(English version bellow on page 7)

Titre

Deep Learning pour l'Immersion et l'Interaction 3D Naturelle dans les Expériences de Réalité Mixte intelligente

Mots clés

Réseaux de neurones convolutifs, réseaux de neurones récurrents, détection multi-objets, tracking temps réel, interaction 3D, interaction intelligente naturelle, immersion virtuelle avancée.

Résumé

Le projet de thèse vise à proposer et à développer une nouvelle approche de vision intelligente permettant l'immersion virtuelle et l'interaction naturelle pour les expériences avancées en Réalité Mixte. Ce système doit pouvoir offrir une interface immersive avec une interaction 3D libre qui favorise la communication naturelle entre l'utilisateur et la représentation virtuelle qui l'entoure dans son environnement mixte.

Différents aspects et axes de recherche seront explorés et étudiés pour offrir une cohérence perceptive dans l'interface de réalité mixte. En effet, le système de réalité mixte développé augmentera le monde réel avec des hologrammes et des projections synthétiques en procurant à l'utilisateur le sentiment de présence à travers la navigation et la manipulation des objets virtuels dans le monde mixte sans aucune astreinte cognitive telle que la perception incohérente, la restriction de déplacement ou la modification de comportement.

Le projet de thèse repose sur l'utilisation des approches de l'intelligence artificielle pour analyser l'expérience utilisateur et étudier l'adaptabilité de l'interface de réalité mixte. Cela étant effectué à l'aide d'un module de détection multi objets qui permet la reconnaissance des objets d'intérêt par le bais d'un processus d'apprentissage automatique. Par la suite, l'intelligence artificielle est utilisée pour la mise en œuvre d'un système interactif intelligent afin de mener efficacement des expériences immersives temps réel, et permettre également la communication et le partage de connaissances à l'aide des interactions 3D naturelles. Le système développé devrait permettre l'amélioration de la perception, la présence et la cognition de l'environnement par le développement d'une interface adaptée au mouvement libre, au comportement naturel et à l'interaction intuitive humaine.

Thématiques / domaines

Deep learning, machine learning, réalité mixte, réalité augmentée, interaction 3D naturelle, computer vision, computer graphics, intelligence artificielle.

Contexte

Le projet de recherche s'articule autour de la réalisation des systèmes de réalité mixte intelligents. La réalité mixte a connu un développement très important ces dernières années grâce aux progrès scientifique et technologique. Toutefois, la reconnaissance d'images reste une problématique difficile et les chercheurs continuent de proposer et de développer des approches pour élaborer des systèmes performants d'identification d'objets. En effet, les SDK les plus populaires proposent des modules de reconnaissance d'image qui se basent sur des marqueurs visuels ou des objets 2D plan (planar objects). La détection d'objet 2D/3D naturel en utilisant l'apprentissage automatique est une problématique complexe qui est assujettis à des processus chronophages qui sont : la collecte et l'étiquetage des données, le monitoring de la convergence et de la précision des modèles, les ressources matérielles et logicielles nécessaires pour l'apprentissage et enfin, l'optimisation des modèles pour des prédictions en temps réel.

Les thématiques de recherche qui seront abordées dans cette thèse pour contribuer à la détection d'objets naturels, reposent sur l'utilisation du deep learning. L'identification d'objet sera basée sur l'utilisation des CNN et R-CNN pour obtenir une meilleure robustesse et précision de détection d'objets libres en environnement réel. Par ailleurs, il est important d'optimiser le temps de calcul pour tester et déployer le module de reconnaissance dans un flux de caméra en temps réel. Ce module pourra être utilisé pour divers cas d'usage et scénarios applicatifs.

Par la suite le système de réalité mixte estimera la localisation de la caméra et l'ajout de graphiques virtuels sur la scène réelle en permettant l'interaction naturelle 3D à travers les mouvements et la gestuelle de l'utilisateur. L'interaction naturelle peut se réaliser à travers l'analyse de mouvement corporels, toutefois cela nécessite la détection du corps et l'analyse 3D de la scène. Dans ce cas, une caméra de profondeur est requise pour l'analyse de la scène afin de déterminer des mouvements basiques. Une approche basée sur le deep learning permettra la reconnaissance de mouvement à l'aide d'un flux standard de la caméra et l'identification de gestes élaborés, par le biais d'un apprentissage séquentiel (CNN) et/ou récurrent (LSTM, GRU).

Dans ce contexte, différents axes de recherche sont étudiés et plusieurs étapes sont nécessaires pour contribuer à l'aboutissement du système :

- Une approche de perception pour l'acquisition et le traitement des données.
- Un module décisionnel de l'intelligence artificielle qui permet l'apprentissage et la classification de données pour des prédictions précises et robustes.
- Une couche applicative d'immersion et d'interaction naturelle pour l'augmentation visuelle et l'interaction naturelle avec le monde virtuel.

- Un module d'optimisation de performances qui permet d'assurer : la pertinence, la mobilité, le temps réel, la robustesse et aussi l'amélioration de l'expérience en réalité mixte en procurant le sentiment de présence et le haut niveau cognitif.

Objectifs de la thèse

Plusieurs thématiques de recherche doivent être prospectées et développées pour élaborer ce système de réalité mixte intelligent, les objectifs visés sont :

- Développer des méthodes de détection multi objets basées sur le deep learning et l'intelligence artificielle pour résoudre la problématique complexe de la reconnaissance de cibles naturelles dans un espace intérieur/extérieur non contrôlé.
- Mixer le réel et le virtuel dans un environnement composite avec une qualité de réalisme et de performance très élevées : temps réel, précision, stabilité et cohérence de la représentation de synthèse. Le but recherché est de vérifier la fiabilité du rendu sensoriel pour la synchronisation réel-virtuel.
- Permettre l'interaction naturelle et non contrainte dans les espaces réel et virtuel à travers des modèles d'apprentissage. L'objectif étant la manipulation des objets virtuels avec une navigation en totale liberté de mouvement pour accroitre le sentiment de présence.

Résultats attendus

1) Veille scientifique sur les modèles d'apprentissage utilisant les réseaux de neurones

Étudier et comparer les architectures les plus performantes et populaires, telles que : AlexNet, VGG16, VGG19, GoogLeNet, ResNet, SqueezeNet, DenseNet, ENet, VGG16, Inception, Xception, MobileNet, ResNet50, etc.

2) Conception et construction des modèles d'apprentissage

- Étudier les modèles R-CNN, tels que : Fast R-CNN, Faster-RCNN, YOLO, etc. Tester ces architectures par un transfert learning.
- Développer des approches de reconnaissance d'objets en utilisant les réseaux de neurones convolutifs basés sur la région (R-CNN).
- Étudier les modèles récurrents RNN tels que le LSTM et GRU sur des séquences d'images pour la reconnaissance de mouvement et de geste dans un processus d'interaction naturelle.
- Construire des modèles performants : rapide et précis. Les modèles seront déployés pour les applications de réalité mixte qui ont l'exigence du temps réel et de précision.

3) Collecte, analyse et ingénierie des données

- Recueillir un volume important de données, réaliser les opérations de nettoyage, étiquetage, augmentation de données, prétraitement (transformation, redimensionnement, etc).
- Vérifier et équilibrer la distribution des différentes catégories d'images, utiliser les techniques de downsampling et upsampling pour obtenir les meilleures proportions de données.

4) Entrainement, optimisation et validation des modèles

- Entrainer les modèles à l'aide des bibliothèques dédiées (TensorFlow ou PyTorch).
- Analyser les performances des modèles et optimiser les hyperparamètres pour améliorer les résultats.
- Paramétrer et superviser le processus d'apprentissage des architectures étudiées et proposées.
- Tester et évaluer les modèles, déterminer les paramètres de performances.
- Sauvegarder et restaurer les modèles en utilisant des checkpoints pour améliorer leurs convergences ultérieurement.

Précision sur l'encadrement

Samir OTMANE, Professeur, Laboratoire IBISC, Univ Evry, Université Paris-Saclay
Madjid MAIDI, Chercheur Associé à IBISC, Enseignant-Chercheur à L'ESME

Conditions scientifiques matérielles (conditions de sécurité spécifiques) et financières du projet de recherches

Plateforme et matériel :

- Serveur de traitement pour l'apprentissage des modèles en deep learning : HP Z4 Intel Xeon 10 x 3.3Ghz / RAM 256Go / GPU NVIDIA Quadro P4000 8GB, 1792 Cores
- Matériels de la plateforme EVR@ : casques de réalité mixte, smartphones, Ecrans 4K https://www.ibisc.univ-evry.fr/plate-formes/realite-virtuelle-et-augmentee-et-robotique

Objectifs de valorisation des travaux de recherche du doctorant :

- 2 conférences internationales par an de rang A ou B (ISMAR, CVPR, ICCV, VRST, ICIP, MMSP, ICMR, etc)
- 1 ou 2 revues de bonne qualité (CVIU, IVC, PR, MVA, VR, PIA, AR, AI, PRML, etc)

Collaborations envisagées

IBISC (Université d'Evry) et ESME (Ecole d'Ingénieurs)

Profil et compétences recherchées

- **Formation et connaissances scientifiques/théoriques**
  - Ingénieur informatique ou master 2 recherche à dominante informatique.
  - Solides compétences en Computer Vision (classification, détection, traitement et analyse vidéo, tracking, géométrie multi-vues) et en Computer Graphics (transformations projectives, modelview, shaders, mesh).
  - Connaissances approfondies en Machine Learning (classification, régression, imputation, clustering, composantes principales) et Deep Learning (descente de gradient, backpropagation, monitoring des performances, optimisation et supervision des hyperparamètres, construction des modèles : CNN, RNN, R-CNN, LSTM, GRU, VAE, GAN).
- **Compétences techniques**
  - **Librairies :** sklearn, tensorflow/PyTorch, Unity, OpenGL, freeglut, OpenCV, ARCore/Vuforia.
  - **Langages :** C++, Python, C#, Java (la connaissance de cuda ou OpenCL serait un plus).
  - **Systèmes :** Windows, Linux, Android.
- **Qualités personnelles**
  - Motivation pour la recherche.
  - Force de proposition, esprit critique, capacité de synthèse, bonne communication orale et écrite, aptitude à convaincre et à structurer son argumentation.
  - Sens d'organisation, gestion de projet, persévérance, créativité, travail prospectif et démarche innovante.

Références bibliographiques

[1] J. Redmon, S. Kumar Divvala, R. B. Girshick and A. Farhadi. « You Only Look Once: Unified, Real-Time Object Detection ». IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016

[2] S. Ren, K. He, R. B. Girshick and J. Sun. « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks ». IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, N° 6, pp. 1137-1149, 2017

[3] R. B. Girshick, J. Donahue, T. Darrell and J. Malik. « Rich Feature Hiérarchies for Accurate Object Detection and Semantic Segmentation ». IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014

[4] R. Girshick. « Fast R-CNN », IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015

[5] J. Guo, P. Chen, Y. Jiang, H. Yokoi and S. Togo. « Real-time Object Detection with Deep Learning for Robot Vision on Mixed Reality Device ». IEEE Global Conference on Life Sciences and Technologies (LifeTech), pp. 82-83, 2021

[6] K. B. Park, S. H. Choi, J. Y. Lee, Y. Ghasemi, M. Mohammed and H. Jeong. « Hands-Free Human–Robot Interaction Using Multimodal Gestures and Deep Learning in Wearable Mixed Reality ». IEEE Access, vol. 9, pp. 55448-55464, 2021

[7] J. Lalonde. « Deep Learning for Augmented Reality ». Workshop on Information Optics (WIO), pp. 1-3, 2018

[8] Q. Cheng, S. Zhang, S. Bo, D. Chen and H. Zhang. « Augmented Reality Dynamic Image Recognition Technology Based on Deep Learning Algorithm ». IEEE Access, vol. 8, pp. 137370-137384, 2020

[9] C. H. Lin, Y. Chung, B. -Y. Chou, H. Y. Chen and C. Y. Tsai. « A Novel Campus Navigation APP with Augmented Reality and Deep Learning ». IEEE International Conference on Applied System Invention (ICASI), pp. 1075-1077, 2018

[10] J. G. Ko, S. Lee, S. Lee and J. Lee. « Lightweight Deep Learning based Intelligent Mobile Augmented Reality ». IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1-3, 2021

[11] C. Li, S. Sun, X. Min, W. Lin, B. Nie and X. Zhang. « End-to-end Learning of Deep Convolutional Neural Network for 3D Human Action Recognition ». IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 609-612, 2017

[12] J. Yan and K. Mei. « Attention Residual Network with 3D convolutional neural network for 3D Human Pose Estimation ». IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 1266-1271, 2021

[13] R. Siriak, I. Skarga-Bandurova and Y. Boltov. « Deep Convolutional Network with Long Short-Term Memory Layers for Dynamic Gesture Recognition ». IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp. 158-162, 2019

Title

Deep Learning for Immersion and Natural 3D Interaction within Intelligent Mixed Reality Experiences

Keywords

Convolutional neural networks, recurrent neural networks, multi-object detection, real-time tracking, 3D interaction, natural intelligent interaction, advanced virtual immersion.

Summary

The thesis project aims to propose and develop a new intelligent vision approach allowing virtual immersion and natural interaction for advanced experiences in mixed reality. This system will offer an immersive interface with a free 3D interaction that enables natural communication between the user and the virtual representation surrounding him in the mixed environment.

Different aspects and lines of research will be explored and studied to ensure perceptual coherence in the mixed reality interface. Indeed, the developed mixed reality system will augment the real world with holograms and synthetic projections by providing to the user the feeling of presence through the navigation and manipulation of virtual objects in the mixed world without any cognitive strain such as inconsistent perception, restriction of movement or behavior modification.

The thesis project is based on the use of artificial intelligence approaches to analyze the user experience and study the adaptability of the mixed reality interface. This is made possible using a multi-object detection module that allows the recognition of objects of interest through a machine learning process. Subsequently, artificial intelligence is used for the implementation of an intelligent interactive system to efficiently conduct immersive real-time experiences and enabling communication and knowledge sharing using natural 3D interactions. The system should improve the perception, presence and cognition of the environment through the development of an interface well adapted to free movement, natural behavior and intuitive human interaction.

Thematics / areas

Deep learning, machine learning, mixed reality, augmented reality, natural 3D interaction, computer vision, computer graphics, artificial intelligence

Context

The research project in centered around the realization of intelligent mixed reality systems. Mixed reality has experienced a very important development in recent years thanks to scientific and technological progress. However, image recognition remains a difficult issue and researchers continue to propose and develop approaches to develop efficient object identification systems. Indeed, the most popular SDKs offer image recognition modules that are based on visual markers or planar objects. The detection of natural 2D/3D objects using machine learning is a complex issue that is subject to time-consuming processes that are : data collection and labeling, monitoring the convergence and accuracy of models, hardware and software resources required for learning and also the optimization of models for real-time predictions.

The research themes that will be addressed in this thesis to detect natural objects are based on deep learning. Object identification will rely upon the use of CNN and R-CNN to achieve better robustness and accuracy of natural object detection in real environments. In addition, it is important to optimize the computation time to test and deploy the recognition module in a real-time camera stream. This module can be used for various use cases and application scenarios.

Subsequently, the mixed reality system will estimate the location of the camera and overlay virtual graphics on the real scene, allowing natural 3D interaction through the movements and gestures of the user. Natural interaction can be achieved through body motion analysis, however this requires body detection and 3D analysis of the scene. In this case, a depth camera is required for scene analysis to determine basic movements. An approach based on deep learning will enable motion recognition using a standard camera flow and advanced gesture recognition through sequential (CNN) and/or recurrent (LSTM, GRU) learning.

In this context, different lines of research are studied, and several steps are necessary to contribute to the completion of the system:

- A perception approach for data acquisition and processing.
- An artificial intelligence decision-making module that enables data learning and classification for accurate and robust predictions.
- An application layer of immersion and natural interaction for visual augmentation and natural interaction with the virtual world.
- A performance optimization module that ensures: relevance, mobility, real-time, robustness and also the improvement of the experience in real life by providing the feeling of presence and the high cognitive level.

Goals of the thesis

Several research themes must be explored to develop this intelligent mixed reality system, the objectives are:

- Develop multi-object detection methods based on deep learning and artificial intelligence to solve the complex problem of natural target recognition in an uncontrolled indoor/outdoor environment.
- Mix the real and the virtual in a composite environment with a very high quality of realism and performance: real time, precision, stability and consistency of the virtual representation. The goal is to verify the reliability of the sensory rendering for real-virtual synchronization.
- Enable natural and unconstrained interaction in real and virtual spaces through learning models. The objective is the manipulation of virtual objects with navigation in total freedom of movement to increase the feeling of presence.

Expected results

5) Scientific monitoring and review of learning models using neural networks

Study and compare the most powerful and popular architectures, such as: AlexNet, VGG16, VGG19, GoogLeNet, ResNet, SqueezeNet, DenseNet, ENet, VGG16, Inception, Xception, MobileNet, ResNet50, etc.

6) Design and construction of learning models

- Study R-CNN models, such as: Fast R-CNN, Faster-RCNN, YOLO, etc. Test these architectures through transfer learning.
- Develop object recognition approaches using region-based convolutional neural networks (R-CNN).
- Study RNN recurring models such as LSTM and GRU on image sequences for motion and gesture recognition in a natural interaction process.
- Build high-performance models: fast and accurate. These models will be deployed for mixed reality applications that require real-time and accuracy.

7) Data collection, analysis and engineering

- Collect a large volume of data, carry out cleaning operations, labeling, data augmentation, pre-processing (transformation, resizing, etc.).
- Check and balance the distribution of different image categories, use downsampling and upsampling techniques to get the best proportions of data.

8) Training, optimization and validation of models

- Train models using dedicated libraries (TensorFlow or PyTorch).
- Analyze model performance and optimize hyperparameters to improve results.
- Configure and supervise the learning process of the proposed architectures.
- Test and evaluate models, determine performance parameters.

- Back up and restore models using checkpoints to improve their convergences later.

Work supervisors

Samir OTMANE, Professor, IBISC laboratory, Univ Evry, Université Paris-Saclay
Madjid MAIDI, Associate Researcher at IBISC, Teacher-Researcher at ESME

Material, scientific and financial conditions of the research project

Platform and hardware:

- Processing server for deep learning models: HP Z4 Intel Xeon 10 x 3.3Ghz / RAM 256GB / GPU NVIDIA Quadro P4000 8GB, 1792 Cores
- EVRA platform: mixed reality headsets, smartphones, 4K displays https://www.ibisc.univ-evry.fr/plate-formes/realite-virtuelle-et-augmentee-et-robotique

Valorization objectives of the doctoral student's research work:

- 2 international conferences per year of rank A or B (ISMAR, CVPR, ICCV, VRST, ICIP, MMSP, ICMR, etc.)
- 1 or 2 good quality journals (CVIU, IVC, PR, MVA, VR, PIA, AR, AI, PRML, etc.)

Collaborations

IBISC (University of Evry, Université Paris-Saclay) and ESME (Engineering School)

Profile and skills sought

- **Education and scientific/theoretical knowledge**
  - Computer Engineer or MSc in computer science area.
  - Strong skills in Computer Vision (classification, detection, video processing and analysis, tracking, multi-view geometry) and Computer Graphics (projective transformations, modelview, shaders, mesh).
  - In-depth knowledge of Machine Learning (classification, regression, imputation, clustering, principal components) and Deep Learning (gradient descent, backpropagation, performance monitoring, optimization and supervision of hyperparameters, model construction: CNN, RNN, R-CNN, LSTM, GRU, VAE, GAN)
- **Technical skills**
  - **Libraries:** sklearn, tensorflow/PyTorch, Unity, OpenGL, freeglut, OpenCV, ARCore/Vuforia
  - **Languages:** C++, Python, C#, Java (knowledge of cuda or OpenCL would be a plus).
  - **Systems:** Windows, Linux, Android.

- **Personal qualities**
  - Motivation for research.
  - Strength of proposal, critical thinking, ability to synthesize, good oral and written communication, ability to convince and structure one's argumentation.
  - Sense of organization, project management, perseverance, creativity, forward-looking work and innovative approach.

References

[1] J. Redmon, S. Kumar Divvala, R. B. Girshick and A. Farhadi. « You Only Look Once: Unified, Real-Time Object Detection ». IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016

[2] S. Ren, K. He, R. B. Girshick and J. Sun. « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks ». IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, N° 6, pp. 1137-1149, 2017

[3] R. B. Girshick, J. Donahue, T. Darrell and J. Malik. « Rich Feature Hiérarchies for Accurate Object Detection and Semantic Segmentation ». IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014

[4] R. Girshick. « Fast R-CNN », IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015

[5] J. Guo, P. Chen, Y. Jiang, H. Yokoi and S. Togo. « Real-time Object Detection with Deep Learning for Robot Vision on Mixed Reality Device ». IEEE Global Conference on Life Sciences and Technologies (LifeTech), pp. 82-83, 2021

[6] K. B. Park, S. H. Choi, J. Y. Lee, Y. Ghasemi, M. Mohammed and H. Jeong. « Hands-Free Human–Robot Interaction Using Multimodal Gestures and Deep Learning in Wearable Mixed Reality ». IEEE Access, vol. 9, pp. 55448-55464, 2021

[7] J. Lalonde. « Deep Learning for Augmented Reality ». Workshop on Information Optics (WIO), pp. 1-3, 2018

[8] Q. Cheng, S. Zhang, S. Bo, D. Chen and H. Zhang. « Augmented Reality Dynamic Image Recognition Technology Based on Deep Learning Algorithm ». IEEE Access, vol. 8, pp. 137370-137384, 2020

[9] C. H. Lin, Y. Chung, B. -Y. Chou, H. Y. Chen and C. Y. Tsai. « A Novel Campus Navigation APP with Augmented Reality and Deep Learning ». IEEE International Conference on Applied System Invention (ICASI), pp. 1075-1077, 2018

[10] J. G. Ko, S. Lee, S. Lee and J. Lee. « Lightweight Deep Learning based Intelligent Mobile Augmented Reality ». IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1-3, 2021

[11] C. Li, S. Sun, X. Min, W. Lin, B. Nie and X. Zhang. « End-to-end Learning of Deep Convolutional Neural Network for 3D Human Action Recognition ». IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 609-612, 2017

[12] J. Yan and K. Mei. « Attention Residual Network with 3D convolutional neural network for 3D Human Pose Estimation ». IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 1266-1271, 2021

[13] R. Siriak, I. Skarga-Bandurova and Y. Boltov. « Deep Convolutional Network with Long Short-Term Memory Layers for Dynamic Gesture Recognition ». IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp. 158-162, 2019