



École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Fatima Zahra JOBBID

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le 10 Octobre 2026 à 11h00 au Grand Amphi à l'ENSIAS

Intitulé de la thèse

AI-BASED APPROACHES FOR INTERPRETABLE STUDENT ENGAGEMENT ANALYSIS: FROM 2D DEEP VISUAL MODELS TO 3D SPATIO-TEMPORAL ARCHITECTURES

Président :

Pr. Mohamed ESSAAIDI, PES, ENSIAS, UM5

Directeur de thèse :

Pr. Hassan BERBIA, PES, ENSIAS, UM5

Rapporteurs :

Pr. Raddouane CHIHEB, PES, ENSIAS, UM5

Pr. Mohammed BERRADA, PES, ENSA de Fes

Pr. Mohammed TALIBI ALAOUI, PES, Faculté des Sciences et Techniques de Fés

Examineur(s) :

Pr. Aissam BERRAHOU, MCH, ENSIAS, UM5

Pr. Abderrahim MESBAH, MCH, ENSIAS, UM5

Pr. Yasser EL MADANI EL ALAMI, MCH, ENSIAS, UM5.

Résumé :

L'engagement des étudiants constitue un déterminant majeur des apprentissages, de la persévérance et de la réussite académique, mais son évaluation objective et à grande échelle reste un défi considérable dans les environnements d'apprentissage médiés par la technologie. En s'appuyant sur les travaux théoriques qui définissent l'engagement comme un construit multidimensionnel comportemental, émotionnel et cognitif, cette thèse étudie comment des méthodes d'intelligence artificielle (IA) basées sur la vision peuvent fournir des estimations robustes, interprétables et en temps réel de l'engagement à partir de données vidéo. La démarche adoptée est progressive et structurée : elle part d'une analyse comparative et d'une modélisation visuelle 2D pour aboutir à des architectures 3D spatio-temporelles fondées sur des réseaux de neurones graphes et des Transformers. La première contribution consiste en une étude comparative structurée des approches d'IA pour la mesure de l'attention étudiante. Onze méthodes sont évaluées selon deux grandes familles : les modalités neurophysiologiques (EEG, fNIRS, fréquence cardiaque, réponse galvanique cutanée) et les modalités non neurophysiologiques (reconnaissance d'expressions faciales, estimation de la pose de tête, suivi oculaire, mouvements corporels). Cette analyse comparative examine la précision, le coût et la faisabilité pratique de chaque approche, et conclut que la reconnaissance d'expressions faciales et l'estimation de la pose de tête offrent le meilleur compromis entre performance et déployabilité en conditions réelles, motivant ainsi les contributions techniques qui suivent. La deuxième contribution propose un cadre visuel profond combinant reconnaissance d'émotions faciales (FER) et estimation de la pose de tête pour prédire l'engagement à partir de vidéos de classe. Un module FER de type ResNet50, enrichi par un double mécanisme d'attention spatiale et canal et entraîné sur les jeux de données RAF-DB et FER2013, atteint 90,44 % de précision sur RAF-DB et 69,01 % sur FER2013, dépassant plusieurs approches récentes de l'état de l'art. Ce module est fusionné avec un estimateur léger de pose de tête basé sur des points de repère géométriques, atteignant une erreur absolue moyenne de 3,98° sur le benchmark AFLW2000-3D. Les deux flux sont intégrés dans un perceptron multi-couche qui prédit l'engagement binaire à partir de caractéristiques visuelles multimodales. Évalué sur le Student Engagement Dataset (SED), cette fusion atteint 88 % de précision et un F1-score macro de 0,93 pour la classe engagé, surpassant nettement les approches unimodales et restant compétitive face aux méthodes à base de Transformers, tout en permettant une inférence en temps réel à 23 images par seconde sur du matériel standard. Des analyses d'explicabilité par SHAP identifient l'expression neutre et l'angle de lacet (yaw) comme les deux prédicteurs dominants de l'engagement, en cohérence avec les théories neurocognitives de l'attention en contexte éducatif. La troisième contribution porte sur des données faciales 3D statiques. En exploitant le jeu de données DAD-3DHeads, la thèse définit des labels d'engagement proxy fondés sur la neutralité de l'expression, l'orientation frontale de la tête et l'absence d'occlusion, puis construit des vecteurs de caractéristiques interprétables en fusionnant les angles d'Euler de la pose de tête avec des Action Units (AU) faciaux extraits via OpenFace. Des modèles d'apprentissage supervisé classiques Random Forest, SVM, XGBoost et perceptrons multicouches sont évalués sur cet espace de caractéristiques. XGBoost obtient les meilleures performances avec 77 % de précision et un F1-score de 0,82. L'analyse d'importance des variables confirme le caractère complémentaire des indices géométriques et d'apparence : l'angle de tangage (pitch) et l'AU06 sont les prédicteurs les plus discriminants. Ce pipeline entièrement reproductible et interprétable constitue une alternative pratique aux approches boîte noire pour l'estimation de l'engagement à partir de données faciales 3D. La quatrième

contribution introduit une architecture 3D spatio-temporelle unifiée opérant directement sur des maillages faciaux paramétrés par FLAME. L'information spatiale est encodée par deux couches de réseaux de neurones graphes avec attention (Graph Attention Networks) qui traitent la topologie FLAME à 5023 sommets, capturant à la fois les relations géométriques locales et les dépendances globales de la structure faciale. La dynamique temporelle est modélisée par un encodeur Transformer à deux couches, et un mécanisme d'attention croisée entre modalités fusionne les représentations spatiale et temporelle en un plongement commun. Une tête d'apprentissage multi-tâches prédit conjointement l'engagement comportemental, la valence affective et un niveau cognitif global. L'architecture présente une convergence stable, caractérisée par des normes de gradient comprises entre 6,4 et 16,7, et est validée sur le jeu de données DAD-3DHeads comprenant 44 898 scans faciaux 3D haute résolution annotés avec des labels d'engagement multidimensionnels. Ce modèle GNN-Transformer surpasse systématiquement les approches 2D et les pipelines basés sur des caractéristiques manuelles, et produit des cartes d'attention spatiales et temporelles fines qui renforcent l'interprétabilité et soutiennent une analyse de l'engagement ancrée dans la physiologie. Dans l'ensemble, la thèse apporte : (i) une étude comparative structurée des approches d'IA pour la mesure de l'attention étudiante ; (ii) un système 2D profond et temps réel de détection de l'engagement combinant FER et estimation de la pose de tête avec explicabilité SHAP ; (iii) une chaîne 3D interprétable et reproductible fondée sur l'apprentissage automatique classique et la fusion de caractéristiques multimodales ; et (iv) une architecture 3D GNN-Transformer pour la modélisation multidimensionnelle de l'engagement à partir de données de maillage facial. Ces contributions font progresser l'état de l'art en matière de mesure de l'engagement par IA et ouvrent la voie à des systèmes d'analytique de l'apprentissage multimodaux, explicables et éthiquement responsables, capables de soutenir un enseignement adaptatif et un retour personnalisé à grande échelle.

Mots-clés:

engagement étudiant, informatique affective, apprentissage profond, reconnaissance d'émotions faciales, estimation de la pose de tête, réseaux de neurones graphes, Transformers, IA explicable, analytique de l'apprentissage.

Abstract:

Student engagement is a central determinant of learning, persistence, and academic success, yet its objective and scalable measurement remains a significant challenge in technology-mediated learning environments. Building on theoretical work that conceptualises engagement as a multidimensional construct encompassing behavioural, emotional and cognitive components, this dissertation investigates how visual artificial intelligence (AI) methods can provide robust, interpretable and real-time estimates of engagement from video data. The thesis follows a progressive and structured methodology, starting from comparative analysis and 2D visual modelling, and culminating in 3D spatio-temporal architectures based on graph neural networks and Transformers. The first contribution consists of a structured comparative study of AI-based approaches to student attention measurement. Eleven methods are evaluated across neurophysiological modalities (EEG, fNIRS, heart rate, galvanic skin response) and non-neurophysiological modalities (facial emotion recognition, head pose estimation, eye tracking, body movements). This comparative analysis examines the accuracy, cost and practical deployability of each approach, and concludes that facial emotion recognition and head pose estimation offer the most favourable balance between performance and real-world feasibility, thereby motivating the technical contributions that follow. The second contribution proposes a deep visual framework that combines facial emotion recognition (FER) and head pose estimation to predict engagement from classroom videos. An enhanced ResNet50-based FER module incorporating a dual spatial and channel attention mechanism, trained on the RAF-DB and FER2013 datasets, achieves 90.44% accuracy on RAF-DB and 69.01% on FER2013, outperforming several recent state-of-the-art methods. This module is fused with a lightweight landmark-based head pose estimator that achieves a mean absolute error of 3.98° on the AFLW2000-3D benchmark. The two streams are integrated into a multilayer perceptron that predicts binary engagement from multimodal visual features. Evaluated on the Student Engagement Dataset (SED), this fusion achieves 88% accuracy and a macro F1-score of 0.93 on the engaged class, significantly outperforming unimodal baselines and remaining competitive with recent Transformer-based approaches, while enabling real-time inference at 23 FPS on standard hardware. Explainability analyses using SHAP identify neutral facial emotions and head yaw angle as the two dominant predictors of engagement, in alignment with neurocognitive theories of attention in educational contexts. The third contribution focuses on 3D static facial data. Leveraging the DAD-3DHeads dataset, the thesis defines proxy engagement labels based on neutral facial emotion, frontal head orientation and absence of occlusion, and constructs interpretable feature vectors by fusing head pose Euler angles with facial Action Units (AUs) extracted using OpenFace. Classical supervised models Random Forest, SVM, XGBoost and multilayer perceptrons are benchmarked on this feature space. XGBoost achieves the best performance with 77% accuracy and an F1-score of 0.82. Feature importance analysis confirms the complementary roles of geometric and appearance-based cues, identifying pitch angle and AU06 as the most discriminative predictors. This fully reproducible and interpretable pipeline provides a practical alternative to black-box deep learning approaches for engagement estimation from 3D facial data. The fourth contribution introduces a unified 3D spatio-temporal architecture that operates directly on FLAME-parameterised facial meshes. Spatial information is encoded using two-layer Graph Attention Networks that process the 5,023-vertex FLAME topology, capturing both local geometric relationships and global facial structure dependencies. Temporal dynamics are modelled by a two-layer Transformer encoder, and a novel cross-modal attention mechanism fuses spatial and temporal representations into a joint embedding. A multi-task learning head jointly predicts

behavioural engagement, affective valence and coarse cognitive load. The architecture exhibits stable convergence, characterised by gradient norms between 6.4 and 16.7, and is validated on the DAD-3DHeads dataset comprising 44,898 high-resolution 3D facial scans annotated with multi-dimensional engagement labels. This GNN-Transformer model consistently outperforms 2D and handcrafted baselines, and produces fine-grained spatial and temporal attention maps that enhance interpretability and support physiologically grounded engagement analysis. Overall, the dissertation contributes (i) a structured comparative study of AI-based approaches to student attention measurement, (ii) a real-time 2D deep visual engagement system combining FER and head pose estimation with SHAP-based explainability, (iii) an interpretable and reproducible 3D engagement pipeline based on classical machine learning and multimodal feature fusion, and (iv) a novel 3D GNN-Transformer architecture for multidimensional engagement modelling from facial mesh data. These contributions advance the state of the art in AI-based engagement measurement and lay the groundwork for future multimodal, explainable and ethically grounded learning analytics systems that can support adaptive teaching and personalised feedback at scale.

Keywords:

student engagement, affective computing, deep learning, facial emotion recognition, head pose estimation, graph neural networks, Transformers, explainable AI, educational data analysis, learning analytics.