DOSSIER DE CANDIDATURE A UNE ALLOCATION DE RECHERCHE POUR LA RENTREE 2016

Titre de la thèse : Représentation parcimonieuse et apprentissage actif pour l'écologie marine

Laboratoire d'accueil ULCO : ULCO/LISIC, Laboratoire d'Informatique Signal Image de la Côte d'Opale.

Web : http://www-lisic.univ-littoral.fr/

Directeur de thèse ULCO : Professeur Denis Hamad

Partenaire étrangers si identifié (noms de la structure de recherche et du codirecteur étranger) : Université Libanaise, Faculté des Sciences Economiques et de Gestion, Département : Informatique de Gestion. Codirecteur Professeur Ali Kalakech, Encadrante : Mariam Kalakech (MCF).

-Thématique : milieux aquatiques

***LABORATOIRE D'ACCUEIL**

Nom du laboratoire d'accueil : Laboratoire d'Informatique Signal et Image de la Côte d'Opale (LISIC).

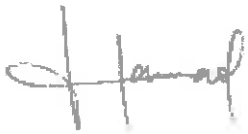
Nombre de HDR dans le laboratoire : 12
Nombre de thèses encadrées dans le laboratoire (rentrée 2014) : 26 **(dont 15 cotutelles)**
Durée moyenne des thèses soutenues dans le laboratoire, sur la période 2010-2014 : **39 mois**

ENCADREMENT

Nom, Prénom du directeur de laboratoire : Christophe Renaud
Nom, Prénom du directeur de thèse (si différent du directeur de laboratoire) : Denis Hamad
Nombre de doctorats en préparation sous la direction du directeur de thèse : 1

Avis détaillé du directeur de thèse :
Le sujet rentre dans l'axe prioritaire de l'ULCO et en accord avec le GIS Campus de la Mer. Notre possédons une bonne expérience dans le domaine d'analyse de données en particulier en écologie marine en collaboration avec l'IFREMER, le LOG dans différents projets Interreg Dymaphy, BQR PhytoClas et FlowCAM (Laboratoire-Écologie Numérique des Milieux Aquatiques UMons-Belgique). Dans le contexte de l'écologie marine, les données sont issues de nombreux capteurs installés sur sites ou embarqués et de caméras hautes résolutions. L'organisation de ces données et l'extraction d'informations pertinences, utiles et robustes sont de véritables défis. En effet, les données sont abondantes, arrivent en continu, comprennent des mesures bruitées, manquantes, voire aberrantes. L'ambition du sujet est d'utiliser des approches originales de représentation parcimonieuse et d'apprentissage actif (projet régional REPAR) pour la reconnaissance et la discrimination des espèces de phytoplancton, dénombrer voire détecter des espèces pouvant être nuisibles.

Signature du directeur de thèse

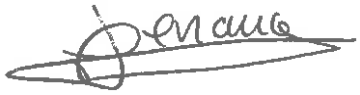


Avis détaillé du directeur de laboratoire :

le projet de thèse proposé se situe dans le domaine «milieux aquatiques » de cet appel et correspond à une réelle compétence du Professeur Hamad. Ce dernier développe depuis plusieurs années des recherches dans le domaine de l'analyse d'images de phytoplanctons, au travers de divers projets importants soulignés dans ce dossier. La thématique est en lien avec les spécialités de l'équipe d'accueil, qui a été évaluée de manière très favorable par l 'AERES fin 2013. Elle est également l'un des thèmes prioritaires de notre Université, au travers de son

GIS Campus de la Mer. J'émets donc un avis très favorable quant au financement de ce sujet.

Signature du directeur de laboratoire



PROJET DE THESE

Intitulé du projet de thèse : Représentation parcimonieuse et apprentissage actif pour l'écologie marine.

Domaine scientifique : surveillance de l'environnement marin

Résumé (1/2 page maxi.) :

Le sujet de thèse se veut de développer des techniques avancées en traitement de données par classification automatique active et en réduction de la dimension avec intégration de connaissances contextuelles du domaine d'application. Les résultats de la classification seront interprétés et comparés avec la classification manuelle réalisée par les taxonomistes dans un but d'amélioration des performances.

Les performances des algorithmes de classification dépendent de la qualité des données à disposition. Mais, aussi de l'intégration, même partielle, de connaissances de plus haut niveau dans leur conception.

L'originalité de notre proposition est l'intégration, dans le processus de sélection mais aussi bien de la classification active, des connaissances sous forme de comparaisons ou préférences simples à formaliser par un opérateur non spécialiste. Celles-ci seront générées automatiquement ou d'une manière interactive par retour d'expérience.

Projet de thèse (5 pages maxi.) :

Développer sur cinq pages :

- ***Le sujet de recherche choisi et son contexte scientifique***
- ***L'état du sujet dans le laboratoire et l'équipe d'accueil***
- ***Le programme et l'échéancier de travail***
- ***Les retombées scientifiques et économiques attendues***
- ***Les collaborations prévues et une liste de 10 publications maximum portant directement sur le sujet***

Dans le cas d'une demande PMCO, préciser en quoi les travaux menés au cours de la thèse répondent aux problématiques des territoires du littoral Côte d'Opale

1- Le sujet de recherche choisi et son contexte scientifique

Le Phytoplancton joue un rôle important dans l'évaluation de la qualité des eaux marines. Cette évaluation s'effectue, de plus en plus, par des mesures hautes fréquences.

Dans le contexte de l'écologie marine, les données sont issues de capteurs divers installés sur sites ou embarqués, de caméras hautes résolutions, voire des informations qualitatives ou quantitatives acquises par expérience. La gestion et l'organisation de ces données dans l'objectif d'extraction d'informations pertinentes, utiles et robustes est un véritable défi. En effet, ces données sont abondantes, bruitées, manquantes, voire aberrantes. Notre ambition est d'utiliser des approches originales de représentation parcimonieuse et d'apprentissage actif pour la reconnaissance et la discrimination des espèces de phytoplancton, la recherche d'états d'efflorescence, de dénombrer voire détecter des espèces pouvant être nuisibles. Une attention particulière sera portée à la qualité des données. Il est clair que, dans un contexte de données abondantes, il n'y a « pas de résultats de qualité sans données de qualité ».

2- L'état du sujet dans le laboratoire et l'équipe d'accueil

Nous avons acquis une expérience de plus de 8 ans dans le domaine de l'écologie marine avec deux thèses soutenues (Wacquet 2011 et Rousseeuw 2014) et des articles publiés dans des journaux et conférences internationaux. Nous travaillons en étroite collaboration avec l'IFREMER, l'Agence des Eaux Artois Picardie et le laboratoire LOG-ULCO. Au niveau international, nous travaillons avec l'Université de Mons-Belgique dans le cadre du comité de pilotage de l'Action FlowCAM / ZooPhytoImage. Aussi, nous avons travaillé avec CEFAS-UK et RWS-The Netherlands dans le cadre du projet Interreg Dymaphy. Nous avons participé à différents projets (DYMAPHY, PhytoClas, FlowCAM, Marel-Carnot, REPAR). Dans ce sens, nous possédons une bonne expérience dans la compréhension de la problématique et le traitement des données marines. Nous sommes membres du conseil scientifique du GIS Campus de la Mer. Au niveau régional nous sommes responsables du thème 1 du GIS GRAISyHM : traitement du signal et de l'image, depuis 12 ans. Nous portons le projet régional REPAR 2015-2017 « Représentation Parcimonieuse et Apprentissage Dynamique pour le signal et l'image ». Par ailleurs, nous possédons une expérience de 10 ans d'enseignement en Master 2 ISIDIS du module Data Mining.

Les membres du projet ont participé à différents Workshop dans le domaine de traitement de données marines :

- Workshop on data analysis in Flow Cytometry, Woerden, The Netherlands, January 16 - 17, 2014.
- 4ème Workshop du GIS 3SGS : Groupement d'Intérêt Scientifique, Surveillance, Sûreté et Sécurité des Grands Systèmes. Valenciennes, 12-13 Octobre, 2011.
- Workshop on data analysis in Flow Cytometry, Delft, The Netherlands, March 23-24, 2010.
- Atelier du RNSM sur les mesures à Haute fréquence dans l'environnement marin qui a eu lieu à Wimereux, 22 - 23 octobre 2009.

3- Le programme et l'échéancier de travail

Les performances des algorithmes d'apprentissage sont souvent pénalisées par la très grande dimension de l'espace de données, Dans ce sens, il est souvent plus judicieux de construire un espace de dimension réduite. C'est l'idée de base de notre approche de sélection spectrale d'attributs et d'instances pour la catégorisation de données de grande dimension. En effet, les approches spectrales constituent une démarche unifiée pour traiter des problèmes de classification aussi bien que pour la sélection d'attributs de représentation.

Les performances des algorithmes de classification dépendent évidemment de la qualité des données à disposition mais aussi de l'intégration de connaissances contextuelles même partielles de plus haut niveau dans leur conception. Certaine srelations (objets x objets), dites contraintes de comparaison entre paires d'objets, sont faciles à formaliser par l'analyste : deux données sont similaires et doivent donc être groupées ensemble, ou non similaires et doivent appartenir à des groupes distincts.

L'originalité de notre proposition est l'intégration, dans le processus de sélection mais aussi bien de classification, des connaissances sous forme de comparaisons ou préférences simples à formaliser par un opérateur non nécessairement spécialiste. Celles-ci seront générées d'une manière interactive par retour d'expérience lors de l'interprétation des résultats de classification.

Le sujet de thèse se veut de développer des techniques avancées de classification automatique et de réduction de la dimension avec éventuellement l'intégration de connaissances contextuelles du domaine d'application. Les résultats de la classification seront interprétés et comparés avec la classification manuelle réalisée par les taxonomistes dans un but d'amélioration des performances.

Le sujet de thèse s'appuie sur la représentation parcimonieuse pour :

1. la représentation et l'organisation des données :
   - sélection d'attributs,
   - sélection d'échantillons ou instances,
   - réduction de la dimension de l'espace attributs,
   - réduction de la dimension de l'espace observations.

2. l'apprentissage spectral :
   - représentation par graphe des données,
   - introduction de connaissances a priori dans le graphe,
   - classification spectrale avec connaissances partielles.

Compétences complémentaires des partenaires libanais

Avec Mariam Kalakech, nous avons publié deux articles dans des journaux internationaux et plusieurs contributions dans des conférences internationales. Ses compétences en réduction de la dimensions et sélection d'attributs et en analyse de textures d'images assureront le succès du projet. Mariam est enseignante invitée au LISIC-ULCO en 2015 et en 2016.

Le Professeur Ali Kalakech est spécialiste en bases de données, réseaux de communication, modélisation et programmation objet et apprentissage. Il apportera ses compétences en apprentissage et en gestion et organisation des bases de données.

4- Les retombées scientifiques et économiques attendues

A l'issu de campagne de mesures hautes fréquences, l'utilisateur se trouve vite face à une grande quantité de données dont il faut analyser afin d'en extraire les informations utiles.

Plusieurs difficultés sont à surmonter dans ce projet :

3. le volume très important des données,
4. l'invariance des attributs extraits,
5. la robustesse par rapport à des données aberrantes ou manquantes,
6. l'aspect évolutif du système avec données non stationnaires.

Dans le cadre de cette thèse, un logiciel avec interface conviviale serait mis à disposition de la communauté en particulier nos collaborateurs du domaine de l'écologie marine. L'interface permettrait une interaction utilisateur conviviale afin de corriger ou affiner la classification obtenue. Dans le futur, nous devrons être en mesure de suivre l'évolution des espèces en temps réel dans de campagnes en mer lors de prélèvements en haute résolution voire en quasi continu.

5- Bibliographie relative au sujet

Thèses soutenues relatives au sujet

1. Guillaume Wacquet. « Classification spectrale semi-supervisée. Application à la surveillance de l'écosystème marin ». Allocation de recherche ministérielle. Thèse de doctorat ULCO soutenue le 8 décembre 2011. Directeur D. Hamad.

2. Mariam Kalakech. « Sélection semi-supervisée d'attributs : application à la classification de textures couleur ». Thèse Soutenue le 08 juillet 2011 à Lille 1. Directeurs L. Macaire et D. Hamad.

3. Kevin Rousseeuw., « Prédiction des efflorescences Pytoplanctonique dans les rivières et les écosystèmes marins côtiers ». de doctorat ULCO soutenue le 11 décembre 2014. Directeur D. Hamad.

Chapitres de livre

1. Mansour A., Leblond I., Hamad D., and Artigas L. F. Sensor Networks for Underwater Ecosystem Monitoring and Port Surveillance Systems. Chapter 19. Sensor Networks for Sustainable Development. Pages 431–468. CRC Press. DOI: 10.1201/b17124-25. 2014.
2. Kévin Rousseeuw, Émilie Caillault, Alain Lefebvre, Denis Hamad. Modèle de Markov Caché hybridé pour la surveillance de l'environnement marin. Chapitre d'ouvrage. Editions CNRS, à paraître.

Journaux

1. Porebski A., Vandenbroucke N. Macaire L. and Hamad D. A new benchmark image test suite for evaluating color texture classification schemes. Multimedia Tools and Applications Journal, Volume 70, Issue 1, pp. 543-556, 2014
2. Rousseeuw K., Poisson-Caillault E., Lefebvre A., and Hamad D. Hybrid Hidden Markov Model for Marine Environment Monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Vol. 8, Issue 1, pp. 204-213, 2014.
3. Kalakech M., Biela Ph., Hamad D. and Macaire L. Constraint score evaluation for spectral feature selection. Neural Processing Letters, vol. 38, issue 2, pp. 155-175, Oct. 2013.
4. Wacquet G., Poisson Caillault E., Hamad D. and Hébert P.A., Constrained Spectral Embedding for K-Way Data Clustering. Pattern Recognition Letters, Volume 34, Issue 9, pp. 1009-1017, 2013.
5. Kalakech M., Biela Ph., Macaire L., Hamad D. Constraint scores for semi-supervised feature selection: A comparative study. Pattern Recognition Letters vol. 32, pp. 656-665, April 2011.

Conférences internationales

1. Hamad D. Constrained graph embedding. Invited speak in International Symposium on 3D Imaging, Metrology, and Data Security, September 26th - 29th, 2015, Shenzhen, China.
2. Kalakech M., Porebski A., Vandenbroucke N., D. Hamad. A New LBP Histogram Selection Score for Color Texture Classification. 5th Int. Conference on Image Processing Theory, Tools and Applications (IEEE-IPTA'15), pp. 242-247, Orléans (France), November 2015.
3. Porebski, A. Vandenbroucke N, Hamad D., A fast embedded selection approach for color texture classification using degraded LBP. 5th Int. Conference on Image Processing Theory, Tools and Applications (IEEE-IPTA'15), pp. 254-259, Orléans (France), November 2015.
4. Nasser A. and Hamad D., Detection and visualisation of outliers using Kernel Principal Components. The Fifth Int. Conference on Digital Information and Communication Technology and its Applications (DICTAP2015), Faculty of Engineering - Lebanese University, Lebanon, April 29 - May 1, 2015.
5. Rousseeuw K., Lefebvre A., Poisson Caillault E., Hamad D., Detection of contrasted physico-chemical and biological environmental status using unsupervised classification tools. 5th FerryBox Worshop, Helsinki, Finland, 24-25 April 2013.
6. Kalakech M., Biela Ph., Hamad D., Macaire L., Semi-supervised evaluation of constraint scores for feature selection. International Conference on Neural Computation Theory and Applications, pp. 175-182, Paris, 24-26 octobre 2011.
7. Wacquet G. Hebert P.A., Caillault Poisson E. and Hamad D., Semi-Supervised K-Way Spectral Clustering using Pairwise Constraints. Int. Conference on Neural Computation Theory and Applications, pp. 72-81, Paris, 24-26 octobre 2011.
8. Kalakech M., Porebski A., Biela Ph., Hamad D., Macaire L., Constraint score for semi-supervised selection of color texture features. 3rd Int. Conference on Machine Vision, Hong Kong, China, December 28 - 30, ICMV 2010.