

# THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641

*Mathématiques et Sciences et Technologies du numérique  
de l'Information et de la Communication*

Spécialité : *Signal, Image, Vision*

Par

**Mathieu RIAND**

**Reconnaissance d'actions à partir d'un faible nombre de vidéos  
2D + profondeur : approche par construction de graphes de scène  
et apprentissage auto-supervisé**

**En vue de la soutenance de Thèse à Polytech Nantes, le 23 juin 2023**

**Unité de recherche : UMR6004 – LS2N**

**Thèse N° : « si pertinent »**

## **Rapporteurs avant soutenance :**

Alexandre BENOIT Professeur des universités, Université Savoie Mont Blanc  
Kévin BAILLY Maître de conférences, Sorbonne Université

## **Composition du Jury :**

Président :	Prénom NOM	Fonction et établissement d'exercice ( <i>à préciser après la soutenance</i> )
Examineurs :	Alice CAPLIER	Professeure des universités, Grenoble INP
	Benoît FURET	Professeur des universités, Nantes Université
Dir. de thèse :	Patrick LE CALLET	Professeur des universités, Nantes Université
Co-dir. de thèse :	Laurent DOLLE	Ingénieur chercheur, CEA Tech Pays de la Loire

## **Invité(s) :**

Anthony MOURAUD Ingénieur chercheur, Lhyfe

**Titre :** Reconnaissance d'actions à partir d'un faible nombre de vidéos 2D + profondeur : approche par construction de graphes de scène et apprentissage auto-supervisé

**Mot clés :** Reconnaissance d'actions, graphes de scène, GNN, apprentissage auto-supervisé, attention humaine

**Résumé :** L'apprentissage par démonstration peut permettre de rendre la robotique plus accessible en ayant simplement à réaliser une tâche devant un robot pour que celui-ci la reproduise ; cependant, il peut être difficile d'apprendre des tâches complexes depuis des démonstrations brutes. Une manière de simplifier ce processus est de séparer les tâches en actions simples que le robot pourra apprendre indépendamment. Dans cette thèse, nous proposons donc d'extraire les actions élémentaires effectuées par des humains dans des vidéos ; pour cela, nous représentons les scènes sous la forme de graphes symboliques dans lesquels chaque nœud est un objet de la démonstration, et nous les classifions grâce

à un GNN (*Graph Neural Network*). Nous explorons plusieurs stratégies de conception des graphes et montrons qu'il est possible d'améliorer la qualité de la reconnaissance d'actions en choisissant la bonne représentation. Puisque les démonstrations sont rarement annotées, nous proposons également des techniques d'apprentissage auto-supervisé appliquées aux graphes permettant de faire usage de données non labellisées pour augmenter encore les performances de notre modèle. Enfin, nous enregistrons 760 démonstrations et récoltons l'attention humaine sur une partie de ces vidéos afin de la comparer à l'attention de notre GNN.

**Title:** Action recognition from a small number of 2D + depth videos : an approach using scene graphs and self-supervised learning

**Keywords:** Action recognition, scene graphs, GNN, self-supervised learning, human attention

**Abstract:** Learning from demonstration can ease the access to robotics by simply having to do a task in front of a robot for it to be able to reproduce the same task ; however, it can be hard to learn complex tasks from raw demonstrations. A way to simplify this process it to cut tasks in simple actions that the robot will learn independently. In this thesis, we propose to extract elementary actions performed by humans in videos ; in order to do this, we represent our scenes as graphs where each node is an object from the demonstration, and we classify them using a GNN

(*Graph Neural Network*). We explore several design strategies for our graphs and show that it is possible to improve the quality of action recognition by carefully choosing the right representation. Since video demonstrations are rarely annotated, we also propose to apply self-supervised techniques to our graphs in order to make use of our annotated data to further improve our model's performance. Finally, we record 760 demonstrations and gather human attention on some of them to compare it to the attention of our GNN.