Fouille vidéo orientée objet, une approche générique

Jonathan Weber*, Sébastien Lefèvre**
Pierre Gançarski*

*Université de Strasbourg
j.weber@unistra.fr,gancarski@unistra.fr
https://lsiit.u-strasbg.fr/
**Université Bretagne-Sud
sebastien.lefevre@univ-ubs.fr
http://www-valoria.univ-ubs.fr/

Résumé. Actuellement, le média vidéo est une des premières sources d'information, mais aussi une des plus volumineuses. Pour traiter cette masse d'informations, les systèmes actuels de fouille vidéo font face à un problème de fossé sémantique : il existe une différence entre la signification sémantique du contenu des séquences vidéos et l'information numérique codée dans les fichiers associés. Ce fossé peut être en partie comblé par l'utilisation des objets réels (du point de vue de l'utilisateur) présents dans les séquences. Cependant la fouille vidéo orientée objet nécessite l'introduction d'informations sémantiques, que ce soit pour l'extraction des objets ou pour la fouille de ces objets. Nous proposons d'introduire de telles informations par le biais d'une interaction avec l'utilisateur. Cette interaction consiste en un mécanisme de retour de pertinence. Le système propose à l'utilisateur un échantillon des résultats obtenus, puis l'utilisateur valide, invalide ou corrige ces résultats. Ces informations de validation/invalidation/correction sont alors utilisées pour guider le système et lui permettre d'améliorer les résultats qu'il produit. Cet article ne propose pas un système complètement opérationnel mais explore certaines pistes pour arriver à un tel système.

1 Introduction

Après l'augmentation massive des données textuelles, et plus récemment des images, disponibles dans des bases de données et sur le Web, nous observons aujourd'hui une augmentation dans le domaine de la vidéo. La vidéo est en train de devenir une des principales sources d'informations. L'aspect temporel des vidéos empêche un parcours rapide et efficace de telles bases de données. Cependant, l'aspect temporel est peu utilisé dans les algorithmes existants liés à la fouille vidéo, sauf dans le cas de la segmentation en plans où l'information temporelle joue un rôle presque exclusif (Koprinska et Carrato, 2001). La fouille de données (Cios et al., 2007) est le processus d'extraction d'informations et de connaissances d'une masse de données. La fouille vidéo (Rosenfeld et al., 2002) est donc l'application de ce processus à des données vidéo, c'est-à-dire des séquences temporelles d'images, éventuellement couplées à

des données audio. Cependant, dans cet article, nous considérerons uniquement les données visuelles. Selon ces définitions, un Système de Fouille Vidéo est un système capable d'extraire de l'information à partir d'une grande base de séquences vidéo.

Plusieurs auteurs se sont essayés à dresser un état de l'art relatif à la fouille vidéo. Cependant, aucun de ces travaux ne présente l'ensemble du domaine de la fouille vidéo, chaque contribution concernant plutôt un sous-domaine (par exemple l'indexation). Parmi les premiers travaux, Idris et Panchanathan (1997) présentent quelques méthodes d'indexation vidéo et précisent que le niveau normal pour analyser le contenu visuel devrait être l'objet. Brunelli et al. (1999) présentent également des systèmes d'indexation vidéo et notent, en 1999, que la détection des objets génériques ne peut pas être réalisée avec les méthodes actuelles. Dix ans plus tard, Brezeale et Cook (2008) étudient le niveau auquel les informations font l'objet d'une classification dans le domaine vidéo : la plupart des méthodes étudiées proposent de travailler au niveau global, quelques-unes utilisent le plan ou la scène et, surtout, aucune n'utilise l'objet. Money et Agius (2008) ont étudié les systèmes de résumé de vidéo. Ils proposent d'utiliser des informations propres à l'utilisateur pour produire des résumés personnalisés et plus riches sémantiquement. Ren et al. (2009) se sont concentrés sur l'utilisation d'informations spatio-temporelles pour la recherche de vidéos. Ils remarquent l'efficacité de l'utilisation des relations spatio-temporelles entre les objets pour résoudre le problème de la recherche de vidéos. Snoek et Worring (2009) présente la recherche de vidéo basée sur des concepts via une étude de 300 articles. Les auteurs insistent sur l'importance de l'efficacité computationnelle des méthodes et de disposer d'un très large panel de détecteurs de concepts permettant d'aborder la diversité des contenus vidéos. Contrairement à ces études, nous ne nous concentrons pas sur un objectif spécifique de fouille vidéo, mais nous souhaitons prendre en considération tous les objectifs possibles. Comme Idris et Panchanathan (1997), nous pensons que le niveau de l'objet est le plus adapté pour la fouille vidéo. Dans ce document, nous nous concentrons sur le rôle de l'objet dans le processus de fouille de données vidéo et discutons de la position de l'utilisateur comme élément fondamental du processus visuel d'exploitation.

Dans cet article, nous introduisons en premier lieu une nouvelle taxonomie pour caractériser les différents systèmes de fouille de données vidéo (SFV). Puis, nous étudions des SFV récents en utilisant la taxonomie introduite précédement. Nous déterminons ensuite les caractéristiques d'un SFV orienté objet, présentons le problème de l'extraction d'objets au sein de vidéo et étudions l'introduction de la sémantique au sein d'un tel système. Enfin, nous proposons un cadre générique qui permettra la création de SFV orienté objet.

2 Caractéristiques des SFV

De nombreux aspects sont à prendre en compte lorsque l'on conçoit un nouveau SFV. Dans cette section, nous allons identifier ces aspects et introduire certains termes qui peuvent être utilisés pour caractériser les SFV.

2.1 Objectifs des SFV

Les objectifs accomplis par un SFV sont variés et dépendent des besoins de ses utilisateurs. Les bases de vidéo nécessitent de grandes capacités de stockage et la fouille manuelle de ces bases est fastidieuse. Des SFV ont donc été développés pour accomplir de façon automatique

les tâches jusqu'alors accomplies par des êtres humains. Le *résumé de vidéo* (Res) vise à produire de courts et représentatifs extraits de vidéo dans le but de permettre aux utilisateurs de retrouver leurs thèmes et leurs contenus sans avoir à regarder la vidéo en entier. L'*indexation de vidéo* (Ind) est la caractérisation d'une vidéo afin d'être capable de la retrouver rapidement ultérieurement en utilisant des requêtes spécifiques. La *classification de vidéo* (Cla) vise à regrouper les vidéos dans des catégories prédéfinies afin d'identifier leur contenu. La *recherche basée sur le contenu* (Rec) permet aux utilisateurs de retrouver des vidéos similaires à une autre vidéo donnée en requête.

2.2 Propriétés des SFV

Les SFV sont caractérisées par différentes propriétés relatives à la nature des données et des informations à traiter, aux descripteurs à extraire et à l'échelle à laquelle les calculer, et au rôle de l'utilisateur. Dans cette section, nous introduisons et décrivons ces différentes propriétés.

Données

Un SFV peut avoir à traiter différents types de données vidéo. Une vidéo peut être compressée (C) par différents algorithmes ou disponible sous forme brute (B). La compression permet un stockage moins couteux en mémoire mais requiert un processus de décompression avant une visualisation et peut induire une perte d'information. Traiter des vidéos compressées est plus rapide car le volume de données à traiter est moindre, cependant l'extraction de concepts visuels est plus complexe, alors que l'analyse du contenu visuel peut être effectué directement dans le cadre de vidéos brutes (avec un cout de calcul plus élevé). Les SFV peuvent également être dédiés au traitement de types de vidéos spécifiques (S) ou génériques (G). Traiter des vidéos spécifiques permet d'obtenir de meilleurs résultats car l'on peut utiliser les connaissances du domaine dans le processus de fouille. Par contre, la prise en compte de vidéos génériques par un SFV facilite la réutilisation et l'adaptation de ce dernier à des contextes variés.

Éléments

Quel que soit l'objectif d'un SFV, il peut être suivi en considérant différentes éléments, de la vidéo dans son intégralité au simple pixel. La première étape en fouille vidéo consiste généralement à extraire l'élément à traiter. La *vidéo intégrale* (Vid) est l'élément classique et le plus simple à traiter car il ne nécessite aucune extraction. Néanmoins, si la vidéo contient des scènes très différentes, cet élément peut ne pas être très significatif. Une *scène* (Sce) est composée de plusieurs plans dans un contexte spatio-temporel identique et peut être difficile à extraire. Un *plan* (Pla) est un segment de vidéo délimité par deux transitions, le problème de son extraction est un sujet très étudié et de nombreuses méthodes ont été proposées pour le résoudre (Lefèvre et al., 2003). La *trame* (Tra) est l'unité temporelle d'une vidéo, une vidéo étant une séquence temporelle de trames. Un *objet* (Obj) est, selon nous, l'élément qui comporte le plus de sémantique mais son utilisation est limitée par la dificulté d'extraire un objet réel à un instant donné, ou au cours du temps. Une *région* (Reg) est un ensemble de pixels connexes qui (au contraire de l'objet) ne repose pas sur un concept sémantique. Enfin, le *pixel (Pix)* est le

plus petit élément et, pris isolément, il n'apporte que peu voire pas d'information. La figure 1 représente les différents éléments.

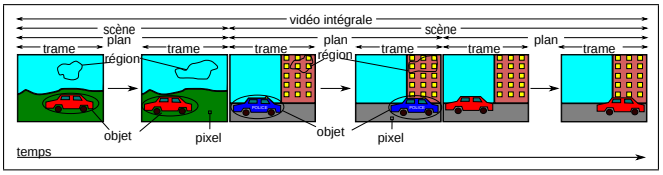

FIG. 1 – *Les différents éléments d'un SFV.*

Descripteurs

Il existe de nombreux descripteurs pour décrire le contenu d'une vidéo. Il est possible de décrire le *mouvement* (Mou) mais également la *couleur* (Coul). Il existe des moyens de caractériser une *texture* (Tex) présente dans les données visuelles. Nous pouvons abstraire les formes et les contours (For) afin de décrire la morphologie des éléments présents dans la vidéo. Outre ces descripteurs classiques, il existe de nombreux autres descripteurs spécifiques régulièrement proposés dans la littérature (Rui et al., 1999). Choisir le descripteur le plus adapté à un SFV précis n'est pas trivial car chaque descripteur vise à caractériser un contenu vidéo selon un point de vue particulier.

Échelles

Afin d'analyser et caractériser les éléments par les différents descripteurs, il est nécessaire de choisir l'échelle à laquelle les descripteurs vont être calculés sur les éléments. L'échelle est liée en partie à l'élément utilisé et au descripteur choisi. À une échelle *globale* (Glo), les descripteurs vidéos sont appliqués sur l'intégralité de la vidéo. Considérant l'échelle du *bloc* (Blo), la vidéo est divisée en blocs suivant une grille spatiale, les descripteurs sont alors calculés dans chaque bloc indépendamment. A la différence de l'échelle bloc, l'échelle *region* ne divise pas la vidéo en blocs mais en régions de tailles et formes variées par une étape de segmentation. Les descripteurs sont ensuite associés à chaque région indépendamment. L'échelle *objet* (Obj) consiste à définir les descripteurs pour les objets réels présents dans les éléments de la vidéo. L'échelle *point d'intérêts* consiste à calculer les descripteurs sur des points (et leur voisinage) dont le voisinage est particulier. Enfin l'échelle *pixel* (Pix) est la plus petite possible : les descripteurs ne servant alors qu'à décrire un pixel, cette échelle ne semble pas très utile. La figure 2 représente ces notions.

2.3 Implication de l'utilisateur

L'implication de l'utilisateur est un point critique dans un SFV. Il y a quatre niveaux d'implication possibles pour un utilisateur. Celle-ci peut être *Nulle* (Nul) si le système est totalement automatique et que l'utilisateur n'intervient pas dans le processus. Elle est *Supervisée* (Sup) lorsque l'utilisateur doit fournir un ensemble complet de données étiquetées afin

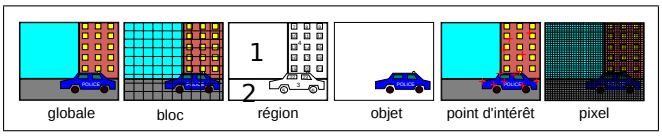

FIG. 2 – *Les différentes échelles de descripteurs possibles pour un SFV.*

de configurer le système pour traiter un jeu de données spécifiques. L'implication est dite *Semi-supervisée* (S-sup) quand l'utilisateur doit fournir moins de données étiquetées et/ou doit valider/invalider certains résultats pour guider le processus de fouille. Enfin, l'implication est appelée *Paramétrique* (Param) lorsque l'utilisateur doit fixer différents paramètres du système.

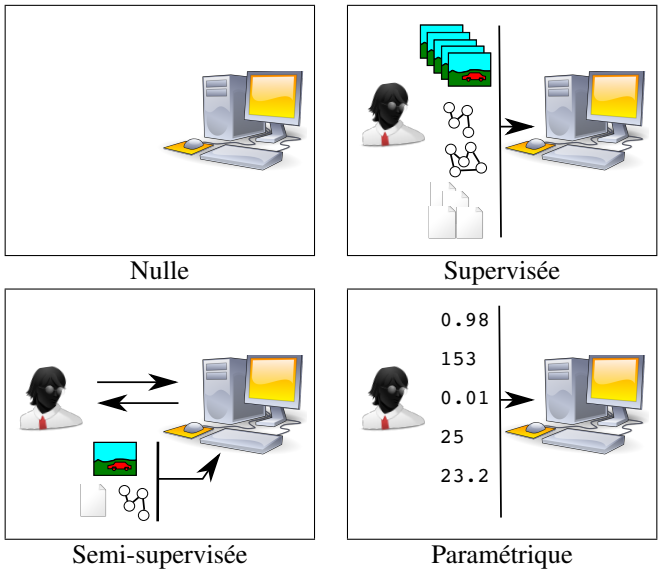

FIG. 3 – *Les différentes implications possibles de l'utilisateur dans un SFV.*

La taxonomie que nous avons introduite dans cette section permet de décrire et caractériser un SFV. Nous allons l'illustrer dans la section suivante en caractérisant les principaux SFV de la littérature.

3 Utilisation de l'objet dans la fouille vidéo

Dans cette section, nous caractérisons des travaux récents concernant de près ou de loin la fouille vidéo orientée-objet, afin de mettre en lumière les tendances actuelles dans ce domaine. Le tableau 1 résume les différentes caractéristiques des systèmes étudiés, selon la taxonomie introduite dans la section 2.

Méthodes	Tâches	Données	Elément	Descripteur	Echelle	Implication
Anjulan et Canagarajah (2007b)	Rec	B,G	Obj	LIR,SIFT	Reg	Param
Anjulan et Canagarajah (2007a)	Cla	B,G	Obj	LIR,SIFT	Reg	Param
de Avila et al. (2008)	Res	B,G	Vid	Col,LP	Glo	Param
Basharat et al. (2008)	Rec	B,G	Vid	SIFT,Col,Tex,Mot	Reg	Nul
Chevalier et al. (2007)	Rec	C,G	Obj	RAG	Reg	Param
Gao et al. (2009)	Rec	B,G	Pla	OFT	Blo	Param
Liu et Chen (2009)	Rec	B,G	Vid	Diverse	Obj	Param
Ren et Zhu (2008)	Res	B,G	Vid	PLR,ECR,HCC	Glo	Param
Sivic et Zisserman (2008)	Rec	B,G	Obj	SIFT	Reg	Param
Teixeira et Corte-Real (2009)	Cla	B,S	Obj	SIFT	Obj	Sup
Zhai et al. (2007)	Res	B,G	Vid	KNNG	Glo	Param

TAB. 1 – *Caractérisation des approches récentes en fouille vidéo.*

Ces articles récents traitent majoritairement des vidéos génériques non-compressées. La recherche est l'objectif le plus fréquent. En effet, la demande principale d'un utilisateur de SFV est certainement de retrouver les vidéos dont il a besoin, surtout dans le cas où celles-ci sont noyées dans une grande quantité de données. Le résumé de vidéo suscite également l'intérêt de la communauté, puisqu'il a pour but de permettre à l'utilisateur de connaître le contenu d'une vidéo sans avoir à la regarder dans son intégralité, ce qui se traduit par un gain de temps important. Mis à part dans le cas du résumé de vidéo, l'élément le plus courant semble être l'objet. Cependant, l'objet est loin d'être l'échelle la plus utilisée, les échelles globale et région sont les plus communes : en effet, produire une segmentation sémantique (en objets) de façon automatique reste aujourd'hui encore un problème ouvert. Les descripteurs utilisés sont variés et souvent combinés afin d'obtenir de meilleurs résultats. Enfin, la grande majorité des SFV demande à l'utilisateur de fixer des paramètres, ce qui est une tâche relativement peu intuitive. Notons qu'un SFV est supervisé tandis qu'un autre est complètement automatique. Aucun des SFV étudiés n'implique l'utilisateur de façon semi-supervisée, ce qui nous semble pourtant un moyen fiable et léger pour guider le système.

4 Vers une fouille vidéo orientée-objet

L'étude des tendances récentes dans la fouille vidéo montre que si les échelles objet et région semblent être adoptées, la vidéo intégrale et les plans sont toujours les éléments les plus couramment traités (excepté pour la recherche où de nombreuses méthodes utilisent l'objet comme élément de base). Pourtant, dans le contexte de l'analyse vidéo, les informations sont apportées principalement au travers des objets et de leur évolution temporelle. En exploitant l'environnement des objets, comme le fond ou les objets adjacents, il est également possible d'introduire une certaine sémantique. De la même façon, les relations spatio-temporelles entre les objets peuvent être utilisées pour enrichir le processus de fouille. Si la fouille vidéo orientée-objet semble pertinente, elle pose cependant le problème de l'extraction des objets qui n'est pas possible sans introduction de sémantique. Notons que cette démarche peut également s'appliquer au processus de fouille afin d'exploiter au mieux les objets. Dans cette section, nous expliquons dans quelle mesure les caractéristiques d'un Système de Fouille Vidéo Orienté-Objet (SFV orienté objet) sont différentes d'un SFV n'étant pas centré sur l'objet.

4.1 Caractéristiques d'un SFV orienté objet

Choisir l'objet comme élément d'un SFV a une influence importante sur les autre caractéristiques (sauf pour les objectifs car a priori ils peuvent tous être accomplis en s'appuyant sur l'objet comme élément).

Données

L'impact sur le type de données est relativement faible. Même s'il n'est pas trivial d'extraire des objets depuis un flux vidéo compressé, des solutions existent, citons notamment les travaux de Babu et al. (2004), Toreyin et al. (2005) et Hsu et al. (2006). De plus, l'approche orientée-objet est adaptée à n'importe quel type de vidéos. Mais, intuitivement, il semble plus simple de traiter des vidéos spécifiques puisque la variété d'objets considérés sera plus limitée. Au contraire, utiliser l'objet comme élément rend plus difficile le traitement de vidéos génériques en l'absence de méthode d'extraction adaptée à tout type d'objet.

Échelle

L'approche orientée-objet entraîne un changement d'échelle. En effet, l'utilisation de l'objet comme élément nous amène à considérer deux types d'échelles complémentaires, une échelle pour l'objet et une échelle pour le contexte, tel qu'illustré en figure 4. Les deux trames présentées comportent le même objet, la navette spatiale *Discovery*. Dans le cadre d'une base contenant de nombreuses vidéos de cette navette spatiale, l'utilisateur pourrait vouloir différencier les différentes situations dans laquelle se trouve cette navette (par exemple celles présentées dans les deux vidéos). Décrire seulement l'objet, qui est ici identique dans les deux vidéos, ne permettrait pas de les distinguer. Il est donc nécessaire de pouvoir également décrire l'environnement propre à l'objet afin de pouvoir distinguer la navette sur sa rampe de lancement de la navette en pleine ascension.

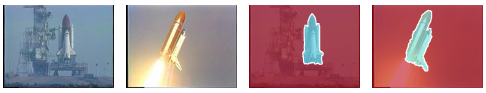

FIG. 4 – *Deux trames extraites de la séquence vidéo STS-53 Launch and Landing, segment 02 of 5 de The Open Video Project (2010) (gauche) et leurs segmentations respectives (droite) de la navette Discovery (bleu) et de son environnement (rouge).*

Descripteurs

Tous les descripteurs peuvent être considérés dans un SFV orienté objet, mais leur usage est différent de celui suivi par les autres types de SFV. En effet, ils peuvent être utilisés pour décrire l'objet et/ou son environnement, et non plus seulement pour décrire l'élément. En outre, les descripteurs de mouvement peuvent être exploités pour décrire le mouvement général de l'objet mais aussi son mouvement interne dans le cas d'objet complexes. De façon plus générale, les descripteurs utilisés dans une approche objet doivent être sémantiquement discriminants, c'est-à-dire que la différence ou la similarité qu'ils mettent en valeur doit avoir une signification sémantique (par exemple la couleur).

Implication de l'utilisateur

Dans un SFV orienté objet, le rôle de l'utilisateur est prédominant de par la sémantique associée au concept d'objet. Un système totalement automatique ne sera pas capable de fouiller sémantiquement les objets, et aura besoin des connaissances de l'utilisateur. De plus, la perception d'un objet est subjective et peut donc être différente d'un utilisateur à l'autre. En effet, chaque utilisateur désire obtenir un résultat personnalisé. L'utilisateur doit donc être particulièrement impliqué dans le processus de fouille afin de pouvoir guider ce dernier. Cependant, même si cette intervention est fondamentale pour le SFV orienté objet, elle doit rester intuitive et légère afin d'être efficace et peu couteuse en temps. Ces propriétés peuvent être assurées au travers de la mise en place d'un retour de pertinence, tel que présenté dans la section 4.3.

4.2 Extraction des objets

Afin d'extraire les objets d'une vidéo, la plupart des méthodes intègrent une étape de segmentation. Seules font exception les méthodes dédiées aux vidéos compressées selon un schéma orientée-objet, voire celles basées sur des points d'intérêt. Pour les SFV, la segmentation vidéo consiste la plupart du temps en un découpage en plans (Lefèvre et al., 2003). Au contraire, pour les SFV orienté objet, l'étape d'extraction doit produire des objets et décrire leur évolution temporelle. Comme nous l'avons souligné dans la section 1, la principale difficulté rencontrée ici est de combler le fossé sémantique séparant les données brutes des objets. Cette extraction peut être effectuée pendant la phase d'encodage de la vidéo dans le cas des données compressés, ou plus généralement avec une segmentation.

Nous représentons une vidéo dans un espace tri-dimensionnel (X,Y,T). une segmentation spatio-temporelle est une partition de cet espace en volumes, chacun représentant un objet spatio-temporel (ou, autrement dit, la définition spatiale de cet objet couplée à son évolution temporelle). Puisqu'un objet est supposé posséder une sémantique, la segmentation nécessite des méthodes intégrant de telles informations. Plus généralement, l'introduction de sémantique est un point important des SFV orienté objet que nous détaillons dans la section suivante.

4.3 Introduire de la sémantique

Un SFV orienté objet nécessite d'introduire de la sémantique dans le processus de segmentation ainsi que dans le processus de fouille. Les descripteurs bas-niveau présentés dans la section 2.2 fournissent des représentations numériques mais ne sont pas capables de donner une perception sémantique de l'objet comme le font les êtres humains. Le fossé sémantique n'est donc toujours pas comblé. Nous pensons néanmoins qu'il puisse l'être en introduisant des connaissances humaines dans un SFV orienté objet.

Pour ce faire, il serait possible de fournir des exemples pour chaque objet potentiellement présent dans une vidéo, mais cette approche ne serait évidemment pas réaliste. Exploiter un mécanisme tel que le retour de pertinence (Ruthven et Lalmas, 2003) semble être une solution plus pertinente. À l'issue du processus de fouille, l'utilisateur évalue un échantillon du résultat qu'il peut également corriger si nécessaire (à l'instar d'un apprentissage par renforcement). En fonction de cette évaluation, le processus peut être relancé pour tenir compte de la connaissance introduite par l'utilisateur (via l'évaluation et la correction de l'échantillon). Ce processus itératif est moins couteux en temps que la production d'exemples complets nécessaire dans

une approche supervisée. Cela garantit également la personnalisation du résultat. De plus, le retour de pertinence peut également être appliqué à l'étape de segmentation (dans le but, ici aussi, de l'améliorer) selon le principe suivant : meilleure sera la segmentation, plus facile la fouille sera. Finalement, créer des descripteurs dédiés aux objets considérés pourrait être également une solution intéressante mais ce problème reste ouvert dans le contexte de vidéos génériques.

4.4 VOMF : Video Object Mining Framework

Nous proposons dans cette section un cadre générique pour les SFV orienté objet, VOMF (Video Object Mining Framework).

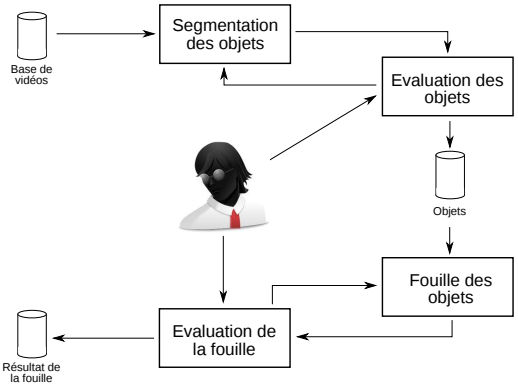

FIG. 5 – *VOMF : Video Object Mining Framework*

Le cadre proposé par VOMF est présenté dans la figure 5 et commence par l'extraction des objets présents dans les vidéos d'une base. Un échantillon des objets obtenus est évalué par l'utilisateur via un système de retour de pertinence. Si les segmentations de l'échantillon sont approuvées par l'utilisateur, l'ensemble des objets est transmis à l'étape de fouille. Dans le cas contraire, les erreurs de segmentation sont identifiées par l'utilisateur, qui introduit de la sémantique par ce biais. Une nouvelle segmentation est alors construite en s'appuyant sur les segmentations courantes et la sémantique apportée par l'utilisateur. Ce cycle est répété jusqu'à ce que l'utilisateur soit satisfait par les objets obtenus. Cependant, il faut veiller à ce que le cycle ne soit pas répété de trop nombreuses fois pour que le temps nécessaire à l'utilisateur soit acceptable. Les résultats de la fouille vidéo sont également évalués par l'utilisateur, via un retour de pertinence sur un échantillon du résultat. Si l'échantillon évalué est satisfaisant, le traitement est terminé. Sinon, à l'instar de la segmentation, l'utilisateur peut corriger l'échantillon. Dans ce cas, la fouille vidéo est relancée et exploite les corrections de l'utilisateur pour améliorer le résultat. L'utilisateur est placé au centre du système. Il supervise le processus de fouille à travers le retour de pertinence et introduit de la sémantique en corrigeant les résultats inappropriés. Pour être efficace, le retour de pertinence ne doit pas être exhaustif. Il faut au contraire que quelques évaluations/corrections suffisent pour influencer profondément les processus de segmentation et de fouille. Ce point est développé dans la prochaine section.

4.5 Un retour de pertinence pour évaluer et guider la fouille vidéo

VOMF est composé de deux étapes, l'extraction des objets et la fouille vidéo. Chacune d'elle dispose de son propre retour de pertinence pour évaluer et guider les processus.

L'extraction des objets est, comme indiqué précédemment, un point critique de VOMF. En fait, sans une extraction d'objets de qualité, il sera particulièrement délicat d'effectuer le processus de fouille. Extraire les objets dans tous les types de vidéos n'est pas une tâche triviale. Le retour de pertinence permet une évaluation directe : le système montre les objets à l'utilisateur et lui demande si ceux-ci correspondent à des objets réels (ou, en d'autres termes, à ceux recherchés par l'utilisateur). Ce retour de pertinence est simple mais très couteux en temps si l'utilisateur doit valider tous les objets extraits. De plus, quel doit être le comportement du système en cas d'insatisfaction de l'utilisateur ? Nous tenons compte de ce problème et proposons la solution suivante. Le système présente un échantillon des objets extraits. L'utilisateur a alors trois possibilités. Il peut valider les objets s'ils représentent ce qu'il recherche. Il peut les corriger. Il peut également les rejeter s'ils ne répondent absolument pas à ses attentes. Les décisions de validation/correction/rejet sont réinjectées dans le système pour guider et améliorer l'extraction des autres objets.

La fouille d'objets est basée sur leurs descriptions mais nécessite également un retour de pertinence. Celui-ci consiste à présenter un échantillon des résultats à l'utilisateur. Par exemple, si l'objectif est la classification, le système présente quelques objets et leur classification. Pour corriger ces objets, l'utilisateur doit changer la classe à laquelle ils appartiennent. Ainsi, l'utilisateur guide le processus de fouille et, à l'itération suivante, cette information est utilisée pour améliorer le résultat.

5 Conclusion

Les systèmes de fouille vidéo (SFV) récents s'appuient sur une description des vidéos réalisée à l'échelle des objets ou des régions, mais sont appliqués sur des éléments tels que les plans ou les vidéos intégrales. Dans cet article, nous avons introduit une nouvelle taxonomie pour caractériser les SFV et l'avons utilisée pour étudier et comparer les SFV actuels. Nous avons également montré que l'objet devrait être l'élément à considérer par les SFV, et nous avons justifié notre proposition en présentant les avantages des systèmes de fouille vidéo orienté objet. Nous avons discuté les répercussions du choix de l'objet comme élément sur les autres caractéristiques définies dans notre taxonomie. L'importance de la segmentation a été soulignée, et nous avons suggéré comment pouvait être introduites des informations de nature sémantique dans les SFV orienté objet. Enfin, nous avons proposé VOMF, un cadre générique pour la fouille vidéo orientée-objet. VOMF offre de nouvelles perspectives, la fouille vidéo étant plus pertinente si les objets considérés sont les objets réels (du point de vue de l'utilisateur) présents dans les vidéos.

Nos futurs travaux incluent l'utilisation de VOMF pour construire des SFV orienté objet pour différents objectifs. Dans cette optique, nous travaillons actuellement sur le problème du clustering des objets. Le but de ces travaux est d'obtenir des groupes d'objets similaires depuis une base de vidéo. Nous disposons actuellement de prototypes pour la fouille guidée par l'util-

isateur d'une part, et pour l'amélioration de segmentation vidéo guidée par l'utilisateur d'autre part. Ces prototypes ont donné des premiers résultats prometteurs mais ne sont aujourd'hui pas totalement aboutis, et nécessitent encore certains travaux de recherche.

Remerciements

Ce travail a été soutenu par Ready Business System et l'Association Nationale de la Recherche et de la Technologie (ANRT). Nous remercions particulièrement Christian Dhinaut de RBS pour sa contribution.

Références

Anjulan, A. et N. Canagarajah (2007a). A novel video mining system. In *14th IEEE International Conference on Image Processing*, pp. 185–188. IEEE.

Anjulan, A. et N. Canagarajah (2007b). Object based video retrieval with local region tracking. *Signal Processing : Image Communication 22*(7-8), 607–621.

Babu, R., K. Ramakrishnan, et S. Srinivasan (2004). Video object segmentation : A compressed domain approach. *IEEE Transactions on Circuits and Systems for Video Technology 14*(4), 462–474.

Basharat, A., Y. Zhai, et M. Shah (2008). Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding 110*(3), 360–377.

Brezeale, D. et D. Cook (2008). Automatic video classification : A survey of the literature. *IEEE Transactions on Systems, Man and Cybernetics-part C : Applications and Reviews 38*(3), 416–430.

Brunelli, R., O. Mich, et C. Modena (1999). A survey on automatic indexing of video data. *J. of Visual Communication and Representation 10*(2), 78–112.

Chevalier, F., J.-P. Domenger, J. Benois-Pineau, et M. Delest (2007). Retrieval of objects in video by similarity based on graph matching,. *Pattern Recognition Letters 28*(8), 939–949.

Cios, K., W. Pedrycz, R. Swiniarski, et L. Kurgan (2007). *Data Mining A Knowledge Discovery Approach*. Springer.

de Avila, S., A. da Luz, et A. de Araujo (2008). Vsumm : A simple and efficient approach for automatic video summarization. In *15th International Conference on Systems, Signals and Image Processing*, pp. 449–452.

Gao, X., X. Li, J. Feng, et D. Tao (2009). Shot-based video retrieval with optical flow tensor and HMMs. *Pattern Recognition Letters 30*(2), 140–147.

Hsu, C.-C., H. Chang, et T.-C. Chang (2006). Efficient moving object extraction in compressed low-bit-rate video. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, Washington, DC, USA, pp. 411–414. IEEE Comp. Soc.

Idris, F. et S. Panchanathan (1997). Review of Image and Video Indexing Techniques. *Journal of Visual Communication and Image Representation 8*(2), 146–166.

Koprinska, I. et S. Carrato (2001). Temporal video segmentation : A survey. *Signal Processing : Image Communication 16*(5), 477–500.

Lefèvre, S., J. Holler, et N. Vincent (2003). A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging 9*(1), 73–98.

Liu, D. et T. Chen (2009). Video retrieval based on object discovery. *Computer Vision and Image Understanding 113*(3), 397–404.

Money, A. et H. Agius (2008). Video summarisation : A conceptual framework and survey of the state of the art. *J. of Visual Communication and Image Representation 19*(2), 121–143.

Ren, W., S. Singh, M. Singh, et Y. Zhu (2009). State-of-the-art on spatio-temporal information based video retrieval. *Pattern Recognition 42*(2), 267–282.

Ren, W. et Y. Zhu (2008). A video summarization approach based on machine learning. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Los Alamitos, CA, USA, pp. 450–453. IEEE Comp. Soc.

Rosenfeld, A., D. Doermann, et D. DeMenthon (Eds.) (2002). *Video Mining*. Springer.

Rui, Y., T. Huang, et S. Chang (1999). Image retrieval : current techniques, promising directions, and open issues. *J. of Visual Communication and Image Representation 10*(4), 39–62.

Ruthven, I. et M. Lalmas (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review 18*(2), 95–145.

Sivic, J. et A. Zisserman (2008). Efficient visual search for objects in videos. *Proceedings of the IEEE 96*(4), 548–566.

Snoek, C. G. M. et M. Worring (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval 4*(2), 215–322.

Teixeira, L. F. et L. Corte-Real (2009). Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters 30*(2), 157–167.

The Open Video Project (2010). http ://www.open-video.org/.

Toreyin, B., A. Cetin, A. Aksay, et M. Akhan (2005). Moving object detection in wavelet compressed video. *Signal Processing : Image Communication 20*(3), 255–264.

Zhai, S., B. Luo, J. Tang, et C.-Y. Zhang (2007). Video abstraction based on relational graphs. In *Proc. of the Fourth Int. Conf. on Image and Graphics*, pp. 827–832. IEEE Comp. Soc.

Summary

Today, video is becoming one of the primary sources of information. Current video mining systems face the problem of the semantic gap (i.e., the difference between the semantic meaning of video contents and the digital information encoded within the video files). This gap can be bridged by relying on the real objects present in videos because of their semantic meaning. But video object mining needs some semantics, both in the object extraction and in the object mining steps. We think that the introduction of semantics during these steps can be ensured by user interaction. We then propose a generic framework to deal with video object mining.