Proposition de projet - Pôle MSTIC

Apprentissage Parallèle pour l'Indexation Multimédia Sémantique

APIMS

Responsable scientifique : Georges Quénot, LIG/MRIM.

1. Introduction

La quantité de documents image et vidéo numériques croît de manière exponentielle depuis de nombreuses années et cette tendance devrait se poursuivre encore longtemps grâce aux progrès technologiques dans ce domaine. L'indexation par concepts des documents image et vidéo est une nécessité pour gérer de manière efficace les masses de données correspondantes. En effet, les mots-clés nécessaires pour la recherche par le contenu n'y sont pas explicitement présents comme dans le cas des documents textuels. La recherche à partir d'exemples ou à partir de caractéristiques dites « de bas niveau » présente également de sérieuses limitations : les exemples nécessaires ne sont généralement pas disponibles et les caractéristiques de bas niveau ne sont pas aisément manipulables et interprétables par un utilisateur. Par ailleurs, une similarité au niveau de ces caractéristiques ne correspond pas forcément à une similarité au niveau sémantique. L'indexation par concepts est un grand challenge en raison du « fossé sémantique » séparant le contenu brut de ces documents (pixels, échantillons audio) et les concepts qui on un sens pour un utilisateur.

Des progrès importants ont été accomplis ces dernières années, notamment dans le cadre des campagnes d'évaluation TRECVID [1]. Ces campagnes annuelles organisées par le National Institute of Standards and Technologies (NIST) américain fournissent des données en quantité importante, des tâches bien définies, des « vérités terrain », des métriques et des outils d'évaluation associés. Elles contribuent largement à fédérer les recherches dans le domaine de l'indexation et de la recherche par le contenu des documents vidéo.

Les méthodes fonctionnant le mieux actuellement sont des méthodes statistiques fonctionnant par apprentissage supervisé à partir d'exemples annotés manuellement. Des caractéristiques dites de bas niveau sont extraites à partir du signal audio ou image brut (des histogrammes de couleur ou des transformées de Gabor par exemple) et sont ensuite envoyées à des classifieurs qui sont entraînés à partir d'exemples positifs et négatifs des concepts à reconnaître. Pour obtenir de bons résultats, il est nécessaire de multiplier les caractéristiques utilisées et de les combiner en utilisant des techniques de fusion appropriées. Un gain supplémentaire est obtenu en utilisant les relations entre les concepts comme les relations statistiques (cooccurrences) ou logiques (générique-spécifique par exemple).

Les principes généraux étant les mêmes, les différences entre les approches concernent les choix sur les caractéristiques, sur les outils de classification et/ou de fusion, et sur la façon de prendre en compte le contexte. La qualité et la quantité des exemples positifs et négatifs utilisés fait également une différence importante. L'état de l'art actuel est l'extraction conjointe de plusieurs centaines de concepts définis dans l'ontologie LSCOM [2]. Cependant, malgré les efforts très importants fournis par un grand nombre d'équipes (plus de 30 équipes ont participé à la tâche d'extraction de concepts dans les plans vidéo lors de la campagne

TRECVID 2007), la précision moyenne des meilleurs systèmes ne dépasse pas 20% (13% seulement pour la campagne 2007).

L'équipe MRIM du LIG a développé des méthodes et des outils pour l'extraction automatique de concepts dans les plans vidéo et a obtenu des résultats un peu supérieurs à la moyenne dans les campagnes TRECVID 2005 à 2007 [3]. L'objectif de ce projet est d'améliorer de manière importante ces méthodes et de leur faire rejoindre voire définir l'état de l'art dans le domaine. Pour cela, il faut d'une part les optimiser en prenant en compte tous les facteurs importants et de leur ajouter un certain nombre d'innovations comme l'utilisation de concepts de niveau intermédiaire, la combinaison de méthodes génériques et spécifiques, et l'apprentissage actif pour l'amélioration de la quantité et qualité de l'annotation servant à l'entraînement des systèmes.

Un des facteurs limitants est la puissance de calcul nécessaire. Il faut en effet entraîner et évaluer les systèmes sur plusieurs centaines de concepts et sur plusieurs dizaines de milliers d'images ou de plans vidéo. Il faut en outre faire cela en étudiant de multiples combinaisons de caractéristiques de bas et moyen niveau, de méthodes de classification et de méthodes de fusion. Nous envisageons pour cela d'utiliser les ressources du projet GRID 5000 [4] afin de pouvoir étudier à grande échelle l'influence combinée de ces différents facteurs. Dans sa version simple, le problème se parallélise assez facilement (on peut faire faire l'apprentissage et l'évaluation d'un concept sur un processeur) mais lorsqu'on veut utiliser le contexte, c'est-à-dire les relations statistiques ou ontologiques des concepts entre eux, il y a lieu de faire coopérer les différents processus entre eux et cela devient un réel problème de programmation parallèle. L'équipe MESCAL du LIG dispose d'une grande expertise dans ce domaine et participera à l'étude et à la mise en œuvre des versions parallèles des méthodes d'extraction de concepts.

L'utilisation de la multimodalité naturellement présente dans les documents vidéo est également essentielle pour la performance des systèmes d'indexation par concepts. L'équipe GETALP du LIG dispose de compétences dans le domaine du traitement du signal audio et de parole et participera à la définition et à l'optimisation des caractéristiques de bas et moyen niveau pour l'indexation des concepts à partir de la piste audio. De même, l'équipe GPIG de GIPSA-Lab dispose de compétences dans l'analyse et l'indexation du mouvement dans les documents vidéo et participera à la définition et à l'optimisation des caractéristiques de bas et moyen niveau pour l'indexation des concepts à partir du mouvement dans la piste image.

L'équipe MRIM assurera la direction du projet et l'intégration des contributions des autres partenaires. Elle s'occupera de l'étude et de la mise en œuvre des méthodes d'optimisation globale des différents facteurs. Enfin, elle travaillera sur de nouvelles extensions comme la combinaison d'approches génériques et spécifiques.

Références

[1] Smeaton, A. F., Over, P., and Kraaij, W. TRECVID: evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the 12th Annual ACM international Conference on Multimedia*, New York, NY, USA, October 10-16, 2004.

[2] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann and J. Curtis, Large-Scale Concept Ontology for Multimedia, *IEEE Multimedia* **13**(3), pp. 86-91, 2006.

[3] Stéphane Ayache, Georges Quénot and Jérôme Gensel, CLIPS-LSR Experiments at TRECVID 2006, *TRECVID'2006 Workshop*, Gaithersburg, MD, USA, November 13-14, 2006.

[4] Bolze, R. et al, Grid'5000: a large scale and highly reconfigurable experimental Grid testbed *International Journal of High Performance Computing Applications*, 20(4), pp 481-494, 2006.

2. Programme de travail

2.1 Contenu

La première partie du travail consistera à mettre en œuvre des versions parallèles des méthodes de classification développées dans l'équipe MRIM et à utiliser ces versions parallèles pour optimiser conjointement les différents éléments (jeux de caractéristiques, opérateurs de classification et opérateurs de fusion) intervenant dans celles-ci. Cette optimisation devra être faire de manière aussi systématique que possible. Compte tenu de l'aspect hautement combinatoire et du coût de calcul (même sur une architecture parallèle) de celle-ci, des méthodes heuristiques appropriées devront être étudiées et mises en œuvre afin d'obtenir le meilleur résultat dans un temps donné.

Parallèlement, nous étudierons des caractéristiques supplémentaires (descripteurs) en vue de leur intégration :

- Le mouvement est encore peu exploité dans notre approche (nous n'utilisons pour l'instant que des statistiques globales). Nous proposons d'adapter et de généraliser un algorithme d'analyse de saut d'athlétisme développé dans l'équipe GPIG qui repose sur la reconnaissance de séquences d'états. Cette version étendue devrait permettre d'extraire des plans vidéo différents évènements constitués de plusieurs actions élémentaires ou des mouvements répétitifs comme la marche ou la course.
- Au niveau de l'audio, nous prévoyons d'adapter des méthodes existantes dans l'équipe GETALP autour de la reconnaissance automatique de la parole multilingue, pour l'indexation de contenus vidéos. L'incertitude sur la langue parlée dans les documents à traiter peut être abordée en appliquant, d'une part, des multiples reconnaisseurs monolingues en parallèle (ce qui est envisageable sur une plateforme de calcul parallèle) et d'autre part, en utilisant un reconnaisseur multilingue unique, incluant des modèles acoustiques proposant une large couverture des langues du monde. Ces deux approches seront évaluées dans le projet.

Dans une deuxième partie, nous mettrons en œuvre des approches intégrées pour la reconnaissance simultanée de plusieurs centaines de concepts en prenant en compte dès les premiers niveaux de l'apprentissage les corrélations existant entre ceux-ci. Nous intègrerons également les nouvelles caractéristiques de bas et moyen niveaux développées dans la première partie.

Ces travaux seront, dans la mesure du possible, planifiés en fonction des évaluations TRECVID sur la détection de concepts dans les plans vidéo. Les expérimentations on lieu en général pendant l'été (juillet-août) et les campagnes s'étendent de février à novembre de l'année en cours. L'objectif est de pouvoir évaluer lors des campagnes 2008 et 2009 ce qu'il est prévu de développer dans la première et la deuxième partie décrites ci-dessus. Notre travail et nos expérimentations ne se réduiront toutefois pas à une participation aux campagnes TRECVID.

Une part importante du travail de parallélisation et d'optimisation sera effectuée par un post-doc recruté pour ce projet. Ce post-doc sera accueilli dans l'équipe MRIM et il travaillera également en étroite relation avec les équipes MESCAL, GPIG et GETALP.

Nous créerons un site web dès le début du projet et nous en assurerons régulièrement la mise à jour pour assurer la visibilité des projets labellisés et aider à assurer celle du pôle et donc de l'établissement. Ce site web servira aussi pour l'organisation du travail entre les partenaires.

2.2 Calendrier (T0 = 1er janvier 2009)

T0 à T0+6 : Mise en œuvre d'une version parallèle, première optimisation des caractéristiques, des opérateurs de classification et des opérateurs de fusion (MRIM, MESCAL).

T0+6 à T0+9 : Expérimentations TRECVID 2009 (MRIM, MESCAL).

T0+9 à T0+12 : Analyse des résultats et optimisations complémentaire (MRIM, MESCAL).

T0 à T0 +12 : Étude de caractéristiques additionnelles pour le mouvement (GPIG) et l'audio (GETALP).

T0+12 : Rapport d'activité intermédiaire (tous).

T0+12 à T0+18 : Étude et intégration de méthodes d'apprentissage prenant directement en compte les corrélations (relations statistiques) entre concepts (MRIM, MESCAL). Intégration des caractéristiques additionnelles (MRIM, GPIG, GETALP).

T0+18 à T0+21 : Expérimentations TRECVID 2010 (tous).

T0+21 à T0+24 : Analyse des résultats et optimisations complémentaire (tous).

T0+24 : Rapport d'activité final (tous).

2.3 Budget prévisionnel

12 mois de post-doc : 45 k€

Fonctionnement (missions, vacations, petit matériel) : 20 k €, répartition par année et par équipe ;

Équipe	2009	2010
LIG/MRIM	3 k€	2 k€
LIG/GETALP	3 k€	2 k€
LIG/MESCAL	3 k€	2 k€
GIPSA/GPIG	3 k€	2 k€

3. Mise en perspective

Au moins une trentaine d'équipes travaillent dans le monde sur ce sujet et évaluent leurs approches dans le cadre des campagnes TRECVID. Au niveau mondial, les meilleures équipes sont : Mediamill (Hollande) [5] en Europe, CMU et IBM [6] en Amérique du nord et l'université de Tsinghua [7], l'université de sciences et technologie de Chine, et Microsoft Asie [8] en Asie. Ces équipes disposent de gros moyens et en particulier d'architectures parallèles pour l'entraînement de leurs systèmes [9]. Elles utilisent des approches systématiques pour identifier les meilleures combinaisons de caractéristiques et de classifieurs. Elles mettent également en œuvre des solutions algorithmiques sophistiquées comme l'annotation simultanée de nombreux concepts en tenant compte des corrélations existant entre eux dès le premier stade de la classification [8] ou l'utilisation du contexte spatial [7].

En France, plusieurs équipes travaillent sur ce thème et participent également aux campagnes d'évaluation TRECVID : le LABRI à Bordeaux, l'IRIT à Toulouse, Eurecom à Sophia-Antipolis, le LIP6 à Paris, et les équipe LIG/MRIM et GIPSA/GPIG à Grenoble. Bien que nous ayons obtenu de bons résultats et eu des participations honorables à TRECVID [3], nous avons encore un certain retard par rapport aux équipes internationales précédemment citées. Notre objectif est donc de combler ce retard et de rejoindre le « groupe de tête » formé par ces équipes. Une coopération renforcée au niveau Grenoblois mais aussi au niveau national, notamment dans le cadre d'une action au niveau du GDR ISIS (action IRIM), se met en place dans ce but. Le présent projet irait dans ce sens en nous permettant de recruter un post-doc pour travailler sur la parallélisation de nos méthodes d'indexation, sur l'exploration systématique des combinaisons de caractéristiques et de classifieurs, et sur l'amélioration de nos méthodes.

Au niveau local, ce projet correspond à un des thèmes affichés par PILSI (Loisir et Multimédia). Les équipes MRIM et GETALP viennent par ailleurs de collaborer sur la compétition de recherche d'information selon le contenu « Star Challenge » (http://www.thestarchallenge.sg), et font partie des 5 équipes (sur 56 équipes de 17 pays au départ) qualifiées pour la finale qui aura lieu le 23 Octobre à Singapour.

Références

[5] Cees G.M. et al., Adding Semantics to Detectors for Video Retrieval, in *IEEE Transactions on Multimedia*, 9(5):975-986, August 2007.

[6] Jun Yang, Rong Yan and Alexander G. Hauptmann, Cross-Domain Video Concept Detection Using Adaptive SVMs, in *ACM Multimedia*, pages 188-197, Augsburg, 24-29 Sep. 2007.

[7] Jinhui Yuan, Jianmin Li and Bo Zhang, Exploiting Spatial Context Constraints for Automatic Image Region Annotation, in *ACM Multimedia*, pages 595-604, Augsburg, 24-29 Sep. 2007.

[8] Guo-Jun Qi et al. Correlative Multi-Label Video Annotation, in *ACM Multimedia*, pages 17-26, Augsburg, 24-29 Sep. 2007.

[9] Frank J. Seinstra et al., High-Performance Distributed Image and Video Content Analysis with Parallel-Horus *IEEE Multimedia*, 14(4):64-75. October-December 2007.

4. Partenaires du projet

Le projet regroupe quatre équipes de deux laboratoires :

Laboratoire d'Informatique de Grenoble (LIG, UMR 5217) :

- Équipe Modélisation et Recherche d'Information Multimédia (LIG), chef de file,
- Équipe Groupe d'Etude en Traduction/Traitement des Langues et de la Parole (GETALP),
- Équipe Middleware Efficiently SCALable (MESCAL).

Laboratoire Grenoblois de l'Image, de la Parole, du Signal et de l'Automatique (GIPSA-Lab, UMR 5216) :

- Équipe Géométrie, Perception, Images, Gestes (GPIG).

Participants :

Nom	Prénom	Fonction	Organisme	Laboratoire	Équipe
Quénot	Georges	Chargé de Recherche	CNRS	LIG	MRIM
Mulhem	Philippe	Chargé de Recherche	CNRS	LIG	MRIM
Besacier	Laurent	Maître de Conférence	UJF	LIG	GETALP
Richard	Olivier	Maître de Conférence	UJF	LIG	MESCAL
Rombaut	Michèle	Professeur	UJF	GIPSA	GPIG
Pellerin	Denis	Professeur	UJF	GIPSA	GPIG

L'équipe MRIM apporte sa compétence en indexation par concepts des documents image et vidéo et assure la direction du projet.

L'équipe GETALP apporte sa compétence en traitement de l'audio.

L'équipe MESCAL apporte sa compétence en programmation parallèle.

L'équipe GPIG apporte sa compétence en analyse du mouvement dans les documents vidéo.