AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mme **LASRI Khadija**
Soutiendra : **le Samedi 04/04/2026** à **09H30**
Lieu : ***FSDM – Centre Visioconférence***

Une thèse intitulée :

Human Action Recognition using deep learning Architectures

En vue d'obtenir le ***Doctorat***

FD : **Sciences et Technologies de l'Information et de la Communication**
Spécialité : **Informatique**

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
TAIRI Hamid	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Président
TAIME Abderazzak	École Supérieure de Technologie, Khénifra	MCH	Rapporteur
HAJJI Tarik	Ecole Nationale Supérieure des Arts et Métiers, Meknès	MCH	Rapporteur
YAHYAOUY Ali	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Rapporteur
EL FAZAZY Khalid	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Examinateur
MAHRAZ Mohamed Adnane	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Examinateur
RIFFI Jamal	Faculté des Sciences Dhar EL Mahraz, Fès	MCH	Directeur de thèse

Résumé :

La reconnaissance des actions humaines (Human Action Recognition, HAR) constitue un défi fondamental en vision par ordinateur, avec des applications critiques dans les domaines de la surveillance, de la santé et de l'interaction homme–machine. Malgré des avancées significatives, les approches existantes demeurent confrontées à des défis majeurs, notamment la modélisation spatio-temporelle complexe, la présence d'arrière-plans encombrés, la fusion multimodale et les contraintes de performance en temps réel.

Cette thèse propose un cadre global fondé sur de nouvelles architectures d'apprentissage profond et des approches multimodales, visant à traiter systématiquement ces limitations à travers des contributions majeures.

Dans un premier temps, l'analyse des vidéos RGB est améliorée grâce à l'architecture CBAM-I3D, qui intègre des mécanismes d'attention aux convolutions 3D inflatées et atteint une précision de 92,4 % sur le jeu de données UCF-101. Cette approche est ensuite renforcée par un cadre Multimodal SlowFast (MM-SF), permettant une fusion efficace des voies spatiales et temporelles et atteignant une précision de 96,7 %.

Pour la reconnaissance basée sur les squelettes, la thèse développe quatre réseaux convolutifs sur graphes innovants conciliant efficacité et performance. LST-AGCN est proposé pour un traitement léger adapté aux dispositifs embarqués, tandis que DPCA-GCN, une architecture à double voie intégrant une attention croisée, atteint 93,2 % sur le benchmark NTU RGB+D 60. Afin de modéliser des topologies adaptatives, DTR-GCN (Dynamic Temporal-Relational GCN) est introduit pour un raffinement canal-par-canal, et TST-GCN exploite des transformeurs de segmentation temporelle adaptative afin de capturer les dépendances temporelles à longue portée.

Afin d'assurer une intégration efficace entre les modalités, la thèse présente MMAF (Multimodal Attention Fusion), un cadre reposant sur une attention inter-modale permettant une fusion dynamique des données squelettiques et RGB, atteignant des performances de l'état de l'art avec 96,8 % sur UCF-101. Enfin, la problématique de la détection spatio-temporelle des actions en temps réel est abordée à travers STAD (YOWOv4), un cadre unifié combinant des transformeurs visuels hiérarchiques et des méthodes modernes de détection d'objets, offrant une localisation précise avec un frame-mAP de 91,2 % tout en maintenant une cadence de 30 images par seconde. Des expérimentations approfondies sur plusieurs jeux de données de référence confirment la robustesse et l'efficacité des approches proposées

Mots clés :
Reconnaissance des actions humaines, Apprentissage profond, Vision par ordinateur, Modélisation spatio-temporelle, Réseaux de graphes, Fusion multimodale, Mécanismes d'attention, Détection en temps réel

HUMAN ACTION RECOGNITION USING DEEP LEARNING ARCHITECTURES

Abstract :

Human Action Recognition (HAR) has emerged as a fundamental challenge in computer vision, with critical applications spanning surveillance, healthcare, and human-computer interaction. Despite significant advances, existing approaches struggle with challenges including complex spatiotemporal modeling, background clutter, multimodal fusion, and real-time performance requirements. This dissertation presents a comprehensive framework of novel deep learning architectures and multi-modal approaches to systematically address these limitations through eight major contributions.

First, we advance RGB-based video analysis by proposing the CBAM-I3D architecture, which integrates attention mechanisms with inflated 3D convolutions to achieve 92.4\% accuracy on the UCF-101 dataset. We further enhance this with a Multimodal SlowFast (MM-SF) framework that effectively fuses spatial and temporal pathways, reaching 96.7\% accuracy.

For skeleton-based recognition, we develop four innovative graph convolutional networks to balance efficiency and performance. We introduce LST-AGCN for lightweight processing on edge devices, and DPCA-GCN, a dual-path architecture with cross-attention that achieves 93.2\% on the NTU RGB+D 60 benchmark. To capture adaptive topologies, we propose DTR-GCN (Dynamic Temporal-Relational GCN) for channel-wise refinement, and TST-GCN, which utilizes adaptive temporal segmentation transformers to model long-range dependencies.

To bridge the gap between modalities, we present MMAF (Multimodal Attention Fusion), a framework that dynamically integrates skeleton and RGB data through cross-modal attention, achieving state-of-the-art results of 96.8\% on UCF-101. Finally, we address the challenge of real-time spatio-temporal action detection with STAD (YOWOv4), integrating hierarchical vision transformers with modern object detection. This unified framework delivers precise localization with a frame-mAP of 91.2\% while maintaining real-time performance at 30 FPS. Extensive experiments across multiple benchmark datasets validate the robustness and efficiency of these proposed methodologies.

Key Words :
Human Action Recognition, Deep Learning, Computer Vision, Attention Mechanisms, Graph Convolutional Networks, Multimodal Fusion, Spatio-temporal Modeling, Real-time Detection