Reconnaissance d'actions basée sur des modèles de segmentation

C. Huyghe[1,2], N. Ihaddadene[2], T. Haessle[3], C. Djeraba[1]

[1] Université de Lille, [2] Yncréa Hauts-de-France, [3] CareClever SAS

catherine.huyghe@yncrea.fr, nacim.ihaddadene@yncrea.fr, thaessle@cutii.io, chabane.djeraba@univ-lille.fr

Résumé

La reconnaissance d'actions humaines dans les vidéos est un problème important en vision par ordinateur. Nous proposons une approche basée sur l'intégration de la segmentation sémantique globale ou partielle du corps humain dans le processus de classification. Le modèle destiné à un robot d'assistance ambiante et l'expérimentation sur des datasets publics seront présentés.

Mots-clés

Reconnaissance d'actions, Assistance ambiante.

1 Introduction

La reconnaissance d'actions humaines à partir de séquences d'images trouve ses applications dans de nombreux domaines, allant de l'indexation des contenus à la surveillance intelligente et la collaboration homme-machine. De plus, les modèles sont souvent embarqués sur des périphériques contraints (Caméras, robots, objets connectés, ...), ce qui réduit considérablement leurs ressources. Les méthodes basées sur les modèles convolutionnels se sont développées ces dernières années.

2 État de l'art

Nous nous intéressons dans notre étude aux techniques basées sur l'utilisation de réseaux de neurones profonds. Différentes architectures pour la reconnaissance d'actions sont proposées dans la littérature [1]. Elles se distinguent par :

Les flux d'entrée : Certaines méthodes utilisent uniquement la séquence d'images en entrée, tandis que d'autres l'augmentent par des séquences du flux optique pour localiser et caractériser le mouvement. Le traitement des deux flux distincts est fusionné par la suite.

La dimension spatio-temporelle : Cette dimension a été traitée de différentes manières. Dans une approche, l'analyse des images par des convolutions 2D est combinée à l'utilisation des LSTM dans les couches ultérieures de l'architecture. Une autre approche consiste à utiliser des convolutions 3D sur les vidéos. Ainsi le noyau des convolutions 3D capture la dimension séquentielle. L'inconvénient est que les convolutions 3D nécessitent plus de paramètres qu'en 2D, ce qui les rend difficiles à entraîner et à déployer.

Sur l'ensemble des méthodes, on constate un faible focus sur l'analyse de la personne (analyse de l'ensemble de l'image) et l'absence du traitement de l'immobilité.

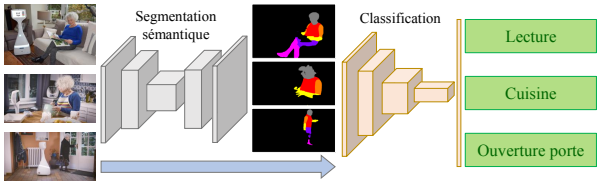

FIGURE 1 – Approche en deux phases pour la reconnaissance d'actions humaines

3 Approche et Expérimentation

Notre objectif est de développer des modèles déployables sur un robot d'assistance ambiante capable de reconnaître différents types d'actions (activités quotidiennes, exercices physiques de réhabilitation, chutes ou immobilité).
Des méthodes de segmentation sémantique du corps humain (totale ou partielle) se sont développées ces dernières années [2], sans arriver au niveau sémantique des actions. Nous proposons dans notre approche d'intégrer cette segmentation sémantique dans une première phase. Les images obtenues sont combinées avec les données de base en entrée d'un modèle convolutionnel 2D avec des couches récurrentes aux derniers niveaux.
L'apprentissage des modèles a été réalisé sur les datasets UCF101, HMDB et Charades. Les résultats seront détaillés. Ils démontrent que la reconnaissance d'actions basée sur la segmentation sémantique permet un gain de performances, un focus sur les zones de présence des personnes ainsi qu'un meilleur traitement de l'immobilité.
Nous prévoyons des tests sur les périphériques embarqués, dont un robot mobile d'assistance ambiante, ainsi que des tests de scénarios d'applications dans un contexte concurrentiel multi-objectifs.

Remerciements : Nous tenons à remercier le FEDER et l'entreprise CareClever pour leur soutien de ce projet.

Références

[1] J. Carreira, A. Zisserman. "Quo Vadis, Action Recognition ? A New Model and the Kinetics Dataset". *IEEE Computer Vision and Pattern Recognition, 2017.*

[2] F. Xia, P. Wang, X. Chen and A. L.Yuille, "Joint Multi-person Pose Estimation and Semantic Part Segmentation,". *IEEE Computer Vision and Pattern Recognition, 2017.*