Thèse DGA – Aix Marseille Université

Titre du sujet de thèse :	Compréhension Multimodale – vers des traitements joints audio/image pour la compréhension multimodale de documents vidéo
Domaine :	Ingénierie de l'Information et Robotique
Thématique:	Traitement de l'information complexe
Mots clés :	*traitement automatique de la parole, traitement automatique de la langue, traitement d'image, apprentissage automatique, recherche d'information*
Laboratoire:	Laboratoire d'Informatique Fondamentale de Marseille Aix Marseille Université - LIF/CNRS (UMR 7279)
Contact :	Frédéric Béchet (frederic.bechet@univ-amu.fr)

Description

La multiplication des documents vidéo accessibles sur Internet a rendu nécessaire le développement d'outils permettant d'effectuer des requêtes complexes sur le contenu multimodal de ces documents. Les communautés scientifiques du traitement de l'image et du traitement du signal audio ont chacune proposé de nombreux descripteurs afin de caractériser ces documents multimédia, mais chacune dans leurs modalités : identification de locuteurs, transcription de parole pour l'audio ; détection de « concepts visuels », classification d'image, reconnaissance d'objets ou de scènes pour la vidéo.

Récemment des campagnes d'évaluation telles que le *Défi-Repère*[1] (2011-2014), co-financés par la DGA et l'ANR et dont l'équipe TALEP du LIF a coordonné l'un des consortiums participants, se sont intéressés à la multimodalité à travers le problème de la détection de *présence* de personnes dans plusieurs modalités à l'intérieur d'un document vidéo : *Qui voit-on ? Qui parle ? Quel nom de personne est écrit à l'écran ? Quel nom de personne est prononcé ?*

Les systèmes développés qui ont concouru dans cette évaluation comportent tous à la fois des modules monomodaux tels que des briques d'identification de locuteurs ou de reconnaissance faciale et des modules de fusion multimodale agrégeant les différents descripteurs obtenus [Bredin12 ; Favre13]. Les résultats obtenus [Galibert13] lors des différentes phases du *Défi-Repère* ont souligné d'une part la force des briques monomodales, permettant d'obtenir de bons résultats quelle que soit la modalité dans les cas « faciles » (par exemple : la personne la plus visible à l'écran est aussi celle qui s'exprime ; un « cartouche » texte est incrusté dans l'image sous une personne et l'identifie) ; et d'autre part la nécessité de développer d'autres processus pour les cas plus complexes.

Le sujet de thèse proposé s'inscrit dans ce cadre en ayant pour but de prolonger les études sur les modèles d'analyse multimodaux qui ont été réalisés par l'équipe lors de notre participation au Défi-Repère. Le but est de concevoir des modèles de Compréhension Multimodale de vidéos permettant d'analyser de manière jointe à la fois le flux audio à travers les étapes de segmentation de signal et de transcription/compréhension de la parole ; et

[1] http://www.defi-repere.fr

d'autre part le flux d'image par rapport à l'analyse fondée sur des *grammaires visuelles* caractérisant les situations à l'écran.

En utilisant comme point de départ les corpus et les annotations développés lors du Défi-Repère, cette thèse vise à dépasser le cadre de la détection et de l'identification de personnes en améliorant les performances des applications d'indexation et de recherche d'information multimédia, non seulement sur des corpus provenant d'émissions télévisées, mais plus généralement sur tout document vidéo nécessitant une analyse de scène multimodale.

Contexte de la thèse

Cette thèse sera financée par la DGA (*Direction Générale de l'Armement*) et l'Université d'Aix Marseille (*AMU*). Elle sera effectuée au sein de l'équipe *Traitement Automatique de la Langue Ecrite et Parlée* (TALEP) du *Laboratoire d'Informatique Fondamentale de Marseille* (LIF-CNRS, UMR 7279) de l'AMU. Une des spécialités de cette équipe est l'analyse linguistique robuste sur des corpus de langue non-canonique tels que des transcriptions de parole spontanée ou des textes provenant de médias sociaux.

L'équipe TALEP est actuellement impliquée dans de nombreux projets collaboratifs (ANR, Européen, internationaux) ayant un lien avec le sujet proposé. En particulier, l'équipe coordonne le consortium *PERCOL* regroupant l'Université d'Aix Marseille, l'Université d'Avignon, l'Université de Lille ainsi que le laboratoire Orange Labs de Lannion afin de participer à la campagne d'évaluation *Défi-Repère* co-organisé par la **DGA** et l'ANR (2011-2014). L'équipe TALEP a organisé à Marseille en 2013 le premier workshop ***SLAM*** : *Speech Language and Audio in Multimedia* http://slam2013.lif.univ-mrs.fr . Ce premier workshop, entièrement dédié au traitement de données multimédia, a été l'occasion de présenter les travaux réalisés dans le cadre du Défi-Repère.

Cette proposition de thèse sera encadrée par Frédéric Béchet, Professeur des Universités à l'Université d'Aix Marseille depuis 2009, coordonnateur du projet ANR PERCOL (Défi-Repère DGA-ANR). Elle sera co-encadrée par Benoît Favre, Maître de conférences à l'Université d'Aix Marseille, spécialisé dans les méthodes de traitement automatique de données audio et vidéo.

Bibliographie récente de l'équipe en lien avec la thèse

- *Sur le traitement de données multimédia dans le cadre du Défi Repère :*

[Favre13] Benoit Favre, Géraldine Damnati, Frederic Bechet, Meriem Bendris, Delphine Charlet, Rémi Auguste, Stéphane Ayache, Benjamin Bigot, Alexandre Delteil, Richard Dufour, Corinne Fredouille, Georges Linarès, Jean Martinet, Gregory Senay, Pierre Tirilly, "*PERCOLI: a person identification system for the 2013 REPERE challenge*", SLAM, Marseille (France) – 2013

[Bendris13] Meriem Bendris, Benoit Favre, Delphine Charlet, Geraldine Damnati, "*Unsupervised Face Identification in TV Content using Audio-Visual Sources*", CBMI, Veszprém (Hungary) – 2013

- *Sur le traitement et la segmentation de données vidéos :*

[Rouvier13] Mickael Rouvier, Georges Linares, Bernard Merialdo, Benoit Favre, "*Searching Segments of Interest in Single Story Web-Videos*", to appear in WIAMIS, Paris (France) – 2013

- *Sur l'analyse de la parole conversationnelle dans un cadre multi-vue/multimedia:*

[Koco12] Sokol Koço, Cécile Capponi, Frédéric Béchet « *Applying multiview learning algorithms to human-human conversation classification* »13th Annual Conference of the International Speech Communication Association ISCA Interspeech 2012, Portland, USA

[Bechet12] Frédéric Béchet, Benoit Favre, Géraldine Damnati, "*Detecting Person Presence in TV Shows with Linguistic and Structural Features*", IEEE ICASSP'12 - 2012

Bibliographie générale

[Bendris09] *Introduction of quality measures in audio-visual identity verification,* Meriem Bendris, Delphine Charlet and Gérard Chollet, actes de la conférence IEEE ICASSP'09, 2009.

[Bendris10] *Talking faces indexing in TV-content*, Meriem Bendris, Delphine Charlet and Gérard Chollet, actes de la conférence CBMI 2010.

[Bredin12] *Fusion of Speech, Faces and Text for Person Identification in TV Broadcast.* Hervé Bredin, Johann Poignant, Makarand Tapaswi, Guillaume Fortier, Viet Bac Le, Thibault Napoleon, Hua Gao, Claude Barras, Sophie Rosset, Laurent Besacier, Jakob Verbeek, Georges Quénot, Frédéric Jurie, Hazim Kemal Ekenel. ECCV 2012.

[Bojanowski13] *Finding Actors and Actions in Movies*. Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, Josef Sivic. ICCV 2013 - IEEE International Conference on Computer Vision.

[Damnati11] *Robust speaker turn role labeling of TV Broadcast News shows*, Géraldine Damnati, Delphine Charlet , actes de la conference IEEE ICASSP 2011: 5684-5687

[Dumont12] *Automatic Story Segmentation for TV News Video Using Multiple Modalities*, Emilie Dumont, Georges Quénot, Int. J. Digital Multimedia Broadcasting 2012 (2012)

[Gaidon13] *Temporal Localization of Actions with Actoms.* Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2013, 35 (11), pp. 2782-2795

[Galibert13] *The First Official REPERE Evaluation*, Olivier Galibert, Juliette Kahn, Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

[Guinaudeau12] *Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation*, Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Computer Speech & Language 26(2): 90-104 (2012)

[Wang12] *Broadcast News Story Segmentation Using Conditional Random Fields and Multimodal Features*, Xiaoxuan Wang, Lei Xie, Mimi Lu, Bin Ma, Engsiong Chng, Haizhou Li, IEICE Transactions 95-D(5): 1206-1215 (2012)