau Internet ou à l'aide d'appareils de captures qui coûtent de moins en moins cher ? De nos jours, les disques de stockage ont de très grandes capacités ; ce qui donne la possibilité de stocker un grand nombre de photos et de vidéos sur nos machines. Comment alors retrouver en un temps raisonnable, la donnée recherchée ?

Les systèmes de gestion de bases de données traditionnelles sont inadaptés pour effectuer des requêtes sur les documents multimédia. On constate un réel problème de recherche d'informations multimédia et plus particulièrement sur les vidéos personnelles. Que faut-il faire ?

C'est en tentant de répondre à toutes ces questions que les chercheurs de ce domaine se sont penchés sur ce problème et tentent encore aujourd'hui d'apporter des solutions adéquates. Ainsi, la meilleure solution, serait d'indexer nos documents multimédia. Dans le cas des données purement textuelles, plusieurs moteurs de recherche et d'indexation, assez puissants, ont vu le jour, comme par exemple : les moteurs de recherche Google, Yahoo, etc. Cependant, dans le cas des données multimédia, les techniques d'indexation deviennent plus complexes et surtout dans le cas d'une vidéo. La recherche basée sur le texte seulement n'exprime pas toujours les besoins de l'utilisateur car les mots sont parfois insuffisants pour exprimer un contenu visuel.

Ces dernières années plusieurs systèmes de recherche et d'indexation de vidéos ont vu le jour et leurs développements se poursuivent. L'intérêt étant de développer des systèmes adaptés pour faire face à une technologie galopante.

Dans les prochains paragraphes, avant d'aborder les techniques d'indexation proprement dites, nous parlerons de quelques domaines d'application faisant appel à l'indexation de la vidéo.

1.2 Les domaines d'application de l'indexation de la vidéo

La gestion des fichiers vidéo s'effectue de nos jours dans de nombreux domaines. Pour faciliter la recherche d'une vidéo donnée, il faut nécessairement l'indexer. Vu la facilité de se procurer des vidéos en grande quantité, même à domicile, cette tâche est devenue une nécessité pour gérer ses propres vidéos personnelles. Ainsi, on peut rencontrer l'indexation de la vidéo dans :

- Le domaine cinématographique où on peut trouver des films de plusieurs catégories tels que les films humoristiques, les films dramatiques, les films d'actions, etc. Afin de retrouver un film donné dans un bref délai, il est donc nécessaire de faire de l'indexation.

- La sécurité : beaucoup d'organismes utilisent de nos jours des cameras de surveillance dans le cadre de la sécurité. L'indexation permet de retrouver rapidement une vidéo qui aurait été prise lors d'un incident quelconque.

- Le domaine du sport : par exemple, résumer un match de soccer de façon automatique, extraire les faits saillants d'un match de hockey, etc.

- L'information : par exemple, produire le résumé d'un journal télévisé.

- etc.

Selon [8], on trouverait les vidéos dites de "tous les jours" c'est-à-dire les vidéos personnelles, la télévision, le cinéma, et celles utilisées par les spécialistes dans les domaines scientifiques et artistiques. Le domaine d'application est alors un atout important dans le cadre de l'indexation car les termes qui sont utilisés pour annoter une vidéo peuvent varier d'un domaine à l'autre. Les vidéos peuvent donc être classées selon l'*intention*, le *contenu*, la *production* et l'*usage*.

1.3 Comment indexe-t-on les vidéos ?

Avant de voir comment se fait l'indexation de la vidéo, nous allons voir ce que signifie ce terme. L'indexation consiste à représenter et à organiser efficacement le contenu des documents d'une base de données [9]. Ceci permet, par la suite, d'en faciliter l'accès et l'utilisation. L'indexation de la vidéo est donc le traitement que l'on fait subir à des documents vidéo, afin de pouvoir :

a) les retrouver à l'aide d'une recherche avec des requêtes,

b) les organiser en créant des catalogues de navigation,

c) extraire l'essentiel du document : soit par un résumé, une scène particulière, des personnages ou objets particuliers ou tout simplement des faits saillants.

Traitements communs à tous les types d'indexation

Que ce soit pour la recherche, le résumé ou la création d'un index de navigation, on doit toujours effectuer un certain nombre de traitements, tel qu'illustré à la Figure 1. Ces traitements sont :

1. La segmentation : qui consiste à découper une longue séquence vidéo en de petites séquences afin d'en faciliter le traitement.

2. La représentation et la classification : qui consiste à extraire les caractéristiques (appelées aussi signatures, attributs ou descripteurs) à partir des séquences vidéo. Ce sont ces séquences qui seront utilisées pour représenter les vidéos par la suite afin de les retrouver ou de les organiser.

3. La création d'un index : qui consiste à classifier les caractéristiques extraites afin d'accélérer la recherche.

4. La comparaison : elle consiste à comparer les caractéristiques de la requête et celles des séquences vidéo de la base de données.

5. L'interactivité : on crée une interface qui permet à l'utilisateur de formuler sa requête et qui servira à afficher les résultats de la recherche. Pour la navigation, l'utilisateur a besoin d'une interface lui permettant de communiquer avec l'index.

Tous ces points seront abordés en détail dans les sections subséquentes.

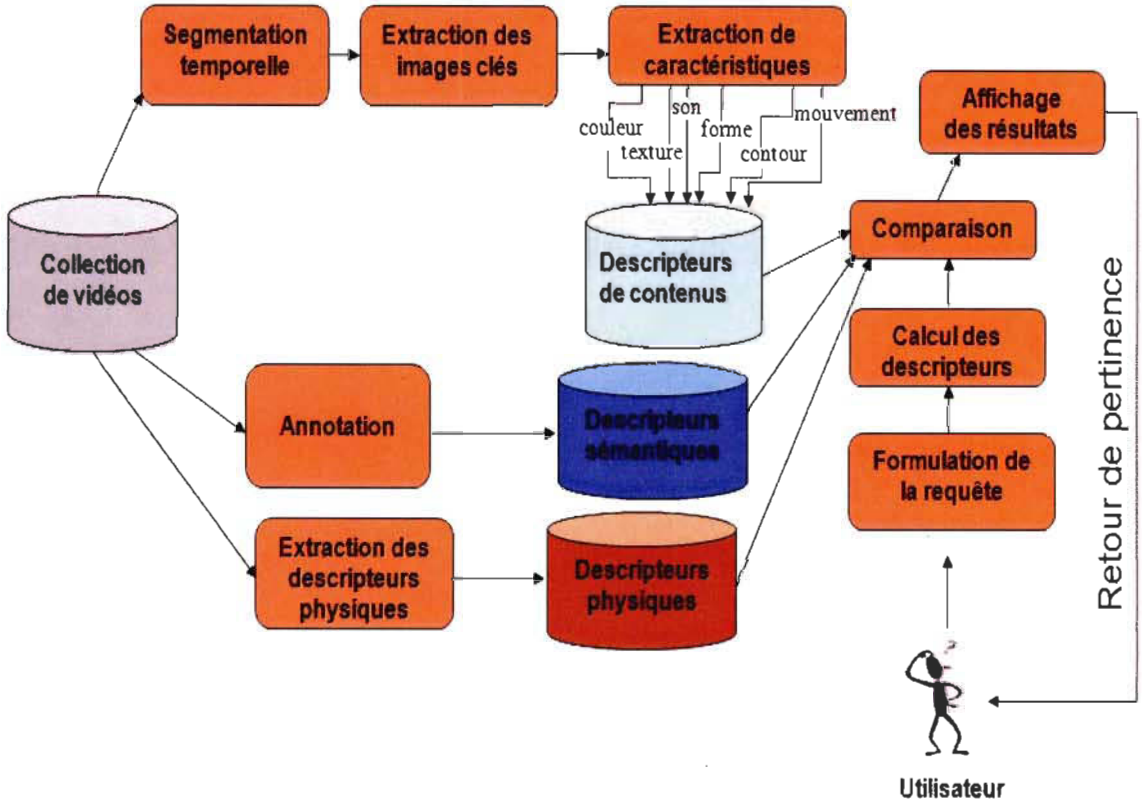


Figure 1 : Structure d'un système d'indexation et de recherche de vidéos

1.4 La segmentation

La segmentation de la vidéo consiste à diviser celle-ci en des séquences homogènes de courtes durées appelées des plans (*shots* en anglais) ou en des scènes. La segmentation est souvent la première phase à effectuer dans l'analyse et la recherche de documents audiovisuelles. C'est une technique qui permet, par exemple, de détecter les points de montage d'un film ou d'une émission télévisée ou de documentaires. Cette détection est très fastidieuse lorsqu'elle est faite manuellement. Plusieurs méthodes automatiques sont utilisées pour segmenter la vidéo. Entre autres, on retrouve dans la littérature des méthodes basée sur les différences de pixel à pixel, la comparaison d'histogrammes de

couleur, et l'estimation du mouvement. Notons cependant, qu'une étude détaillée des méthodes existantes dépasse le cadre de ce mémoire. Le lecteur peut trouver plus de détails dans [9,12, 13,15].

1.4.1 Pourquoi segmenter les vidéos ?

Une longue séquence vidéo contient trop d'informations (plusieurs thèmes, différentes personnes, plusieurs lieux, plusieurs objets, etc.). Conséquemment, on ne peut pas la décrire telle qu'elle est. Afin de pouvoir l'indexer, on doit d'abord la diviser en des scènes et chaque scène en des plans courts relativement homogènes. Une fois la vidéo segmentée, on peut décrire chaque scène ou chaque plan avec des caractéristiques tel que détaillé dans la section 1.5. Cette description permettra par la suite d'effectuer la recherche, l'organisation et de multiples opérations.

1.4.2 Comment segmenter une vidéo ?

La segmentation concrète d'une vidéo consiste à découper celle-ci. Pour analyser ou pour indexer le contenu d'une vidéo, on procédera à sa segmentation temporelle (Figure 2 [19]) qui consiste à la découper en scènes et les scènes en plans. Une scène sera définie comme un ensemble de plans. D'après [9], la scène peut se définir comme étant la plus petite unité sémantique d'un film. Cela peut être à titre d'exemple, *un dialogue*, *une conversation téléphonique*, une *action observée de plusieurs points de vue*, etc. Le plan quant à lui, peut se définir, comme étant une prise de vues sans interruption. L'assemblage des plans permet d'obtenir une scène ou une séquence vidéo complète. Les plans et les scènes permettent, entre autres, d'extraire des images-clés qui résument l'essentiel de ces plans ou de ces scènes.

Les images, les plans, les scènes et les séquences forment respectivement les unités hiérarchiques d'une vidéo.

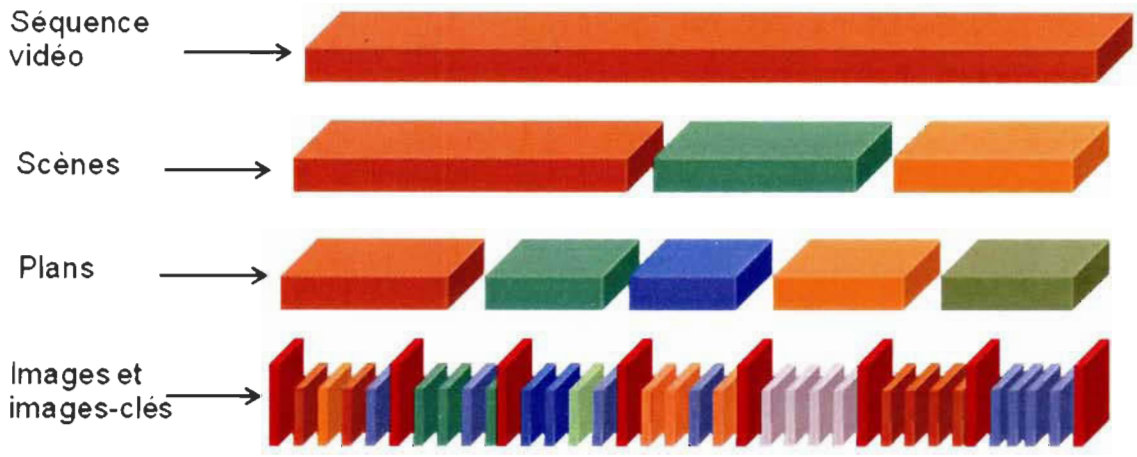


Figure 2 : La segmentation temporelle d'une séquence vidéo

Les caractéristiques utilisées pour la segmentation sont généralement les mêmes que les caractéristiques finales utilisées pour la recherche et l'indexation.

Dans ce qui suit, nous allons voir ce qu'est la segmentation en plans et ce qu'est la segmentation en scènes.

1.4.2 La segmentation en plans

Un plan est défini comme une séquence d'images durant laquelle l'acquisition du signal est continue. La durée d'un plan est généralement courte, de quelques secondes à quelques minutes. La segmentation en plans est la première étape à effectuer afin de procéder à une bonne analyse et à l'indexation du contenu d'une vidéo.

Une détection de changement de plan peut se faire soit en effectuant une différence d'images successives ou en utilisant le mouvement dans les images.

Ainsi, après avoir découpé une séquence vidéo en plans, on a la possibilité soit d'extraire les caractéristiques directement des plans, soit de diviser chacun de ces plans en images-clés, et finalement extraire les caractéristiques à partir de ces dernières. Les caractéristiques extraites serviront à l'indexation et à la recherche. Parmi ces caractéristiques, on retrouve la couleur, la texture, la forme et le mouvement.

Notons que même si la "segmentation en plans" est souvent incorrectement confondue à la "segmentation en scènes", les plans sont des sous-parties d'une scène.

1.4.3 La segmentation en scènes

La scène est définie comme un ensemble de plans vidéo et constitue une unité logique ayant un sens, c'est-à-dire compréhensible lors d'une visualisation; tandis que le plan peut s'avérer non significatif.

On note que l'on peut extraire les caractéristiques aussi bien des scènes que des plans qui les composent, voire même des images-clés contenues dans ces plans. Hanjalic et al. [13] indiquent qu'au lieu d'un plan, une scène peut aussi être utilisée comme unité élémentaire dans le cadre de l'indexation de la vidéo par le contenu.

1.5 La représentation

La vidéo étant une composition d'images et de son, les signatures qui en sont extraites sont les caractéristiques des images, du son ou du texte utilisé pour annoter la vidéo. D'une manière générale, la représentation d'une vidéo consiste à attribuer à cette vidéo une signature numérique qui la décrit de façon précise. Dans le cas d'une recherche, c'est cette signature qui permettra de localiser une séquence. Dans le cas d'une navigation (index), elle permettra de l'attribuer à la bonne classe et dans le cas d'un résumé, elle permettra d'en extraire l'essentiel.

Les caractéristiques doivent être extraites *a priori* (*offline*) afin que la recherche ne soit pas trop lente. Une fois qu'elles sont extraites, on les sauvegarde dans des fichiers qui seront utilisés lors de la recherche. Parfois, on organise les caractéristiques extraites au sein d'un index hiérarchique qui sera utilisé aussi bien pour la recherche que pour le catalogage.

On peut identifier trois grandes familles de caractéristiques utilisées pour l'indexation de la vidéo [18] :

- les caractéristiques physiques d'une séquence vidéo : sa taille, sa date de création, sa durée, etc.

- les caractéristiques sémantiques (appelées aussi caractéristiques de haut niveau) caractérisant la sémantique des vidéos; Elles sont généralement extraites du texte associé à ces vidéos.

- les caractéristiques visuelles (appelées également des caractéristiques de bas niveau) : les traits visuels que l'on peut extraire d'une vidéo tels que la couleur, la texture, le mouvement, etc. Elles permettent d'effectuer l'indexation par le contenu.

1.5.1 Les caractéristiques physiques d'une vidéo

Les caractéristiques physiques de la vidéo sont des caractéristiques alphanumériques qui peuvent être extraites automatiquement et qui donnent les informations de base sur la vidéo. Parmi ces caractéristiques, on retrouve le nom de la vidéo, son chemin, sa taille, sa date et l'heure de modification, le nombre total d'images de la vidéo, le format de compression utilisé, la longueur de la séquence en secondes, la résolution des images, le nombre d'images par secondes, etc.

Contrairement aux caractéristiques sémantiques, qui peuvent être parfois subjectives, les caractéristiques physiques sont objectives, dès qu'elles n'indiquent que des informations relatives à la vidéo et ne prêtent pas sujet à interprétation.

1.5.2 Les caractéristiques sémantiques d'une vidéo

Les caractéristiques sémantiques de la vidéo, aussi appelées caractéristiques de haut niveau, sont des informations qui donnent la description de cette vidéo selon l'interprétation humaine. Comme souligné dans [20], cette interprétation dépend de deux éléments :

- le niveau de connaissance et la perception de l'interprétateur,
- l'objectif que dégage l'interprétateur lors de l'observation.

Notons que les caractéristiques sémantiques sont étroitement liées au domaine d'utilisation et à la compréhension de l'interprétateur ; ce qui donne une certaine subjectivité aux caractéristiques.

Le texte, duquel sont extraites les caractéristiques sémantiques, peut provenir de plusieurs sources :

- texte entourant les vidéos dans le cas où les vidéos se trouvent à l'intérieur d'un autre document (ex. une vidéo se trouvant sur une page Web). On exploite alors le texte qui entoure la vidéo pour l'indexer. Ce texte peut être un ensemble de mots décrivant la vidéo, le titre de la vidéo, le nom de l'auteur, l'URL (*Uniform Ressource Locator*) de la page Web, etc. ;

- texte extrait des vidéos dans le cas de vidéos qui contiennent du texte à l'intérieur (ex. vidéos dont certaines scènes sont décrites avec du texte). On peut extraire ce texte et s'en servir pour l'indexation ;

- Annotation : on peut attribuer une sémantique à une vidéo en l'annotant. L'annotation, est un procédé indispensable dans l'indexation de la vidéo. Il s'agit d'affecter à chaque vidéo un certain nombre de mots-clés ou de phrases qui décrivent au mieux ladite vidéo, et cela de façon manuelle, automatique ou semi-automatique. Le texte utilisé pour l'annotation peut à son tour servir à l'indexation.

1.5.3 Les caractéristiques relatives au contenu

Ce sont des caractéristiques extraites directement du contenu de la vidéo sans recourir à du texte. Ces descripteurs utilisés pour la représentation du contenu multimédia sont des descripteurs dont les valeurs sont numériques [14]. Deux types de caractéristiques sont généralement utilisés : les caractéristiques qui décrivent l'aspect visuel de la vidéo et celles qui décrivent l'aspect sonore. Notons que ces caractéristiques sont plus objectives que les caractéristiques textuelles.

1.5.3.1 Les caractéristiques visuelles

Les caractéristiques visuelles les plus utilisées pour décrire un contenu vidéo sont généralement :

- **La couleur**

  La couleur est : l'« *impression que fait sur l'œil la lumière réfléchie par la surface des corps et qui nous les rend diversement sensibles* [16].» C'est l'une des caractéristiques la plus importante utilisée dans l'indexation par le contenu. La description de la couleur repose sur des techniques utilisées dans différents espaces de couleur. Dans la littérature, on rencontre le plus souvent les espaces de couleur RGB (*Red Green Blue* pour RVB = Rouge Vert Bleu) ou HSV (*Hue Saturation Value* pour TSV=Teinte Saturation Valeur). Notons qu'il est possible de passer d'un espace à l'autre à l'aide de certaines conversions.

  Les caractéristiques utilisées pour décrire la couleur sont entre autre : les histogrammes de couleurs, les moments de la couleur, etc.

- **La texture**

  La texture peut se définir comme étant un motif basique qui se répète dans l'image, ou des caractéristiques fréquentielles [40]. La texture est aussi un descripteur visuel très utilisé dans l'indexation de la vidéo par le contenu.

  Plusieurs techniques sont utilisées depuis quelques années pour l'analyser. L'une des méthodes les plus connues pour l'analyse de la texture est la matrice de cooccurrences de Haralick [13]. Plusieurs caractéristiques peuvent être extraites de cette matrice, dont par exemple : la moyenne, la variance, l'entropie, etc.

- **La forme**

  La forme peut être définie par le contour qui délimite une région de l'image. De ce fait la notion de forme est étroitement liée à celle du contour. Elle caractérise les objets contenus dans l'image. L'une des caractéristiques de détection de forme qu'on rencontre souvent est le descripteur de Fourier.

  Notons que le contour représente des points de l'image numérique qui correspondent à une variation brutale de niveau de gris ou une variation de couleurs des pixels. Pour plus de détails sur les caractéristiques de détection du contour, on peut se référer à [17].

- **Le mouvement**

  Comme la couleur, la texture ou la forme, le mouvement est aussi un descripteur très utilisé dans l'indexation et la recherche de vidéos par le contenu. Le mouvement peut être considéré comme le déplacement des objets ou le déplacement de la caméra lors de la prise. En même temps qu'il est considéré comme un descripteur de contenu, il sert aussi à faciliter la segmentation des vidéos.

  Pour estimer le mouvement des objets dans une vidéo, plusieurs techniques sont utilisées. Dans [21] par exemple, l'auteur étudie plusieurs méthodes d'estimation de mouvement telles que : le principe de conservation de l'intensité des pixels, la mesure du gradient, etc. La méthode la plus simple serait par exemple de comparer les images (frames) entre elles pixel par pixel en effectuant leur différence à l'aide de l'équation suivante :

$$d_{ij} = \begin{cases} 1, si \left| f(x,y,t_i) - f(x,y,t_j) \right| > T, \\ 0, \text{sinon}, \end{cases}$$

  où $f(x,y,t_i)$ et $f(x,y,t_j)$ sont des images consécutives d'une vidéo, x et y sont les coordonnées d'un pixel à un temps t_i donné et T est un seuil fixé.

Il y a mouvement d'objet lorsque la différence entre 2 pixels i et j est supérieure à 0. Lorsque la différence est nulle, cela indique qu'il n'y a pas eu de déplacement d'objet.

1.5.3.2 Les caractéristiques sonores

La bande sonore est également une caractéristique à part entière dans l'indexation de la vidéo. Pour la traiter, il faut d'abord la décomposer en ses composantes de base qui sont : la parole, la musique et le bruit.

Pour l'indexation de la bande sonore, on rencontre parmi les techniques utilisées : les différences acoustiques entre la parole et la musique, le taux de passage à zéro, la variation du flux spectral, etc. [22].

A la Figure 3, on a un aperçu rapide du mode d'indexation du signal sonore à l'aide de deux étapes : le prétraitement et la reconnaissance (Pinquier et al. dans [11]) :

Le prétraitement

Également appelé étape de paramétrisation, le prétraitement est une étape de recherche des caractéristiques les plus significatives. Le prétraitement acoustique est la phase de définition de l'espace de représentation de la parole ou de la musique. Ces deux caractéristiques sont en effet considérées comme des classes différentes et décomposées en Parole/Musique selon un modèle Classe/Non Classe (Figure 3).

La reconnaissance

Tel qu'indiqué dans [11], cette étape est constituée de deux sous-étapes qui sont la classification et le lissage. La classification est la phase qui succède au prétraitement acoustique. Il s'agit de classifier tous les paramètres (parole/musique) et les assembler

en segments. Le lissage quant à lui permettra par la suite de supprimer tous les segments qui ne seront pas pertinents à l'indexation.

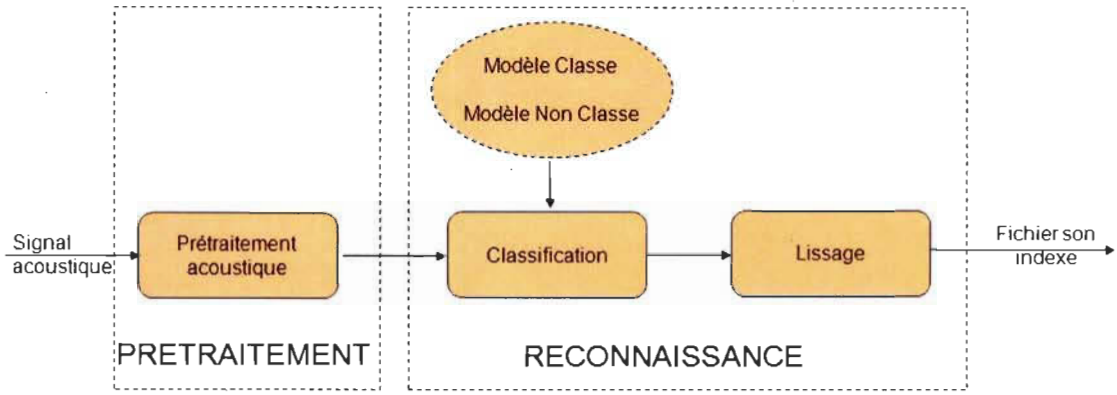


Figure 3 : Schéma général d'un système d'indexation basé sur le son [11]

Bien que le contenu de la vidéo inclue l'image et le son, nous nous limitons dans ce travail au contenu visuel. L'utilisation du son dépasse le cadre de notre spécialité. Elle est abordée par les chercheurs qui travaillent sur le traitement de la parole.

1.6 Mesures de similarité et comparaison

Les mesures de similarité sont indispensables pour toute indexation de vidéos. En effet, quand on fait de la recherche, les mesures de similarité sont utilisées pour comparer une requête et chaque vidéo de la collection afin de juger si cette dernière ressemble à la requête. Quand on crée un catalogue pour la navigation, on a également besoin de mesures de similarité qui permettent, dans le cas où des séquences vidéo se ressemblent, de les regrouper au sein d'une même classe.

Dépendamment des caractéristiques utilisées, différentes mesures de similarité peuvent être utilisées selon le texte, les attributs physiques ou les descripteurs de contenu. Cela est détaillé dans ce qui suit.

1.6.1 Mesures de similarité utilisées avec le texte

Pour retrouver des vidéos à l'aide du texte, on a besoin de mesurer la similarité entre le texte qui accompagne la requête et le texte qui accompagne les vidéos de la base de données. D'une manière générale, on utilise l'appariement (*matching* en anglais) binaire qui consiste à vérifier si un mot de la requête est présent ou absent dans une image de la BD.

En général, on rencontre des défis à surmonter :

- les synonymes : en composant une requête avec le mot « voiture», le moteur de recherche doit être capable de retrouver des vidéos dont la description comporte le mot « automobile » mais pas le mot « voiture », si l'automobile est l'objet de la recherche.
- la langue : un bon moteur de recherche doit permettre de répondre correctement à des requêtes formulées dans une langue autre que celle utilisée dans l'annotation des vidéos. Pour ce faire, il faut bien sûr recourir à des dictionnaires ou des traducteurs automatiques.

1.6.2 Mesures de similarité utilisées avec les attributs physiques

Avec les attributs physiques, on emploie souvent l'appariement (*matching*) étant donné que les attributs physiques sont des valeurs exactes et objectives (ex. : la date de création d'une vidéo). Notons cependant que l'utilisateur peut formuler sa requête avec des valeurs précises, par exemple « Je cherche une vidéo qui a été prise le 01/05/2005 » ou avec des intervalles de valeurs, par exemple, « Je cherche une vidéo qui a été prise entre le 01/05/2005 et le 28/05/2005 ».

1.6.3 Mesures de similarité utilisées avec le contenu

Les mesures de similarité qu'on rencontre dans l'indexation par le contenu sont entre autres : des mesures de distance, des mesures probabilistes, etc.

- **Les mesures de distance**

Les distances nous permettent d'exprimer une certaine ressemblance ou dissemblance entre les concepts comparés. Une distance mathématique doit vérifier les axiomes suivants : la symétrie, la séparation et l'inégalité du triangle.

De façon formelle, une distance d(x,y) sur un ensemble E, est une application $d : E \times E \rightarrow \Re^+$ telle que :

$$d(x,y) = d(y,x) \text{ (Symétrie)}, \ \forall x, y \in E$$

$$d(x,y) \geq 0 \, et \, d(x,y) = 0 \Leftrightarrow x = y \text{ (Séparation)}, \ \forall x, y \in E$$

$$d(x,z) \leq d(x,y) + d(y,z) \text{ (Inégalité du triangle)}, \ \forall x, y \in E$$

Les mesures de distance souvent utilisées dans l'indexation par le contenu sont soit la distance euclidienne ou la distance pondérée.

La distance euclidienne est définie comme la racine de la somme des carrés des différences des coordonnées,

$$d(x_1, x_2) = \left(\sum_{j=1}^{p} (x_1^j, x_2^j)^2 \right)^{1/2},$$

la distance pondérée est définie par :

$$d(x_1, x_2) = \left(\sum_{j=1}^{p} m_j (x_1^j, x_2^j)^2 \right)^{1/2},$$

où les x_1 et x_2 sont deux vecteurs de caractéristiques des vidéos à comparer. Les coefficients $m_{j>0\ (j=1,\ldots p)}$, strictement positifs, pondèrent l'influence de la $j^{ème}$ variable. Par exemple, les vecteurs peuvent être des vecteurs de moyennes des bandes R, G et B d'une vidéo.

- **Les mesures probabilistes**

Parmi les mesures probabilistes utilisées dans la recherche et l'indexation, on rencontre la divergence de Kullback-Leibler qui est considérée dans la littérature comme une très bonne mesure de similarité pour la recherche d'information dans les grandes bases de données.

Pour deux jonctions de densité P et Q de variables aléatoires discrètes, la divergence de Kullback-Leibler de Q à P est donnée par :

$$D_{KL}(P/Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

où les P(i) sont les différentes composantes de la distribution P et les Q(i) sont les différentes composantes de la jonction de densité Q.

1.7 La création de l'index

Une fois les caractéristiques des vidéos extraites, il faut procéder à l'indexation. Il s'agit d'organiser les séquences vidéo au sein d'une structure hiérarchique généralement appelée index.

Cet index sera utilisé à la fois pour la recherche et le catalogage. Concernant le catalogage, il est nécessaire de créer une interface qui permet de naviguer dans les différentes classes de l'index. En ce qui à trait à la recherche, l'index permet de réduire le temps de recherche puisque le moteur n'aura plus à visiter toutes les classes de la BD, il se limitera plutôt aux classes les plus vraisemblables. L'index aide également à améliorer la précision de la recherche puisque les classes visitées contiennent peu de bruit. Le principe utilisé est similaire à celui qui consiste à consulter la table des matières d'un livre, au lieu de le parcourir en entier pour trouver l'objet de notre recherche.

Nous verrons quelques techniques d'indexation dans le chapitre portant sur l'état de l'art.

1.8 La formulation de la requête

La formulation de la requête est une phase de communication entre l'utilisateur et le système de recherche. Selon [7], cette tâche, renferme une *problématique délicate* et un *énorme défi* :

1) Chez l'utilisateur, le défi étant de décrire les vidéos souhaitées *avec le peu d'outils dont il dispose;*

2) Au niveau du système, le défi est de bien interpréter la requête formulée par l'utilisateur.

Dans le contexte d'une recherche textuelle, lors de la formulation de la requête, l'utilisateur utilise des mots-clés ou des expressions. Afin d'obtenir les meilleurs résultats, les mots et expressions utilisés doivent décrire au mieux l'attente de l'utilisateur. Il se peut donc que la recherche donne des résultats non pertinents si la requête est mal formulée.

Dans le cas d'une recherche par le contenu, la requête peut être formulée de différentes manières. La plus connue étant la requête par l'exemple où l'utilisateur exprime ses besoins en fournissant un exemple de vidéo. Le système doit retourner toutes les séquences dont le contenu est similaire.

Dans le cas de la recherche par les attributs physiques, l'utilisateur formule sa requête en spécifiant soit le nom du fichier, la durée de la séquence, la date de la modification ou un intervalle de dates, un intervalle de durées, la taille, etc. Par exemple, l'utilisateur peut chercher toutes les vidéos dont la durée ne dépasse pas 20s ou des vidéos dont la taille est de 10Mo. Dans une requête formulée seulement avec les attributs physiques, on obtient des résultats lorsque la base de données renferme effectivement des vidéos qui répondent aux critères de la requête. Cependant, il se peut que le contenu de ces vidéos

soit complètement différent de ce à quoi s'attendait l'utilisateur comme on le verra dans le chapitre traitant des expérimentations.

1.9 Recherche et raffinement

C'est la phase d'interrogation de la base de données pour retrouver les vidéos à retourner à l'utilisateur. Après que l'utilisateur ait formulé sa requête, le système parcourt les fichiers de caractéristiques créés auparavant, ou l'index s'il y en a un, à la recherche de vidéos qui correspondent à cette requête. Ces vidéos sont ensuite présentées à l'utilisateur. S'il n'est pas satisfait des résultats, l'utilisateur peut raffiner sa recherche. C'est ce qu'on appelle dans la littérature le retour de pertinence, ou «*Relevance Feedback* » [5]. Ce processus oblige l'utilisateur à fournir plus de détails sur ce qu'il cherche et conséquemment le système à exploiter ces détails pour trouver des résultats plus pertinents.

1.10 Conclusion

Dans ce premier chapitre, nous avons vu quelques concepts généraux utilisés dans l'indexation et l'annotation de la vidéo. Nous avons vu, par exemple, que faire de l'indexation de la vidéo est surtout motivé par le besoin grandissant des utilisateurs qui ne cesse de s'accroitre à cause de la facilité par laquelle on peut se procurer des données audiovisuelles. On a vu également comment se fait l'indexation de la vidéo, comment s'établissent les différentes caractéristiques qui composent une séquence vidéo, et surtout reconnaître l'importance de l'annotation visant une meilleure indexation, car les moteurs de recherche exploitant seulement le contenu ne sont pas appropriés pour effectuer des requêtes textuelles.

Dans le prochain chapitre, nous discutons des différents types d'annotation, de notre choix d'annotation, et nous présenteront de quelques travaux réalisés dans ce domaine.

CHAPITRE 2

ÉTAT DE L'ART

2.1 Introduction

Nous venons de voir dans le chapitre précédent que pour faire une bonne indexation d'une vidéo, il est nécessaire d'effectuer de l'annotation, car la recherche par le contenu à elle seule n'est pas suffisante, et que pour combiner celle-ci à une recherche par le texte, la base de données doit être annotée. Dans ce chapitre, nous commencerons d'abord par expliquer en quoi consiste l'annotation d'une vidéo, ensuite nous verrons les techniques d'annotation existantes à savoir : l'annotation manuelle, l'annotation automatique et l'annotation semi-automatique. Pour terminer, nous verrons quelques travaux apparentés au nôtre.

2.2 Qu'est-ce que l'annotation de la vidéo ?

L'annotation de la vidéo est le processus par lequel on affecte à une vidéo des informations textuelles. Ces informations permettent par la suite de retrouver cette vidéo. Ces informations textuelles, qu'on appelle aussi métadonnées, sont des mots-clés se rapportant essentiellement à la vidéo annotée. Par exemple, on peut annoter une vidéo à l'aide de sa date de création, sa taille, la durée de la séquence, le nombre total d'images, le type de compresseur utilisé, le lieu géographique où la vidéo a été prise,

l'auteur, le temps qu'il fait au moment de la prise, l'évènement durant lequel la vidéo a été prise, etc.

D'une manière générale, l'annotation d'une vidéo consiste à associer à celle-ci, une sémantique afin de la référencer. Selon [8], l'annotation exprime deux aspects distincts à savoir : *la description* et *l'interprétation*. Dans la description de la vidéo, on peut retrouver tout ce qui explique le contenu de la vidéo (les objets, les personnes, les scènes qu'on y rencontre, les lieux, les événements, etc.). L'interprétation quant à elle permet de donner un avis pour expliquer une séquence donnée ou toute autre partie de la vidéo. L'interprétation est subjective et dépend surtout de la personne qui interprète (c'est par exemple, l'idée générale que l'on a par rapport au contenu d'une vidéo).

Nous verrons dans ce qui suit, la signification des termes « annotation manuelle », « annotation automatique » et « annotation semi-automatique ».

2.3 L'annotation manuelle

L'annotation manuelle, comme son nom l'indique, est faite manuellement par un humain dont le mandat est d'attribuer à chaque vidéo un ensemble de mots-clés. C'est une technique d'annotation qui permet d'effectuer des recherches assez précises, puisque les mots-clés sont directement entrés par un humain et peuvent être corrigés si le besoin se présente. Notons cependant que pour une base de données de grande taille, cette technique devient très fastidieuse voire impossible à effectuer par un être humain. Nous allons donc vite constater que l'annotation manuelle présente un certain nombre de limites. Mis à part la difficulté d'annoter une grande base de données vidéo, le texte qui est introduit par l'utilisateur peut être subjectif. Cela est dû au fait que l'utilisateur qui effectue l'annotation peut faire une interprétation différente de celle d'un autre utilisateur qui à son tour interprétera la même vidéo avec d'autres mots-clés. De plus,

cette méthode rencontre des limites dans la formulation des requêtes lors de la recherche, car l'utilisateur final peut formuler sa requête avec des mots-clés différents de ceux qui ont été utilisés par l'annotateur [36].

Pour pallier au problème de l'annotation manuelle, il est donc nécessaire de recourir à une autre méthode.

2.4 L'annotation automatique

L'annotation automatique est une tâche effectuée par la machine et elle permet d'alléger le travail de l'utilisateur. En allant dans le même sens que l'annotation automatique des images [35], on peut définir l'annotation automatique d'une vidéo comme étant le procédé par lequel un système informatique assigne automatiquement une légende ou des mots-clés à cette vidéo. Dans le cas de l'indexation et de la recherche par le contenu multimédia c'est un défi majeur, contrairement au cas des fouilles de données (*Datamining* en anglais), cette méthode connaissant de grandes avancées [9].

Pour effectuer l'annotation automatique d'une vidéo, il faut dans un premier temps extraire les caractéristiques qui décrivent ladite vidéo. Ces caractéristiques peuvent être des attributs physiques, sémantiques ou en rapport avec le contenu de la vidéo. Par la suite, il faut utiliser les techniques de propagation pour annoter d'autres vidéos considérées similaires. Pour ce faire, on peut utiliser, par exemple, l'apprentissage automatique [35].

L'annotation automatique a l'avantage d'être plus objective par rapport à l'annotation manuelle du moment où le travail est essentiellement fait par la machine. Alors pour économiser du temps et pour faciliter le travail de l'utilisateur, cette méthode d'annotation serait la mieux indiquée lorsqu'il s'agit de vidéos simples à annoter (c'est-à-dire contenant quelques objets) et qui sont faciles à reconnaître. Cependant, dans le cas

de vidéos complexes (beaucoup d'objets, de mouvements, de personnes) difficiles à reconnaitre, cette méthode atteint vite ses limites.

Selon une étude faite sur une grande base d'images (COREL [27]), l'annotation automatique a donné des résultats très variables pour un modèle donné [28]. Ce qui permet d'affirmer que cette méthode n'est pas précise dans certains cas.

On remarque qu'avec des images, la méthode ne donne pas de bons résultats, or la vidéo est encore plus complexe qu'une image et ne produit pas de bons résultats.

Parmi les techniques utilisées d'annotation automatique, on peut citer : la classification statistique, les réseaux de neurones, etc.

- **La classification**

La classification, comme son nom l'indique, est l'action de ranger par classe. C'est donc un procédé qui nous permet de savoir si un objet donné appartient à telle ou telle classe. Pour faire de la classification, on utilise le plus souvent des algorithmes d'apprentissage automatique. L'apprentissage automatique est un ensemble de procédés qui permet à l'individu de produire des règles automatiques à partir d'une base de données d'apprentissage. Ainsi, à partir d'exemples déjà traités, il est possible de prédire des situations similaires aux cas déjà traités.

Dans l'indexation et la recherche de fichiers multimédias (images et sons) cette technique est aussi utilisée, même si elle n'est pas encore très développée. Par exemple, on trouve des systèmes utilisant des algorithmes d'apprentissage automatique basés sur des modèles de Markov cachés dans la segmentation de vidéos [23, 38]. On a aussi l'exemple du modèle de classification utilisant la mixture de gaussiennes (GMM = *Gaussian Mixture Model* en anglais). Dans ce type de classification, on considère que

notre échantillon de données suit une loi de probabilité définie comme une somme pondérée de plusieurs distributions gaussiennes [9].

Entre autres méthodes utilisées dans la classification automatique, on peut aussi mentionner la méthode des k plus proches voisins (*k nearest neighbours* en anglais).

- **Les Réseaux de Neurones**

Les systèmes utilisant les réseaux de neurones s'inspirent des systèmes biologiques. Dans la littérature, les réseaux de neurones artificiels sont définis comme suit : « *des cellules interconnectées, où à chaque connexion entre 2 neurones on a attaché ce que l'on appelle un poids, qui n'est qu'un simple nombre réel. De plus, chaque neurone va traiter les informations grâce à une fonction d'activation* » [24]. Les poids sont ajustés par une méthode d'apprentissage couplée à une minimisation des erreurs de classification.

Les réseaux de neurones sont utilisés dans différents domaines et permettent, par exemple, d'effectuer la classification des vidéos, la reconnaissance des formes, la détection automatique des visages, etc.

Dans l'annotation automatique de la vidéo, les réseaux de neurones sont utilisés pour faciliter la classification des vidéos ou la reconnaissance des formes, etc. On peut citer, par exemple, les systèmes de reconnaissance de formes qui utilisent les réseaux de neurone pour identifier les plaques d'immatriculation. A l'aide de certains procédés mathématiques et la construction d'algorithmes, ces systèmes permettent de détecter l'image d'une voiture provenant d'une caméra, et ainsi de reconnaître le numéro de la plaque d'immatriculation de la voiture [25, 26]. Dans [40], les auteurs utilisent les réseaux de neurones pour faire l'analyse de la couleur ou de la texture afin d'identifier les différentes scènes d'une vidéo.

2.5 L'annotation semi-automatique

L'annotation manuelle, comme nous l'avons vu, exige beaucoup de temps et donc un coût de réalisation très élevé. L'annotation automatique par contre, facilite la tâche de l'utilisateur en lui permettant d'annoter assez rapidement plusieurs données. Seulement cette dernière méthode n'est pas très précise quant aux résultats escomptés comme nous l'avons vu précédemment.

Pour réduire l'effort de l'utilisateur et obtenir des résultats de recherche assez précis, il faut donc trouver une autre méthode d'annotation. La solution la plus simple pour surmonter les problèmes de l'annotation automatique et ceux de l'annotation manuelle, est de combiner les deux méthodes, d'où le terme : annotation semi-automatique.

Cette méthode nécessite l'intervention de l'utilisateur qui sera chargé d'annoter une vidéo donnée et d'effectuer, à l'aide d'un algorithme, une propagation des mots-clés au reste des vidéos de la base de données considérées comme étant visuellement similaires.

D'une manière générale, on distingue deux approches au niveau de l'annotation semi-automatique [29] :

- la "classification statistique d'images" (ou vidéos) qui, à partir d'une base pré-annotée, cherche à construire, par apprentissage, un modèle statistique, généralement un "prédicteur" dédié à un mot-clé donné. Ce modèle décidera par la suite, s'il faut ou non, annoter d'autres vidéos avec le même mot-clé.
- le "raisonnement par cas" qui, toujours à partir d'une base pré-annotée, permet de propager les mots-clés au reste des vidéos qui sont visuellement similaires à celles annotées. Dans ce cas, le système retourne à l'utilisateur le résultat de la recherche, et ce dernier pourra raffiner la recherche à partir d'un retour de pertinence (*Relevance Feedback*). L'utilisateur indique alors la pertinence du mot-clé suggérée

par la machine, ce qui lui permet de raffiner progressivement le processus d'annotation.

2.6 Travaux pertinents

Dans cette section, nous allons parler de quelques travaux effectués dans le domaine de l'annotation semi-automatique des vidéos. On parlera entre autres, des méthodes portant sur l'apprentissage actif, la détection des faits saillants dans une vidéo, etc.

2.6.1 Annotation semi-automatique avec apprentissage actif

L'apprentissage actif consiste à utiliser un algorithme d'apprentissage qui permet, à partir d'une base de données dont seulement une partie est annotée, de développer un modèle statistique qui permettra par la suite d'associer ces mots-clés au reste des vidéos de la base de données. En clair, il consiste à utiliser un système de classification qui sélectionne les échantillons les plus informatifs pour l'entraînement des systèmes [30].

Dans [31], les chercheurs utilisent l'apprentissage actif pour améliorer la performance de leur système. Ils utilisent la méthode d'annotation dite "cachée" (*Hidden annotation*). Selon les auteurs, cette méthode réduit considérablement l'écart entre les caractéristiques visuelles dites de bas niveau et les caractéristiques sémantiques dites de haut niveau. Dans un premier temps, un échantillon de la base de données est annoté à l'aide du module d'apprentissage qui se chargera par la suite d'annoter automatiquement le reste des données de la base. L'interface d'annotation se compose de 3 parties :

- une liste des objets de la base de données à annoter,
- une vue permettant la visualisation des images proposées par le système,

- un bouton "*Annotate*", qui permet d'annoter la liste choisie à partir d'un modèle existant.

Dans [32], Song et al. proposent un système qui fonctionne à l'aide de l'apprentissage actif avec plusieurs prédicateurs. Ils utilisent principalement deux ensembles de caractéristiques : un pour décrire la couleur, l'autre pour décrire la forme. Le processus de l'apprentissage actif fonctionne comme suit :

- Les vidéos sont d'abords segmentées en plans selon la similarité visuelle ou selon l'ordre temporel.
- A partir de chaque plan, on extrait un certain nombre d'images clés.
- Un échantillon de plans est annoté manuellement avant d'effectuer la propagation au reste des données.
- Deux prédicateurs sont entrainés en utilisant les caractéristiques (couleur et forme) extraites des vidéos utilisées pour l'apprentissage. Ces prédicateurs sont basés sur des mixtures de gaussiennes (GMM). Les paramètres de ces GMM sont estimés en utilisant l'algorithme "Expectation-Maximisation [32]". Cet algorithme permet d'annoter l'ensemble des plans restants.
- Pour finir, les plans de test sont pré-classifiés sur la base de la similarité visuelle et l'ordre temporel.

Dans des travaux antérieurs [37], Song et al présentent une méthode d'annotation semi-automatique de vidéos personnelles à l'aide de l'apprentissage actif. C'est une méthode où les caractéristiques de bas niveau suivantes sont aussi extraites : l'histogramme de la couleur, les moments de la couleur et l'histogramme de la couleur aux alentours des points de contour. Comme dans la méthode précédente, dans un premier temps, un échantillon de plans vidéo est annoté manuellement et par la suite, un algorithme d'apprentissage supervisé basé sur le modèle GMM permet d'effectuer la propagation aux autres plans. L'utilisation des plans au lieu des séquences vidéo permet de réduire considérablement le nombre de séquences ayant besoin de correction de la part de

l'annotateur dans le processus de l'apprentissage actif. La figure 4 résume cette procédure d'annotation.

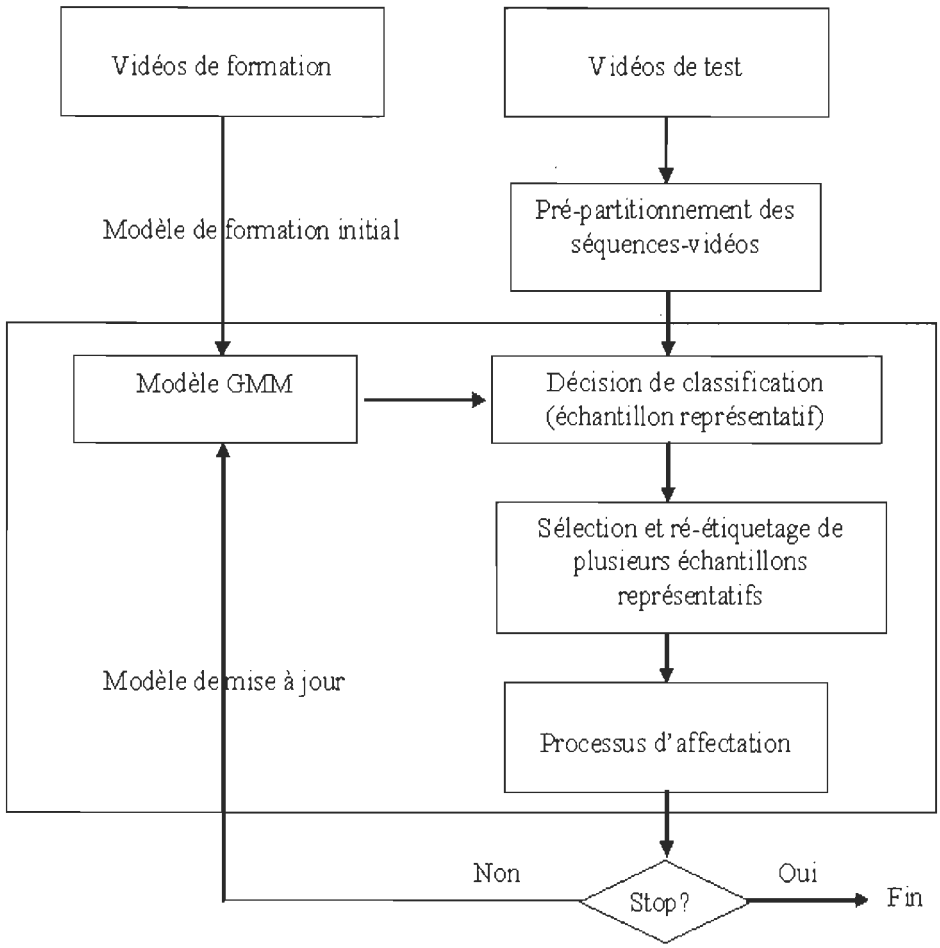


Figure 4 : Système d'annotation semi-automatique [37]

On rencontre dans la littérature d'autres méthodes d'annotation semi-automatique, comme [33], utilisant l'apprentissage actif avec pré-classification. L'algorithme de [33] effectue dans un premier temps la classification d'un groupe d'échantillons les plus représentatifs et ensuite, les informations sont propagées aux autres objets à l'aide d'un

modèle qui ne permet pas de répéter l'annotation d'un même groupe d'échantillons qui a déjà été annoté.

2.6.2 Détection des faits saillants dans une vidéo

Les faits saillants d'une vidéo sont des segments vidéo considérés significatifs. Dans un match de « soccer », par exemple, les faits saillants peuvent être : l'entrée de la balle dans les filets, certaines occasions de marquer, lorsque l'arbitre attribut une sanction, etc.

Les systèmes traditionnels d'édition de la vidéo suivent généralement les étapes suivantes afin de détecter les faits saillants d'une vidéo :

- parcourir la vidéo afin d'identifier les séquences les plus significatives,
- lire les séquences identifiées en avant et en arrière, afin de localiser les limites,
- effectuer un zoom, image par image, pour déterminer les limites de chacune.

Des travaux ([34]), présentent un système de navigation et d'indexation de vidéos personnelles où les vidéos sont organisées selon les images-clés (keyframes), les plans ou les scènes. Ceci permet de faciliter l'accès et la détection des faits saillants.

Dans [6], l'auteur propose une interface afin d'aider les utilisateurs à annoter les séquences vidéos de façon semi-automatique en un temps acceptable. Il développe un environnement qui permet d'éditer les séquences vidéo. L'objectif de cette édition est de détecter les faits saillants des vidéos. L'interface est composée de trois parties :

- une partie pour le contrôle des vidéos et des faits saillants.
- une partie pour les régions candidates (c'est-à-dire la possibilité de sélectionnée les régions significatives).

- une partie pour visualiser les images candidates.

La détection des faits saillants est faite comme suit :

- pendant qu'il visionne une vidéo, l'utilisateur peut cliquer sur un bouton pour dire que l'image (frame) actuelle appartient à un fait saillant.
- à partir de cette image, le système évalue les plans voisins pour en choisir quelques candidats susceptibles de renfermer des faits saillants.

2.6.3 Système d'annotation semi-automatique utilisant la norme MPEG-7

La norme MPGEG-7 est un standard ISO qui a été développé par les chercheurs du groupe « Moving Picture Experts Group (MPEG) » afin de pallier au problème de la recherche des contenus multimédias. MPEG-7, appelé « Multimedia Content Description Interface», est un système d'encodage permettant de décrire le contenu des données multimédias.

Dans [39], Vezzani et al. utilisent cette norme comme standard dans l'annotation des clips vidéos. Ils présentent un système de segmentation à structure hiérarchique et d'annotation automatique de vidéos par le biais de la normalisation des caractéristiques de bas niveau. Le fonctionnement de ces systèmes est brièvement décrit ci-dessous :

Étant donné une base de données de vidéos numériques, on suppose qu'il est possible de partitionner les clips vidéo en un ensemble de L classes $\{C_1, C_2, \ldots, C_L\}$ selon le contenu ou selon les différentes vues de la caméra. Compte tenu du grand nombre de clips utilisés dans l'apprentissage, on assigne à chacun d'eux une classe C_k, et de là, on effectue l'annotation automatique. Un clip inconnu, peut être classifié en utilisant la méthode des plus proches voisins ou d'autres mesures de similarité telle que la distance par exemple.

2.7 Conclusion

Dans ce chapitre, nous avons vu trois grandes familles de méthodes d'annotation : l'annotation manuelle, automatique, et semi-automatique.

En plus de sa subjectivité, l'annotation manuelle, n'est pas utilisable dans le cas de grandes bases de données puisqu'elle nécessite beaucoup de temps.

L'annotation automatique quant à elle n'est pas encore au point et ne fonctionne que dans le cas de vidéos très simples.

L'annotation semi-automatique semble être la meilleure piste puisqu'elle permet de combiner les avantages des deux méthodes précédentes et de pallier à leurs inconvénients.

Motivés par cela, nous avons concentré nos efforts sur le développement d'un outil d'annotation semi-automatique présenté au chapitre suivant.

CHAPITRE 3

NOTRE OUTIL POUR L'ANNOTATION SEMI-AUTOMATIQUE ET LA RECHERCHE

3.1 Introduction

L'indexation d'une vidéo peut s'appliquer à diverses catégories de vidéos, notamment les vidéos se trouvant sur le web, les vidéos biomédicales, les vidéos personnelles, les matchs de soccer ou encore les émissions de télévision. Puisque les techniques utilisées dépendent du type de vidéos visées, nous avons limités notre travail à une collection de vidéos personnelles. Les vidéos personnelles sont, entre autres, des vidéos familiales, des vidéos que nous avons prises nous-mêmes à diverses occasions, comme par exemple lors d'un camping, lors d'une randonnée ou lors d'une sortie avec des amis.

Nous avons développé notre approche avec le progiciel Matlab. Matlab est une contraction du terme anglais « *matrix laboratory* ». Matlab est un langage de programmation et il inclus un environnement de développement. Ce logiciel, complètement dédié aux calculs scientifiques et à la visualisation de données, a été développé par la société « *The MathWorks* »[1].

Notre travail se divise principalement en deux grandes parties, à savoir : la création d'un module de recherche qui permet à l'utilisateur d'effectuer et de raffiner sa recherche, et

[1] « The MathWorks développe et commercialise l'environnement de développement MATLAB, standard mondial des logiciels pour le calcul scientifique et technique. » http://www.mathworks.fr

la création d'un module d'annotation, que nous appellerons "fiche descriptive", à travers laquelle tout utilisateur pourra annoter directement des vidéos lors de leur sauvegarde.

Après la description du schéma de fonctionnement de notre système, nous décrirons les étapes que nous avons suivies pour élaborer nos deux modules et nous décrirons leurs fonctionnalités. Pour ce faire, nous décrirons l'application à l'annotation dans deux sections : une première pour l'annotation automatique et une deuxième pour l'annotation semi-automatique. Ensuite, nous décrirons les différentes options de recherche offertes par l'outil développé dans ce projet.

3.2 Schéma de fonctionnement

Le schéma présenté à la figure 5 explique le fonctionnement de notre application.

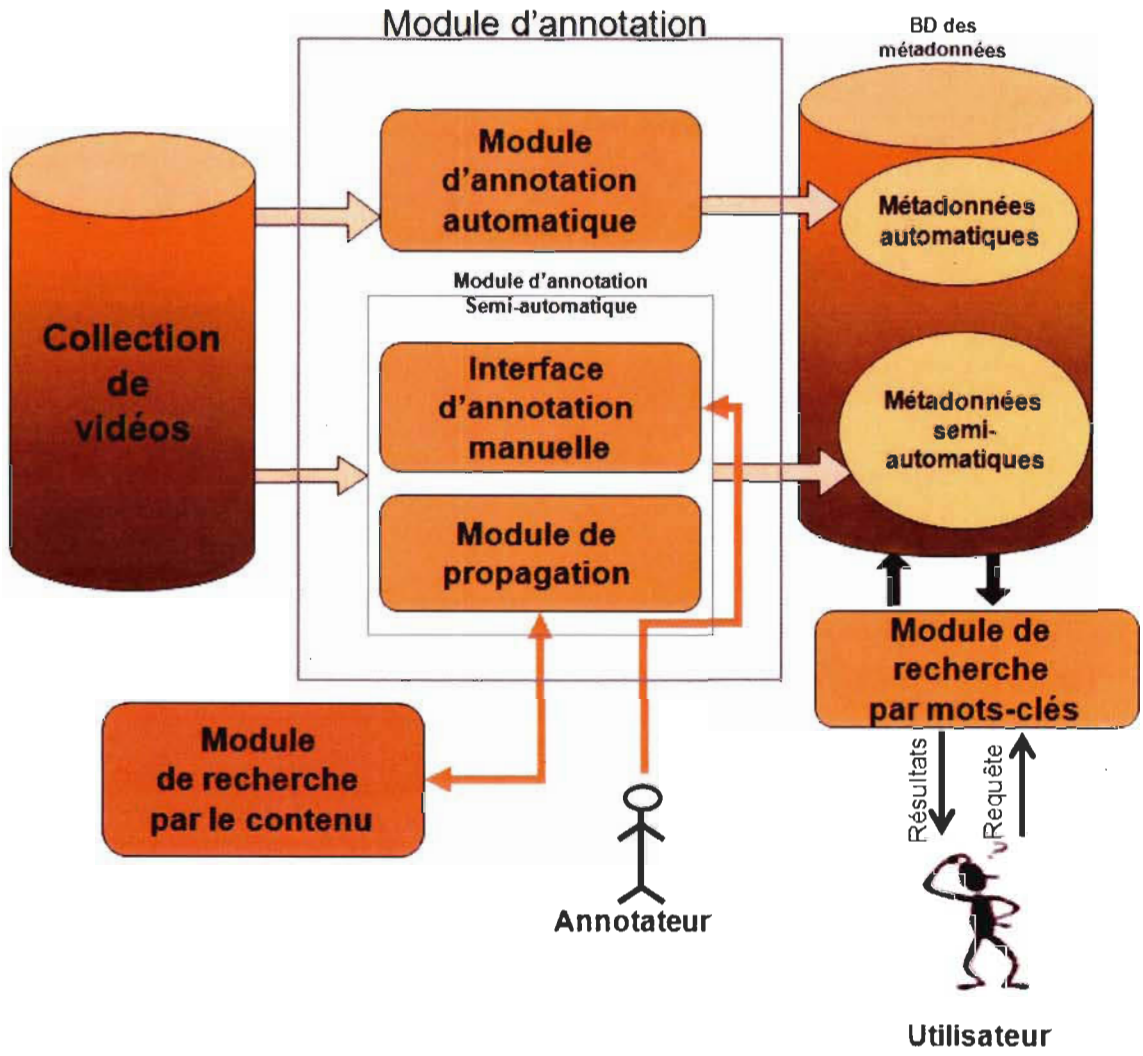


Figure 5 : Schéma de fonctionnement

Description détaillée du schéma de fonctionnement

- **Collection de vidéos.** Il s'agit de notre base de données de vidéos. Ce sont des familles de vidéos personnelles gratuites que nous avons collectées pour notre expérimentation.

- **BD des métadonnées.** Cette base de données est destinée à contenir les métadonnées physiques extraites par le module d'annotation automatique et les métadonnées semi-automatiques extraites par le module d'annotation semi-automatique.

- **Module d'annotation automatique.** Ce module effectue l'extraction des données fournies automatiquement par le système. Ces données physiques sont décrites dans la section 3.3.1 de ce chapitre.

- **Module d'annotation semi-automatique.** Ce module comporte deux sous modules : un pour l'annotation manuelle et un autre pour la propagation des mots-clés. Notons que le module de propagation des mots clés fait appel au module de recherche par le contenu comme on l'expliquera dans la section 3.3.2.2.

- **Module de recherche par le contenu.** Ce module de recherche est utilisé lors de la comparaison basée sur le contenu. A l'aide des caractéristiques tels la couleur, la texture, la forme, le mouvement, etc., ce module nous permet, à l'aide d'une vidéo, dite vidéo exemple, de retrouver dans la collection, toutes les vidéos visuellement similaires.

- **Module de recherche par les mots-clés.** Ce module permet à l'utilisateur d'effectuer des recherches basées sur les mots-clés.

- **Annotateur.** La personne qui attribue une sémantique aux vidéos.

- **Utilisateur.** La personne qui utilise l'application pour effectuer ses recherches.

3.3 Application à l'annotation

Afin d'assurer à l'utilisateur une certaine convivialité, nous avons élaboré une interface d'annotation très simple d'utilisation qui lui permet d'effectuer facilement l'annotation des vidéos (Figure 6). A partir de cette interface, l'utilisateur peut effectuer une annotation semi-automatique. Quant à l'annotation automatique, il peut la faire directement à partir de l'interface de recherche (Figure 10). Dans le chapitre 2, nous avons abordé les notions qui portent sur l'annotation. Nous avons vu, par exemple, que l'annotation manuelle, bien que facile d'utilisation, présente certaines limites compte tenu qu'il n'est pas aisé d'attribuer une sémantique à une grande collection de vidéos. L'annotation automatique, quant à elle, facilite le travail de l'utilisateur ; cependant elle n'est pas précise pour l'annotation d'une base de données multimédia. Pour l'annotation semi-automatique, nous avons vu qu'elle combine les techniques des deux types d'annotations précédentes, et de ce fait, elle est plus adéquate.

Dans les prochaines sous-sections, nous répondrons aux questions suivantes ;

- Comment accomplir l'annotation automatique afin d'extraire les métadonnées physiques des vidéos ?
- comment l'annotation semi-automatique, nous permet-elle d'attribuer les caractéristiques sémantiques à l'ensemble de nos vidéos ? Notons que cette annotation comporte deux étapes : l'annotation manuelle d'un échantillon de vidéos et la propagation des mots-clés au reste des vidéos de la BD à partir du moteur de recherche.

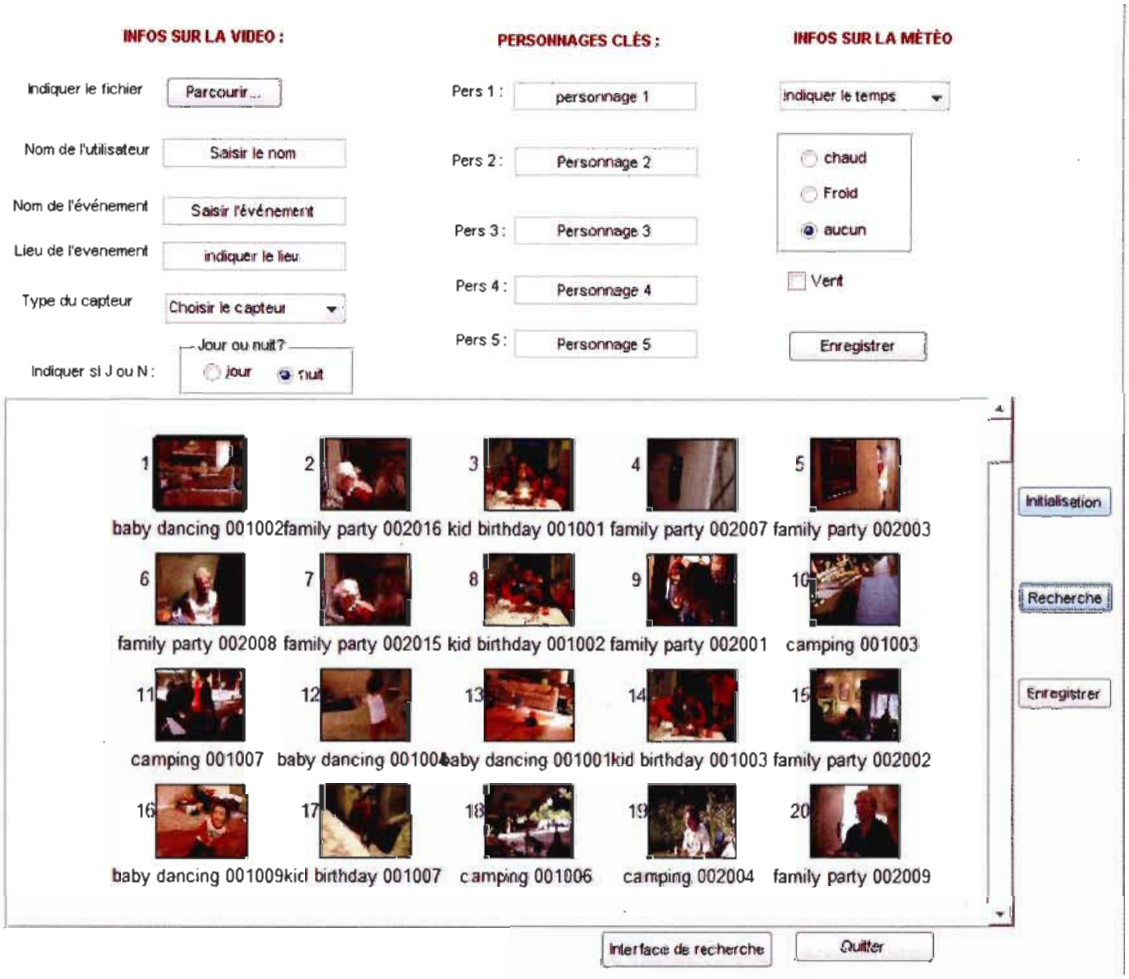


Figure 6 : Fiche descriptive permettant d'attribuer une sémantique aux vidéos.

3.3.1 Annotation automatique

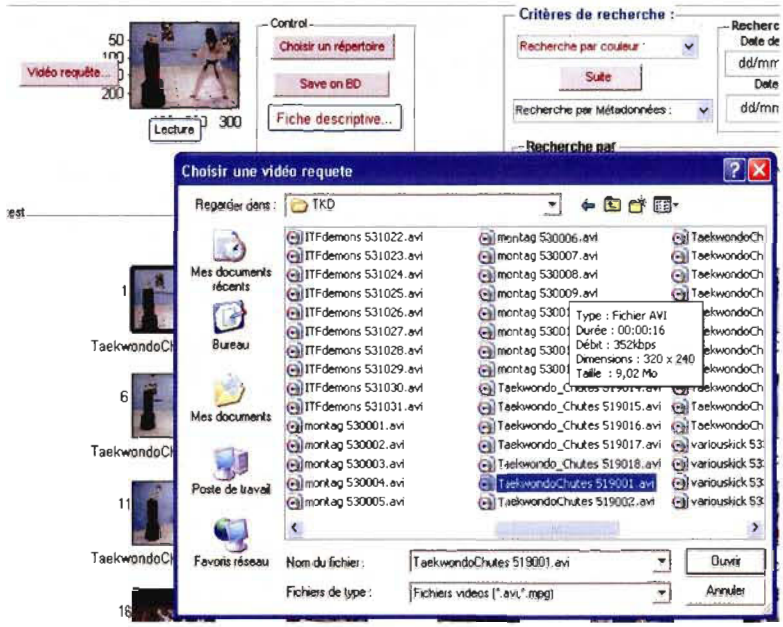
Dans cette section, nous discuterons de l'annotation automatique dans notre projet recherche. Dans notre projet, l'annotation automatique se limitera aux attributs physiques; une tâche réalisée par l'ordinateur.

Les métadonnées que nous extrayons automatiquement sont les caractéristiques physiques de la vidéo, soit :

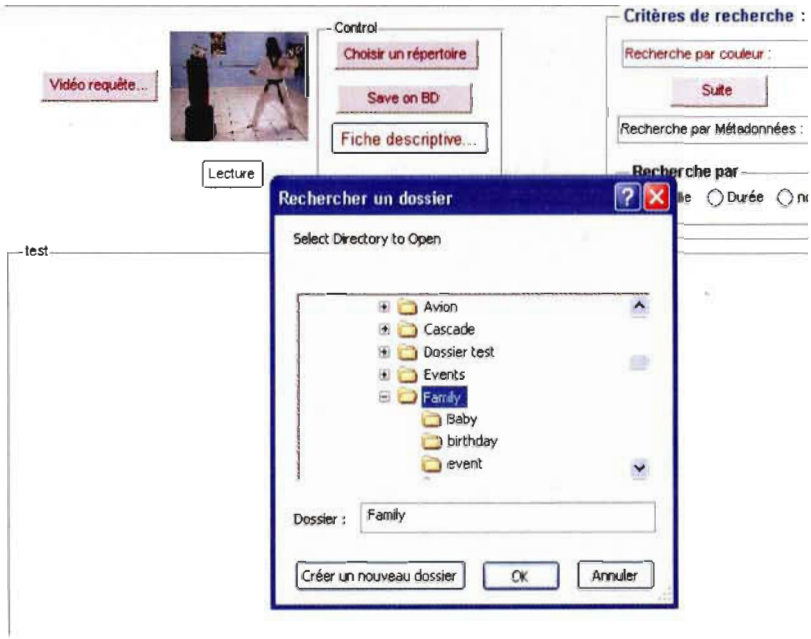
- **le nom du fichier vidéo** : nom donné à la vidéo,
- **le chemin** : chemin qui indique l'emplacement physique de la vidéo dans l'ordinateur,
- **la taille** : la dimension du fichier vidéo sur le disque,
- **la date de modification** : dernière date à laquelle le fichier vidéo a été modifié,
- **le nombre total d'images de la vidéo** : nombre total d'images contenues dans la séquence vidéo.
- **le format de compression utilisé** : mode de représentation des données de la vidéo. A titre d'exemple, on peut rencontrer les formats *.avi, .mpg, .mov*, etc.,
- **la durée** : période de temps correspondant à la durée du fichier vidéo, soit la longueur de la séquence exprimée en secondes,
- **la résolution des images** : qualité des images contenues dans la vidéo, exprimée en pixels par pouce (ppp),
- **le type de compression** : norme de compression utilisée,

- **le nombre d'images par seconde** : débit de la séquence vidéo, soit le flux d'images en par seconde.

La procédure que nous utilisons pour effectuer l'annotation automatique est la suivante : étant donnée une base de vidéos, le programme nous permet de choisir une vidéo cible (Figure 7 (a)) ou une famille de vidéos (Figure 7 (b)) et de sauvegarder les métadonnées physiques (chemin du fichier, nom, date de modification, le nombre total d'images, le nombre d'images par seconde, la taille, la durée de la séquence) dans la base de données des descripteurs. Ces informations serviront de critères de recherches.



(a) Annotation d'une vidéo



(b) Annotation d'un répertoire de vidéos

Figure 7 : Annotation automatique d'une vidéo ou d'un répertoire de vidéos

3.3.2 Annotation semi-automatique

L'annotation semi-automatique est effectuée en deux étapes : dans un premier temps, on effectue l'annotation manuelle d'un échantillon de vidéos de la base de données ; ensuite, on propage les mots-clés de chaque vidéo au vidéo qui lui sont visuellement similaires.

3.3.2.1 Annotation manuelle

Comme indiqué dans le chapitre sur l'état de l'art, l'annotation manuelle consiste à associer manuellement des mots-clés aux vidéos cibles. Pour ce faire, l'annotateur doit utiliser une fiche descriptive (Figure 6) pour faire l'annotation.

L'interface d'annotation manuelle (Figure 6) que nous appelons « fiche descriptive » nous permet de recueillir les métadonnées suivantes :

- **Le nom de l'annotateur**. Nom de la personne ayant effectuée l'annotation.
- **Le nom de l'événement**. Identification de l'événement durant lequel la vidéo est prise. Le nom de l'événement peut être, à titre d'exemple, « randonnée au bord de la mer » si tel est le cas traité.
- **Le type de l'événement**. Ajout de détails liés à l'événement. Un type d'événement peut par exemple être « découverte de la mer », « compétition », etc. L'ajout d'un tel descripteur peut aider à préciser l'événement.

- **Le lieu de l'événement**. Lieu de l'événement où la vidéo a été prise; une ville, un pays ou tout lieu qui peut s'avérer utile lors de la recherche.

- **Les personnages-clés de la vidéo**. Personnes ou objets importants contenus dans la vidéo et permettant de la décrire.

- **La météo**. Temps qu'il faisait lors de la prise de la vidéo; pluie, neige, vent, temps nuageux ou beau-temps.

- **Vidéo prise de jour ou de nuit**. Vidéo prise le jour ou la nuit.

- **Vidéo prise à l'intérieur ou à l'extérieur**. Vidéo prise à l'air libre ou dans une pièce.

Le bouton « parcourir » de la Figure 8 (a) permet à l'utilisateur de parcourir manuellement ses répertoires et de choisir le fichier qu'il souhaite annoter. Par la suite, l'utilisateur peut attribuer une sémantique à la vidéo en saisissant les informations nécessaires proposées par l'interface.

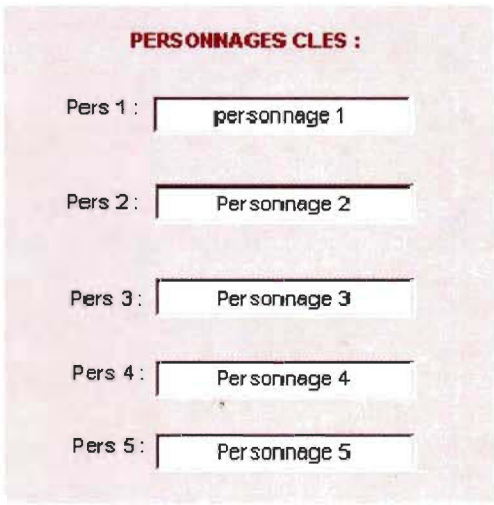
- Pour les informations générales relatives à la vidéo, l'utilisateur aura à saisir les informations demandées à la Figure 8 (a).

- Pour décrire les personnages-clés de la vidéo, l'utilisateur aura à saisir les informations demandées à la Figure 8 (b).

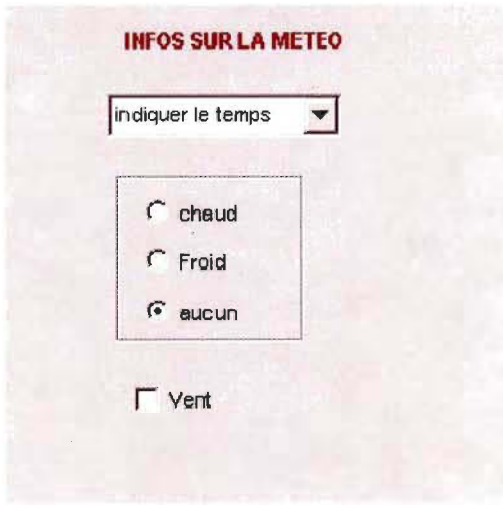
- Concernant la météo, il aura à fournir les informations demandées à la Figure 8 (c).

- Une fois les sémantiques introduites, le bouton «Enregistrer » de la fenêtre principale (Figure 6) permet les sauvegarder dans la base de données des descripteurs.

(a) Choix d'une vidéo cible, saisie d'informations de base, et choix du temps.



(b) Saisie des personnages clés



(c) annotation manuelle de la météo

Figure 8 : Annotation manuelle d'une vidéo

3.3.2.2 Propagation des mots-clés

Le module de recherche et de propagation nous permet d'effectuer une recherche en sélectionnant une vidéo cible (Figure 9). Une fois la vidéo cible sélectionnée, le bouton « Recherche » permet à l'utilisateur de rechercher et d'afficher toutes les vidéos qui sont visuellement similaires à la vidéo requête. Le bouton « Enregistrer » permet de propager les mots-clés qui accompagnent la vidéo cible à toutes les vidéos qui lui sont similaires à l'aide de mesures de similarité que nous verrons à la section 3.4.1.1. Avant la sauvegarde, l'utilisateur peut, s'il le souhaite, visualiser les vidéos obtenues de la recherche à l'aide du lecteur intégré à l'interface. Ne saurait été la contrainte de temps que cela peut engendrer, la visualisation permettrait à l'utilisateur d'être certain d'avoir introduit la bonne sémantique à toutes les vidéos annotées.

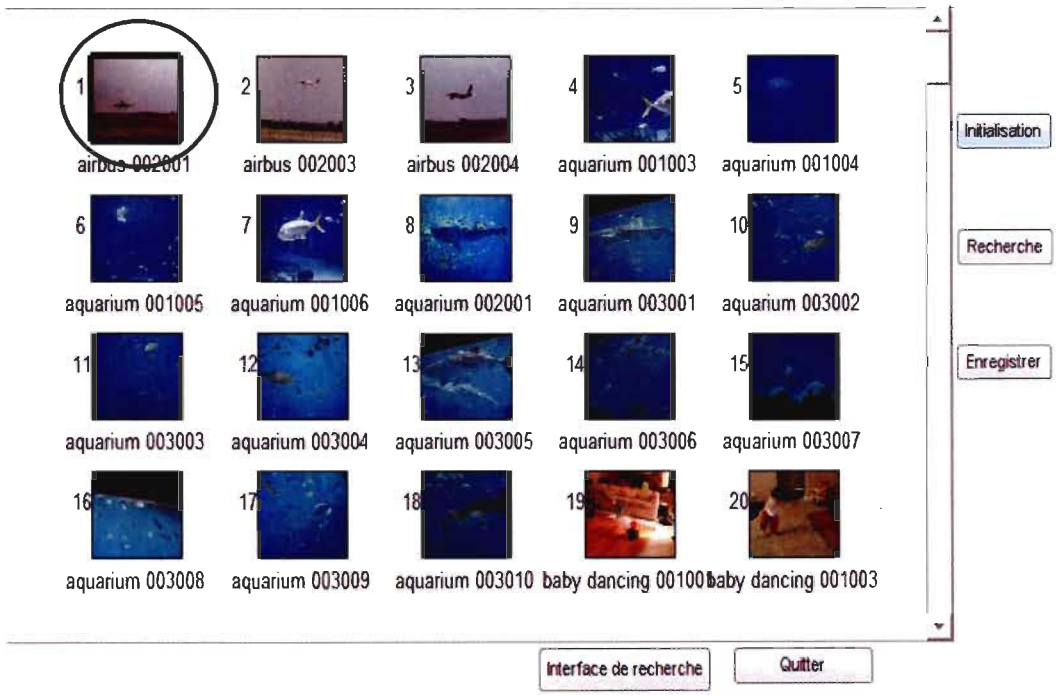


Figure 9 : Annotation semi-automatique d'une vidéo

3.4 Application à la recherche

Après avoir décrit l'outil développé d'annotation, nous décrirons l'outil de recherche (Figure 10). Toutes les vidéos, ayant été annotées par mots-clés à partir de l'interface d'annotation, sont accessibles à travers l'interface de recherche selon des critères que nous allons décrire dans les prochaines sections.

Notons qu'à partir de l'interface de recherche, nous pouvons accéder directement à l'interface d'annotation automatique à l'aide du bouton « Fiche Descriptive ». Également, une fois une vidéo requête choisie, ou un répertoire de vidéos choisi, nous pouvons effectuer l'annotation automatique à partir de cette interface à l'aide du bouton « Sauvegarder dans la BD », ce qui permet, lors d'une recherche, d'avoir la possibilité d'annoter ou de raffiner l'annotation.

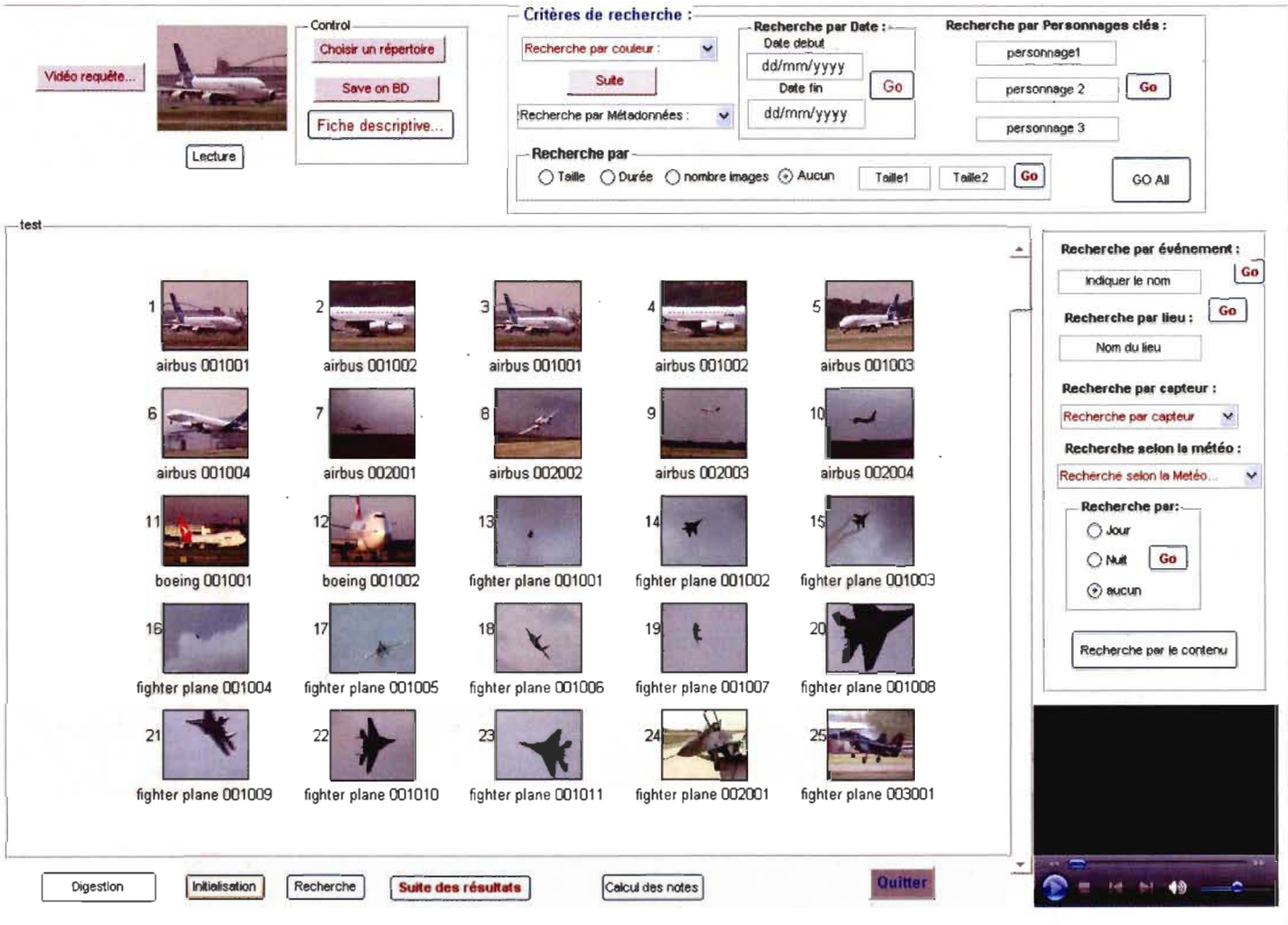


Figure 10 : Interface de recherche du système

3.4.1 Fonctionnement de l'interface de recherche

Dans cette section, nous décrivons les options de recherche de l'interface, ainsi que leur fonctionnement.

L'outil développé offre les modes de recherche suivantes :

- la recherche par le contenu par vidéo-exemple,
- la recherche par métadonnées physiques (date, taille, etc.),
- la recherche par métadonnées sémantiques (personnages-clés, événement, météo, lieu d'événement, dispositif d'acquisition de la vidéo, etc.).

Chacun de ces modes de recherche sera détaillé plus loin. Notons finalement que notre outil permet à l'utilisateur de combiner plusieurs modes de recherche au sein de la même requête.

3.4.1.1 Recherche par le contenu

La recherche par le contenu est essentiellement basée sur l'aspect visuel des vidéos. Pour comparer une vidéo requête au reste de la collection, nous utilisons, les caractéristiques suivantes :

- les moments de la couleur RGB,
- l'histogramme de la couleur RGB et HSV,
- l'histogramme de la couleur aux alentours des points de contour,
- le pourcentage des points de contours,

- la texture décrite par la matrice de cooccurrence (la moyenne, la variance, l'homogénéité et l'entropie),

- le mouvement (la variance de la moyenne des couleurs).

Pour effectuer une recherche par le contenu, on procède comme suit :

1) l'affichage d'un échantillon des vidéos présentes dans la base de données sélectionnée à l'aide d'un bouton d'initialisation activé par l'utilisateur.

2) l'utilisateur sélectionne une des vidéos sélectionnées comme requête puis lance la recherche à l'aide du bouton « Recherche » (Figure 11). Par exemple, on sélectionne la vidéo encerclée à la Figure 11 comme vidéo requête (R1), l'outil cherche toutes les vidéos qui lui ressemblent visuellement et affiche les résultats (Figure 12); la mesure de similarité utilisée étant la distance euclidienne.

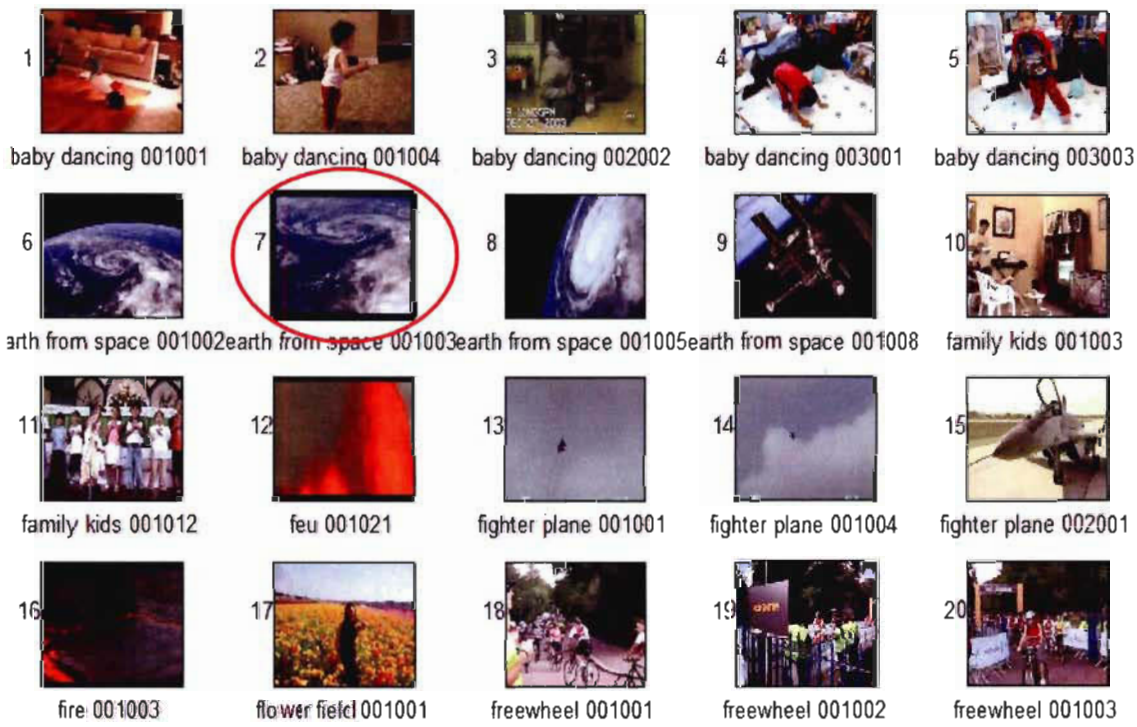


Figure 11 : Initialisation et affichage d'un échantillon de vidéos de la BD

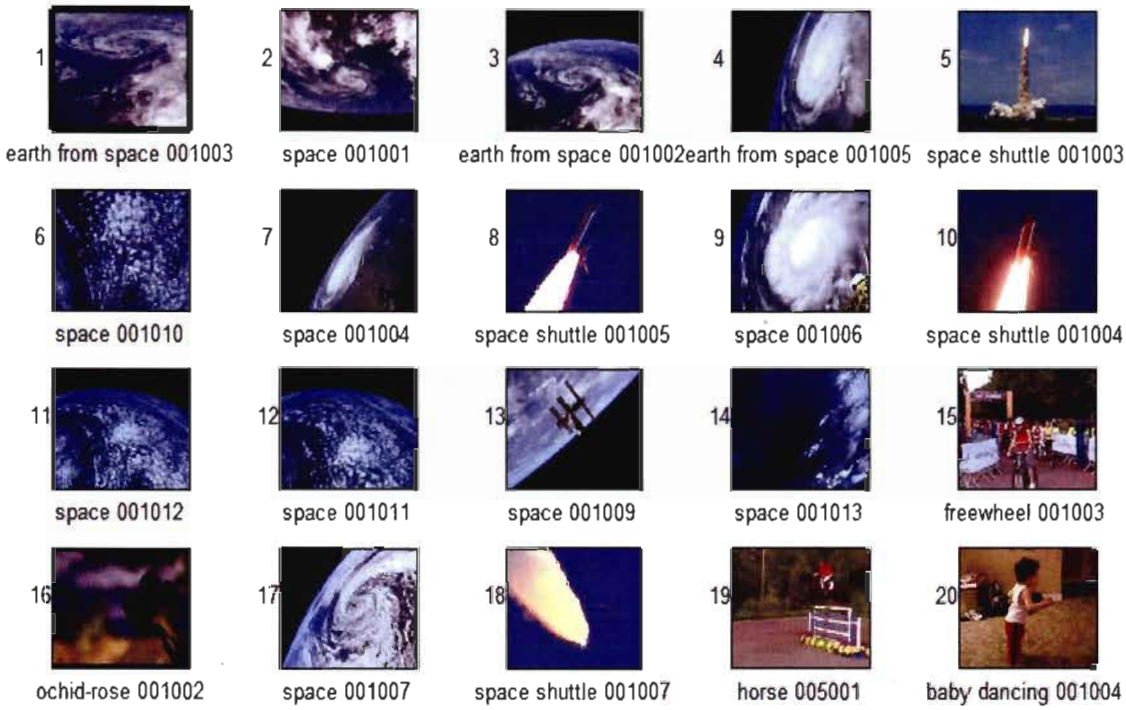


Figure 12 : Affichage des résultats de la requête R1

3) Finalement si la recherche est non satisfaisante, l'utilisateur peut requérir du système de raffiner le choix en sélectionnant une autre vidéo requête.

3.4.1.2 Recherche par attributs physiques

L'interface permet à l'utilisateur de formuler une requête à l'aide de plusieurs options, dont la recherche par métadonnées automatiques (physiques). L'utilisateur formule une requête selon un critère donné et le système lui retourne les résultats voulus.

A titre d'exemple, pour faire une recherche selon la date (Figure 13), l'utilisateur choisit une date précise ou un intervalle de dates (une date de début et une date de fin). Ensuite,

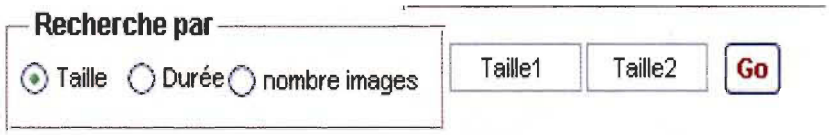
le système effectue une recherche dans la base de données des descripteurs et affiche les résultats.

Figure 13 : Recherche par date.

L'utilisateur peut aussi choisir d'effectuer une recherche à l'aide d'autres attributs physiques tels que : le nombre total d'images de la vidéo, le format de compression, la longueur de séquence en secondes, la résolution des images ou encore le nombre d'images par secondes. Il effectue son choix dans un menu déroulant (Figure 14(a)). L'algorithme effectue un parcours de la base de données des descripteurs et affiche les résultats escomptés.

Relativement à la Figure 14(b), l'utilisateur a le choix d'effectuer une recherche selon un intervalle de valeurs. Lors d'une recherche basée sur la taille de la vidéo (en Mo) ou sur la durée (en s) ou sur le nombre total d'images constituant la vidéo, l'utilisateur propose un intervalle et le système recherche les séquences qui correspondent aux critères soumis.

(a) Selon un tri croissant



(b) Recherche selon un intervalle de données.

Figure 14 : Recherche par métadonnées automatiques

3.4.1.3 Recherche à l'aide des caractéristiques sémantiques

- **La recherche par personnages-clés**

Les personnages-clés sont des personnes ou des objets représentatifs, permettant de décrire une vidéo. Lors de la recherche, l'utilisateur introduit le personnage (ou mot-clé) de recherche (Figure 15); ensuite le système recherche, dans la base de données, toutes les vidéos qui sont annotées par le terme proposé et affiche les 20 premières vidéos respectant les critères soumis.

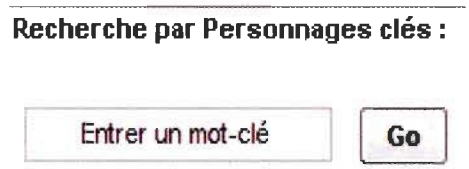


Figure 15 : Recherche par personnages-clés.

- **La recherche par événement**

La fonction de recherche par événement permet à l'utilisateur d'effectuer une recherche en indiquant le nom de l'événement (Figure 16). L'algorithme de recherche parcourt la base des descripteurs et compare le nom entré par l'utilisateur à ceux qui ont été utilisés lors de l'annotation. En cas de correspondance, toutes les vidéos décrites par le même événement sont affichées.

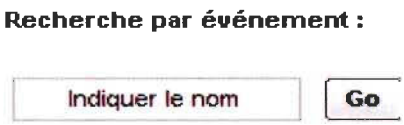


Figure 16 : Recherche par événement

- **Recherche selon la météo**

Il se trouve souvent que lors d'une recherche, l'utilisateur souhaite retrouver des vidéos prises pendant qu'il neigeait, qu'il pleuvait, prises lorsqu'il faisait beau temps ou par temps nuageux. L'outil développé, dispose de ces options de recherche (Figure 17) et permet à l'utilisateur de retrouver assez rapidement l'objet de sa requête. Toutes les vidéos annotées avec ces termes, seront affichées comme résultat à la recherche.

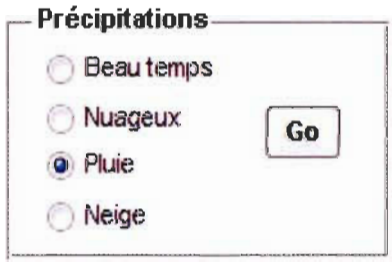


Figure 17 : Recherche selon la météo

- **Recherche par vidéo prise de jour ou de nuit**

L'utilisateur peut soumettre une requête selon que les vidéos sont prises de jour ou de nuit. Ainsi, dans notre base de descripteurs, nous avons des vidéos annotées selon ce critère. Pour formuler ce type de requête, l'utilisateur fait son choix (Figure 18) et lance la recherche.

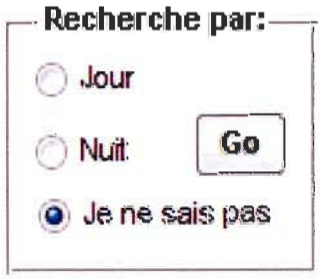


Figure 18 : Recherche selon une vidéo prise de jour ou de nuit.

3.4.1.4 Recherche par combinaison de caractéristiques

La recherche par combinaison de caractéristiques, consiste juxtaposer nos différentes caractéristiques de recherche entre-elles, c'est-à-dire les caractéristiques sémantiques, celles du contenu et les caractéristiques physiques. Cette méthode de recherche permet à l'utilisateur de mieux raffiner ses requêtes et permet d'obtenir de meilleurs résultats, ce dont nous traiterons à la section 4.5.

3.5 Conclusion

Il existe principalement trois méthodes de recherche : textuelle, iconique, par l'exemple [2]. Nous avons vu au chapitre sur l'« État de l'art », que la méthode de recherche basée seulement sur le contenu est insuffisante compte tenu du fossé sémantique qui existe et nous avons vu également que la recherche basée seulement sur le texte présente une certaine subjectivité.

Dans ce chapitre, nous avons présenté un outil facile d'utilisation et qui permet à l'utilisateur d'effectuer une recherche textuelle ou une recherche par l'exemple. Cet outil offre deux interfaces dont l'une permet de faire l'annotation des vidéos et l'autre la recherche de vidéos.

Nous avons également présenté les différentes méthodes que nous avons utilisées pour réaliser l'annotation automatique et semi-automatique de nos vidéos. Nous avons aussi vu les différentes options de recherche dont dispose l'outil.

Notons, qu'à partir de l'interface d'annotation, l'utilisateur peut visualiser les vidéos avant même de les annoter, ce qui permet au système de retourner de très bons résultats lors des requêtes. Un point important est le fait que l'utilisateur peut raffiner sa requête en choisissant comme requête l'une des vidéos résultante.

Dans le prochain chapitre intitulé « Expérimentations », nous pourrons vérifier la performance de notre outil selon les évaluations que nous ferons à partir de tests.

CHAPITRE 4

EXPÉRIMENTATIONS

4.1 Introduction

Dans ce chapitre, nous allons expérimenter notre système d'annotation et de recherche. Nous faisons les expériences sur une base de données constituée d'environ 700 vidéos gratuites que nous avons téléchargées à partir d'Internet, et que nous avons nous-mêmes découpées et subdivisées en 21 familles selon leur contenu.

Afin d'évaluer les performances de n'importe quel système de recherche, on a besoin de répondre à deux questions; à savoir : la vérité terrain (« *ground truth* » en anglais) et les critères de performances (« *performance criteria* ») [1]

A propos de la vérité terrain, elle peut être établie de différentes façons dont : le jugement humain, une base de données pré-classifiée par un ordinateur, etc. En ce qui nous concerne, et compte tenu de notre système, nous avons opté pour une base de données classifiée par un ordinateur.

En ce qui a trait aux critères de performance, ceux qui sont le plus utilisés sont principalement la précision (Pr = *Precision* en anglais) et le rappel (Re = *recall* en anglais). La précision nous donne la proportion des résultats pertinents par rapport au nombre total de résultats affichés. Elle est notée comme suit :

$$\mathrm{Pr} = \frac{\sum(\textit{vidéos pertinentes affichées})}{\sum(\textit{vidéos affichées})}.$$

Le rappel nous fournit la proportion des résultats pertinents affichés par rapport au nombre total de résultats pertinents dans la BD. Cette proportion (rappel) est :

$$\mathrm{Re} = \frac{\sum(\textit{vidéos pertinentes affichées})}{\sum(\textit{vidéos pertinentes de la BD})}.$$

Dans la littérature [1,2, 3], on constate qu'il est difficile de calculer le rappel surtout avec de grandes bases de données. Ceci est dû au fait que le nombre total de résultats pertinents à une requête donnée est difficile à estimer.

Dans notre analyse, nous calculerons alors la précision en fonction du scope (Sc). Par définition, le scope est le nombre de vidéos retournées à l'utilisateur. On note $Sc = \sum(\textit{vidéos affichées})$. Ainsi la fonction $\mathrm{Pr} = \mathrm{f}(\mathrm{Sc})$ nous donnera, pour différentes valeurs du scope, la précision du nombre de vidéos qui seront retournées.

Dans les sections suivantes, nous évaluerons la recherche par attributs physiques, la recherche par attributs sémantiques, la recherche par attributs visuels, et finalement, la recherche par attributs combinés.

4.2 Recherche par attributs physiques

Tel qu'indiqué dans les chapitres précédents, nous entendons par attributs physiques la date de modification, la taille, le nombre d'images, la durée, etc.

Dans ce qui suit, nous présentons les résultats d'une recherche à l'aide de certains de ces attributs physiques, en l'occurrence la taille des fichiers, leurs dates de modification, leurs durées, et le nombre total d'images. Par la suite, nous analyserons les résultats obtenus.

4.2.1 Quelques exemples de recherche avec les attributs physiques

Recherche selon la taille

Une première approche consiste à permettre à l'usager de formuler des requêtes basées sur la taille des fichiers vidéo. Dans ce cas, l'usager a le choix entre soit donner une taille précise ou donner un intervalle de tailles. Le système se charge par la suite de trouver les vidéos qui répondent à ces critères. Notons également que l'usager peut choisir entre l'affichage des résultats selon un ordre croissant ou décroissant des tailles.

À titre d'exemple, la Figure 19 illustre les résultats de la recherche par taille de fichier où la requête est « Afficher les vidéos selon la taille dans un ordre croissant ».

La Figure 20, illustre les résultats suite à la requête « Trouvez toutes les vidéos dont la taille du fichier est comprise entre 1.5 Mo et 2 Mo ».

L'utilisation des intervalles de tailles, permet à l'utilisateur de raffiner sa requête. Par exemple, le fait de commencer avec un intervalle large, puis de le rétrécir, permet de mieux cerner la recherche.

Notons qu'en l'absence de la connaissance du nom de la vidéo, ni d'aucune autre caractéristique de recherche et que seulement une approximation de la taille de la vidéo est connue, alors cet attribut lui permettrait d'effectuer une requête.

Figure 19 : Résultats de recherche selon la taille dans un ordre croissant

Figure 20 : Résultats de recherche de vidéos dont la taille est comprise entre 1.5 et 2 Mo

Recherche selon la date

Comme dans le cas de la taille, l'utilisateur a la possibilité, avec l'attribut « date », d'afficher les résultats de ses recherches selon un ordre croissant ou décroissant, en précisant une date quelconque ou en choisissant un intervalle de dates. Toutes ces méthodes d'affichage ont leur utilité propre selon l'objectif visé par l'utilisateur.

Le choix d'affichage selon un ordre précis, donne à l'utilisateur, la possibilité de retrouver aisément les vidéos les plus anciennes lorsque l'affichage est dans un ordre croissant, ou de retrouver les plus récentes lorsque l'affichage est effectué dans un ordre décroissant.

Le choix de l'affichage selon une date précise est la plus adéquate si l'utilisateur a une connaissance préalable des dates auxquelles les vidéos qu'il cherche ont été prises. C'est le cas de l'exemple à la Figure 21 où la requête est « affichez toutes les vidéos prises le 29/11/2007 ». Les vidéos affichées sont alors pertinentes. Le problème qu'on peut rencontrer à ce niveau, est qu'il est plus difficile de donner une date précise que de donner un intervalle. Le choix d'affichage selon un intervalle de dates aide à résoudre ce problème. Cependant, en utilisant un intervalle de dates, le nombre de résultats affichés est beaucoup plus élevé. La Figure 22 illustre un exemple d'affichage selon un intervalle de dates où la requête est « Affichez toutes les vidéos dont les dates sont comprises entre le 10/01/2007 et le 30/12/2007.» Ensuite pour retrouver les vidéos qu'il souhaite, l'utilisateur peut rétrécir de plus en plus l'intervalle de départ.

Figure 21 : Résultats d'une recherche avec une date précise.

Figure 22 : Résultats d'une recherche avec un intervalle de dates

Recherche selon la durée

Nous avons défini l'attribut « durée de la vidéo » comme étant la période de temps, en secondes, que dure le fichier vidéo. Le système de recherche permet l'affichage des résultats selon la durée dans un ordre croissant, décroissant ou selon un intervalle. L'affichage selon un ordre croissant permet de retrouver les séquences courtes. Celui selon un ordre décroissant permet de retrouver les séquences longues. En choisissant l'affichage selon un intervalle donné, l'utilisateur augmente la précision dans ses

recherches. A titre d'exemple, la Figure 23 illustre la requête «Affichez les vidéos dont la durée est comprise entre 4 et 5s ». Avec un intervalle plus restreint, la précision est encore meilleure, comme l'indique les résultats à la Figure 24 où la requête est : «Afficher toutes les vidéos dont la durée est comprise entre 4 et 4.3s ». Dans les requêtes avec l'attribut « durée », on peut remarquer que l'utilisateur pourrait augmenter ses chances de retrouver une vidéo donnée s'il a au préalable une idée sur la durée anticipée.

Figure 23 : Affichage des vidéos ayant une durée comprise entre 4 et 5s

Figure 24 : Affichage des vidéos ayant une durée comprise entre 4 et 4.3s

Recherche avec les autres attributs physiques

A l'instar des attributs tels que la date, la taille et la durée, les autres attributs physiques comme le nombre total d'images et le nombre d'images par seconde, suivent la même logique de recherche. L'objectivité de ces attributs fait en sorte que le résultat retourné répond aux critères de la requête.

4.2.2 Évaluation de la recherche par attributs physiques

Pour mesurer la précision des résultats, nous avons besoin du paramètre "*vérité terrain*". Nous le posons comme suit : si une image retournée répond aux critères de la requête, on lui attribue la cote 1 sinon 0.

Nous avons effectué un certain nombre de recherches avec les attributs physiques. Pour chaque attribut, nous avons formulé une requête puis mesuré la précision des résultats obtenus. Finalement, nous avons tracé la courbe Pr = f(Sc).

La Figure 25 illustre la précision de l'attribut « Taille » en fonction du scope. L'allure de la courbe de précision indique que toutes les vidéos résultantes sont pertinentes. La conclusion est la même pour l'attribut physique « Date » (Figure 26.

En ce qui a trait à la précision des résultats, puisque la recherche avec les attributs physiques se fait de façon binaire (c'est-à-dire comprise ou pas comprise entre deux tailles, prise à une date donnée ou non, etc.) alors elle est toujours concordante. On peut déduire que pour tous les autres attributs physiques, les courbes de précision seront identiques à celles des attributs « Taille » et « Date ».

Malgré leur objectivité, nous constatons que les attributs physiques donnent des résultats qui peuvent être très différents les uns des autres en termes de contenu. Par exemple, à la Figure 23, la requête consiste à demander seulement d'afficher toutes les vidéos dont la

durée est comprise entre 4 et 5s; on constate que ce sont plusieurs familles de vidéos qui sont affichées. Cependant, ces attributs peuvent être très utiles pour raffiner les résultats de la recherche lorsqu'on les combine avec d'autres caractéristiques, comme nous le verrons à la section 4.5.

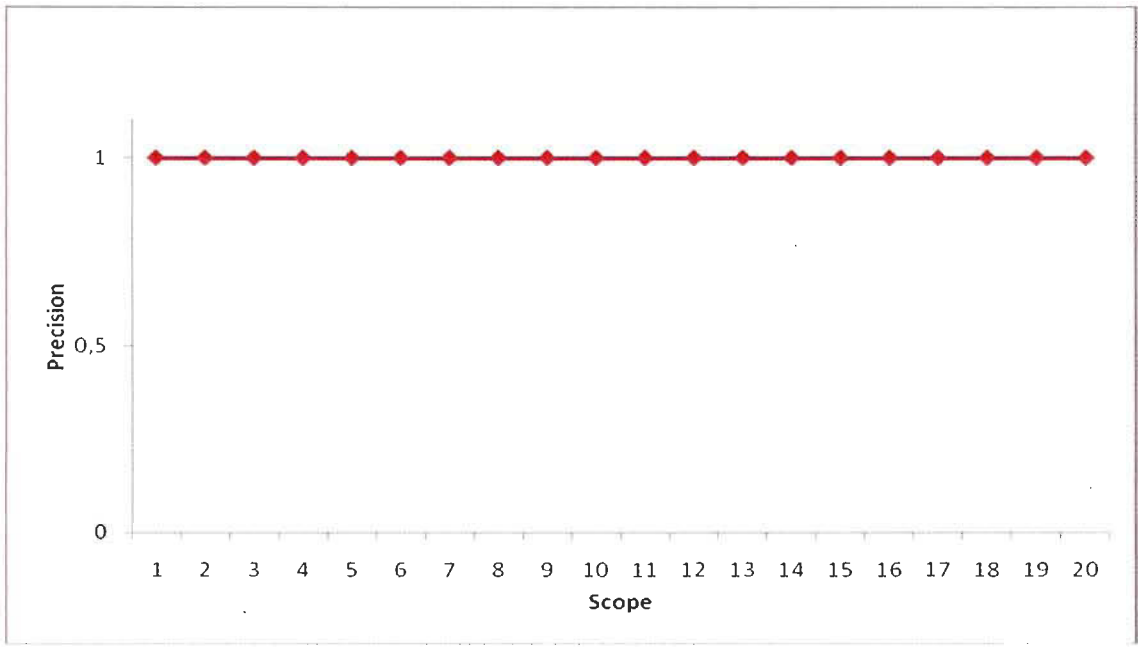


Figure 25 : Tracé de la courbe de précision de l'attribut « Taille »

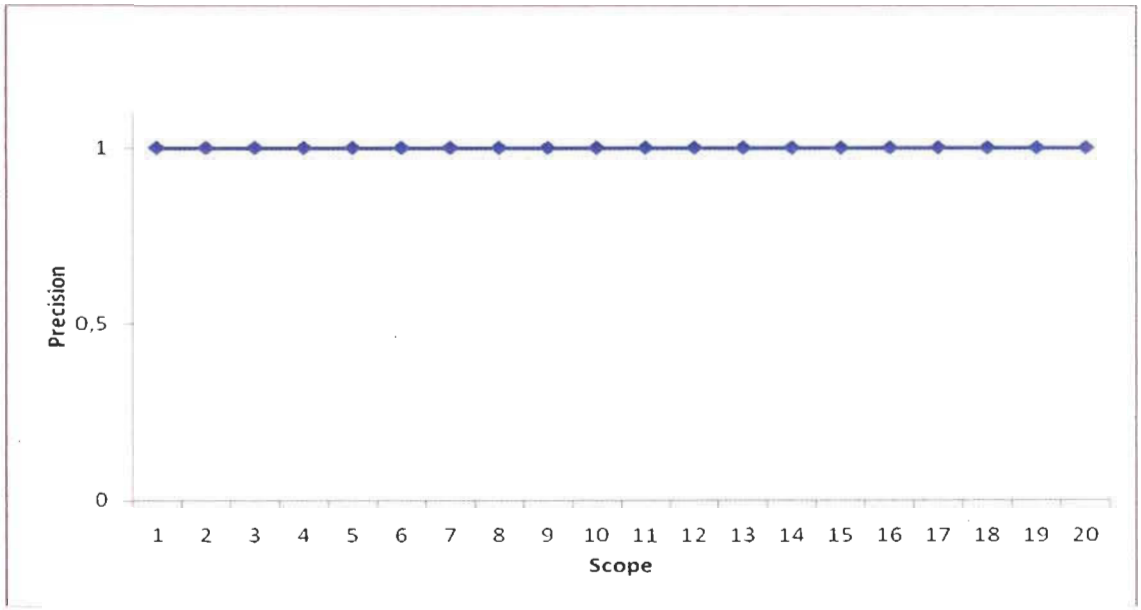


Figure 26 : Tracé de la courbe de précision de l'attribut Date

4.3 Recherche par attributs sémantiques

L'outil d'annotation nous a permis d'attribuer une sémantique à toutes les vidéos de la collection. Nous avons défini les attributs sémantiques comme étant les textes qui accompagnent les vidéos et qui permettent de retrouver ces dernières lors d'une recherche. Dans ce qui suit, nous allons effectuer quelques exemples de recherche ou à l'aide de ces attributs avant de procéder à leur évaluation. Par la suite, nous interpréterons les résultats obtenus.

4.3.1 Quelques exemples de recherche

- **Recherche des vidéos prises de jour ou de nuit**

Ce critère de recherche est utile lorsque l'utilisateur a besoin de rechercher seulement les vidéos qui ont été prises pendant le jour ou celles qui ont été prises pendant la nuit. Ce critère à lui seul peut être utilisé pour formuler une requête comme illustre les Figures 27 et 28 où on cherche respectivement toutes les vidéos prises de jour et celles prises de nuit. On constate à la Figure 27 que la majorité des vidéos sont prises de jour mis à part quelques unes dont les images-clés apparaissent sombres.

Figure 27 : Exemple de résultats de vidéos prises de jour

Figure 28 : Exemple de résultats de vidéos prises de nuit

- **Recherche par personnages-clés**

Tel que défini antérieurement, nous désignons par « personnages-clés », tout ce qui est représentatif dans le contenu d'une séquence vidéo. Cela peut être des personnes ou des objets particuliers qui peuvent représenter ladite séquence. On peut, tout simplement, utiliser un mot ou une expression qui donne l'idée générale du contenu de la séquence. C'est le cas de l'exemple à la Figure 29 (a) et (b) où les requêtes sont respectivement : « affichez toutes les vidéos annotées avec le mot "démonstration " » et « affichez toutes les vidéos indexées par le personnage "acrobate" ».

(a) résultats de la recherche avec le mot clé « démonstration ».

(b) résultats de la recherche avec le mot clé « acrobate »

Figure 29 : Résultats de la recherche par personnages-clés

Dans le premier exemple (Figure 29 (a)), les vidéos affichées comme résultats, répondent toutes à la requête effectuée. Dans le deuxième exemple (Figure 29 (b)), les vidéos affichées répondent aussi au critère de la requête, mais on remarque qu'elles ne sont pas de la même famille. Notamment les vidéos encerclées, n'étaient pas souhaitées dans cette requête, mais elles sont affichées car elles ont aussi été annotées avec le terme « acrobate ».

- **Recherche par événement**

La recherche par événement s'effectue à l'aide de requêtes basées sur des mots-clés comme dans le cas de la recherche par personnages-clés. A ce niveau, on peut effectuer une recherche selon un nom d'événement ou selon un lieu où s'est tenu un événement donné. A titre d'exemple de résultats d'affichage, la Figure 30 illustre la requête suivante : « affichez toutes les vidéos prises lors d'un camping » ou encore l'exemple à la Figure 31 où l'on recherche toutes les vidéos prises au « Maroc ».

Figure 30 : Recherche de vidéos prises dans un camping.

Figure 31 : Recherche de vidéos prises au Maroc.

- **Recherche selon la météo**

Rappelons que la recherche selon la météo est basée sur les conditions suivantes : pluie, neige, beau-temps, temps nuageux, chaleur et froid. En plus d'être des critères de recherche, ces caractéristiques, à l'instar des caractéristiques physiques, permettent de mieux raffiner les requêtes lorsqu'elles sont combinées avec d'autres critères. La Figure 32 illustre un exemple de résultats de recherche avec l'attribut "pluie".

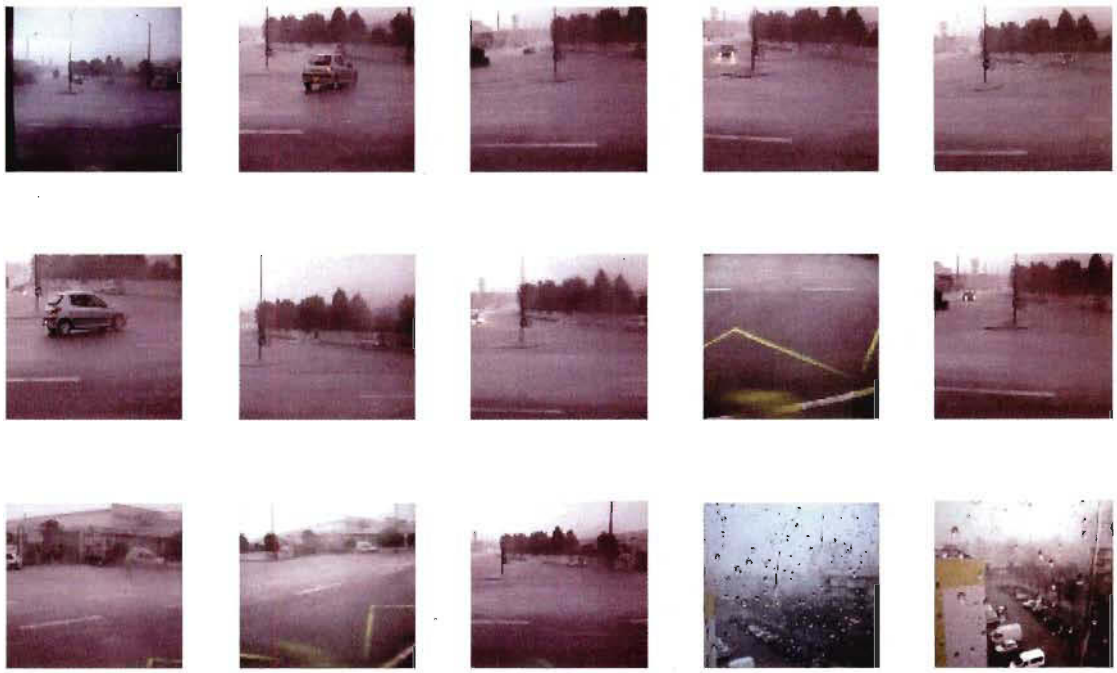


Figure 32 : Exemple de résultats de recherche avec l'attribut météo

4.3.2 Évaluation de la recherche par attributs sémantiques

Une recherche effectuée avec les caractéristiques sémantiques est plus intuitive pour l'utilisateur qu'une recherche par le contenu. L'utilisateur a la possibilité d'utiliser ses propres mots pour effectuer ses requêtes. Lorsque l'utilisateur effectue sa requête et que le terme existe dans la base, les vidéos correspondantes sont affichées. Dans le cas où le terme n'est pas connu, le système retourne un message de type : "ce nom est inexistant dans la BD" ou "veuillez reformuler votre requête".

Dans cette section, nous allons tenter de mesurer la performance de la recherche avec chacun des attributs sémantiques cités auparavant, à savoir : la météo, jour ou nuit, et personnages-clés. Nous avons besoin de "*vérité terrain*" qui permet de juger la pertinence des résultats retournés. Nous avons adopté la stratégie suivante : si une vidéo résultante provient de la même famille que la requête, on juge qu'elle est pertinente, sinon elle ne l'est pas. Pour l'ensemble des familles de vidéos dont nous disposons, nous posons $C = \{C_1, C_2, \ldots, C_n\}$, où C est l'ensemble des classes et les C_i représentent différentes familles de classes.

Nous avons effectué plusieurs recherches, chacune avec un nouveau mot-clé comme requête. Si une vidéo résultante fait partie de la même classe que la requête, elle reçoit automatiquement comme cote 1; sinon c'est la cote est 0. On peut traduire cela par : $\Pr(v) = \begin{cases} 1, si\ v \in C_q \\ 0, si\ v \notin C_q \end{cases}$, où v représente une vidéo résultante quelconque, q la requête et C_q la classe à laquelle appartient q.

Ensuite, nous calculons la moyenne des résultats pour chaque valeur du scope et finalement, nous traçons la courbe de précision Pr=f(Sc).

Évaluation de l'attribut Jour ou nuit

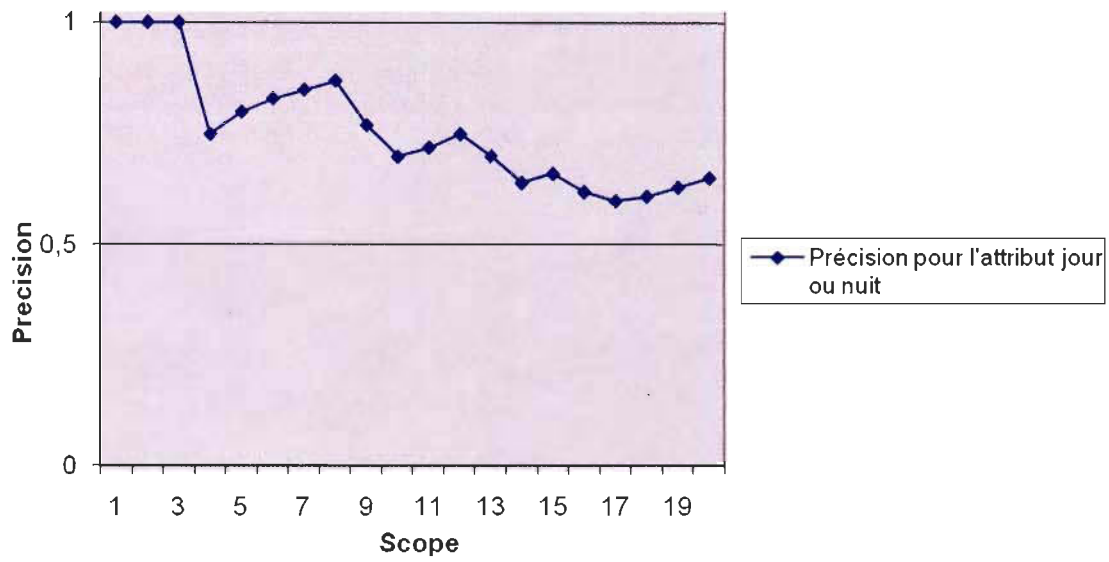


Figure 33 : Courbe de précision de l'attribut jour ou nuit

Discussion des résultats

On tire de la Figure 33 que la précision commence à son maximum (1) et ne décroit pas rapidement. Comme on peut le constater à la Figure 33, la valeur minimale de la précision est de 0,65 même après 20 résultats. On peut conclure que la précision est relativement bonne.

D'autre part, nous constatons que la précision diminue quand on augmente le scope. Ceci est dû à l'imperfection de l'annotation semi-automatique. En effet, l'attribut jour/nuit a été rajouté aux vidéos par le module de propagation semi-automatique de mots-clés. Ainsi, une vidéo claire peut être considérée par le système comme une vidéo prise de jour, et une vidéo sombre peut être considérée comme étant prise de nuit. Or, une vidéo claire peut avoir été prise la nuit sous de bonnes conditions d'illumination, de

même qu'une vidéo sombre peut avoir été prise le jour sous de mauvaises conditions d'illumination.

Évaluation de l'attribut personnages-clés

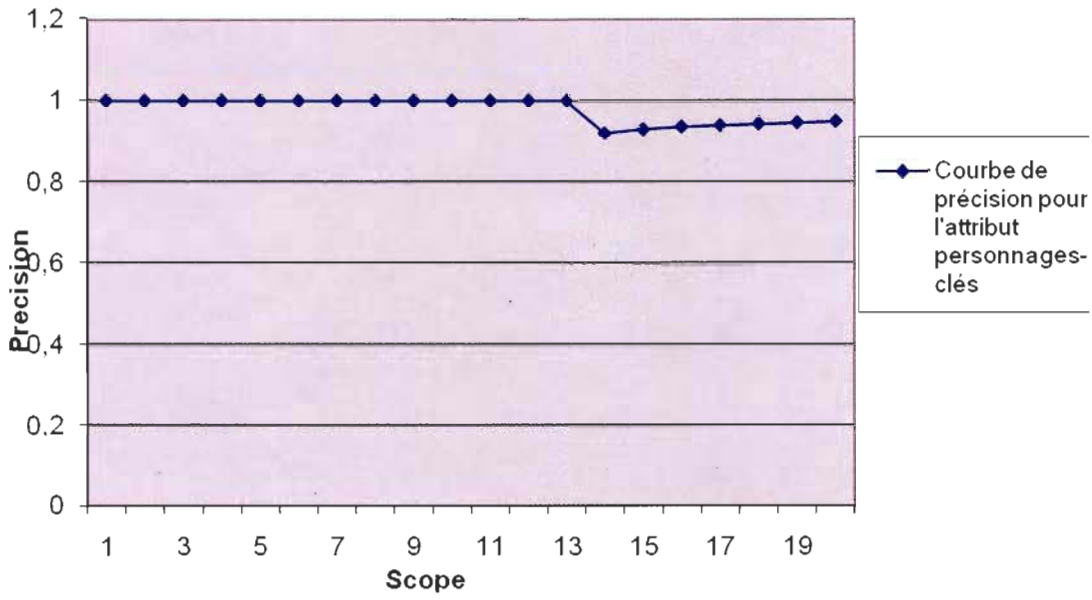


Figure 34 : Courbe de précision de l'attribut personnages-clés.

Discussion des résultats

La courbe de la précision moyenne de l'attribut « personnage-clé » est globalement bonne. Elle commence avec un maximum de 1 et ne dégénère pas rapidement tel qu'indiqué à la Figure 34. En effet la valeur minimale de la précision est de 0,78. Toutefois, si on augmente le scope il y aura du bruit, compte tenu de l'imperfection de l'annotation semi-automatique, tel qu'indiqué entièrement.

Évaluation de l'attribut « Événement »

Pour évaluer cet attribut, nous avons effectué des tests sur quelques familles de vidéos. Pour chaque famille, nous avons calculé la valeur moyenne de la précision. Par la suite, nous avons représenté la précision de chaque famille tel qu'illustré à la Figure 35.

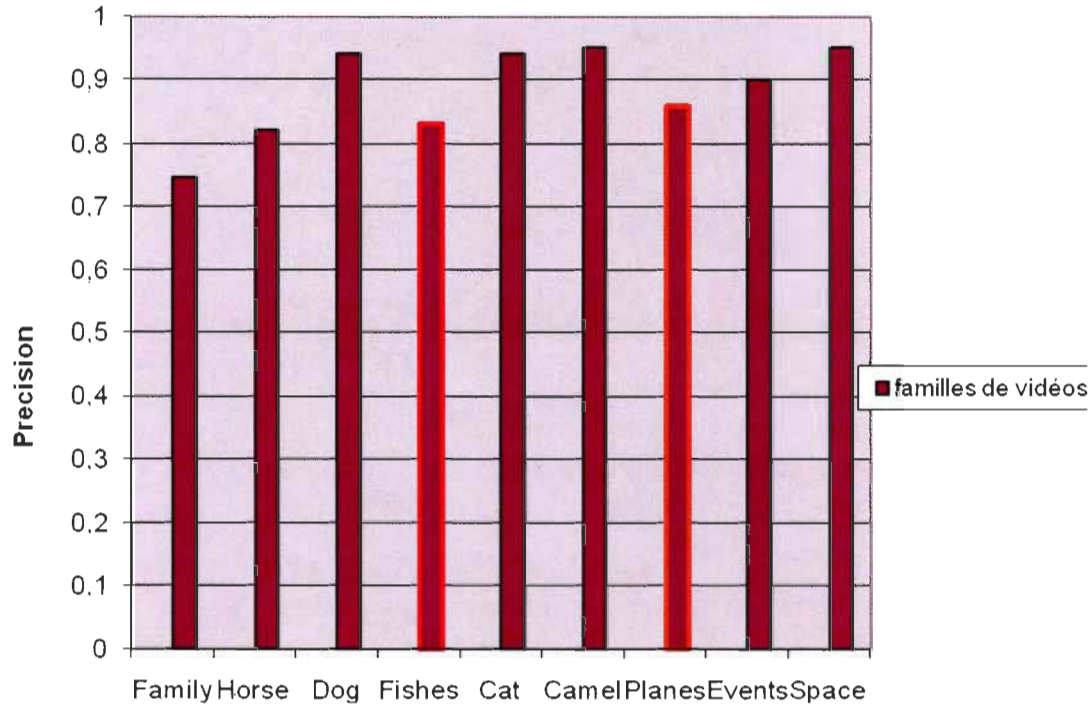


Figure 35 : Évaluation de l'attribut « événement » pour quelques familles de vidéos.

Discussion des résultats

Globalement, l'histogramme de précision indique des résultats assez satisfaisants pour les familles testées. Les valeurs de la précision moyenne sont comprises entre 0,85 et 0,66 par rapport au maximum qui est de 1. Dans les résultats des requêtes effectuées, on rencontre souvent des vidéos qui se ressemblent visuellement mais qui n'appartiennent pas nécessairement aux mêmes classes prédéfinies. Cela signifie que ces vidéos peuvent

être prises lors d'événements différents, mais elles ont été annotées avec les mêmes noms compte tenu de leur similarité visuelle comme le cas des familles de vidéos « Fishes » et « Planes » à la Figure 35. Or, si la vidéo affichée ne fait pas partie des classes souhaitées, elle reçoit automatiquement une cote égale à 0. Nous reparlerons de ce problème à la section 4.4.

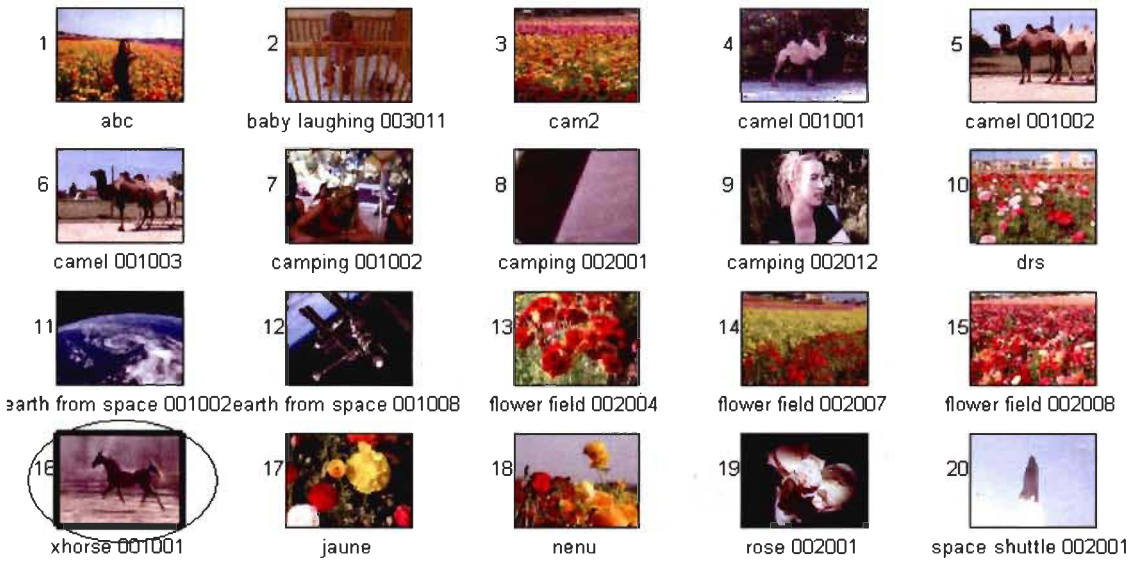
4.4 Recherche par le contenu

4.4.1 Exemple de recherche par le contenu

Afin que l'utilisateur soit capable d'effectuer des recherches par le contenu, le moteur de recherche commence par lui proposer un ensemble de vidéos qui représentent un échantillon du contenu de la base de données. La Figure 36(a) illustre un exemple de vidéos affichées initialement à l'utilisateur.

Ensuite, l'utilisateur formule sa requête en sélectionnant l'une de ces vidéos. Dans l'exemple de la Figure 36(a), l'utilisateur a choisi la vidéo « xhorse001001 » (entourée d'un cercle) comme requête.

Par la suite, le moteur recherche, dans la base de données, toutes les vidéos qui répondent à cette requête. La Figure 36(b) illustre les résultats de la recherche de cette requête.



(a) Fichiers initiaux et choix de la requête

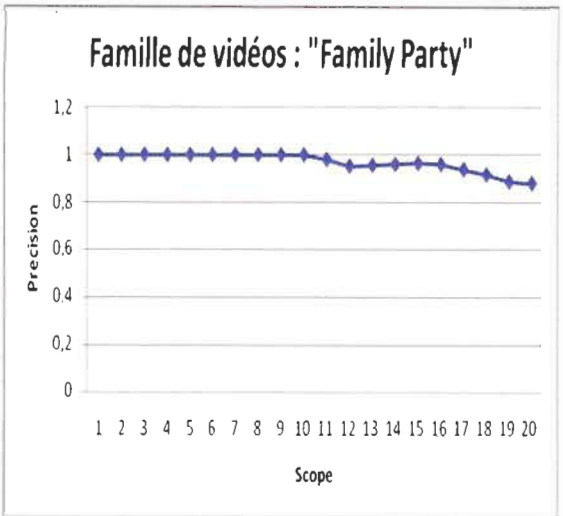
(b) résultats de la recherche

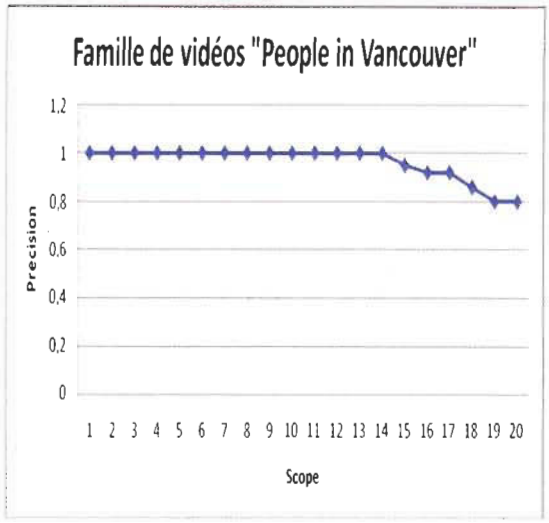
Figure 36 : Exemple de recherche par le contenu

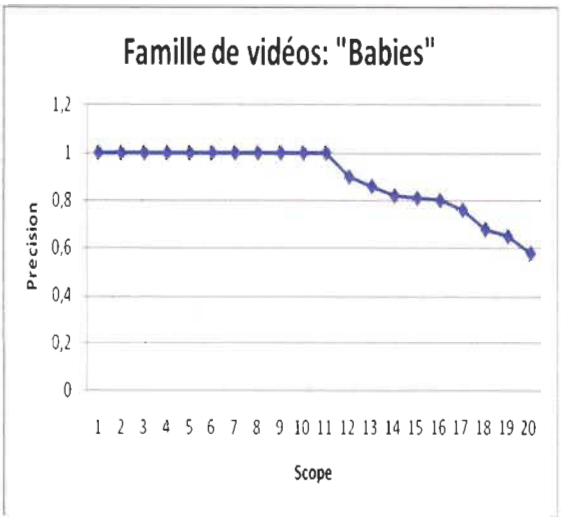
4.4.2 Évaluation de la recherche par le contenu.

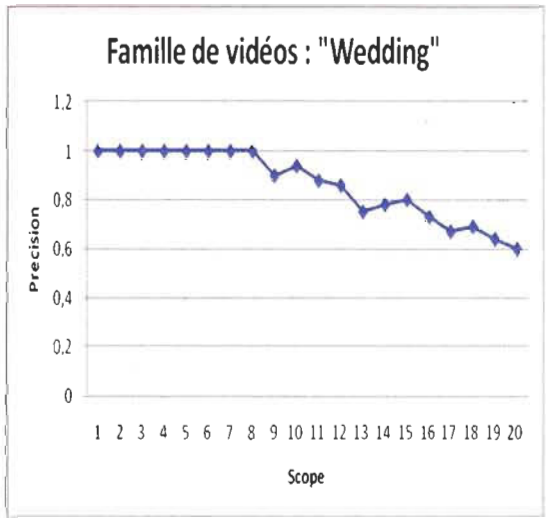
Afin d'évaluer la recherche par le contenu, nous avons effectué des requêtes sur quelques familles de vidéos. Pour chaque famille, nous avons choisi plusieurs vidéos comme requêtes, puis nous avons calculé la moyenne des précisions. Quant à la "vérité terrain", elle a été obtenue comme suit : une vidéo résultante qui provient de la même classe que la requête obtient le score 1 alors qu'une vidéo provenant d'une classe différente obtient le score 0.

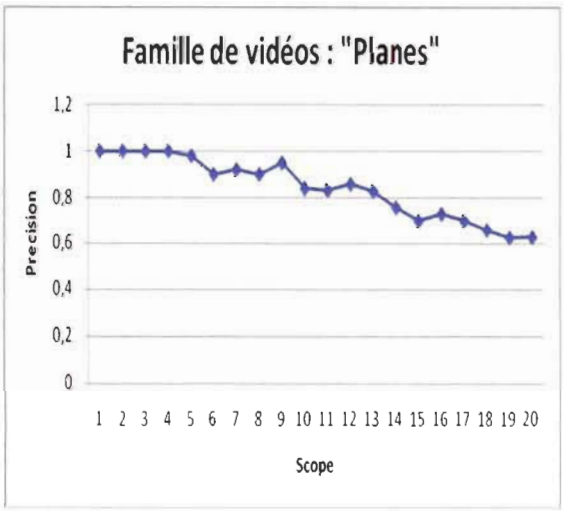
Les courbes suivantes représentent les valeurs moyennes de la précision en fonction du scope pour quelques catégories de vidéos analysées.

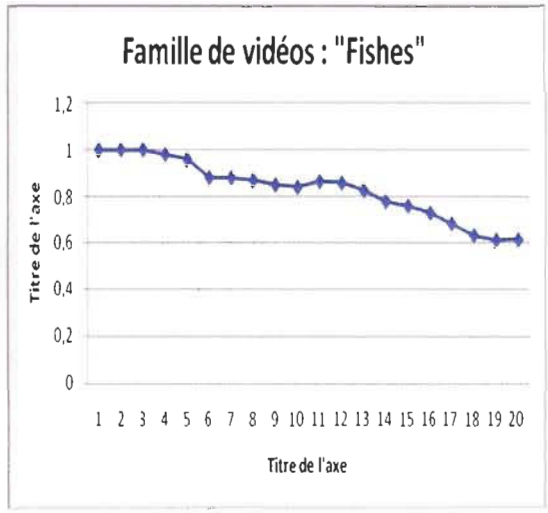

Famille de vidéos : "Family Party"
Precision
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scope


Famille de vidéos "People in Vancouver"
Precision
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scope


Famille de vidéos: "Babies"
Precision
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scope


Famille de vidéos : "Wedding"
Precision
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scope


Famille de vidéos : "Planes"
Precision
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scope


Famille de vidéos : "Fishes"
Titre de l'axe
1,2
1
0,8
0,6
0,4
0,2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Titre de l'axe

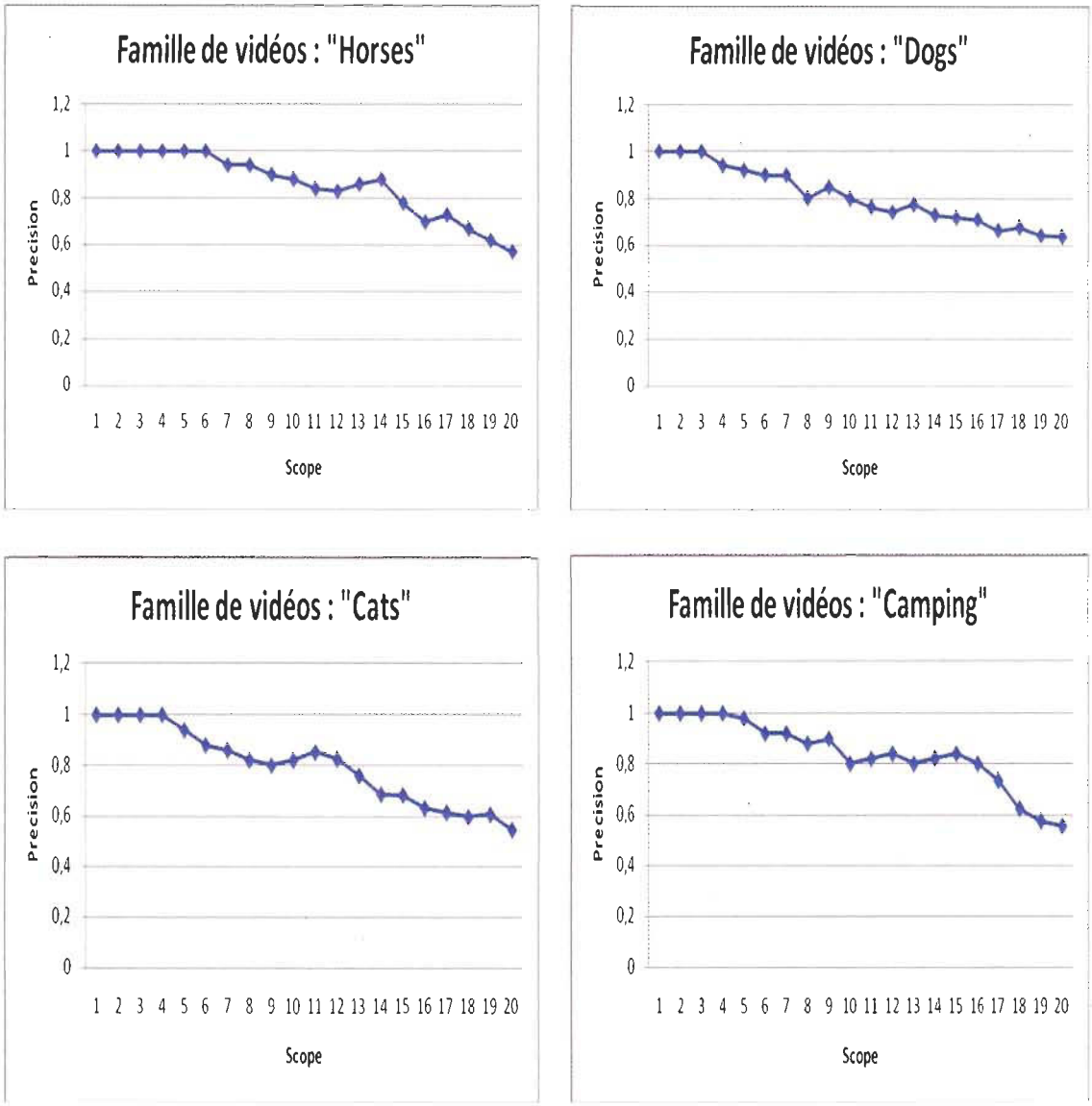


Figure 37 : Courbes de précision pour la recherche par le contenu de quelques familles de vidéos

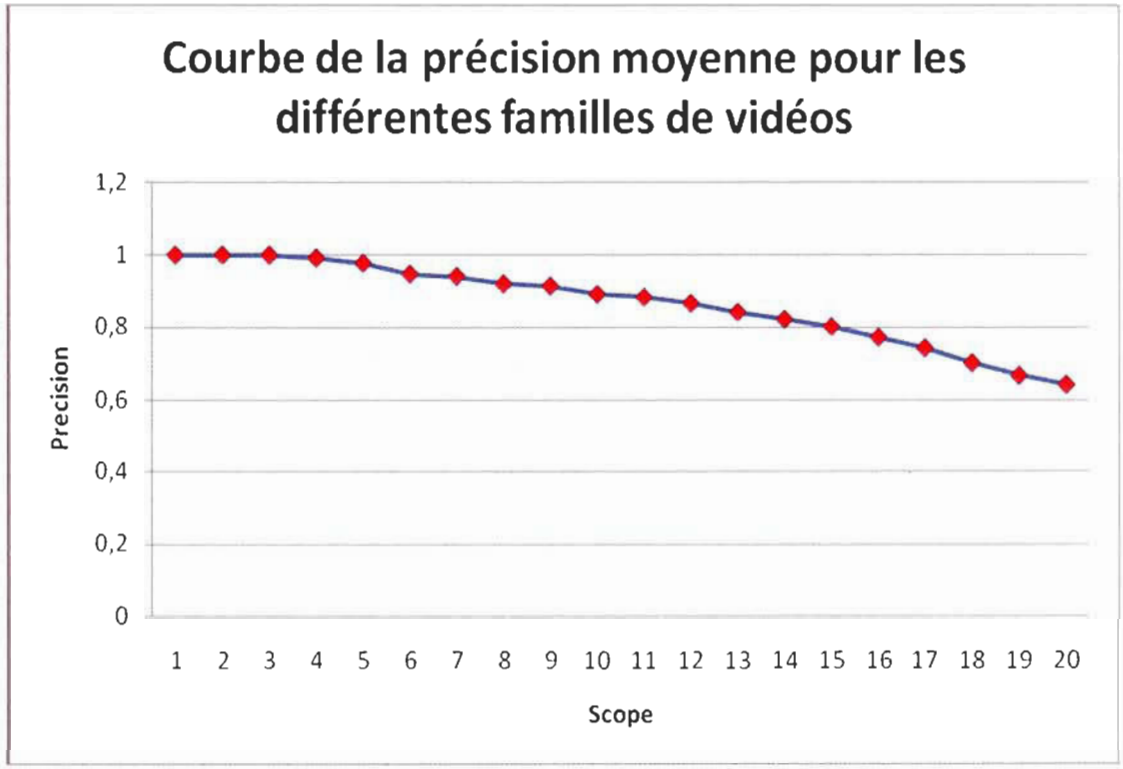


Figure 38 : Courbe de la précision moyenne pour les familles de vidéos analysées

Dsicussion des résultats

Nous remarquons pour l'ensemble de ces résultats que la précision est bonne. Elle commence avec un maximum de 1 et ne décroit pas rapidement comme l'illustre la Figure 38 qui indique la précision moyenne en fonction du scope et ce pour toutes les familles de vidéos analysées. Notons cependant que cette précision depend surtout de la famille considérée. Comme l'indiquent les courbes à la Figure 37, on remarque, par exemple, que pour certaines familles de vidéos, comme "Family Party" ou "People in Vancouver", la précision est très bonne avec un maximum de 1 et un minimum d'environ 0,80. Cette precision est satisfaisante car les scènes des vidéos de chaque famille considérée, sont semblables les unes aux autres au sein d'une même famille, et différentes des scènes des autres familles. D'autre part, pour certaines familles de vidéos, comme la famille "Planes" et "Fishes", la précision diminue avec le scope. Cela

est dû au fait que ces deux familles de vidéos ont un contenu visuellement similaire, soit une grande zone bleue (l'océan pour "Fishes" et le ciel pour "Planes") comme l'illustre la Figure 39. Cette figure présente des vidéos representant des avions dans le ciel ou des poissons dans l'océan. Ainsi, lors d'une requête effectuée avec une vidéo appartenant à l'une ou l'autre de ces deux familles, on retrouve dans les resultats affichés, des vidéos appartenant aux deux familles.

On peut aussi remarquer, que dans le cas des familles de vidéos comme "Cats" et "Camping", la précision commence avec un maximum de 1 et décroit jusqu'à une valeur minimale proche de 0,55. Cela est dû au fait que le décor dans le contenu de ces familles change à des moments donnés ; ce qui explique la baisse de la précision.

Figure 39 : Exemple de familles de vidéos différentes avec un contenu visuellement similaire

4.5 Recherche par combinaison de caractéristiques

Dans cette section, nous allons effectuer des recherches en combinant divers critères de recherche, notamment les attributs sémantiques et visuels. Les critères physiques quant à eux nous permettront de raffiner les requêtes grâce à leur objectivité. Nous évaluerons par la suite cette méthode de recherche à travers plusieurs requêtes et en traçant la courbe de la moyenne des précisions. Pour finir, nous discuterons les résultats que nous aurons obtenus.

4.5.1 Exemples de recherche avec la combinaison des critères

Dans les exemples des figures suivantes, nous allons effectuer des requêtes qui montrent :

- L'affichage des vidéos selon l'attribut sémantique seulement (Figure 40) où la requête est : « Affichez toutes les vidéos annotées avec le mot-clé "danse" »;
- L'affichage des vidéos selon leur contenu visuel (Figure 41), où la première vidéo affichée (entourée d'un cercle) a été utilisée comme vidéo requête;
- L'affichage des vidéos dont la requête combine l'attribut visuel et sémantique (Figure 42) où la requête est : « Affichez toutes les vidéos ressemblant à celle entourée d'un cercle (Figure 42), et annotées avec le mot-clé "danse" »;
- Le raffinement de l'affichage en utilisant les critères physiques. Figure 43, montre les résultats de la requête « Affichez les vidéos similaires à la vidéo encerclée à la Figure 41, qui sont annotées avec le mot-clé "danse", et dont la date est comprise entre le 30/10/2007 et le 30/12/2007 ».

Figure 40 : Recherche avec le mot clé « danse »

Figure 41 : Recherche avec l'attribut visuel seul

Figure 42 : Combinaison de l'attribut visuel + mot-clé "danse"

Figure 43 : Requête combinant les attributs visuels, sémantiques et physiques

4.5.2 Évaluation de la recherche par combinaison d'attributs

Afin d'évaluer cette méthode de recherche, nous avons procédé comme dans les cas précédents, soit : si une vidéo résultante fait partie de la classe souhaitée, elle reçoit comme précision une cote de 1 sinon la cote 0. Nous avons tracé la courbe de précision et nous avons interprété les résultats obtenus dans la section « Discussion des résultats ».

Étant donné que les attributs physiques nous permettent de raffiner nos requêtes, nous avons combiné seulement les attributs visuels et sémantiques comme dans l'exemple de la Figure 42. Nous avons formulé des requêtes combinant les critères tels que : « personnages-clés », «événement » ou « météo » d'une part et l'attribut visuel d'autre part. Les requêtes que nous avons effectuées sont donc du type : « Affichez toutes les vidéos annotées avec le mot-clé "bébé" et ayant un contenu visuel semblable à celui de la vidéo choisie ». Nous avons tracé ensuite la courbe de la précision moyenne de la recherche par combinaison des caractéristiques, que nous comparons à la recherche par le contenu à lui seul et à celle par le texte seul (Figure 44).

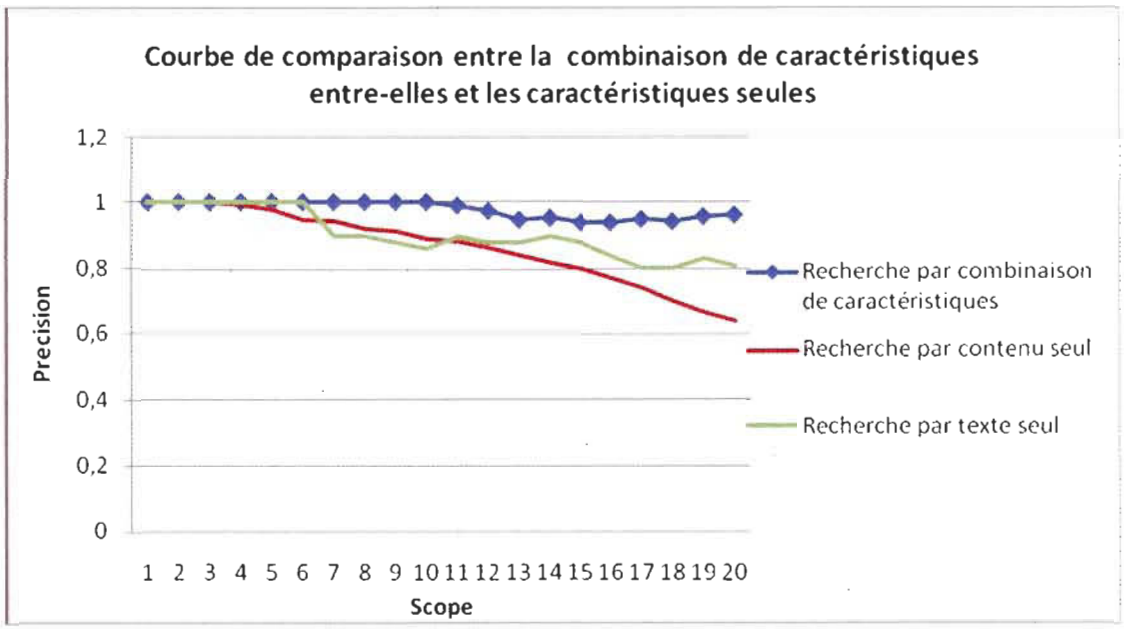


Figure 44 : Courbe de précision de la recherche par combinaison d'attributs vs les attributs seuls

4.5.3 Discussion des résultats de la recherche par combinaison de critères

La courbe de précision de la recherche par combinaison d'attributs (Figure 44) montre que globalement, la combinaison donne de meilleurs résultats. En effet, on remarque que la courbe est à son maximum pour un certain nombre de valeurs, avant de baisser très légèrement. Comme l'illustre la Figure 44, nous comparons ce critère de recherche à celui de la recherche par le texte seul et la recherche par le contenu seul.

Nous remarquons, que le texte à lui seul n'est pas suffisant pour obtenir de très bons résultats de recherche, tout comme la recherche par le contenu à lui seul ne l'est pas. Pour combler ce vide qu'on appelle fossé sémantique, on constate que la meilleure solution, est de combiner la recherche par le contenu et celle par le texte, comme l'indique la courbe de la recherche par combinaison de caractéristiques de la figure précédente.

La combinaison de caractéristiques visuelles, sémantiques et physiques est indispensable pour obtenir une bonne performance de la méthode.

4.5 Conclusion

Dans ce dernier chapitre, nous venons de voir quelques exemples de formulation d'une requête selon les attributs physiques, sémantiques et visuels. Nous avons vu également comment se fait l'affichage des vidéos résultantes et la méthode que nous avons utilisée pour effectuer l'évaluation de notre outil de recherche. Pour chacun des critères utilisés, nous avons donné une interprétation des résultats obtenus. Nous avons constaté que pour chacun des attributs de recherche, les résultats répondent au critère de la requête. Cependant, pour le cas de la recherche par attributs physiques seuls, même si les résultats sont objectifs, ils demeurent parfois loin de ce à quoi l'utilisateur peut s'attendre. Dans la section 4.5, nous montrons l'importance de combiner les différents attributs de recherche entre eux; notamment les attributs visuels, sémantiques et physiques. Nous remarquons, que lorsqu'il s'agit de combiner tous ces attributs entre eux, la moyenne de la précision est très satisfaisante et avoisine le maximum de l'échelle du critère.

CONCLUSION GÉNÉRALE

Dans ce mémoire, nous avons présenté essentiellement le problème d'annotation et d'indexation des documents multimédias et son application à la recherche de vidéos.

Nous avons vu que le besoin de développer de nouveaux systèmes d'annotation et d'indexation est grandissant de nos jours. Cela est dû à la nécessité de pouvoir bien gérer la quantité abondante de données multimédias qu'on peut facilement stocker sur un ordinateur.

Dans le premier chapitre, nous avons vu quelques aspects généraux de l'indexation de la vidéo. Nous avons présenté principalement de l'utilité de l'indexation de la vidéo et comment procéder pour l'effectuer. Pour cela, nous avons présenté différentes caractéristiques de la vidéo telles que : la couleur, la texture, la forme, le mouvement, le son; et comment procéder pour les analyser. Nous avons terminé ce premier chapitre en indiquant le besoin de faire de l'annotation afin de procéder à une bonne indexation.

Dans le deuxième chapitre, nous avons vu les différentes méthodes d'annotation existantes à savoir l'annotation manuelle, automatique, et semi-automatique. L'annotation manuelle, qui est faite essentiellement par un humain, consomme temps et ressources. L'annotation automatique, faite essentiellement par la machine, est objective et bien appropriée pour des petites bases de données contenant de vidéos simples. Cependant, elle devient inapplicable dès qu'il s'agit de vidéos un peu complexes. Partant de la constatation que l'annotation manuelle et l'annotation automatique présentent chacune des insuffisances, nous avons concentré nos efforts sur le développement d'un nouveau mécanisme d'annotation semi-automatique qui tente de tirer profit des avantages des deux autres méthodes.

Dans le troisième chapitre, nous avons présenté le fonctionnement de notre système qui se compose de deux modules : un premier module pour l'annotation des vidéos et un

second pour la recherche. Le premier module permet d'attribuer du texte aux vidéos de façon semi-automatique. Le second module, quant à lui, permet d'effectuer des recherches en partant d'une requête formulée par l'utilisateur. La requête peut être basée sur les attributs physiques (date de modification, taille de la vidéo, le nombre d'images total constituant la vidéo, la durée de la séquence, etc.); les attributs sémantiques (les personnages et objets-clés de la vidéo, l'événement durant lequel la vidéo a été prise, les informations sur la météo, etc.); les attributs visuels (le contenu visuel de la vidéo tel que la couleur, la texture, la forme, etc.). Pour une recherche plus précise, nous avons permis utilisé la combinaison des différents types d'attributs.

L'évaluation de notre système de recherche, présenté au quatrième chapitre, indique que la combinaison de caractéristiques visuelles et sémantiques, aide à combler le fossé sémantique auquel font face les systèmes de recherche par le contenu seul et ceux par le texte seul.

Pour conclure, les vidéos peuvent être caractérisées par différents types d'attributs : visuels, textuels et physiques. Notre hypothèse est que pour obtenir de bons résultats durant la recherche, il faut exploiter tous les attributs au sein du moteur de recherche. Conséquemment, nous avons opté pour le développement d'un système qui combine les différents types d'attributs. Les attributs physiques et les attributs visuels ont été extraits de façon automatique à partir des séquences vidéo. Pour ce qui est des attributs textuels, nous avons développé un mécanisme d'annotation semi-automatique dont le principe est simple mais efficace : l'utilisateur commence par annoter manuellement un échantillon de vidéos, ensuite le module propage les mots-clés à toutes les vidéos qui sont visuellement similaires. Cette façon de procéder est plus rapide que l'annotation manuelle et plus précise que l'annotation automatique. Les expériences que nous avons conduites démontrent que notre mécanisme d'annotation est efficace et que notre outil de recherche est précis. Elles confirment également nos attentes quant à la nécessité de combiner les différents types d'attributs afin d'obtenir de bons résultats.

Le domaine de l'indexation et l'annotation des vidéos étant très vaste, il reste encore beaucoup d'efforts à fournir afin de développer des outils efficaces. Parmi les pistes prometteuses qu'il faudrait explorer, il y a l'utilisation du son dans l'indexation de la vidéo. En effet, le son est une composante fondamentale de la vidéo et si on arrive à bien l'exploiter, cela devrait contribuer grandement à l'amélioration de la précision des recherches. Une autre piste qui semble prometteuse est le développement de techniques efficaces pour l'annotation automatique. La reconnaissance d'objets et de scènes peut être utilisée pour cette fin.

BIBLIOGRAPHIQUE

[1] J. R. SMITH and S. CHANG, "Searching for Images and Videos on the World Wide Web", Columbia University, Center for Image Technology for New Media, New-York, August 1996.

[2] B. FURHT, S. W. SMOLIAR, and H. ZHANG, "Video Processing in Multimedia Systems", Kluwer International Series In Engineering And Computer Science, pp. 377, October 1995.

[3] M. L. KHERFI AND D. ZIOU, "Image Retrieval from the World Wide Web: Issues, Techniques, and Systems", dans le journal ACM Computing Surveys, 36(1):35-67, Mars 2004.

[4] Y. RUI, T. HUANG, and S. CHANG, "Image Retrieval: Current Techniques, Promising Directions and Open Issues", Journal of Visual Communication and Image Representation, vol. 10, pp. 39–62, Avril 1999.

[5] Y. RUI, T. S. HUANG and S. MEHROTRA, Content-based Image Retrieval with Relevance Feedback in MARS, Beckman Institute and Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, pp. 815-818, 1997.

[6] P. WU, "A Semi-automatic Approach to Detect Highlights for Home Video Annotation", IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Quebec, Canada, vol. 5, pp. 957-960, Mai 2004.

[7] M. L. KHERFI, "Review of Human-Computer Interaction Issues in Image Retrieval", Chapitre du livre : Advances in Human-Computer Interaction, (Shane Pinder Eds.), Intech, octobre 2008. ISBN : 978-953-7619-15-2.

[8] A. SALWAY, "Video Annotation: the Role of Specialist Text", Thesis, Department of Computing, School of Electronic Engineering, Information Technology and Mathematics, University of Surrey, Guildford, United Kingdom, pp. 188, December 1999.

[9] F. SOUVANNAVONG, "Indexation et Recherche de Plans Vidéo par le Contenu Sémantique", Thèse sur le Traitement du Signal et des Images, École Nationale Supérieure des Télécommunications, Paris, pp. 141, Juin 2005.

[10] M. J. CAREY, E .S. PARRIS, H. LLOYD-THOMAS, "A Comparison of Features for Speech, Music and Discrimination", IEEE, Monmouthshire, U.K, pp. 149-151, 1999.

[11] J. PINQUIER, C. SÉNAC, R. ANDRE-OBRECHT, " Indexation de la Bande Sonore : Recherche des Composantes Parole et Musique", Institut de Recherche en Informatique de Toulouse - Université Paul Sabatier, Toulouse, France, Septembre 2001.

[12] S. LEFEVRE, J. HOLLER, N. VINCENT, "Segmentation Temporelle de Séquences d'Images en Couleurs", Laboratoire d'Informatique, Université de Tours, France.

[13] A. HANJALIC, R. L. LAGENDIJK, and J. BIEMOND, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems", IEEE Transactions on Circuits and Systems for Video Technology, pp. 580–588, Juin 1999.

[14] T. URRUTY, F. BELKOUCH, C. DJERABA, "Kpyr, une Structure Efficace d'Indexation de Documents Vidéo", Laboratoire d'Informatique Fondamentale de Lille, France.

[15] R, BRUNELLI, O. MICH, and C. M. MODENA, "A Survey on the Automatic Indexing of Video Data", Journal of Visual Communication and Image Representation, ITC-irst, I-38050 Povo, Trento, Italy, pp. 78-112, Juin 1999.

[16] Mediadico, [en ligne], http://www.mediadico.com/dictionnaire/definition/couleur/1 consulté le 12 mars 2009.

[17] A. HANJALIC, "Content–Based Analysis of Digital Video", Kluwer Academic Publisher, Massachuesetts, USA, pp. 187, 2004.

[18] M. SCUTURICI, "Contribution aux Techniques Orientées Objet de Gestion des Séquences Vidéo pour les Serveurs Web", Thèse en Informatique, Institut National des Sciences Appliquées de Lyon, pp. 164, Décembre 2002.

[19] M. A. BOURENANE, " Indexation des Vidéos Personnelles par le Contenu", Mémoire, Université du Québec à Trois-Rivières, 2008.

[20] BOUCHER and T. LE, "Comment Extraire la Sémantique d'une Image", 3rd International Conference : Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia, Mars 2005.

[21] Y. CHEN, True Motion Estimation Théory, Application, And Implementation. Thesis, presented to the faculty of Princetown University in Candidacy for the Degree of Doctor of Philosophy, November 1998.

[22] M. J. CAREY, E. S. Parris and H. Lloyed-Thomas, A Comparison of Features for Speech, Music Discrimination, IEEE, pp. 149-152, Mars 1999.

[23] J. S. BORECZHY and L. D. WILCOX, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features", IEEE International Conference, Seattle, WA, USA ,vol. 6, pp. 3741-3744, Mai 1998.

[24] A. MESTAN, "Introduction aux Réseaux de Neurones Artificiels Feed Forward", [en ligne] "ftp://ftp-developpez.com/alp/tutoriels/intelligence-artificielle/reseaux-de-neurones/reseaux-de-neurones.pdf "

[25] S. DRAGHICI, "A Neural Network Based Artificial Vision System for Licence Plate Recognition", Dept. of Computer Science, Wayne State University.

[26] O. MARTINSKY, "Algorithmic and Mathematical Principles of Automatic Number Plate Recognition systems", 2007.

[27] P. TIRILLY, V. CLAVEAU, And P. GROS, "Annotation d'Images sur de Grands Corpus Réels de Données - Limites des Histogrammes de Couleurs", Saint-Etienne, pp. 473–478, mars 2007.

[28] H. MÜLLER, S. MARCHAND-MAILLET, and T. PUN, "The Truth about Corel - Evaluation in Image Retrieval", Computer Vision Group, Proceedings of the International Conference on Image and Video Retrieval, University of Geneva, pp. 38-49, Jan. 2002.

[29] B. P. Bertrand, "Vision et Morphologie", Annotation Semi-automatique d'Images [En ligne], Disponible sur : « http://www.ensmp.fr/ingenieurcivil/Options/Options2006/VM.pdf », (consulté le 15.03.2008).

[30] G. QUENOT, "Apprentissage actif pour l'Indexation des Images et des Vidéos", [en ligne]. Disponible sur : « http://clips.imag.fr/mrim/georges.quenot », (consulté le 10.04.2008).

[31] C. ZHANG and T. CHEN, "Annotating Retrieval Database with Active Learning", IEEE International Conference on Image Processing, Dept, of Electrical and Computer Engineering, USA, vol.3, pp. 595-8, Sept. 2003.

[32] Y. SONG, X. HUA, L. DAL, and R. WANG, "Semi-Automatic Video Semantic Annotation Based on Active Learning with Multiple Complementary Predictors", International Multimedia Conference, Singapore, pp. 97-104, November 2005.

[33] H. T. NGUYEN and A. SMEULDERS, "Active Learning Using Pre-Clustering", International Conference on Machine Learning, Banff, Canada, Vol. 69, pp. 97, July 2004.

[34] W. YING and H. ZHANG, "An Indexing and Browsing System for Home Video", Hewlett-Packard Laboratories, Proceedings of Eight IEEE International Symposium on Multimedia, Palo Alto, CA, 2006.

[35] Wikipédia, "Annotation automatique d'images" [en ligne]. Disponible sur : <http://fr.wikipedia.org/wiki/Annotation_automatique_d'images>, (consulté le 15.07.2008).

[36] R. FABLET " Modélisation Statistique non Paramétrique et Reconnaissance du mouvement dans des Séquences d'Images ; Application à l'Indexation de la Vidéo" Thèse sur le Traitement du Signal, Université de Rennes, 242 p, Juillet 2001.

[37] Y. SONG, X. HUA, L. DAI and R. WANG, "Semi-Automatic Video Semantic Annotation Based on Active Learning", Visual Communications and Image Processing 2005, Vol. 5960, pp. 251-258, SPIE, Bellingham, WA, 2005.

[38] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS, "Structuration Multimodale d'une Vidéo de Tennis par Modèles de Markov Cachés", Colloque sur le Traitement du Signal et des Images, France, 2003.

[39] R. VEZZANI, C. GRANA, D. BULGARELLI, and R. CUCCIARA, "A Semi-automatic Video Annotation Tool with MPEG-7 Content Collections", IEEE International Symposium on Multimedia, Washington, pp. 742-745, December 2006.

[40] J. KITTLER, K. MESSER, W. J. CHRISTMAS, B. LEVENAISE-OBADIA, D. KOUBAROULIS, "Generation of Semantic Cues for Sports Video Annotation", International Conference on Image Processing, Centre for Vision, Speech and Signal Processing, School of Electronics, Computing and Mathematics, University of Surrey, Guildford, UK, vol.3, pp. 26-29, 2001.