



# HABILITATION À DIRIGER LES RECHERCHES

De la Vision par Ordinateur aux  
Modèles Profonds

**Diane LINGRAND**

Laboratoire d'Informatique, de Signaux et Systèmes de Sophia Antipolis (I3S)  
UMR7271 Université Côte d'Azur - CNRS

Présentée en vue de l'obtention de  
l'**habilitation à diriger les recherches**  
en Informatique  
Soutenue le : bientôt

Devant le jury, composé de :  
Samia AINOUI, Pr, INSA Rouen  
Isabelle BLOCH, Pr, Université Sorbonne  
Marc CHAUMONT, MCF HdR, Université  
Bretagne Sud  
... .., Univ ...  
... .., Fr,  
Univ ...





# DE LA VISION PAR ORDINATEUR AUX MODÈLES PROFONDS

---

*From Computer Vision to Deep Models*

**Diane LINGRAND**



## **Jury :**

### **Rapporteurs**

Samia AINOUS, Pr, INSA Rouen

Isabelle BLOCH, Pr, Université Sorbonne

Marc CHAUMONT, MCF HdR, Université Bretagne Sud

### **Examineurs**

... .., Univ ...

Diane LINGRAND

***De la Vision par Ordinateur aux Modèles Profonds***

ix+71 p.

# **De la Vision par Ordinateur aux Modèles Profonds**

## **Résumé**

Résumé en 100 pages de 25 ans d'activité.

**Ceci n'est pas le document définitif mais  
un brouillon en cours de rédaction.**

**Mots-clés :** Apprentissage Automatique, Vision par Ordinateur.

# **From Computer Vision to Deep Models**

## **Abstract**

Enjoy

**Keywords:** Machine Learning, Computer Vision.



# Remerciements

---

Merci !





# Table des matières

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Quand l'apprentissage n'était pas automatique</b>	<b>3</b>
2.1	La Vision par Ordinateur à la fin du XXème siècle . . . . .	5
2.2	Passage à l'échelle : utilisation de grilles de calcul . . . . .	7
2.2.1	Gridification d'algorithmes . . . . .	8
2.2.2	(Re-)soumission efficace de calculs . . . . .	10
<b>3</b>	<b>Apprentissage automatique pour les données images</b>	<b>15</b>
3.1	Gastronomie : classification subjective et incertitude dans les labels . . . . .	17
3.1.1	Annotation d'images par oculomètre . . . . .	18
3.1.2	Indépendance de l'estimateur GBIE aux classes cibles . . . . .	20
3.1.3	Gestion de l'incertitude des labels . . . . .	21
3.2	Classification de vidéos . . . . .	21
3.2.1	Un descripteur de mouvement : les SINGLETs . . . . .	22
3.2.2	Une vidéo est un tenseur (2D+T) . . . . .	23
3.2.3	Déformations dynamiques de vidéos . . . . .	32
<b>4</b>	<b>Apprentissage profond</b>	<b>37</b>
4.1	Biodiversité en biologie marine . . . . .	37
4.1.1	Travaux préliminaires : reconnaissance de poissons dans leur état naturel	38
4.1.2	Collaboration avec le laboratoire ECOSEAS . . . . .	39
4.2	Archeology . . . . .	46
4.2.1	Previous work . . . . .	46
4.2.2	ANR AIWOOD . . . . .	48
4.3	Navigation autonome . . . . .	54
4.3.1	Impact des cartes routières pour la prédiction de trajectoire . . . . .	56
4.3.2	Prise en compte d'anciennes cartes routières pour faciliter l'estimation de cartes actualisées. . . . .	56
<b>5</b>	<b>Réflexions et perspectives</b>	<b>59</b>
5.1	Reflexions concernant les travaux passés . . . . .	59
5.2	Et maintenant ? . . . . .	60
5.3	Prise en compte des connaissances . . . . .	61
	<b>Bibliographie</b>	<b>63</b>



# CHAPITRE 1

---

## Introduction

### A COMPLETER

Dans ce manuscrit, je vais décrire mes travaux depuis ma thèse jusqu'à aujourd'hui. Le premier chapitre est rapidement consacré à la période de ma thèse jusqu'en 2011 environ et concerne la vision par ordinateur, l'imagerie médicale ainsi que la recherche large d'hyper-paramètres pour des chaînes de traitement de données médicales. Dans un second chapitre, je décris une période plus récente, de 2013 à 2023, pendant laquelle j'ai étudié les algorithmes d'analyse d'images et de vidéos, que ce soit pour la représentation des données (descripteurs) ou leur classification ou détection par apprentissage automatique, outil plus efficace de recherche d'hyper-paramètres. Le troisième chapitre décrit des travaux plus récents avec de l'apprentissage profond concernant des domaines d'application dont mes recherches avaient déjà débuté auparavant. Je conclus par quelques réflexions et des propositions de perspectives de ces travaux.



# CHAPITRE 2

---

## Quand l'apprentissage n'était pas automatique

*Dans les années 90, Mike Brady disait :*

*“Pour un stage de DEA, il faut que ça marche avec une image.  
Pour une thèse, il faut que ça marche avec deux images.”*

*Dans les années 2000, on est passé aux grilles de calculs puis à l'apprentissage automatique et à l'apprentissage profond.*

---

<b>2.1</b>	<b>La Vision par Ordinateur à la fin du XXème siècle . . . . .</b>	<b>5</b>
<b>2.2</b>	<b>Passage à l'échelle : utilisation de grilles de calcul . . . . .</b>	<b>7</b>
2.2.1	Gridification d'algorithmes . . . . .	8
2.2.2	(Re-)soumission efficace de calculs . . . . .	10

---





## 2.1 La Vision par Ordinateur à la fin du XXème siècle

Je suis née dans la période d'invention du détecteur de contours de Sobel, j'ai passé mon baccalauréat la même année que la publication du détecteur de points d'intérêts de Harris et Stephens ([Harris & Stephens, 1988](#)). Ce que je découvris plus tard, c'est que j'ai grandi dans une période où chaque image nécessitait un nouvel algorithme pour détecter ses contours ou sa texture : type de filtre, coefficients des filtres, seuillage postérieur ... J'ai découvert le traitement d'images et la vision par ordinateur vers la fin de mes études d'ingénieur à une époque où les images numériques étaient encore rares et seulement accessibles dans les milieux professionnels. Les appareils photographiques numériques grand public allaient arriver quelques années plus tard. Ce domaine en pleine effervescence m'a tout de suite plu : ma décision d'orientation en vision par ordinateur était prise.

Pendant ma thèse ([Lingrand, 1999](#)), effectuée à l'INRIA Sophia Antipolis, dans l'équipe RobotVis, sous la direction de Thierry Viéville, j'ai étudié le problème de la vision par ordinateur concernant des séquences d'images monoculaires non calibrées. Ce problème est bien formalisé mais, dans le cas général, il n'est pas possible de retrouver tous les paramètres concernant la caméra (paramètres intrinsèques et extrinsèques) et la structure géométrique de la scène 3D. J'ai étudié les cas particuliers physiques (modèle de caméras, évolution des paramètres internes de la caméra, déplacement des objets dans la scène, de la caméra, structure de la scène) conduisant à des équations spécifiques, de façon hiérarchique. Les singularités permettent de retrouver, selon les cas, plus ou moins d'éléments que dans le cas général sur le mouvement ou la structure, mais toujours avec plus de précision car moins de paramètres sont en jeu ([Viéville & Lingrand, 1999](#); [Viéville, Lingrand, & Gaspard, 2001](#); [Lingrand, 2002b, 2002a](#)). On pouvait conclure qu'un modèle général est important mais que utiliser des informations ou connaissances supplémentaires est un atout.

En post-doctorat, au MNI (Institut Neurologique de Montréal, McGill), l'équipe de Jean Gotman s'intéressait à la fusion des signaux EEG et fMRI concernant des patients épileptiques résistants aux traitements connus. J'ai observé que les périodes d'acquisition simultanée EEG et fMRI était longue et qu'il était impossible de rester complètement immobile dans le scanner IRM. Pour tout traitement de signaux fMRI, il est nécessaire de recalibrer les images 3D dans un repère commun. En prenant en compte l'information que les mouvements sont de faible amplitude, et donc potentiellement approximable au premier ordre, j'ai pu adapter mes travaux de thèse au recalage rigide d'images 3D et ainsi améliorer ce recalage. J'ai pu vérifier l'hypothèse de départ qui était que la plupart du temps on est dans une situation de mouvement particulier et ainsi améliorer la précision du recalage ([Lingrand, Montagnat, Collins, & Gotman, 2001](#)).

De retour en France, à mon recrutement Maîtresse de Conférences au laboratoire I3S, Sophia Antipolis, j'ai participé aux travaux sur les contours actifs par ensembles de niveaux (*levelsets*) et courbes paramétrées de type B-splines. Pour ces méthodes, il était important de bien régler les hyper-paramètres tels que les forces internes et externes qui sont l'expression des contraintes sur les formes segmentées et des propriétés des objets à segmenter telles que la texture par exemple. Plus particulièrement, je me suis intéressée à la segmentation du myocarde dans des images TEMP (Tomographie à Émission Mono-Photonique) volumiques et temporelles (3D+T). Les cardiologues avaient présenté leur méthode de détection d'anomalie de l'épaisseur du muscle cardiaque sur une coupe à quelques instants du cycle cardiaque et il était apparu que cette vue partielle ne permettait pas de détecter tous les types d'anomalies. Une segmentation en 3D et temporelle a permis de mieux visualiser mais aussi de mieux quantifier les différents volumes ([Charnoz, Lingrand,](#)

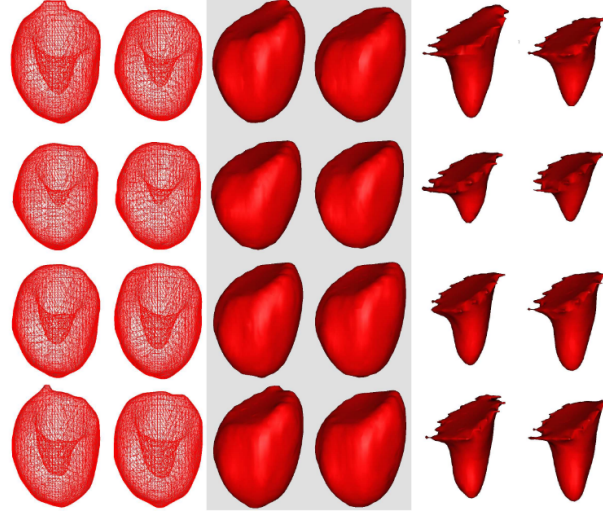


Figure 2.1 – A gauche : la segmentation vue par maillage. Au milieu : la membrane extérieure. A droite : la chambre intérieure. A l'intérieur de chaque partie, de gauche à droite et de haut en bas sont présentés les instants 0 à 7 du cycle cardiaque. Grâce à cette segmentation, la phase de diastole (3 à 7) et la phase de systole (0 à 2) sont bien dissociées.

& Montagnat, 2003; Lingrand, Charnoz, Koulibaly, Darcourt, & Montagnat, 2004; Lingrand & Montagnat, 2005). Un exemple de segmentation est donnée figure 2.1. C'est par une discussion avec les experts du domaine médical que de tels travaux ont été possibles.

Le modèle d'évolution des courbes de niveau que nous avons établi pour ce problème est basé sur l'équation :

$$\frac{\partial u_n}{\partial t} = \left( \lambda_{S_{i,n}} (I_n - \mu_{in_{S_{i,n}}}) - \lambda_{out} (B - I_n) - \lambda_c \kappa_n \right) \|\nabla u_n\|$$

où :

$n$  : correspond à l'indice temporel

$S_{i,n}$  : est une des 5 régions  $i$ ,  $S_0$  étant la région de l'apex du ventricule et  $S_4$  la région au delà de la base

$B$  : fond de l'image (hors ventricule)

$\lambda_{S_{i,n}}$ ,  $\lambda_{out}$  et  $\lambda_c$  : sont des hyper-paramètres qu'il faut choisir de façon à converger au mieux vers la solution qui doit donc être connue. Lors de cette étude, nous avons à la fois travaillé sur des données simulées NCAT (dites de fantôme) et sur de vraies données issues de l'hôpital Pasteur. Dans ces deux cas, les jeux de paramètres optimaux étaient bien différents.

En 2003, on écrivait en conclusion de (Charnoz et al., 2003) :

L'étude des paramètres mathématiques nécessite maintenant d'être poursuivie sur des données cliniques. Une validation plus profonde, incluant des comparaisons des résultats de segmentation par des experts médicaux sur des bases de données plus grandes, est nécessaire. Des contraintes supplémentaires prenant compte de la physiologie du cœur telles que la quasi incompressibilité du muscle devront être ajoutées. La méthode pourra alors être utilisée pour l'extraction quantitative de paramètres cliniques.

Néanmoins, peu d'images étaient disponibles à cette époque et même si leur nombre grandissait, on ne savait pas encore bien gérer de grosses quantités de données ni adapter automatiquement les différents algorithmes de traitement d'images à des variations. Ainsi, même si des applications industrielles existaient, elles étaient liées à des conditions bien spécifiques d'utilisation : éclairage fixe, objets restreints, paramètres des caméras fixe ...

Parmi les travaux que je viens de citer :

- le modèle de mes travaux de thèse fonctionne pour les caméras qui étaient disponible sur le système robotique de mon équipe, le tout fonctionnant dans une salle fixe avec éclairage fixe,
- les travaux de post-doctorat fonctionnent avec le scanner IRM qui était disponible pour l'hôpital dans la journée et les expériences en soirée,
- les travaux de segmentation cardiaque fonctionnent bien pour les images fournies par l'hôpital Pasteur à Nice.

Si on souhaite changer les conditions de fonctionnement, différents hyper-paramètres doivent être considérés : les seuils sur les détecteurs de points d'intérêts, les critères d'arrêts, les paramètres d'évolution ...

Cela nécessite d'avoir accès à une grande diversité de données mais aussi d'explorer un espace d'hyper-paramètres qui peut s'avérer très voir trop grand. La section suivante concerne le passage à l'échelle en imagerie biomédicale.

## 2.2 Passage à l'échelle : utilisation de grilles de calcul

Ces travaux ont été réalisés dans le cadre du projet Européen EGEE\* (Enabling Grids for E-sciencE) qui a fournit l'infrastructure de calcul (200000 CPU à la fin du projet, en 2010, répartis en plus de 250 centres de calculs avec plus de 13000 utilisateurs). Ils ont également été financés par le projet ANR Neurolog† et le projet ONCO-MEDIA‡ (ONtology and COntext related MEDical image Distributed Intelligent Access) du programme régional ICT-ASIA.

L'idée principale était de permettre la validation d'algorithmes à grande échelle en utilisant davantage de données, davantage d'algorithmes et de variations de ces algorithmes ainsi que des ressources de calcul partagées et disponible via le projet EGEE. Ce projet d'envergure concernait différents domaines répartis par organisations virtuelles (VO), chacune d'elles ayant ses problématique propres. Dans la 'VO biomed' à laquelle j'appartenais, nous nous intéressions aux applications biomédicales et notamment, les applications visées dans le projet ANR Neurolog : segmentation du cerveau, recalage 3D, détection de régions, classification de tissus et suivi temporel pour la sclérose en plaques, les attaques cérébrales, tumeurs cérébrales et maladie d'Alzheimer. Ces applications sont en réalité composée de différents algorithmes élémentaires composés sous la forme de flot de traitement (ou *workflow*). Un flot de traitement comporte différents algorithmes qui peuvent chacun présenter différentes variations : que ce soit dans la nature de l'algorithme ou dans sa paramétrisation.

La figure 2.2 présente l'architecture du projet Neurolog avec les différentes parties étudiées afin de permettre le déploiement d'algorithme sur les grilles de calculs permettant des recherches exhaustives d'hyper-paramètres sur de grandes quantité de données réparties.

---

\*. <https://eu-egee.org/>

†. ANR-06-TLOG-024 : <https://neurolog.i3s.unice.fr/neurolog>

‡. <http://www.onco-media.com>

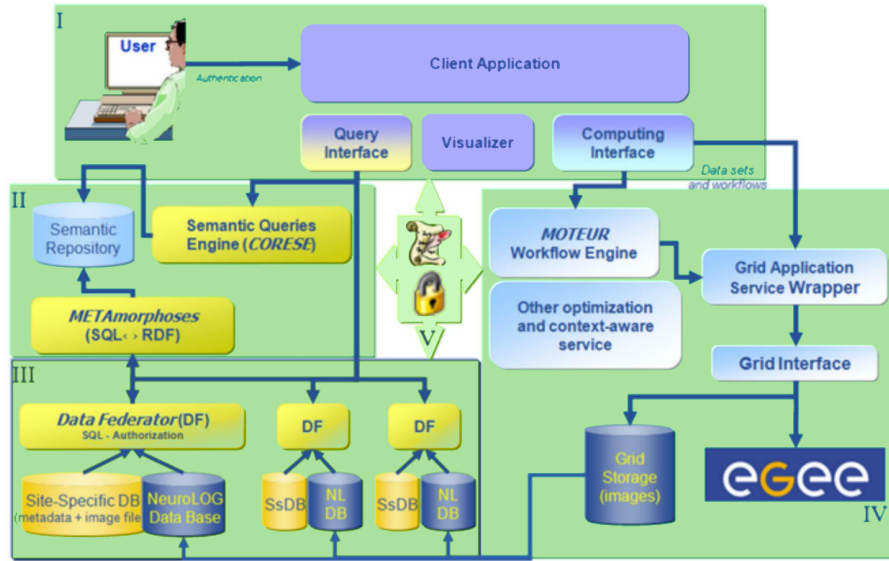


Figure 2.2 – Architecture du projet Neurolog. J’ai contribué aux composantes “MOTEUR Workflow Engine” et “Other optimization and context-aware services”

Mes contributions ont concernés une partie du moteur de flots de traitements et de données ainsi que des optimisations pour l’exécutions de travaux sur grille.

### 2.2.1 Gridification d’algorithmes

A ce moment, différents laboratoires disposaient d’applications médicales testées localement sur des petits jeux de données. Afin de permettre à ces applications d’être exécutées sur des grilles de calcul nous avons élaboré un moteur de déploiement de flots de traitements avec des flots de données pour les grilles : MOTEUR (oMe-made OpTimisEd scUfl enactoR) (Glatard, Montagnat, Lingrand, & Pennec, 2008). Ce moteur permet d’exprimer clairement les flots de traitement et de données afin d’optimiser à la fois le parallélisme et la composition de traitements et de données (Montagnat, Glatard, & Lingrand, 2006).

Une vue générale de l’utilisation du moteur de flots est présentée sur le poster en figure 2.3 décrit dans (Rojas Balderrama et al., 2008). La chaîne de traitements est exprimée par les experts en analyse d’images médicales puis convertie dans un langage de flots de traitements, Scufi. L’application est ensuite gridifiée sur la grille EGEE afin d’obtenir des réglages d’hyper-paramètres sur les images de plusieurs patients.

Sur la figure 2.4, nous présentons un flot de traitements et données simplifié que nous avons réalisé à partir d’une application développée par l’équipe Asclépios de l’INRIA pour la segmentation du cerveau (Pernod, Souplet, Rojas Balderrama, Lingrand, & Pennec, 2008). La gridification de cette application sur la grille EGEE grâce à MOTEUR a permis d’étudier un hyper-paramètre, la proportion de voxels à prendre en compte lors de l’algorithme EM, et de d’établir que 1% permettait d’accélérer l’algorithme sans baisse de qualité du résultat.

Ce premier déploiement sur EGEE, à l’aide de MOTEUR, nous a permis de valider cette approche et de mettre en évidence la nécessité de mieux gérer les différents types de parallélisme

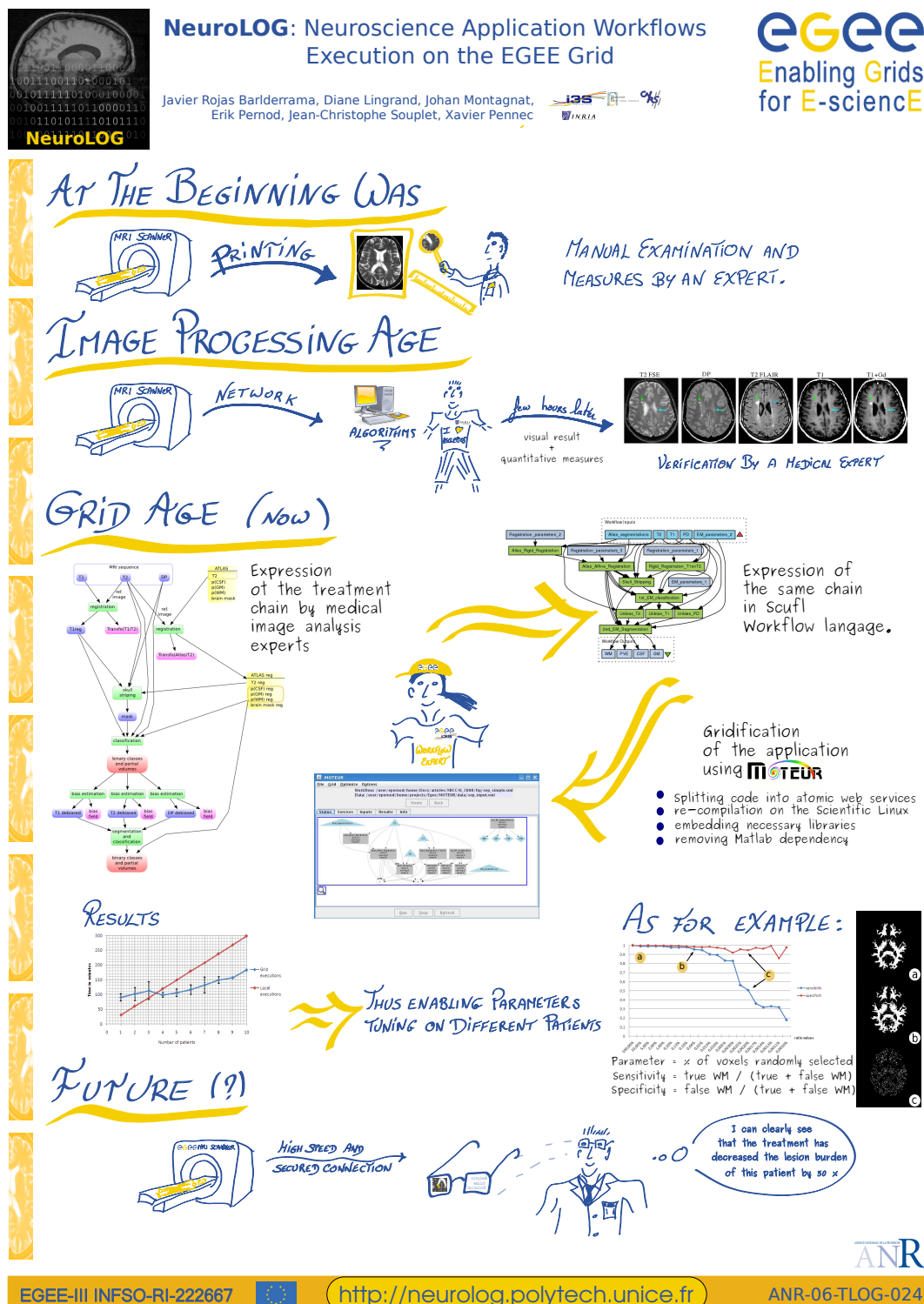


Figure 2.3 – Présentation d'une partie des travaux du projet Neurolog à la conférence EGEE 2008.

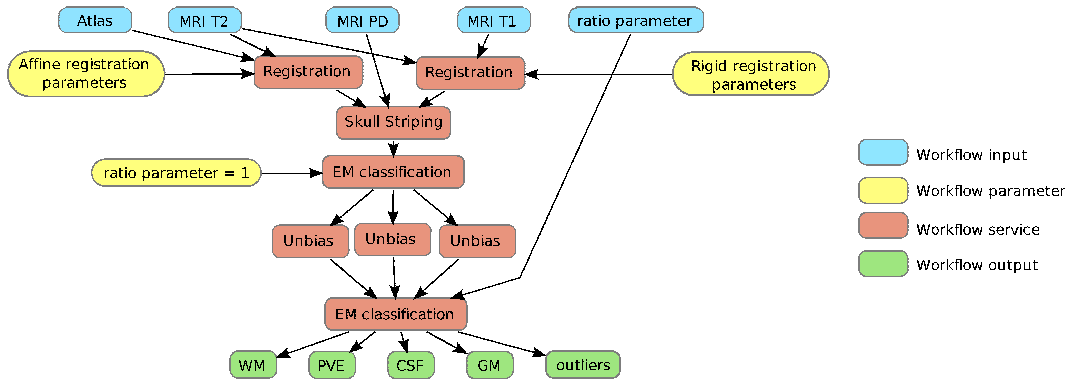


Figure 2.4 – Workflow simplifié pour la segmentation du cerveau : matière blanche (WM), fluide céphalo-rachidien (CSF), matière grise (GM)

afin de réduire le temps pour obtenir les résultats finaux. En effet, des latences très variables ont été observées avec un lourd impact sur le temps global d'exécution.

D'autres déploiements ont eu lieu dans le cadre de ces projets (Olabarriaga, Lingrand, & Montagnat, 2008) mais je me suis ensuite focalisée sur la gestion des soumissions de travaux sur la grille afin de minimiser les latences observées.

### 2.2.2 (Re-)soumission efficace de calculs

La grille EGEE consiste en une fédération de nombreux centres de calculs répartis dans le monde. Sur une grille aussi complexe que hétérogène, la variabilité de latence des soumissions de calculs est très importante. Cette latence (temps mesuré entre la demande d'exécution et le démarrage effectif du calcul sur un noeud terminal) a différentes origines (pannes électriques, pannes réseaux, mise à jour ...) et sa distribution est à queue lourde : un nombre conséquent de travaux ne se terminent jamais. Si les grilles de calculs pouvaient permettre d'accélérer des applications biomédicales à une échelle supérieure à ce qui pouvait être réalisé sur un cluster de CPU (Grid'5000 par exemple), la queue lourde des latences était un vrai problème. Du point de vue de l'utilisateur d'une grille de calcul, ce système apparaît comme complexe tant du point de vues des files d'attente que des politiques de priorité ou des systèmes de gestion des flots. De nombreux utilisateurs ont commencé naïvement à dupliquer les calculs afin d'avoir plus de chances de récupérer tous les résultats en un temps fini, surchargeant ainsi encore plus que nécessaire un système déjà beaucoup sollicité.

Gérer ces problèmes de latence a été une spécificité du groupe d'applications biomédicales car celles-ci peuvent se décomposer en de multiples sous-parties alors que d'autres communautés scientifique, comme par exemple en physique, décomposent beaucoup moins les calculs qui sont par contre beaucoup plus long, rendant ainsi la latence généralement négligeable par rapport au temps de calcul global.

Dans le cadre des travaux sur MOTEUR, il a par exemple été montré que grouper les tâches séquentielles ou optimiser la granularité des tâches permet de limiter les impacts des latences (Glatard, Montagnat, & Pennec, 2006). Dans (Glatard, Montagnat, & Pennec, 2007), les auteurs ont modélisé la latence afin d'optimiser le délai d'expiration des tâches pour se prémunir de valeurs extrême de latence. Ainsi, au bout d'un certain temps  $t_{\infty}$ , un travail non terminé correctement est re-soumis. Ce temps est calculé à partir de la distribution des latences qui a été



mesurée en soumettant sur EGEE, à intervalles constants, des travaux de durée quasiment nulle (commande `/bin/hostname` sur le noeud final). La latence est mesurée par la durée totale entre la soumission et le retour du résultat. A partir de la distribution des latences,  $t_\infty$  est calculé pour minimiser la latence totale incluant les re-soumissions.

J’ai étendu ces travaux par :

- l’acquisition de davantage de données
- l’utilisation de données des utilisateurs grâce à l’observatoire des grilles
- la modélisation plus fine de la latence en fonction du contexte d’exécution
- l’étude de différents algorithmes de re-soumission
- la différenciation dans la modélisation entre les travaux perdus et les erreurs pouvant survenir en un temps fini

**Dépendance du contexte d’utilisation** Dans (Glatard, Lingrand, Montagnat, & Riveill, 2007; Lingrand, Glatard, & Montagnat, 2009), nous avons montré que le modèle de latence utilisé impacte la valeur de  $t_\infty$  et que plusieurs améliorations peuvent améliorer la latence totale. En premier, il convient de mettre à jour régulièrement le modèle de latence mesuré. En second, plusieurs éléments contextuels (le *Resource Broker*, le site de calcul et le jour de la semaine) ont une influence mesurable. Notamment, on a effectué le partitionnement (*clustering*) des différentes distributions de latence afin d’utiliser une distribution différente par *cluster* et donc de temps de re-soumission  $t_\infty$  et montré ainsi une réduction de l’espérance de la latence.

**Utilisation de données des utilisateurs - Observatoire des grilles** Si l’utilisation de soumissions de travaux simples pour mesurer les latences était très simple, avoir des informations pertinentes sur tous les travaux soumis sur la grille qui permettent de découpler le temps d’exécution et la latence par exemple pour notre cas, n’était pas possible au départ. Une équipe de l’ICL (*Imperial College of London*) a commencé à proposer un outil permettant de recueillir toutes les traces des travaux, point de départ du développement ultérieur de l’Observatoire des Grilles<sup>§</sup> (Germain-Renaud et al., 2011).

En collaboration avec l’équipe de l’ICL, nous avons pu analyser de façons plus dense la latence sur une période de plusieurs mois entre septembre 2005 et juin 2007 (Lingrand, Montagnat, Martyniak, & Colling, 2009). Travailler avec une base de donnée conséquente (plus de 33 millions d’enregistrements) nécessite un long travail de pré-traitement : nettoyage des données, standardisation des échelles de mesure, homogénéisation des champs. Cette étude a montré qu’il était important de considérer le ratio de travaux abandonnés (*outliers*) (environ 16%) et celui de travaux se terminant par une erreur (environ 20%). L’information concernant une erreur est le plus souvent connue avant celle des travaux abandonnés. Un meilleur modèle évoluant avec le temps a montré son intérêt (Lingrand, Montagnat, Martyniak, & Colling, 2010).

**Prise en compte des erreurs et des travaux abandonnés** (Lingrand & Montagnat, 2010) Pour une variable aléatoire  $X$ , sa fonction de densité de probabilité (*pdf*) est notée  $f_X$  tandis que la fonction de distribution cumulative (*cdf*) est notée  $F_X$ .  $R$  est la latence d’un travail soumis qui se termine correctement,  $F$  le temps de détection d’une erreur et  $L$  la latence avec re-soumission sans délai.  $L$  dépend de la distribution des temps de détection des erreurs. En notant  $\rho$  la fraction de travaux abandonnés,  $\phi$  celle de travaux se terminant sur une erreur, il reste  $(1 - \rho - \phi)$  de travaux se

§. Grid Observatory <http://www.grid-observatory.org/>

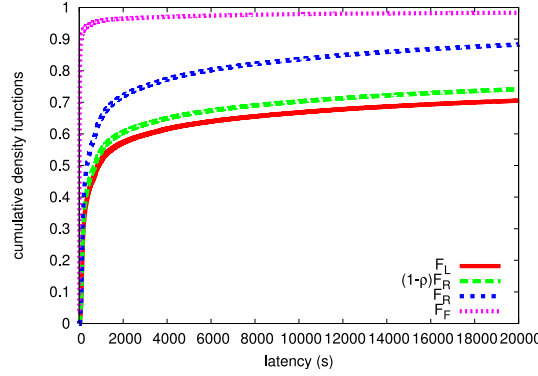


Figure 2.5 – Latence de détection des fautes ( $F_F$ ), latence des tâches réussies ( $F_R$ ), latence des tâches nécessitant des re-soumissions en cas d'échec ( $F_L$ ).

terminant correctement. Un travail rencontre une latence  $L < t$  si ce n'est pas un travail abandonné et que (i) soit le travail se termine par une erreur à  $t_0 < t$  avec une probabilité  $\phi f_F(t_0)$  et le travail re-soumis rencontre une latence de  $L < (t - t_0)$ , (ii) soit le travail se termine correctement avec une probabilité de  $(1 - \rho - \phi)$  et une latence  $R < t$  avec une probabilité  $P(R < t) = F_R(t)$ . La distribution cumulative de la latence  $L$  est alors définie récursivement par :

$$F_L(t) = (1 - \rho - \phi)F_R(t) + \phi \int_0^t f_F(t_0) \cdot F_L(t - t_0) dt_0$$

Cette équation peut être discrétisée par pas de une seconde qui est la précision des mesures. On considère également que aucun travail réussit n'a de latence nulle. Ainsi, on obtient :

$$F_L(0) = 0 \tag{2.1}$$

$$F_L(t) = \frac{1}{1 - \phi f_F(0)} \left[ (1 - \rho - \phi)F_R(t) + \phi \sum_{u=1}^{t-1} f_F(t - u) F_L(u) \right] \tag{2.2}$$

**Les stratégies de re-soumission** Les 3 stratégies sont décrites dans (Lingrand, Montagnat, & Glatard, 2009) et illustrées sur la figure 2.6 :

**SR (Simple re-soumission)** : Quand la latence dépasse un seuil ( $t_\infty$ ), la tâche est annulée et re-soumise

**MR (Multiples soumissions)** : Au départ,  $b$  copies de la même tâche sont soumises. Si l'une d'entre-elles se termine correctement avant  $t_\infty$ , alors toutes les autres sont annulées. Sinon, à  $t_\infty$ , à nouveau  $b$  copies de la même tâche sont soumises.

**DR (Re-Soumission avec délai)** : Au départ, une tâche est soumise. Si à  $t_0$  elle ne s'est pas terminée correctement, une copie de cette tâche est également soumise. Si à  $t_\infty$  la première tâche n'est pas terminée, elle est annulée. Et ainsi de suite jusqu'à ce qu'une tâche se termine correctement.

Les différents calculs permettant de calculer l'espérance du sur-coût d'exécution d'une tâche sur la grille (on ne compte pas le temps d'exécution sur le noeud final) sont détaillés dans (Lingrand, Montagnat, & Glatard, 2009) et conduisent aux résultats suivants :

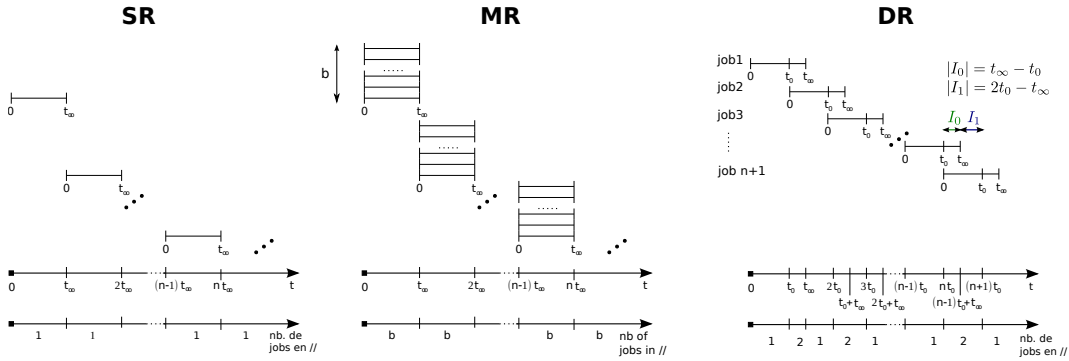


Figure 2.6 – Les trois types de re-soumissions étudiées : Re-Soumissions simples (SR), multiples (MR) ou avec délai (DR).

$$\begin{aligned}
 \text{SR : } E_J(t_\infty) &= \frac{1}{F_L(t_\infty)} \int_0^{t_\infty} (1 - F_L(u)) du \\
 \text{MR : } E_J(t_\infty) &= \frac{1}{1 - (1 - F_L(t_\infty))^b} \int_0^{t_\infty} (1 - F_L(u))^b du \\
 \text{DR : } E_J(t_0, t_\infty) &= \frac{1}{F_L(t_\infty)} \int_0^{t_\infty} u f_L(u) du \\
 &+ \frac{F_L(t_0)}{F_L(t_\infty)} \int_0^{t_\infty - t_0} u f_L(u) du + \frac{t_0}{F_L(t_\infty)} \\
 &+ t_0 \frac{F_L(t_\infty - t_0)}{F_L(t_\infty)} + t_0 \frac{F_L(t_0) F_L(t_\infty - t_0)}{F_L^2(t_\infty)} \\
 &- t_0 + \int_0^{t_\infty - t_0} u f_L(u) du \\
 &- \frac{t_0}{F_L(t_\infty)^2} \int_0^{t_\infty - t_0} f_L(u + t_0) \cdot f_L(u) du \\
 &- \frac{1}{F_L(t_\infty)} \int_0^{t_\infty - t_0} u f_L(u + t_0) \cdot f_L(u) du
 \end{aligned} \tag{2.3}$$

C'est cette espérance et sa variance (non présentée ici) que l'on souhaite minimiser tout en minimisant l'impact sur la charge supplémentaire sur la grille. On calcule la différence de charge sur la grille considérée comme le coût de changement d'algorithme de re-soumission. On compte le nombre de tâche en parallèle multiplié par le ratio entre le temps d'exécution global avec un algorithme de re-soumission et le temps d'exécution sans re-soumission :

$$\Delta_{cost} = N_{//} * \frac{E_J(\text{with } N_{//})}{E_J(\text{with } b = 1)} \tag{2.4}$$

Sans surprise, le coût de la stratégie de soumission multiple (MR) est le plus fort dès que le nombre de tâches vaut 2 mais plus nombre de tâches augmente, plus vite le résultat est obtenu. Néanmoins, si tout le monde utilisait cette méthode, la charge de la grille pourrait être tellement importante qu'au final personne n'y gagnerait. La stratégie avec délai (DR) permet de bien améliorer la latence par rapport à la stratégie simple (SR) sans trop augmenter l'impact sur la charge de la grille qui reste autour de 1.

Le résultat global est qu'il existe un algorithme optimum permettant de réduire la latence sans impact important sur la charge de la grille.

Ces différentes études ont ouvert la voie à un changement d'échelle dans les algorithmes de traitement d'images : plus de variabilité pouvait maintenant être étudié, plus de données allaient être utilisées pour valider ces algorithmes. A la fois la façon de traiter les données des grilles pour améliorer les re-soumission commençait à s'approcher de techniques d'apprentissage automatique mais aussi la voie était ouverte pour trouver de nouveaux algorithmes plus évolutifs en traitement d'images. J'ai alors commencé à m'intéresser à l'apprentissage automatique pour les algorithmes concernant les images.

Retour à cette thématique en partie : soumission du Projet Sage-HPC, NumPEx 2025. Responsable WP3.

# CHAPITRE 3

---

## Apprentissage automatique pour les données images

*A partir de 2012, j'ai commencé à étudier les algorithmes avec apprentissage automatique pour les images.*

---

---

<b>3.1</b>	<b>Gastronomie : classification subjective et incertitude dans les labels</b>	<b>17</b>
3.1.1	Annotation d'images par oculomètre . . . . .	18
3.1.2	Indépendance de l'estimateur GBIE aux classes cibles . . . .	20
3.1.3	Gestion de l'incertitude des labels . . . . .	21
<b>3.2</b>	<b>Classification de vidéos</b> . . . . .	<b>21</b>
3.2.1	Un descripteur de mouvement : les SINGLETs . . . . .	22
3.2.2	Une vidéo est un tenseur (2D+T) . . . . .	23
3.2.2.1	Représentation tensorielle d'un ensemble de données	26
3.2.2.2	HOPLS ( <i>High order partial least square</i> ) . . . . .	27
3.2.2.3	Extraction de représentation communes et individuelles CIFE/CIFA/COBE . . . . .	29
3.2.3	Déformations dynamiques de vidéos . . . . .	32

---





Une nouvelle ère commençait. En 2012, il n'était plus possible de publier avec des résultats sur quelques images : une validation sur des bases d'images devenait obligatoire. La base ImageNet ([Deng et al., 2009](#)), contenant 1000 classes de plus de 1,2 millions d'image, était connue depuis déjà quelques années pour la compétition de classification d'images ILSVRC (ImageNet Large Scale Visual Recognition Challenge) mais d'autres bases d'images continuaient d'être partagées chaque année comme par exemple les bases LifeCLEF concernant la bio-diversité.

Mes premiers travaux ont concerné les compétitions liés aux bases de données LifeCLEF. J'ai commencé avec la base de données de photos de plantes issue du projet PlantNet. Nous avons testé différents descripteurs d'images : sac de descripteurs SIFT en faisant varier le type de SIFT (monochrome ou couleur) mais aussi les descripteurs (SURF, HoG, LPB), les méthodes de clustering (nombres de partitions, affectation dure ou souple pour les partitions (*soft or hard assignment*)). Nous avons également incorporé des méta-données disponibles à la description. Ces travaux nous ont montré les limitations des approches de l'époque et la nécessité d'avoir accès à de meilleurs descripteurs : les descriptions profonds issus des réseaux convolutionnels (CNN). C'est aussi en 2012 qu'un CNN, AlexNet, a remporté la compétition ILSVRC.

Plus les différences entre les classes d'images sont faibles et la variabilité à l'intérieur d'une classe est forte, plus le descripteur image doit être suffisamment performant pour extraire les informations nécessaires à la classification ultérieure. En 2012, on était au début d'une longue suite de réseaux convolutionnels qui allaient révolutionner la classification d'images et ainsi l'extraction de descripteurs performants. Ces réseaux ont même dépassé les capacités de classification visuelle humaine sur ImageNet en 2016. Néanmoins, ce qui m'a intéressé, c'est que dans des applications pratiques, tout était loin d'être résolu.

Après cette période de montée en compétences sur l'apprentissage automatique, j'ai rejoint l'équipe Mind (Sparks, I3S) dirigé par Frédéric Precioso et j'ai participé à 3 projets de recherche en classification : subjective d'images, de vidéos et de nuages de points 3D.

### 3.1 Gastronomie : classification subjective et incertitude dans les labels

Ces travaux se situent dans le cadre du projet ANR VISIIR\* qui a financé la thèse de Stéphanie Lopez ([Lopez, 2017](#)) à l'encadrement de laquelle j'ai participé.

Ce projet avait pour but d'explorer de nouvelles méthodes pour l'annotation sémantique d'images en comblant le fossé sémantique entre des données images brutes et les concepts présents dans ces images. Les réseaux convolutionnels, même si performants, nécessitaient de grandes bases de données d'images et de nombreuses classes sont très différentes. Ici, on s'intéresse à une classification à grain fin car toutes les images concernent des plats gastronomiques.

Dans cette thèse, nous avons considéré un problème de classification binaire d'images pour lequel il n'existait pas de base d'annotation car cette dernière était subjective : une image correspond-elle au goût de l'utilisateur ou non ?

Ce sujet nous a conduit à étudier 2 questions : (i) comment construire une base de donnée minimale avec le minimum d'efforts et (ii) effectuer une classification subjective de données. Nous avons finalement concentré tous les efforts sur la base de données en utilisant un oculomètre (*eye-tracker*) mis à disposition par le partenaire industriel Tobii et en présentant aux utilisateurs des paires d'images afin qu'ils expriment, par le regard, leur préférence.

---

\*. ANR-13-CORD-0009

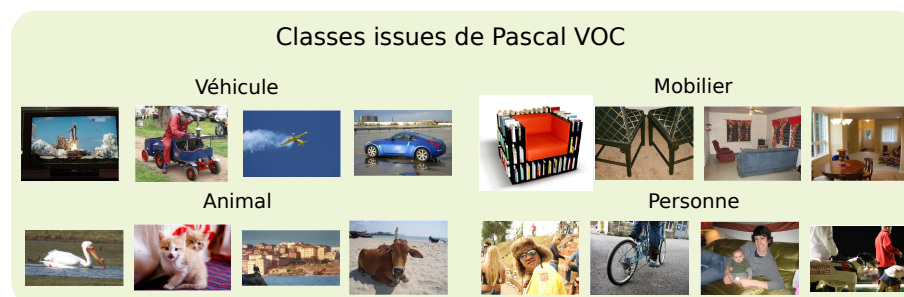


Figure 3.1 – Quelques images choisies dans PascalVOC.

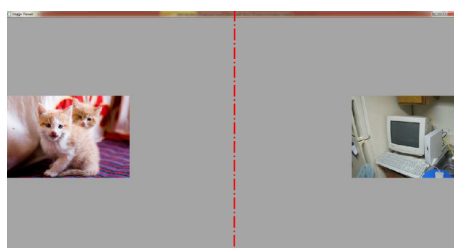


Figure 3.2 – L'utilisateur doit s'intéresser à une cible. Ici : les animaux. Il doit donc sélectionner l'image de gauche avec son regard.

### 3.1.1 Annotation d'images par oculomètre

Dans un premier temps, nous nous sommes intéressés à l'annotation par oculomètre en partant d'une partie d'une base de donnée bien connue pour laquelle les classes sont connues et sont objectives : Pascal VOC 2007<sup>†</sup> (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010). L'application visée était une classification binaire. Nous avons étudié quatre problèmes de classification binaire : Est-ce un véhicule ou non ? Est-ce un animal ou non ? Est-ce une personne ou non ? Est-ce un mobilier ou non ? (voir figure 3.1).

Dans (Lopez, Revel, Lingrand, & Precioso, 2015), nous avons défini un protocole d'acquisition du regard basé sur le paradigme de préférence visuelle : présentation de 2 images côte à côte horizontalement et détermination de la préférence par l'étude des saccades et fixations visuelles (voir figure 3.2).

Nous avons ensuite effectué différentes expériences avec 86 sujets sur deux sites : La Rochelle et Nice pour lesquels différents oculomètres ont été utilisés et avons enregistré les données brutes de ces oculomètres. Les sujets avaient des tâches très simples : chercher quelle image comporte un animal, un véhicule, un mobilier ou une personne et ce, avec 40 images par classe (le protocole précis est détaillé dans (Lopez et al., 2015)).

A partir de ces données, nous avons cherché un estimateur simple de la décision du sujet en fonction de caractéristiques des données oculométriques. Nous avons comme contrainte que la décision doit être prise rapidement (avant 1 seconde) afin de proposer d'autres données à anno-

<sup>†</sup>. <http://host.robots.ox.ac.uk/pascal/VOC/>

<b>n</b>	<b>caractéristiques dans la littérature</b>	<b>caractéristiques adaptées à notre cas</b>
1	line number	<i>idem</i>
2,3	max. size of pupilla on left, right image	<i>idem</i>
4	total fixation number (F)	<i>idem</i>
5	F/(F + number of saccades)	<i>idem</i>
6,7	left image : spread in x, y	<i>idem</i>
8,9	right image : spread in x, y	<i>idem</i>
10,11	both images : spread in x, y	<i>idem</i>
12	average distance between fixations	<i>idem</i>
13,14	left image : spread in x, y for fixations	<i>idem</i>
15,16	right image : spread in x, y for fixations	<i>idem</i>
17,18	both images : spread in x, y for fixations	<i>idem</i>
19,20	first and last seen image	first seen image and mean of $x$ at $t_0$
21	image label with maximal pupilla size	<i>idem</i>
22,23	first and last fixated image	first fixated image and mean of $x$ at $t_1$
24,25	duration of first and last fixation	<i>idem</i>
26,27	number of fixations during 1st and last visit	<i>idem</i>
28	total fixation duration	<i>idem</i>
29,30	number of fixations on left (right) image	<i>idem</i>

TABLE 3.1 – Caractéristiques du regard étudiées dans la littérature et adaptées à notre cas.

ter à l'utilisateur. Le cadre global concernait l'apprentissage actif permettant d'annoter moins de données mais les plus importantes.

Dans la figure 3.1, nous présentons les différentes caractéristiques du regard disponibles dans la littérature en 2015 dans la colonne du milieu.

Nous avons étudié la classification des images en utilisant ces critères oculométriques à l'aide d'un arbre de décision afin de mesurer également la pertinence des décisions. Les caractéristiques les plus mises en évidence étaient, par ordre décroissant : la dernière image vue (numéro 20), la dernière image fixée (numéro 23) et l'amplitude horizontale de positions pour l'image de gauche (6), celle de droite (8) ainsi que les deux images (10). Dans l'optique de prendre une décision rapide, on ne souhaitait pas attendre trop longtemps pour prendre la décision (le temps moyen d'observation des sujets est de 1840 ms). Les notions de dernière image et de dernière image fixée (caractéristiques 20 et 23) ne pouvaient donc pas être utilisées. Par contre, ayant remarqué l'importance de la position horizontale, on s'est intéressé à la position horizontale moyenne. En injectant cette valeur dans les caractéristiques, elle est devenue la caractéristique la plus pertinente. On a alors observé les profils de plusieurs sujets et couples d'images. Deux profils principaux sont ressortis et sont présentés figure 3.3. Sur cette figure on observe en premier que la décision ne peut pas être prise avant 480ms, période pendant laquelle la persiste la fixation initiale de la croix au milieu de l'écran (initialisation du regard). On ne peut pas non plus se satisfaire d'une seule valeur moyenne dans le cas de profils variés. Ainsi, deux valeurs de moyenne vont être utilisées à deux instants notés  $t_0$  et  $t_1 > t_0$ . On a décidé de limiter l'attente de la décision à 960ms ce qui conduit à une contrainte sur  $t_1$  :  $t_1 \leq 960$ . Selon les sites d'expérimentations, les valeurs optimales de  $t_0$  et  $t_1$  fluctuent légèrement avec une différence de l'ordre de 160ms. Les caractéristiques 20 et 23 ont

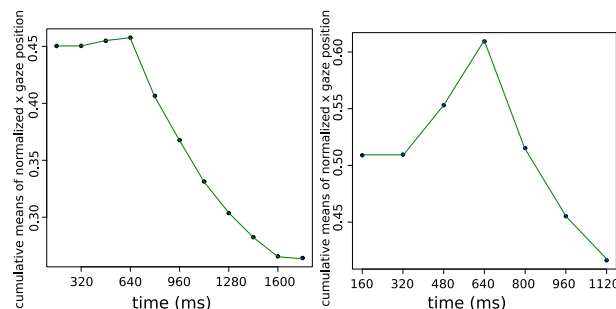


Figure 3.3 – Moyenne cumulée de la position horizontale du regard. La cible est l’image de gauche ( $x$  normalisé doit être inférieur à 0.5). On observe que selon le profil, des décisions doivent être prises différemment.

été remplacées par des moyennes cumulées à deux instants  $t_0$  et  $t_1$ . Toutes ces modifications sont présentées dans la dernière colonne du tableau 3.1. Avec ce nouveau jeu de caractéristiques, la qualité de la prise de décision est sensiblement inférieure (88%) à celle correspondant au premier jeu de données qui nécessitait d’attendre plus longtemps (95%) mais reste néanmoins acceptable.

Le résultat de ce premier travail, outre la base de données oculométriques rendue publiquement disponible, est l’établissement d’un estimateur de l’intention par le regard (GBIE : *Gaze-Based Intention Estimator*) dépendant de 2 paramètres  $t_0$  et  $t_1$ . A l’aide des deux populations de sujets (S1 : Nice et S2 : La Rochelle), nous avons pu montrer l’indépendance aux utilisateurs même si quelque fluctuations mineures peuvent être observées. Il convenait alors de vérifier que ces résultats étaient généralisables à d’autres problèmes de classification avec des classes plus proches et de domaines différents ainsi que

### 3.1.2 Indépendance de l’estimateur GBIE aux classes cibles

Nous avons extraits deux ensembles de 4 classes (voir figure 3.4) dans la base *UPMC Food-101* déjà étudiée par (X. Wang, Kumar, Thome, Cord, & Precioso, 2015) :

**F1** : carpaccio de boeuf, salade de betteraves, cannoli, glace

**F2** : apéritif, dessert, plat contenant du citron, plat contenant des fruits rouges

afin de mesurer l’indépendance des l’estimateur GBIE aux classes cibles. Lors des expériences de ces ensembles F1 et F2 (Lopez et al., 2016), les images non cibles étaient également des images de nourriture des classes non ciblées mais du même groupe de classes.

Ces ensembles de classes de nourriture présente des difficultés supplémentaires par rapport aux classes précédentes, et notamment F2 car par exemple, le concept de dessert dépend de la culture gastronomique d’origine. On a pu mesurer la difficulté de choix pour F2 notamment par le temps moyen pris pour une décision qui a augmenté de 266ms pour F2 par rapport à F1. Un compromis acceptable pour les hyper-paramètres est de  $t_0 = 800\text{ms}$  et  $t_1 = 960\text{ms}$ . La précision d’annotation pour F1 est alors de l’ordre de 81% mais chute à 55% pour F2. Il devient important de gérer les erreurs dans la labellisation si on veut apprendre un modèle de classification automatique à partir de ces données et annotations par le regard.



Figure 3.4 – Quelques images des ensembles de classes F1 et F2

### 3.1.3 Gestion de l’incertitude des labels

Nous avons élaboré un estimateur d’intention par le regard (GBIE), calculable en temps réel, indépendant de l’utilisateur et de la catégorie cible. Cette annotation implicite est meilleure qu’une annotation aléatoire mais reste incertaine. L’application finale étant une classification des images, nous allons utiliser une représentation des images issue d’un CNN et les labels incertains. Cette classification va prendre en compte l’incertitude sur les labels. Pour cela, nous avons adapté la méthode P-SVM proposée dans (Niaf, Flamary, Rouvière, Lartizien, & Canu, 2014) combinant classification SVM classique et régression SVM avec tolérance  $\epsilon$ . En entrée, cet algorithme nécessite de distinguer les labels les plus fiables des plus incertains. Nous avons testé différentes stratégies afin d’établir un critère de pertinence pour discriminer les labels les plus fiables, utilisés pour la classification, des labels les plus incertains, utilisés pour la régression. En plus de la mesure d’incertitude produite par notre GBIE, nous avons associé deux autres métriques : la confiance majoritaire et la représentativité de l’image. La confiance majoritaire correspond à la proportion de labels communs par GBIE tandis que la représentativité est fonction de la distance de la représentation profonde de la donnée à la fonction de décision SVM. La précision du P-SVM est évaluée dans différents contextes (centré utilisateur et validation par comité) et peut atteindre les performances d’un algorithme de classification standard entraîné avec les labels certains. Ces évaluations ont tout d’abord été menées sur un benchmark standard pour se comparer à l’état de l’art, et dans un second temps, sur une base d’images de nourriture (Lopez, Revel, Lingrand, & Precioso, 2017).

## 3.2 Classification de vidéos

Quand la thèse de Katy Blanc a débuté (Blanc, 2018), la classification d’images fixes avait beaucoup progressé. Les réseaux CNN tels que VGG16 ou bientôt la famille des ResNet et In-

ception affichaient de bonnes performances sur des bases de données comme ImageNet. Pour la classification de vidéo, il était possible de classer certaines actions par la présence d'objets liés aux actions sur les images fixes composant la vidéo. Par exemple, pour une vidéo d'équitation, il suffit de détecter une image avec un cavalier et un cheval. D'autres actions ne peuvent pas être reconnues à partir des images fixes : ce sont les actions caractérisées par le mouvement. Par exemple des actions spécifiques lors d'un match de football, différents signes de la langue des signes . . . . Cela explique pourquoi, à cette époque, le bond de progression en classification d'images fixes ne s'est pas répercuté sur la classification de vidéos : la dimension temporelle exigeait une attention particulière.

Parmi les différentes approches, on peut citer les descripteurs adaptés à la dimension temporelle comme les STIP (Laptev & Lindeberg, 2003), analogues au détecteur de Harris pour repérer les changements spatiaux ou les IDT (*Improved Dense Trajectories*) (H. Wang, Kläser, Schmid, & Cheng-Lin, 2011), composés d'un descripteur spatial local HOG, d'un descripteur spatial du flot optique HOF et d'un descripteur des changements spatiaux du flot optique MBH.

D'autres part, différentes architecture de réseaux de neurones se différenciaient principalement par (i) les entrées : série d'images couleur ou non, avec flot optique précalculé ou non, (ii), convolution image par image (2D) ou globale (3D), et (iii) la méthode d'agrégation temporelle, simple (moyenne, maximum) ou récurrente (RNN). Les réseaux de neurones récurrents introduisent une temporalité mais ils ont une mémoire limitée dans le temps. Il est à noter que combiner des descripteurs tels que IDT avec des réseaux de neurones permet d'augmenter la reconnaissance, montrant ainsi que certains aspects de la dimension temporelle ne sont pas captés par des réseaux neuronaux même profonds.

La dimension temporelle possède une élasticité propre, différente des dimensions spatiales. Elle peut être déformée localement : une dilatation partielle provoquera un ralentissement visuel de la vidéo sans en changer la compréhension, à l'inverse d'une dilatation spatiale sur une image qui modifierait les proportions des objets.

Dans ces travaux on s'est intéressé à la problématique d'une description robuste de vidéo en considérant l'élasticité de la dimension temporelle sous trois angles différents. Dans un premier temps, nous avons décrit localement et explicitement les informations de mouvements. Des singularités sont détectées sur le flot optique, puis traquées et agrégées dans une chaîne pour décrire des portions de vidéos. Nous avons utilisé cette description sur du contenu sportif. Puis nous avons extrait des descriptions globales implicites grâce aux décompositions tensorielles. Les tenseurs permettent de considérer une vidéo comme un tableau de données multi-dimensionnelles. Les descriptions extraites sont évaluées dans une tâche de classification. Pour finir, nous avons étudié les méthodes de normalisation de la dimension temporelle. Nous avons utilisé les méthodes de déformations temporelles dynamiques des séquences. Nous avons montré que cette normalisation aide à une meilleure classification.

### 3.2.1 Un descripteur de mouvement : les SINGLETs

Une start-up de Sophia, Wildmoka, nous a exposé sa volonté de créer automatiquement des résumés d'événements sportifs comme les matchs de football. Différentes méthodes sont pertinentes pour cette détection mais aucune n'est complètement satisfaisante. Par exemple, une hausse soudaine de l'activité des réseaux sociaux pendant un match peut correspondre à un événement important. Nous avons mis au point un descripteur de singularités du mouvement, SINGLET (Blanc, Lingrand, & Precioso, 2017), afin de rechercher des événements saillants dans des vidéos de



matchs sportifs. Ce descripteur permet de détecter des zooms, des ralentis ou autres moments saillants. Nous avons produit une base de données à partir de 4 matchs de football de la coupe du monde FIFA 2014 que nous avons annotés avec les buts, les fautes, les corners et les moments marquants afin d'en faire un résumé. Nous détectons correctement 88,2% des moments marquants à l'aide de cette base de données. Afin de mettre en évidence la généralisation de notre approche, nous testons notre système sur le match final du championnat du monde de handball 2015 sans aucun réentraînement, affinage ou adaptation.

Ce descripteur, SINGLET, correspond au mouvement des singularités des mouvements par l'étude du flot optique à différentes résolutions, suivi tout au long de la vidéo. Ce descripteur nous a permis de détecter les portions intéressantes des vidéos de matchs de football afin de construire automatiquement des résumés (détection de zoom, de pics d'activité ...).

Ce travail est inspiré de Kihl *et al* (Kihl, Tremblais, & Augereau, 2008) qui extrait des singularités du mouvement dans le domaine des fluides mécaniques. Les 2 composantes du flot optiques,  $U$  et  $V$  sont projetés sur l'espace des polynômes de Legendre de degré  $d$ . Ils sont ensuite exprimés dans la base canonique. En restreignant les approximations au premier degré, on obtient une expression du flot linéaire :

$$\begin{pmatrix} U \\ V \end{pmatrix} \simeq \mathbf{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \mathbf{b} \quad (3.1)$$

Les singularités apparaissent lorsque le flot optique s'annule donc en position  $(x_1 \ x_2)^T = -\mathbf{A}^{-1}\mathbf{b}$  et leur type (voir figure 3.5 dépend à la fois de la trace de  $\mathbf{A}$  et de  $\Delta(\mathbf{A})$  avec :

$$\Delta(\mathbf{A}) = (\text{tr}(\mathbf{A}))^2 - 4 \det(\mathbf{A}) \quad (3.2)$$

Pour détecter plusieurs éventuelles singularités en un même point, plutôt que d'augmenter le degré  $d$  d'approximation, nous avons préféré conserver le degré 1 mais effectuer plusieurs détections à différentes échelles spatiales. Un seuillage nous permet de garder les singularités de mouvement et éliminer celles qui concernent du bruit, par exemple de compression.

Les SINGLETs représente un suivi de singularités sur plusieurs images consécutives selon un critère combinant à la fois la proximité spatiale et la ressemblance des singularités. Un exemple de suivi est présenté figure 3.6 et un exemple de SINGLETs est proposé figure 3.7.

Les SINGLETs, caractéristiques haut-niveau des vidéos, nous ont permis de détecter des moments saillants dans des matchs de football afin de créer des résumés de ces derniers. Nous avons considéré les zooms, les ralentis ainsi que l'agitation globale. Concernant les zooms, indications fiables de moments saillants dans un match nous avons pu comparer notre méthode basée sur les SINGLETs à l'état de l'art du moment. Deux méthodes principales existaient alors pour les zooms : Estimation Globale du Mouvement (GME) (Ye, Huang, Gao, & Jiang, 2005) et celle de Duon (Duan, Xu, Tian, Xu, & Jin, 2005) que nous avons surpassé (voir figure 3.8). Nous avons également de bonnes détections sur d'autres éléments comme par exemple les ralentis.

Nous avons ainsi montré l'importance de la description du mouvement. Afin de prendre en compte à la fois les informations de mouvements mais aussi de contenu d'image, nous nous sommes ensuite intéressés à la vidéo comme un tenseur en 3 dimensions afin d'étudier le couplage entre les informations spatiales et temporelles.

### 3.2.2 Une vidéo est un tenseur (2D+T)

Nous avons étudié les tenseurs afin d'extraire les relations entre les dimensions spatiales et temporelles. Cette étude a été motivée par l'intérêt des mouvements dans la reconnaissance d'ac-

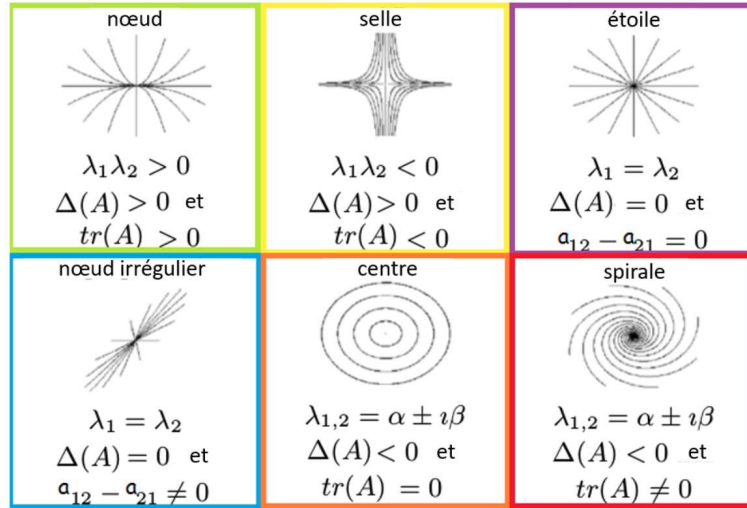


Figure 3.5 – Classification des singularités selon les valeurs de  $A$  (illustration construite à partir d’une illustration de (Kihl et al., 2008))

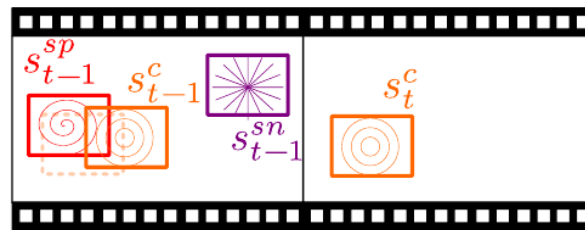


Figure 3.6 – Deux flots optiques successifs : recherche d’une correspondance pour la singularité  $s_t^c$  parmi les singularités du flot optique précédent. Vis à vis du ratio de chevauchement,  $s_{t-1}^{sp}$  et  $s_{t-1}^c$  sont candidates alors que  $s_{t-1}^{sn}$  ne l’est pas. La singularité correspondante est la singularité la plus proche en fonction de la position et du type. Il s’agit ici de :  $s_{t-1}^{sp}$ . Les types de singularités  $sp$ ,  $sn$  et  $c$  correspondent respectivement à spirale, étoile et centre.



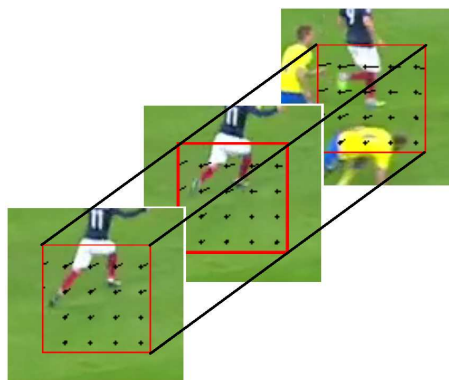


Figure 3.7 – SINGLETs : illustration du suivi des singularités extraites du flot optique sur 3 images consécutives dans un match de foot. C'est une singularité spirale comme cela peut être vu dans le flot.

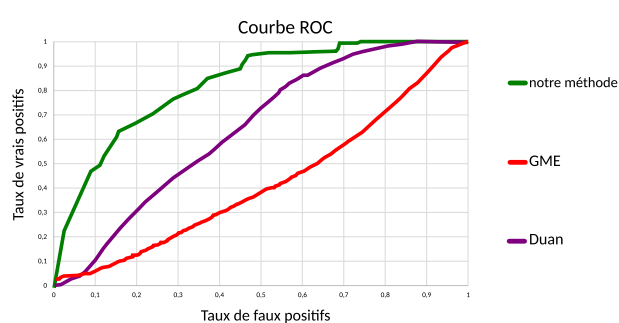


Figure 3.8 – Courbes ROC pour chaque méthode pour la détection de zoom : notre méthode surpasse la méthode de Duan (Duan et al., 2005) et la méthode GME (Ye et al., 2005).

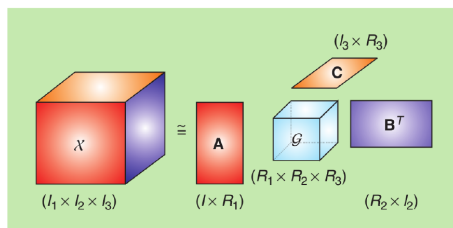


Figure 3.9 – Décomposition de Tucker d’un tenseur du troisième ordre :  $\mathcal{X} = \mathcal{G} \times_1 A \times_2 B \times_3 C$ . Les espaces des colonnes A, B et C représentent les sous-espaces de la données pour les trois modes. Le tenseur  $\mathcal{G}$  représente les interactions complexes possibles parmi les composantes du tenseur (illustration de (Cichocki et al., 2015))

tions mais aussi par des travaux antérieurs qui ont exploré différemment les dimensions tensorielles ainsi que des outils de manipulation de tenseurs qui ont vu le jour. Par exemple, dans (Lopes, Oliveira, de Almeida, & de A. Araujo, 2009) les auteurs ont montré que des descripteurs 2D, par exemple TIF, de coupes 2D du tenseur 3D (X et T, Y et T, X et Y) pouvaient apporter des informations pertinentes en comparaison de descripteurs purement 3D (X,Y,T) comme les STIPs (Laptev & Lindeberg, 2003).

La décomposition des tenseurs selon HOSVD (De Lathauwer, De Moor, & Vandewalle, 2000), figure 3.9, est une décomposition de Tucker avec des contraintes d’orthogonalité, permettant de rechercher des valeurs propres de données multidimensionnelles. Il est ainsi possible de généraliser des outils initialement destinés à des matrices (tenseurs d’ordre 2) vers des tenseurs d’ordre supérieurs. Parmi ces outils, la PCA est généralisée en TPCA (*Tensor Principal Component Analysis*) permettant de réduire les dimensions des tenseurs initiaux.

De même, des méthodes de classification SVM ont été généralisée à des données d’ordre supérieur (Kotsia, Guo, & Patras, 2012) : on cherche alors l’espace vectoriel le plus discriminant pour classer des tenseurs.

Cependant, les tentatives de généralisation des classifieurs linéaires (SVM, LDS) aux ordres de supérieurs, éventuellement couplés à des réductions de dimensions linéaires (TPCA), sont gourmandes en calculs et nécessitent l’introduction de non linéarité pour concurrencer l’état de l’art.

### 3.2.2.1 Représentation tensorielle d’un ensemble de données

Le tenseur peut représenter une donnée comme vu précédemment mais également un ensemble de données.

Vasilescu et al. furent les premiers à populariser l’application de la TPCA sur du contenu multimédia grâce à leur analyse des matrices facteurs sur une base de données de visages : TensorFaces (Vasilescu & Terzopoulos, 2002). Dans cet article, les auteurs construisent un tenseur  $\mathcal{D}$  stockant la base d’images de sorte qu’une dimension représente la donnée (les pixels) et les autres leurs attributs (le sujet, la luminosité, l’expression faciale et la position du sujet). Ce tenseur  $\mathcal{D}$  est d’ordre 5 et de taille 28x5x3x3x7945.

La décomposition TPCA est ainsi de la forme :  $\mathcal{D} = \mathcal{Z} \times_1 U_{\text{sujet}} \times_2 U_{\text{point de vue}} \times_3 U_{\text{illumination}} \times_4 U_{\text{expression}} \times_5 U_{\text{pixels}}$ .

Les auteurs ont choisi un tenseur noyau  $\mathcal{Z}$  de même taille que  $\mathcal{D}$ . Par conséquent, les matrices facteurs sont toutes carrées. Ils ont ensuite analysé les matrices facteurs extraites et le tenseur

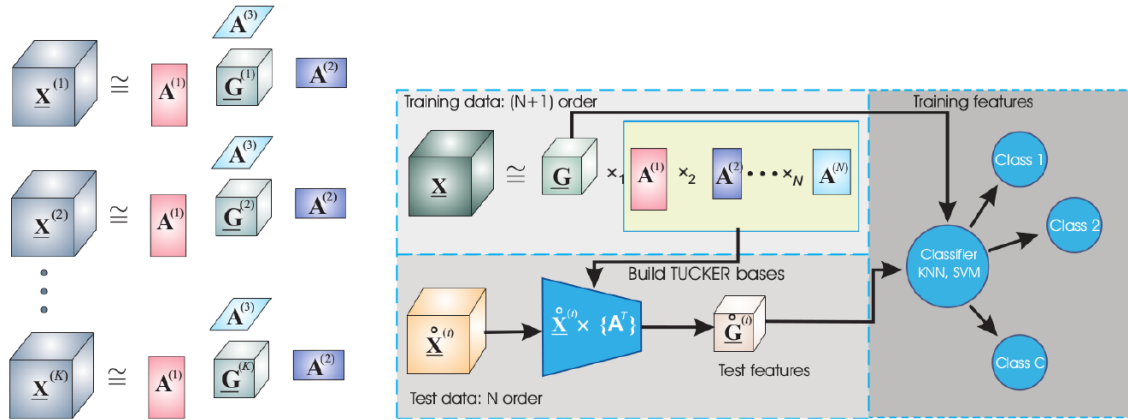


Figure 3.10 – (a) Illustration d'une extraction de représentation par TPCA. (b) Le système classique de classification par TPCA sur le tenseur contenant l'ensemble des données (illustration provenant de (Phan & Cichocki, 2010)).

noyau.  $U_{pixels}$  contient les vecteurs propres liés aux dimensions spatiales, correspondant à ceux d'une PCA réalisée en dimension 2 sur l'ensemble des images, sans prendre en compte les autres propriétés.  $\mathcal{Z} \times_5 U_{pixels}$  est présenté comme le tenseur permettant d'analyser les vecteurs propres selon les modes en ajoutant la dimension spatiale grâce à  $U_{pixels}$ . On peut également obtenir les vecteurs propres spécifiques à un point de vue en multipliant  $\mathcal{Z} \times_5 U_{pixels}$  par la colonne de  $U_{point\ de\ vue}$  correspondant au point de vue choisi. Dans une perspective de classification selon le sujet, les 45 vecteurs propres contenus dans  $\mathcal{B} = \mathcal{Z} \times_2 U_{point\ de\ vue} \times_3 U_{illum} \times_4 U_{express} \times_5 U_{pixels}$  forment la base de chaque combinaison d'illumination, point de vue et expression ce qui permet d'exprimer un sujet  $p$  avec n'importe quelle illumination, point de vue et expression. Pour déterminer la classe d'un nouveau sujet, toutes les projections de ce dernier sur tous les points de vue, illumination et expression sera calculé et le plus proche voisin permet de déterminer sa classe. Cependant, cette approche nécessite de disposer de toutes les variations de point de vue, illumination et expression pour tous les sujets et ne permet pas non plus d'utiliser plusieurs images du même sujet dans les mêmes conditions.

Par la suite TPCA sera utilisée comme une projection de la donnée multi-dimensionnelle et c'est le noyau qui sera la représentation de la donnée. Cette utilisation de la PCA rejoint l'utilisation de la décomposition de Tucker. Différentes approches, comme (Phan & Cichocki, 2010), ajoutent des contraintes sur la décomposition afin d'être discriminants selon leur classe ou labels.

La méthode TPCA est toujours employée sur des bases contenant peu de données, elles-mêmes étant de petites tailles et très corrélées spatialement : des visages, des mains ... En effet, la décomposition de Tucker demande d'effectuer des SVD sur les matricisations de ce tenseur contenant toute la base. Il est donc préférable d'avoir un tenseur initial de taille raisonnable.

### 3.2.2.2 HOPLS (High order partial least square )

Cette méthode a été introduite par (Zhao et al., 2013) comme une méthode généralisée de régression multi-linéaire. L'objectif est de mettre en relation la décomposition de deux tenseurs



$\mathcal{Y} \in \mathbb{R}^{N \times 10}$ . Puis la décomposition HOPLS de ces tenseurs est calculée de sorte à optimiser l'approximation par rapport aux exemples d'apprentissage :

$$\begin{aligned}\mathcal{X} &= \sum_{r=1}^R \mathcal{G}_r \times_1 t_r \times_2 P_r^{(1)} \times_3 P_r^{(2)} \times_4 P_r^{(3)} + \mathcal{E}_R \\ \mathcal{Y} &= \sum_{r=1}^R \mathcal{D}_r \times_1 t_r \times_2 Q_r^{(1)} + \mathcal{F}_R\end{aligned}$$

où les vecteurs latents sont notés  $t_r \in \mathbb{R}^N$  et les matrices facteurs  $P_r^{(i)} \in \mathbb{R}^{20 \times i} (1 \leq i \leq 3)$  et  $Q_r^{(1)} \in \mathbb{R}^{91}$ . En phase de test, les noyaux,  $\mathcal{G}_r$  et  $\mathcal{D}_r$  ( $1 \leq r \leq R$ ), ainsi que les matrices facteurs liées aux modes provenant de la donnée, soit  $P_r^{(i)}$  et  $Q_r^{(1)}$ , qui correspondent aux dimensions spatiales, à la dimension temporelle et à la dimension des classes sont réutilisés. Uniquement les vecteurs latents  $t_r$  sont calculés en test. Les vecteurs latents sont d'ailleurs également des matrices facteurs mais cette fois liées au mode 1 qui est de dimension  $N$  et qui correspond à la dimension des numéros d'échantillons, soit la seule dimension partagée entre les tenseurs initiaux à lier. Pour une nouvelle vidéo  $x$ , on utilise les éléments fixes de la première décomposition pour estimer ses vecteurs latents  $t_r$ . Puis on utilise les éléments fixes de la seconde décomposition pour estimer le vecteur de label  $y$ .

Les hyper-paramètres que nous avons étudié sont  $R$ , le nombre de vecteurs latents,  $L_1$ ,  $L_2$ ,  $L_3$ , les dimensions du tenseur noyau lié aux données ( $L_1$  correspond à la dimension temporelle) et  $K_1$ , la dimension du tenseur noyau lié aux labels. Chaque modèle a été défini par les 4 valeurs  $L_1$ ,  $L_2$ ,  $L_3$  et  $K_1$  tandis que  $R$ , le nombre de vecteurs latent, a été choisi comme la plus petite valeur pour laquelle l'erreur de classification a atteint un minimum.

Bien qu'HOPLS extrait des filtres permettant de lier la donnée et son label, il n'atteint pas des taux de reconnaissance suffisants sur une base de vidéos simples. Une contrainte récurrente des méthodes tensorielles est le besoin de considérer, lors de la décomposition, tous les échantillons de la base dans un même tenseur ; ce qui engendre rapidement des problèmes de stockage selon la taille de la base. C'est pourquoi nous avons redimensionné les données en taille 20x20x20. Dans le but d'expérimenter l'extraction d'information implicite par les méthodes tensorielles sur des bases de données plus fournies, nous passons à l'expérimentation de la méthode CIFA. Contrairement à HOPLS, cette méthode n'explore qu'une classe à la fois et demande ainsi moins de stockage.

### 3.2.2.3 Extraction de représentation communes et individuelles CIFE/CIFA/COBE

Cette méthode a été créée dans le but d'exploiter la nature liée de groupes de données et de leurs dimensions. Elle consiste en une décomposition sous la forme de facteurs communs et de facteurs individuels. En organisant les modes pour que ce soit le premier qui soit partagé, on obtient, pour une donnée  $n$  d'ordre 3 :

$$\mathcal{X}_n = \mathcal{G}_n \times_1 A^{(1,n)} \times_2 A^{(2,n)} \times_3 A^{(3,n)}$$

où la décomposition du premier mode est  $A^{(1,n)} = [\bar{A}^{(1)} \check{A}^{(1,n)}]$ . On note que le premier terme de cette décomposition ne comporte plus d'indice  $n$  car il est commun à l'ensemble des données.

A partir de l'ensemble  $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N\}$ , on obtient  $C$  éléments communs, qui, après des étapes de normalisation sont stockés dans  $\bar{F}$ . Cette méthode de classification est à préférer pour des données ayant plusieurs attributs pour classer les exemples en sous-catégories (comme

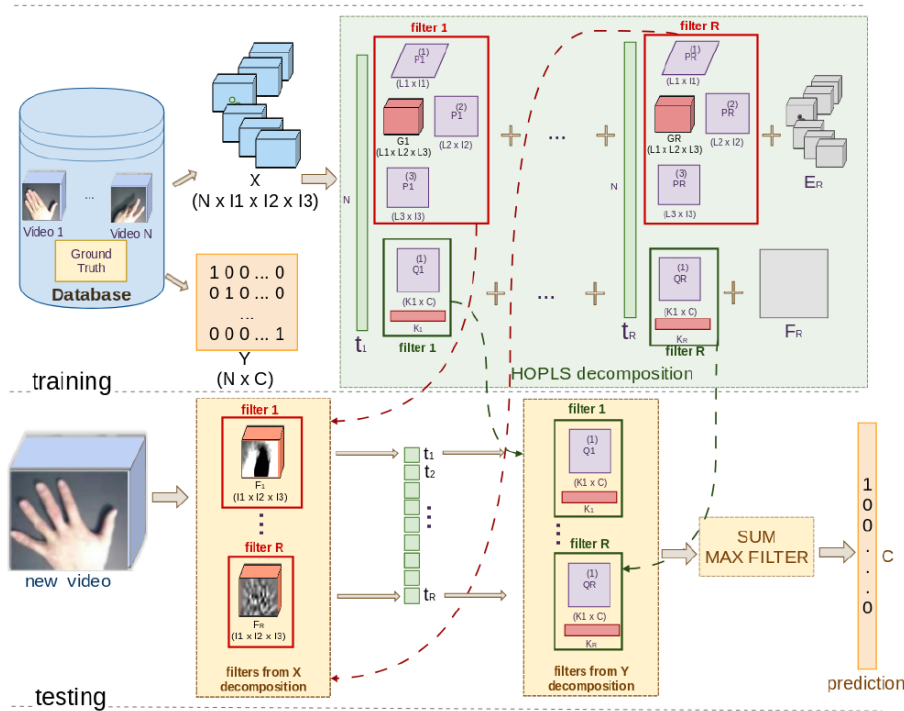


Figure 3.12 – Système d’utilisation d’HOPLS sur Cambridge dans un objectif de prédiction.

dans (Vasilescu & Terzopoulos, 2002)) même si elle peut s’adapter à la classification classique. Chaque classe est représentée par quelques représentants. A l’arrivée d’un nouvel élément, celui-ci n’est pas projeté mais comparé directement aux représentants de chaque classe afin de déterminer sa classification. La figure 3.13 présente un exemple sur une base de visage.

Nous avons effectué des expérimentations sur différentes bases : Yale (la base utilisée dans la publication d’origine), Cambridge mais aussi d’autres bases de vidéos d’actions : KTH, UCF101 et IsoGD. Les résultats ont été intéressants sur Cambridge mais décevants sur les bases d’actions humaines. Les vidéos de mains de la base Cambridge sont très normalisées alors que pour les autres bases d’actions, les sujets filmés se déplacent. On retrouve ces variations spatiales dès les premiers représentants communs extraits par l’aspect texturisé des frames. En visualisant ces représentants vidéos, on perçoit un mouvement régulier et continu de l’action représentée.

Il semble que l’alignement spatial et temporel des données soit une étape nécessaire avant l’extraction des éléments communs par CIFE. D’ailleurs, les auteurs n’ont évalué leur méthode que sur des données très corrélées spatialement. Cet effet est sûrement dû au caractère linéaire de la méthode CIFE. Un autre élément est que cette méthode se concentre sur les variations à l’intérieur des classes. Ajouter de la discrimination entre classes pourrait aider également à la classification.

Ces méthodes, HOPLS et CIFA, n’ont pas apporté les améliorations visées en classification de mouvement. Même si la séparation de l’influence de caractéristiques telles que l’illumination est intéressante pour la classification, on a pu observer que la nature linéaire des décompositions les rendent sensibles aux variations spatiales et temporelles. Les méthodes d’optimisation globale gèrent également mal le bruit et les données erronées. Cependant, ces méthodes ont pourtant un

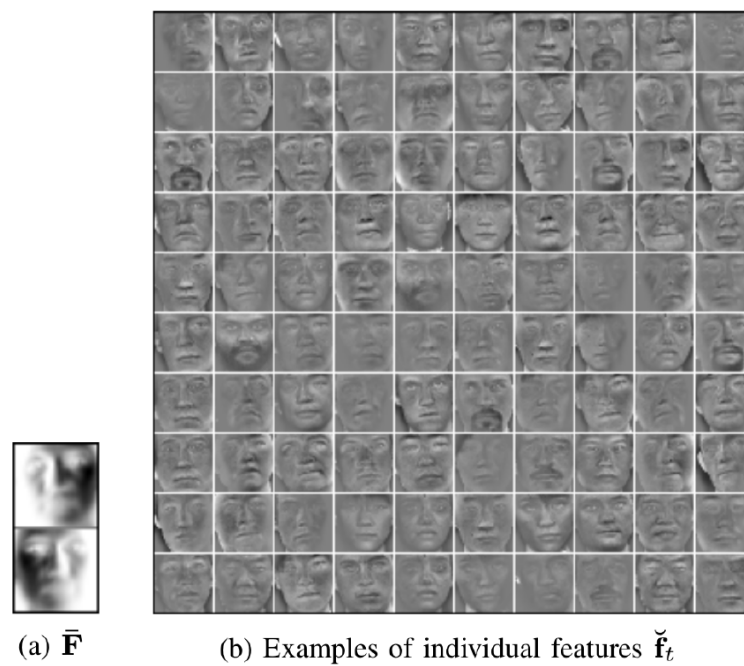


Figure 3.13 – Extraction de CIFE sur une base de données de visages : (a) éléments communs et (b) éléments individuels obtenus en enlevant les éléments communs (illustration de (G. Zhou et al., 2016)).

pouvoir de compression, de complétion, et même d'extraction de connaissance certain. Ces méthodes doivent être donc utilisées sur des données normalisées spatialement et temporellement. Dans la section suivante, nous allons donc nous intéresser à la normalisation temporelle de la vitesse d'exécution sur un ensemble de données d'une même classe.

Avec le recul des années, d'autres approches ont tenté de séparer des caractéristiques de données images comme les autoencodeurs variationnels, certains contraignant la représentation latente au minimum ( $\beta$ -VAE (Higgins et al., 2017)) afin de correspondre à des caractéristiques (orientation du visage, éclairage, expression, âge...) ou les réseaux adversaires génératifs (GAN). Plus récemment, différentes approches tentent d'exhiber des concepts depuis des images afin de représenter des images et d'expliquer les décisions des algorithmes de classification. Les principaux avantages de ces dernières approches basées sur des réseaux de neurones profonds sont (i) la prise en compte de non-linéarité, (ii) la possibilité d'utiliser de grandes bases de données et (iii) les méthodes d'optimisation permettant de limiter l'impact du bruit et des données erronées.

### 3.2.3 Déformations dynamiques de vidéos

Dans cette partie, nous cherchons à faciliter la classification des vidéos en réduisant la variabilité au sein d'une même classe par la réduction de l'élasticité temporelle. La méthode de base utilisée pour l'alignement temporel des séquences est DTW (*Dynamic Time Warping*), décrite figure 3.14. DTW permet de déterminer les transformations temporelles qui maximisent la corrélation entre les représentations de deux séquences vidéos ainsi transformées.

DTW minimise :

$$\min_{\{p_x, p_y\} \in \Psi} J_{DTW} = \sum_{t=1}^l \|x_{p_t^x} - y_{p_t^y}\|^2$$

où  $l$  est la longueur de la vidéo alignée,  $p_x \in \{1 : n_x\}^l$  et  $p_y \in \{1 : n_y\}^l$ . Le  $i$ ème élément de  $X$ ,  $x_i$ , est aligné avec le  $j$ ème élément de  $Y$ ,  $y_j$ , si il existe un instant  $t$  tel que  $p_t^x = i$  et  $p_t^y = j$ .

Cette méthode a été étendue et généralisée à d'autres cas. Notamment, grâce à l'analyse de corrélations canoniques (CCA), CTW utilise la projection linéaire des données dans un espace latent maximisant la corrélation. Cette méthode est nommée la déformation temporelle canonique (CTW). Cette projection linéaire est un moyen de hiérarchiser l'importance de chaque composant de la représentation dans le calcul de la corrélation et ainsi de différencier ce qui est corrélé de ce qui ne l'est pas. DCTW est une extension de CTW en remplaçant les projections linéaires par des projections non-linéaires modélisées par des réseaux de neurones. Afin de pouvoir traiter plus que des paires de vidéos, GCTW (F. Zhou & De la Torre, 2016) modélise la corrélation d'un ensemble de séquences par la somme de la corrélation de chaque paire.

Soit un ensemble de  $m$  séquences,  $\{X_i\}_{i=1}^m$ , GCTW cherche pour tout  $X_i = [x_1^i \dots x_{n_i}^i] \in \mathbb{R}^{d_i \times n_i}$  une transformation spatiale linéaire  $V_i \in \mathbb{R}^{d_i \times d}$  et une transformation temporelle non-linéaire  $W_i = W(p_i) \in [0, 1]^{n_i}$  paramétrée par  $p_i \in 1 : n_i^l$ , telle que les séquences en sortie  $V_i^T X_i W_i \in \mathbb{R}^{d \times l}$  sont alignées les unes aux autres, minimisant :

$$\begin{aligned} \min_{\{V_i\}_{i \in \Phi}, \{p_i\}_{i \in \Psi}} J_{gctw} &= \sum_{i=1}^m \sum_{j=1}^m \frac{1}{2} \|V_i^T X_i W_i - V_j^T X_j W_j\|_2^F \\ &+ \sum_{i=1}^m (\phi(V_i) + \psi(p_i)) \end{aligned} \quad (3.3)$$



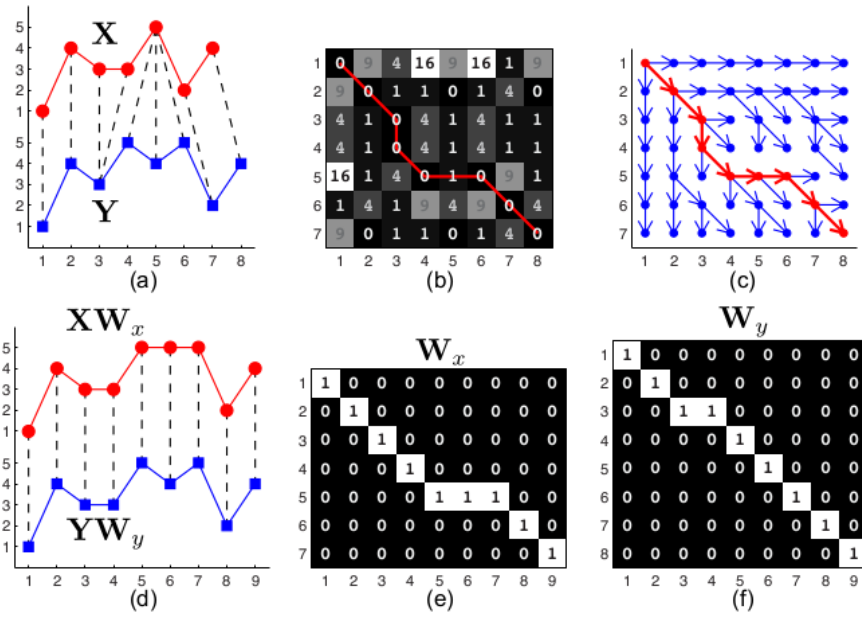


Figure 3.14 – Algorithme DTW pour aligner deux séquences. (a) Deux séquences 1D ( $n_x = 7$  et  $n_y = 8$ ) et l'alignement optimal par DTW illustré en pointillés. (b) Matrice des distances euclidiennes entre les éléments des séquences, la courbe rouge est le chemin optimal ( $l = 9$ ). (c) La pratique de la programmation dynamique illustrée par les possibilités de déplacements contraignant l'optimisation. (d) Le résultat des déformations sur les signaux. (e) et (f) Les matrices de déformation temporelle de chaque signal qui construites à partir des chemins  $p$  (illustration provenant de (F. Zhou & De la Torre, 2016)).

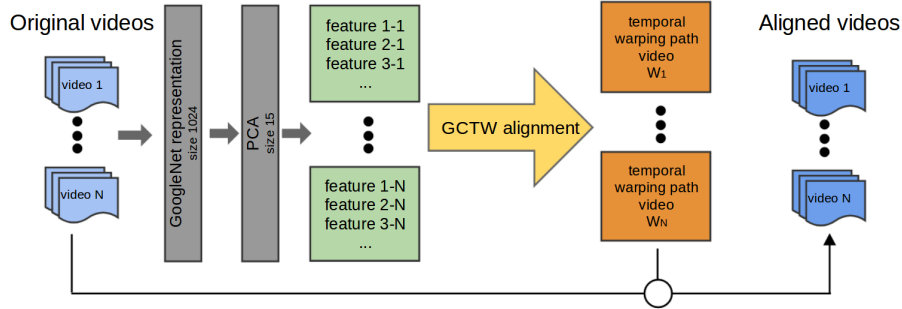


Figure 3.15 – Notre algorithme d’alignement de vidéos en utilisant GCTW

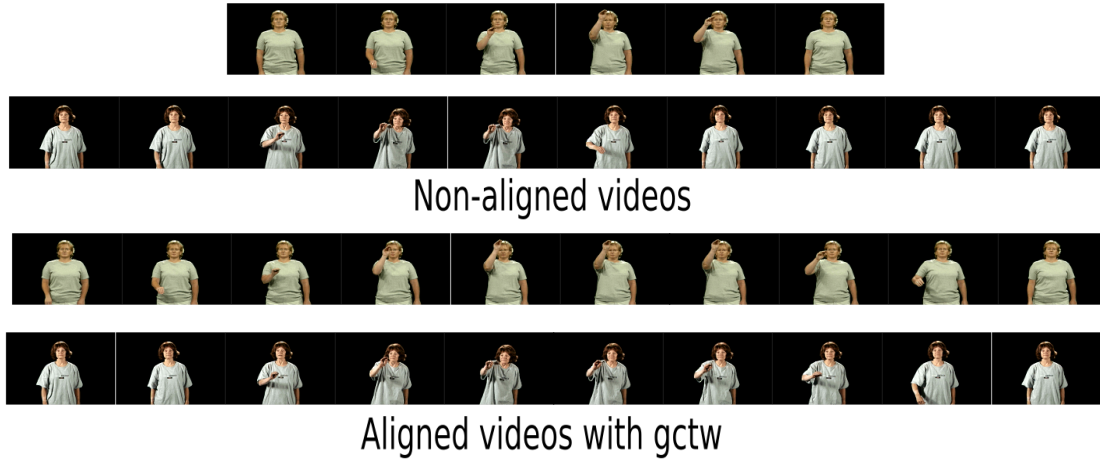


Figure 3.16 – Deux exemples de vidéos du même signe de la base ASL. En haut : avant alignement. En bas : après alignement. La première vidéo est une des plus courtes de la base ASL.

où  $\phi(\cdot)$  et  $\psi(\cdot)$  sont des termes de régularisation des transformations spatiales ( $V_i$ ) et temporelles ( $W_i$ ). Les contraintes d’orthogonalité ont été omises. Les auteurs de GCTW ont choisi de modéliser les déformations temporelles  $p$  par une combinaison non négative de déformations monotones prédéfinies :  $p = Qa$

Dans (Blanc et al., 2019), nous avons utilisé la partie temporelle de GCTW et l’avons appliquée pour aligner des vidéos du langage des signes de deux bases de données : ASL[REF 25] et IsoGD (Wan et al., 2016) (voir figure 3.16). L’algorithme que nous avons mis en place est illustré par la figure 3.15. Pour représenter les vidéos, nous avons choisi d’extraire les caractéristiques d’un CNN, GoogleNet, pour chacune des frames (1024 paramètres) suivi d’une PCA pour réduire encore la taille des données. Les séquences de chaque classe sont alignées par GCTW en utilisant uniquement la composante temporelle  $W_i$ . Chaque donnée  $X_i$  est remplacée par  $X_i W_i$ . Ainsi, l’élasticité temporelle est calibrée pour chaque classe.

Classification	ASL		IsoGD	
protocole	préc. top-1	préc. top-5	préc. top-1	préc. top-5
Baseline	76.7	96.2	45	89.05
Protocole 1	14.5	37.9	41.65	71.03
Protocole 2	91.7	98.7	81.27	97.01
Protocole 3	50.4	92	81.34	97.11

TABLE 3.2 – Précisions *Top-1* et *Top-5* sur les bases ASL et IsoGD en test, selon les différents protocoles.

Nous avons mesuré l'intérêt d'un tel alignement en comparant les résultats de classification d'un réseau de neurone C3D (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) avec ou sans alignement selon différents protocoles :

**base :** apprentissage et test sur les données non alignées

**protocole 1 :** apprentissage sur les données alignées et test sur les données non alignées

**protocole 2 :** apprentissage et test sur les données alignées (on suppose connaître la classe des données de test pour les aligner avec la bonne classe)

**protocole 3 :** apprentissage sur les données alignées et test sur les données alignées par rapport à chacune des classes (on ne fait pas de supposition pour la classe des données de test : on essaie chacune des classes)

Les résultats de classification sont présentés dans le tableau 3.2. Tout d'abord, on remarque que le protocole 1 fait chuter la reconnaissance, comparé à la référence, que ce soit sur ASL ou sur IsoGD. Le fait que C3D ait des difficultés à reconnaître des actions exécutées à des vitesses non-observées durant l'entraînement indique que la variation temporelle entrave effectivement la classification par des réseaux de neurones convolutionnels vidéo. Le protocole 2, même s'il n'est pas réaliste en pratique car il nécessite de connaître la classe par rapport à laquelle calculer l'alignement, montre l'intérêt de l'alignement temporel pour la classification. Le protocole 3, variation réaliste du protocole 2 n'a pas les mêmes résultats sur les 2 bases vidéos. En effet, la base ASL est très cadrée (même fond, même vêtements) par rapport à la base IsoGD qui présente des éléments différents. L'alignement GCTW ne prenant en compte qu'une classe à la fois, la variabilité inter-classes a été réduite dans le cas des données ASL, les éléments pris en compte pour l'alignement étant probablement les mêmes pour les classes confondues. On montre ainsi qu'éliminer l'élasticité temporelle intra-classes permet d'améliorer la classification à condition d'une variabilité suffisante dans les vidéos.

L'ajout d'une normalisation indépendante de la donnée et de sa classe et l'ajout d'un critère de discrimination sont des améliorations envisagées. Notre travail préliminaire sur l'extension du principe du réseau de transformation spatiale (Spatial Transformer Network) (Jaderberg, Simonyan, Zisserman, & Kavukcuoglu, 2015) à un réseau de transformation temporelle semble prometteur. Le réseau de transformations spatiales est en fait un réseau contenant un module peu profond servant à une transformation spatiale. Ce module, appelé module spatio-temporel (ST), est composé de deux parties : une permettant de localiser des données dans l'entrée et de prédire la transformation, et une permettant d'effectuer cette transformation spatiale sur l'entrée. La partie de localisation du module peut être composée de couches convolutionnelles ou fully-connected

mais doit se terminer par une couche de régression pour pouvoir prédire au mieux les valeurs de la transformations  $\theta$ . C'est à cette étape que les informations communes à la classe et nécessaires à l'alignement seront sélectionnées. La deuxième partie du module concerne la transformation de la donnée par les paramètres de déformation  $\theta$  prédits. Ce module est appliqué aux cubes de sorties de la couche précédente. Cette partie du réseau peut donc être placée autant au début du réseau directement sur les images qu'après quelques couches de neurones sur les cartes de descriptions (feature maps). L'idée est alors de s'inspirer de ce réseau pour construire un réseau de normalisation temporelle.

## Apprentissage profond

L'apprentissage est devenu profond quand il est devenu possible de combiner la représentation des données et les tâches d'apprentissage finales (classification, régression, segmentation ...). Les fonctions mathématiques apprises sont devenues suffisamment complexes pour prendre en compte la représentation de données diverses et les tâches visées.

Les premiers travaux ont concernés, dans le domaine des images, principalement des tâches de classification entre des classes bien différentes. Plus le domaine a progressé, plus on a pu attaquer des problèmes plus difficiles : de la classification fine (des classes proches, par exemple dans CUB ou LifeCLEF), des classes déséquilibrées, des problèmes avec peu de données ....

Ce qui m'intéresse c'est de partir d'applications concrètes et d'en extraire des problématiques d'apprentissage nouvelles afin d'apporter de nouvelles méthodes non pas pour améliorer les résultats sur des bases (sur-)exploitées mais pour permettre à des utilisateurs de différents domaines d'application de résoudre des problèmes grâce à l'apprentissage automatique.

Dans les trois prochaines parties, nous allons nous intéresser (i) à la biologie marine avec de vraies photos et vidéos en conditions réelles afin de proposer aux biologistes des outils leur permettant de suivre l'évolution de la biodiversité en fonction des éléments qui sont pertinents pour ces études et non pas des métriques standards d'apprentissage automatique, (ii) à l'archéologie en proposant notamment une aide à la détermination de l'espèce ou du genre en anthracologie ainsi (iii) qu'à la navigation autonome en nous focalisant sur l'utilisation de connaissances, même inexacts, afin de simplifier et d'améliorer l'apprentissage.

### 4.1 Biodiversité en biologie marine

En collaboration avec le laboratoire ECOSEAS, nous nous intéressons à l'automatisation du suivi de la biodiversité, notamment dans les aires marines protégées (*MPA : Marine Protected Area*), à partir d'images et de vidéos. Il s'agit d'étudier les variations de présence d'espèces de poissons dans certaines zones géographiques au cours du temps. Les deux propriétés principales de notre approche sont :

- pas d'images propres d'aquarium mais de vraies images de la mer avec du bruit, du mouvement, ... bref en milieu naturel
- la prise en compte des intérêts des biologistes plutôt que des métriques usuelles de machine learning

Nous allons détailler ces points par la suite.



Figure 4.1 – De gauche à droite, exemples des espèces *Acanthurus nigrofasciatus*, *Chromis margaritifer* et *Dascyllus reticulatus*. En haut, la visualisation encyclopédique, en bas la façon dont ces espèces apparaissent dans la base de données LifeCLEF2014. [REF Spampinato]

#### 4.1.1 Travaux préliminaires : reconnaissance de poissons dans leur état naturel

Notre premier essai a eu lieu en 2014 avec notre participation à la compétition de reconnaissance de poissons (*Fish*) de la base LifeCLEF2014. Nous avons été motivés par cette base de vidéos qui comportait des images et vidéos en conditions naturelles et non pas en condition d’aquarium. Les données visuelles sont de mauvaise qualité (voir figure 4.1 extraite de (Spampinato et al., 2016)), enregistrées dans des conditions réelles et concerne 10 espèces de poissons de la zone biogéographique de l’Indo-Pacifique oriental. Ces données vidéo sont plutôt médiocres en termes de résolution, assez difficiles à exploiter en raison des phénomènes naturels à prendre en compte (eau trouble, algues déplacées par le courant, etc.) et de la quantité considérable de données à traiter (400 vidéos de 10 minutes, 30000 images).

Nous avons conçu une chaîne de traitement basée sur la segmentation de l’arrière-plan, la sélection de points clés à l’aide d’une échelle adaptative, la description avec OpponentSift et l’apprentissage de chaque espèce par un classificateur binaire linéaire Support Vector Machines (voir figure 4.2). Par rapport à la base de référence conçue par les organisateurs du défi LifeCLEF, notre approche (Blanc, Lingrand, & Precioso, 2014) atteint une meilleure précision mais un rappel moins bon.

Nous avons participé à la synthèse de cette compétition (Spampinato et al., 2016) pour laquelle nous étions les seuls concurrents à produire des résultats sur les vidéos tandis qu’une autre équipe avait produit des résultats sur les images. Les deux approches ont montré des performances élevées (pour certaines espèces de poissons, la précision et le rappel étaient proches de un) dans la classification des objets et ont surpassé les méthodes de pointe. De plus, malgré le fait que l’ensemble de données soit déséquilibré en termes de nombre d’images par espèce, les deux méthodes semblent assez robustes face au problème de la longue traîne des données, affichant les meilleures performances sur les classes d’objets les moins représentées.

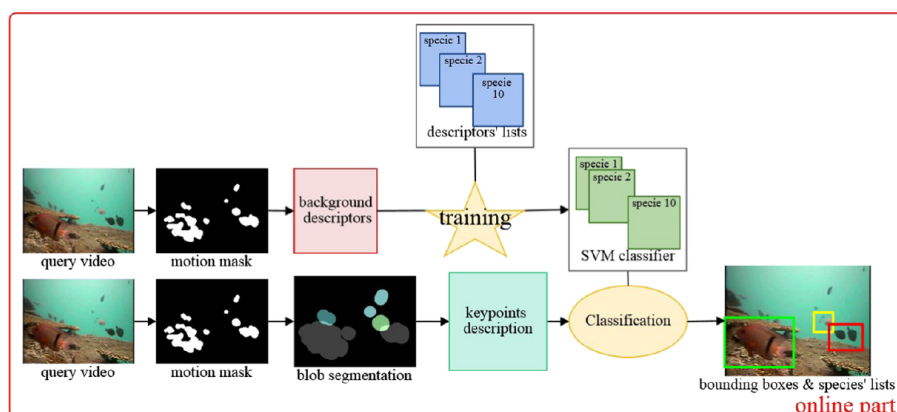


Figure 4.2 – Notre algorithme de détection d’espèces de poissons sur la base LifeCLEF 2014.

#### 4.1.2 Collaboration avec le laboratoire ECOSEAS

Pour contrer différentes menaces de l’environnement marin telles que le sur-tourisme ou la pêche intensive, différents outils ont été mis en place tels que les aires marines protégées (MPA : *Marine Protected Area*), zones dans lesquelles le mouillage et la pêche sont limités ou interdits. Afin de mesurer l’impact réel des ces aires marines protégées, il faut pouvoir évaluer les espèces présentes, leur nombre et leur évolution dans le temps. Pour cela, des méthodes d’évaluation fiables et impartiales doivent être mises en place.

Dans la plupart des cas, le recensement visuel sous-marin (UVC : *Underwater Visual Census*) est la méthode privilégiée, mais elle peut s’avérer coûteuse en termes d’efforts humains et est limitée par des facteurs météorologiques et logistique (temps de plongée limité, profil de plongée à respecter...). De plus, la présence de plongeurs peut également avoir un impact sur la faune aquatique. Les progrès technologiques permettent d’utiliser des méthodes d’enregistrement vidéo plus autonomes (par exemple, des véhicules télécommandés (ROV : *Remotely Operated Vehicle*), qui pallient ces limites.

Le but de la thèse de Kilian Bürgi (Bürgi, 2025) était d’étudier le suivi de la biodiversité marine par des techniques modernes notamment à partir de vidéos sous-marines acquises par drone sous-marin. J’ai participé à l’encadrement de sa thèse, co-dirigée par Cécile Sabourault (laboratoire ECOSEAS) et Charles Bouveyron (équipe MAASAI).

Les contributions que nous allons détailler par la suite concernent :

- la validation d’une chaîne de traitement permettant la détection de 20 espèces de poissons méditerranéens présentes sur des aires marines protégées étudiées par le laboratoire ECOSEAS (Bürgi, Bouveyron, et al., 2025)
- Métrique automatique permettant d’évaluer la quantité de poissons de 3 espèces choisies en fonction de caractéristiques différentes sur une trajectoire linéaire, cohérente avec les besoins des biologistes marins (Bürgi et al., 2026).
- Utilisation de ces travaux pour le suivi des aires marines protégées sur plusieurs saisons (papier 3)
- Mise à disposition d’un outil de mesure pour les biologistes marins (papier accepté en démo à ECAI 2025) (Bürgi, Petiot, et al., 2025)

**Preuve de concept : présence et absence d'espèces de poissons dans les vidéos** Dans un premier temps, nous avons eu accès à des vidéos réalisées par des plongeurs (DOV : *Diver Operated Video*) auxquelles était associée une évaluation des populations de 20 espèces de poissons par le plongeur : on avait ainsi accès à la vérité terrain vue d'un plongeur pour chaque vidéo sous la forme de présence ou absence de chaque espèce. Nous avons mis en place une chaîne de traitement entièrement automatisée, décrite figure 4.3. Les couleurs de chaque frame sont corrigées par le modèle pré-entraîné UIEC<sup>2</sup>-Net (Y. Wang, Guo, Gao, & Yue, 2021). Nous avons ensuite testé différents algorithmes de détection d'instances d'objets et le meilleur compromis en terme de rapidité, précision et rappel est le modèle YOLOv7 (C.-Y. Wang, Bochkovskiy, & Liao, 2023). Chaque modèle était initialement pré-entraîné sur la base COCO (*Common Objects in Context*), puis adapté aux datasets marins DeepFish et OzFish pour lesquels nous avons modifié les labels afin de détecter seulement la classe des poissons. Enfin, l'apprentissage du modèle a été adapté à la base de donnée du laboratoire ECOSEAS concernant 20 espèces de poissons méditerranéens sur 8 zones de la Côte d'Azur, entre 1 et 37 mètres de profondeur, réparties entre les saisons chaude et froide, avec différents environnement (sables, posidonies, roches ...). A partir des détections sur chaque image, on peut reconstruire différentes données comme la présence ou l'absence d'une espèce sur une vidéo.

Il est intéressant de noter que notre approche s'est éloignée de l'état de l'art (au moment de l'étude YOLOv10 affichait de meilleurs performances que YOLOv7 sur des bases classiques) car des modèles meilleurs selon des métriques usuelles du domaine (mAP par exemple) et des bases de données classiques telles que COCO n'étaient pas aussi performant dans notre cas précis. Cela est dû à la nature des images (fonds marins, souvent bruitées), à la nature des classes (des espèces de poissons qui peuvent se ressembler) mais surtout à la tâche finale qui n'est pas tant la précision des boîtes englobantes mais la détection ou non d'un poisson. Dans notre cas, le rappel avec le modèle YOLOv10 est catastrophique :

Architecture	Rappel	Précision	Score F1
Fast R-CNN + ResNet50 + RPN + FPN	0.50	0.62	0.56
Faster R-CNN + ResNeXt101 + FPN	0.51	0.67	0.61
RetinaNet + ResNet101 + FPN	0.23	0.62	0.34
YOLOv7	0.62	0.64	0.63
YOLOv10	0.28	0.85	0.52

Cette première étude (Bürgi, Bouveyron, et al., 2025) nous a permis de valider notre approche sur des données réelles pour lesquelles les résultats concordent avec les mesures manuelles et a pu convaincre les biologistes marins de réaliser plus de campagnes d'acquisition de données afin de correspondre davantage à l'évaluation UVC traditionnelle.

**Estimation de la biodiversité automatique** La présence ou l'absence d'espèces de poissons est une information intéressante mais pas suffisante pour étudier l'évolution de la biodiversité. Compter le nombre d'espèces présentes mais aussi évaluer leur nombre permet d'étudier la dynamique des populations et l'équilibre de la biodiversité (Pinna et al., 2023). Selon les espèces, le comptage est difficile car certains poissons comme *Epinephelus marginatus* sont solitaires et sont souvent immobiles et camouflés dans les roches. Le comptage systématiques des espèces par les biologistes marins étant une tâche longue et fastidieuse, nous nous sommes limité à 3 espèces issues de niches écologiques différentes :

***Epinephelus marginatus* (Merou brun)** : solitaire, pouvant se cacher de façon immobile en imitation de la roche



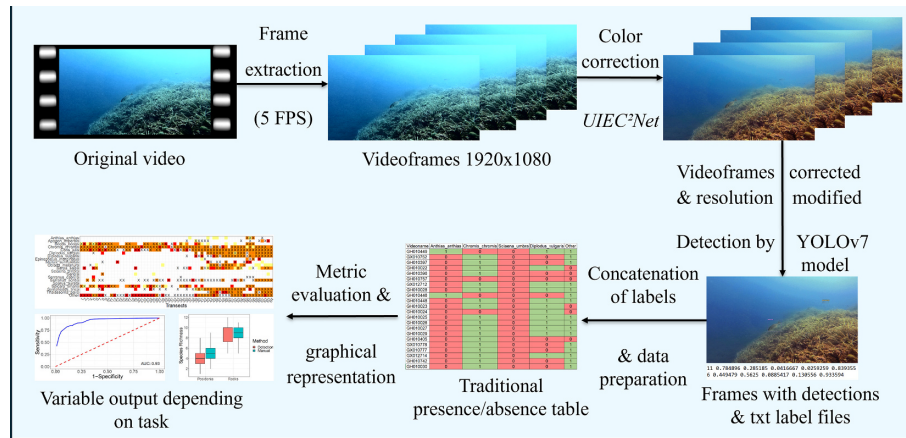


Figure 4.3 – Chaîne de traitement des vidéos : extraction des frames, pré-traitement, détection des poissons et fusion des détections selon les sorties nécessaires.

***Sciaena umbra* (Corb commun) :** en petits groupes proches avec beaucoup de ressemblances et d’occlusion

***Diplodus vulgaris* (Sar à tête noire) :** en bancs de poissons nombreux variant en nombres

Une mesure couramment utilisé en biologie marine,  $N_{max}$ , correspond au nombre maximum de poissons vus sur une seule frame dans une vidéo prise par une caméra fixe et un appât pour attirer les poissons. Seulement, pour évaluer en milieu naturel sans appât, les plongeurs réalisent des *transects* ou plongée selon une trajectoire linéaire pendant lesquels des poissons peuvent arriver et repartir de ces trajectoires : la mesure  $N_{max}$  sous-estime largement le nombre réel de poissons. Nous avons tout de suite écarté l’idée de réaliser un suivi des poissons afin de compter le nombre de poissons au fil de la séquence vidéo pour des raisons de coût car nous visons à terme de pouvoir embarquer ce type d’algorithme mais aussi pour des raisons de précisions car pour certaines espèces, il est impossible de reconnaître les individus au sein d’une même espèce. Nous avons élaboré 3 mesures simples et rapides pour améliorer la métrique  $N_{max}$  (Bürgi et al., 2026) :

$N_{cluster}$  : au lieu de ne prendre que le nombre maximum de poissons, la somme des maximums de chaque cluster (obtenu par k-mean) du profil de comptage de chaque poisson est utilisée et doit correspondre à la somme des bancs de poissons observés pendant la trajectoire.

$N_{heuristic}$  : est similaire à la méthode précédente en utilisant des heuristiques propres à chaque espèce concernant la détermination des bancs de poissons (nombre de poisson minimal et distance entre bancs).

$N_{TCN}$  : en utilisant un réseau de neurones de type TCN, il est possible de prendre en compte la nature dynamique des bancs de poissons

Nous avons pu mesurer la qualité de ces estimations sur les détections vraies des poissons dans les différentes frames mais aussi dans des conditions réalistes par détection automatique des poissons selon une chaîne de traitement similaire de celle du paragraphe précédent 4.4. Comme déjà étudié par d’autres auteurs,  $N_{max}$  sous-estime la quantité de poisson. Les méthodes  $N_{TCN}$  et  $N_{heuristic}$  sont celles qui se rapprochent le plus des évaluations manuelles par les plongeurs. Avec

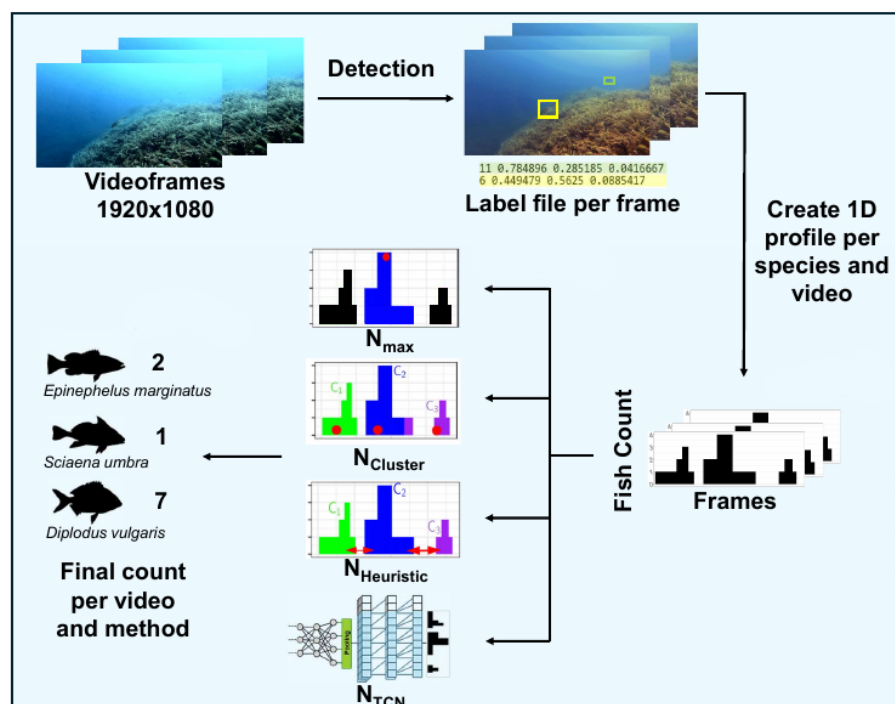


Figure 4.4 – Chaîne de traitement des vidéos : extraction des frames, pré-traitement, détection des poissons et comptage par frame et par espèce.

une variation absolue comparable à la variation inter-plongeurs, nous avons montré qu’il s’agit de méthodes fiables pour quantifier le nombre de poissons pour ces trois espèces méditerranéennes différentes. La méthode  $N_{heuristic}$  nécessite une estimation des paramètres pour chaque nouvelle espèce tandis que  $N_{TCN}$  obtient de meilleurs résultats lorsque la situation est plus complexe et le nombre de poissons plus élevé.

**Suivi de la biodiversité au cours du temps dans les aires marines protégées** Grâce aux méthodes développées précédemment, nous avons étudié les données de plusieurs aires marines protégées de la Côte d’Azur (Corniche Varoise, Esterel et Cap Ferrat), sur différents substrats (herbiers marins et roches), à des profondeurs variables, entre Octobre 2023 et Avril 2025 (Bürgi, Sun, et al., 2025). Les 3 espèces sont les mêmes afin de simplifier la tâche d’annotations.

Les principales contributions de cette étude sont les suivantes :

- Impact de la répartition des données d’apprentissage en fonction des densités de poissons et des saisons répartition sur l’entraînement des modèles et évaluation de la faisabilité d’exploiter les données de différentes saisons et années pour la généralisation intersaisonnière des modèles. Nous avons ainsi pu proposer un guide de répartition optimale des données à annoter pour réaliser un apprentissage généralisable aux années ultérieures.
- Identification des espèces les plus touchées par des facteurs environnementaux spécifiques, fournissant des recommandations fondées sur des données pour optimiser les efforts de protection et de conservation dans les zones de conservation étudiées

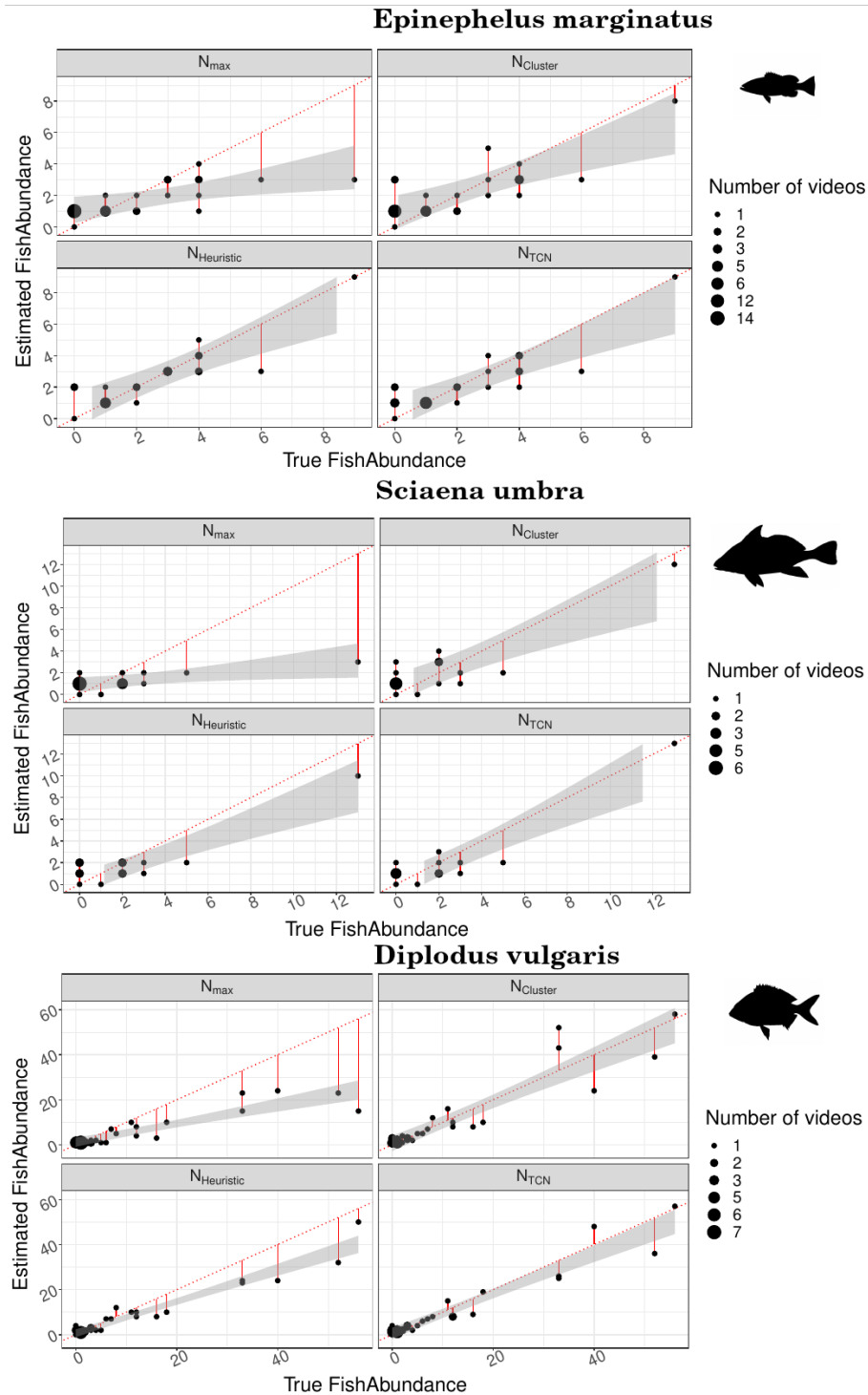


Figure 4.5 – Regression linéaire entre le nombre exact de poissons et son estimation par les 4 méthodes différentes à partir des détection de la chaîne de traitements 4.4. En gris, l'intervalle de confiance à 95%. La taille des points est relative au nombre de vidéos sauf pour les 36/45/26 vidéos ne comportant pas de poissons (0) qui ont été réduites à la taille 1. La ligne pointillée rouge indique une estimation exacte : les points en dessous correspondent donc à une sous-estimation. Au total : 56 individus *Epinephelus marginatus* pour 19 vidéos. *Sciaena umbra* : 33 individus pour 9 vidéos. *Diplodus vulgaris* : 334 individus pour 28 vidéos.

- Recherche des facteurs environnementaux ayant le plus d’impact sur la fréquence et la densité des espèces

Dans le cadre de cette étude, nous avons pris en compte le fait que pour l’étude de la biodiversité, il est plus important de détecter les poissons peu nombreux que d’oublier un petit nombre de poisson dans un grand banc de poissons. Ainsi, une même différence dans le décompte des poissons doit avoir un impact plus grand quand le nombre de poissons est petit. Nous avons ainsi modifié la métrique de MAE (*Mean Absolute Error*) en un nouveau critère CMARE :

$$\text{CMARE} = \frac{1}{n} \sum_{i=1}^n k * \frac{|y_i - \hat{y}_i|}{k + y_i}$$

pour lequel  $n$  est le nombre de vidéos,  $y_i$  le nombre exact de poissons dans la vidéo  $i$  et  $\hat{y}_i$  son estimation. Le facteur de correction  $k$  a été déterminé à 234 par le fait que les biologistes marins estiment que l’erreur doit représenter 70% de la vraie erreur (CMARE pour une erreur de 1 poisson sur 100 vaut 0.7).

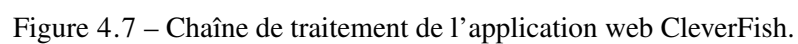
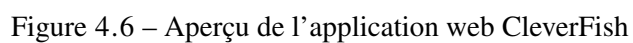
Nous avons identifié des variations saisonnières chez *Diplodus vulgaris* qui pourraient être perturbées par le changement climatique. De plus, nous avons confirmé les préférences en matière de profondeur et de substrat chez toutes les espèces cibles, ce qui corrobore la littérature existante. Nos résultats démontrent que la vision par ordinateur permet une extraction rapide et moins biaisée des données à partir de grands ensembles de données, facilitant ainsi une surveillance efficace de la biodiversité.

Nous prévoyons que cette étude servira de base à l’extension des applications de la vision par ordinateur à d’autres régions marines, espèces et scénarios environnementaux. L’efficacité accrue du traitement des données permettra des évaluations plus fréquentes et plus approfondies de la biodiversité, renforçant ainsi les efforts de conservation grâce à une meilleure compréhension écologique.

**Partage de données et partage d’outils** Les travaux présentés dans le cadre de cette collaboration avec le laboratoire ECOSEAS et au travers de la thèse de Kilian Bürgi sont soit publiés soit en cours de publication. Les données sont publiques ainsi que les codes sur des sites `github` liés à chacune des publications. De plus, une interface destinée aux biologistes marins, CleverFish (Bürgi, Petiot, et al., 2025), est disponibles sur le site du 3IA Côte d’Azur\* et a été présenté à la 3ème Conférence des Nations Unies sur l’Océan à Nice en Juin 2025 (UNOC3). Elle est disponible gratuitement sur inscription et permet à un biologiste marin d’importer ses propres vidéos afin d’obtenir la détection et le décompte de 19 espèces méditerranéennes par frame et par vidéo (voir un aperçu de l’application sur la figure 4.6). Outre la visualisation à l’écran, il est possible d’exporter les résultats dans des formats qui conviennent à cette communauté (voir la chaîne de traitement sur la figure 4.7). Une vidéo de présentation est disponible sur YouTube†.

\*. CleverFish : <https://3ia-demos.inria.fr/cleverfish/en/>

†. <https://www.youtube.com/watch?v=yAX3ECZ-azU>



## 4.2 Archeology

### 4.2.1 Previous work

Le projet européen Digiart<sup>‡</sup> (Revolutionizing Cultural Heritage Through Digital Art And 3D Technology), du programme Horizon H2020, 2015-2018, concerne la préservation et la promotion auprès d'un large public de sites culturels et archéologiques à l'aide de nouvelles technologies. Les participants de ce projet étaient experts dans des différents domaines : archéologie, art numérique, modélisation 3D, informatique afin de construire ensemble des expériences immersives liant l'histoire ancienne et la technologie moderne. Cette multitude de compétences dans le projet en a fait sa richesse et sa complexité : les outils fournis devaient être utilisés par des non spécialistes de l'informatique.

Dans le cadre de ce projet, nous nous sommes intéressés à la reconstruction de scènes 3D à partir d'acquisition Lidar et à la reconnaissance d'objets à partir de nuages de points 3D sur des sites archéologiques. Nos contributions ont porté principalement sur deux points que nous détaillons dans les paragraphes suivants :

- le recalage entre scans de haute résolution (nombreux points) afin de recréer une scène 3D complète acquise depuis différents points de vue. Pour cela nous avons mis au point un détecteur et descripteur de paires de points d'intérêts adapté à ce problème : KPPF (Malleus et al., 2017)
- un outil à destination des archéologues leur permettant de scanner un objet, même partiel et leur fournissant des objets proches à partir de catalogue de leur domaine (Samoun, Fisichella, Lingrand, Malleus, & Precioso, 2018)

**KPPF (*Keypoint-based Point-Pair-Feature*) : un descripteur pour le recalage de nuages de points 3D** L'un des défis les plus importants dans le domaine du traitement des données 3D consiste à pouvoir reconstruire une scène 3D complète avec une grande précision à partir de plusieurs captures. Ce processus se déroule généralement en deux phases principales : une étape d'alignement grossier, puis une étape d'alignement fin. Dans cet article (Malleus et al., 2017), nous proposons une méthode d'enregistrement global automatique et évolutive (c'est-à-dire sans pose arbitraire du capteur) sous les contraintes suivantes : sans marqueurs, données à très grande échelle (plusieurs millions de points par scan), peu de chevauchement entre les scans, plus de deux ou trois douzaines de scans, sans connaissance a priori des 6 degrés de liberté.

Nous ne traitons ici que le recalage grossier et considérons que l'étape de recalage fin est prise en charge par des approches existantes dédiées telles que l'ICP (Iterative Closest Point). Nous évaluons de manière approfondie notre méthode sur notre propre ensemble de données composé de 33 scans à grande échelle d'un bâtiment intérieur. Les données présentent quelques paires de scans avec très peu de chevauchement, des défis architecturaux (un patio et une rotonde ouverts sur plusieurs niveaux du bâtiment, des vitres, des reflets...), plusieurs millions de points par scan. Nous rendons cet ensemble de données public dans le cadre d'un benchmark mis à la disposition de la communauté.

Dans un premier temps, des points d'intérêts (SIFT3D), répétables, sont extraits et groupés par paires afin de calculer notre descripteur, léger et compact, de paires 4.8. Seuls les points pour lesquels les normales sont stables (surface pouvant être localement approximée par un plan) sont conservés. Ces descripteurs sont ensuite utilisés pour estimer les transformations rigides. Comme

---

<sup>‡</sup>. <http://digiart-project.eu/>

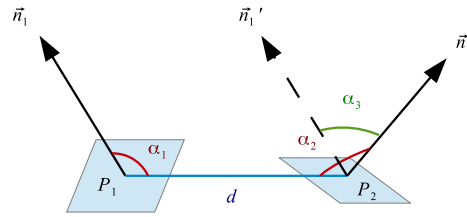


Figure 4.8 – Le descripteur : angles entre chaque normale ( $n_1$  et  $n_2$ ) et le segment défini par les points  $P_1$  et  $P_2$  ( $\alpha_1$  et  $\alpha_2$ ) et angle entre les normales  $n_1$  et  $n_2$  ( $\alpha_3$ )

il existe de nombreux points aberrants, les transformations ainsi calculées sont accumulées dans un espace de vote afin d'estimer la zone de plus grande densité, à la façon d'une transformée de Hough.

Nous avons ainsi évalué la précision de notre méthode, son évolutivité par rapport au nombre initial de points et sa robustesse face aux occlusions, au faible chevauchement des scans et aux défis architecturaux.

### Un système interactif de reconnaissance d'objets 3D pour utilisation sur site archéologique

Dans le cadre du projet Digiart, nous avons conçu un système interactif de recherche de formes 3D basé sur le contenu (CB3DR) adapté aux archéologues et paléontologues (Samoun et al., 2018). Notre solution CB3DR vise à scanner un objet à la volée à l'aide d'un capteur 3D peu coûteux (par exemple Structure Sensor de Occipital) et à récupérer des formes similaires dans une base de données à partir du nuage de points 3D acquis. Notre système répond aux besoins des archéologues qui souhaitent pouvoir acquérir des artefacts sans avoir de connaissances préalables en matière de numérisation, puis effectuer facilement des requêtes à partir des bases de connaissances sur le terrain relatives au patrimoine culturel, et ainsi récupérer des artefacts (c'est-à-dire des objets ou des parties d'objets) de forme similaire sans avoir à transporter ni même à déplacer l'artefact trouvé sur le site. Il s'agit clairement d'un contexte de recherche plutôt que de classification, car l'artefact trouvé peut être inconnu.

En première étape, nous avons étudié les différents descripteurs existants, avec ou sans apprentissage, afin de trouver celui qui sera le plus adapté à notre contexte de basse résolution, de formes 3D incomplètes et de systèmes d'acquisition variables tout en restant rapide à calculer sur une tablette munie d'un capteur 3D. Nous avons considéré des descripteurs plus anciens, ne nécessitant pas d'apprentissage (GSHOT, ESF ...) ainsi que des descripteurs modernes et profonds comme PointNet.

Afin d'inclure les archéologues dans le processus nous avons mis en place une chaîne traitement avec de l'apprentissage actif. Un archéologue fait l'acquisition intégrale ou partielle d'un objet 3D. Notre système calcule le descripteur et une recherche par similarité permet de sélectionner  $K$  objets de notre base de données les plus proches. Il s'agit de l'initialisation de l'apprentissage actif par SVM. A chaque itération, quelques exemples positifs et négatifs sont proposés à l'utilisateur qui doit éventuellement apporter ses corrections. Le classifieur SVM est réentraîné entre chaque itération. La figure 4.9 montre un exemple d'interface.



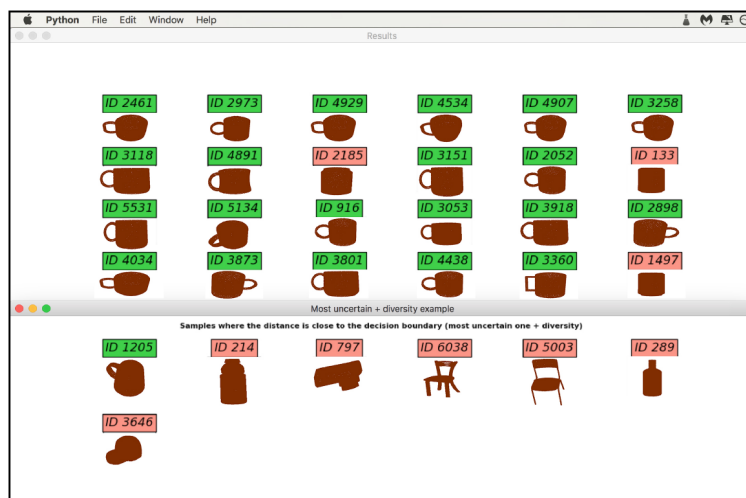


Figure 4.9 – Après 5 itérations. En haut, le résultat proposé à l'utilisateur. Les échantillons correctes sont en vert, les autres en rouge. En bas, les exemples comportant le plus d'incertitude que l'utilisateur pourra corriger à la prochaine itération.

Nous avons évalué notre système sur des bases de données standards ainsi que sur notre base de données maison comportant des objets simples et disponibles en nombre au laboratoire : tasse, chaise, table. . . . Cela nous a permis de sélectionner les 2 meilleurs descripteurs : PointNet et ESF. PointNet, avec ou sans raffinement de l'apprentissage est généralement meilleur que ESF qui ne nécessite pas d'apprentissage et est aussi plus rapide. Le nombre d'itération en actif est entre 10 et 20 selon les bases de données. Nous avons également montré que sur une base d'os du petit bassin (pelvis) acquises par notre système, les résultats sont meilleurs avec le descripteur ESF. Ainsi, même si l'état de l'art à ce moment indiquait que PointNet ou d'autres réseaux de neurones avaient les meilleures performances, dans notre cas précis, un descripteur plus ancien et plus simple permettait d'obtenir les meilleurs résultats car les particularités de nos données ne correspondaient pas à celle des bases sur lesquelles les algorithmes modernes étaient comparés.

## 4.2.2 ANR AIWOOD

Je suis responsable du partenaire I3S dans le projet ANR AI-WOOD<sup>§</sup>. Ce projet est porté par le CEPAM (Cultures et Environnements Préhistoire, Antiquité, Moyen Âge, UMR 7264 CNRS-Université Côte d'Azur), avec pour partenaire Muséum National d'Histoire Naturelle. Il a pour but le développement de nouvelles approches d'apprentissage profond visant à réaliser l'identification taxonomique (c'est-à-dire la classification au niveau de l'espèce, du genre ou de la famille) du bois et du charbon de bois à partir d'images microscopiques en 2D. Le projet a un intérêt principal d'un point de vue archéologique, l'idée principale étant d'entraîner un classificateur pour l'identification des espèces et des familles sur une collection moderne (environ 120 espèces) et de l'utiliser ensuite pour identifier des charbons de bois anciens.

Les anthracologues (c.-à-d. les archéologues spécialisés dans l'identification et l'analyse des charbons de bois anciens) effectuent cette identification en s'appuyant sur l'anatomie comparée et

<sup>§</sup>. ANR-23-CE38-0013 <https://www.cepam.cnrs.fr/programmes-recherche/anr-ai-wood/>



Espèces	Nb. arbres	Images/Coupe			Nb. images
		R	T	Tr	
<i>Picea abies</i>	20	95	48	50	193
<i>Juniperus communis</i>	19	143	81	86	310
<i>Larix decidua</i>	19	62	41	40	143
<i>Pinus sylvestris</i>	23	121	65	65	251
Total	<b>81</b>				<b>897</b>

TABLE 4.1 – Pour les 4 premières espèces, nombre d’individus et d’images par type de coupe (ou section).

sur les caractéristiques anatomiques établies par l’IAWA <sup>¶</sup> qu’ils construisent manuellement par observation microscopique. Outre le fait qu’elle est longue et fastidieuse, cette routine d’identification n’est pas entièrement satisfaisante, aussi en raison de la proximité anatomique de certaines essences. Si les espèces de certains charbons archéologiques ont pu être identifiées avec certitude, ce n’est pas le cas de tous les charbons.

Le but de ce projet est donc d’explorer le potentiel de l’apprentissage profond pour identifier directement le taxon d’un spécimen à partir de l’observation microscopique et éventuellement d’améliorer la routine d’identification. Bien que certaines tentatives dans ce sens ont été faites dans la littérature (Rosa da Silva, Deklerck, Baetens, & et al., 2022), il y a encore une marge d’amélioration considérable.

**Etude de faisabilité sur les charbons modernes et travail sur la fusion d’images sans compromis sur la résolution.** Avec Dieu-Donné Fangnon, doctorant recruté sur ce projet, et mon collègue Marco Corneli, nous avons mis en place un modèle d’apprentissage profond (Théry-Parisot et al., 2025) permettant d’obtenir une précision de l’ordre de 80 % sur 4 classes de charbons modernes pour lesquelles nous disposons de suffisamment de données (voir tableau 4.1. Il est à noter que nous avons respecté la contrainte que les images utilisées en test pour évaluer les performances ne proviennent d’aucun individu présent dans la base d’entraînement. Cela paraît évident mais n’a pas toujours été respecté dans les publications précédentes.

Un arbre individuel est représenté par une ou plusieurs images pour chacune des coupes radiales, tangentielle et transverse (voir figure 4.10. Pour une même coupe, les différentes images s’expliquent par le fait qu’elles peuvent correspondre à différentes parties de l’arbre ou bien, le champ de vue au microscope étant réduit, à plusieurs parties d’une même coupe physique. Un anthracologue recherche des caractéristiques anatomiques qui ne sont pas forcément toutes présentes sur la même image. On peut également se trouver dans un cas de figure où un anthracologue aurait vu certaines propriétés anatomiques sans qu’elles soient présentes dans aucune image.

La chaîne de traitement que nous avons mise en place est globalement présentée sur la figure 4.11 puis détaillée au niveau de l’extraction des caractéristiques sur la figure ???. Nous avons choisi de concaténer la description de chacune des types de coupes car, d’après les experts anthracologues, ce ne sont pas toujours les mêmes coupes qui apportent des informations sur les espèces. Remplacer cette concaténation par une autre méthode d’agrégation serait envisageable. Nous avons pris l’exemple des coupes radiales pour détailler l’extraction de caractéristiques de cette coupe sur la figure 4.12. Le même algorithme est appliqué sur les 2 autres types de coupes. On

¶. <https://www.iawa-website1.org/>

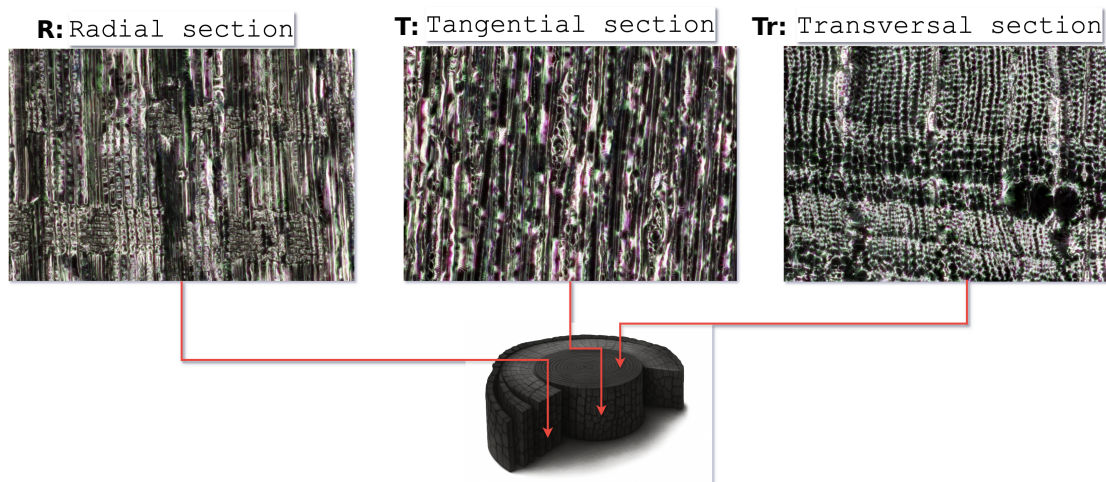


Figure 4.10 – Les trois types de coupes d'un charbon

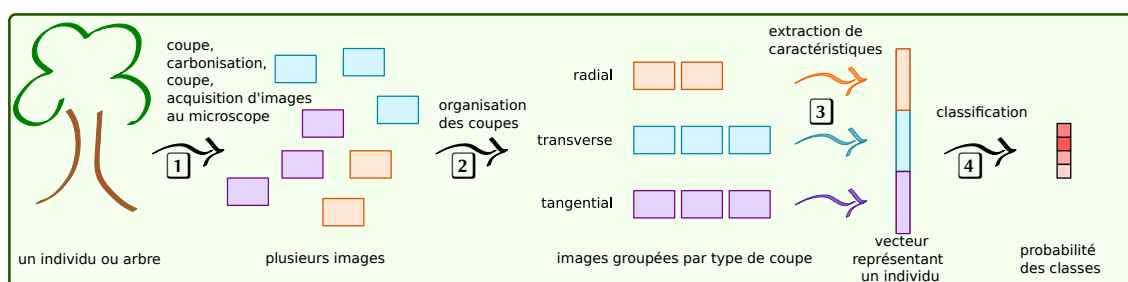


Figure 4.11 – Une vue générale de la chaîne de traitement pour la classification des arbres. Pour chaque arbre individuel, une ou plusieurs parties sont coupées, puis brûlées pour obtenir du charbon, découpé selon 3 orientations possible : radiale, tangentielle ou transversale. Ensuite, une ou plusieurs images sont acquises au microscope [1]. Les images sont ensuite groupées par type de coupe [2] et un vecteur de caractéristiques par coupe est extrait [3]. Ces 3 vecteurs sont alors concaténés dans un but de classification [4]. L'extraction d'un vecteur pour les coupes radiales est détaillé sur la figure 4.12.

remarque qu'il s'agit d'une fusion au niveau des caractéristiques de chaque sous-image. La fusion à cet étage a été motivée par des études précédentes : les auteurs de ont montré, dans le cadre de tâches de classification vidéo, qu'il est préférable de fusionner au niveau de la carte d'activation plutôt qu'au niveau des scores softmax. Ils empilent les résultats de deux branches du réseau, puis utilisent une convolution 1x1. Plusieurs articles (Feichtenhofer, Pinz, & Zisserman, 2016; Barbosa, Marinho, Martin, & Hovakimyan, 2020; Seeland & P., 2021) s'accordent à dire que la fusion tardive donne de meilleurs résultats, plus précisément au niveau des représentations. Une étude approfondie des différentes façons d'utiliser plusieurs vues a été menée dans (Seeland & P., 2021). Les auteurs ont étudié différents points de fusion dans un réseau de neurones : fusion à l'entrée, à la fin de la partie descripteur du réseau ou au niveau des scores de prédiction. Ils ont également montré qu'une fonction de fusion apprise donne de meilleurs résultats.

Différentes méthode de fusion (ou *pooling*) peuvent être considérées : la somme, la moyenne, le maximum ou encore une combinaison de ces mesures qui sont des méthodes sans apprentissage.

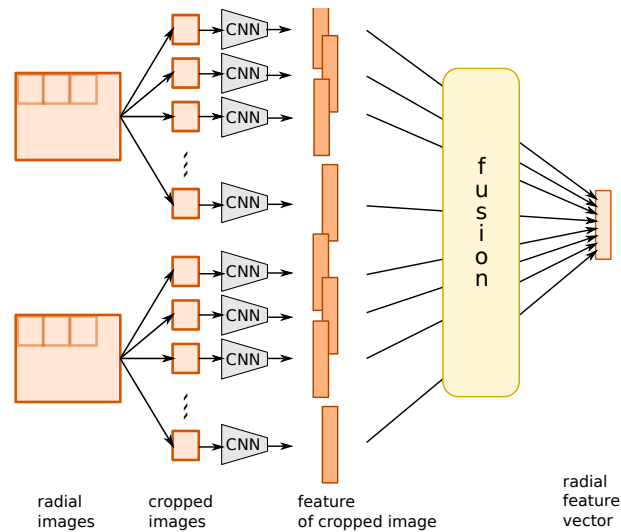


Figure 4.12 – Pour chaque image radiale du même individu, on découpe des sous-images afin de les donner en entrée à un CNN (sans tête). La dimension de ces sous-images dépend du type de CNN, pré-entraîné sur ImageNet. Après un GAP ou aplatissement, on obtient un vecteur caractéristique pour chaque sous-image. Un algorithme de fusion (*pooling*) permet de fusionner tous les vecteurs caractéristiques en un unique vecteur commun.

Dans (Seeland & P., 2021), ces méthodes ont été étudiées et comparées avec des méthodes par apprentissage. Ces dernières conduisent à de meilleurs résultats. Cependant, dans cette étude, le nombre de vues était fixe ce qui permettait ainsi de concaténer les cartes d’activation et d’utiliser une couche dense pour la fusion. Une autre méthode, moins efficace utilisait la convolution 1x1.

Dans notre cas, il n’était pas possible d’utiliser directement les méthodes de fusion issues de la littérature car nous disposons d’un nombre variable d’images par coupes, selon les individus. De plus, nous découpons, sans chevauchement, des sous-images de dimensions correspondant aux dimensions requises pour les modèles convolutifs pré-entraînés : un changement d’échelle des images ferait disparaître les détails fins observés au microscope permettant de différencier des espèces.

Sans apprentissage, nous avons comparé la moyenne, le maximum, la variation autour de la moyenne (écart type) ainsi que la concaténation de ces trois valeurs. Nous avons également mis en place plusieurs méthodes de fusion par apprentissage et avons conservé la meilleure (voir figure 4.13). Nous avons cherché à exprimer la contrainte de variabilité de dimension en entrée ainsi que l’invariance à l’ordre des entrées. De façon similaire à PointNet (Qi, Su, Mo, & Guibas, 2017), nous avons proposé d’utiliser une fonction simple (maximum) qui nous enrichissons en transformant les entrées par MLP. Nous avons ajouté des connexions résiduelles afin d’obtenir des performances au moins équivalents à celle d’un maximum seul et de faciliter l’apprentissage. Dans le tableau 4.2 nous pouvons observer que la fusion apprise surpasse les autres fusions. Nous avons également étudié différentes résolutions et modèles pré-entraînés.

En dehors des modèles convolutionnels pré-entraînés, nous avons également testé deux modèles de représentation par transformeurs : DINO v2 et v3. Pour DINO v3, nous avons testé deux pré-entraînements différents : l’un sur des images courantes (LVD), l’autre sur des images satellites. Nous pensions que le modèle entraîné sur des images satellites aurait de meilleurs per-

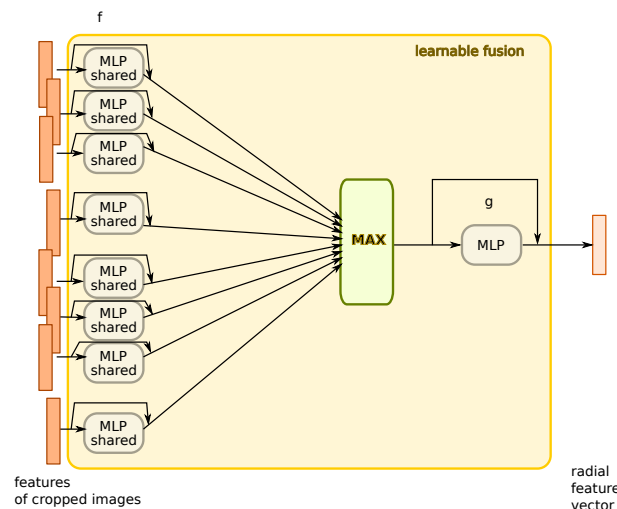


Figure 4.13 – Illustration de la fusion par apprentissage que nous proposons : chaque vecteur de caractéristiques de chaque sous-images d’un type de coupes est transformé par la fonction  $f$  (avec connexion résiduelle), modélisée par un MLP partageant ses poids, avant une fonction élémentaire de fusion (max) elle-même suivie d’une autre fonction  $g$  (avec connexion résiduelle) modélisée par un autre MLP.

Dimension sous-images	Méthode de fusion	Modèle			
		ResNet-50	EfficientNet-B7	EfficientNet-V2L	ConvNeXt-XLarge
Original	Mean	75.37± 0.08	83.90± 0.06	79.12± 0.05	86.40± 0.07
	Max	79.04± 0.08	<b>86.40±0.04</b>	80.29±0.09	86.40± 0.07
	Std	75.37± 0.08	85.22± 0.07	79.12± 0.09	76.54± 0.08
	Mean-Max-Std	77.79± 0.09	<b>86.40± 0.04</b>	79.12±0.09	85.15±0.08
	fusion MLP	77.87± 0.08	82.72± 0.07	85.15± 0.05	76.54± 0.09
224x224	Mean	80.22±0.04	83.97±0.11	84.04±0.11	80.22±0.07
	Max	85.07±0.10	81.54±0.13	85.07±0.10	80.15±0.10
	Std	86.32±0.12	83.97±0.09	83.90±0.10	80.22±0.08
	Mean-Max-Std	86.32±0.12	<b>86.40±0.12</b>	85.15±0.11	81.40±0.11
	fusion MLP	81.47±0.08	79.56±0.11	81.54±0.08	76.54±0.09
300x300	Mean	80.22±0.08	85.15±0.14	<b>86.47±0.09</b>	81.54±0.10
	Max	86.32±0.12	81.54±0.13	85.15±0.11	77.79±0.08
	Std	85.15±0.11	83.97±0.11	82.72±0.09	81.47±0.08
	Mean-Max-Std	85.15±0.11	85.22±0.12	83.97±0.11	82.65±0.09
	fusion MLP	80.22±0.07	75.20±0.04	83.82±0.10	75.37±0.11
500x500	Mean	81.54±0.07	86.32±0.12	85.29±0.12	82.72±0.09
	Max	85.15±0.09	83.82±0.10	87.65±0.10	83.90±0.10
	Std	81.54±0.08	85.07±0.10	86.40±0.08	77.87±0.09
	Mean-Max-Std	85.15±0.09	85.07±0.10	<b>90.15±0.09</b>	83.90±0.13
	fusion MLP	80.22±0.07	77.79±0.06	<b>91.32±0.08</b>	75.29±0.06

TABLE 4.2 – Comparaison de méthodes de fusion selon différents modèles pré-entraînés et différentes résolutions (chaque résolution correspond approximativement à la résolution demandée par chaque modèle).

formances car il a du apprendre des éléments discriminants de texture. Nos expérimentations n'ont malheureusement pas été concluantes. Dans notre cas, les modèles convolutionnels semblent mieux capter les éléments discriminants pour la classification.

Dim. ss img	Backbone		
	DINOv2-ViTb14	DINOv3-ViTb16(lvd)	DINOv3-ViTl16(sat)
224×224	80.15±0.10	<b>83.97±0.10</b>	73.97±0.07
518×518	<b>81.32±0.13</b>	80.15±0.10	78.90±0.06

TABLE 4.3 – Précision moyennes (%) et écart type de la validation croisée utilisant la représentation des images par le jeton CLS de deux versions récentes de DINO, avec notre méthode de fusion par apprentissage.

Pendant cette étude, la base de données a continué de s'étoffer et 21 espèces disposent au 1er janvier 2026 d'au moins une quinzaine d'individus. Nous avons aisément étendu notre modèle à ces 21 espèces contenant 520 individus au total, tout en conservant les mêmes conclusions et une précision de l'ordre de 80% pour chacune des espèces. Il est à noter que cette précision monte à environ 85% au niveau du genre et 94% au niveau de la famille.

Notre méthode de fusion a été soumise à la conférence ICIP 2026. Elle nous permettra d'initier des études sur les sous-images utiles ou non dans la prise de décision : celles qui ont participé ou non au maximum des caractéristiques. Ces travaux sont en cours.

**Intérêts en apprentissage automatique** Ce projet, en lien avec la communauté anthracologique, est intéressant car il présente différentes problématiques ou défis :

**un nombre variable de données par individu pour leur identification** Les images sont des images de coupes de différentes résolutions microscopiques mais pas du même volume 3D comme on peut le rencontrer en imagerie médicale par exemple car lorsqu'un charbon est coupé dans un sens, il n'est pas recolé pour être coupé orthogonalement. Il convient d'étudier comment utiliser les différentes coupes afin de prendre une décision de classification. D'autre part, afin de conserver la résolution initiale des images, celles-ci sont découpées en sous-images. Nous avons déjà proposé une méthode de fusion permettant un nombre variable d'images par coupe et par individu.

**des classes hiérarchique** Les classes elles-mêmes proviennent d'une hiérarchie. Par exemple, le pin cembro est de la famille des Pinaceae, du genre Pinus. On devra envisager de déterminer l'espèce, le genre ou la famille ou bien utiliser une information des anthracologues sur la famille ou le genre afin d'aider la détermination de l'espèce. Les échelles d'acquisition microscopiques pourraient également être utilisées de façon hiérarchique : des détails fins sont nécessaires pour différencier des espèces mais pas forcément au niveau de la famille ou du genre.

**une distribution des classes non uniforme** Il existe de nombreuses espèces mais créer une base de données d'images de charbons modernes prend du temps. Si on veut être exhaustif sur une région, il est nécessaire de prélever de nombreuses espèces. On obtient généralement suffisamment d'individus pour un apprentissage classique que pour peu d'espèces mais il convient de se pencher sur la questions de l'apprentissage avec peu de données pour d'autres espèces (*Few Shot Learning*) ou même encore d'apprentissage sur des don-

nées très déséquilibrées et à longue traine (*Long tailed Learning*). Le recrutement d'un IE sur ce projet va renforcer cet axe.

**des connaissances des experts** Les anthracologues utilisent des critères d'identification IAWA, internationalement reconnus mais qui ne permettent pas d'identifier tous les taxons en général. Ceci est d'autant plus vrai pour les charbons archéologiques pour lesquels tous les critères anatomiques ne peuvent pas être identifiés. Néanmoins, nous envisageons d'utiliser cette connaissance selon différents axes mais aussi, par recherche de l'explication de nos décisions, de pouvoir expliciter de nouveaux critères d'identification non connus aujourd'hui. Utiliser la distance entre les ensembles de critères pour revisiter la hiérarchie des classes est aussi une piste à étudier.

**un changement de domaine entre l'époque moderne et archéologiques** La vérité des espèces des charbons archéologiques n'est pas connue. Elle a pu être déterminée pour certains d'entre-eux mais pas pour tous. La base de données de charbons de bois modernes en cours de création comporte des labels certains (quelqu'un a vu l'arbre qu'il a coupé et a pu bien plus aisément, et avec certitude, l'identifier). Même si on construit un modèle pertinent pour des charbons modernes, il conviendra d'étudier le transfert de domaine vers l'archéologique avec peu de données labelisées.

### 4.3 Navigation autonome

Je suis responsable du partenaire I3S dans le projet ANR Multitrans<sup>||</sup>. Ce projet a démarré au printemps 2022. Il est porté par le laboratoire LITIS (INSA Rouen) et a pour partenaires l'I3S et le laboratoire Valeo.ai. Ce projet concerne la conduite autonome avec une approche originale qui utilise le transfert d'apprentissage (voir figure 4.14) entre (i) un modèle complètement simulé (MuSHR), (ii) un modèle réduit (circuit à une échelle 1/10e) et avec des véhicules équipés de différents capteurs et (iii) un modèle à taille réelle (voitures Valeo, navette Milla à l'IMREDD). Dans le cadre de ce projet, nous nous intéressons à l'apprentissage fédéré permettant à différents véhicules autonomes de partager de l'information, au passage de l'apprentissage dans un domaine simulé vers un domaine réel (adaptation de domaines) et insérer de la connaissance, même imprécise, pour apprendre mieux.

Dans un premier temps, nous avons construit un circuit modulable à l'échelle 1/10e et construit différentes générations de véhicule à base de véhicule en modèles réduits que nous avons équipés de carte GPU (Jetson Nano), de multiples capteurs visuels (Lidar, RGB-D, caméras stéréoscopiques). Cette plateforme existe également en virtuel afin de simuler également des expérimentations et apprentissages virtuels. Différents projets étudiants en cycle ingénieur à Polytech mais aussi en Master 2 DSAI ont travaillé sur les différentes versions de cette plateformes. Nous avons notamment pu tester l'apprentissage fédéré entre différentes voitures explorant différentes parties du circuit. Dans la perspective d'une compétition à plus grande envergure, nous avons fait un premier test en organisant un Hackathon sur 6 jours avec les étudiants en IA de l'école Centrale de Marseille en Mars 2023 4.15.

Dans le cadre de ce projet, outre les différents projets étudiants encadrés, nous avons recruté une ingénieure, Li Yang, ainsi qu'un post-doc, Rémy Sun, qui a commencé en Janvier 2023. Aujourd'hui, Rémy Sun est devenu chercheur permanent dans l'équipe MAASAI et Li Yang a terminé son contrat.

---

<sup>||</sup>. ANR-21-CE23-0032 <https://anr-multitrans.github.io/>



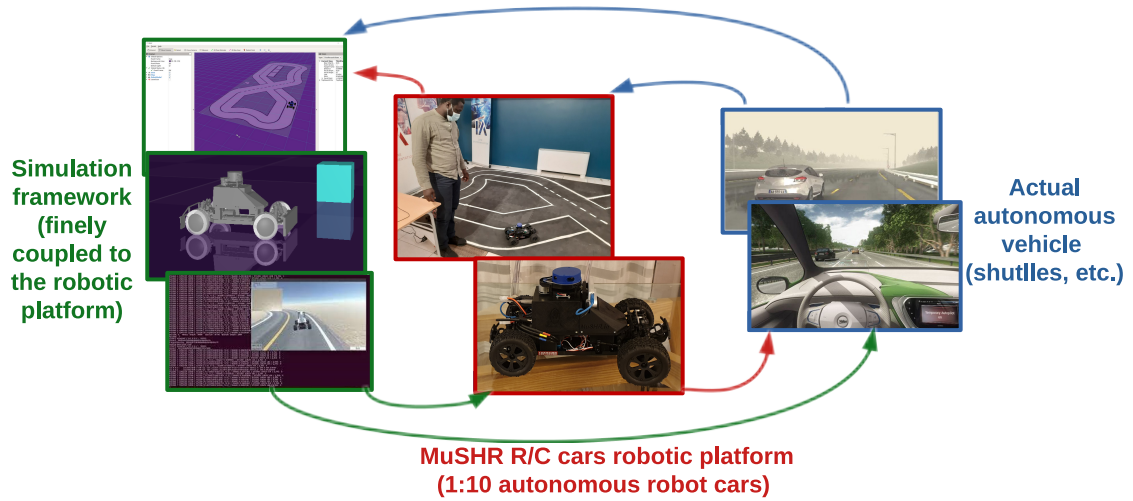


Figure 4.14 – Adaptation de domaine entre le monde simulé, les modèles réduits et les vrais véhicules.



Figure 4.15 – Hackthon à l'école Centrale de Marseille sur notre plateforme (Mars 2023).

J’ai également eu l’occasion de présenter le projet Multitrans au Pôle SCS Smart Mobility en octobre 2022 et de participer à une table ronde sur les expérimentations en conduite autonome dans les Alpes Maritimes au cours du SophIA Summit 2022 (26 novembre 2022).

Dans le cadre du projet ANR Multitrans, nous avons apporté plusieurs contributions concernant les cartes nécessaires à la navigation que nous allons détailler dans les paragraphes suivants. Ces travaux vont continuer dans le cadre de plusieurs projets. Le projet LogIA (LOGistique robotique renforcée par l’IA générative) accepté en 2025 dans l’appel à projets “AAP-IA Generative” de 2024 est géré par Enchanted Tools, a pour partenaires NXP, INRIA, ISIR (Sorbonne Université) et l’Université d’Avignon ainsi que la participation hors demande d’aide de Hugging Face. Un autre projet est en cours de définition et demande de financement : BONSAI (Building Open Networks for Sustainable Artificial Intelligence), porté par Paulo MOURA (IMREDD, Nice), coPI moi-même, avec pour partenaires des membres du réseau universitaire Ulysseus : Université Technique de Košice en Slovaquie (TUCE), Université de Gênes en Italie (UniGE) ainsi que l’entreprise SII (Sophia Antipolis). Nous sommes également impliqués dans le montage d’un projet BPI sur l’appel CORAM avec l’entreprise de navettes autonomes MILLA.

### 4.3.1 Impact des cartes routières pour la prédiction de trajectoire

La prévision de trajectoire pour la conduite autonome nécessite de prendre en compte différentes informations, notamment une carte de l’environnement en haute précision (*HDM*). Depuis plusieurs années, la représentation de ces cartes sous forme de graphe s’est généralisée ([Gao et al., 2020](#); [Liang et al., 2020](#)) car il s’agit d’une représentation simple et facile à utiliser pour des données qui peuvent s’avérer complexes. Ces travaux concernent des bases de données du domaine, généralement fixes. Or, dans la vraie vie, rien n’est statique : des travaux de voirie sont courants conduisant à des modifications régulières d’amplitudes variables (déplacement d’un trottoir, modification d’une intersection en rond-point...). Nous nous sommes intéressés au processus de construction de ces graphes à partir de capteurs pour la prédiction de trajectoire. Nous avons notamment étudié dans ([Sun, Lingrand, & Precioso, 2023](#)) l’impact de la résolution spatiale, les relations entre les graphes et les prédictions de trajectoires ainsi que l’ajout de connaissances dans ces graphes.

À cette fin, nous avons réalisé des expériences approfondies basées sur des graphes (PGP ([Deo, Wolff, & Beijbom, 2022](#)) et LAformer ([Liu et al., 2024](#)) dont les codes sources sont disponibles) sur l’ensemble de données nuScenes ([Caesar et al., 2020](#)). Nous avons montré expérimentalement que la résolution du graphe semble principalement affecter les prévisions de trajectoires plus longues et que des informations du graphe (en particulier les coordonnées des nœuds) sont très utiles pour affiner les trajectoires. Nous avons également conduit des expériences sur les données Argoverse 1.1 ([Chang et al., 2019](#)) avec les méthodes LAformer et LaneGCN ([Liang et al., 2020](#)) qui ont montré le bénéfice de l’apprentissage de certaines caractéristiques persistantes des nœuds (au niveau de la ville ou au niveau des voies routières elles-mêmes) qui peuvent être partagées entre plusieurs scénarios.

### 4.3.2 Prise en compte d’anciennes cartes routières pour faciliter l’estimation de cartes actualisées.

Si les cartes routières en haute définition constituent un élément essentiel de la conduite autonome. Cependant, leur acquisition et leur maintenance sont coûteuses. L’estimation de ces cartes



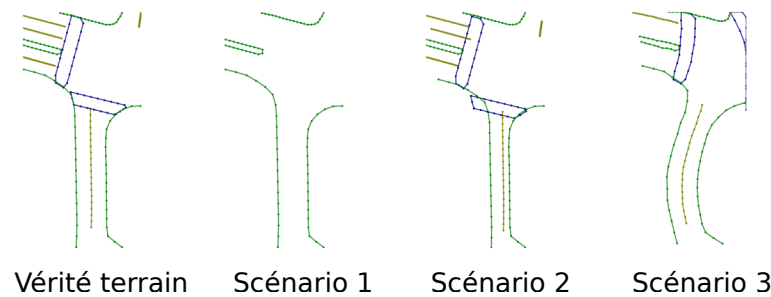


Figure 4.16 – Les scénarios de détériorations des cartes routières haute-définition.

à partir de capteurs promet donc d’alléger considérablement les coûts. Les méthodes précédentes se contentaient au mieux de géolocaliser des cartes de mauvaise qualité ou de prendre en compte un jeu de données générale de cartes connues. Nous avons proposé dans (Sun, Yang, Lingrand, & Precioso, 2025) de prendre en compte les cartes existantes, qu’elles soient imprécises, bruitées ou obsolètes.

Nous avons défini des scénarios raisonnables de détériorations de cartes existantes (voir figure 4.16), correspondant à 3 types de défauts d’anciennes cartes et les avons appliqué à des cartes hautes définitions :

**Scénario 1** : carte ne possédant que les bords des routes

**Scénario 2** : carte bruitée (bruit sur la localisation des éléments et bruit en interne aux éléments sur la localisation des points les définissant).

**Scénario 3** : changements sur le terrain : tracé modifié (déformations), ajout ou retrait d’éléments (passages piétons, délimiteurs de voies)

Nous avons construit MapEX (voir figure 4.17 à partir d’un algorithme classique de détection d’objets dont la version la plus performante à ce moment était MapTRv2 (Liao et al., 2025)). Ces modèles classiques encodent les données des capteurs ainsi que des requêtes (pour chaque objet sa classe et sa localisation, représentée par une liste de  $L$  points 2D.). Un mécanisme d’attention croisé précède la prédiction d’une liste de paires (classe, localisation). Un appariement entre les objets détectés et la vérité terrain permet de calculer un coût à la fois sur la classification et la localisation afin d’apprendre les différentes composantes de ces modèles. Afin de prendre en compte les cartes existantes, nous avons tout d’abord ajouté un module nommé **EX query encoding** dont la fonction est d’encoder de façon statique (sans apprentissage) les éléments existants sous la forme (classe, localisation) et de les concaténer aux requêtes classiques. Les étapes d’attention croisée et de prédiction sont similaires mais le mécanisme d’**Attribution** a également été modifié en ajoutant une étape avant l’appariement habituel par la méthode Hongroise (*Hungarian matching algorithm*) : les éléments existants de cartes sont appariés si la distance entre la prédiction et la précédente valeur est inférieure à un seuil (1 mètre dans notre étude) et sont alors exclus de l’appariement hongrois. Sinon, on considère que cet élément a soit été trop transformé ou a peut-être disparu et il est alors potentiellement apparié par l’algorithme hongrois comme les autres éléments classiques.

Expérimentalement, MapEX apporte des améliorations significatives sur l’ensemble de données nuScenes. Par exemple, MapEX, à partir de cartes bruitées, apporte une amélioration de 38% par rapport au détecteur MapTRv2 sur lequel il est basé et de 8% par rapport à l’état de l’art à ce moment (table 4.18).

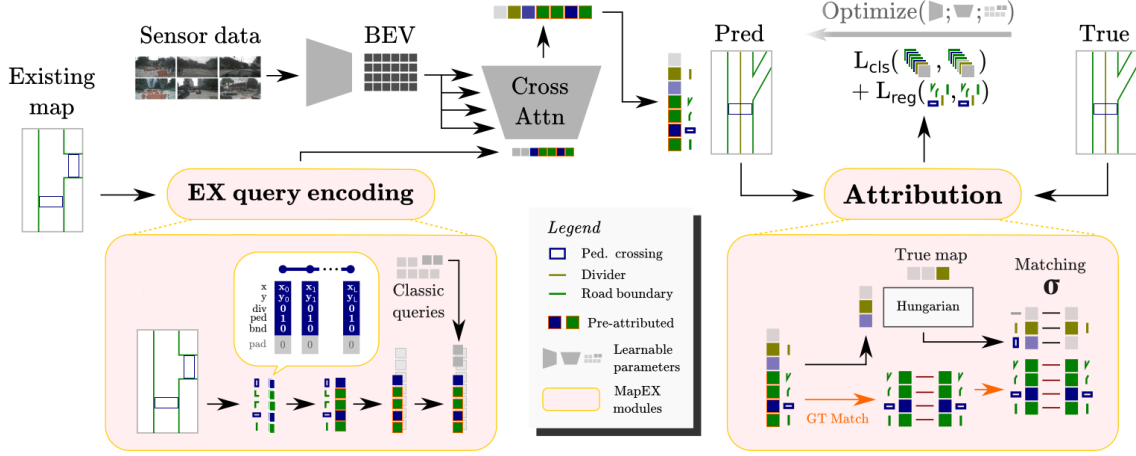


Figure 4.17 – Notre modèle d’estimation de nouvelles cartes routières haute définition à partir de capteurs et en utilisant des cartes existantes. Au modèle MapTRv2 (Liao et al., 2025), nous avons ajouté 2 modules : **EX query encoding** qui encode les cartes existantes sous la forme de requêtes (*query*) qui sont concaténées aux autres requêtes classiques et **Attribution** qui concerne une pré-attribution des des prédictions connues aux éléments de cartes par un algorithme Hongrois. Les éléments grisés sur ce schéma nécessitent un apprentissage.

Method	Backbone	Epoch	Extra info	Average Precision at {0.5m, 1.0m, 1.5m}			
				$AP_{divider}$	$AP_{ped}$	$AP_{boundary}$	$mAP$
Previous methods							
MapTRv2 <sup>†</sup>	R50	24	✗	62.4	59.8	62.4	61.5
MapTRv2	V2-99	110	Depth pretrain	73.7	71.4	75.0	73.4
MapNeXT	II-H	110	Foundation backbone	79.3	77.4	78.8	78.5
Our method							
MapEX-S1	R50	24	Map w/ only boundaries	66.1 ± 0.6	62.5 ± 0.4	<b>99.9 ± 0.1</b>	76.2 ± 0.1
MapEX-S2a	R50	24	Map w/ element shift	82.5 ± 1.0	78.4 ± 0.8	93.5 ± 0.4	84.8 ± 0.3
MapEX-S2b	R50	24	Map w/ point noise	78.4 ± 0.1	62.1 ± 0.6	72.4 ± 0.4	70.9 ± 0.3
MapEX-S3a	R50	24	Outdated maps	84.6 ± 0.3	74.1 ± 0.6	99.1 ± 0.1	85.9 ± 0.2
MapEX-S3b	R50	24	50% outdated maps	<b>92.8 ± 0.1</b>	<b>87.2 ± 0.1</b>	<u>99.3 ± 0.2</u>	<b>93.1 ± 0.1</b>

Figure 4.18 – Comparaison des métriques selon les éléments routiers pour le modèle de base (MapTRv2<sup>†</sup>) et 2 autres modèles concurrents avec notre méthode MapEx selon les différents scénarios envisagés. Dans tous les scénarios MapEx obtient les meilleures performances. Selon les scénarios, l’état de l’art peut-être dépassé (meilleurs résultats en gras, second meilleur souligné)

# CHAPITRE 5

## Réflexions et perspectives

Lorsque l'on enseigne l'apprentissage profond, la limitation en temps et en ressources de calcul nous conduit souvent à réaliser les travaux pratiques sur des datasets ou des extraits de datasets bien connus : MNIST, Pascal VOC, ImageNet, COCO quand ce n'est pas sur des datasets synthétiques générés pour l'occasion, éventuellement artificiellement bruités. Ces datasets facilitent des expérimentations simples et permettent de découvrir le fonctionnement d'algorithmes ou modèles dans un cadre relativement simple.

Mais si on veut se confronter à des utilisations réalistes, le travail est bien différent, que ce soit au niveau des annotations ou des données elles-mêmes : différents aspects de l'apprentissage doivent être remis en cause et adaptés. C'est aussi de vraies applications que se dégagent des problématiques nouvelles. Je vais détailler ici quelques réflexions issues de travaux précédents.

### 5.1 Réflexions concernant les travaux passés

**Un écart majeur entre les bases de données classiques et les données de terrain :** Lors des études en biologie marine, on a pu observer que la labellisation des poissons dans les vidéos par des biologistes marins est différente de celle qu'on rencontre habituellement en vision par ordinateur dans des datasets comme COCO par exemple : il n'est pas utile de labelliser/entourer chaque poisson d'un volumineux banc de poisson alors qu'un spécimen solitaire ou rare ne devra pas être oublié. Les méthodes et métriques d'évaluation classiques doivent être modifiées et adaptées à cette configuration. Labelliser des images de charbons de bois est une tâche que seuls des anthracologues peuvent réaliser. Sur des charbons modernes, puisque l'arbre a pu être observé, la labélisation est certaine alors que pour des charbons archéologiques, certaines espèces, genres et même familles ne peuvent pas être déterminés, même par des anthracologues. La construction d'une base moderne d'images est une tâche longue et fastidieuse d'autant plus qu'un arbre ne peut pas être uniquement représenté par trois coupes mais il faut généralement plusieurs images pour capter les différentes caractéristiques anatomiques observables au microscope. Pour être exhaustif, même en se limitant à une région géographique, le nombre d'espèces est très important. Sachant que de plus, certaines espèces d'arbres sont plus accessibles que d'autres, la collecte d'échantillons est déséquilibrée et par conséquent la base de données également.

**Le meilleur modèle de l'état de l'art n'est pas forcément le meilleur modèle pour un cas pratique précis :** Il existe actuellement des procédés classiques de transfert d'apprentissage d'un

modèle, par exemple appris sur la base ImageNet, vers une autre base d'images : on récupère la partie descriptive à laquelle on ajoute un tête de classification à apprendre, avant d'éventuellement affiner la partie descriptive. Les différents modèles descriptifs des images sont testés sur des ensemble de données grandissant. Chaque nouveau modèle (CNN ou transformeur) publié doit comparer des métriques par rapport aux modèles précédents sur les mêmes ensembles de données. Lorsque l'on étudie, par exemple, la suite des modèles de détection d'objets YOLO, chaque version est ainsi meilleure que la précédente, sur des métriques et données communes. Cependant, cela ne signifie pas que ces modèles soient toujours meilleur, quelque soit la métrique ou les données. Certains modèles YOLO plus ancien peuvent s'adapter plus facilement à des types de données différentes. Nous en avons fait l'expérience lors de la thèse de Kilian Bürgi ([Bürgi, 2025](#)). Des modèles descriptifs par CNN pour la classification d'images de charbons archéologiques se sont révélés plus pertinents que d'autres modèles plus performants sur des images naturelles ou que des modèles récents basés sur des transformeurs comme DINO v2 ou v3 ([Oquab et al., 2024](#); [Siméoni et al., 2025](#)).

**Des efforts pour prendre en compte des bases de données non équilibrées :** Lorsque peu de données sont disponibles, différentes méthodes (*Few Shot Learning*) ont été étudiées. Depuis quelques années, des travaux considèrent des apprentissages dont les classes sont très déséquilibrées allant de classes aux données très nombreuses à des classes comportant très peu de données. On parle même d'apprentissage à longue traine lorsque que 20% des classes comportent 80% des données.

**Les interactions avec des experts d'autres domaines : difficiles mais enrichissantes :** J'ai eu la chance d'interagir avec des spécialistes nombreux domaines pendant ma carrière : neurologie, imagerie cardiaque, vidéos événementielles, notamment sportives, biologie marine, archéologie ou industrie de la navigation autonome. Diriger les recherches vers des applications concrètes permet de donner naturellement un sens à ces recherches. Se confronter à d'autres domaines n'est pas toujours aisé : les attentes sont différentes, la communication peut être difficile lorsqu'un même vocabulaire peut être utilisé avec des sens différents, lorsque les échelles de temps ne sont pas les mêmes. De plus, il n'est pas forcément simple d'accéder aux connaissances des experts : même si seul un médecin peut interpréter des examens médicaux, il n'a pas forcément le temps d'annoter toutes les images que l'on souhaiterait. Mais cela est enrichissant et permet d'étudier des aspects qui ne l'auraient pas été sans cela et permettre d'utiliser ces approches ensuite dans d'autres domaines.

## 5.2 Et maintenant ?

- apprendre aussi bien en utilisant moins de données
- construire des modèles ingérant des connaissances :
  - pour apprendre plus vite (du coup moins de données)
  - des connaissances, même pas toujours exactes
  - des connaissances hiérarchiques
- comprendre ce que fait le modèle

Depuis plusieurs années, les modèles en apprentissage automatique grossissent, entraînant des coûts en données, en mémoire et en énergie toujours plus importants alors qu'une prise de

conscience de la nécessité de restreindre nos consommations en minéraux rares et énergétiques se répand également. L'IA doit rester un outil et non pas une croyance en une divinité numérique.

Comme déjà dit auparavant, diriger les recherches par les applications donne un sens naturel aux recherches menées. Cela doit se faire en collaboration avec les experts du domaine d'application pour garantir une compréhension des problèmes à résoudre et une adéquation des travaux avec les buts poursuivis (on ne comptera jamais exactement toutes les anchois d'un volumineux banc de d'anchois). La richesse des domaines d'applications se situe au niveau de la connaissance des experts. Cependant, l'utilisation de ces connaissances n'est pas directe dans les modèles d'apprentissage. La marche d'amélioration importante dans les années 2010 en classification d'images par réseaux convolutionnels (compétition ImageNet) et leur facile adaptation à d'autres domaines en imagerie (images de coloscopie par exemple) a fait négliger la connaissance du domaine.

De plus, à une époque où nombreux sont ceux qui méprise l'acquisition de connaissances au profit d'agents conversationnels basés sur de grands modèles fondations, revaloriser la connaissance humaine me semble important.

La difficulté à représenter des connaissances vient également du fait que ces dernières peuvent être de nature et de précision très différente. On a vu dans nos études sur la navigation autonome l'utilisation de différentes informations même imprécises, bruitées ou dépassées concernant les cartes routières. Dans l'étude des charbons, notre architecture dépend d'informations fournies par les anthracologues : c'est en leur demandant d'identifier des espèces devant nous que nous avons compris leur processus d'identification. Par exemple, les informations étant très différentes entre les 3 types de couches, nous avons choisi et conserver ces trois informations. Mais ce n'est qu'un exemple. L'équilibre entre connaissances des experts et aide pour les experts est un équilibre difficile.

Parmi les différentes thématiques que je souhaite explorer prochainement, je vais détailler la prise en compte des connaissances du domaine dans l'apprentissage et l'adéquation de nos modèles aux besoins des applications.

Lors de la thèse de Kilian Bürgi (?), nous avons mis en place un critère d'évaluation correspondant aux annotations des biologistes marins et également

## 5.3 Prise en compte des connaissances

**la prise en compte des relations hiérarchiques entre les classes :** le mélèze d'Europe (*Larix decidua*) est une espèce du genre *Larix* qui fait partie de la famille des Pinaceae. Certains détails anatomiques sont communs aux Pinaceae, d'autres aux *Larix* et enfin d'autres sont uniques pour l'espèce.

**les détails anatomiques sont connus** et répertoriés sous la forme de critères IAWA\* (International Association of Wood Anatomists)

PEPR NumPEX AAP accepté en novembre 2025 mené par Laetitia Grimaldi SAGE-HPC Smart strateGies for multi-fidelity optimization in Exascale HPC Environments Partenaires : INRIA () et Université de Strasbourg (équipe Semosis) Résolution d'un problème d'optimisation issu de la Physique. Simplification de la fonction de coût. Approximation de la fonction de coût. Exascale Comment orchestrer tout cela (niveau de fidélité, méthodes d'optimisation, allocation de ressources exascale.)?

---

\*, <https://www.iawa-website2.org/>



# Bibliographie

---

- Barbosa, A., Marinho, T., Martin, N., & Hovakimyan, N. (2020). Multi-Stream CNN for Spatial Resource Allocation : a Crop Management Application. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (p. pp. 258-266). Retrieved from [https://openaccess.thecvf.com/content\\_CVPRW\\_2020/papers/w5/Barbosa\\_Multi-Stream\\_CNN\\_for\\_Spatial\\_Resource\\_Allocation\\_A\\_Crop\\_Management\\_Application\\_CVPRW\\_2020\\_paper.pdf](https://openaccess.thecvf.com/content_CVPRW_2020/papers/w5/Barbosa_Multi-Stream_CNN_for_Spatial_Resource_Allocation_A_Crop_Management_Application_CVPRW_2020_paper.pdf) doi:10.1109/CVPRW50498.2020.00037
- Blanc, K. (2018). *Description de contenu vidéo : mouvements et élasticité temporelle* (Theses, COMUE Université Côte d’Azur (2015 - 2019)). Retrieved from <https://theses.hal.science/tel-02010091>
- Blanc, K., Lingrand, D., Paladini, A., Coviello, L., Mitrev, D., Söhler, E., ... Precioso, F. (2019, May). Analysis of temporal alignment for Video Classification. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. Lille, France: IEEE. Retrieved from <https://hal.science/hal-05173364> doi: 10.1109/FG.2019.8756622
- Blanc, K., Lingrand, D., & Precioso, F. (2014, November). Fish Species Recognition from Video using SVM Classifier. In *ACM Workshop on Multimedia Analysis for Ecological Data in conjunction with ACM Multimedia*. Orlando, United States. Retrieved from <https://hal.science/hal-01323227> doi: 10.1145/2661821.2661827
- Blanc, K., Lingrand, D., & Precioso, F. (2017, July). SINGLETS : Multi-Resolution Motion Singularities for Soccer Video Abstraction. In *Proceedings of the Workshop CVsports (in conjunction with CVPR)*. Honolulu (Hawaii), United States. Retrieved from <https://hal.science/hal-01540342>
- Bürgi, K. (2025). *Détection et Suivi de la Biodiversité Marine par Intelligence Artificielle* (Theses, Université Côte d’Azur, France). Retrieved from <https://theses.hal.science/tel->
- Bürgi, K., Bouveyron, C., Lingrand, D., Derijard, B., Precioso, F., & Sabourault, C. (2025, March). Towards a fully automated underwater census for fish assemblages in the Mediterranean Sea. *Ecological Informatics*, 85, 102959. Retrieved from <https://hal.science/hal-04896273> doi: 10.1016/j.ecoinf.2024.102959
- Bürgi, K., Petiot, S., Sabourault, C., Sun, R., Lingrand, D., Derijard, B., & Bouveyron, C. (2025, October). CleverFish : An AI-driven Platform to Monitor and Explore Marine Ecological Resources. In *ECAI 2025 - European Conference of Artificial Intelligence*. Bologna, Italy. Retrieved from <https://hal.science/hal-05200321>
- Bürgi, K., Sun, R., Bouveyron, C., Lingrand, D., Dérijard, B., Precioso, F., & Sabourault, C. (2026). Automated Counting of Fish in moving Diver Operated Videos (DOV) for Biodiversity Assessments. *Methods in Ecology and Evolution*. Retrieved from <https://hal.science/hal-04865293> doi: 10.1111/2041-210x.70283

- Bürge, K., Sun, R., Bouveyron, C., Lingrand, D., Dériard, B., Precioso, F., & Sabourault, C. (2025). Leveraging computer vision for efficient and scalable biodiversity monitoring in marine ecosystems : A multi-year study on ecologically important species. *to be submitted soon to Global Change Biology*.
- Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... Beijbom, O. (2020, June). nuScenes : A Multimodal Dataset for Autonomous Driving . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 11618-11628). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from <https://doi.ieeeecomputersociety.org/10.1109/CVPR42600.2020.01164> doi: 10.1109/CVPR42600.2020.01164
- Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., ... Hays, J. (2019). Argoverse : 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 8740-8749). doi: 10.1109/CVPR.2019.00895
- Charnoz, A., Lingrand, D., & Montagnat, J. (2003, June). A levelset based method for segmenting the heart in 3D+T gated SPECT images. In *Proceedings of the Second International Workshop on Functional Imaging and Modeling of the Heart* (p. 50-59). Lyon, France. Retrieved from <https://hal.science/hal-00691656>
- Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., & Phan, H. A. (2015). Tensor decompositions for signal processing applications : From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2), 145-163. doi: 10.1109/MSP.2013.2297439
- De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4), 1253-1278. Retrieved from <https://doi.org/10.1137/S0895479896305696> doi: 10.1137/S0895479896305696
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet : A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848
- Deo, N., Wolff, E., & Beijbom, O. (2022, 08–11 Nov). Multimodal trajectory prediction conditioned on lane-graph traversals. In A. Faust, D. Hsu, & G. Neumann (Eds.), *Proceedings of the 5th Conference on Robot Learning* (Vol. 164, pp. 203–212). PMLR. Retrieved from <https://proceedings.mlr.press/v164/deo22a.html>
- Dojat, M., Péligrini-Issac, M., Ahmad, F., Barillot, C., Batrancourt, B., Gaignard, A., ... Wali, B. (2011, June). NeuroLOG : A framework for the sharing and reuse of distributed tools and data in neuroimaging. In *Organization for Human Brain Mapping (HBM'11)*. Quebec, Canada. Retrieved from <https://inria.hal.science/hal-00813796>
- Duan, L.-Y., Xu, M., Tian, Q., Xu, C.-S., & Jin, J. (2005). A unified framework for semantic shot classification in sports video. *IEEE Transactions on Multimedia*, 7(6), 1066-1083. doi: 10.1109/TMM.2005.858395
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88, 303-338. Retrieved from <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html> doi: 10.1007/s11263-009-0275-4



- Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020, June). VectorNet : Encoding HD Maps and Agent Dynamics From Vectorized Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Germain-Renaud, C., Cady, A., Gauron, P., Jouvin, M., Loomis, C., Martyniak, J., ... Sebag, M. (2011, May). The Grid Observatory. In I. C. S. Press (Ed.), *IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*. Newport Beach, United States. Retrieved from <https://inria.hal.science/inria-00586502>
- Gibaud, B., Ahmad, F., Barillot, C., Michel, F., Wali, B., Batrancourt, B., ... Pélérini-Issac, M. (2011, June). A federated system for sharing and reuse of images and image processing tools in neuroimaging. In *Computer Assisted Radiology and Surgery (CARS'11)*. Berlin, Germany. Retrieved from <https://hal.science/hal-00690921>
- Glatard, T., Lingrand, D., Montagnat, J., & Riveill, M. (2007, May). Impact of the execution context on Grid job performances. In *International Workshop on Context-Awareness and Mobility in Grid Computing (WCAMG'07)* (pp. 713–718). Rio de Janeiro, Brazil: IEEE. Retrieved from <https://hal.science/hal-00459267> doi: 10.1109/CCGRID.2007.62
- Glatard, T., Montagnat, J., Emsellem, D., & Lingrand, D. (2008, July). A Service-Oriented Architecture enabling dynamic services grouping for optimizing distributed workflows execution. *Future Generation Computer Systems*, 24(7), 720–730. Retrieved from <https://hal.science/hal-00459808> doi: 10.1016/j.future.2008.02.011
- Glatard, T., Montagnat, J., Lingrand, D., & Pennec, X. (2008, August). Flexible and efficient workflow deployment of data-intensive applications on grids with MOTEUR. *International Journal of High Performance Computing Applications*, 22(3), 347–360. Retrieved from <https://hal.science/hal-00459130> (Special issue on Workflow Systems in Grid Environments) doi: 10.1177/1094342008096067
- Glatard, T., Montagnat, J., & Pennec, X. (2006, February). Probabilistic and dynamic optimization of job partitioning on a grid infrastructure. In *Proceedings of the 14th Euromicro Conference on Parallel, Distributed and network-based Processing* (p. 231–238). Montbéliard-Sochaux, France. Retrieved from <https://hal.science/hal-00683203> doi: 10.1109/PDP.2006.61
- Glatard, T., Montagnat, J., & Pennec, X. (2007, May). Optimizing jobs timeouts on clusters and production grids. In *Proceedings of the International Symposium on Cluster Computing and the Grid* (p. 100–107). Rio de Janeiro, Brazil. Retrieved from <https://hal.science/hal-00683172> doi: 10.1109/CCGRID.2007.78
- Harris, C., & Stephens, M. (1988). A Combined Corner and Edge Detector. In *Alvey Vision Conference (vol. 15)*.
- Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., ... Lerchner, A. (2017). beta-VAE : Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26*. OpenReview.net.
- Jaderberg, M., Simonyan, K., Zisserman, A., & kavukcuoglu, k. (2015). Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett

- (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. Retrieved from [https://proceedings.neurips.cc/paper\\_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf)
- Kihl, O., Tremblais, B., & Augereau, B. (2008). Multivariate orthogonal polynomials to extract singular points. In *2008 15th IEEE International Conference on Image Processing* (p. 857-860). doi: 10.1109/ICIP.2008.4711890
- Kim, T.-K., & Cipolla, R. (2007). Gesture Recognition Under Small Sample Size. In Y. Yagi, S. B. Kang, I. S. Kweon, & H. Zha (Eds.), *Computer Vision – ACCV 2007* (pp. 335–344). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-76386-4\_31
- Kotsia, I., Guo, W., & Patras, I. (2012). Higher rank support tensor machines for visual recognition. *Pattern Recognition*, 45(12), 4192-4203. Retrieved from <https://www.sciencedirect.com/science/article/pii/S003132031200218X> doi: <https://doi.org/10.1016/j.patcog.2012.04.033>
- Laptev, & Lindeberg. (2003). Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision* (Vol. 1, p. 432-439). doi: 10.1109/ICCV.2003.1238378
- Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Proc. of European Conference on Computer Vision – ECCV 2020* (pp. 541–556). Cham: Springer International Publishing.
- Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., ... Wang, X. (2025). MapTRv2 : An End-to-End Framework for Online Vectorized HD Map Construction. *International Journal of Computer Vision (IJCV)*, 133, 1352–1374. doi: <https://doi.org/10.1007/s11263-024-02235-z>
- Lingrand, D. (1999). *Analyse adaptative du mouvement dans des séquences monoculaires non calibrées* (Theses, University of Nice - Sophia Antipolis, France). Retrieved from <https://theses.hal.science/tel-00459424>
- Lingrand, D. (2002a). An Exhaustive Study of Particular Cases Leading to Robust and Accurate Motion Estimation. *Computer Vision and Image Understanding*, 85(3), 159–188. Retrieved from <https://hal.science/hal-00459253> doi: 10.1006/cviu.2002.0966
- Lingrand, D. (2002b). Minimal parameterization of Fundamental Matrices using motion and camera properties. *Robotics and Autonomous Systems*, 39(3-4), 169–179. Retrieved from <https://hal.science/hal-00459258> doi: 10.1016/S0921-8890(02)00202-6
- Lingrand, D., Charnoz, A., Koulibaly, M., Darcourt, J., & Montagnat, J. (2004, May). Toward accurate segmentation of the LV myocardium and chamber for volumes estimation in gated SPECT sequences. In *Proceedings of the European Conference on Computer Vision* (Vol. LNCS 3024, p. 1-10). Prague, Czech Republic. Retrieved from <https://hal.science/hal-00691648>
- Lingrand, D., Glatard, T., & Montagnat, J. (2007, August). Taking into account the execution context to optimize medical image databases indexation. In *Second Singaporean-French Biomedical Imaging Workshop*. Lyon, France. Retrieved from <https://hal.science/hal-00461626>
- Lingrand, D., Glatard, T., & Montagnat, J. (2009, October). Modeling the latency on production grids with respect to the execution context. *Parallel Computing*, 35(10-11), 493–511. Retrieved from <https://hal.science/hal-00459261> doi: 10.1016/j.parco.2009.07.003

- Lingrand, D., & Montagnat, J. (2005, June). Levelset and B-spline deformable model techniques for image segmentation : a pragmatic comparative study. In *14th Scandinavian Conference on Image Analysis* (pp. 25–34). Joensuu, Finland. Retrieved from <https://hal.science/hal-00460711> doi: 10.1007/11499145\_4
- Lingrand, D., & Montagnat, J. (2010, December). Efficient resubmission strategies to design robust grid production environments. In *Proceedings of the IEEE e-Science (e-Science)* (p. 198-205). Brisbane, Australia: IEEE. Retrieved from <https://hal.science/hal-00677824> doi: 10.1109/eScience.2010.11
- Lingrand, D., Montagnat, J., Collins, L. D., & Gotman, J. (2001, June). Compensating Small Head Displacements for an accurate fMRI Registration. In *Proceedings of the Scandinavian Conference on Image Analysis* (p. 10-16). Bergen, Norway. Retrieved from <https://hal.science/hal-00691688>
- Lingrand, D., Montagnat, J., & Glatard, T. (2008, February). Estimation of latency on production grid over several weeks. In *ICT4Health* (p. 4). Manila, Philippines. Retrieved from <https://hal.science/hal-00461615>
- Lingrand, D., Montagnat, J., & Glatard, T. (2009, June). Modeling user submission strategies on production grids. In *International Symposium on High Performance Distributed Computing* (p. 121-130). Munchen, Germany. Retrieved from <https://hal.science/hal-00459073> doi: 10.1145/1551609.1551633
- Lingrand, D., Montagnat, J., Martyniak, J., & Colling, D. (2009, May). Analyzing the EGEE production grid workload : application to jobs submission optimization. In *14th Workshop on Job Scheduling Strategies for Parallel Processing* (Vol. LNCS 5798, pp. 37–58). Roma, Italy: Springer. Retrieved from <https://hal.science/hal-00459077> doi: 10.1007/978-3-642-04633-9
- Lingrand, D., Montagnat, J., Martyniak, J., & Colling, D. (2010, March). Optimization of jobs submission on the EGEE production grid : modeling faults using workload. *Journal of Grid Computing*, 8(2), 305-321. Retrieved from <https://hal.science/hal-00677775> doi: 10.1007/s10723-010-9151-2
- Liu, M., Cheng, H., Chen, L., Broszio, H., Li, J., Zhao, R., ... Yang, M. Y. (2024). Laformer : Trajectory prediction for autonomous driving with lane-aware scene constraints. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (p. 2039-2049). doi: 10.1109/CVPRW63382.2024.00209
- Lopes, A. P. B., Oliveira, R. S., de Almeida, J. M., & de A. Araujo, A. (2009). Comparing alternatives for capturing dynamic information in bag-of-visual-features approaches applied to human actions recognition. In *2009 IEEE International Workshop on Multimedia Signal Processing* (p. 1-6). doi: 10.1109/MMSP.2009.5293303
- Lopez, S. (2017). *Content based images retrieval based on implicit gaze annotations* (Theses, COMUE Université Côte d’Azur (2015 - 2019)). Retrieved from <https://theses.hal.science/tel-01724391>
- Lopez, S., Revel, A., Lingrand, D., & Precioso, F. (2015, September). One gaze is worth ten thousand (key-)words. In *IEEE International Conference on Image Processing (ICIP), 2015* (pp. 3150–3154). Québec, Canada. Retrieved from <https://hal.science/hal-01323204> doi: 10.1109/ICIP.2015.7351384

- Lopez, S., Revel, A., Lingrand, D., & Precioso, F. (2017, September). Handling noisy labels in gaze-based CBIR system. In *Proceedings of the Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Antwerpen, Belgium. Retrieved from <https://hal.science/hal-01635028>
- Lopez, S., Revel, A., Lingrand, D., Precioso, F., Dusaucy, V., & Giboin, A. (2016, June). Catching Relevance in One Glimpse : Food or Not Food ? In *ACM Advances in Visual Interfaces (AVI)* (p. 324-325). Bari, Italy. Retrieved from <https://hal.science/hal-01330873> doi: 10.1145/2909132.2926078
- Malleus, L., Fisichella, T., Lingrand, D., Precioso, F., Gros, N., Noutary, Y., ... Samoun, L. (2017, October). KPPF : Keypoint-based Point-Pair-Feature for scalable automatic global registration of large RGB-D scans. In *ICCV 2017 proceedings*. Venice, Italy. Retrieved from <https://hal.science/hal-01635013>
- Michel, F., Gaignard, A., Ahmad, F., Barillot, C., Batrancourt, B., Dojat, M., ... Wali, B. (2010, June). Grid-wide neuroimaging data federation in the context of the NeuroLOG project. In T. S. . I. B. . V. B. . T. G. . Y. Legré (Ed.), *Proceedings of the HealthGrid* (Vol. 159, p. 112-123). Paris, France: IOS press. Retrieved from <https://inserm.hal.science/inserm-00512799> doi: 10.3233/978-1-60750-583-9-112
- Montagnat, J., Frohner, A., Jouvenot, D., Pera, C., Kunszt, P., Koblitiz, B., ... Farkas, Z. (2008, March). A Secure Grid Medical Data Manager Interfaced to the gLite Middleware. *Journal of Grid Computing*, 6(1), 45-59. Retrieved from <https://hal.science/hal-00683989> doi: 10.1007/s10723-007-9088-2
- Montagnat, J., Gaignard, A., Lingrand, D., Rojas Balderrama, J., Collet, P., & Lahire, P. (2008, June). NeuroLOG : a community-driven middleware design. In *HealthGrid* (pp. 49-58). Chicago, United States: IOS Press. Retrieved from <https://hal.science/hal-00461611>
- Montagnat, J., Glatard, T., & Lingrand, D. (2006, June). Data composition patterns in service-based workflows. In *Proceedings of the Workshop on Workflows in Support of Large-Scale Science* (p. 1-10). Paris, France. Retrieved from <https://hal.science/hal-00683193>
- Niaf, E., Flamary, R., Rouvière, O., Lartizien, C., & Canu, S. (2014, March). Kernel-Based Learning From Both Qualitative and Quantitative Labels : Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging. *IEEE Transactions on Image Processing*, 23(3), 979 - 991. Retrieved from <https://hal.science/hal-00977046> doi: 10.1109/TIP.2013.2295759
- Olabarriaga, S. D., Lingrand, D., & Montagnat, J. (2008, September). MICCAI-Grid Workshop. In S. D. Olabarriaga, D. Lingrand, & J. Montagnat (Eds.), *Medical imaging on grids : achievements and perspectives*. New-York, United States. Retrieved from <https://inria.hal.science/hal-00694970>
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2024). DINOv2 : Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Retrieved from <https://openreview.net/forum?id=a68SUt6zFt>
- Pernod, E., Souplet, J.-C., Rojas Balderrama, J., Lingrand, D., & Pennec, X. (2008, September). Multiple Sclerosis Brain MRI Segmentation Workflow deployment on the EGEE

- Grid. In *MICCAI-Grid Workshop* (pp. 55–64). New York, United States. Retrieved from <https://hal.science/hal-00459138>
- Phan, A. H., & Cichocki, A. (2010). Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear Theory and Its Applications, IEICE*, 1(1), 37-68. doi: 10.1587/nolta.1.37
- Pinna, M., Zangaro, F., Saccomanno, B., Scalone, C., Bozzeda, F., Fanini, L., & Specchia, V. (2023). An overview of ecological indicators of fish to evaluate the anthropogenic pressures in aquatic ecosystems : from traditional to innovative dna-based approaches. *Water*, 15(5). doi: <https://doi.org/10.3390/w15050949>
- Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017, July). PointNet : Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rojas Balderrama, J., Lingrand, D., Pernod, E., Souplet, J.-C., Pennec, X., & Montagnat, J. (2008, September). NeuroLOG : Neuroscience Application Workflows Execution on the EGEE Grid. In *EGEE conference*. Istanbul, Turkey. Retrieved from <https://hal.science/hal-00461625>
- Rojas Balderrama, J., Montagnat, J., & Lingrand, D. (2010, July). jGASW : A Service-Oriented Framework Supporting High Throughput Computing and Non-functional Concerns. In *Proceedings of the IEEE International Conference on Web Services* (p. 691-694). Miami, United States: IEEE. Retrieved from <https://hal.science/hal-00677819> doi: 10.1109/ICWS.2010.59
- Rosa da Silva, N., Deklerck, V., Baetens, J., & et al. (2022). Improved wood species identification based on multi-view imagery of the three anatomical planes. *Plant Methods*, 18(79). doi: <https://doi.org/10.1186/s13007-022-00910-1>
- Samoun, L., Fisichella, T., Lingrand, D., Malleus, L., & Precioso, F. (2018, October). An Interactive Content-Based 3D Shape Retrieval System for on-Site Cultural Heritage Analysis. In *ICIP 2018 25th IEEE International Conference on Image Processing* (p. 1043-1047). Athens, France: IEEE. Retrieved from <https://hal.science/hal-05173388> doi: 10.1109/ICIP.2018.8451546
- Seeland, M., & P., M. (2021). Multi-view classification with convolutional neural networks. *PLoS ONE*, 16(1). doi: <https://doi.org/10.1371/journal.pone.0245230>
- Seitz, L., Montagnat, J., Pierson, J.-M., Oriol, D., & Lingrand, D. (2005, April). Authentication and autorisation prototype on the microgrid for medical data management. In *Technology and Informatics* (Vol. 112, p. 222-233). Oxford, United Kingdom. Retrieved from <https://hal.science/hal-00691618>
- Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., ... Bojanowski, P. (2025). *Dinov3*. Retrieved from <https://arxiv.org/abs/2508.10104>
- Spampinato, C., Palazzo, S., Joalland, P.-H., Paris, S., Glotin, H., Blanc, K., ... Precioso, F. (2016, February). Fine-Grained Object Recognition in Underwater Visual Data. *Multimedia Tools and Applications*, 75(3), 1701-1720. Retrieved from <https://hal.science/hal-01323129> doi: 10.1007/s11042-015-2601-x
- Sun, R., Lingrand, D., & Precioso, F. (2023, October). Exploring the Road Graph in Trajectory Forecasting for Autonomous Driving. In *IEEE Xplore* (p. 71-80). Paris, France: IEEE. Retrieved from <https://hal.science/hal-04385182> doi: 10.1109/ICCVW60793.2023.00014



- Sun, R., Yang, L., Lingrand, D., & Precioso, F. (2025, February). Mind the map ! Accounting for existing maps when estimating online HDMaps from sensors. In *Winter conference on Applications of Computer Vision - WACV 2025*. Tucson (USA), United States. Retrieved from <https://hal.science/hal-04385135> doi: 10.48550/arXiv.2311.10517
- Théry-Parisot, I., Corneli, M., Lingrand, D., Thiebault, S., Tengberg, M., Garberi, P., ... Fangnon, D.-D. (2025, June). Artificial intelligence for identification of wood and charcoal in archaeological and palaeological perspectives. In *Of People & Trees : New Directions in Anthracology and the Archaeological History of Human-woodlands Interactions. 10th World Archaeological Congress Darwin, Australia*. Darwin, Australia. Retrieved from <https://hal.science/hal-05131718>
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015, December). Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 4489-4497). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from <https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.510> doi: 10.1109/ICCV.2015.510
- Vasilescu, M. A. O., & Terzopoulos, D. (2002). Multilinear Analysis of Image Ensembles : TensorFaces. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *European Conference on Computer Vision (ECCV 2002)* (pp. 447–460). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-47969-4\_30
- Viéville, T., & Lingrand, D. (1999). Using Specific Displacements to analyze Motion without Calibration. *International Journal of Computer Vision (IJCV)*, 31(1), 5–29.
- Viéville, T., Lingrand, D., & Gaspard, F. (2001, October). Implementing a multi-model estimation method. *International Journal of Computer Vision*. Retrieved from <https://inria.hal.science/inria-00000172>
- Wan, J., Li, S. Z., Zhao, Y., Zhou, S., Guyon, I., & Escalera, S. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (p. 761-769). doi: 10.1109/CVPRW.2016.100
- Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7 : Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *2023 IEEE/CVF conference on computer vision and pattern recognition (cvpr)* (p. 7464-7475). doi: 10.1109/CVPR52729.2023.00721
- Wang, H., Kläser, A., Schmid, C., & Cheng-Lin, L. (2011, June). Action Recognition by Dense Trajectories. In *CVPR 2011 - IEEE Conference on Computer Vision & Pattern Recognition* (p. 3169-3176). Colorado Springs, United States: IEEE. Retrieved from <https://inria.hal.science/inria-00583818> doi: 10.1109/CVPR.2011.5995407
- Wang, X., Kumar, D., Thome, N., Cord, M., & Precioso, F. (2015, June). RECIPE RECOGNITION WITH LARGE MULTIMODAL FOOD DATASET. In *IEEE International Conference on Multimedia & Expo (ICME), workshop CEA*. Turin, Italy. Retrieved from <https://hal.science/hal-01196959> doi: 10.1109/ICMEW.2015.7169757
- Wang, Y., Guo, J., Gao, H., & Yue, H. (2021). UIEC<sup>2</sup>-Net : CNN-based underwater image enhancement using two color space. *Signal Processing : Image Communication*, 96, 116250. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0923596521001004> doi: <https://doi.org/10.1016/j.image.2021.116250>

- Ye, Q., Huang, Q., Gao, W., & Jiang, S. (2005). Exciting event detection in broadcast soccer video with mid-level description and incremental learning. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (p. 455–458). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1101149.1101250> doi: 10.1145/1101149.1101250
- Zhao, Q., Caiafa, C. F., Mandic, D., Chao, Z. C., Nagasaka, Y., Fujii, N., ... Cichocki, A. (2013, July). Higher Order Partial Least Squares (HOPLS) : A Generalized Multilinear Regression Method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7), 1660–1673. Retrieved from <https://doi.org/10.1109/TPAMI.2012.254> doi: 10.1109/TPAMI.2012.254
- Zhou, F., & De la Torre, F. (2016). Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2), 279-294. doi: 10.1109/TPAMI.2015.2414429
- Zhou, G., Cichocki, A., Zhang, Y., & Mandic, D. P. (2016). Group component analysis for multiblock data : Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11), 2426-2439. doi: 10.1109/TNNLS.2015.2487364







# De la Vision par Ordinateur aux Modèles Profonds

Diane LINGRAND

## Résumé

Résumé en 100 pages de 25 ans d'activité.

**Ceci n'est pas le document définitif mais  
un brouillon en cours de rédaction.**

**Mots-clés :** Apprentissage Automatique, Vision par Ordinateur.

## Abstract

Enjoy

**Keywords:** Machine Learning, Computer Vision.