Recherche efficace de motifs spatio-temporels dans des grands cubes de données satellites

Proposition de thèse CIFRE (démarrage prévu fin 2023)
Discipline : Informatique, Traitement du Signal et des Images
Mots-clefs : Traitement d'image, télédétection, séries temporelles, morphologie mathématique, apprentissage profond

Environnement

Entreprise : CLS `https://www.cls.fr/`

CLS (Collecte Localisation Satellites), filiale du CNES (Centre National d'Études Spatiales) et de la CNP (Compagnie Nationale à Portefeuille), est une société internationale spécialisée dans la fourniture de solutions d'observation et de surveillance de la Terre depuis 1986.

CLS est une entreprise française leader dans plusieurs domaines à l'international comme au niveau national, développant des solutions innovantes fondées sur l'observation de la Terre et des Océans. Depuis 2021, CLS est devenue une société à mission, en inscrivant dans ses statuts le fait de se dépasser au quotidien dans l'innovation, le développement de solutions pour une Planète durable. Dans ce cadre, elle développe et favorise des solutions efficientes pour répondre aux problèmes de recherche centrés sur l'Observation de la Terre.

Un accompagnement de la part de l'unité de R&D du pôle Terre et Hydrologie du groupe CLS sera assuré lors de la thèse. L'unité est située à Villeneuve d'Ascq, près de Lille (59), France. La thèse sera effectuée en relation étroite avec le projet Horizon Europe Evoland et ses membres (CNES, DLR, Cesbio entre autres).

CLS mettra à disposition son infrastructure de calcul pourvue de clusters CPU et GPU et d'un gros volume de stockage permettant l'analyse de très gros volumes de données (de l'ordre du peta-octet) en conditions de production dans le cadre de la thèse.

Laboratoire : IRISA, équipe OBELIX `https://www.irisa.fr/fr/equipes/obelix`

L'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires, `www.irisa.fr`), est une unité mixte de recherche (UMR 6074) en informatique, en traitement du signal et des images et en robotique. Sur la conjonction de ces thématiques, l'IRISA se positionne comme le grand laboratoire de recherche en Bretagne avec une présence affirmée sur les campus de Rennes (35), Vannes (56), et Lannion (22). Elle s'organise autour de 7 départements, dont celui "Signal, Image, Langage" dans lequel s'inscrivent les travaux de l'équipe OBELIX.

L'équipe OBELIX (Observation de l'environnement par imagerie complexe) mène ses recherches dans le domaine de l'intelligence artificielle (apprentissage automatique et vision par ordinateur) appliquée à l'observation de la Terre. Elle conçoit ainsi des solutions informatiques originales, efficientes et robustes pour faire face à la complexité inhérente aux données environnementales. En particulier, l'équipe a développé ces dernières années une expertise méthodologique dans les domaines de l'apprentissage profond, du transport optimal, et des hiérarchies morphologiques.

Cette équipe est principalement localisée à Vannes (56), et dispose de moyens de calcul (clusters CPU et GPU) qui pourront être mobilisés pendant la thèse. Elle porte également le parcours de Master 2 GeoData Science du Master Erasmus Mundus "Copernicus Master in Digital Earth".

Contexte industriel

Depuis 2011, le Copernicus Land Monitoring Service (CLMS `https://land.copernicus.eu/`) fournit des produits pour la surveillance de l'état, des changements et des caractéristiques de la couverture/utilisation des terres végétalisées, non végétalisées, des variables biophysiques, des conditions de l'eau et de la cryosphère. Cette cartographie à large échelle utilisent les données d'observation à haute résolution de la terre (10 m de résolution spatiale, une acquisition toutes les semaines) gratuites et libres d'accès telles que Sentinel (1 et 2), ainsi que des données commerciales à très haute résolution spatiale (1 m de résolution spatiale, tous les ans).

Dans ce contexte, il est important de développer des algorithmes, des méthodes et procédés semi-automatiques (voire automatiques) afin de limiter au strict nécessaire le recours à des traitements humains au regard des masses de données manipulées. Des interventions humaines, même de courte durée, répétées à l'échelle de l'Europe voire du globe ont un effet majeur sur la capacité à fournit les produits Copernicus en un

temps raisonnable. De plus, il est nécessaire d'optimiser les traitements informatiques pour réduire leur coût financier et énergétique. Dans ce contexte, l'unité de R&D du pôle Terre et Eau du groupe CLS cherche à concevoir des solutions (semi-)automatiques efficaces pour analyser de grands volumes de données d'observation de la Terre.

La recherche d'automatisation a donné lieu à une collaboration entre CLS et OBELIX depuis plusieurs années, afin de concevoir et déployer des solutions efficientes de cartographie automatisée large-échelle.

Contexte scientifique

Ainsi, dans le cadre d'une demande de l'Agence Européenne de l'Environnement, CLS et OBELIX ont conçu et déployé une chaîne de production originale de la cartographie des trames vertes à l'échelle continentale pour le compte du programme Copernicus. Pour faire face au volume de données à traiter (38 000 images, soit 120 To), et à la diversité des scènes étudiées, la solution développée s'est appuyée sur des algorithmes efficaces de caractérisation multi-échelle des pixels (profils d'attributs) à l'aide des hiérarchies morphologiques, et de classification semi-supervisée par une approche ensembliste de forêts aléatoires [3]. Elle a été implantée à l'aide de composants logiciels C++ diffusés sous licence libre : TRISKELE [1] et Broceliande [2]. Une attention particulière est portée à l'optimisation systématique de toutes les étapes du processus, y compris l'extraction des descripteurs [4]. Cette étape, centrale dans le processus de cartographie automatique, est souvent mise en oeuvre à l'aide des profils d'attributs calculés efficacement à l'aide des hiérarchies morphologiques [6].

Dans un autre contexte, l'équipe OBELIX a collaboré avec le CNES dans le cadre d'une étude R&T pour développer une solution efficace de recherche automatique par l'exemple dans des bases d'images satellites. Pour cela, elle a exploité les hiérarchies morphologiques pour calculer des histogrammes de formes (ou Pattern Spectra) qui permettent de mettre en oeuvre des algorithmes efficaces de recherche par l'exemple. La solution ainsi développée rend possible la recherche de motifs spatiaux de taille variable (et non connue a priori) dans une base de très grandes images [5]. Contrairement aux approches populaires en vision par ordinateur basées sur l'apprentissage automatique ou profond, elle ne s'appuie pas sur un entraînement préalable d'un modèle prédictif, et fonctionne sans recourir à des données annotées. Ce travail a abouti au démonstrateur Korrigan (Fig. 1).

Au vu de la pertinence des hiérarchies morphologiques et des outils qui en découlent (profils d'attributs, histogrammes de formes) pour élaborer des solutions efficaces d'analyse semi-automatique d'images satellites, leur extension aux séries temporelles d'images satellites a également été étudiée, au travers d'une thèse de doctorat conduite conjointement par l'équipe OBELIX, le CNES, et CLS [7].

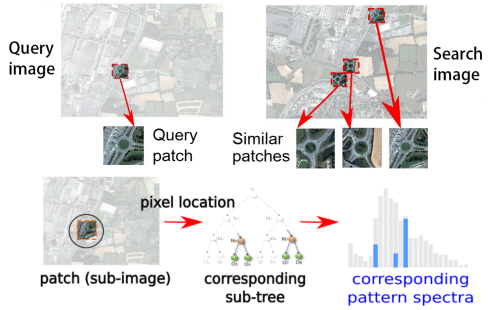

Figure 1 – Illustration d'une recherche par l'exemple avec Korrigan, une approche déterministe efficace développée par l'équipe OBELIX et le CNES [5]. Un opérateur sélectionne une zone d'intérêt dans une image satellite (*query image*) et souhaite obtenir des résultats semblables dans une image cible (*search image*). Pour ce faire, le modèle de recherche (*patch*) produit une représentation hiérarchique de laquelle sont extraits des sous-arbres (*corresponding sub-tree*). Ils produiront un sous-histogramme (*corresponding pattern spectra*) servant de signature à la recherche.

Objectifs de la thèse

En dépit des progrès récents en intelligence artificielle appliquée à l'observation de la Terre, illustrés par des performances toujours accrues sur des jeux de données standardisés, son utilisation dans un contexte de cartographie opérationnelle reste confrontée à différents verrous, que la thèse cherchera à lever.

En premier lieu, les approches modernes d'analyse d'image requièrent, dans leur grande majorité, de disposer au préalable d'une grande quantité d'exemples afin d'entraîner les modèles prédictifs comme les réseaux de neurones profonds. Les solutions interactives, permettant à un utilisateur de fouiller ses données afin d'en extraire les informations recherchées et d'en découvrir de nouvelles, restent peu étudiées.

De plus, l'avènement de l'apprentissage profond a amené un besoin toujours plus important en ressources informatiques : capacité de calcul sur CPU ou GPU, mémoire vive. La sobriété numérique est devenue aujourd'hui une question sociétale majeure, au-delà des intérêts économiques qu'elle peut procurer.

Enfin, la majorité des développements récents portent sur l'identification de motifs soit purement spatiaux (segmentation sémantique, détection d'objets) soit purement temporels (classification de séries temporelles). Les séries temporelles d'images satellites, disponibles en masse avec l'avènement de missions telles que Landsat ou Sentinel, nécessitent de porter une attention conjointe aux dimensions spatiale et temporelle.

La prise en compte de ces différents verrous s'effectuera au travers d'une problématique scientifique originale : la recherche efficace de motifs spatio-temporels dans des grands cubes de données satellites. Cette recherche, conduite de façon interactive et itérative par un utilisateur, s'appuiera sur un nombre restreint d'exemples, sur la base desquels une fouille d'un cube de données spatio-temporelles sera effectuée afin d'en extraire les motifs les plus similaires.

Bien que le paradigme de la recherche par l'exemple ou par le contenu ait été largement étudié en analyse d'image, y compris en observation de la terre, son application à des exemples spatio-temporels reste originale. Elle permettrait pourtant d'offrir de nombreux cas d'utilisation, comme par exemple l'identification d'inondations, de feux de foret, de fauchages non conformes de prairie, etc.

Ce mécanisme de fouille interactive permettra également de constituer facilement des ensembles de données de référence, qui pourront être par la suite utilisés pour entraîner des modèles IA dont la pertinence reste avérée lorsque les phénomènes étudiés peuvent être observés en amont.

Afin de mettre en oeuvre un tel mécanisme, plusieurs paradigmes peuvent être explorés, et nous souhaitons comparer l'intérêt des approches stochastiques et déterministes dans un tel contexte. Alors que les premières sont généralement basées sur un apprentissage et font aujourd'hui office de référence dans des tâches usuelles de classification (réseaux de neurones profonds), les secondes présentent l'avantage de pouvoir être implantées à l'aide d'algorithmes particulièrement efficaces, comme les hiérarchies morphologiques par exemple. Dans tous les cas, une attention particulière sera portée à l'efficience, au passage à l'échelle, et à la robustesse de la méthode en présence de peu d'exemples.

Déroulement de la thèse

Le contexte CIFRE de ce projet doctoral nécessite de réaliser à la fois une étude théorique, des développements méthodologiques, et une validation expérimentale en portant une attention toute particulière à l'industrialisation des solutions proposées et à leur acceptabilité par les utilisateurs finaux.

Il en découle le programme prévisionnel suivant, qui pourra être adapté en fonction des intérêts du candidat et de l'évolution de l'état de l'art :

- Appropriation du sujet, analyse des données disponibles et des besoins, et cas d'utilisation : 3 mois
- État de l'art : 3 mois
- Définition du protocole de test, évaluation expérimentale des solutions existantes identifiées dans l'état de l'art : 3 mois
- Élaboration d'une solution stochastique basée sur l'apprentissage profond : 7 mois
- Élaboration d'une solution déterministe basée sur les hiérarchies morphologiques : 7 mois
- Intégration dans un processus interactif de recherche de motifs spatio-temporels : 7 mois
- Validation dans différents contextes applicatifs, retour d'expérience et optimisation de l'interaction homme/machine : 3 mois
- Rédaction de la thèse : 3 mois

Valorisation

Les résultats obtenus au cours de la thèse (nouveaux procédés efficaces de recherche par l'exemple dans des bases d'images de haute résolution) seront diffusés dans la communauté scientifique au travers de publications dans les meilleures revues et conférences des domaines concernés : analyse d'image, vision par ordinateur, apprentissage profond et télédétection. Une partie importante des travaux sera valorisée dans le cadre du projet de recherche Horizon Europe Evoland (2023-2025) et présentée de manière continue aux partenaires du projet en France et à l'étranger, ainsi que lors de conférences et d'évènements liés aux activités Copernicus. Les résultats seront ensuite potentiellement utilisés en production pour les futurs produits Copernicus à l'échelle pan européenne et globale.

Il est également envisagé une collaboration avec l'équipe du Prof. Begüm Demir (`https://rsim.berlin`) dont les travaux en recherche d'image satellite par le contenu font référence dans le domaine. Cette collaboration pourra prendre la forme d'une visite de plusieurs mois à Berlin au cours de la seconde ou troisième année de thèse (si le candidat est intéressé).

En outre, les méthodes qui seront développées seront diffusées prioritairement sous la forme de code open-source. Le faible nombre de logiciels libres efficaces dans le domaine de l'analyse de séries temporelles d'images satellites devrait permettre de maximiser l'impact technologique des travaux de thèse.

Informations pratiques

Profil recherché

Le candidat devra être titulaire d'un Master ou d'un Diplôme d'Ingénieur prioritairement en Informatique, ou à défaut en Traitement du Signal et des Images, ou en Mathématiques Appliquées. Il devra être capable d'aborder les différents aspects du sujet, tels que la conception et l'optimisation d'algorithmes efficaces, la mise en oeuvre de réseaux de neurones profonds au travers de frameworks existants, l'implantation et l'expérimentation dans des environnements informatiques complexes, la maîtrise des fondements scientifiques des méthodes étudiées.

En particulier, les compétences suivantes sont attendues :

— excellentes compétences en algorithmique et programmation (C++, Python)
— expérience du traitement d'image et/ou de l'apprentissage profond
— intérêt marqué pour les problématiques liées à l'observation de la terre (des connaissances en télédétection seront appréciées)
— maîtrise de l'anglais à l'oral et à l'écrit
— curiosité et rigueur scientifiques
— esprit d'analyse et de synthèse
— communication et esprit d'équipe

Contacts

— Antoine MASSE, amasse@groupcls.com
— François MERCIOL, Francois.Merciol@irisa.fr
— Sébastien Lefèvre, Sebastien.Lefevre@irisa.fr

Rémunération

Selon le barème de l'entreprise (à discuter lors de l'entretien, minimum brut annuel : 33 k€)

Lieux d'activité

Les travaux se dérouleront majoritairement dans les locaux de CLS à Villeneuve d'Ascq à proximité de Lille (59) avec un accompagnement de l'équipe OBELIX (UMR 6074 IRISA) à Vannes (56). L'inscription académique s'effectuera au sein de l'Université Bretagne Sud (UBS) et de l'École Doctorale MathSTIC – Bretagne Océane.

La thèse sera dirigée par Sébastien Lefèvre (Professeur, UBS) et co-encadrée par François Merciol (Maître de Conférences, UBS) et Antoine Masse (Responsable Département R&D, CLS).

Processus de sélection

— **les candidatures seront évaluées au fil de l'eau**
— **date limite de dépôt du dossier : 28 juillet 2023**
— pièces à joindre au dossier : CV + lettre de motivation + lettres de recommandation + relevés de notes
— chaque candidature pré-sélectionnée fera l'objet d'un entretien

Références

[1] François Merciol and al. Tree representations of images for scalable knowledge extraction and learning for earth observation. 2017. `https://gitlab.inria.fr/obelix/triskele/`.

[2] François Merciol and al. Broceliande is a tools for classification base on triskele and random forest. 2018. `https://gitlab.inria.fr/obelix/broceliande/`.

[3] François Merciol, Loïc Faucqueur, Bharath Bhushan Damodaran, Pierre-Yves Rémy, Baudouin Desclée, Fabrice Dazin, Sébastien Lefèvre, Antoine Masse, and Christophe Sannier. Geobia at the terapixel scale : Toward efficient mapping of small woody features from heterogeneous vhr scenes. *ISPRS International Journal of Geo-Information*, 8(1) :46, 2019.

[4] François Merciol, Minh-Tan Pham, Deise Santana, Antoine Masse, and Christophe Sannier. BROCELIANDE : a comparative study of attribute profiles and featureprofiles from different attributes. In *ISPRS*, Nice, France, June 2020.

[5] Behzad Mirmahboub, Jérôme Moré, David Youssefi, Alain Giros, François Merciol, and Sébastien Lefèvre. Fast Pattern Spectra using Tree Representation of the Image for Patch Retrieval. In *Discrete Geometry and Mathematical Morphology 2021*, pages 107–119, May 2021.

[6] Deise Santana Maia, Minh-Tan Pham, Erchan Aptoula, Florent Guiotte, and Sébastien Lefèvre. Classification of remote sensing data with morphological attributes profiles : a decade of advances. *IEEE geoscience and remote sensing magazine*, 9(3) :43–71, 2021.

[7] Caglayan Tuna. *Morphological Hierarchies for Satellite Image Time Series*. PhD thesis, Université Bretagne Sud, 2020.