Extraction et gestion d'informations pour la construction d'une base vidéo d'apprentissage

Alain Simac-Lejeune*

*Litii
alain.simaclejeune@litii.com,
http://www.litii.com

Résumé. Indexer une vidéo consiste à rattacher un ou plusieurs concepts à des segments de cette vidéo, un concept étant défini comme une représentation intellectuelle d'une idée abstraite. L'indexation automatique se base sur l'extraction automatique de caractéristiques fournies par un système de traitement d'images. Cependant, il est nécessaire de définir les index ou concepts. Pour cela il faut définir le lien qui existe entre ces caractéristiques et ces concepts. Ce qui sépare les caractéristiques extraites sur lesquelles se base l'indexation automatique et les concepts est appelé fossé sémantique qui est le manque de concordance entre les informations que les machines peuvent extraire depuis les documents numériques et les interprétations que les humaines en font. La définition d'un concept peut être faite automatiquement si l'on dispose d'une base d'apprentissage liée au concept. Dans ce cas, il est possible "d'apprendre" le concept de manière statistique. Mais la construction de cette base d'apprentissage nécessite de faire intervenir un utilisateur ou un expert applicatif. En fait, il s'agit de s'appuyer sur ses connaissances pour extraire des segments vidéo représentatifs du concept que l'on souhaite définir. On peut lui demander d'indexer manuellement la base d'apprentissage, mais cette opération est longue et fastidieuse. Dans cet article, nous proposons une méthode qui permet d'extraire l'expertise pour que l'implication de l'expert soit la plus simple et la plus limitée possible.

1 Introduction

Actuellement, de nombreuses recherches traitent de l'analyse et de la reconnaissance d'activité. Ces recherches portent sur le développement de traitements automatiques de vidéo en terme d'extraction d'informations, de modélisation de ces informations et de reconnaissance automatique. La construction et l'évaluation de ces traitements requièrent des corpus vidéo représentant de nombreuses heures. Ces vidéos sont généralement annotées manuellement. Dans le cas de vidéo de longue durée, l'annotation manuelle est source d'erreurs, non reproductible et chronophage. De plus, la qualité des annotations dépend grandement de l'expertise de l'utilisateur. L'association de cette expertise à des traitements automatiques facilite cette tâche et représente un gain de temps.
La difficulté d'intégration de l'utilisateur dans la boucle de définition ne réside pas directement

dans l'utilisateur mais dans la manière de lui représenter les informations extraites c'est à dire dans le saut du fossé sémantique qu'il est nécessaire d'effectuer entre les données extraites et ce qu'il souhaite définir. C'est pourquoi nous proposons d'utiliser un modèle de données simple mais à la sémantique plus élevée que des primitives images ainsi qu'un système de questions-réponses permettant la construction d'un modèle de concept.
Après avoir présenté rapidement l'état de l'art, nous décrivons d'abord l'architecture et le fonctionnement de notre approche, puis nous présenterons les différentes étapes de constitution de la base d'apprentissage : extraction d'informations, modélisation de ces informations, création des modèles de 'concept'. Enfin, nous présentons l'évaluation de cette approche en la comparant à une méthode manuelle.

2 État de l'art et approche

La création d'une base d'apprentissage est une tâche qui s'effectue de manière manuelle la plupart du temps. Pour cela, un ou plusieurs utilisateurs regardent, généralement en accéléré, l'intégralité, un résumé ou des segments et annotent ceux-ci. Mais cette méthode demande beaucoup de temps et l'annotation reste d'une qualité variable puisque dépendante de l'annotateur. Pour ces raisons, des propositions ont été formulées afin d'automatiser ou tout au moins, minimiser l'implication humaine, de l'annotation. Ainsi, des méthodes (Rehatschek et Müller (1999); Correira et Chambel (1999); Kokkoras et al. (2002); Lin et al. (2003)) ont vu le jour pour assister un utilisateur. Dernièrement, la méthode la plus utilisée est celle intitulée "annotation collaborative"(Ayache et Quénot (2008)) qui consiste à utiliser les annotations successives pour améliorer la technique de sélection des séquences à annoter. Celle-ci permet de diminuer le temps d'annotation de manière significative tout en assurant une qualité minimale indispensable pour assurer le bon fonctionnement des algorithmes d'apprentissage. Notre approche se

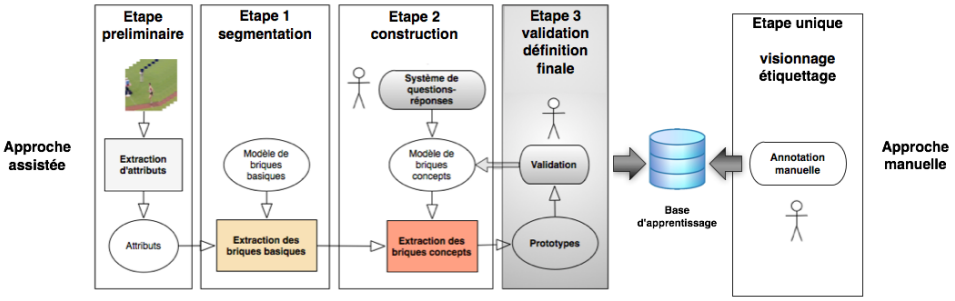

FIG. 1 – *Processus proposé.*

décompose en deux étapes (figure 1) :

- une première étape 'offline' (étape préliminaire et 1) d'extraction d'informations et de combinaisons des primitives extraites pour la construction d'attributs de niveau sémantique plus important,
- une seconde étape 'online' (étape 2 et 3) de construction de modèles spatio-temporels permettant la définition d'un concept et l'annotation des séquences correspondantes.

3 Extraction d'informations

La première étape est l'extraction d'informations des images formant les séquences afin de diminuer la quantité d'informations tout en sélectionnant les informations les plus pertinentes. Nous avons utilisé deux types d'informations : celles permettant la caractérisation des mouvements et celles permettant la caractérisation de l'environnement dans lequel évoluent les objets.

3.1 Extraction de primitives

Afin d'extraire des primitives intéressantes, nous avons utilisé 8 d'extracteurs classiques d'informations qui sont (figure 2) : les points d'intérêt spatiaux (Harris et Stephens (1988), SIP dans la suite) qui sont des coins spatiaux dans l'image ; les points d'intérêt spatio-temporels (Laptev (2005); Simac-Lejeune et al. (2010), STIP dans la suite) qui représentent des coins spatiaux ayant un mouvement discontinue dans le temps ; les couleurs dominantes Ravishankar et al. (1999) obtenues par analyse de mode dans un histogramme Teinte-Saturation-Luminance (TLS dans la suite) ; le flot optique (Ranchin et Dibos (2005)) calculé sur un ensemble de points particuliers (ici, les points d'intérêt spatiaux) et non sur une grille régulière ; les droites caractéristiques (Duda et Hart (1972)) extraites en utilisant un extracteur de Hough classique ; les caractéristiques du mouvement dominant calculées en utilisant le modèle d'Odobez-Bouthemy (Odobez et al. (1994)) ; le masque des objets en mouvement dans l'image obtenu par analyse du mouvement dominant dans le cas d'une caméra mobile et par extraction du fond (Boykov et al. (1999); Martel-Brisson et Zaccarin (2005)) dans le cas d'une caméra fixe.

FIG. 2 – *Exemple de primitives extraites : image compensée, flot optique, lignes caractéristiques, puis sur la seconde ligne, points d'intérêt spatio-temporels, masque des objets en mouvement et histogramme HLS.*

3.2 Construction d'attributs

Ces primitives extraites nous permettent ensuite de créer par combinaison et par construction, 24 caractéristiques de plus haut niveau sémantique. Ces combinaisons ont été définies empiriquement :

Caractéristiques	**Méthode d'obtention**
boite englobante d'un objet en mouvement	calcul du masque
nombre de lignes	somme des lignes extraites (Hough)
couleurs dominantes	couleur des 3 modes principaux de l'histogramme HLS
type de caméra (fixe ou mobile)	fixé par l'utilisateur
orientation et zoom de la caméra	mouvement dominant
position horizontale d'un objet	milieu vertical boite englobante
position verticale d'un objet	milieu horizontal boite englobante
compacité d'un objet	rapport hauteur/largeur de l'objet
taille relative d'un objet	rapport de la boite englobante avec celui de l'image
intensité du flot optique	vecteurs flot optique
orientation du flot optique	vecteurs flot optique
quantité de SIP/STIP par image/objet/cadran	nombre de points d'intérêt
nombre de changements (mouvement)	détection de mode dans le suivi du nombre de STIP

4 Modèle de représentation

4.1 Définition

Indexer un document vidéo consiste à associer un segment vidéo, c'est à dire une suite d'images, à un ou plusieurs concepts signifiants pour l'utilisateur. Le concept peut concerner l'image complète, ou simplement une partie spécifique de l'image, celle-ci étant souvent segmentée en objets. Au final, pour un concept donné, il s'agit de déterminer les suite d'images pour lesquelles ce concept apparaît. Nous proposons d'appeler 'briques' ces suites d'images du fait de leur représentation 3D (les deux dimensions de l'image et la dimension temporelle), elles représentent une suite d'images consécutives dans le plan et correspondent à la validité de certaines propriétés en terme d'attributs.

4.2 Les briques

A partir de la définition d'une brique spatio-temporelle et des attributs dont nous disposons, on peut construire de nombreux modèles de briques. Dans le cas général, un concept signifiant pour l'utilisateur correspond le plus souvent à un modèle mettant en jeu plusieurs attributs parfois combinés de manière évoluée (présence de règles notamment). Nous proposons donc une structure hiérarchique à deux niveaux, un premier niveau où l'on extrait des briques dites de base, puis un deuxième niveau où l'on combine celles-ci afin d'extraire celles correspondantes au concept.

Premier niveau : les briques de base sont directement définies à partir des caractéristiques extraites du document vidéo : une caractéristique particulière vérifie une propriété particulière. Ainsi une brique de base sera caractérisée par un couple caractéristique/propriété, ce couplage pouvant prendre différentes formes.
Deuxième niveau : le concept est défini à partir d'une combinaison de caractéristiques vérifiant certaines propriétés (règles de combinaison, de présence, de succession, d'interdiction, etc.). On définit les briques combinées correspondant à ces combinaisons de caractéristiques jusqu'à la brique concept.
La structure en briques nécessite 2 étapes :

- une première étape de modélisation consistant à définir les modèles de représentation des briques de base et des briques combinées c'est à dire la manière de les définir, de les construire, de les représenter, de les assembler et enfin de les utiliser,
- une seconde étape d'extraction consistant à extraire les briques de séquences d'images. Cette extraction est réalisée en utilisant les caractéristiques extraites de chaque séquence et des modèles de briques définis.

4.3 Modèle de brique basique

Un modèle de briques de base correspond à la définition d'un type de brique. Deux éléments le composent :

- une caractéristique extraite de la séquence d'images comme le nombre d'objets en mouvement ou un attribut bas-niveau construit à partir d'une caractéristique extraite comme la vitesse et l'orientation de la caméra (construit à partir du flot optique) ;
- une propriété de validation sur la caractéristique qui peut être soit une valeur soit un ensemble de définition.

Par exemple : une *brique* concernant le '**type de caméra**' peut prendre deux valeurs : fixe ou mobile (2 valeurs possibles) ; une *brique* '**compacité**' peut prendre toutes valeurs comprises entre 0 et 1 (une infinité de valeurs possibles). Chaque couple caractéristique/propriété définit donc une brique de base. L'ensemble des briques ainsi définies permet de segmenter une séquence vidéo.

4.4 Les opérateurs de combinaison

A partir d'un ensemble de modèles de briques de base liées directement aux caractéristiques extraites, il est nécessaire de définir des briques élaborées résultats de la combinaison d'autres briques (de bases ou elles-mêmes combinées) afin de représenter un concept. Pour répondre au besoin de la combinaison, nous proposons l'utilisation d'opérateurs temporels complétés par des opérateurs logiques.
La difficulté rencontrée est la représentation et la modélisation des trois éléments clés pour l'analyse temporelle : la date, la durée des événements et le délai entre événements. Il existe différents modèles permettant de modéliser le temps et se déclinent en différents aspects (ponctuel ou intervalle, temps ordonné ou non, discret ou dense...) - (Le Ber et al. (2007)). Nous avons choisi d'utiliser l'algèbre d'intervalles de Allen (1981)(voir figure 3) qui propose 13 relations : b (avant), m (rencontre), o (se recouvre), s (fait démarrer), d (pendant), f (termine), eq (est égale) et leurs relations inverses bi, mi, oi, si, di, fi. La durée n'étant pas représentée dans ce modèle, nous l'avons complété à l'aide des opérateurs 'interval and duration' dit INDU

(Pujari et al. (1999)) : 3 relations permettant de comparer la durée de deux intervalles X et Y : X est de durée inférieure à Y (X<Y), supérieure (X>Y) ou égale (X=Y). Enfin, 3 opérateurs complètent cet ensemble : le OU exclusif, la répétition et choix multiple entre. Certains combinaisons étant possibles, on obtient au total 25 opérateurs.

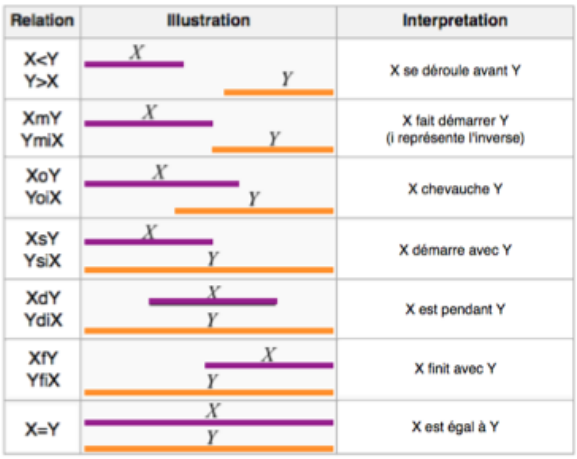

FIG. 3 – *Les opérateurs de Allen.*

4.5 Exemple

On rappelle que la compacité se calcule en utilisant les dimensions de la boite englobante et correspond au minium du rapport hauteur/largeur ou largeur/hauteur.
En utilisant les briques de compacité : compacité faible $(c-f)$, compacité moyenne $(c-m)$, compacité forte $(c-F)$, orientation du flot optique unique $(fo-*)^1$, intensité du flot optique moyenne $(fo-io-mo)$ et les opérateurs : m (suivi de), d (en même temps), on peut construire une définition de la course à pied :

$$(c-f \ \mathbf{m} \ c-m \ \mathbf{m} \ c-F \ \mathbf{m} \ c-m) \ \mathbf{d} \ (fo-io-m) \ \mathbf{d} \ (fo-*)^1 \qquad (1)$$

Ce que l'on peut traduire par "succession de compacité faible/moyenne/forte avec un vecteur moyen d'intensité moyenne du flot optique, orienté dans une seule direction tout en ayant une brique de compacité '*faible*' plus longue que la moyenne et une '*forte*' plus longue que la moyenne". La succession est donnée par l'opérateur m utilisé sur les briques de compacité. La présence simultanée est obtenue par l'usage de l'opérateur d entre le groupe de la compacité, de l'intensité et de l'orientation uniforme du flot optique.
Au final, on obtient une définition du concept 'course à pied'. C'est définition n'est pas unique. Une autre définition, utilisant des briques différentes serait tout à fait envisageable. Le principal est que la définition choisie permette la récupération du concept recherché.

5 Construction de concepts

Afin d'assister l'utilisateur dans sa construction de la définition d'un concept, notre approche est basée sur l'utilisation d'un système de questions-réponses avec lequel l'utilisateur va interagir afin de choisir les éléments constitutifs de la définition.
Les systèmes de questions/réponses que l'on rencontre habituellement sont construits afin d'aider l'utilisateur à construire une recherche sur un ensemble de données préparées (étiquetées, indexées, classées, annotées, etc.).
Le système de questions/réponses (Q/R) représente les connaissances de l'expert en traitement d'images sur les caractéristiques extraites de la vidéo et sur l'interprétation générique qui peut en être faite. Afin qu'il soit efficace, ce système doit être bien structuré.

5.1 Définition

Nous avons choisi d'utiliser un formalisme objet organisé sous forme de graphe de type arbre (planaire enraciné ce qui permet un parcours des sommets dans l'ordre lexicographique en utilisant un parcours en profondeur préfixé). Cette représentation permet de modéliser les questions/réponses où chaque entité est composée d'une question, de réponses associées, et de relations vers des briques, des opérateurs ou d'autres questions mais aussi de représenter la hiérarchie entre questions et le regroupement thématique des questions. Chaque question/-réponses est un objet composé d'un objet question et d'objets réponse. Les informations sur chaque question sont représentées par les attributs de l'objet et les relations entre objets.
Le système de questions/réponses doit répondre à un certain nombre de problématiques au regard de la structuration en briques décrite dans le chapitre précédent :

- il doit permettre de lier des réponses à des briques de base ;
- il doit permettre de définir les liaisons entre briques afin de définir des briques composées ;
- il doit assurer une certaine cohérence dans le processus de questionnement (éviter les questions sans objet, maintenir une suite "logique" dans les questions).

5.2 Les questions

Les questions sont de trois types : celles permettant d'obtenir des informations sur le concept (elles sont reliées aux briques, c'est le cas de la plupart d'entre elles), celles permettant d'obtenir des informations sur les liaisons (elles sont reliées aux opérateurs) et enfin celles permettant d'obtenir des informations de navigation (elles donnent des informations d'accessibilité sur les questions utilisables après la question courante).

5.3 Liaisons avec les attributs

Les questions à "réponses-briques" (la plupart des questions sont de ce type) permettent la sélection de briques de base liées au concept à définir. Une réponse est liée à une ou plusieurs briques de base.
Les questions à "réponses-liaison" sont plus difficiles à construire. Il s'agit de questions qui permettent la sélection d'opérateurs de liaison entre briques. Ce qui est particulièrement délicat

n'est pas le choix d'un opérateur mais le choix des briques à lier à l'aide de l'opérateur choisi. On distingue trois cas dans la mise en place de l'opérateur :

- la réponse sélectionne simultanément un opérateur de liaison et des briques de base ainsi la liaison est directe.
- la réponse sélectionne uniquement un opérateur sans sélectionner de briques et sans contrainte ou règle. C'est le système qui cherchera à appliquer l'opérateur sur les briques sélectionnées au moment de la définition finale.
- la réponse sélectionne un opérateur et une règle de liaison qui définit les briques admissibles à l'usage de l'opérateur. Ainsi, si les briques admissibles sont déjà sélectionnées, elles sont liées par l'application de l'opérateur sinon, on attend leur sélection. Dans le cas où elles ne sont pas sélectionnées à la fin du processus, l'opérateur n'est pas utilisé dans la définition finale.

5.4 Fonctionnement du système de Q/R

Le système a pour objectif la création d'un modèle de concept en utilisant une interaction avec l'utilisateur. Le système de questions/réponses permet la sélection de briques constitutives et d'opérateurs de liaison. Une définition complète est effectuée en répondant à une série de questions (en général, autour de 15-20 questions) sélectionnées en utilisant les réponses précédemment données :

1. *Quel est le type d'activité ? réponse choisie : déplacement => activation des thèmes "mouvement', "déplacement" et "environnement".*
2. *Y a-t-il un déplacement par rapport au sol ? réponse oui => sélection des briques caméra (cam-mobile) et translation*
3. *Est-ce un mouvement régulier, ponctuel, unique, séquentiel ? réponse séquentiel => sélection de l'opérateur de séquentialité (m)*

A la fin du processus, on recherche dans la base de données vidéo de définition, les séquences susceptibles de contenir le concept défini. L'utilisateur valide celles correspondantes et invalide celles non correspondantes. Les séquences validées permettent éventuellement d'ajuster la définition initiale pour obtenir la définition finale du concept : c'est le processus de validation.

6 Base d'apprentissage

6.1 Les données

Afin de caractériser les performances des approches proposées, nous avons utilisé les séquences d'une base de vidéos existante et disponible pour tous [1] qui est une base constituée par l'Université de Floride (University of Central Florida) : la base UCF Sports Action 50. Elle est composée de 5000 séquences collectées sur Youtube et réparties sur 50 catégories à raison de 100 séquences par catégorie (liste figure 4). Chaque séquence dure 4 secondes à raison de 25 images par seconde (100 images), dans une résolution de 320 pixels par 240 pixels. Pour

1. http ://server.cs.ucf.edu/ vision/data.html

l'ensemble des séquences de cette base, il est important de noter qu'aucune ne présente de transition. Il s'agit donc de plans.

6.2 Bases d'apprentissage et de test

A partir de la base de 5000 séquences, nous avons construit deux bases de 2500 séquences disposant d'égales répartitions sur les différentes catégories (50 séquences de chacune des 50 catégories). La première base sera la base à annoter. Une fois annotée, elle servira à l'apprentissage des 50 concepts par notre système d'apprentissage.
La seconde base servira de base de test en permettant l'indexation des séquences en utilisant les modèles appris par le système.

6.3 Apprentissage

Nous avons utilisé le système WEKA [2] (Waikato Environment for Knowledge Analysis) et l'algorithme de Machine à Vecteurs de Support (SVM) en mode multi-classe. Les données d'entrée sont sous format ARFF (pour Attribute-Relation File Format) importées en CSV et sont composées de l'ensemble des valeurs des 24 attributs présentés (section 3.2).

Pour chaque séquence, composée de plusieurs images, on dispose des attributs pour chaque image. Afin d'obtenir un seul groupe d'attribut par séquence, les valeurs de chaque attribut sont moyennées dans le temps. Ainsi, chaque ligne correspond aux valeurs moyennes des 24 attributs ainsi que la classe annotée à la main ou fournie par le système de questions/réponses. Au final, dans les deux méthodes, les valeurs des attributs sont identiques, le biais engendré par le calcul de la moyenne n'avantage aucune des deux méthodes.

7 Évaluation de la base

7.1 Protocole

Dans le but d'évaluer notre approche, nous avons annoté la base d'apprentissage une fois manuellement et une fois via notre système. Cette base d'apprentissage a ensuite été utilisée dans un classifieur SVM et le résultat est enfin utilisé pour faire l'indexation de la base de test. La comparaison de l'indexation de la base de test permet l'évaluation de la qualité de la préparation de la base d'apprentissage, seul élément différent entre les deux essais.
La comparaison s'effectue en terme de précision/rappel/f-mesure entre chacune des cinquante catégories ainsi qu'en terme de temps d'annotation.
On rappelle que le **rappel** est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données et que la **précision** est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé par le système pour une requête donnée. La **f-mesure** est une mesure qui combine la précision et le rappel est leur pondération, aussi appelée F-score, et représentant 2 fois le produit de la précision et du rappel sur la somme de la précision et du rappel.

2. http ://www.cs.waikato.ac.nz/ml/weka/

7.2 Résultats : temps et qualité

En terme de **temps**, la préparation des deux bases d'apprentissage prend 6 heures et 56 minutes pour celle annotée manuellement et 3 heures et 20 minutes pour celle annotée de manière assistée. L'annotation manuelle est effectuée selon les modalités suivantes (qui sont les conditions habituelles d'annotation) : une annotateur est chargé de 5 concepts, il regarde l'intégralité des séquences, en vitesse normale (généralement, la vitesse est accélérée mais pour des séquences courtes (4 secondes), l'accélération rend impossible à l'annotateur de saisir le concept).
L'annotation assistée correspond à l'utilisation du système de questions-réponses sur l'ensemble des concepts en utilisant en moyenne 18 questions et en visionnant et validant en moyenne 20 prototypes. Cette annotation assistée est ainsi effectuée en 4 minutes en moyenne.
Le tableau 4 présente les résultats obtenus en terme de **qualité** sous forme de précision, de rappel et de f-mesure sur la même base de test en utilisant deux bases d'apprentissage distinctes : la base annotée manuellement et la base annotée de manière assistée.

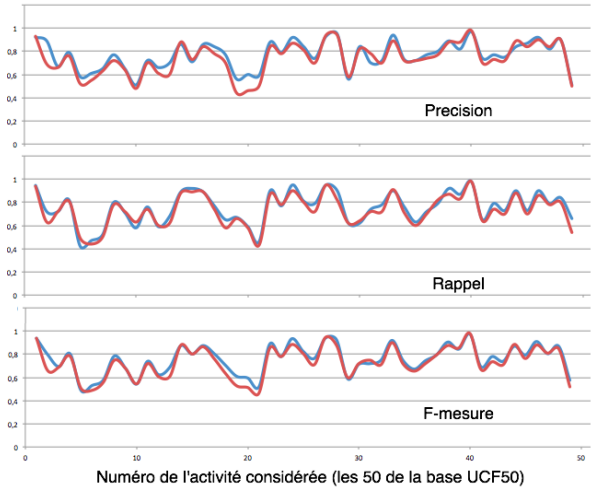

FIG. 4 – *Précision, rappel et f-mesure sur les deux méthodes (manuelle et automatique) pour les 50 catégories.*

7.3 Analyse

A partir de la figure 4, on peut observer que bien que les résultats sur la base assistée soient généralement un peu moins bons que la base manuelle, il reste à un niveau intéressant et sont très proches tant en précision, qu'en rappel ou qu'en f-mesure (souvent autour de 0.8). On remarque également que certaintes catégories obtiennent des résultats identiques comme 'ski', 'escalade' ou 'baseball'.
Au final, la précision moyenne s'établit à 0,766 pour la version manuelle contre 0,738 pour l'automatique ; le rappel moyen à 0,754 contre 0,732 et la f-mesure à 0,757 contre 0,732. Les

résultats sont donc très proches bien que moins bons pour la méthode automatique.
Enfin, le temps nécessaire à l'annotation a été divisé par deux dans notre exemple. Cette diminution dépend étroitement du nombre de catégories et du nombre de séquences. Plus le nombre de séquences est grand et plus l'apport est important alors que plus le nombre de catégories augmente et moins l'apport est important.

8 Conclusion

L'annotation assistée par un système de questions-réponses de concepts spatio-temporels modélisés par un modèle de briques combinées montre des résultats probants. Toutefois, elle présente quelques limitations. La première concerne le paramétrage des briques qui est difficile à établir bien qu'il soit envisageable d'effectuer une boucle de rétroaction afin de modifier le paramétrage. La seconde correspond à la création des liaisons entre questions/réponses et caractéristiques/opérateurs puisqu'elle soit effectuée manuellement et empiriquement. Une méthode pouvant permettre de créer ses liaisons est également une stratégie de rétroactions. Après une série de réponses et donc une sélection de briques, on demande à l'utilisateur de choisir des séquences correspondant à la définition. Il est ensuite nécessaire d'effectuer un apprentissage. Cependant, on perd de l'intérêt du système d'assistance à l'annotation. La dernière limite porte sur la généricité du système proposé. En effet, les résultats ont été obtenus sur 50 concepts simples, plutôt bien distincts les uns des autres et présentant chacun des spécificités dans les mouvements. Il serait nécessaire de l'évaluer sur une base hétérogène présentant un contenu d'une difficulté plus importante. Cependant, la force de cette approche, par rapport à d'autres approches (par exemple, la compensation de mouvement), est qu'elle permet d'obtenir une qualité d'annotation acceptable puisqu'elle permet d'obtenir des résultats la plupart du temps proche ou identique à l'annotation manuelle. De plus, cette approche permet de diminuer de manière très intéressante le temps nécessaire à l'utilisateur pour créer une base d'apprentissage.

Références

Allen, J. F. (1981). An interval-based representation of temporal knowledge. In P. J. Hayes (Ed.), *IJCAI*, pp. 221–226. William Kaufmann.

Ayache, S. et G. Quénot (2008). Video corpus annotation using active learning. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, et R. W. White (Eds.), *ECIR*, Volume 4956 of *Lecture Notes in Computer Science*, pp. 187–198. Springer.

Boykov, Y., O. Veksler, et R. Zabih (1999). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 2001.

Correira, N. et T. Chambel (1999). Active video watching using annotation. In *ACM Multimedia (2)*, pp. 151–154.

Duda, R. O. et P. E. Hart (1972). Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM 15*, 11–15.

Harris, C. et M. Stephens (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151.

Kokkoras, F., H. Jiang, I. P. Vlahavas, A. K. Elmagarmid, E. N. Houstis, et W. G. Aref (2002). Smart videotext : a video data model based on conceptual graphs. *Multimedia Syst. 8*(4), 328–338.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision 64*(2-3), 107–123.

Le Ber, F., G. Ligozat, et O. Papini (2007). Raisonnements sur l'espace et le temps : des modèles aux applications. *VISUAL*, 383–390.

Lin, C., B. Tseng, et J. Smith (2003). Video collaborative annotation forum : Establishing ground-truth labels on large multimedia datasets. In *Proc. of the TRECVID Workshop*.

Martel-Brisson, N. et A. Zaccarin (2005). Moving cast shadow detection from a gaussian mixture shadow model. In *CVPR (2)*, pp. 643–648. IEEE Computer Society.

Odobez, P. Bouthemy, et P. Temis (1994). Robust multiresolution estimation of parametric. *Journal of Visual Communication and Image Representation 6*, 348–365.

Pujari, A. K., G. V. Kumari, et A. Sattar (1999). Indu: An interval and duration network. In N. Y. Foo (Ed.), *Australian Joint Conference on Artificial Intelligence*, Volume 1747 of *Lecture Notes in Computer Science*, pp. 291–303. Springer.

Ranchin, F. et F. Dibos (2005). Segmentation d'objets en mouvement par utilisation du flot optique. *ORASIS*, 383–390.

Ravishankar, K. C., B. G. Prasad, S. K. Gupta, et K. K. Biswas (1999). Dominant color region based indexing for cbir. In *ICIAP*, pp. 887–892. IEEE Computer Society.

Rehatschek, H. et H. Müller (1999). A generic annotation model for video databases. In D. P. Huijsmans et A. W. M. Smeulders (Eds.), *VISUAL*, Volume 1614 of *Lecture Notes in Computer Science*, pp. 383–390. Springer.

Simac-Lejeune, A., M. Rombaut, et P. Lambert (2010). Spatio-temporal block model for video indexation assistance. *Knowledge Discovery and Information Retrieval*, 475–480.

Summary

Video indexing is link to one or more concepts with specific segments of the video. A concept is defined as a mental description of an abstract idea. Automatic indexing is based on the automatic features extraction provided by an image processing system. However, it is neccessary to define the index or concepts. It requires to definine the relationship between these characteristics and concepts. The semantic gap represents the notion which separates the extracted features based on the automatic indexing and concepts. It is the mismatch between the information extracted by machines from the digital documents and human interpretations. A concept can be automatically defined if a learning database linked to the concept is available. In this case, it is possible to statistically "learn" the concept. But the building of this base of learning needs to involve an user or an expert in application. In fact, it is based on his knowledge to extract video segments representing the defined concept. The base of learning can be manually indexed by the expert but it is a long and boring operation. In this paper, we propose a method allowing to automatically extract the expertise in order that the expert implication is as simple and as limited as possible.