Reconnaissance d'actions humaines par vision pour la robotique d'assistance à l'autonomie à domicile

Catherine Huyghe

Université de Lille

Thèse De Doctorat

Reconnaissance d'actions humaines par vision pour la robotique d'assistance à l'autonomie à domicile

Préparée par :

Catherine Huyghe

Soutenue publiquement le 06/12/2023, devant le jury composé de :

Hichem Sahbi	Chargé de recherche, Université Sorbonne	**Rapporteur**
Yannick Benezeth	Maître de conférences, Université de Bourgogne	**Rapporteur**
Abdenour Bouzouane	Professeur, Université du Québec à Chicoutimi	**Président**
Chaabane Djeraba	Professeur, Université de Lille	**Directeur de thèse**
Nacim Ihaddadene	Maître de conférences, Junia, Lille	**Co-encadrant**

Ce projet est cofinancé par l'Union européenne avec le Fonds européen de développement régional.

Remerciements

Tout d'abord, je remercie mon directeur de thèse, Chaabane Djeraba, pour sa guidance, sa patience et son soutien tout au long de cette thèse.

Je remercie mon encadrant, Nacim Ihaddadene, pour nos discussions, ses conseils et son soutien tout au long de cette thèse.

Mes remerciements vont également à mes rapporteurs, Hichem Sahbi et Yannick Benezeth, pour avoir consacré leur temps et leur expertise à l'évaluation de ma thèse.

Je remercie Clarisse Dhaenens et Abdenour Bouzouane pour avoir accepté d'être examinateur de ma thèse.

Je tiens à reconnaître le soutien financier crucial apporté par l'Union européenne avec le Fonds européen de développement régional, la région Haut-de-France et l'entreprise CareClever (Cutii). Leur contribution a permis de réaliser cette thèse.

Je remercie l'ensemble des membres de Cutii et tout particulièrement Thomas Haessle et Luis Parada pour leur accompagnement, leur soutien et nos discussions riches dans le but de permettre au robot Cutii d'évoluer et d'aider le plus grand nombre de personnes.

Un grand merci à toute l'équipe FOX qui m'a accompagné, soutenu et aidé pendant ces années.

Enfin, je souhaite remercier Junia pour avoir fourni l'environnement et les ressources nécessaires pour mener à bien cette thèse.

Merci à ceux qui m'ont soutenu et encouragé pendant ces années de thèse.

Ce travail de thèse n'aurait pas été possible sans l'aide précieuse de chacune de ces personnes et institutions. Je suis profondément reconnaissante envers vous tous.

Abstract

In recent years, the development of robotics has enabled the use of robots in active assisted living (AAL). They are used at various levels to help vulnerable individuals, due to age, illness, or disability, achieve a certain level of autonomy. They can be used, for example, to assist with daily tasks or to alert caregivers or family members in case of dangerous or abnormal situations. Falls or immobility on the floor are important examples of such dangerous or abnormal situations.

Our thesis falls within the field of computer vision, and in particular that of human action recognition, which could be deployed in the context of home assistance robotics. We propose an approach for human action recognition based on semantic segmentation of human body parts, with a focus on the human body to address partial or slow movements and immobility. Indeed, in the context of home assistance robotics, it is essential to consider the movements of individuals, periods of immobility, as well as camera movements caused by the robot's motion. The literature pays little attention to immobility and camera movements.

As part of validating our approach, many datasets for human action recognition are available, covering a wide range of general or daily activities, but they provide less coverage for abnormal situations. When specific actions are available, they are often not adapted to the context of home assistance robotics. To validate our approach, we also propose a new dataset that covers specific situations useful in the context of home assistance.

In our thesis, we investigate the optimal pairing of input modalities and architectures for human action recognition in the context of home assistance robotics. Our study resembles an optimization problem, where each architecture is associated with a subset of input modalities. Our goal is to maximize the accuracy of the action recognition system. Optimization is based on an objective function that evaluates the performance of each modality-architecture combination in terms of

accuracy. In this process, we take into account the specificities of different architectures. Our experimental study will allow us to determine the best modality-architecture combinations, thus contributing to the development of a robust and efficient solution for home assistance robotics.

Résumé

Ces dernières années, le développement de la robotique a permis l'utilisation de robots dans l'assistance au maintien à domicile (AAD). Ils sont utilisés à différents niveaux pour permettre aux personnes vulnérables, en raison de leur âge, d'une maladie ou d'un handicap, d'acquérir un certain niveau d'autonomie. Ils peuvent être utilisés, par exemple, pour aider à accomplir des tâches quotidiennes ou pour alerter les soignants ou les membres de la famille en cas de situation dangereuse ou anormale. La chute ou l'immobilité sur le sol sont des exemples importants de ces situations dangereuses ou anormales.

Notre thèse s'inscrit dans le domaine de la vision par ordinateur, et en particulier dans celui de la reconnaissance des actions humaines qui serait déployé à terme dans le contexte de la robotique d'assistance au maintien à domicile. Nous proposons une approche pour la reconnaissance des actions humaines basée sur la segmentation sémantique des parties du corps humain, afin de se concentrer sur le corps humain et de traiter les mouvements partiels ou lents ainsi que l'immobilité. En effet, dans le contexte de la robotique d'assistance au maintien à domicile, il est essentiel de considérer les mouvements des personnes, les périodes d'immobilité, ainsi que les mouvements de la caméra provoqués par le déplacement du robot. La littérature accorde peu d'attention aux immobilisations et aux mouvements de caméra.

Dans le cadre de la validation de notre approche, de nombreux ensembles de données destinés à la reconnaissance des actions humaines sont disponibles et couvrent un large éventail d'activités générales ou quotidiennes, mais ils couvrent moins le cas des situations anormales. Lorsque des actions spécifiques sont disponibles, elles ne sont souvent pas adaptées au contexte de la robotique d'assistance. Pour valider notre approche, nous proposons également un nouveau jeu de données qui couvre certaines situations spécifiques utiles dans le contexte de l'assistance au maintien à domicile.

Dans notre thèse, nous étudions le meilleur couplage entre les modalités d'entrée et les architectures de reconnaissance d'actions humaines pour une application dans un robot d'assistance au maintien à domicile. Notre étude s'apparente à un problème d'optimisation, où chaque architecture est associée à un sous-ensemble de modalités d'entrée. Notre objectif est de maximiser la précision du système de reconnaissance d'actions. L'optimisation repose sur une fonction objectif évaluant les performances de chaque combinaison modalité-architecture en termes de précision. Dans ce processus, nous prenons en considération les spécificités des différentes architectures. Notre étude expérimentale nous permettra de déterminer les meilleures associations modalité-architecture, contribuant ainsi à la création d'une solution robuste et efficace pour la robotique d'assistance au maintien à domicile.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Motivations générales

En 2050, les personnes de plus de 65 ans représenteront 1/3 de la population française [2]. 90% d'entre elles estiment qu'il est préférable de vivre à domicile pour bien vieillir [3]. Cependant l'isolement social, qui concerne aujourd'hui plus de 8,5 millions de personnes, engendre des phénomènes de dépression, de perte d'autonomie et de repli sur soi [4] sans oublier le manque grandissant de personnel de soins et d'aidants.

D'un autre côté, on dénombre 2 millions de chutes chaque année chez les plus de 65 ans. Leurs conséquences peuvent être graves. Elles entraînent plus de 130 000 hospitalisations et 12 000 décès. Une chute sur deux a lieu à domicile lors d'activités quotidiennes. Aussi, 40% des personnes hospitalisées après une chute ne peuvent plus retourner vivre chez elles, occasionnant une éventuelle institutionnalisation [5][6][7].

1.1.1 Aide au maintien à domicile

La technologie permet de contribuer à une meilleure qualité de vie pour les bénéficiaires tout en apportant un soutien technique aux professionnels de santé et aux aidants familiaux. Les solutions technologiques d'aide au maintien à domicile et d'assistance à la personne sont en cours de développement. Elles permettent de couvrir de nombreux besoins pour les personnes en perte d'autonomie. Il existe

aujourd'hui de nombreuses solutions adaptées à chaque besoin suivant le degré de dépendance et les pathologies. Parmi ces technologies d'aide au maintien à domicile déployées actuellement, nous pouvons citer :

— les solutions de téléassistance ou de téléalarme qui permettent de sécuriser le logement et les habitants, tout en favorisant la communication avec les proches, les aidants ou les secours ;
— les solutions de détection d'anomalies ou les bracelets connectés qui permettent de lancer des alertes en cas de chutes ou d'accidents domestiques. Ces solutions permettent aussi de réaliser un suivi de certaines constantes de santé des utilisateurs.

Pour beaucoup de produits commercialisés les fonctionnalités sont limitées, en particulier lorsqu'elles sont appliquées aux personnes fragilisées par l'âge, la maladie ou le handicap [8] [9]. Ainsi les systèmes basés sur des accéléromètres peuvent se déclencher lors d'un mouvement brusque ou se déclencher de manière inappropriée lors d'une chute lente. Nous pouvons aussi citer l'exemple des systèmes basés sur les boutons d'appel qui peuvent être déclenchés malencontreusement, ou à l'inverse, la personne en danger peut ne pas être dans la capacité d'appuyer sur le bouton après une chute.

1.1.2 La robotique d'assistance

Certaines solutions technologiques pour l'aide au maintien à domicile se basent sur la robotique d'assistance.

La robotique d'assistance permet aux personnes vulnérables d'acquérir un certain niveau d'autonomie tout en offrant une présence rassurante qui permet de réduire leur sentiment de solitude [10]. Certains robots d'assistance sont déjà utilisés dans les maisons de retraite et les hôpitaux pour soutenir le personnel médical.Par exemple, le robot ROMEO peut aider à la mobilité en fournissant un support physique lors des déplacements ou en aidant à réaliser des gestes quotidiens. Le robot PARO, quant à lui, est un robot de compagnie thérapeutique en forme de phoque, capable de réagir aux caresses et aux paroles. Il apporte du réconfort aux

personnes âgées.

La smart surveillance, intégrée à la robotique d'assistance peut renforcer l'efficacité des dispositifs d'aide au maintien à domicile. Grâce à des capteurs avancés et des algorithmes intelligents, ces robots peuvent analyser l'environnement et détecter les comportements inhabituels ou les situations anormales. Lorsque des comportements ou des situations inattendus sont détectés, les robots d'assistance peuvent lancer des alertes instantanées aux aidants ou aux services médicaux appropriés. Ces alertes permettent une intervention rapide et efficace, réduisant ainsi les risques pour la sécurité et la santé des personnes âgées.

Ainsi, la robotique d'assistance au maintien à domicile répond aux enjeux sociaux et sociétaux en offrant une assistance complète et une surveillance proactive. Elle devient un véritable assistant de vie, contribuant à améliorer la qualité de vie et la sécurité des personnes âgées.

1.1.3 Les contraintes des robots d'assistance à la personne

Les systèmes robotiques pour l'assistance au maintien à domicile doivent répondre à plusieurs contraintes. Ils doivent, dans un premier temps, être facilement acceptés par les utilisateurs [11]. Ils doivent aussi être dans la capacité d'évoluer au sein de leur espace de fonctionnement tout en récoltant des informations. En ce qui concerne la partie interaction avec leurs utilisateurs, ils doivent avoir des fonctionnalités attractives, et être rassurants pour être plus facilement adoptés. Le respect de la vie privée des utilisateurs est donc essentiel. Il permet la bonne acceptation du système robotique, et la protection des données utilisateurs [12].

Dans ce contexte, le traitement en local des données permet de réduire les risques de divulgation de données personnelles suite aux transferts et aux stockages externes. Ce traitement des données en local impose les contraintes liées aux systèmes embarqués. En effet, les systèmes robotiques ont des ressources matérielles et énergétiques limitées. Si les données sont traitées au sein du robot, les traitements ne doivent pas être coûteux (mémoire, calculs) au risque de limiter

fortement la durée de fonctionnement du système robotique.

Un autre aspect concerne les précisions des systèmes. Les systèmes robotiques pour l'aide à l'assistance à domicile devraient être capables de détecter et de lancer une alerte en présence d'une situation anormale requérant une assistance. Cependant, ils ne devraient pas pour autant lancer des alertes lorsque cela n'est pas nécessaire. Il est donc très important de réduire le nombre de fausses alertes tout en évitant de manquer une réelle situation dangereuse.

Pour pouvoir comprendre, interpréter et détecter les situations à risque, les robots doivent être capable de modéliser l'environnement qui les entoure et suivre son évolution, tout en prenant en compte la présence et le comportement humain.

1.1.4 Contexte industriel

Cette thèse a été cofinancée par l'entreprise CareClever et le Fonds européen de développement régional (FEDER).

L'entreprise CareClever développe un robot d'assistance pour les personnes âgées et celles ayant des besoins particuliers.

Ce robot nommé Cutii, est un robot mobile et autonome permettant d'améliorer la qualité de vie de ses utilisateurs en fournissant une assistance aux personnes fragilisées. Il a pour but d'empêcher l'isolement social et de renforcer la sécurité de ses utilisateurs.

Ainsi, ce robot compagnon est capable de fournir une assistance quotidienne, de stimuler la communication sociale et de promouvoir le bien-être émotionnel. Grâce à sa polyvalence, Cutii peut aider les utilisateurs à accomplir diverses tâches, telles que la gestion des rendez-vous médicaux, la diffusion de vidéos éducatives, le maintien des fonctions cognitives via des exercices, la communication avec la famille via des appels vidéos, et même encourager des activités physiques pour maintenir la santé.

Il est déployé dans des maisons de retraite, des établissements de santé et des domiciles de personnes fragilisées par l'âge, la maladie ou le handicap.

L'objectif de cette thèse dans ce contexte industriel est de permettre à Cutii d'effectuer de la surveillance intelligente pour assurer la sécurité de ses utilisateurs. La tâche principale est la conception et l'intégration dans le robot Cutii d'une IA avancée permettant au robot, via ses caméras, de détecter les situations anormales qui pourraient survenir, telles que les chutes, les malaises ou les situations d'urgence médicale. Cette intégration est cruciale pour assurer la sécurité et la tranquillité d'esprit des utilisateurs et de leurs familles.

La thèse a donc été réalisée en étroite collaboration avec l'entreprise pour déterminer les situations à risque, concevoir un jeu de données répondant à leurs attentes, effectuer un benchmark sur les cartes conçues pour embarquer de l'IA dans des systèmes robotiques, concevoir et entraîner une IA capable de fonctionner au sein de leur robot et l'intégrer dans le robot.

1.2 La compréhension des actions humaines

La compréhension des actions humaines est un domaine de recherche en intelligence artificielle qui vise à développer des systèmes capables de reconnaître et d'interpréter les actions humaines à partir de données vidéos. Elle permet aux systèmes numériques de voir et de comprendre ce qui les entourent de la même manière qu'un être humain.
Cela peut être utile dans de nombreux contextes, comme la surveillance des lieux publics, la reconnaissance de mouvements pour les jeux vidéo ou encore la robotique d'assistance.

Le problème de la compréhension des actions humaines peut être divisé en deux sous-domaines : la segmentation des actions et la reconnaissance des actions (voir la figure 1.1).

La segmentation des actions est un problème en traitement d'image qui consiste à diviser une vidéo en segments correspondant à des actions distinctes. Plutôt que

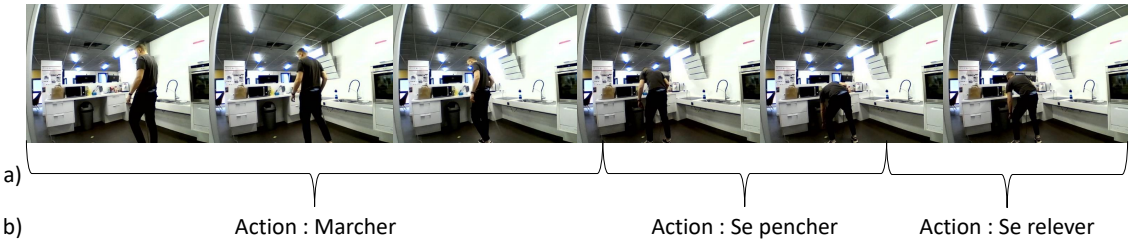

FIGURE 1.1 – Les deux sous-domaines de la compréhension des actions : la segmentation des actions (a) et la reconnaissance des actions (b).

de simplement identifier l'action globale, la segmentation des actions permet de localiser précisément les instants où une action commence et se termine dans la séquence. Par exemple, dans le contexte de l'aide au maintien à domicile, la segmentation des actions pourrait être utilisée pour détecter les moments précis où une personne commence à préparer un repas et le moment où elle termine cette tâche, en fournissant une délimitation temporelle précise pour chacune des actions. Cela peut être utile pour identifier les différentes actions qui se produisent dans une vidéo et pour les classer en différentes catégories. La segmentation des actions peut être difficile en raison de la complexité des scènes et des mouvements des personnes. Il peut être difficile de distinguer les différentes actions dans la vidéo et de les séparer en segments cohérents.

La reconnaissance d'actions humaines (RAH) consiste, à partir de vidéos et donc de séquences d'images, à décrire, suivre, analyser et reconnaître les activités et les mouvements effectués par une ou plusieurs personnes dans les segments.

Elle trouve de nombreuses applications. Par exemple, dans le domaine de la surveillance intelligente, elle permet la détection des comportements suspects dans un environnement et peut être utilisée pour prévenir les intrusions ou les incidents indésirables.

Elle peut également être utilisée pour l'indexation de contenu dans des vidéos, permettant une recherche et une récupération facile de séquences spécifiques en fonction des actions humaines qu'elles contiennent.

Dans le domaine de la sécurité, elle permet de détecter des comportements suspects ou anormaux, tels que l'errance dans des zones restreintes ou des mouvements brusques. Cela peut contribuer à renforcer la sécurité dans les aéroports,

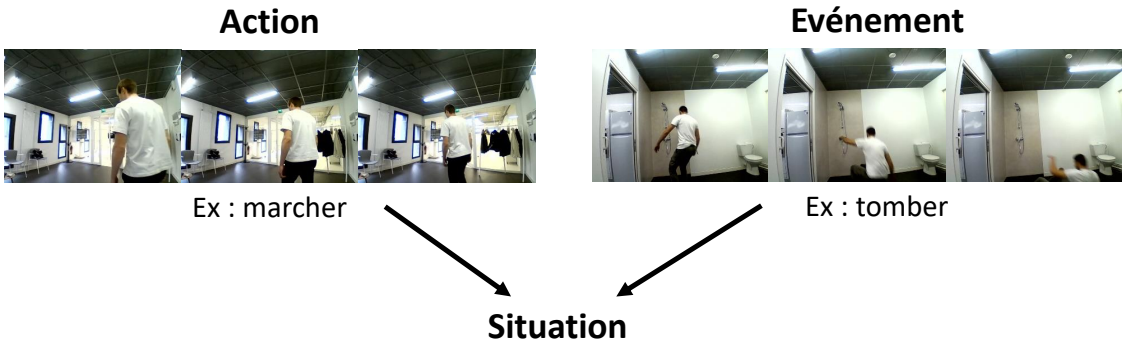

FIGURE 1.2 – Distinction entre une action, un événement et une situation.

les bâtiments gouvernementaux ou les installations sensibles.

Dans le contexte de l'aide au maintien à domicile, la reconnaissance d'actions humaines pourrait détecter et classifier des activités telles que se lever d'une chaise, préparer un repas, prendre des médicaments, etc.

Pour simplifier, on peut considérer qu'il existe une distinction claire entre les statuts d'événement et d'action comme cela est illustré sur la figure 1.2.

Événement : L'événement est un phénomène se produisant dans la nature sous l'effet d'une cause. Par exemple, une personne qui fait un malaise soudainement est un événement qui peut être causé par des facteurs tels que des problèmes de santé, la chaleur excessive ou le stress.

Action : L'action est la conduite d'un humain (ou d'une entité anthropomorphisée) doté d'une raison d'agir (motif) et d'une intention. Les actions peuvent être physiques, telles que marcher, parler ou écrire, ou mentales, telles que réfléchir, décider ou planifier.

On conclut de cette dichotomie que l'événement peut être expliqué par des lois alors que l'action humaine ne peut être que comprise, c'est-à-dire interprétée.

La compréhension de l'action humaine est cependant moins transparente qu'on ne pourrait le penser et résulte toujours d'un processus interprétatif complexe. À ce propos, deux éléments doivent être pris en compte.

Toute conduite humaine n'est pas forcément totalement intentionnelle et il existe des degrés de motivation et de responsabilité de l'agent humain.

L'action humaine n'est pas la production d'un acteur solitaire, mais s'inscrit toujours dans un contexte historique, social et culturel régi par des normes.

Situation : Une situation est un état ou une condition spécifique dans lequel une personne ou un groupe se trouve. Elle peut être liée à l'environnement physique, aux circonstances sociales, aux émotions ou aux relations entre les individus. Les situations peuvent être positives, négatives ou neutres, et elles peuvent varier en termes de complexité et de dynamisme. Les situations peuvent également être influencées par des facteurs internes et externes, par des événements et des actions passées ou en cours.

Une situation peut évoluer avec le temps, les actions effectuées et les événements en cours. Elle regroupe à la fois les événements et les actions.

Dans cette thèse, nous prenons en considération les situations. Nous considérons donc à la fois les actions et les événements.

1.3 Les défis

Dans cette thèse nous nous sommes concentrés sur le problème de la reconnaissance d'actions humaines pour une application dans un système robotique pour l'assistance au maintien à domicile.

La robotique d'assistance au maintien à domicile se développe, avec des fonctionnalités de support et d'assistance permettant aux personnes fragilisées d'acquérir un meilleur niveau de vie et de vivre dans un environnement plus sécurisant. L'un des enjeux de ces robots d'assistance est de pouvoir alerter les soignants ou les membres de la famille en cas de situations dangereuses ou anormales. Les chutes ou l'immobilité sont des exemples de ces situations rares mais importantes.

Les robots mobiles d'assistance au maintien à domicile intègrent une ou plusieurs caméras pour la localisation, l'évitement des obstacles et l'interaction avec l'utilisateur. Ces capteurs sont de plus en plus utilisés pour comprendre le comportement de l'utilisateur. Cependant, certaines contraintes de la robotique mobile rendent cette tâche plus compliquée. Comme illustré sur la figure 1.3, le mouvement du robot entraîne des changements de rotation (figure 1.3-a), de translation (figure 1.3-b) ou des deux en même temps (figure 1.3-c) dans le flux vidéo.

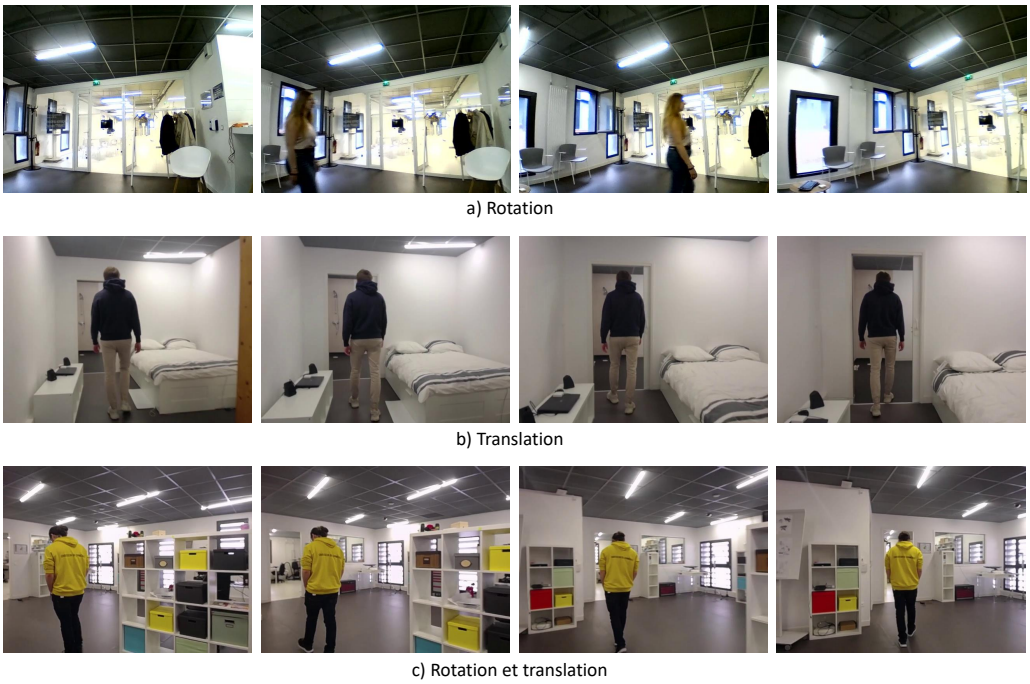

FIGURE 1.3 – Présentation des différents mouvements du robot qui induisent des déplacements de la caméra dans le flux vidéo. Les images sont espacées d'une seconde.

Un autre défi est la prise en compte des situations statiques, comme dormir, être assis ou être allongé sur le sol, qui ne génèrent aucun mouvement de la part des humains. Ces situations compliquent le processus de reconnaissance d'actions humaines et montrent les limites des approches par flux optique, qui sont actuellement les plus répandues dans l'état de l'art. De plus, les modèles sont souvent intégrés dans des dispositifs contraints (mémoire, puissance de calcul et énergie limitées), ce qui réduit considérablement leurs ressources.

Il existe un grand nombre de jeux de données disponibles pour former et tester les modèles de reconnaissance d'actions humaines. Certains de ces jeux de données couvrent un large éventail d'actions liées aux gestes quotidiens. D'autres jeux de données permettent la détection de chute. Cependant, malgré cette grande quantité de données existantes, la plupart des ensembles de données ne présentent pas de situations rares et anormales autres que les chutes. Sur les centaines de milliers de vidéos disponibles correspondant à des milliers d'actions, il y a un manque de

données pour les événements rares tels que les évanouissements, les immobilités ou les postures anormales et dangereuses. Il existe très peu de classes qui pourraient être exploitées pour l'assistance au maintien à domicile (AAD). Ce problème et cette thématique ont été soulevés dans le workshop AVA (Accessibility, Vision, and Autonomy Meet) de CVPRW 2023 (Computer Vision and Pattern Recognition workshop).

1.4 Les contributions de la thèse

Notre thèse s'inscrit dans le domaine de la vision par ordinateur et en particulier le domaine de la reconnaissance des actions humaines. La problématique abordée dans cette thèse porte sur la reconnaissance de situations normales et anormales dans le domaine spécifique de la robotique d'assistance au maintien à domicile. L'objectif est de répondre aux contraintes propres à la robotique et aux exigences de l'aide au maintien à domicile des personnes fragilisées.

L'enjeu principal consiste à explorer et à expérimenter des approches novatrices de reconnaissance d'actions humaines afin de détecter des situations dangereuses pouvant survenir à domicile, telles que les chutes, les attaques cardiaques ou encore les immobilités.

Cette thèse aborde le problème d'optimisation qui consiste à trouver le meilleur couple entre une ou plusieurs modalités d'entrée et une architecture de reconnaissance d'actions humaines. Elle s'attache à réaliser une comparaison approfondie des diverses modalités existantes pour la reconnaissance d'actions humaines, notamment l'analyse des mouvements, l'exploitation des postures et différentes segmentations sémantiques. Dans le contexte de la robotique d'assistance au maintien à domicile, cette étude examine les performances de chaque modalité en utilisant différentes architectures de classification afin de déterminer le meilleur couple modalité-architecture.

Les résultats de cette recherche mettent en évidence que l'utilisation de la segmentation sémantique des parties du corps humain fusionnée avec l'image RGB offre les meilleures performances parmi les diverses modalités de reconnaissance

étudiées. La segmentation sémantique des parties du corps humain permet d'identifier de manière précise les différentes parties du corps humain et d'assigner des étiquettes sémantiques à chaque région, ce qui facilite grandement la reconnaissance des actions. L'image RGB permet de garder le contexte. Cette approche se révèle particulièrement efficace dans le contexte spécifique de la robotique d'assistance au maintien à domicile, où la prise en compte des immobilités et des mouvements du robot revêt une importance cruciale.

Notre première contribution, par rapport à l'état de l'art, est donc l'utilisation de la segmentation sémantique des parties du corps humain fusionné avec les images RGB pour la reconnaissance d'actions humaines.

Notre deuxième contribution, par rapport à l'état de l'art, est la création d'un jeu de données spécifiquement dédié à la robotique d'assistance au maintien à domicile. Ce jeu de données comprend des données cruciales pour la détection de situations dangereuses, telles que des chutes, des attaques cardiaques ou encore des immobilités avec des mouvements de caméra induit par les mouvements du robot. Ce nouveau jeu de données permettra la reconnaissance d'un ensemble prédéfini de situations dangereuses dans le cadre de la robotique d'assistance au maintien à domicile et permettra d'enrichir les données disponibles pour la recherche dans ce domaine.

1.5 Plan du manuscrit

Dans le chapitre suivant, nous allons présenter un état de l'art sur les différentes modalités existantes pour la reconnaissance d'actions humaines à partir de données vidéos. Nous explorerons ainsi les modalités basées sur le mouvement et sur l'humain. Nous discuterons ensuite des limites des modalités existantes dans le contexte de notre application.

Puis nous explorerons dans le chapitre 3 les différentes architectures existantes pour la reconnaissance d'actions humaines.

Dans le chapitre 4, nous proposerons notre approche. Pour cela nous allons montrer les différentes modalités et les différentes architectures que nous avons

expérimentées. Ensuite nous exposerons notre approche de résolution du problème d'optimisation qui permet de déterminer le meilleur couple entre les modalités d'entrée et les architectures prises en compte.

Dans le chapitre 5, nous présenterons un nouveau jeu de données. Pour cela, nous passerons en revue les différents jeux de données de la littérature. Nous commencerons par explorer les jeux de données génériques pour la reconnaissance d'actions humaines. Après avoir présenté leurs limites, nous présenterons les jeux de données dédiés aux situations anormales et ceux dédiés à la détection de chutes.

Puis nous présenterons un nouveau jeu de données pour la reconnaissance d'actions humaines dans le cadre de la robotique d'assistance au maintien à domicile. Dans ce jeu de données, nous nous concentrerons sur trois aspects :

— le premier est la capture des vidéos du point de vue du robot d'assistance (en déplaçant et en faisant tourner les caméras pour certaines vidéos) ;
— le second correspond à des situations anormales comme la chute ou l'allongement au sol, et des actions qui pourraient être confondues avec elles en générant de fausses alertes (être allongé dans un lit est différent d'être allongé à terre) ;
— enfin, le dernier correspond à l'aspect dynamique ou statique des situations et du robot. Ainsi certaines vidéos présentent des sujets statiques avec une caméra en mouvement ou immobile, et d'autres vidéos présentent des sujets en mouvement avec une caméra en mouvement ou immobile.

Dans le chapitre 6, nous expérimenterons les différentes approches. Après avoir expérimenté les différentes modalités sur différentes architectures, nous montrerons que l'utilisation de la segmentation partielle ou totale pour la reconnaissance d'actions humaines, qui n'est pas largement considérée dans la littérature, présente de meilleurs résultats que les méthodes basées sur le mouvement dans le contexte de la robotique d'assistance au maintien à domicile.

Enfin nous conclurons dans le dernier chapitre.

La liste des contributions est visible dans le chapitre 8.

Chapitre 2

Différentes modalités d'entrée pour la reconnaissance d'actions humaines

La reconnaissance des actions humaines (RAH) est une tâche fondamentale en vision par ordinateur. Elle vise à analyser une vidéo pour pouvoir identifier et définir le type d'actions effectuées par une ou plusieurs personnes. Elle trouve de nombreuses applications telles que les interactions homme-machine, la sécurité ou encore la prévention.

La reconnaissance d'actions humaines se base sur des séquences d'images formant une vidéo.

Les premiers travaux se sont concentrés sur l'utilisation de vidéos RGB comme entrée pour la reconnaissance d'actions humaines, en raison de sa popularité et de sa facilité d'accès. Avec le développement de différents types de capteurs et l'augmentation des puissances de calcul, il existe différents types de modalités d'entrée [13].

Dans le cadre de cette thèse, nous nous concentrons uniquement sur les modalités liées à la vision par ordinateur. Les derniers travaux, basés sur la vision par ordinateur, utilisent différentes modalités d'entrée en pré-traitant les séquences d'images.

Ainsi, dans la littérature, beaucoup de méthodes d'apprentissage pour la reconnaissance d'actions humaines se basent sur des séquences d'images pré-traitées.

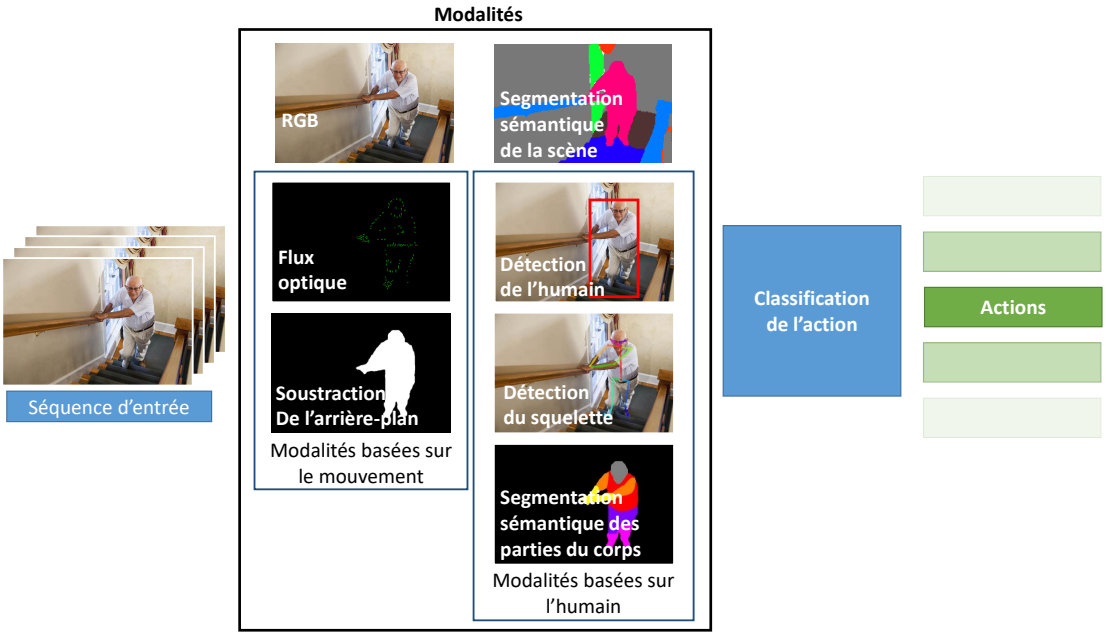

FIGURE 2.1 – Différents types de modalités d'entrée pour la reconnaissance d'actions humaines.

Il existe différentes modalités d'entrée comme illustré sur la figure 2.1. Il existe la modalité RGB et la segmentation sémantique de la scène. Il existe également des modalités basées sur le mouvement comme la soustraction de l'arrière-plan ou l'estimation de flux optique et des modalités basées sur l'humain comme la détection de l'humain ou l'estimation de la pose humaine ou encore la segmentation sémantique des parties du corps humain.

Ces différentes modalités sont présentées dans ce chapitre.

2.1 RGB

Les images RGB sont généralement les plus faciles à collecter puisqu'elles font référence aux images ou aux vidéos capturées par les caméras RGB. Elles sont très proches de ce que voient les yeux humains. Elles contiennent beaucoup d'informations sur l'environnement et le contexte de la scène d'action. Cependant, cette richesse rend la reconnaissance d'actions humaines à partir de données RGB difficiles. En effet, les données RGB contiennent souvent beaucoup de variations de l'arrière-plan, des points de vue, d'échelles et des conditions d'éclairage. De plus,

les séquences d'images RGB ont généralement une grande taille de données due à la masse d'informations capturées. Cela entraîne des coûts de calcul élevés lors de la modélisation de l'image spatio-temporelle.

D'autres approches peuvent également être utilisées, comme la reconnaissance d'actions humaines à partir de données de profondeur, qui utilisent des capteurs de profondeur pour fournir des informations sur la distance des objets à la caméra. Ces approches peuvent être moins sensibles aux variations de l'arrière-plan et des points de vue, mais elles peuvent être moins précises que la reconnaissance d'actions humaines à partir de données RGB.

2.2 Segmentation sémantique de la scène

La segmentation sémantique de la scène est une tâche de vision par ordinateur qui consiste à diviser une image en régions sémantiques cohérentes. Elle permet d'attribuer à chaque pixel une étiquette de classe. Comme la prédiction se fait pour chaque pixel de l'image, cette tâche est communément appelée prédiction dense.

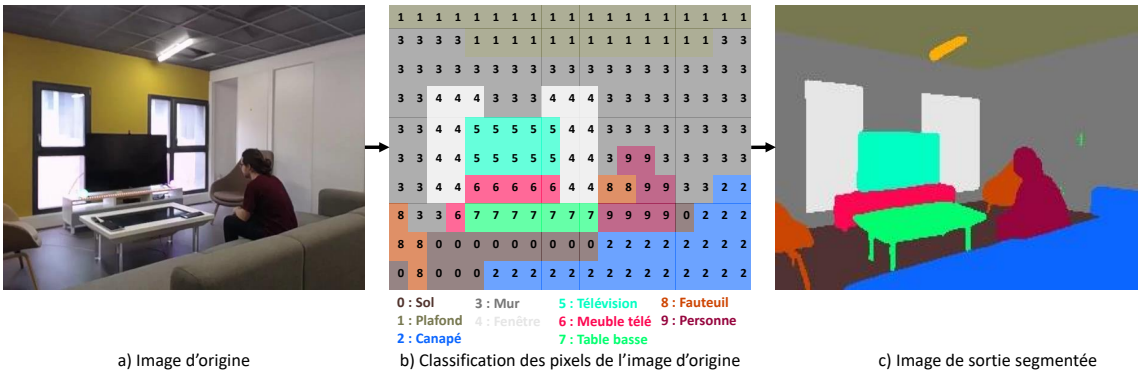

FIGURE 2.2 – Présentation de la segmentation sémantique de la scène. Elle consiste à classer chaque pixel de l'image d'entrée dans des classes définies.

Les réseaux de neurones convolutionnels (CNN) ont été les premiers modèles d'apprentissage profond à être utilisés pour la segmentation sémantique. En 2014, Long et al. ont proposé un réseau de neurones convolutionnels appelé FCN pour *Fully Convolutional Network* [14] pour la segmentation sémantique de la scène de l'image. FCN présente certaines limites, comme l'absence de contexte global et le

fait qu'il ne peut pas être utilisé en temps réel.

Depuis lors, de nombreuses variantes de FCN ont été proposées comme ParseNet [15] qui améliore le FCN en prenant en compte le contexte global. Cette architecture utilise des cartes de caractéristiques globales normalisées et dégroupées pour améliorer les résultats de la segmentation sémantique.

Les architectures codeur-décodeur ont également été largement utilisées pour la segmentation sémantique. Par exemple, le *Learning Deconvolution Network for Semantic Segmentation* [16] utilise un encodeur basé sur un réseau convolutionnel pour générer des cartes de caractéristiques, tandis que le décodeur utilise des architectures de déconvolution pour produire une carte de segmentation pixel par pixel. Cette approche permet d'étendre la carte des caractéristiques sans perdre d'informations importantes.

SegNet [17] est un autre modèle basé sur une architecture codeur-décodeur. Il stocke des informations sur les indices de mise en commun lors de l'étape de mise en commun maximale et les utilise pour l'échantillonnage ascendant, améliorant ainsi la précision de la segmentation.

Les modèles multi-échelles et basés sur les réseaux pyramidaux ont également contribué aux progrès de la segmentation sémantique. Le Feature Pyramid Network (FPN) [18] utilise une voie ascendante pour créer des cartes de caractéristiques à plusieurs échelles, une voie descendante pour suréchantillonner et des connexions latérales pour fusionner les informations. PSPNet [19] utilise un module de regroupement pyramidal pour capturer les informations contextuelles à différentes échelles.

Les architectures basées sur R-CNN, comme PANet [20], ont également été utilisées pour la segmentation d'instances. Ces modèles étendent l'architecture R-CNN en introduisant une branche de sortie de segmentation binaire et en utilisant des couches de mise en commun adaptatives pour améliorer la précision.

Les architectures de DeepLab comme le DeepLabV3 [21], ont également contribué de manière significative à la segmentation sémantique. Ces modèles utilisent des stratégies telles que l'*Atrous Spatial Pyramid Pooling (ASPP)* [22] pour maintenir une grande résolution sans augmenter le nombre de paramètres, et des modules en cascade pour capturer les objets à différentes échelles.

Plus récemment, les Transformers ont également été utilisés pour la segmentation sémantique. Parmi eux, le modèle *Vision Transformer (ViT)* [23] a montré

des performances prometteuses.

2.3 Modalités basées sur le mouvement

La principale différence entre les vidéos et les images est que les vidéos ont une structure temporelle en plus de la structure spatiale que l'on trouve dans les images. Cela signifie que les informations contenues dans une vidéo sont codées non seulement dans l'espace (c'est-à-dire dans les objets ou les personnes présentes dans la vidéo), mais aussi de manière séquentielle et dans un ordre spécifique.

Les méthodes basées sur les images RGB apprennent les caractéristiques basées sur le contexte de l'image et non sur l'action inhérente. Cela signifie qu'ils n'utilisent pas la représentation du mouvement comme caractéristique de classification, mais qu'ils apprennent à utiliser des indices spatiaux pour comprendre les informations temporelles contenues dans la vidéo.

D'autres modalités permettent de donner la priorité au mouvement en tant que caractéristique clé de la classification. C'est le cas des méthodes basées sur la soustraction de l'arrière-plan ou de l'estimation de flux optique.

2.3.1 Soustraction de l'arrière-plan

Certaines méthodes de reconnaissance d'actions humaines se basent sur la soustraction de l'arrière-plan [24]. La soustraction de l'arrière-plan permet d'extraire le premier plan (objet en mouvement) et l'arrière-plan (objet fixe) de l'image en vue d'un traitement ultérieur. Cela permet de concentrer l'apprentissage sur les objets en mouvement, donc sur l'humain effectuant l'action, tout en éliminant les autres informations.

Les premières méthodes de soustraction de l'arrière-plan calculaient la différence entre l'image actuelle et une image d'arrière-plan de référence. Les pixels qui ont des valeurs différentes entre l'image de référence et l'image en cours d'acquisition sont considérés comme des objets en mouvement. Ces méthodes sont simples à mettre en place mais elles sont sensibles aux variations d'éclairage et aux objets

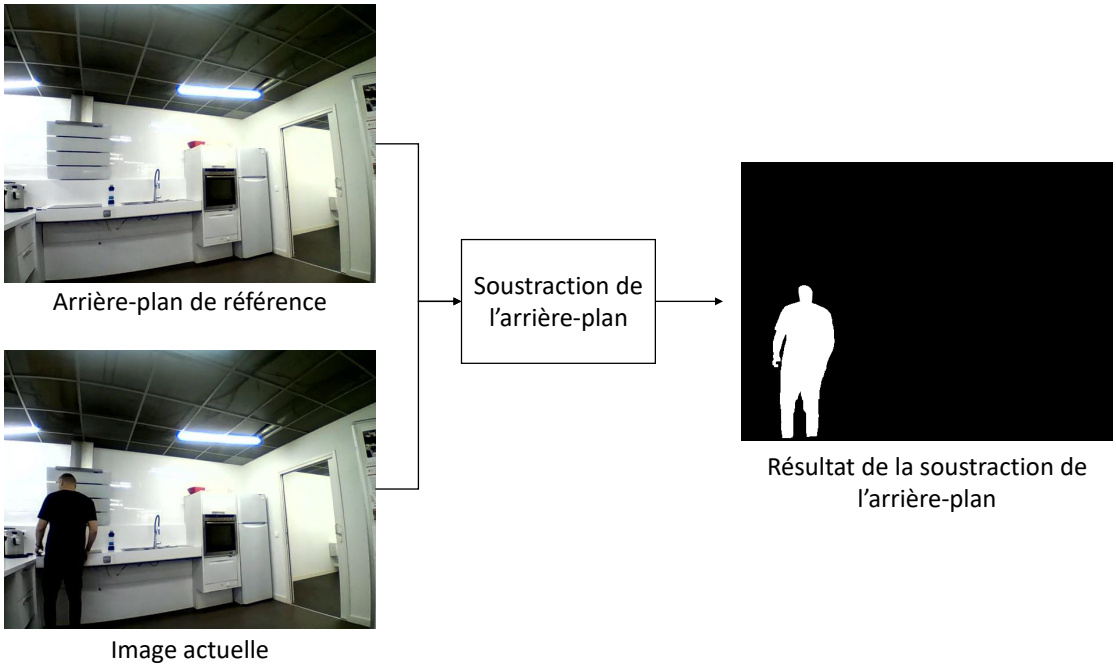

FIGURE 2.3 – Présentation de la soustraction de l'arrière-plan. C'est la différence entre une image actuelle et une image d'arrière-plan de référence.

en mouvement rapide. Ces méthodes supposent que l'arrière-plan est toujours statique. Elles ne peuvent être utilisées que dans des scénarios d'environnement fixe et de caméra fixe.

Pour pouvoir utiliser la soustraction de l'arrière-plan dans des contextes réels (variations d'éclairage, des mouvements de l'arrière-plan ou de la caméra, etc), les modèles d'apprentissage automatique se sont développés ces dernières années.

Différents modèles d'apprentissage automatique ont été utilisés pour la modélisation de l'arrière-plan et la détection du premier plan comme les modèles SVM (Support Vector Machine) [25], les modèles d'apprentissage flou [26] , et les modèles d'apprentissage subspatial [27].

Le déploiement des réseaux neuronaux convolutifs (Convolutional Neural Network (CNN)) et des réseaux neuronaux profonds (Deep Neural Network (DNN)) [28] ont permis d'améliorer considérablement les performances de la soustraction de l'arrière-plan. Les méthodes d'apprentissage profond basées sur les réseaux neuronaux profonds (DNN) et les réseaux neuronaux convolutifs (CNN) ont la capa-

cité de pallier les inconvénients du paramétrage inhérent aux réseaux neuronaux classiques. Les DNN ont permis l'intégration des techniques de détection des changements [29].

La soustraction de l'arrière-plan permet d'obtenir des informations sur les objets en mouvement tels que les silhouettes et les contours humains.

Cependant, elle est essentiellement utilisée dans un contexte de caméra fixe car il est difficile d'obtenir des silhouettes et des contours précis dans le cas de scènes complexes et de mouvements de caméra dus à la nécessité d'avoir une image d'arrière-plan de référence. Les méthodes de soustraction de l'arrière-plan basées sur l'apprentissage profond sont généralement plus complexes à mettre en place et à adapter à des applications réelles en raison de la complexité du calcul. Elles ont également une certaine sensibilité aux bruits dans les images ou les vidéos, ce qui peut entraîner des erreurs de détection.

2.3.2 Flux optique

Une autre méthode permettant de mettre en évidence les mouvements dans une vidéo est l'estimation de flux optique. L'estimation de flux optique est une technique couramment utilisée en traitement d'image pour suivre les mouvements des objets dans une vidéo. L'objectif est d'estimer les vecteurs de déplacement entre deux images, montrant comment les pixels d'un objet dans la première image peuvent être déplacés pour former le même objet dans la seconde image. Il s'agit d'une sorte d'apprentissage par correspondance, car si les pixels correspondants d'un objet sont connus, le champ de flux optique peut être estimé.

Pour estimer le flux optique, les méthodes supposent une constance de la luminosité. Elles essaient d'estimer comment la luminosité des pixels se déplace au fil du temps.

Le problème du flux optique peut être exprimé de la manière suivante :

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{2.1}$$

où I est l'intensité du pixel à la position x, y au temps t.

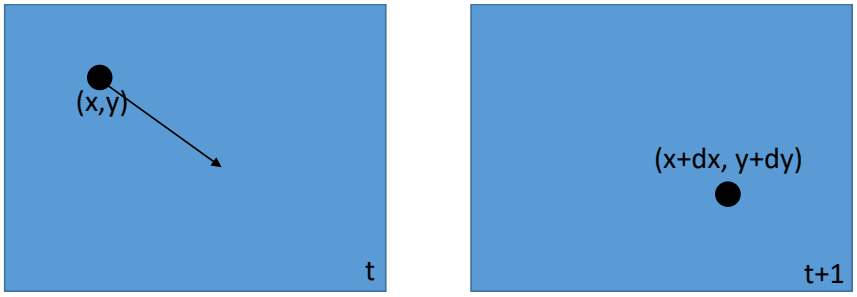

FIGURE 2.4 – Principe du flux optique. Les deux images successives illustrent le déplacement du point noir de (dx, dy).

Les caractéristiques du pixel au moment t sont les mêmes que celles du pixel au moment dt, mais à un endroit différent (désigné par dx et dy). C'est ce changement d'emplacement, illustré sur la Figure 2.4 qui est prédit par l'estimation de flux optique.

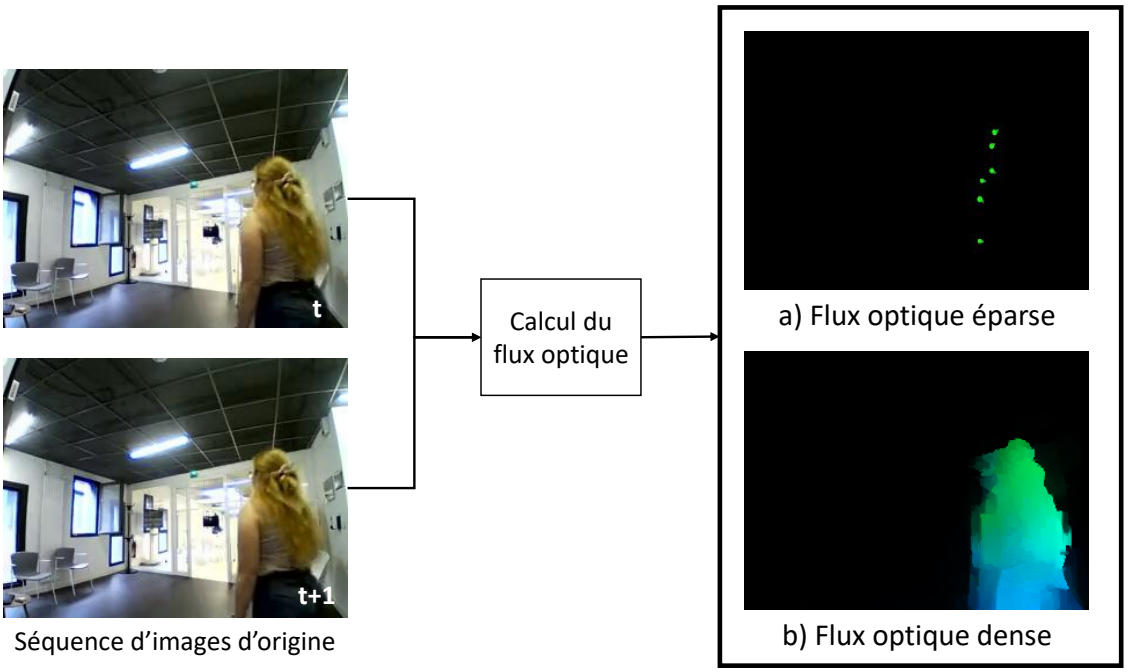

FIGURE 2.5 – Différents types d'estimation de flux optique. En a) le flux optique épars donne les vecteurs de flux des caractéristiques intéressantes. En b) le flux optique dense donne les vecteurs de flux de l'ensemble des pixels de l'image.

Il existe différents types de flux optique comme illustré sur la figure 2.5.

Le flux optique épars donne les vecteurs de flux de certaines caractéristiques intéressantes comme par exemple certains pixels représentant les bords ou les coins d'un objet. Le flux optique épars permet alors de suivre ces vecteurs et de modéliser le mouvement. Les caractéristiques extraites sont transmises à la fonction de flux optique d'une image à l'autre pour s'assurer que les mêmes points sont suivis. Il existe plusieurs implémentations du flux optique épars, notamment la méthode Lucas-Kanade [30] ou la méthode Horn-Schunck [31].

Le flux optique dense donne les vecteurs de flux de l'image entière, c'est-à-dire de l'ensemble des pixels. Il tente d'estimer le vecteur de flux optique pour chaque pixel de chaque image. Bien qu'une telle estimation puisse être plus lente, le flux optique dense donne un résultat plus précis et plus dense. Il existe plusieurs implémentations du flux optique dense. L'implémentation la plus populaire est celle de Farneback [32].

Si le problème du flux optique est historiquement un problème d'optimisation, des approches récentes utilisant l'apprentissage profond ont donné des résultats impressionnants. Ces approches prennent deux images successives de la séquence vidéo en entrée pour produire le flux optique dense (image codée en couleur), qui peut être exprimé comme suit :

$$(u, v) = f(I_{t-1}, I_t) \tag{2.2}$$

Où u est le mouvement dans la direction x, v est le mouvement dans la direction y, et f est le réseau neuronal qui prend en entrée deux images consécutives I_{t-1} (image au temps $t-1$) et I_t (image au temps t).

Il existe de nombreuses méthodes de résolution des problèmes de flux optique par l'apprentissage profond.

DeepFlow [33] combine des techniques d'apprentissage en profondeur avec des méthodes classiques de vision par ordinateur pour estimer le flux optique.

FlowNet [34] est une architecture de réseau de neurones convolutifs profonds spécialement conçue pour l'estimation du flux optique.

RAFT [35] utilise une approche itérative et récursive pour affiner le flux optique,

en exploitant des blocs de neurones pour modéliser les relations spatiales entre les pixels.

FlowFormer [36] est une méthode récente qui adopte une architecture basée sur des Transformers et utilisant les flux optiques pour estimer le mouvement entre les images.

GMFlow [37] repose sur une combinaison de modèles graphiques et de réseaux neuronaux, exploitant des représentations denses pour estimer le flux optique.

Il est à noter que la récupération de flux optique peut se faire via un capteur de vision dynamique (DVS). Une caméra DVS permet de capturer les changements d'intensité lumineuse. Ces changements d'intensité peuvent ensuite être représentés sous forme de vecteur de déplacement comme dans l'estimation de flux optique.

L'estimation de flux optique peut être utile pour la reconnaissance d'actions humaines, car elle permet de mettre en évidence les parties en mouvement dans une vidéo [38]. Seules les parties en mouvement sont visibles.

Cependant l'apprentissage est basé essentiellement sur le mouvement des personnes. En se concentrant uniquement sur le mouvement des personnes, il peut ignorer d'autres informations importantes de la scène, comme la posture générale et la position des personnes. Pour cette raison, il peut être utile de combiner l'estimation de flux optique avec d'autres modalités pour obtenir une meilleure reconnaissance d'actions humaines.

2.4 Modalités basées sur l'humain

Les méthodes basées sur le flux optique ou la soustraction de l'arrière-plan se focalisent essentiellement sur le mouvement des objets dans la scène, sans tenir compte de l'humain en tant qu'objet distinct. Elles font de la reconnaissance d'actions humaines sans prendre en compte l'humain. Cela peut limiter leur capacité à reconnaître les actions humaines dans des contextes complexes avec la présence de mouvement lent et partiel ou encore la présence d'immobilité.

Depuis quelques années se sont développées des approches qui utilisent des re-

présentations du corps humain pour la reconnaissance d'actions humaines. Ces approches peuvent utiliser des techniques de détection de personnes pour repérer les personnes dans la scène, ou utiliser la modélisation du squelette humain pour suivre les mouvements des membres et de la tête des personnes. Ces approches peuvent fournir des informations plus précises sur la position et la posture des personnes dans la scène, ce qui peut améliorer la reconnaissance d'actions humaines.

2.4.1 Détection de l'humain

Les méthodes de détection de l'humain sont des approches utilisées en traitement d'image qui vise à détecter et à suivre les humains dans des images ou des vidéos. Ces approches peuvent utiliser des algorithmes de reconnaissance d'objets pour détecter les personnes dans la scène.

L'objectif est de déterminer l'emplacement des humains dans les images. Les premières méthodes utilisent des boîtes d'ancrage. Pour sélectionner les zones d'intérêt, on utilisait une fenêtre coulissante qui se déplaçait sur l'image. Cette fenêtre avait une taille et une forme fixes. À chaque position, elle extrayait des caractéristiques pour déterminer la présence ou l'absence d'une personne.

Cependant, les personnes peuvent avoir différentes apparences et tailles. De plus, la taille de l'image affecte également la taille effective de la fenêtre. Si on utilise un réseau de neurones convolutif (CNN) ou un réseau de neurones profonds (DNN) pour classifier les images à chaque position, ce processus devient extrêmement lent.

Dans FasterR-CNN [39] (basé sur FastR-CNN [40] et sur R-CNN [41]), au lieu de fournir les propositions de régions au CNN, c'est l'image d'entrée qui est fournie au CNN pour générer une carte de caractéristiques. Puis un réseau séparé est utilisé pour prédire les propositions de régions. Les propositions de régions prédites sont ensuite remodelées à l'aide d'une couche de mise en commun RoI *(Region of Interest)*, qui est ensuite utilisée pour classer l'image dans la région proposée et prédire les valeurs de décalage pour les boîtes de délimitation.

Tous les algorithmes de détection d'objets précédents utilisent des régions pour localiser l'objet dans l'image. Le réseau ne regarde pas l'image complète. Il examine plutôt les parties de l'image qui ont de fortes probabilités de contenir l'objet.

YOLO [42] est un algorithme de détection d'objets très différent des algorithmes basés sur les régions. Un seul réseau convolutif prédit les boîtes englobantes et les probabilités de classe pour ces boîtes. Il commence par diviser l'image en une grille SxS. Dans chacune des cases de la grille, sont prises m boîtes de délimitation. Pour chaque boîte de délimitation, le réseau produit une probabilité de classe et des valeurs de décalage pour la boîte de délimitation. Les boîtes de délimitation dont la probabilité de classe est supérieure à une valeur seuil sont sélectionnées et utilisées pour localiser l'objet dans l'image. Cette technique améliore considérablement les temps de calcul.

D'autres méthodes permettent de détecter les objets sans utiliser les boîtes englobantes.

Le modèle CornerNet[43] prédit les coins supérieurs gauches et inférieurs droits des boîtes englobantes pour chaque pixel, ainsi qu'un encastrement. Les encastrements de chaque coin sont comparés pour déterminer à quel objet ils appartiennent. Une fois tous les points mis en correspondance, il est facile de récupérer la boîte englobante.

ExtremeNet [44] est basée sur CornerNet [43], mais au lieu de prédire les coins, elle prédit le centre des objets ainsi que les points les plus éloignés à gauche, à droite, en haut et en bas. Ces "points extrêmes" sont ensuite mis en correspondance sur la base de leur géométrie, ce qui permet de les faire correspondre ensemble et de récupérer la boîte englobante.

Quelle que soit la méthode utilisée, la détection de l'humain permet de situer l'humain dans la vidéo pour concentrer l'apprentissage sur l'humain et ainsi faire abstraction du reste [45]. Cependant, ces approches peuvent également avoir des limitations, comme la difficulté à détecter les personnes dans des scènes complexes ou lorsque les personnes sont occultées.

2.4.2 Détection du squelette humain

La détection du squelette ou l'estimation de la pose humaine [46] consiste, à partir d'images, à localiser les articulations humaines.

Cela permet de modéliser le squelette sous forme de graphe. Chaque nœud du graphe correspond à une articulation (coude, genou, poignet, etc), et chaque arête du graphe correspond à un membre du corps humain (bras, torse, cou, etc).

Les données de squelette peuvent être collectées à l'aide de systèmes de capture de mouvement ou être calculées sur des données vidéos. En général, l'estimation de la pose humaine est sensible aux variations du point de vue. En revanche, les systèmes de capture de mouvement qui sont insensibles à la vue et à l'éclairage peuvent fournir des données de squelette fiables.

Cependant, dans de nombreuses applications, il n'est pas pratique de déployer des systèmes de capture de mouvement. Ainsi, de nombreux travaux récents sur la reconnaissance d'actions humaines ont utilisé des données de squelette obtenues à partir de cartes de profondeur, [47] ou de vidéos RGB [48].

Les premiers travaux se sont concentrés sur l'extraction de caractéristiques spatiales et temporelles créées à la main à partir de séquences de squelettes pour la reconnaissance d'actions humaines. Les méthodes basées sur des caractéristiques créées à la main peuvent être divisées en deux catégories : les méthodes basées sur les articulations [49] et les méthodes basées sur les parties du corps [50], en fonction des techniques d'extraction de caractéristiques utilisées.

Les méthodes classiques ont leurs limites et l'estimation de la pose a été considérablement remodelée par les CNN. Avec l'introduction de DeepPose [51], la recherche sur l'estimation de la pose humaine est passée des approches classiques à l'apprentissage profond.

Il existe deux approches concurrentes illustrées dans la figure 2.6.

L'approche simple consiste à intégrer d'abord un détecteur de personnes, puis à calculer les parties et enfin à calculer la pose de chaque personne. Cette méthode est connue sous le nom d'approche descendante.

Une autre approche consiste à détecter toutes les parties de l'image (c'est-à-

FIGURE 2.6 – Différentes approches pour la détection de squelette humain. a) Approche ascendante : Elle consiste à d'abord détecter les personnes présentes, puis pour chaque personne à détecter les articulations pour construire le squelette. b) Approche descendante : Elle consiste à d'abord détecter l'ensemble des articulations puis à détecter à quelle personne elles appartiennent pour finir par construire les différents squelettes.

dire les parties de chaque personne), puis à associer les parties appartenant à des personnes distinctes. Cette méthode est connue sous le nom d'approche ascendante.

OpenPose [46] est l'une des approches ascendantes les plus populaires pour l'estimation de la pose humaine multi-personnes. Comme pour de nombreuses approches ascendantes, elle détecte d'abord les parties (points clés) appartenant à chaque personne dans l'image. Pour cela elle utilise des cartes de confiance des parties du corps (*Part Confidence Maps*) et un champ d'affinité des parties (*Part Affinity fields (PAF)*), qui est un ensemble de champs vectoriels 2D. Le PAF code le degré d'association entre les parties détectées dans les cartes de confiance. Les cartes de confiance et les champs d'affinité des parties sont ensuite traités par un algorithme glouton pour obtenir les poses de chaque personne dans l'image.

AlphaPose [52] est une méthode descendante populaire d'estimation de la pose. Les auteurs affirment que les méthodes descendantes dépendent généralement de

la précision du détecteur de personnes, car l'estimation de la pose est effectuée sur la région où se trouve la personne. Par conséquent, les erreurs de localisation et les prédictions de boîte de délimitation en double peuvent entraîner une performance sous-optimale de l'algorithme d'extraction de pose.

L'utilisation de données de squelette pour la reconnaissance d'actions présente de nombreux avantages en raison de la structure du corps et des informations de pose qu'elles fournissent, de leur représentation simple et informative, de leur invariance d'échelle et de leur robustesse face aux variations des textures des vêtements et des arrière-plans.

Les changements des points d'articulation humaine entre chaque image sont utilisés pour décrire l'action, y compris les changements de position et d'apparence des points d'articulation [53]. Cependant, la performance de ces méthodes dépend des résultats de l'estimation de la pose humaine. Lorsqu'une occlusion se produit dans la scène, l'estimation des points communs est manquante ou incorrecte, ce qui affecte les résultats de la reconnaissance d'actions humaines.

2.4.3 Segmentation sémantique des parties du corps humain

Les méthodes de segmentation sémantique des parties du corps humain totale ou partielle se sont développées ces dernières années [54] sans arriver au niveau sémantique des actions.

Elles visent à identifier les différentes zones du corps humain en regroupant les pixels d'une image en zones sémantiques. On peut ainsi regrouper les pixels représentant une partie du corps donnée comme les jambes, les bras, la tête, etc, et ceux n'appartenant à aucune partie du corps comme l'arrière-plan, les objets, etc. La segmentation sémantique des parties du corps humain est illustrée dans la figure 2.7.

Le corps humain a une hiérarchie naturelle. A-AOG [55] est un modèle qui représente la décomposition et l'articulation des parties du corps humain. Les réseaux basés sur les graphes utilisent la convolution des graphes pour capturer les

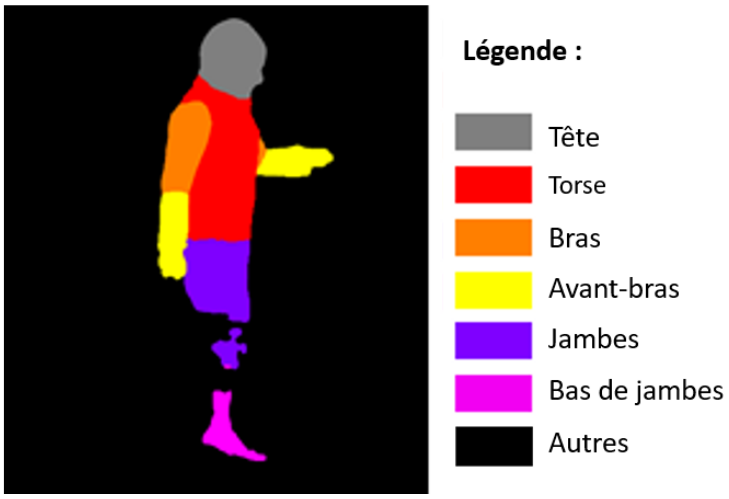

FIGURE 2.7 – Résultat de la segmentation sémantique des parties du corps humain sur une image.

relations sémantiques entre les parties du corps. Graphonomy [56] est un réseau d'analyse syntaxique humaine basé sur les graphes qui utilise la convolution des graphes pour capturer les informations globales et la cohérence sémantique.

Certains réseaux utilisent des cellules de mémoire LSTM pour retenir les informations importantes. LG-LSTM [57] combine les dépendances spatiales à courte et longue distance pour exploiter les contextes locaux et globaux.

D'autres approches combinent des informations auxiliaires telles que l'estimation de la pose humaine [58]. Ces informations sont utilisées pour guider la segmentation et améliorer la reconnaissance des modèles.

D'autres travaux utilisent un modèle de segmentation sémantique comme codeur d'image, et entraînent leur modèle à partir de ces poids initiaux. Cette méthode accélère la convergence du réseau.

Les méthodes de segmentation sémantique s'appuient sur les réseaux entièrement convolutifs comme FCN [14], les modèles basés sur le codeur-décodeur comme SegNet[17], les réseaux multi-échelles comme FPN [18] et la famille DeepLab [21].

Cependant, les méthodes existantes présentent des limites dans les scènes complexes en présence de plusieurs personnes ou d'occlusion.
Une autre limitation est le besoin de grandes quantités de données annotées, qui peuvent être coûteuses et longues à obtenir. Les données annotées sont essentielles pour l'apprentissage des méthodes basées sur l'apprentissage profond, mais l'obten-

tion d'annotations pour plusieurs individus dans la même scène peut être difficile, en particulier lorsque les individus ont des poses complexes ou se cachent les uns les autres.

2.5 Analyse critique

Il est important de noter que les limitations que nous décrivons dans cette section sont liées aux applications de reconnaissance d'actions humaines dans le contexte de l'AAD et de la robotique d'assistance. L'analyse de la prise en compte des contraintes liées à la robotique d'assistance au maintien à domicile des différents prétraitements présentés est visible dans le tableau 2.1.

La reconnaissance d'actions humaines est un domaine riche. La diversité des modalités d'entrée rend difficile l'identification d'architectures efficaces.

Malgré plusieurs modalités de reconnaissance d'actions humaines, nous remarquons que l'accent est mis sur l'analyse de l'image entière et sur la considération du mouvement plus que sur la considération de l'humain.

La reconnaissance d'actions humaines dans un contexte de l'assistance de l'autonomie à domicile (AAD) doit prendre en compte les immobilités, c'est-à-dire des situations sans mouvement de la part des personnes en plus des situations avec du mouvement.

Les méthodes pour la reconnaissance d'actions humaines présentes dans la littérature traitent principalement des actions en présence de mouvements. Toutes les méthodes de la littérature basées sur la détection des mouvements ne peuvent pas être utilisées dans le contexte de l'AAD. Elles ne permettent pas de se focaliser sur la personne mais uniquement sur les mouvements. Ces méthodes sont inutiles en cas d'immobilité et ont de mauvais résultats pour les mouvements partiels. Enfin, cette modalité n'est pas adaptée à la robotique d'assistance en raison du mouvement du robot qui entraîne un mouvement de caméra. Ces mouvements de caméra génèrent un flux optique supplémentaire sur l'arrière-plan et les objets de la scène et peuvent rendre le mouvement de la personne illisible.

Nous considérons qu'il est nécessaire de détecter l'humain dans l'étape de prétraitement et d'extraire une représentation plus riche et plus cohérente avec la compréhension du comportement humain.

Certaines des méthodes sont basées sur l'humain. C'est le cas de la détection de squelette humain. Mais en présence d'occlusion, ces méthodes présentent de mauvais résultats dus à la grande perte des informations sur la position des membres des personnes.

La segmentation sémantique de la scène ou la segmentation sémantique des parties du corps humain peut être une bonne solution pour notre problématique même si elles sont très peu utilisées pour la reconnaissance d'actions humaines.

Modalités	**Immobilité de la personne**	**Mouvement de la personne**	**Arrière-plan dynamique**	**Vue partielle de la personne**
RGB	Oui	Oui	Oui	Oui
Flux optique	Non	Oui	Non	Oui
Soustraction de l'arrière-plan	Non	Oui	Oui/Non	Oui
Détection de l'humain	Non	Oui	Oui	Oui/Non
Détection du squelette	Non	Oui	Oui	Oui/Non
Segmentation sémantique de la scène	Oui	Oui	Oui	Oui
Segmentation sémantique des parties du corps humain	Oui	Oui	Oui	Oui

TABLE 2.1 – Tableau récapitulatif de la prise en compte ou non des contraintes liées à la robotique d'assistance au maintien à domicile pour chaque modalité.

Chapitre 3

Architectures d'apprentissage profond pour la reconnaissance d'actions humaines

Les architectures pour la reconnaissance d'actions humaines ont pour objectif d'extraire les caractéristiques des données d'entrée afin de les classifier parmi des classes d'actions prédéfinies.

Les données vidéos contiennent des informations spatiales comme l'emplacement des personnes, leurs positions, les objets, etc. Elles contiennent également des informations temporelles c'est-à-dire le mouvement. Le mouvement est la différence entre deux images successives d'une vidéo. Ainsi les architectures en apprentissage profond pour la reconnaissance d'actions humaines se basent sur des données spatio-temporelles et doivent apprendre à la fois les caractéristiques spatiales et temporelles.

Les caractéristiques peuvent être apprises suivant différents degrés de granularité. Certains travaux modélisent les informations dans leur ensemble, c'est notamment le cas des approches basées sur le RGB. D'autres auteurs se basent sur des informations plus détaillées comme le squelette.

Beaucoup d'approches se basent essentiellement sur les données temporelles. Le mouvement peut définir à lui seul l'action. C'est une des informations les plus

importantes en reconnaissance d'actions humaines. Il existe différentes manières de modéliser le mouvement. Certains travaux se basent sur le flux optique quand d'autres travaux utilisent les trajectoires définies par la différence d'emplacement des articulations. D'autres utilisent des réseaux permettant de modéliser le mouvement comme les réseaux neuronaux récurrents (*Recurrent Neural Network (RNN)*).

Aujourd'hui, il existe plusieurs architectures différentes pour la reconnaissance d'actions humaines. Cependant, il n'existe pas de référence standard pour la recherche d'architecture. Le choix de l'architecture varie en fonction de l'application et des besoins spécifiques de chaque situation.

Dans ce chapitre nous verrons le fonctionnement des architectures d'apprentissage profond puis nous détaillerons les principales architectures d'apprentissage profond pour la reconnaissance d'actions humaines. Elles peuvent être résumées comme suit : les architectures basées sur la mémoire à long terme (LSTM) [59], les architectures basées sur les réseaux convolutionnels 3D [60], celles basées sur les réseaux convolutionnels à deux flux [61], celles basées sur les mécanismes d'attention [62] qui deviennent majoritaires dans la littérature. Nous étudierons également les méthodes de description des images appliquées à la reconnaissance d'actions humaines.

3.1 Réseaux de neurones convolutifs

Les architectures pour la reconnaissance d'actions humaines peuvent se baser sur des réseaux de neurones convolutifs (CNN pour *Convolutional Neural Network*).

L'élément de base d'un CNN est la couche convolutive [63]. Cette couche apprend et applique un ensemble de filtre à l'image d'entrée afin d'en extraire les caractéristiques. Chaque filtre glisse sur l'image et produit une carte de caractéristiques qui met en évidence la présence de certains motifs visuels dans l'image.

La sortie de la couche convolutive passe ensuite par une fonction d'activation non linéaire, telle que la fonction ReLU (unité linéaire rectifiée), qui introduit la non-linéarité dans le réseau et lui permet de modéliser des relations plus complexes

entre les caractéristiques [64].

Après plusieurs couches convolutives, le réseau comprend généralement une ou plusieurs couches de mise en commun (*pooling* en anglais). Les couches de mise en commun permettent de réduire la dimension spatiale des cartes de caractéristiques tout en conservant les caractéristiques importantes.

Diverses méthodes peuvent être utilisées dans les couches de mise en commun pour réduire l'échantillonnage des cartes de caractéristiques [65]. Une des méthodes consiste à ne garder que les valeurs maximales de chaque fenêtre. Une autre méthode consiste à ne garder que les valeurs moyennes de chaque fenêtre.

Les dernières couches d'un CNN sont généralement des couches entièrement connectées, qui mettent en correspondance la sortie des couches convolutives avec les classes ou les étiquettes de sortie. Ces couches sont entraînées à l'aide de techniques de rétropropagation standard afin d'optimiser les paramètres du réseau pour la tâche donnée.

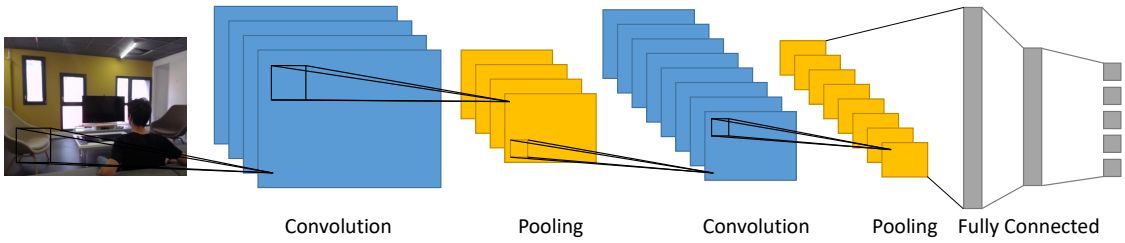

FIGURE 3.1 – Exemple d'un réseau de neurones convolutifs.

Les CNN permettent d'apprendre les caractéristiques spatiales des données d'entrée. Afin d'apprendre également les caractéristiques temporelles, il est nécessaire d'ajouter dans le réseau le traitement de la dimension temporelle.

Il existe plusieurs manières de prendre en compte cette dimension temporelle dans l'apprentissage. Un des premières solutions est d'ajouter une mémoire sur les caractéristiques spatiales apprises sur les données d'entrée successive. Cette approche est connue sous le nom de réseaux de neurones récurrents (*Recurrent Neural Network (RNN)*). Une autre solution est de mesurer les corrélations entre

les images successives dans le temps et dans l'espace. C'est l'approche apportée par la convolution 3D (3D-CNN). Enfin, il est possible de combiner différents flux de réseaux pour apprendre en parallèle les informations spatiales et temporelles avec des réseaux à deux flux (*Two-Stream*).

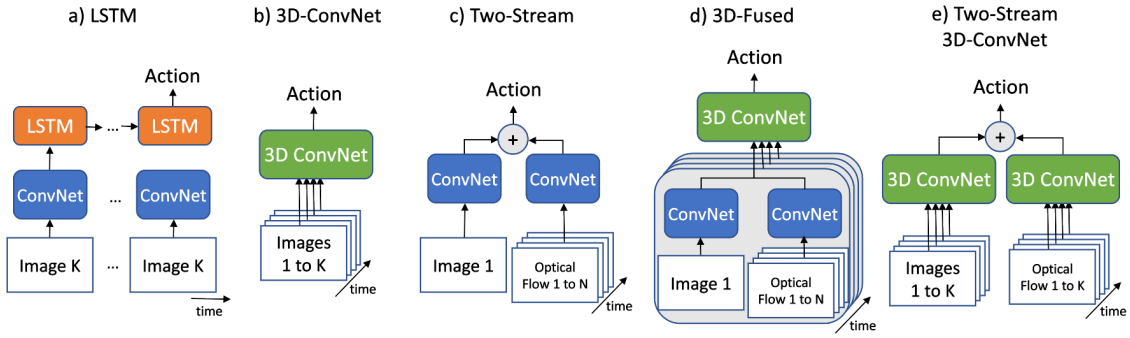

FIGURE 3.2 – Différentes architectures basées sur la convolution pour la reconnaissance d'actions humaines à partir de données vidéos. K est le nombre total d'images dans la vidéo et N représente un sous-ensemble d'images dans la vidéo.

3.1.1 Convolution 2D et RNN

La convolution 2D est une technique utilisée en traitement d'image pour extraire des caractéristiques ou des informations utiles à partir d'une image en appliquant de petites matrices de nombres, appelées filtres ou noyaux.

L'objectif des convolutions 2D et de transformer l'image d'entrée en carte de caractéristiques en fonction des valeurs du filtre via une fonction mathématique de convolution défini par :

$$G(m,n) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{3.1}$$

Où G est la carte de caractéristiques résultante de l'opération de convolution, I est l'image d'entrée ou une couche de caractéristiques, K est le filtre (ou noyau) utilisé pour effectuer la convolution, i et j sont les indices de la sortie de la convolution et m, n sont les indices de l'entrée.

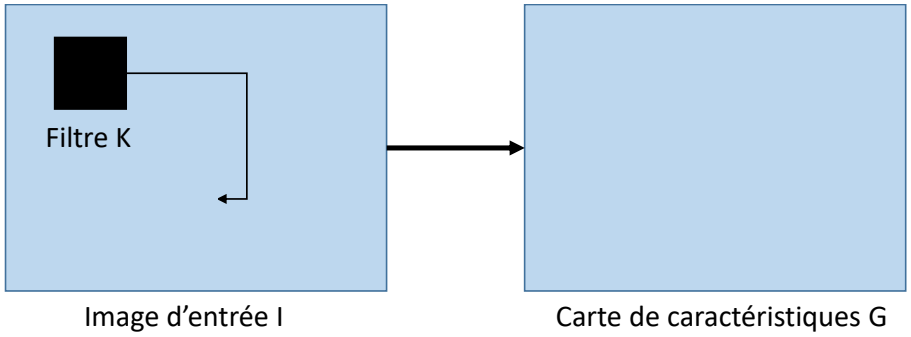

FIGURE 3.3 – Schéma d'une convolution 2D. Elle donne en sortie une carte de caractéristiques G obtenue par la convolution du filtre K sur l'entrée I.

Les filtres ou noyaux se déplacent sur l'image en utilisant une fenêtre glissante comme illustré sur la figure 3.3. A chaque pixel de l'image, l'opération de convolution est appliquée. Les résultats de ces opérations de convolution sont combinés pour créer une nouvelle image appelée carte de caractéristiques. Ces couches de convolution sont souvent combinées avec des couches de mise en commun pour réduire la dimension des caractéristiques, et avec des couches entièrement connectées (*Fully Connected*) pour la classification.

La convolution 2D sur les images est l'une des opérations de base dans les réseaux de neurones profonds. Il est donc facile d'utiliser la convolution 2D sur les séquences d'images pour en extraire les caractéristiques spatiales. Cependant, la convolution 2D ne modélise pas de manière inhérente les caractéristiques temporelles et nécessite une agrégation ou une modélisation supplémentaire de ces informations.

Pour pouvoir modéliser les caractéristiques à long terme, une couche récurrente telle qu'une cellule LSTM (Long Short-Time Memory) est ajoutée au modèle [59]. Cette architecture est illustrée sur la figure 3.2-a.

Le LSTM [66] est une cellule d'architecture de réseau neuronal récurrent (RNN) bien adaptée à la modélisation de données séquentielles, telles que les séries chronologiques ou le langage naturel. Les LSTM ont été inventés pour résoudre le problème de la disparition du gradient. Ce problème donne l'incapacité au réseau de propager les informations utiles du gradient d'un bout à l'autre du réseau. Le

modèle a alors du mal à apprendre à partir de longues séquences de données. Ce problème commun aux RNN traditionnels a été résolu par les LSTM. Les LSTM utilisent une cellule de mémoire spéciale qui peut maintenir son état sur de longues périodes de temps. Cette cellule de mémoire permet au réseau d'apprendre des dépendances à long terme dans les données.

Ainsi, la plupart des méthodes existantes ont adopté des architectures RNN à portes (Gated-RNN), telles que les mémoires à long et court terme comme AR-Net (LSTM) [67]. Cela permet de modéliser la dynamique temporelle à long terme des séquences vidéos.

Dans un réseau, chaque cellule LSTM possède trois composants principaux : une porte d'entrée, une porte de sortie et une porte d'oubli. Ces portes contrôlent le flux d'informations entrant et sortant de la cellule. Elles permettent au LSTM d'apprendre les informations à mémoriser ou d'oublier les informations inutiles. La porte d'entrée détermine les valeurs de la séquence d'entrée qui doivent être transmises à la cellule. La porte de sortie détermine les valeurs de la cellule qui doivent être transmises à la sortie. La porte d'oubli détermine les valeurs de la cellule qui doivent être réinitialisées ou écartées (figure 3.4). Les cellules LSTM fonctionnent avec les équations suivantes :

$$\begin{aligned} i_t &= \sigma(W_{ix} * x_t + W_{ih} * h_{t-1} + b_i) \\ f_t &= \sigma(W_{fx} * x_t + W_{fh} * h_{t-1} + b_f) \\ o_t &= \sigma(W_{ox} * x_t + W_{oh} * h_{t-1} + b_o) \\ g_t &= \tanh(W_{gx} * x_t + W_{gh} * h_{t-1} + b_g) \\ c_t &= f_t * c_{t-1} + i_t * g_t \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{3.2}$$

Où x_t est l'entrée à l'étape t.

h_{t-1} est la sortie cachée à l'étape $t-1$.

c_{t-1} est la mémoire cellulaire à l'étape $t-1$.

i_t, f_t, o_t et g_t sont respectivement les sorties des portes d'entrée, d'oubli, de sortie et de mémoire cellulaire.

W_{ix}, W_{fx}, W_{ox}, W_{gx}, W_{ih}, W_{fh}, W_{oh}, W_{gh} sont respectivement les poids cachés

pour les portes d'entrée, d'oubli, de sortie et de mémoire cellulaire.

b_i, b_f, b_o, b_g sont respectivement les biais pour les portes d'entrée, d'oubli, de sortie et de mémoire cellulaire.

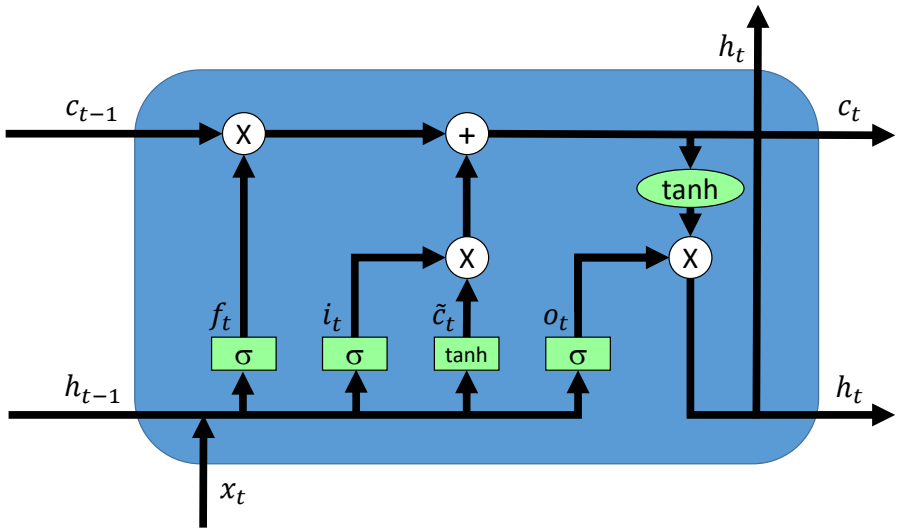

FIGURE 3.4 – Schéma d'une cellule LSTM. Elle prend au temps t une entrée x_t, la sortie cachée h_{t-1} de l'étape $t-1$ et la mémoire cellulaire c_{t-1} de l'étape $t-1$. Elle possède une porte d'entrée i_t, une porte d'oubli f_t, une porte de sortie o_t. Elle donne en sortie, la mémoire cellulaire c_t et la sortie cachée h_t qui seront données en entrée de la prochaine cellules LSTM.

Dans LRCN [59], les auteurs ont introduit le réseau convolutif récurrent à long terme (LRCN pour **Long-term Recurrent Convolutional**). Ce réseau contient un CNN 2D pour extraire les caractéristiques RGB au niveau de l'image. Ce CNN 2D est suivi de LSTM pour prendre en compte les caractéristiques temporelles dans la classification de l'action.

Dans [68] les caractéristiques RGB sont extraites au niveau de l'image et les caractéristiques de flux optique à partir d'un CNN 2D pré-entraîné. Ces caractéristiques sont ensuite transmises à un cadre LSTM empilé.

D'autres travaux comme [69] ont adopté le LSTM bi-directionnel, qui consiste en deux LSTM indépendants pour apprendre les informations temporelles en avant et en arrière, pour la reconnaissance d'actions humaines.

L'ajout de couches LSTM au-dessus de convolution 2D, permet au réseau d'apprendre les caractéristiques spatio-temporelles directement dans un environnement

de formation de bout en bout. Mais cela augmente les besoins en ressources de calculs.

Il est également plus complexe à paramétrer et à entraîner que les architectures plus simples.

3.1.2 Convolution 3D

De nombreuses recherches ont étendu les réseaux de convolution 2D à des réseaux de convolution 3D afin de modéliser simultanément les informations spatiales et temporelles dans les vidéos qui sont cruciales pour la reconnaissance d'actions humaines.

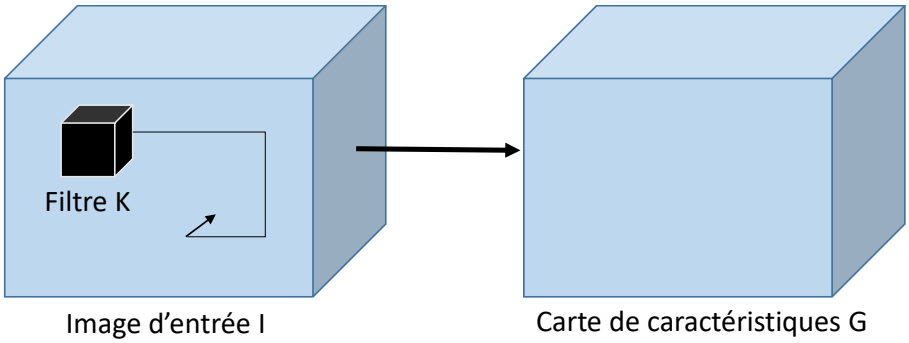

FIGURE 3.5 – Schéma d'une convolution 3D. Elle donne en sortie une carte de caractéristiques G obtenue par la convolution du filtre K sur l'image d'entrée I.

La convolution 3D est un type d'opération de convolution qui s'applique à des données d'entrée tridimensionnelles, telles qu'une image 3D ou une vidéo comme illustrée sur la figure 3.2-b. Comme la convolution 2D, la convolution 3D applique un ensemble de filtres sur les données d'entrée pour produire une série de cartes de caractéristiques. Les cartes de caractéristiques représentent les différentes informations des données d'entrée. La figure 3.5 illustre le fonctionnement des convolutions 3D. Les cartes de caractéristiques sont obtenues via la fonction mathématique de convolution 3D suivantes :

$$G(m,n,o)(I*K)(i,j,k) = \sum_{m}\sum_{n}\sum_{O} I(m,n,o) * G(i-m, j-n, k-o) \tag{3.3}$$

Où G est la carte de caractéristique résultante de l'opération de convolution. I est l'image d'entrée (ou une carte de caractéristique). K est le filtre (ou noyau) utilisé pour effectuer la convolution. i, j et k sont les indices de la sortie de la convolution. m, n, o sont les indices de l'entrée.

Dans C3D [70], un modèle CNN 3D a été introduit pour apprendre les caractéristiques spatio-temporelles des vidéos brutes dans un cadre d'apprentissage de bout en bout. Cependant, ces réseaux sont principalement utilisés pour l'apprentissage avec des clips au lieu d'apprendre à partir de vidéos complètes. Les dépendances spatio-temporelles à longue portée dans les vidéos sont donc ignorées. Par conséquent, plusieurs approches se sont concentrées sur la modélisation des dépendances spatio-temporelles à longue portée dans les vidéos.

Dans T3D [71], un modèle DenseNet [72] a été étendu avec des filtres 3D et des noyaux de mise en commun pour créer un CNN 3D temporel (T3D). La couche de transition temporelle peut alors modéliser des profondeurs de noyau de convolution temporelle variables. Le T3D peut capturer de manière dense et efficace l'apparence et les informations temporelles à court, moyen et long terme.

D'autres travaux ont combiné des CNN 2D et des CNN 3D. Par exemple, l'architecture ECO [73] utilise des CNN 2D pour extraire des caractéristiques spatiales, qui sont empilées puis transmises à des CNN 3D pour modéliser les dépendances à long terme.

Les convolutions 3D permettent aux réseaux neuronaux d'apprendre les caractéristiques spatiales et temporelles des données d'entrée. Cela leur permet de mieux comprendre le contexte et la signification des données 3D [70].

Les convolutions 3D sont plus difficiles à former que celles basées sur la convolution 2D en raison de la dimension supplémentaire qui complexifie les calculs et le paramétrage du réseau. Elles ont donc besoin de plus de puissance de calcul et de données que les convolutions 2D.

3.1.3 Réseaux convolutionnels à deux flux

A. Zisserman a permis la généralisation des réseaux convolutionnels à deux flux [61] [74].

Un réseau à deux flux pour la reconnaissance d'actions humaines est un type de modèle de reconnaissance d'actions humaines qui utilise deux flux de données pour reconnaître les actions humaines dans des vidéos. Le premier flux de données est généralement composé d'une image RGB, tandis que le deuxième flux de données est généralement une séquence d'informations temporelles associées à l'image RGB du premier flux. Cela peut être les mouvements des articulations du corps humain ou l'estimation de flux optique.

Dans ce type d'architecture, les deux flux apprennent les caractéristiques des données d'entrée. Le premier flux apprend les caractéristiques spatiales tandis que le deuxième flux apprend les données temporelles. Ces deux flux sont ensuite fusionnés dans les dernières couches de classification comme illustré sur la figure 3.2-c.

Il existe plusieurs méthodes de fusion des deux flux.

Dans [75], plusieurs stratégies de fusion ont été étudiées comme illustré sur la figure 3.6. On peut distinguer la fusion tardive, la fusion précoce et la fusion lente.

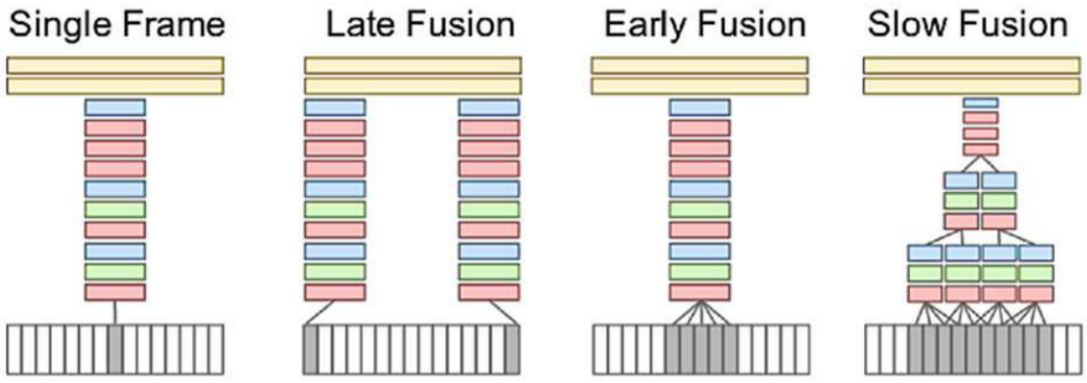

FIGURE 3.6 – Les différentes méthodes de fusion expérimentées dans [75]. Les cases rouges, vertes et bleues indiquent respectivement les couches de convolution, de normalisation et de pooling.

Pour déterminer la meilleure méthode de fusion de deux flux, les différentes méthodes de fusion ont été comparées à une architecture de référence nommée *single frame*. Cette architecture est essentiellement un réseau de classification d'images

sans caractéristiques temporelles. Elle réunit les prédictions sur les images successives pour en reconnaître les actions.

La méthode de fusion tardive permet d'extraire les caractéristiques de chaque modalité d'entrée séparément, puis elle les fusionne après la classification. Cette méthode est efficace mais ne permet pas de capturer les informations spatio-temporelles fines.

La fusion précoce fusionne les caractéristiques extraites de chaque image vidéo en un seul vecteur de caractéristiques avant de l'introduire dans la couche de classification. Cette méthode est coûteuse en termes de calcul, mais elle permet de capturer des informations spatio-temporelles fines.

Enfin, la fusion lente fusionne les caractéristiques de manière hiérarchique à partir d'une pile d'images, de sorte qu'au fur et à mesure que le réseau s'approfondit, davantage de caractéristiques temporelles sont apprises.

Les différentes expérimentations, réalisées dans [75], montrent qu'il est efficace de fusionner les réseaux spatiaux et temporels au niveau de la dernière couche de convolution. Cela réduit le nombre de paramètres tout en conservant la précision.

Bien que les architectures à deux flux aient été couronnées de succès dans de nombreuses applications, elle souffre de certaines limitations. Tout d'abord, elle est coûteuse en termes de calcul, ce qui la rend difficile à mettre à l'échelle pour les applications en temps réel. Deuxièmement, elle dépend de la disponibilité du flux optique, qui peut être difficile à estimer avec précision et inefficace en cas d'immobilité de la personne.

Pour remédier à ces limitations, de nouvelles architectures ont été proposées. Ainsi de nombreuses méthodes utilisent un CNN 3D pour exploiter les informations spatio-temporelles et un flux pour exploiter les informations de mouvement [76]. Le réseau à deux flux obtient une précision de pointe en utilisant des images RGB et de flux optique en entrée. Cependant, chaque flux est généralement entraîné individuellement ce qui augmente les besoins en ressources de calculs. Certaines approches tentent de construire un réseau à deux flux de manière plus efficace.

Dans D3D [77] est introduit le réseau *Distilled 3D Network (D3D)*. Il a obtenu de hautes performances sans calcul de flux optique pendant l'inférence. D3D combine les informations de mouvement du flux temporel dans le flux spatial, ce qui conduit le flux spatial à se comporter comme le flux temporel. D3D entraîne deux réseaux, dont un réseau enseignant et un réseau élève. Le réseau enseignant est un flux temporel appris d'un réseau à deux flux. Le réseau élève est un flux spatial. La connaissance du réseau de l'enseignant est distillée dans le réseau de l'étudiant pendant la phase d'entraînement.

Une amélioration des réseaux convolutionnels à deux flux est de les fusionner par un filtre 3D, comme illustré sur la figure 3.2-d. La convolution 3D permet d'apprendre les correspondances entre les caractéristiques spatiales et temporelles du flux. Elle permet un meilleur apprentissage des caractéristiques spatio-temporelles.

Dans [76], les convolutions 2D de l'architecture à double flux de [61] sont étendues en 3D. L'entrée du flux spatial est alors constituée des séquences d'images de la vidéo au lieu d'une seule image comme dans les architectures à deux flux de base. Cette approche a donné des résultats prometteurs, atteignant des performances de pointe sur plusieurs points de référence. Cependant, elle souffre encore de certaines limitations, telles que la nécessité d'une grande quantité de données pour un entraînement efficace et un coût de calcul élevé.

L'utilisation de deux flux de données permet de mieux comprendre les actions humaines dans des vidéos, en utilisant à la fois des informations visuelles et temporelles.

Ce type de réseau permet une modélisation temporelle à long terme et une augmentation des performances sans augmentation significative de la taille des paramètres.

3.2 ConvLSTM

Les méthodes de reconnaissance d'actions doivent apprendre des caractéristiques spatio-temporelles. Les convLSTMs proposés par Shi et al. dans [78] sont des cellules adaptées aux données spatio-temporelles. Les convLSTMs sont une

variante des cellules LSTM visibles dans la figure 3.7.

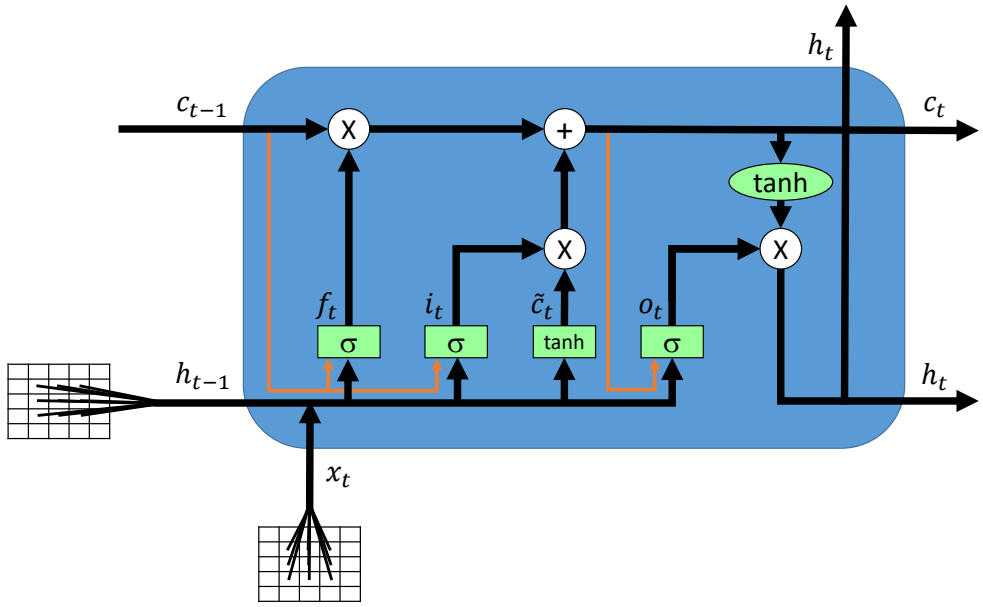

FIGURE 3.7 – Schéma d'une cellule de convLSTM. Elle prend au temps t une entrée x_t, la sortie cachée h_{t-1} de l'étape $t-1$ et la mémoire cellulaire c_{t-1} de l'étape $t-1$. Elle possède une porte d'entrée i_t, une porte d'oubli f_t, une porte de sortie o_t. Elle donne en sortie, la mémoire cellulaire c_t et la sortie cachée h_t qui seront données en entrée de la prochaine cellule.

Les équations clés des cellules convLSTM sont présentées ci-dessous, où "$*$" désigne l'opération de convolution et "$\circ$" désigne le produit Hadamard.

$$\begin{aligned} i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \\ H_t &= o_t \tanh(C_t) \end{aligned} \tag{3.4}$$

La cellule de convLSTM se distingue d'une cellule de LSTM par la présence d'une convolution à l'intérieur de la cellule LSTM. Dans une cellule de convLSTM, la multiplication matricielle entre les matrices de poids W et les entrées X est remplacée par une opération de convolution. Même changement pour la multiplication matricielle entre les matrices de poids W et l'état caché au pas de temps précédent H_{t-1}. Les convLSTM ont une structure convolutive dans les transitions

état-état et entrée-état.

Avec la convolution à l'intérieur du LSTM, l'entrée x, l'activation cellulaire c, la porte d'entrée i, les états cachés h, la porte d'oubli f et la porte de sortie o sont des tenseurs 3D dont la première dimension est le nombre de canaux. Les deux dernières dimensions sont les dimensions spatiales capables de retenir toutes les informations spatiales.

Les cellules convLSTM peuvent capturer les caractéristiques spatiales et temporelles en même temps. Elles sont donc efficaces pour les données spatio-temporelles telles que les vidéos.

3.3 Mécanismes d'attention

Le Transformer est un modèle de séquence à séquence basé sur les mécanismes d'attention. Les mécanismes d'attention [79] proviennent du domaine du traitement du langage naturel. Ils ont été conçus pour traiter des séquences entières. Les mécanismes d'attention peuvent paralléliser certaines opérations contrairement aux RNN, qui sont de nature séquentielle. Les mécanismes d'attention ont ensuite été appliqués au traitement des images [23] avant d'être utilisés dans le traitement des vidéos [62]. Comme les mécanismes d'attention sont conçus pour modéliser des séquences, il est nécessaire de transformer les vidéos d'entrée en un ensemble de séquences, où chaque élément est appelé un token. Ce processus s'appelle la tokenisation. Il existe différents types de tokenisation (fig. 3.8), la tokenisation des patchs 2D ou 3D, des images et des clips.

La plupart des architectures basées sur les mécanismes d'attention pour le traitement vidéo, utilisent une tokenisation de patch 2D. Elle consiste à diviser les images vidéos d'entrée en régions de taille fixe $h * w$. D'autres architectes utilisent plutôt une tokenisation de patch 3D. Elle permet de prendre en compte la dimension temporelle en divisant les vidéos d'entrée en régions de taille fixe $t * h * w$.

Dans le cas de la tokenisation par frame, des architectures de base (*backbones*) apprennent, pour chaque image, les caractéristiques spatiales locales. Cela permet aux mécanismes d'attention de se concentrer sur la modélisation des caractéristiques temporelles. La tokenisation par frame permet de modéliser des vidéos plus

longues.

Dans le cas de la tokenisation par clips, les informations de plusieurs images vidéos (ce qui forme un clip) sont condensées dans chaque token individuel pour réduire la dimension temporelle. Le mécanisme d'attention utilisera plus d'images pour couvrir des étendues temporelles plus longues. La tokenisation des clips est efficace pour les tâches de modélisation à long terme. Mais suivant la taille des clips, les informations à grain fin, comme l'emplacement des objets, peuvent être perdues ou mélangées.

La tokenisation a un impact sur le niveau auquel l'information est modélisée. La tokenisation basée sur les images ou les clips permet une modélisation temporelle plus longue. La tokenisation des patchs 2D ou 3D permet une modélisation spatio-temporelle plus fine. La tokenisation impacte la longueur de la séquence d'entrée et par conséquent, la complexité de la mémoire du modèle. La tokenisation par les patchs 2D ou 3D avec une conception efficace ou la tokenisation sur les images donnent les meilleurs compromis entre la performance et la complexité.

Les tokens sont ensuite intégrés. Le type d'intégration des tokens dépend du type de tokenisation. L'objectif de l'intégration des tokens est de pouvoir donner en entrée du réseau des tokens représentant de plus petites portions de la vidéo d'entrée sous forme de vecteur. Pour une tokenisation par patch 2D ou 3D, il est possible d'effectuer une projection linéaire. Pour une tokenisation par patch 2D ou par images, des convolutions 2D peuvent être effectuées. Pour une tokenisation par patch 3D ou par clip, des convolutions 3D peuvent être effectuées. Pour alléger l'apprentissage, des architectures de base (*backbones*) peuvent également être utilisés.

Les mécanismes d'attention, prenant en entrée des tokens, sont invariants par permutation. Pour pouvoir prendre en compte le biais structurel, c'est-à-dire l'emplacement de chaque token dans la vidéo d'origine, un codage des positions est ajouté aux tokens. Ce codage des positions permet de prendre en compte le positionnement des tokens dans la séquence d'entrée.

Il prend en entrée la séquence d'entrée $X \in \mathbb{R}^{T^{(X)} \times d_m}$ et la séquence de sortie ou la séquence sur laquelle l'attention est appliquée $M \in \mathbb{R}^{T^{(M)} \times d_m}$ où $T^{(\cdot)}$ est la taille de la séquence et d_m la dimension de chaque élément de la séquence, donc

de chaque token.

Dans chaque couche des mécanismes d'attention, X est mis en correspondance avec un ensemble de requêtes $Q \in \mathbb{R}^{T^{(X)} \times d_k}$ tandis que M est mis en correspondance avec un ensemble de paires de clés $K \in \mathbb{R}^{T^{(M)} \times d_k}$ et de valeurs $V \in \mathbb{R}^{T^{(M)} \times d_k}$, où $d_k = d_m/h$ et h est le nombre d'en-têtes. Ainsi, la requête Q représente une caractéristique intéressante, l'ensemble de paires de clés K représente les caractéristiques qui peuvent être pertinentes par rapport à la requête Q et la valeur V représente les caractéristiques d'origine.

Dans ce type de réseau, les cellules d'attention remplacent les cellules de convolution. Dans les problèmes de classification, l'attention permet d'apprendre les caractéristiques des classes. L'attention est définie comme une fonction à 3 variables qui sont Q, K et V.

Dans cette fonction, l'affinité dans l'opération non-locale est instanciée en tant que produit scalaire entre Q et K. Il en résulte une matrice qui est ensuite utilisée pour peser combien chaque valeur contribue à la représentation de sortie de chacune des autres valeurs. Ce produit scalaire permet de déterminer les pertinences de chaque valeur par rapport à une requête donnée et de pondérer en conséquence les interactions entre les valeurs. Pour cela, le produit scalaire est mis à l'échelle par un facteur de normalisation d_k correspondant à la dimension du vecteur k. Ce produit scalaire normalisé est converti en probabilités par la fonction softMax. Les probabilités résultantes sont ensuite utilisées pour mettre à jour les caractéristiques d'entrée V. Le modèle mathématique est le suivant :

$$Att(Q, K, V) = softMax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.5}$$

Similaire aux filtres multiples dans une couche convolutive, le *Multi-Head-Attention (MHA)* a été proposé dans *Attention Is All You Need* [79], où X et M sont mis en correspondance avec différentes représentations, c'est-à-dire différents ensembles de Q, K et V, afin d'effectuer différentes opérations d'attention simultanément. La sortie de chaque tête est concaténée et mappée dans un espace commun à d_m dimensions par une une transformation linéaire $W^{(O)} \in \mathbb{R}^{h.d_k \times d_m}$:

$$\begin{aligned} MHA(X,M) &= Concat(head_1, ...head_h)W^{(O)}, \\ \text{où } head_i &= Att(XW_i^{(Q)}, XW_i^{(K)}, XW_i^{(V)}), \\ et\ W_i^{(.)} &\in \mathbb{R}^{d_m \times d_k} \end{aligned} \tag{3.6}$$

Afin de réduire les temps d'entraînement et stabiliser l'apprentissage, une couche complète de Transformer est composée de deux sous-couches ou plus, suivies d'une connexion résiduelle et d'une couche de normalisation [80].

En pratique, chaque couche du Transformer contient au moins une sous-couche *Multi-Head-Attention (MHA)* et se termine par une sous-couche (*Feed-Forward Network*) finale chargée de transformer les représentations des données.

La sortie du Transformer est donnée à un perceptron multicouche (MLP) qui donne en sortie la classe d'action de la donnée d'entrée.

Bien que les mécanismes d'attention aient obtenu des résultats prometteurs, ils ont besoin de mémoire et de puissance de calcul.

Certains travaux ont permis de réduire la complexité des calculs comme dans SCT [81].

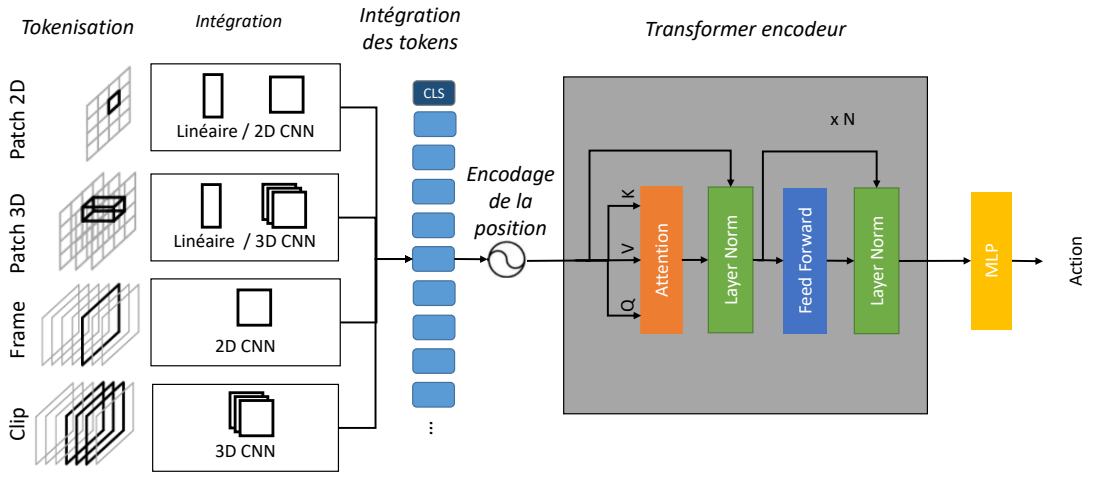

FIGURE 3.8 – Architecture basée sur les mécanismes d'attention pour la reconnaissance d'actions humaines.

3.4 L'apprentissage de la correspondance image-texte

L'apprentissage de la correspondance image-texte est un sujet de recherche qui fait le lien entre les domaines de la vision par ordinateur (CV) et le traitement du langage naturel (NLP). L'objectif est d'apprendre conjointement les représentations du texte et des images pour ensuite pouvoir décrire les images en langage naturel.

L'état de l'art actuel dans ce domaine est dominé par des modèles tels que CLIP (Contrastive Language-Image Pre-Training) [82] et XCLIP [83].

CLIP [82] se concentre sur l'apprentissage de modèles visuels en utilisant une supervision en langage naturel. Cette approche exploite des ensembles de données contenant des images et leurs descriptions textuelles associées. Grâce à l'apprentissage contrastif, le modèle apprend à associer des caractéristiques visuelles et textuelles similaires tout en distinguant les paires dissemblables. Lors de l'inférence, une nouvelle image est donnée au modèle. Le modèle va alors chercher la ou les descriptions textuelles les plus proches de l'image. Le modèle attribue ensuite à l'image la classe associée à la ou aux descriptions les plus proches.

En se basant sur la supervision du langage naturel, le modèle peut généraliser ses connaissances à de nouvelles tâches visuelles sans avoir besoin d'annotations explicites.

XCLIP [83] étend CLIP pour traiter des tâches de reconnaissance sur des vidéos. Il introduit des informations temporelles dans les modèles pré-entraînés afin de comprendre et reconnaître des actions, des événements et des dynamiques dans les vidéos. Les images vidéos sont transformées en un format séquentiel et encodées avec les descriptions textuelles correspondantes. Le modèle apprend ensuite à saisir les relations temporelles entre les images et le contexte linguistique.

ActionCLIP [84] propose un paradigme pour la reconnaissance d'actions humaines en intégrant l'apprentissage contrastif et le pré-apprentissage sur un ensemble de données vidéos à l'aide d'invités en langage naturel. Au lieu d'utiliser

des annotations explicites, il utilise des invités dérivés des sous-titres des vidéos pour guider l'apprentissage des représentations vidéos. ActionCLIP se concentre sur la capture des relations entre les segments vidéos et leurs descriptions textuelles associées, afin d'apprendre des représentations d'actions efficaces.

Ces modèles permettent de décrire les séquences d'images vidéos à l'aide du langage naturel. Ils peuvent être adaptés à différentes situations, car ils n'exigent pas une connaissance préalable de toutes les classes d'actions. C'est un algorithme qui permet ensuite de classifier les descriptions obtenues dans les différentes classes d'actions.

Cependant, ces modèles nécessitent des ressources de calcul importantes pour fonctionner correctement. Ils ouvrent la voie à diverses applications, de la compréhension d'images et de vidéos au traitement du langage naturel et à l'intelligence artificielle multimodale.

3.5 Analyse critique

Architecture	**Année**	**Modalités**	**Accuracy**
LRCN [59]	2015	RGB	82.90
C3D [70]	2015	RGB	85.20
Beyond Short-Snippets [68]	2015	RGB, flux optique	88.6
Two-stream [61]	2014	RGB, flux optique	88.00
3D-Fused [74]	2016	RGB, flux optique	92.50
T3D+TSN [71]	2017	RGB	93.2
ECO [73]	2018	RGB	93.30
I3D [76]	2017	RGB, flux optique	98.00
PERF-Net [85]	2020	RGB, squelette, flux optique	98.60
D3D [77]	2020	RGB	97.00
actionCLIP [84]	2021	RGB	97.10
DB-LSTM [69]	2021	RGB, flux optique	97.30
SCT-L [81]	2022	RGB	98.7

TABLE 3.1 – Résultats obtenus sur le jeu de données UCF101 pour chacune des architectures citées.

La reconnaissance d'actions humaines est un domaine de la vision par ordinateur qui se développe rapidement. Il existe de nombreuses architectures pour

la reconnaissance d'actions humaines. Les résultats obtenus sur le jeu de données UCF101 [86] par les méthodes citées précédemment dans cette section sont visibles dans le tableau 3.1.

D'une part, la méthode idéale et optimale serait de prendre en compte toutes les caractéristiques des différents prétraitements et des différentes architectures ensemble.

D'autre part, malgré plusieurs architectures de reconnaissance d'actions humaines, nous constatons que l'accent est mis sur l'analyse de l'image entière et sur la prise en compte du mouvement plus que sur la prise en compte de l'humain.

Il est important de noter que la meilleure architecture pour la reconnaissance d'actions humaines dépend de nombreux facteurs, tels que les données disponibles, les exigences en matière de précision et de temps de calcul, et les besoins spécifiques de l'application.

Dans le cadre des applications de HAR dans le contexte de l'AAD et de la robotique d'assistance, l'accent doit être mis sur la précision. Toutes les situations à risque doivent être détectées pour pouvoir lancer des alertes dans les temps. Cependant, aucune alerte ne doit être déclenchée en l'absence de situation à risque. Les temps de calcul et les ressources énergétiques doivent également être pris en compte pour que l'application puisse fonctionner en temps réel au sein d'un robot mobile d'assistance.

Malgré les nombreuses solutions de reconnaissance d'actions humaines existantes, les contraintes liées au contexte de l'AAD et de la robotique d'assistance sont très peu prises en compte dans la littérature.

Chapitre 4

Notre approche

Il existe une grande diversité de modalités et d'architectures pour la reconnaissance d'actions humaines. Le meilleur choix de la modalité d'entrée, couplée à une architecture, dépend de nombreux facteurs.

Dans notre contexte de la robotique d'assistance au maintien à domicile, il est important de prendre en compte les mouvements des personnes, les périodes d'immobilité des personnes, ainsi que les mouvements de caméra dus au déplacement du robot. Mais dans la littérature, les immobilités et les mouvements de caméra dans le cadre de la reconnaissance d'actions humaines sont peu pris en compte. Il est donc nécessaire de déterminer quel couplage entre une ou plusieurs modalités d'entrée et une architecture de reconnaissance d'actions humaines convient le mieux pour une application dans un robot d'assistance au maintien à domicile.

Pour cela nous allons expérimenter plusieurs couplages entre différentes modalités d'entrée et différentes architectures.

Dans ce chapitre, nous commencerons par vous présenter les différentes modalités d'entrée et les différentes architectures prises en compte dans nos expérimentations.

Enfin nous montrerons la problématique d'optimisation du couplage entre une ou plusieurs modalités et une architecture.

4.1 Modalités d'entrée

Il existe différentes modalités d'entrée pour la reconnaissance d'actions humaines, chacune ayant ses avantages et ses inconvénients.

L'objectif est d'expérimenter différentes modalités d'entrée, notamment la segmentation sémantique de la scène et la segmentation sémantique des parties du corps humain, qui sont peu utilisées pour la reconnaissance d'actions humaines. Nous souhaitons déterminer la modalité la plus adaptée à la reconnaissance d'actions humaines en présence d'immobilité et de mouvement de caméra, dans le contexte de la robotique d'assistance au maintien à domicile. Ainsi nous allons expérimenter le RGB, l'estimation de flux optique, la détection du squelette, la segmentation sémantique de la scène et la segmentation sémantique des parties du corps humain.

4.1.1 RGB

Le canal RGB capture l'ensemble des informations visuelles de la scène, incluant la couleur, la texture, la luminosité, etc. Cela permet de fournir une représentation visuellement riche des différentes scènes.

4.1.2 Flux optique

Le flux optique est une technique couramment utilisée dans l'analyse des mouvements dans les vidéos, y compris dans le contexte de la reconnaissance d'actions humaines. Cette technique repose sur le calcul des variations locales d'intensité lumineuse entre les images successives d'une séquence vidéo, ce qui permet de modéliser le mouvement. Seuls les mouvements dans la scène sont visibles.

Pour estimer le flux optique, nous utilisons la méthode algorithmique d'estimation de flux optique dense de Gunnar Farneback [32].

La méthode de Farneback utilise une approche pyramidale pour traiter les images à différentes résolutions. Cela implique la construction de pyramides d'images

en réduisant progressivement leur résolution à chaque niveau. Cette approche permet de gérer les mouvements de grands déplacements et les changements d'échelle dans les images, ce qui améliore la robustesse de l'estimation du flux optique.

Ensuite, pour estimer le flux optique dense, la méthode approche les fenêtres locales des images par des polynômes quadratiques par le biais de la transformation d'expansion polynomiale. Cette approximation par des polynômes quadratiques permet de capturer les variations locales du mouvement dans la fenêtre et facilite le calcul des déplacements des pixels.

Après avoir approché les fenêtres locales par des polynômes, la méthode observe comment ces polynômes se transforment sous l'effet de la translation (mouvement) entre les deux images successives. En analysant ces transformations, la méthode définit une approche pour estimer les champs de déplacement à partir des coefficients d'expansion polynomiale. Cela permet de calculer une estimation initiale du flux optique dense pour certains pixels.

Enfin, pour obtenir une estimation plus précise du flux optique dense, la méthode de Farneback effectue une série de raffinements itératifs. Ce processus permet d'améliorer progressivement l'estimation du déplacement en prenant en compte les informations des pixels voisins. En utilisant ce raffinement itératif, le flux optique dense final est calculé avec une précision accrue.

L'avantage de cette méthode est qu'elle est robuste aux variations de contraste et de luminosité entre deux images successives et qu'elle est efficace en temps réel.

4.1.3 Détection du squelette

La reconnaissance d'actions humaines basée sur la détection du squelette est une approche couramment utilisée dans la vision par ordinateur pour analyser et comprendre les mouvements humains dans une vidéo. Cette méthode repose sur l'idée que les informations les plus importantes pour comprendre une action se trouvent dans la structure du corps humain, représentée par un graphe.

Pour pouvoir détecter le squelette humain sur les séquences d'images, nous nous sommes appuyés sur le modèle *Bottom-Up Human Pose Estimation Via Disentan-*

gled Keypoint Regression appelé DEKR [87]. Ce modèle permet de détecter et de modéliser le squelette humain de toutes les personnes présentes dans la scène par une méthode ascendante. Elle commence donc par la détection des articulations individuelles, suivie de leur connexion pour former les poses complètes. La méthode introduit la régression démêlée des articulations, qui permet de prédire les coordonnées des articulations de manière indépendante, tout en prenant en compte les relations spatiales entre les articulations voisines.

La régression démêlée des articulations est réalisée en utilisant un réseau de neurones profonds qui apprend à prédire les coordonnées des articulations à partir des caractéristiques de l'image en entrée. Le réseau est formé de manière à minimiser une fonction de perte qui mesure l'écart entre les prédictions des articulations et leurs coordonnées réelles. En utilisant cette approche, les articulations peuvent être estimées avec précision, tout en résolvant le problème de la dépendance mutuelle entre les articulations. Cette solution contient également une méthode efficace pour la détection initiale des articulations, en utilisant des modèles de détection d'objets pré-entrainés. Ces modèles permettent d'obtenir rapidement une estimation approximative des articulations, qui est ensuite raffinée par la régression démêlée.

Cette méthode se distingue par sa rapidité en termes de temps de calcul avec une précision améliorée par rapport aux autres méthodes.

La détection du squelette se fait directement sur l'image RGB comme cela se fait dans la littérature.

4.1.4 Segmentation sémantique de la scène

La segmentation sémantique de la scène est une technique de vision par ordinateur qui vise à attribuer des étiquettes sémantiques à chaque pixel d'une image ou d'une vidéo. Elle permet de comprendre et de différencier les différentes régions ou objets présents dans la scène.

Dans le contexte de la reconnaissance d'actions humaines, la segmentation sémantique de la scène peut être utilisée pour enrichir l'analyse de la scène en fournissant des informations sur l'humain, l'environnement ou les objets avec lesquels les personnes interagissent.

Pour pouvoir pré-traiter nos données avec la segmentation sémantique de la scène, nous nous sommes appuyés sur le modèle existant de *Vision Transformer Adapter For Dense Predictions* [88].

Ce modèle est une adaptation de *Vision Transformer (ViT)* pour les prédictions denses comme la segmentation sémantique de la scène.

Cette approche utilise un ViT ordinaire [23] avec une tokenisation par patch. Ensuite, le ViT ordinaire est couplé avec un module d'antériorité spatiale conçu pour capturer les caractéristiques spatiales locales des images d'entrée, tout en maintenant l'architecture originale de ViT.

Ensuite, ces caractéristiques sont injectées dans le ViT pour y introduire les priorités spatiales. Enfin, un extracteur de caractéristiques multi-échelles permet d'extraire les caractéristiques hiérarchiques des données obtenues à partir du ViT. Les caractéristiques multi-échelles de granularité fine, nécessaires pour les prédictions denses, sont ainsi reconstruites.

Cette architecture permet de combler l'écart de performance entre le ViT ordinaire et les Transformers spécifiques à la vision pour les tâches de prédiction dense comme la segmentation sémantique de la scène.

4.1.5 Segmentation sémantique des parties du corps humain

La segmentation sémantique des parties du corps humain est une tâche spécifique de la segmentation sémantique qui vise à attribuer des étiquettes sémantiques à chaque pixel correspondant à une partie spécifique du corps humain, telle que la tête, les bras, les jambes, le torse, etc. Cette segmentation fine permet d'obtenir une représentation plus détaillée des mouvements et des interactions entre les différentes parties du corps.

Pour pouvoir pré-traiter nos données avec la segmentation sémantique des parties du corps humain, nous nous sommes appuyés sur le modèle existant de *Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation*

[1].

L'approche de *Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation* [1] utilise la complémentarité des données réelles et synthétiques, c'est-à-dire des variations riches et réalistes des données réelles et les étiquettes faciles à obtenir des données synthétiques, de manière efficace. Cette méthode permet d'apprendre la segmentation sémantique des parties du corps humain sur des images réelles sans aucune étiquette annotée par l'homme.

Pour cela, les informations complémentaires obtenues par deux domaines différents sont considérées : la détection du squelette humain sur les données réelles et synthétiques et la segmentation sémantique des parties du corps humain sur les données synthétiques pour améliorer les performances de segmentation.

L'approche proposée utilise une architecture de réseau à deux flux qui traite séparément les données synthétiques et réelles (voir la figure 4.1). Pour les deux flux, l'image d'entrée passe par une architecture de base (*backbones*) provenant de ResNet101 [89] pour extraire les caractéristiques. Puis les caractéristiques sont utilisées pour détecter le squelette humain. Le modèle de détection du squelette s'appuie sur OpenPose [46]. OpenPose est une approche ascendante basée sur des champs d'affinité (*Part Affinity fields (PAF)*) et des cartes de point-clés.

Le premier flux apprend la segmentation des parties du corps humain sur les données synthétiques en se basant sur la détection du squelette humain.

Le deuxième flux utilise les données réelles qui ne sont pas annotées pour la segmentation sémantique des parties du corps humain. Il détecte le squelette humain sur les données réelles et apprend à corréler la segmentation sémantique des parties du corps humain avec la détection de squelette humain. Pour cela, lors de l'apprentissage, les deux flux sont combinés à l'aide d'une méthode d'apprentissage complémentaire inter-domaine. Cette méthode utilise une fonction de perte complémentaire qui encourage les deux flux à apprendre des représentations complémentaires. Ainsi les paramètres du réseau et les informations apprises par l'un des flux sont partagés avec l'autre flux, ce qui leur permet d'apprendre mutuellement. Cette approche permet au réseau d'exploiter des informations complémentaires et d'améliorer la précision de la segmentation.

L'avantage de cette approche est qu'elle est robuste aux variations des données

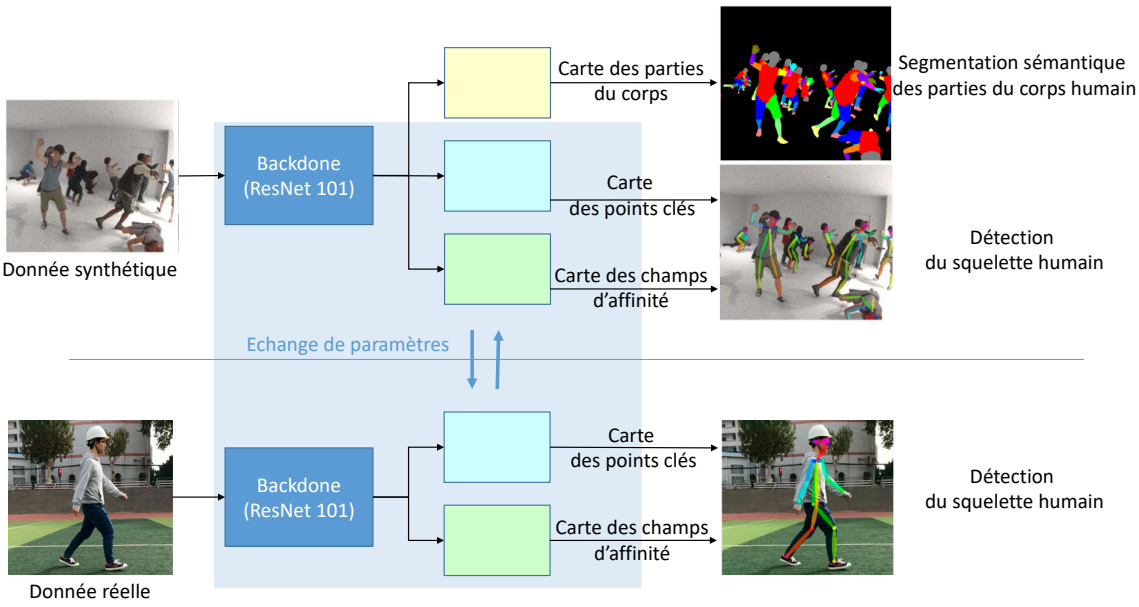

FIGURE 4.1 – Approche de CDCL [1] composée de deux flux. Le premier est l'apprentissage de la segmentation sémantique des parties du corps humain sur les données synthétiques basée sur la détection du squelette humain. Le deuxième est l'entraînement des données réelles via la détection du squelette humain. Les paramètres de l'architecture de base (*backbones*), des cartes des points clés et des cartes des champs d'affinité sont partagés avec le premier flux.

d'entrée telles que les changements de pose, d'éclairage et d'arrière-plan. Elle est aussi capable de segmenter avec précision les parties individuelles du corps dans des scènes complexes avec des occlusions et des parties du corps qui se chevauchent, ce qui constitue un défi pour les méthodes existantes.

4.2 Architectures

Notre objectif est de pouvoir comparer les différentes modalités d'entrée sur différents types d'architecture.

Pour cela, nous avons choisi d'expérimenter quatre architectures. La première architecture utilisée est basée sur les convLSTM et ne prend qu'une modalité en entrée. La seconde est basée sur les convLSTM et prend deux modalités en entrée. La troisième est basée sur les mécanismes d'attention et ne prend qu'une modalité en entrée. La dernière est basée sur la description texte des images et ne prend

qu'une modalité en entrée.

4.2.1 Architecture à un flux basée sur les convLSTM

Les cellules de convLSTM intègrent des capacités de mémoire à long terme en plus des convolutions. Elles sont capables de capturer à la fois l'information spatiale (via les opérations de convolution) et l'information temporelle (via les opérations de récurrence). L'utilisation de cellules de convLSTM permet donc de capturer les caractéristiques spatio-temporelles importantes des actions humaines.

Notre première architecture est donc basée sur les cellules de convLSTM. Elle comporte un seul flux qui est composé de cellules convLSTM comme illustré dans la figure 4.2. Elle ne prend en entrée qu'une seule modalité.

Cette architecture va nous permettre de comparer les modalités une à une et de tester les convLSTM pour la reconnaissance d'actions humaines.

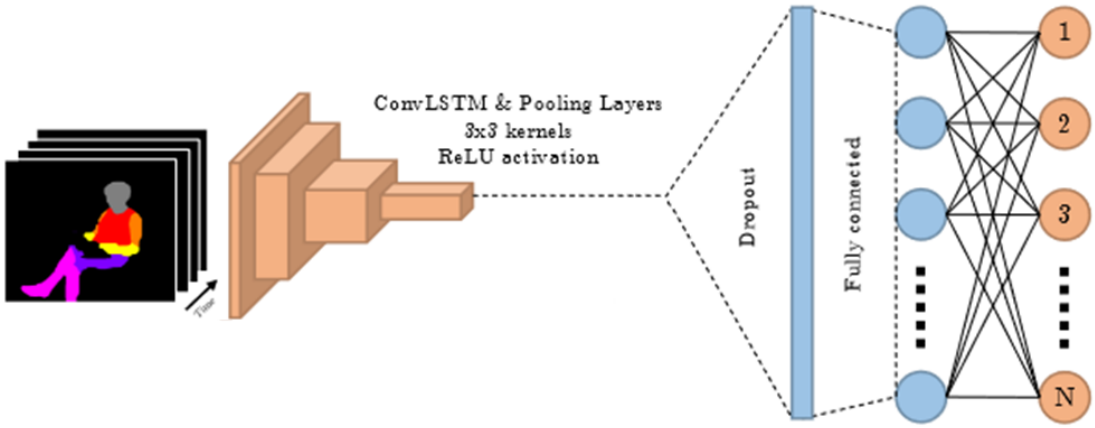

FIGURE 4.2 – Architecture à un flux pour la reconnaissance d'actions humaines basée sur les cellules convLSTM.

4.2.2 Architecture à deux flux basée sur les convLSTM

Les architectures à deux flux sont largement utilisées dans la littérature. L'un des deux flux apprend les données spatiales et l'autre les données temporelles.

Nous avons expérimenté comment l'ajout d'une nouvelle modalité pouvait modifier les résultats obtenus. Son architecture est présentée dans la figure 4.3.

Ainsi une seule image, souvent RGB, est transmise au premier flux. Ce flux permet ainsi d'apprendre les caractéristiques spatiales. L'ensemble des séquences d'images pré-traitées sont transmises au second flux. Il permet d'apprendre les caractéristiques temporelles. Chaque flux forme alors une prédiction et le score de classe est déterminé par leur fusion (figure 4.3).

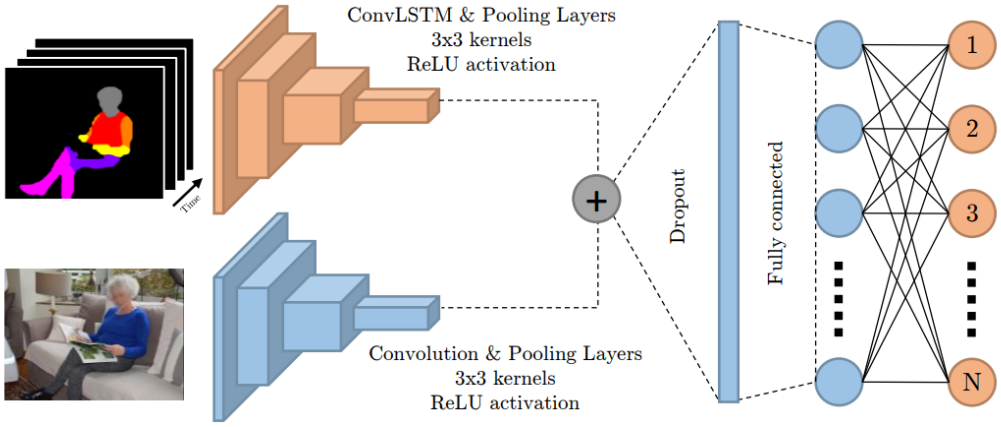

FIGURE 4.3 – Schéma de notre architecture à deux flux basée sur les convLSTM.

4.2.3 Architecture basée sur les mécanismes d'attention

Les mécanismes d'attention permettent au modèle de se concentrer sur les parties spécifiques de l'entrée qui sont les plus pertinentes pour la tâche de reconnaissance d'actions humaines. Cela permet de mettre l'accent sur les parties du corps humain les plus significatives pour l'action en cours, en ignorant les parties moins importantes ou perturbantes. Ainsi, l'attention permet de mieux représenter les caractéristiques discriminantes des actions.

Dans cette architecture, un modèle basé sur le mécanisme d'attention est utilisé

pour classifier les résultats de la modalité d'entrée. Des couches d'auto-attention à têtes multiples sont utilisées pour classifier la séquence de tokens obtenue par incorporation. Plusieurs stratégies d'incorporation sont disponibles, telles que l'échantillonnage de trame uniforme et l'incorporation de *tubelets*. La méthode *Tubelet embedding* permet de mettre en correspondance une séquence d'images avec une séquence de tokens qui alimente les couches de transformation. Il s'agit d'une extension de l'incorporation de *Vision Transformer (ViT)* [23] à la 3D. C'est une représentation intermédiaire linéaire qui conserve l'information spatio-temporelle. Alors que l'incorporation de ViT vise à remplacer les convolutions 2D, l'incorporation de *tubelets* vise à être l'équivalent des convolutions 3D. Un patch est défini par sa largeur, sa hauteur et par le nombre d'images. Chaque patch est projeté et transformé en un jeton. Ces volumes sont ensuite aplatis pour construire la séquence de tokens. La figure 4.4 explique ce processus de tokenisation. Les dimensions de *tubelets* plus petites entraînent un nombre croissant de tokens, ce qui augmente le temps de calcul. Lorsque les informations temporelles de différentes trames sont fusionnées par le Transformer dans le cas d'un échantillonnage de trame uniforme, cette information est maintenue dans le cas de l'intégration des *tubelets* en raison de la nature volumétrique des patchs.

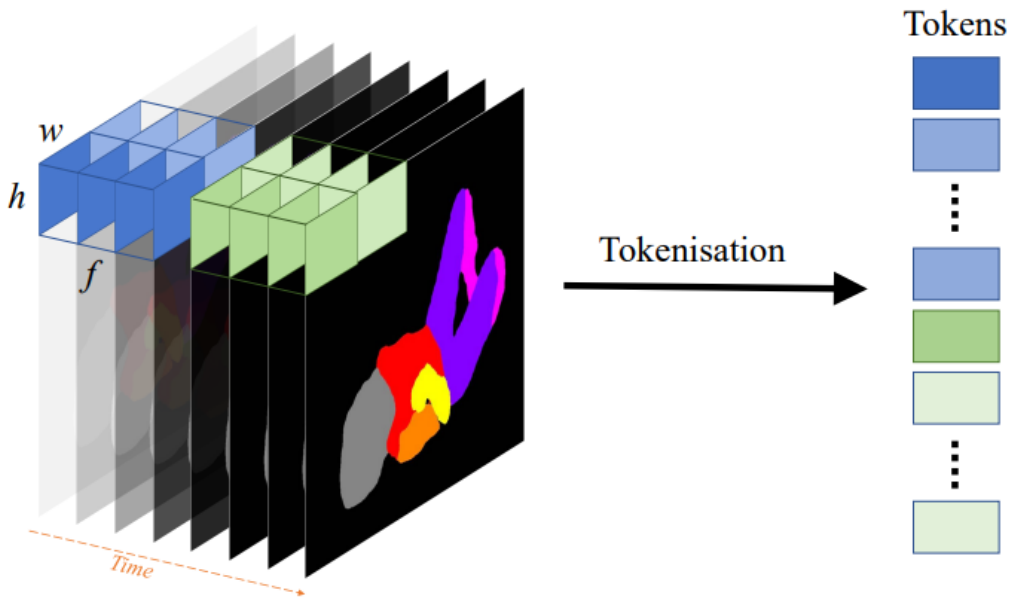

FIGURE 4.4 – Exemple de l'incorporation de *tubelets* dans une séquence d'images segmentées.

4.2.4 Architecture basée sur la correspondance image-texte

La correspondance image-texte nécessite un apprentissage permettant au modèle d'acquérir la capacité de faire correspondre les images à différents textes. Le modèle apprend à décrire les images.

Nous nous sommes appuyés sur le modèle CLIP [82]. CLIP est affiné sur un ensemble de données spécifiques, telles que Kinetics [76] ou UCF101[86], en utilisant une tête de classification linéaire pour prédire l'étiquette de l'action à partir des caractéristiques visuelles et textuelles. Les caractéristiques visuelles sont obtenues en faisant passer les images d'une vidéo à travers un réseau de neurones convolutifs pré-entraîné, tandis que les caractéristiques textuelles sont obtenues en transmettant l'étiquette de l'action sous forme de texte en langage naturel à l'encodeur linguistique du modèle CLIP. Le modèle génère ensuite une représentation conjointe des caractéristiques visuelles et textuelles, qui est utilisée pour prédire l'étiquette de l'action.

4.3 Problème d'optimisation

Notre recherche de la meilleure approche peut être vue comme un problème d'optimisation qui consiste à trouver le meilleur couple entre une ou plusieurs modalités d'entrée et une architecture de reconnaissance d'actions humaines.

Ce problème d'optimisation peut être modélisé de la manière suivante :

Soit M l'ensemble des modalités d'entrée, et A l'ensemble des architectures de reconnaissance d'actions humaines. Chaque architecture a dans A est associée à un sous-ensemble $M_a \subseteq M$.

Afin de maintenir une solution finale suffisamment légère en termes de ressources de calcul pour être utilisée dans un robot d'assistance, chaque sous-ensemble M_a de M contient une ou deux modalités. Dans le cas où un sous-ensemble contient deux modalités, l'une des deux modalités sera le RGB, afin d'éviter d'effectuer deux prétraitements simultanément.

Définissons la fonction objectif $f : M \times A \to \mathbb{R}$ qui évalue les performances du système de reconnaissance d'action pour chaque couple (m, a) où m appartient

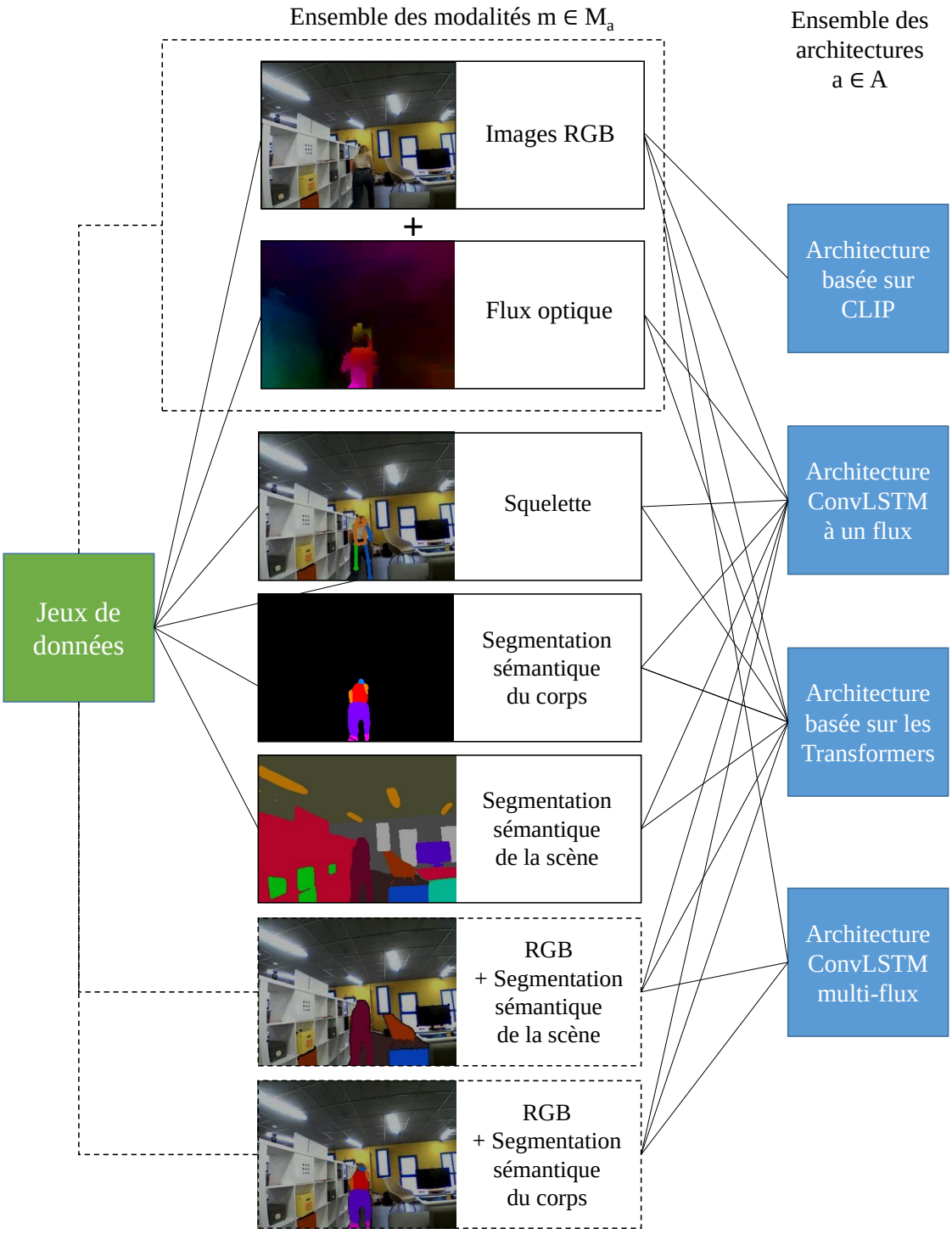

FIGURE 4.5 – L'ensemble des couples (m, a) pris en compte pour un jeu de données.

à M_a pour une architecture a donnée. Cette fonction objectif attribue le score d'exactitude, représentant la précision obtenue avec le couple de modalité d'entrée m et d'architecture de reconnaissance d'actions humaines a.

Le problème d'optimisation consiste à trouver les couples (m^*, a^*) tels que m^* appartienne à M_{a^*} pour chaque a^* dans A, maximisant la fonction objectif f : $(m^*, a^*) = \arg\max_{m \in M_a, a \in A} f(m, a)$

L'ensemble des couples (m, a) est visible dans la figure 4.5.

L'architecture Clip est basée sur l'apprentissage de la description textuelle des images d'entrée. Cette description textuelle des images est entraînée uniquement à partir d'images RGB. Par conséquent, notre architecture Clip ne peut accepter que des images RGB en entrée. L'architecture à deux flux, basée sur les convLSTM, prend forcément deux modalités différentes en entrée. L'une de ces modalités est le RGB, afin de ne pas alourdir le modèle final. Le squelette étant modélisé directement sur le RGB, l'ajout du RGB n'aura que peu d'incidence sur les résultats obtenus. Il ne sera donc pas pris en compte pour l'architecture à deux flux. Pour l'architecture à un flux, basée sur les convLSTM, et celle basée sur les mécanismes d'attention, toutes les modalités seront expérimentées.

4.4 Synthèse

La reconnaissance d'actions humaines implique une variété de modalités et d'architectures et le choix optimal dépend de plusieurs facteurs. Dans le contexte de la robotique d'assistance au maintien à domicile, il est essentiel de considérer les mouvements des personnes, les périodes d'immobilité, ainsi que les mouvements de la caméra provoqués par le déplacement du robot. Cependant, la littérature accorde peu d'attention aux immobilisations et aux mouvements de caméra.

Par conséquent, il est nécessaire de déterminer le meilleur couplage entre les modalités d'entrée et les architectures de reconnaissance d'actions humaines pour une application dans un robot d'assistance au maintien à domicile.

Cette démarche s'apparente à un problème d'optimisation, où chaque architecture est associée à un sous-ensemble de modalités d'entrée, et l'objectif est de

maximiser la précision du système de reconnaissance d'actions. Cette optimisation repose sur une fonction d'objectif évaluant les performances de chaque combinaison modalité-architecture en termes de précision.

Dans ce processus, nous prenons en considération les spécificités des différentes architectures, telles que la nécessité d'utiliser des images RGB pour certaines d'entre elles. Notre démarche expérimentale nous permettra de déterminer les meilleures associations modalité-architecture, contribuant ainsi à la création d'une solution robuste et efficace pour la robotique d'assistance au maintien à domicile.

Chapitre 5

Nouveau jeu de données JARD

Les méthodes actuelles de reconnaissance d'actions humaines sont basées sur l'apprentissage profond. Cependant l'entraînement d'un modèle d'apprentissage profond nécessite un grand nombre de données. Il existe dans la littérature un grand nombre de jeux de données non contrôlés pour la reconnaissance d'actions humaines.

Dans nos recherches, nous avons distingué deux principales catégories de jeux de données. La première concerne les jeux de données génériques. La seconde concerne les jeux de données dédiés aux anomalies et aux situations anormales, telles que les chutes.

Dans ce chapitre nous allons présenter différents jeux de données de la littérature. Nous allons commencer par présenter les jeux de données génériques conçus pour faire de la reconnaissance d'actions humaines génériques. Puis nous allons présenter les jeux de données dédiés aux situations anormales (détection d'événements anormaux et détection de chutes). Après une analyse des jeux de données existants, nous présenterons notre nouveau jeu de données pour la reconnaissance d'actions humaines et la détection d'événements anormaux dans le contexte de la robotique d'assistance au maintien à domicile.

5.1 Jeux de données génériques

Ci-dessous nous décrivons les jeux de données vidéos non contrôlés pour la reconnaissance d'actions humaines présents dans le tableau 5.1.

Jeux de données	**Année**	**Nb de vidéos**	**Nb d'actions**	**Nb de citations**	**Types d'actions**
Sports-1M [75]	2014	1,133,158	487	6,037	Sport
UCF101 [86]	2012	13,320	101	3,716	Sport, interaction humaine, pratique d'un instrument, soin personnel, action quotidienne
HMDB51 [90]	2011	6.766	51	2,883	Sport, interaction humaine, activité quotidienne
Kinetics [76]	2017	650,000	700	1,709	Sport, interaction humaine, pratique d'un instrument, soin personnel, activité quotidienne
Activitynet [91]	2015	28,000	200	1,428	Sport, pratique d'un instrument, soin personnel, action quotidienne
Hollywood2 [92]	2009	3,669	12	806	Action quotidienne, interaction humaine
20BN-something-something [93]	2017	108,499	174	780	interaction humaine
Youtube-8M [94]	2019	8,000,000	4,800	650	
HowTo100M [95]	2019	136,000,000	23,000	550	sport, soin personnel, activité quotidienne
Charades [96]	2016	9,848	157	313	activité quotidienne
HACS (SLAC) [97]	2017	520,000	200	164	sport, pratique d'un instrument, activité quoitidienne

TABLE 5.1 – Liste des principaux jeux de données vidéos non contrôlés pour la reconnaissance d'actions humaines.

Sport-1M : Le jeu de données Sport-1M [75] contient 1,133,158 de vidéos réparties dans 487 classes d'actions. Les données vidéos de l'ensemble de Sport-1M ont été collectées sur YouTube. Il couvre les actions liées aux sports. L'annotation a été faite via l'API des sujets YouTube. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets et dans les conditions d'éclairage.

UCF-101 : Le jeu de données UCF-101 [86] contient 13,320 vidéos réparties dans 101 classes d'actions. Les données vidéos de l'ensemble de UCF101 ont été collectées sur YouTube. Il couvre des actions liées aux sports, aux interactions humaines, à la pratique d'un instrument de musique, aux soins personnels et aux actions quotidiennes telles que des actions liées à la cuisine. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets et dans les conditions d'éclairage.

HMDB : Le jeu de données HMDB [90] contient 6,766 vidéos réparties dans 51 classes d'actions. Les vidéos proviennent de différentes sources comme les films ou les plateformes publiques tels que YouTube. Les données vidéos de HMDB51 contiennent, en plus de l'étiquette de la catégorie d'action, une méta-étiquette décrivant les propriétés du clip comme les parties visibles du corps, le mouvement de la caméra, le point de vue de la caméra, le nombre de personnes impliquées dans l'action et la qualité de la vidéo. Il couvre des actions liées aux sports, aux interactions humaines, et aux actions quotidiennes telles que manger, boire et fumer.

Kinetics : Le jeu de données kinetics [76] contient 650,000 vidéos réparties dans 700 classes d'actions. Les données vidéos ont été collectées sur YouTube. Il couvre des actions liées aux sports, aux interactions humaines, à la pratique d'un instrument de musique, aux soins personnels et aux actions quotidiennes. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets, leur échelle et dans les conditions d'éclairage.

ActivityNet : Le jeu de données ActivityNet [91] contient 28,000 vidéos réparties dans 200 classes d'actions. Les données vidéos de l'ensemble de ActivityNet ont été collectées sur YouTube. Il couvre des actions liées aux sports, aux soins personnels, aux activités ménagères, à la pratique d'un instrument de musique et aux actions quotidiennes. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets, leur échelle et dans les conditions d'éclairage.

Hollywood2 : Le jeu de données Hoolywood2 [92] contient 3,669 vidéos réparties dans 12 classes d'actions dans 10 scènes différentes. Les données vidéos ont été collectées à partir de films hollywoodiens. Les arrière-plans des vidéos sont encombrés et les vidéos peuvent contenir des mouvements de caméra et des plans de coupe.

20BN-something-something : Le jeu de données 20BN-something-something [93] contient 108,499 vidéos réparties dans 174 classes d'actions. Les données vidéos se concentrent sur l'interaction humaine avec des objets quotidiens. Les vidéos sont à la première personne.

Youtube-8M : Il s'agit d'un ensemble de données pour la classification générale des vidéos multi-labels dont la classification des actions [94]. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets, leur échelle et leurs conditions d'éclairage.

HowTo100M : Le jeu de données HowTo100M [95] contient 136M vidéos réparties dans 23k classes d'actions. Les données vidéos de l'ensemble de HowTo100M ont été collectées sur YouTube. Il couvre des actions liées aux sports, aux soins personnels, aux activités ménagères et aux activités quotidiennes. Les vidéos contiennent des mouvements de caméra et différents points de vue, des arrière-plans encombrés et des variations dans l'apparence des objets, leur échelle et dans les conditions d'éclairage.

Charades : Le jeu de données Charades [96] contient 9,848 vidéos réparties dans 157 classes d'actions. Les données vidéos ont été collectées par Amazon mechanical turk. Il couvre les actions liées aux activités quotidiennes à l'intérieur telles que la cuisine ou le ménage.

HACS (SLAC) : Le jeu de données HACS [97] contient 520,000 vidéos réparties dans 200 classes d'actions. HACS, qui est une extension au jeu de données SLAC, permet la reconnaissance et la localisation temporelle des actions. Cet ensemble de données couvre des actions liées aux sports, à la pratique d'un instrument

de musique et aux activités ménagères.

5.2 Limitation des jeux de données génériques

La plupart des vidéos de ces jeux de données de reconnaissance d'actions humaines non contrôlés proviennent de vidéos collectées sur internet ou sur des films. Malgré la grande quantité de données existantes pour la reconnaissance des actions humaines, des données manquent pour des actions et des situations importantes correspondant à des événements anormaux rares tels que les chutes, les malaises et les postures anormales et dangereuses.

Pour pouvoir faire de la détection d'événements anormaux, des jeux de données ont été créés.

5.3 Jeux de données dédiés aux événements anormaux

Il existe très peu de jeux de données permettant de reconnaître des situations anormales comme les chutes ou les malaises. Certains jeux de données de la littérature traitent des situations anormales tandis que d'autres jeux de données se concentrent exclusivement sur la détection de chutes.

Dans cette section nous décrivons les principaux jeux de données dédiés aux situations anormales. La première partie détaille certains jeux de données pour la détection de situations anormales comme les événements ou les comportements anormaux. La deuxième partie décrit certains jeux de données pour la détection de chutes.

5.3.1 Situations anormales

La détection de situations anormales est un domaine qui concerne l'identification d'événements ou de comportements qui sont différents de ce qui est considéré

comme normal ou habituel dans un système ou un environnement donné.

Dans de nombreux domaines, tels que la sécurité, la santé et la surveillance de l'environnement, il est important de pouvoir détecter rapidement les situations anormales afin de pouvoir prendre des mesures appropriées pour prévenir des problèmes plus graves comme déclencher une alarme et une intervention humaine immédiate.

Il existe différents jeux de données pour la détection de situation anormale. Nous allons détailler ceux présents dans le tableau 5.2.

Jeux de données	**Année**	**Nb de vidéos**	**Nb Actions**	**Nb citations**
UCF-Crime [98]	2018	1900	13 anomalies	1110
Avenue [99]	2013	37	-	1036
UCSD Anomaly Detection [100]	2013	98	-	9

TABLE 5.2 – Liste des principaux jeux de données vidéos pour la détection de situations anormales.

UCF-Crime : Le jeu de données UCF-Crime [98] est composé de 1900 vidéos de surveillance du monde réel, longues et non tronquées, avec 13 anomalies réalistes, dont les suivantes : abus, arrestation, incendie criminel, agression, accident de la route, cambriolage, explosion, bagarre, vol, fusillade, vol à l'étalage et vandalisme.

Avenue : Le jeu de données Avenue [99] contient des vidéos capturées dans l'avenue du campus CUHK. Il a été créé dans le but de détecter des événements anormaux comme une action étrange (exemple : courir), un objet anormal dans la scène (exemple : vélo) ou une personne marchant dans une direction inattendue.

UCSD Anomaly Detection : Le jeu de données *UCSD Anomaly Detection* [100] a été créé dans le but de détecter la circulation d'entités non piétonnes dans des allées piétonnes et des schémas anormaux de mouvement des piétons. Les anomalies les plus courantes sont les cyclistes, les patineurs, les petits chariots et les personnes marchant sur une allée ou sur l'herbe qui l'entoure. Quelques cas de personnes en fauteuil roulant ont également été enregistrés.

5.3.2 Jeux de données pour la détection de chutes

La détection de chutes est un sous-domaine de la reconnaissance d'actions humaines. Elle vise à détecter les chutes en temps réel. Cette détection permet alors de lancer des alertes afin de réduire le temps nécessaire au patient pour recevoir des soins médicaux. Cela permet de réduire les conséquences négatives liées aux chutes [101].

Pour pouvoir détecter les chutes, des jeux de données pour la détection de chutes existent dans la littérature. Nous allons détailler ceux présents dans le tableau 5.3.

Jeux de données	**Année**	**Nb de vidéos**	**Nb d'actions**	**Nb de citations**
UR Fall Detection [102]	2014	70	-	467
SDU Fall [103]	2014	-	6	275
UP-Fall [104]	2019	33	11	196
Le2i [105]	2012	250	9	123
Multiple cameras fall [106]	2011	24 scénarios*8 caméras	10	156
HQFSD [107]	2016	72	37	43
CAUCAFALL [108]	2022	100	10	-

TABLE 5.3 – Liste des principaux jeux de données vidéos pour la détection de chutes.

UR Fall Detection : Le jeu de données *UR Fall Detection* [102] contient 70 vidéos. 40 vidéos contiennent des activités de la vie quotidienne et 30 contiennent des chutes. Les chutes sont enregistrées avec deux caméras Microsoft Kinect (RGB et profondeur) et les données accélérométriques correspondantes. Les activités de la vie quotidienne sont enregistrées avec une seule caméra et un accéléromètre. Des exemples sont donnés dans la figure 5.1-a.

SDU Fall : Le jeu de données *SDU Fall* [103] contient 6 actions (tomber, se pencher, s'accroupir, s'asseoir, se coucher et marcher) effectuées par dix sujets. Chaque sujet effectue 30 fois les six actions. A chaque fois, les conditions suivantes sont changées aléatoirement : porter ou ne pas porter un gros objet, allumer ou éteindre la lumière, changer la disposition de la pièce, changer de direction et de position par rapport à la caméra.

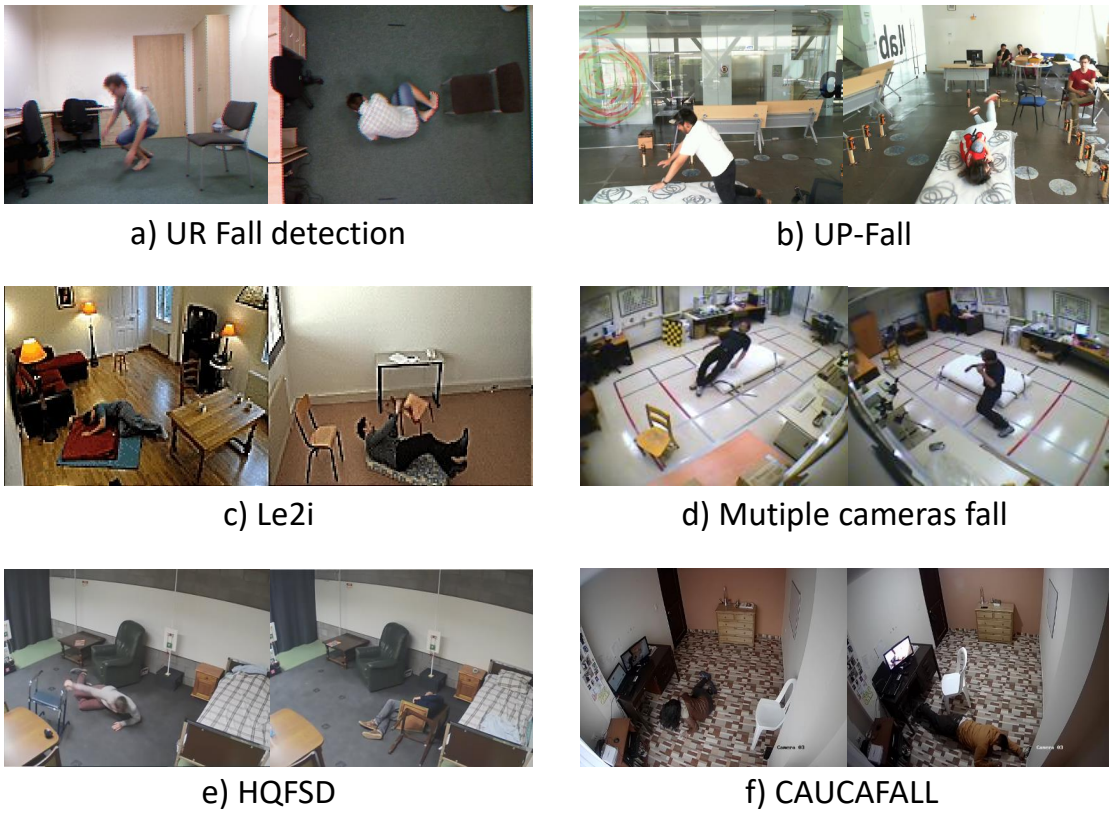

FIGURE 5.1 – Exemples de chutes provenant des jeux de données de la littérature pour la détection de chutes.

UP-Fall : Le jeu de données UP-Fall Detection [104] comprend 11 activités et 3 vidéos par activité. Les 17 sujets ont effectué six activités humaines quotidiennes simples (marcher, se tenir debout, ramasser un objet, s'asseoir, sauter et s'allonger) ainsi que cinq types de chutes (tomber en avant avec les mains, tomber en avant avec les genoux, tomber en arrière, tomber assis sur une chaise vide et tomber sur le côté). Ces données ont été recueillies à l'aide d'une approche multimodale, c'est-à-dire des capteurs portables, des capteurs ambiants et des dispositifs de vision. Des exemples sont présentées dans la figure 5.1-b.

Le2i : Le jeu de données Le2i [105] contient 250 séquences vidéos dans 4 endroits différents (une maison, un café, un bureau et une salle de lecture). 192 vidéos contiennent des chutes (chutes vers l'avant, chutes en position assise inappropriée, perte d'équilibre) et 57 vidéos contiennent plusieurs activités normales (marcher

dans différentes directions, s'asseoir, se lever, s'accroupir, faire le ménage, déplacer une chaise). Il comprend également des variations afin de fournir des exemples des principaux problèmes : variations d'illumination, occlusions, arrière-plans encombrés et texturés. Dans la figure 5.1-c, des exemples de Le2i sont fournis.

Multiple cameras fall : Le jeu de données "Multiple camera fall" [106] contient 24 scénarios enregistrés par 8 caméras. Les scénarios sont réalisés par un seul sujet. Les 22 premiers scénarios contiennent une chute et des activités quotidiennes normales comme marcher ou faire le ménage et des activités pouvant être confondues avec une chute comme s'asseoir ou s'accroupir. Les 2 derniers ne contiennent que des activités quotidiennes normales pouvant ou non être confondues avec une chute. Les vidéos contiennent des occultations ou des objets en mouvement. La figure 5.1-d comporte des exemples de ce jeu de données.

HQFSD : Le jeu de données HQFSD, pour *High quality fall simulation data* [107], contient 72 vidéos avec 10 sujets différents. 55 vidéos contiennent des chutes et 17 vidéos contiennent des actions de la vie quotidienne comme marcher avec et sans aide à la marche, s'asseoir et se lever, manger et boire, se mettre au lit et en sortir, dormir, changer de vêtements, enlever et mettre des chaussures, lire, se transférer du fauteuil roulant au fauteuil et vice versa, faire le lit, tousser et éternuer violemment, ramasser quelque chose sur le sol. Dans les scénarios de chutes, des défis supplémentaires ont été incorporés, tels que des occlusions, une chute partielle ou totale hors du champ de vision de la caméra, et dans un scénario, plus d'une personne se trouvait dans le champ de vision de la caméra. Dans le cadre des activités normales, les mêmes défis ont été incorporés, ainsi que des changements d'éclairage. Des exemples sont présentés dans la figure 5.1-e.

CAUCAFALL : Le jeu de données CAUCAFALL [108] contient les vidéos de 10 sujets effectuant 5 activités quotidiennes qui sont marcher, sauter, ramasser un objet, s'asseoir et s'agenouiller et 5 types de chutes différentes qui sont les chutes vers l'avant, les chutes vers l'arrière, les chutes latérales à gauche, les chutes latérales à droite et les chutes survenant en position assise. Les vidéos contiennent des occultations ou des objets en mouvement, des changements d'éclairage (naturel, artificiel et nocturne). Vous pouvez voir des exemples de CAUCAFALL dans la

figure 5.1-f.

5.4 Analyse critique

L'ensemble des jeux de données diffèrent les uns des autres par le nombre de vidéos et de classes d'actions, le nombre de sujets dans les vidéos, le fond des vidéos, l'apparence et les variations des actions, le mouvement de la caméra, la qualité de la vidéo, etc.

Les jeux de données non contrôlés pour la reconnaissance d'actions humaines restent limités en taille et en diversité. De nombreux jeux de données utilisés dans la littérature traitent d'actions similaires. Par exemple, de nombreux jeux de données traitent d'actions liées au sport, à l'interaction entre plusieurs personnes ou à l'interaction d'une personne avec un objet. Peu de ces jeux de données traitent de situations anormales liées au contexte de l'assistance à l'autonomie à domicile.

Des jeux de données pour la détection de situations anormales ont été créés ces dernières années. Ces situations anormales sont souvent des situations rencontrées en extérieur ou dans des lieux publics et non dans le cadre d'une personne à son domicile. D'autres jeux de données ont été créés pour faire de la détection de chutes. Ces jeux de données contiennent en général des actions de la vie quotidienne et des actions de chutes. Toutes les actions sont basées sur le mouvement. De plus, ces jeux de données conçus pour la détection de chutes ne prennent en compte que les chutes comme situations anormales. La chute doit donc pouvoir être captée par la caméra. Lors de l'utilisation d'un robot d'assistance, il se peut que le robot arrive après la chute et n'ait pas pu voir la personne chuter. Peu de jeux de données prennent en compte l'immobilité de la personne due à certaines situations comme une personne inconsciente étendue à terre dont la chute n'a pas pu être captée par la caméra.

Sur plus de 150 millions de vidéos, correspondant à plus d'une centaine de milliers d'actions, il existe peu de données adaptées pour former des modèles per-

tinents dans le domaine de la robotique d'assistance au maintien à domicile. Certaines situations spécifiques, liées à l'immobilité de la personne et aux mouvements de la caméra dus au déplacement du robot, sont particulièrement sous-représentées dans les jeux de données disponibles. Cela souligne l'importance de développer des jeux de données adaptés à ce contexte spécifique de la robotique d'assistance à domicile.

5.5 Création du jeu de données JARD

Pour surmonter le défi du manque de données en matière de reconnaissance de situations anormales dans le domaine de la robotique d'assistance au maintien à domicile, nous avons créé un nouveau jeu de données pour entraîner, tester et valider des modèles de reconnaissance de situations normales et anormales dans le contexte de la robotique d'assistance au maintien à domicile, nommé JARD.

5.5.1 Paramètre de collecte des données

Ce jeu de données est conçu pour la reconnaissance d'action dans un contexte de robotique d'assistance au maintien à domicile, où l'utilisation se fait principalement au sein d'une habitation.

Ce nouveau jeu de données a été enregistré dans un appartement expérimental avec différentes pièces (cuisine, salon, bureau, chambre, salle de bain, salle à manger). Le panel d'utilisateurs est composé d'une quinzaine de personnes pour la première version du jeu de données. Les utilisateurs ont effectué différentes actions et simulé différents événements. Des exemples sont visibles dans la figure 5.2.

Les capteurs de la caméra ont été montés sur un robot mobile à deux hauteurs différentes, à 30 cm et à 1 mètre 50 du sol. Cette configuration offre une variété d'options pour les positions des caméras et les arrière-plans. De plus, elle permet de déplacer et de faire pivoter les caméras pendant la capture des actions. Ainsi, certaines vidéos présentent des arrière-plans statiques, tandis que d'autres affichent des arrière-plans dynamiques dus aux mouvements du robot (translation, rotation,

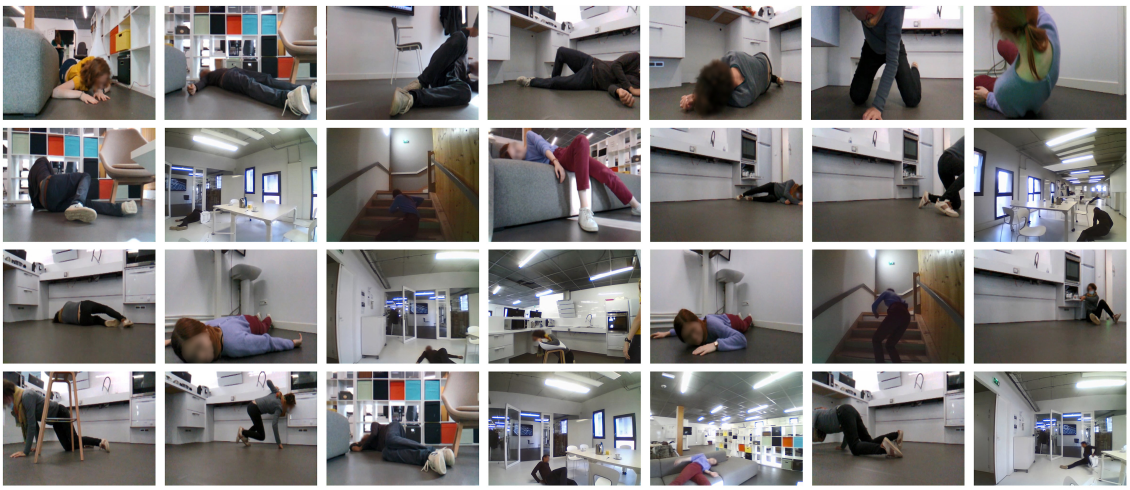

FIGURE 5.2 – Exemples de situations anormales dans le contexte de la robotique d'assistance à l'autonomie à domicile provenant de notre jeu de données JARD.

ou les deux).

5.5.2 Contenu du jeu de données JARD

TABLE 5.4 – Liste des classes du jeu de données JARD.

Action	Mouvement	Anormale	Nb de vidéos	Classe similaire
Tomber	Oui	Oui	184	Se pencher
Attaque cardiaque	Oui	Oui	106	-
Se pencher	Oui	Non	102	Tomber
S'asseoir	Oui	Non	134	Se pencher et Tomber
Être assis	Non	Non	78	-
Être allongé	Non	Non	134	Être allongé à terre
Être allongé à terre	Non	Oui	161	Être allongé sur un lit
Se mettre debout	Oui	Non	106	Se lever
Se lever	Oui	Non	182	Se mettre debout
Marcher	Oui	Non	177	-

La première version du jeu de données contient 10 classes détaillées dans le tableau 5.4. Ces classes ont été choisies pour être adaptées au contexte de l'assistance à l'autonomie à domicile.

FIGURE 5.3 – Exemples d'événements du jeu de données JARD.

Certaines actions représentent des activités et des événements quotidiens normaux, comme s'asseoir, marcher, être assis et être allongé.

Un autre sous-ensemble de données contient des événements et des actions anormales, tels que tomber ou encore être allongé sur le sol. Le jeu de données contient également des situations normales présentant une grande similitude avec les situations anormales afin que le robot puisse les distinguer. Ainsi les classes "tomber" et "se pencher" peuvent être confondues. Dans les deux cas, la personne a un mouvement vers le bas.

Certaines situations présentent des postures similaires, mais le contexte les rend différentes. Par exemple, être allongé sur un lit est différent d'être allongé sur le sol. La première situation est une situation normale de la vie quotidienne alors que la seconde est une situation anormale pour laquelle une alerte doit être lancée.

Apprendre les deux situations permettra de réduire les fausses alarmes liées à la confusion de ces situations.

Dans les classes "être allongé", "être allongé à terre", "être assis" les sujets

peuvent être immobiles. Les classes "être allongé" et "être assis" sont des classes de situation normale de la vie quotidienne. La classe "être allongé à terre" est une classe de situation anormale pour laquelle une alerte doit être lancée. Dans l'ensemble de ces classes, la personne peut être immobile ou en mouvement. Ce sont des classes d'événement. Les sujets ne font pas une action mais sont dans une posture où ils peuvent être mobiles (présence de mouvements), peu mobiles (mouvements très lents, ou très petits) ou complètement immobiles (sans mouvement).

Ainsi une personne allongée dans un lit peut être en train de dormir (présence d'immobilité), en train de lire (peu de mouvements) ou en train de bouger (présence de mouvements). De même qu'une personne allongée à terre peut être dans un état comateux (présence d'immobilité), abasourdi (présence de petits mouvements lents) ou en train de bouger (présence de mouvements). Des exemples d'actions normales ou dangereuses, statiques ou mobiles sont visibles dans la figure 5.3.

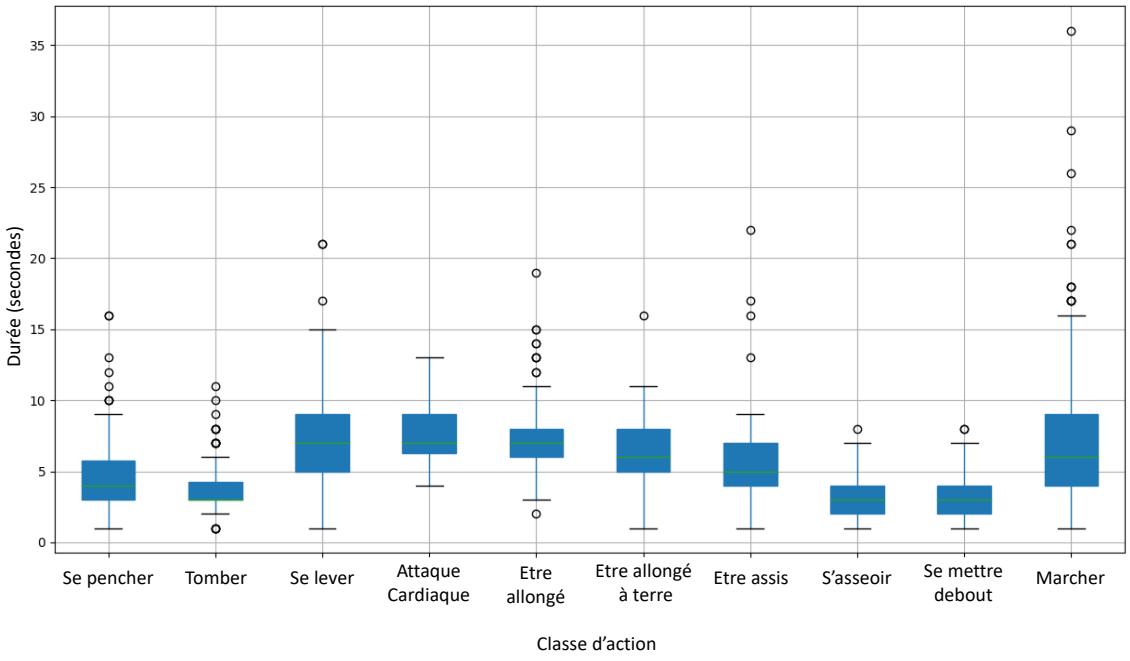

FIGURE 5.4 – Distribution de la durée des actions dans notre jeu de données JARD.

Chaque situation est représentée par plus d'une centaine de vidéos sauf la classe "être assis" qui n'en contient que 78.

Les vidéos ont des durées différentes selon la nature de l'action. Dans la fig 5.4,

nous pouvons voir que la durée des vidéos va de 2 secondes jusqu'à 36 secondes. L'action "marcher" à une grande variation dans la durée des vidéos, comprise entre 2 et 36 secondes. Cela est dû au fait que les personnes peuvent marcher longtemps avec des variations dans leur direction. C'est le cas lorsqu'elles traversent la pièce ou lorsque le robot les suit. Elles peuvent également passer rapidement devant le robot, ce qui donne des vidéos plus courtes. Les actions "s'asseoir", "se mettre debout", et "tomber" sont des actions rapides et brèves. Les vidéos sont donc plus courtes (entre 2 et 15 secondes) et avec une moins grande variation dans la durée. Les actions "être allongé" "être allongé à terre", "être assis" et "se lever" ont des durées et des variations plus grandes. Certaines vidéos des actions "être allongé", "être allongé à terre" et "être assis" contiennent des immobilités avec des mouvements du robot. Les durées sont donc plus importantes. D'autres vidéos contiennent des immobilités avec aucun mouvement du robot, les durées sont donc moins importantes.

Pour être au plus proche de la réalité, l'ensemble des classes d'actions contiennent différentes variations. Il peut y avoir un changement de rythme de la personne réalisant les actions, une différence de lieu et d'arrière-plan au sein d'une même classe d'action, et des différences notables dans la manière de réaliser les actions. Ainsi dans les vidéos de la classe d'action "être allongé", les personnes peuvent être sous les couvertures, sur les couvertures ou dans le fauteuil. Elles peuvent lire, regarder la télé ou encore dormir.

Dans l'ensemble des classes des défis supplémentaires ont été incorporés comme des occultations ou plusieurs personnes présentes dans la scène. Pour les vidéos de chutes, les chutes peuvent être partielles ou totales hors du champ de vision de la caméra. Ainsi la chute peut être visible en totalité ou partiellement sur la vidéo.

5.6 Synthèse

Dans la littérature, de nombreux jeux de données sont présents, mais la plupart se concentrent sur la reconnaissance de situations normales. Certains abordent les situations anormales, mais souvent dans des contextes extérieurs ou publics, plutôt que dans le cadre spécifique de l'assistance au maintien à domicile. De plus, les

jeux de données conçus pour la détection de chutes se limitent généralement aux activités quotidiennes et aux chutes, excluant des situations comme l'immobilité due à des circonstances telles qu'une personne inconsciente allongée au sol, sans qu'une chute soit visible par la caméra.

En fin de compte, malgré la quantité de jeux de données disponibles, il y a un manque de données adaptées pour former des modèles pertinents dans le domaine de la robotique d'assistance au maintien à domicile. Des scénarios spécifiques, impliquant l'immobilité des personnes et les mouvements de la caméra causés par les déplacements du robot, sont particulièrement sous-représentés dans les jeux de données existants. Cela met en évidence l'importance du développement de jeux de données adaptés à ce contexte spécifique de la robotique d'assistance à domicile.

Pour relever ce défi de l'insuffisance de données dans la reconnaissance de situations anormales en robotique d'assistance, nous avons créé un nouveau jeu de données appelé JARD. Il a été spécialement conçu pour se concentrer sur la reconnaissance de situations anormales dans le contexte de la robotique d'assistance, une caractéristique cruciale pour assurer la sécurité et le bien-être des personnes qui dépendent de l'assistance à domicile.

Il permet de distinguer des situations anormales et des situations normales qui pourraient être confondues, évitant ainsi de générer de fausses alertes (par exemple, il différencie être allongé dans un lit de se trouver allongé sur le sol). De plus, il améliore la capacité à détecter les actions humaines et les situations anormales, même lorsque des mouvements de caméra sont provoqués par les déplacements du robot, qu'il s'agisse de rotations ou de translations.

Les vidéos ont été capturées du point de vue d'un robot d'assistance, ce qui rend ce jeu de données particulièrement adapté à la robotique d'assistance.

Il est important de noter que ce nouveau jeu de données JARD peut être utilisé de manière autonome pour l'apprentissage d'une IA ou en complément de jeux de données existants.

Chapitre 6

Expérimentation

Notre principal objectif est de développer une intelligence artificielle capable de discerner les situations normales des situations anormales dans le contexte de la robotique d'assistance au maintien à domicile.

Dans la littérature, il existe plusieurs modalités d'entrée et plusieurs architectures de classification.

Notre démarche vise à identifier le couplage optimal entre une modalité d'entrée et une architecture qui convient le mieux à notre contexte spécifique. Pour atteindre cet objectif, nous avons mené des expérimentations portant sur différents couples (m, a), comme illustré dans la figure 4.5.

Dans ce chapitre nous commençons par présenter le choix et la préparation des jeux de données utilisés dans nos expérimentations.

Puis nous montrerons le résultat des différentes modalités appliquées sur les jeux de données.

Après avoir introduit les métriques d'évaluations, nous présenterons et nous discuterons des résultats obtenus.

6.1 Préparation des jeux de données

Cette recherche de la solution optimale doit se faire sur notre jeu de données dédié à la robotique d'assistance au maintien à domicile.

Il est également intéressant de l'effectuer sur jeu de la littérature. En évaluant nos différentes expérimentations sur un jeu de données de la littérature nous pour-

rons vérifier la robustesse et la capacité de généralisation de notre approche.

En utilisant plusieurs jeux de données, nous pourrons obtenir une évaluation plus complète de notre solution de reconnaissance d'actions et vérifier qu'elle peut être adaptable à différentes conditions et contextes. En effet, chaque jeu de données peut mettre l'accent sur des aspects spécifiques. Ainsi, notre jeu de données met l'accent sur les périodes d'immobilité et sur les mouvements de la caméra, tandis que les jeux de données de la littérature mettent davantage l'accent sur la gestuelle et les mouvements des humains

Nous avons décidé d'effectuer les expérimentations sur le jeu de données UCF101 [86]. Ce jeu de données est intéressant car il contient de nombreux types d'actions génériques que nous pouvons retrouver dans la littérature comme les actions liées au sport, aux interactions humaines, à la pratique d'un instrument de musique, aux soins personnels et aux actions quotidiennes.

Cette diversité dans les types d'actions entraîne une diversité dans les scènes, les postures et les mouvements des sujets. Il va nous permettre de tester la généralisation de notre approche sur l'ensemble de ces types d'actions et de mouvements.

Nous utilisons deux jeux de données pour nos expérimentations, notre jeu de données JARD et le jeu de données UCF101 [86].

60% des vidéos de chaque jeu de données (UCF101 et JARD) sont utilisées pour l'entraînement, 20% pour les tests lors de l'apprentissage et les 20% restants pour la validation du modèle.

Du fait que la longueur des vidéos varie, le nombre d'images diffère d'une vidéo à l'autre, et l'action peut se produire à divers moments au sein de chaque vidéo. Les enregistrements vidéo sont capturés à un rythme de 25 images par seconde. Pour saisir le mouvement de manière adéquate, nous considérons une séquence de 8 images, en les sélectionnant à intervalles de 16 images. Le point de départ de cette sous-séquence de 8 images est choisi de manière aléatoire.

Cela permet au modèle d'apprendre à reconnaître l'action indépendamment de sa position spécifique dans la vidéo. En exposant le modèle à différentes positions de départ, il est exposé à une plus grande diversité d'exemples. Il devient plus robuste et généralisable.

Pour réduire le temps de calcul lors de l'inférence et ainsi permettre une utilisa-

tion en temps réel dans un système robotique, l'ensemble des séquences d'images vidéo est redimensionné à 256px par 256px.

6.2 Modalités d'entrée

Notre objectif est d'évaluer chaque couple entre une modalité d'entrée et une architecture pour la reconnaissance d'actions. Pour cela nous avons pré-traité nos données avec les différentes modalités d'entrée pour ensuite les comparer sur nos différentes architectures.

6.2.1 RGB

La modalité d'entrée la plus simple est le RGB. Il n'y a pas de prétraitement à effectuer. Ce sont dans notre cas les séquences d'images vidéo d'origine.

Nous l'avons utilisée en entrée de l'ensemble des architectures à un flux que nous expérimentons (celle basée sur les ConvLSTM, celle basée sur les mécanismes d'attention et celle basée sur CLIP).

6.2.2 Détection du squelette humain

Pour détecter le squelette humain nous nous sommes appuyés sur le modèle *Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression* appelé DEKR [87].

Des exemples du résultat du prétraitement sur notre jeu de données JARD sont visibles dans la figure 6.1.

Nous pouvons voir que dans certains cas, la détection du squelette humain permet d'identifier facilement la position de la personne. Sur la figure 6.1-a, malgré le mouvement de chute de la personne, nous observons que la détection du squelette humain permet de voir correctement ce mouvement de chute de la personne. Sur la figure 6.1-b où la personne est assise, nous observons que la personne est immobile et que le robot est en mouvement de rotation lent. Malgré cette immobilité de la personne et ce mouvement du robot, le squelette et donc la posture de la

personne reste parfaitement visible. Dans d'autre cas la détection et la modélisation du squelette est plus approximative comme sur la figure 6.1-c où la personne est allongée à terre. Cela est dû notamment à la position de la personne et aux éventuelles occultations dues à cette position allongée. Sur la figure 6.1-d où la personne est allongée sur un lit, les articulations de la personne ne sont pas détectées. Le squelette n'est donc pas visible. La personne est dans une position trop complexe et trop éloignée de la caméra pour pouvoir avoir une détection correcte du squelette humain. L'apprentissage se base alors sur les données RGB.

FIGURE 6.1 – Exemples de la détection du squelette humain appliquée sur notre jeu de données JARD.

Nous avons utilisé cette modalité en entrée de notre architecture à un flux basée sur les ConvLSTM et de notre architecture basée sur les mécanismes d'attention.

6.2.3 Flux optique

Nous avons utilisé l'estimation de flux optique dense en entrée de notre architecture à un flux basée sur les convLSTM et celle basée sur les mécanismes d'attention.

Des exemples de l'estimation de flux optique dense sur notre jeu de données sont visibles dans la figure 6.2. Dans certaines vidéos de notre jeu de données, le robot est immobile. Le mouvement de la personne est alors bien visible. C'est le cas pour la vidéo de la classe "tomber" illustrée sur la figure 6.2-a.

Mais si la personne est immobile et le robot aussi, comme c'est le cas pour la classe "être allongé" illustrée sur la figure 6.2-c, il est impossible de classifier correctement l'action. L'estimation de flux optique retourne une image noire.

Dans d'autres vidéos, le mouvement du robot bruite l'image. Dans certains cas, l'action reste visible. Sur la figure 6.2-b qui correspond à l'action "se pencher", le mouvement de la personne reste visible malgré le bruitage de l'arrière-plan dû au mouvement du robot. Mais sur la figure 6.2-c qui correspond à une vidéo de l'action "Marcher", le mouvement du robot rend difficile la reconnaissance de la personne et donc de l'action réalisée.

6.2.4 Segmentation sémantique de la scène

Pour effectuer la segmentation sémantique de la scène, nous nous sommes appuyés sur le ViT-Adapter [88].

Nous avons utilisé l'entraînement existant sur le jeu de données ADE20K [109]. Ce jeu de données permet de classer chaque pixel de l'image en 150 classes distinctes. Certaines classes correspondent à l'extérieur *(sky, building, sidewalk, etc)*, d'autres correspondent à l'intérieur *(table, sofa, bed, desk, toilet, etc)* et une classe correspond à une personne.

Des exemples de cette segmentation sémantique de la scène sur notre jeu de données sont visibles dans la figure 6.3. Avec la détection du squelette, la personne reste visible, qu'elle soit en mouvement, comme illustré sur la figure 6.3-a, ou immobile, comme illustré sur les figures 6.3-b et c. Cette segmentation sémantique

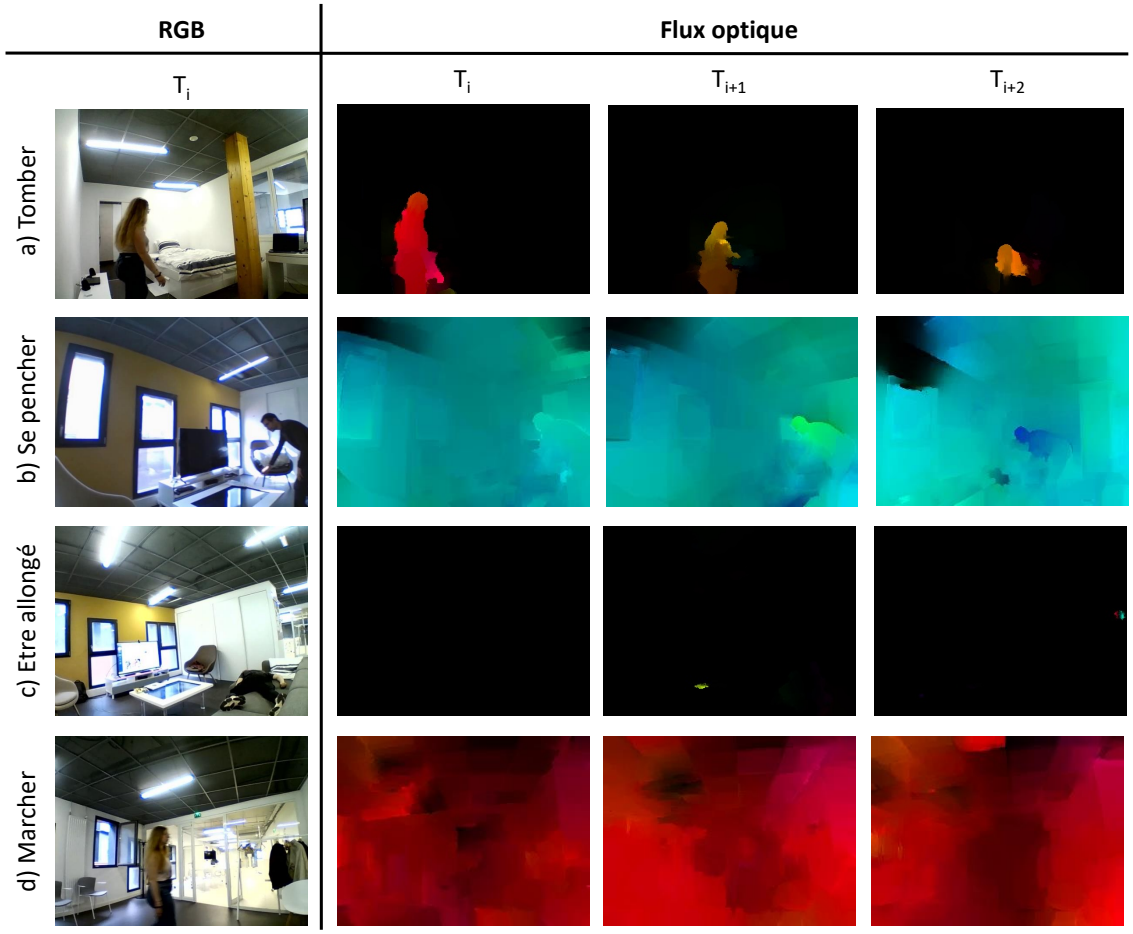

FIGURE 6.2 – Exemples de l'estimation de flux optique dense appliquée sur notre jeu de données JARD.

de la scène nous donne l'emplacement de la personne au sein de son environnement. Ainsi sur la figure 6.3-b, nous voyons que la personne est allongée sur son lit. Sur la figure 6.3-c, nous voyons que la personne est allongée sur le sol. Nous pouvons ainsi déduire facilement l'emplacement de la personne au sein de son environnement.

La segmentation sémantique de la scène cherche à classer l'ensemble des pixels d'une image. Il peut y avoir des erreurs de classification. Ainsi dans la figure 6.3-d, la personne n'est pas reconnue par le modèle et n'a pas été correctement classée.

Nous avons utilisé cette modalité en entrée de notre architecture à un flux basée sur les ConvLSTM et de notre architecture basée sur les mécanismes d'attention.

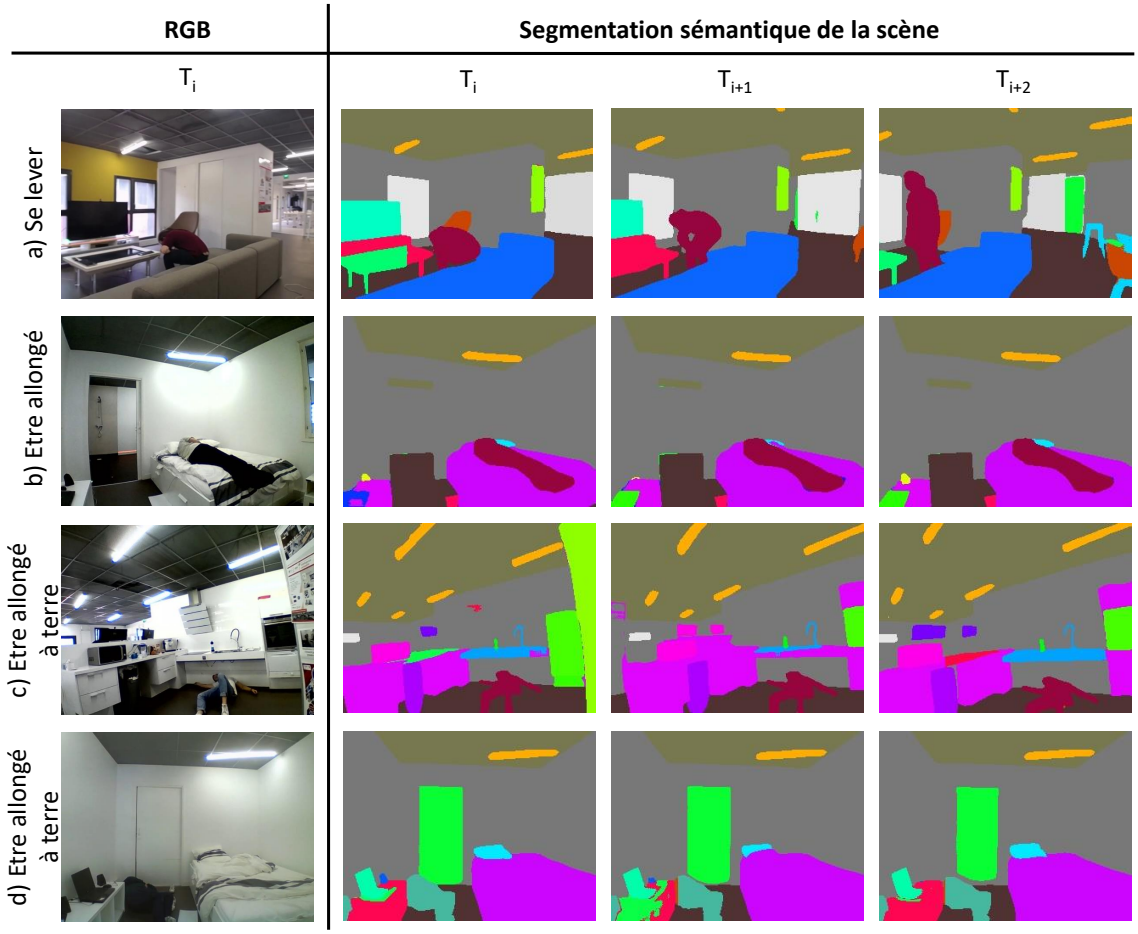

FIGURE 6.3 – Exemples de la segmentation sémantique de la scène appliquée sur notre jeu de données JARD.

6.2.5 Segmentation sémantique des parties du corps humain

Pour effectuer la segmentation sémantique des parties du corps humain, nous nous sommes appuyés sur des modèles d'apprentissage profond existants et en particulier sur *Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation* [1].

Des exemples de la segmentation sémantique des parties du corps humain sur notre jeu de données sont visibles dans la figure 6.4. La segmentation sémantique des parties du corps humain permet de distinguer l'emplacement et la position de

chaque partie du corps visible.

Il arrive dans certains cas que la segmentation sémantique des parties du corps humain ne soit pas complète. C'est le cas dans l'exemple illustré sur la figure 6.4-c. Nous voyons qu'il y a un mouvement de la personne, mais nous n'avons pas les détails de son mouvement. L'ensemble des pixels de la personne n'a pas été reconnu par le modèle. Il est alors difficile de voir si la personne tombe ou si elle se penche.

Dans d'autres cas, aucune partie de la personne n'est reconnue. C'est le cas sur la figure 6.4-d. L'image de sortie est alors une image noire dépourvue d'information. Il est alors impossible de reconnaître l'action.

Les seules informations obtenues par cette modalité sont liées à la position du corps de la personne. Il y a une perte de l'ensemble des informations liées au contexte. Ainsi sur la figure 6.4-b, nous voyons que la personne est allongée. Mais il est impossible de savoir si elle est allongée sur un lit, un fauteuil ou à terre.

Nous avons utilisé cette modalité en entrée de notre architecture à un flux basée sur les ConvLSTM et de notre architecture basée sur les mécanismes d'attention.

6.3 Utilisation de plusieurs modalités d'entrée

Chacune des modalités précédentes ont leurs avantages et leurs inconvénients. Il peut être intéressant d'utiliser plusieurs modalités, afin de combiner leurs avantages et/ou de combler leurs inconvénients.

Certaines modalités peuvent être fusionnées en une seule image. Cette fusion permet d'avoir les informations pertinentes des deux modalités et de continuer à utiliser un réseau à un flux. Il est aussi possible d'utiliser un réseau à deux flux où chaque flux va prendre une modalité différente en entrée.

Dans cette section, nous allons étudier l'influence de l'utilisation de plusieurs modalités sur les résultats. Certaines modalités seront fusionnées sur une seule image quand d'autres seront utilisées sur notre architecture à deux flux.

6.3.1 RGB et flux optique

Dans la littérature, le flux optique est généralement utilisé avec une image RGB.

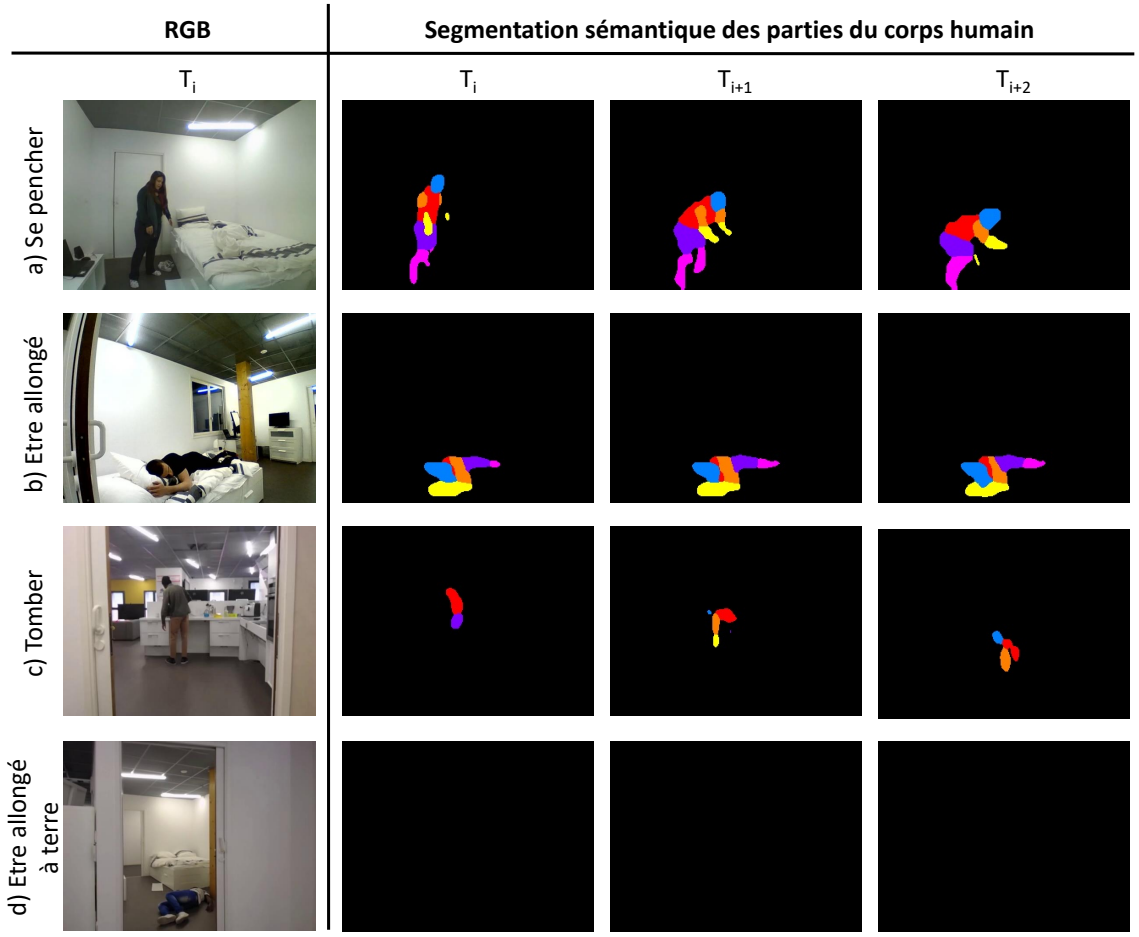

FIGURE 6.4 – Exemples de la segmentation sémantique des parties du corps humain appliquée sur notre jeu de données JARD.

Le flux optique permet d'avoir toutes les informations temporelles des différents mouvements mais perd beaucoup d'informations spatiales. Les séquences d'images RGB ont l'avantage d'être facilement utilisables. Elles conservent toutes les informations d'origine mais peuvent être bruitées.

L'idée est d'utiliser ces deux modalités complémentaires. Nous allons utiliser notre architecture à deux flux. Ainsi le premier flux prend en entrée une image RGB pour modéliser les informations spatiales et donc le contexte. Le second flux prend en entrée la séquence de flux optique correspondante pour modéliser les informations temporelles et ainsi avoir des informations sur les différents mouvements de la scène.

6.3.2 RGB et segmentation sémantique de la scène

Les séquences d'images RGB ont l'avantage d'être facilement utilisables. Elles conservent toutes les informations d'origine mais peuvent être bruitées.

La segmentation sémantique de la scène à l'avantage de mettre en avant tous les différents objets présents dans la scène. Dans l'architecture à deux flux, nous avons donné en entrée au premier flux une seule image RGB et au second flux une séquence d'images segmentées.

Certains objets ont peu d'intérêt dans notre contexte de reconnaissance de situation anormale. Nous avons donc gardé que la classification des objets importants sur notre jeu de données. Nous avons conservé 7 classes : les lits, les sofas, les fauteuils, les chaises, les murs, le sol et les personnes. Nous avons ensuite superposé ces 7 classes sur nos données RGB comme illustré sur la figure 6.5.

Cette fusion permet de mettre en avant les objets importants de la scène avec la segmentation sémantique de la scène tout en gardant les informations RGB. Ainsi s'il y a une mauvaise segmentation, les informations restent visibles avec le RGB.

Nous avons ensuite donné ces données en entrée à notre architecture à un flux basée sur les convLSTM et à notre architecture basée sur les mécanismes d'attention.

6.3.3 RGB et segmentation sémantique des parties du corps humain

La segmentation des parties du corps humain permet d'avoir une bonne vision de l'humain, mais elle génère une perte d'informations liée au contexte et à l'environnement. Afin de compenser cette perte d'informations et ainsi améliorer les résultats, nous l'avons utilisé avec les données RGB. Ainsi, sur notre architecture à deux flux, le premier flux prend en entrée une image RGB et notre second flux la séquence d'images segmentées correspondante.

Pour l'architecture à un flux basée sur les convLSTM et celle basée sur les mécanismes d'attention, les images RGB ont été fusionnées avec les images segmentées comme illustré sur la figure 6.6.

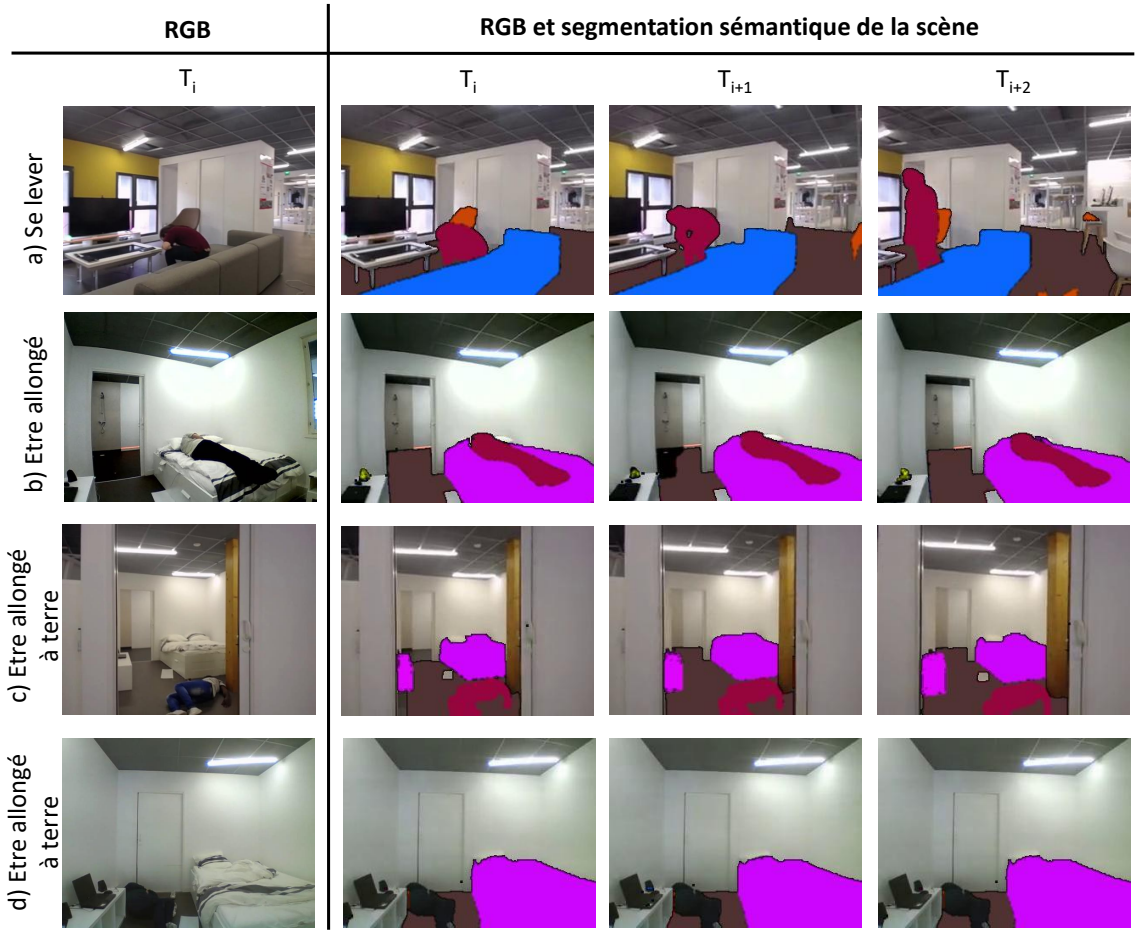

FIGURE 6.5 – Exemples de la fusion du RGB et de la segmentation sémantique de la scène appliquée sur notre jeu de données.

Cette fusion comble le déficit d'informations résultant de la segmentation sémantique des parties du corps humain tout en mettant en évidence la position de la personne dans la scène. Par exemple, sur la figure 6.6-b nous voyons que la personne est allongée sur un lit et non à terre.

Les informations perdues en raison de la mauvaise classification des pixels de l'image sont compensées par les données provenant des images RGB. Par exemple, sur la figure 6.6-d, la personne n'a pas été reconnue par le modèle de segmentation sémantique du corps humain mais elle reste visible sur l'image RGB.

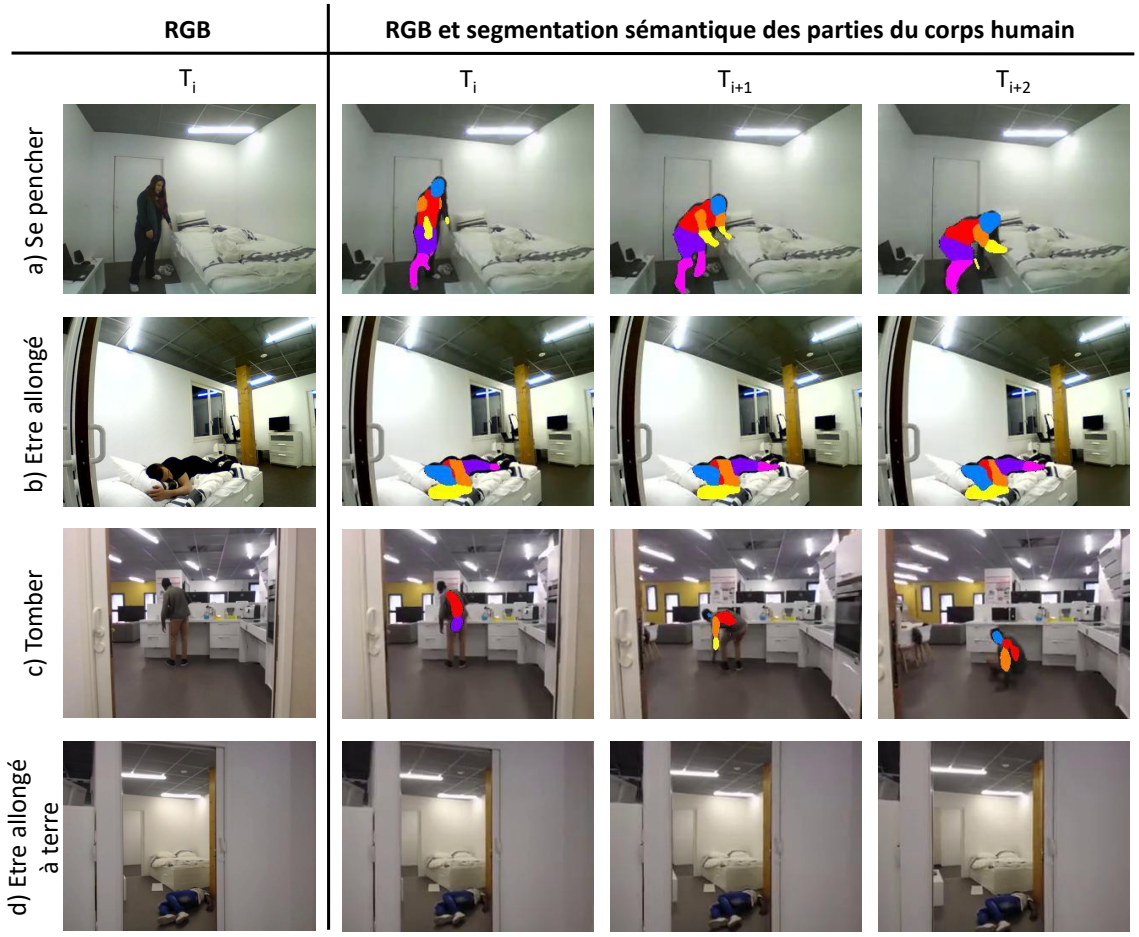

FIGURE 6.6 – Exemples de la fusion du RGB et de la segmentation sémantique des parties du corps humain appliquée sur notre jeu de données.

6.4 Métriques d'évaluation

Pour comparer les différentes approches de la reconnaissance d'actions humaines, il est nécessaire d'utiliser des métriques d'évaluation communes. La métrique d'évaluation des approches de reconnaissance d'actions humaines la plus utilisée est l'exactitude. L'exactitude peut être définie comme le rapport entre le nombre de prédictions correctes et le nombre total d'échantillons en entrée.

Son calcul se base sur la matrice de confusion, un tableau croisé entre les valeurs réelles et les valeurs prédites. Cette matrice est généralement construite à partir d'un ensemble de données de test, où chaque échantillon est associé à une étiquette

connue.

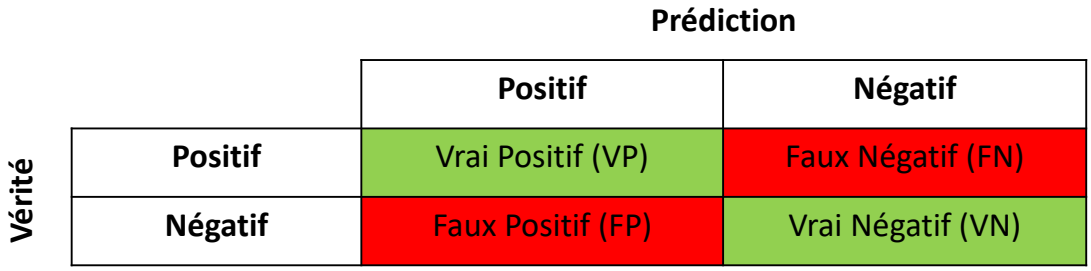

FIGURE 6.7 – Matrice de confusion.

Cette matrice de confusion, visible sur la figure 6.7, est composée de 4 valeurs.

- Vrai positif (VP) : C'est le nombre d'actions correctement prédites comme appartenant à une classe d'action spécifique.

- Faux positif (FP) : C'est le nombre d'actions incorrectement prédites comme appartenant à une classe d'action spécifique.

- Vrai négatif (VN) : C'est le nombre d'actions correctement prédites comme n'appartenant pas à une classe d'action spécifique.

- Faux négatif (FN) : C'est le nombre d'actions incorrectement prédites comme n'appartenant à une classe d'action spécifique.

L'exactitude se calcule à l'aide de la formule suivante :

$$\frac{VP + VN}{VP + VN + FP + FN} \tag{6.1}$$

6.5 Discussion

Dans cette section, nous allons discuter de notre jeu de données et des résultats obtenus dans nos différentes expérimentations.

Les résultats obtenus sur notre jeu de données sont visibles dans le tableau 6.1 et les résultats obtenus sur le jeu de données UCF101 sont visibles dans le tableau 6.2.

TABLE 6.1 – Résultats obtenus sur notre jeu de données.

Architecture	**Modalité**	**Score**
ConvLSTM (un flux)	RGB	75.81
ConvLSTM (un flux)	Détection du squelette	73.63
ConvLSTM (un flux)	Flux optique dense	52.66
ConvLSTM (un flux)	segmentation sémantique de la scène	74.93
ConvLSTM (un flux)	segmentation sémantique des parties du corps humain	65.82
ConvLSTM (un flux)	RGB et segmentation sémantique de la scène	75.52
ConvLSTM (un flux)	RGB et segmentation sémantique des parties du corps humain	**78.31**
Mécanisme d'attention	RGB	79.03
Mécanisme d'attention	Détection du squelette	79.03
Mécanisme d'attention	Flux optique dense	69.76
Mécanisme d'attention	segmentation sémantique de la scène	75.81
Mécanisme d'attention	segmentation sémantique des parties du corps humain	66.13
Mécanisme d'attention	RGB et segmentation sémantique de la scène	77.42
Mécanisme d'attention	RGB et segmentation sémantique des parties du corps humain	**80.24**
ConvLSTM (deux flux)	RGB + Flux optique	76.13
ConvLSTM (deux flux)	RGB et segmentation sémantique de la scène	77.42
ConvLSTM (deux flux)	RGB et segmentation sémantique des parties du corps humain	**79.96**
CLIP	RGB	55.45

6.5.1 Jeu de données

Nous avons expérimenté différentes approches sur notre jeu de données et sur UCF101.

Sur les différents tableaux de résultats (6.1, 6.2), nous pouvons nous apercevoir que les résultats obtenus sur notre jeu de données sont moins bons que ceux obtenus sur le jeu de données d'UCF101.

Notre jeu de données est plus difficile que le jeu de données UCF101. Afin d'enrichir la diversité des actions capturées, rendant l'apprentissage plus réaliste et pertinent pour des situations réelles, certaines vidéos ont été filmées dans de petits espaces de pièces d'habitation. Ainsi, dans certaines vidéos, la distance entre la caméra et les personnes est très faible. La personne se retrouve très près de la caméra et est difficilement distinguable. Dans d'autres vidéos, il n'y a qu'une vue partielle des actions ou de la personne.

TABLE 6.2 – Résultats obtenus sur le jeu de données UCF101.

Architecture	**Modalité**	**Score**
ConvLSTM (un flux)	RGB	87.55
ConvLSTM (un flux)	Détection du squelette	83.84
ConvLSTM (un flux)	Flux optique dense	74.68
ConvLSTM (un flux)	segmentation sémantique de la scène	64.43
ConvLSTM (un flux)	segmentation sémantique des parties du corps humain	55.75
ConvLSTM (un flux)	RGB et segmentation sémantique des parties du corps humain	**88.16**
Mécanisme d'attention	RGB	**97.71**
Mécanisme d'attention	Détection du squelette	97.71
Mécanisme d'attention	Flux optique dense	85.55
Mécanisme d'attention	segmentation sémantique de la scène	76.87
Mécanisme d'attention	segmentation sémantique des parties du corps humain	58.71
Mécanisme d'attention	RGB et segmentation sémantique des parties du corps humain	97.50
ConvLSTM (deux flux)	RGB + Flux optique	86.16
ConvLSTM (deux flux)	RGB et segmentation sémantique de la scène	93.19
ConvLSTM (deux flux)	RGB et segmentation sémantique des parties du corps humain	**94.25**
CLIP	RGB	73.80

Le mouvement du robot peut également bruiter les vidéos en ajoutant des mouvements sur les objets, les personnes et l'arrière-plan. Ces mouvements bruités peuvent rendre plus difficile l'apprentissage des caractéristiques temporelles.

Une autre contrainte est liée à la position des caméras. Les données sont enregistrées à partir de deux positions de caméra différentes, dont l'une est positionnée près du sol. La même action est donc filmée depuis deux points de vue différents. Une même action n'est pas représentée de la même manière selon la position de la caméra. Ainsi, certaines données ne contiennent que la partie inférieure du corps et d'autres ne contiennent que la partie supérieure du corps.

La similitude entre les différentes classes complexifie la reconnaissance d'actions humaines sur notre jeu de données. Malgré cette similitude, la distinction des différentes classes est nécessaire pour le contexte de l'aide à domicile des personnes fragilisées. L'action "se mettre debout" peut facilement être confondue avec l'action "se lever". L'action "tomber" peut facilement être confondue avec l'action "se pen-

cher". La première nécessite une alerte, tandis que la deuxième n'en nécessite pas. L'action "être allongé" peut facilement être confondue avec l'action "être allongé à terre". La première ne nécessite pas forcément d'alerte tandis que la deuxième nécessite une alerte surtout si la personne est immobile ou si elle n'arrive pas à se relever.

Malgré les difficultés supplémentaires de notre jeu de données, il offre une précieuse ressource pour l'apprentissage de la reconnaissance d'actions humaines dans le contexte de la robotique d'assistance au maintien à domicile des personnes fragilisées.

6.5.2 Modalités d'entrée

Nous avons expérimenté différentes modalités d'entrée.

Des exemples des différentes modalités d'entrée utilisées sont visibles dans la figure 6.8.

Chaque modalité d'entrée a ses avantages et ses inconvénients. La modalité RGB est facile à obtenir. Elle a l'avantage de conserver toutes les informations visibles à l'œil nu. Mais elle peut être facilement bruitée, notamment avec les variations d'éclairage.

La détection du squelette humain permet de voir l'emplacement des articulations et des membres des personnes présentes dans la scène. Elle permet de concentrer l'apprentissage sur la personne. Nous pouvons voir dans les tableaux 6.1 et 6.2 que pour certaines architectures, la détection du squelette nous donne des résultats similaires à ceux obtenus avec le RGB. La méthode de détection du squelette n'arrive pas toujours à détecter les articulations pour pouvoir reconstituer le squelette. C'est notamment le cas lorsqu'il y a des occlusions ou lorsque les personnes sont dans des positions trop complexes pour ces modèles de détection du squelette. Ainsi lorsque la personne est allongée, il arrive fréquemment que des articulations ne soient pas détectées. Le squelette étant modélisé sur les images RGB, l'apprentissage se focalise alors sur le RGB plutôt que sur le squelette.

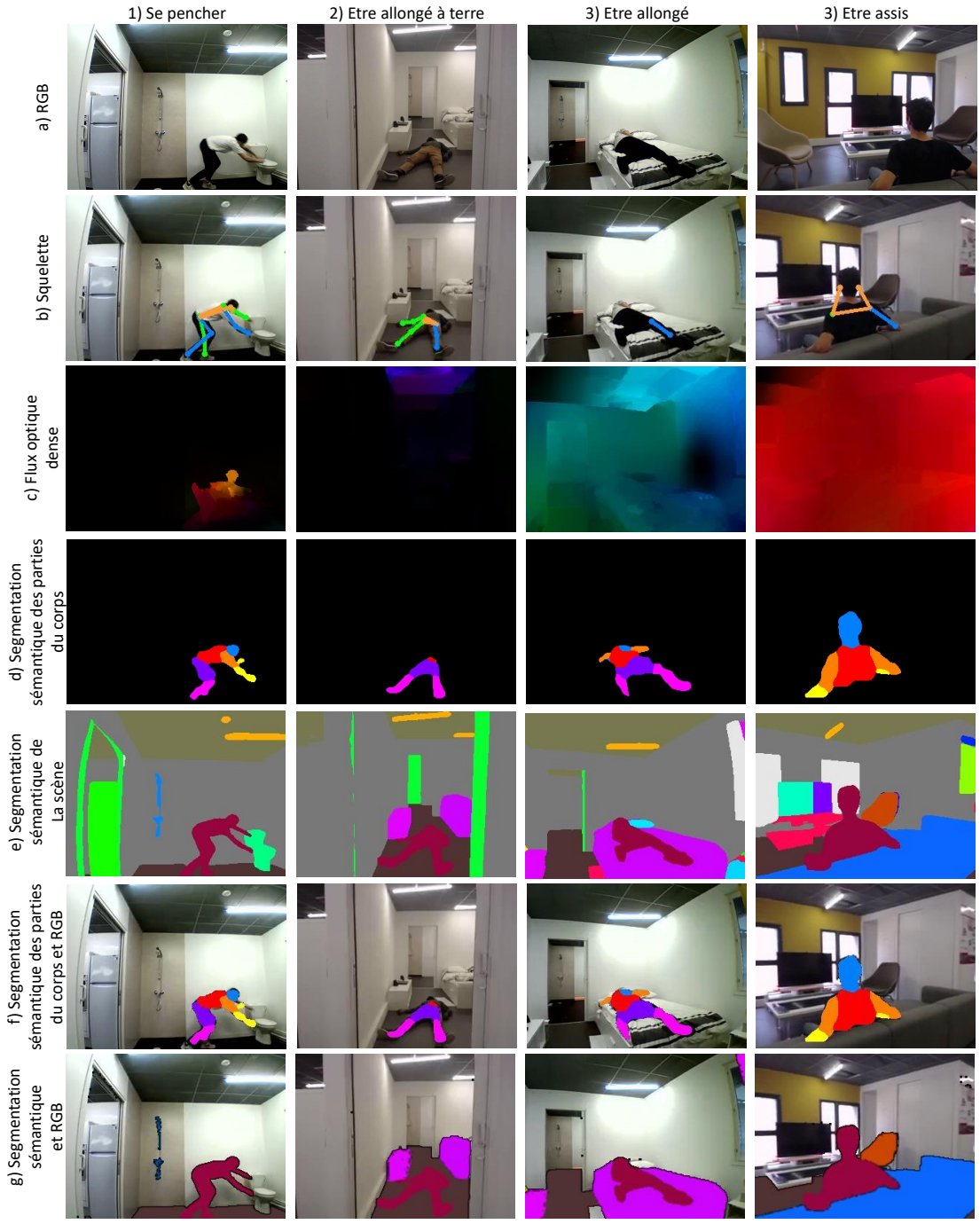

FIGURE 6.8 – Exemples des différentes modalités d'entrée expérimentées sur notre jeu de données JARD.

La segmentation sémantique des parties du corps humain, telle que la détection du squelette, permet de mettre en avant la position des différents membres des personnes présentes dans la scène. Cependant, lorsqu'elle est utilisée seule, elle donne de moins bons résultats que d'autres modalités. Cela est principalement dû à la perte du contexte car seul la personne est visible.

La segmentation sémantique des parties du corps humain peut être fusionnée avec le RGB. L'ajout du RGB apporte des données contextuelles essentielles, générant ainsi de meilleurs résultats par rapport à la simple segmentation sémantique des parties du corps humain.

Sur notre jeu de données, la segmentation sémantique des parties du corps humain fusionnée avec le RGB donne les meilleurs résultats. La segmentation permet d'avoir les informations sur la position des personnes, qu'elles soient immobiles ou en mouvement et que la caméra soit immobile ou en mouvement. Le RGB permet de différencier les actions liées au contexte comme différencier quelqu'un qui est allongé à terre ou dans un lit.

Sur le jeu de données UCF101, avec certaines architectures, la segmentation sémantique des parties du corps humain fusionnées avec le RGB donne les meilleurs résultats. Sur l'architecture avec les mécanismes d'attention, la segmentation sémantique des parties du corps humain donne un résultat légèrement en dessous du RGB et de la détection du squelette (97,50% contre 97,71%).

La segmentation sémantique de la scène permet de classer l'ensemble des pixels dans différentes catégories. Il peut y avoir des erreurs de classification des pixels. Les informations liées au contexte sont alors faussées ce qui entraîne de moins bon résultat.

L'estimation de flux optique dense permet de modéliser les mouvements mais perd toutes les informations liées au contexte. Utilisé seul sur UCF101, il donne de moins bon résultat que le RGB, la détection de squelette ou encore la segmentation sémantique des parties du corps fusionné avec le RGB.

Dans certaines de nos classes d'action, les personnes et la caméra sont immobiles. Il n'y a donc aucune information sur la séquence d'images pré-traitées. Dans d'autres séquences d'images, le mouvement des caméras bruite les données pré-

traitées. Sur notre jeu de données, c'est donc la modalité qui donne les moins bons résultats. L'estimation de flux optique n'est pas adaptée à notre problématique.

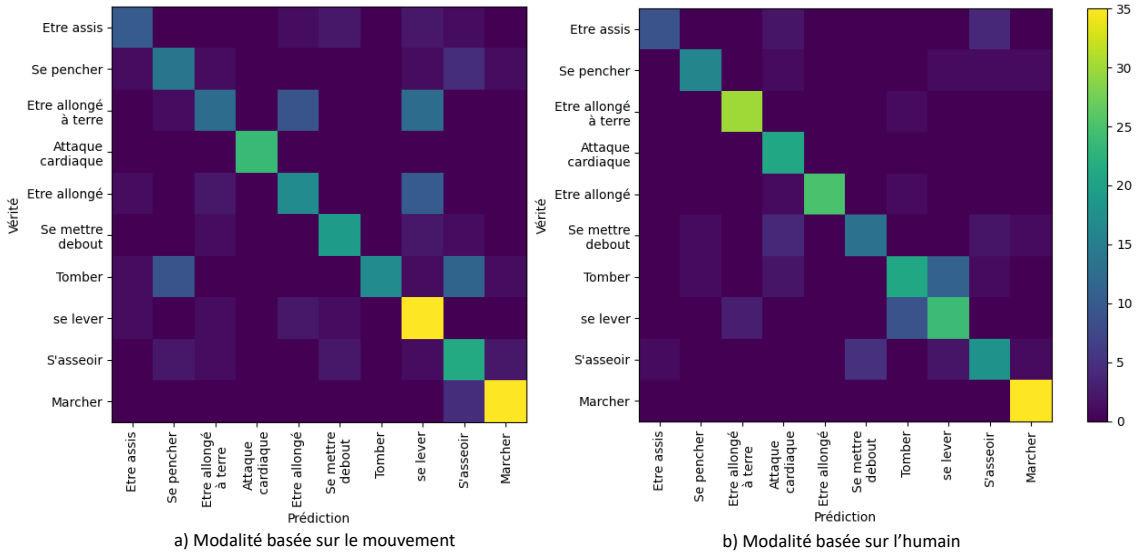

FIGURE 6.9 – Matrices de confusion obtenues sur notre jeu de données.

Sur les matrices de confusion de la figure 6.9, les matrices de confusion révèlent les classes qui sont confondues les unes avec les autres. Dans les situations où l'arrière-plan n'est pas visible, comme dans le flux optique dense, les classes "être allongé" et "être allongé à terre" peuvent être confondues entre elles, ainsi qu'avec la classe "se lever". En effet, dans la classe "se lever", la personne est initialement allongée et se met accroupie ou debout à la fin de la vidéo.

L'action "s'asseoir" présente un mouvement vers le bas similaire à la classe "tomber", ce qui peut entraîner certaines confusions. Dans certaines vidéos de la classe "tomber", la personne marche avant de chuter, ce qui peut également prêter à confusion pour le modèle.

Dans les situations où l'arrière-plan est visible et que l'attention est davantage portée sur la personne, comme dans la segmentation sémantique des parties du corps humain fusionnée avec le RGB, ce sont les classes avec des positions similaires de la personne qui sont le plus souvent confondues. Ainsi, les classes "tomber" et "se lever" sont les plus confondues par le modèle. Dans l'action "tomber", la personne passe souvent de la position debout à la position allongée, tandis que dans

la classe "se lever", c'est l'inverse : la personne passe de la position allongée à la position debout. La même observation s'applique aux classes "s'asseoir" et "se mettre debout".

Ces modalités axées sur l'humain se concentrent davantage sur la position de la personne dans l'espace que sur les caractéristiques temporelles. Elles sont intéressantes pour la reconnaissance d'actions humaines dans le contexte de la robotique d'assistance au maintien à domicile.

6.5.3 Architectures

Nous avons expérimenté nos modalités sur différentes architectures.

CLIP s'appuie sur les caractéristiques spatiales des images pour les décrire textuellement. Le texte obtenu est ensuite utilisé pour prédire la classe de l'action. La classification des actions se fait uniquement sur les données spatiales. Les données temporelles ne sont donc pas prises en compte. Notre architecture de CLIP est donc moins adaptée à notre problématique.

Les mécanismes d'attention et les convLSTM sont deux approches populaires pour les tâches de reconnaissance d'actions humaines. Nous avons expérimenté les deux architectures sur notre jeu de données et sur UCF101. Dans l'ensemble de nos expérimentations, nous pouvons observer que les mécanismes d'attention sont plus performants que les convLSTM.

L'une des raisons possibles de cette différence de performance est la capacité des mécanismes d'attention à se concentrer de manière sélective sur les informations pertinentes d'une séquence vidéo. En se concentrant sur certaines régions ou images, les mécanismes d'attention peuvent mieux capturer la dynamique essentielle de l'action en cours et ainsi mieux la classer. L'architecture composée des convLSTM traitent l'ensemble de la séquence vidéo de manière uniforme, ce qui peut conduire le modèle à être distrait par des informations non pertinentes ou redondantes.

Un autre avantage des mécanismes d'attention est leur capacité à gérer les dé-

pendances à long terme dans une séquence vidéo. Les ConvLSTM sont conçus pour capturer les dépendances spatio-temporelles en étendant les capacités de la LSTM traditionnelle au traitement spatial. Cependant, dans les séquences plus longues, la ConvLSTM peut souffrir de gradients qui s'évanouissent, ce qui rend difficile l'apprentissage des dépendances à long terme par le modèle. Les mécanismes d'attention, quant à eux, peuvent se concentrer sur des images ou des régions importantes de la séquence vidéo, ce qui les rend mieux adaptés à la modélisation de séquences plus longues.

Les mécanismes d'attention sont également moins coûteux en termes de calcul que les convLSTM, car ils ne nécessitent pas autant de paramètres. Cela peut rendre les mécanismes d'attention plus pratiques pour les applications en temps réel ou lorsqu'il s'agit de grands ensembles de données.

Les mécanismes d'attention sont plus performants que les convLSTM dans les tâches de reconnaissance d'actions humaines en raison de leur capacité à s'intéresser de manière sélective aux informations pertinentes et à apprendre les dépendances à long terme. Ils sont également plus adaptés aux applications en temps réel comme le nécessite notre contexte de la robotique d'assistance au maintien à domicile.

6.6 Synthèse

Notre jeu de données permet d'apprendre la reconnaissance de situations normales et anormales à un système robotique pour l'assistance au maintien à domicile.

Les différentes expérimentations permettent de mettre en avant que la segmentation sémantique des parties du corps humain fusionnée avec le RGB donne les meilleurs résultats sur l'ensemble des architectures pour notre jeu de données.

C'est cette modalité couplée à notre architecture basée sur les mécanismes d'attention qui est la solution optimale sur notre jeu de données.

Sur le jeu de données de la littérature UCF101, cette solution arrive après la modalité RGB et la modalité de la détection du squelette, couplées avec cette même architecture.

Notre solution reste néanmoins généralisable aux données généralistes de la littérature.

Chapitre 7

Conclusion

A travers cette thèse, nous avons abordé la problématique de la reconnaissance d'actions humaines dans un contexte d'assistance robotique pour l'aide au maintien à domicile des personnes fragilisées par l'âge, la maladie ou le handicap. L'objectif est de permettre à un robot de reconnaître les différentes situations dangereuses comme les accidents domestiques, les chutes et les malaises pour pouvoir qu'il puisse lancer des alertes. Dans le cadre de notre projet, cette reconnaissance se fait à l'aide d'une caméra RGB embarquée sur un robot d'assistance.

7.1 Contributions

Le travail réalisé dans le cadre de cette thèse nous a permis d'apporter deux principales contributions. La première est la proposition d'une méthode de reconnaissance de situations normales et anormales basée sur l'utilisation de la segmentation sémantique des parties du corps humain. La seconde concerne la création d'un jeu de données pour la reconnaissance de situations anormales dangereuses dans le contexte de l'assistance robotique pour l'aide au maintien à domicile des personnes fragilisées.

7.1.1 Reconnaissance d'actions humaines basée sur la segmentation sémantique des parties du corps humain

Notre problématique se rapproche de la reconnaissance d'actions humaines. Elle permet de reconnaître les actions effectuées par un humain.

La détection de chutes est aussi une partie de notre problématique. Elle permet de détecter la chute par le mouvement qu'effectue l'humain lors de la chute.

Dans les domaines de la reconnaissance d'actions humaines et de la détection de chutes, l'accent est généralement mis sur le mouvement pour reconnaître les situations.

Ces solutions basées sur le mouvement ne permettent pas de reconnaître les situations statiques. La détection et la reconnaissance de situation ne peuvent plus se baser sur le mouvement mais uniquement sur le contexte. Ce sont la détection, la position, et l'emplacement de la personne qui permettent de définir la situation. Quelques travaux ont fondé leur apprentissage sur l'humain. Sur ces méthodes, l'occultation est une limite importante. Il en résulte une perte importante d'informations qui peut empêcher la bonne reconnaissance de la situation.

Dans cette thèse nous avons comparé différentes modalités d'entrée sur différentes architectures afin de déterminer le meilleur couple entre une ou deux modalités d'entrée et une architecture de reconnaissance d'actions humaines pour faire de la reconnaissance de situations normales et anormales dans le contexte de la robotique d'assistance.

Nous avons montré que la segmentation sémantique des parties du corps humain, qui est peu prise en compte dans la littérature, fusionnée avec le RGB permet de prendre en compte les mouvements et les immobilités des personnes, ainsi que les mouvements de la caméra dus aux déplacements du robot d'assistance. Cela permet d'améliorer les résultats (80,24% sur notre jeu de données) par rapport aux modalités basées sur le mouvement comme le flux optique (69,76% sur notre jeu de données).

7.1.2 Nouveau jeu de données

Les jeux de données actuels pour la reconnaissance d'actions humaines comprennent essentiellement des actions de la vie quotidienne ou liées au sport. Les jeux de données pour la détection de chutes comprennent des actions de chutes et des actions de la vie courante. Peu de jeux de données publics tiennent compte des situations statiques et dangereuses pour l'aide au maintien à domicile des personnes fragilisées. Aucun ne permet de détecter des situations anormales et dangereuses comme une personne inconsciente.

Dans cette thèse, nous avons créé un jeu de données pour la reconnaissance de situations anormales et dangereuses dans un contexte d'assistance robotique pour l'aide au maintien à domicile. Ce nouveau jeu de données a été enregistré par des robots mobiles ayant effectué ou non des mouvements lors de l'enregistrement (rotation, translation ou les deux). Ce jeu de données comporte plus de 1250 vidéos réparties sur 10 classes d'actions effectuées par une vingtaine de personnes. Il peut aider la communauté scientifique à progresser dans le domaine de la robotique d'assistance et faciliter la prise en charge des personnes fragilisées.

7.2 Perspectives

Les travaux de cette thèse ouvrent sur divers axes de recherche pour améliorer la reconnaissance d'actions humaines dans le contexte de la robotique d'assistance au maintien à domicile des personnes fragilisées. Les différents axes de recherche concernent l'amélioration des modèles d'apprentissage pour leur permettre de mieux comprendre le contexte, l'allégement des modèles d'apprentissage pour les rendre plus facilement utilisables dans un système robotique en temps réel et l'enrichissement de notre jeu de données.

7.2.1 Mieux inclure le contexte

La connaissance du contexte fait référence à la capacité du robot à comprendre l'environnement dans lequel il opère. Elle est importante pour reconnaître les si-

tuations dangereuses et y réagir, car la gravité de la situation peut dépendre du contexte spécifique de l'environnement. Par exemple, si le robot détecte une personne allongée dans un fauteuil, il doit être capable de déterminer si la personne dort ou s'il s'agit d'une urgence médicale nécessitant une assistance immédiate.

La connaissance du contexte peut également être améliorée en incorporant des indices contextuels tels que le lieu, l'heure de la journée ou les habitudes de la personne. Par exemple, si le robot détecte une personne allongée dans un fauteuil à l'heure habituelle de sa sieste, il peut reconnaître le contexte et tendre plus facilement vers une action normale.

En comprenant mieux le contexte de l'environnement dans lequel il évolue, le robot peut prendre des décisions plus précises et plus efficaces pour devenir un assistant à la sécurité et au bien-être des personnes.

7.2.2 Alléger les modèles

La robotique d'assistance au maintien à domicile est soumise aux contraintes des systèmes embarqués. Le fonctionnement sur batterie, les limites de places et de poids impliquent une réduction de la puissance de calcul par rapport à d'autres systèmes.

Les futurs travaux devront prendre en compte ses contraintes de puissance de calcul et d'énergie pour pouvoir avoir une solution fiable et efficace fonctionnant en local sur le robot.

Une des premières perspectives est de réduire le poids du modèle en supprimant les neurones non nécessaires (pruning) et en optimisant au mieux les algorithmes.

Une autre perspective est de développer des protocoles de communication plus efficaces entre le robot et les systèmes externes, tels que les centres de contrôle à distance ou les installations médicales. En optimisant les protocoles de communication et en minimisant la quantité de données à transmettre, il est possible de réduire les besoins en calcul et en énergie du robot, tout en maintenant un niveau élevé de fonctionnalité et de fiabilité.

Une autre perspective à considérer est l'utilisation de techniques d'apprentissage automatique, telles que l'apprentissage par transfert, pour former des modèles qui peuvent fonctionner efficacement sur des systèmes embarqués. En transférant les connaissances de systèmes plus grands et plus puissants vers des systèmes embarqués, il est possible d'atteindre des niveaux élevés de précision et de performance, tout en respectant les contraintes des systèmes embarqués.

Enfin une dernière perspective est l'apprentissage à faible consommation énergétique basé sur une approche bio-inspirée et en particulier sur les réseaux de neurones à impulsion. Les réseaux de neurones à impulsion sont conçus pour simuler le comportement des neurones biologiques. Chaque neurone accumule les impulsions entrantes au fil du temps et, lorsqu'un seuil est atteint, il génère une impulsion de sortie qui est transmise aux neurones connectés.

Les réseaux de neurones à impulsion sont souvent considérés comme plus économes en énergie que les réseaux de neurones traditionnels, car ils ne nécessitent pas de calculs intensifs à chaque itération. Cela permet de réduire la consommation d'énergie par rapport aux réseaux de neurones traditionnels. Les réseaux de neurones à impulsion répondent aux contraintes des systèmes embarqués et sont donc facilement utilisables dans des systèmes robotiques.

7.2.3 Enrichir le jeu de données

Pour notre jeu de données, plusieurs perspectives sont à explorer. La première est d'enregistrer de nouvelles données sur les classes existantes. Le but est d'enrichir l'ensemble de données existant avec une plus grande diversité d'échantillons. Il faut pour cela collecter des données représentant différents profils ou types de personnes, comme celles qui marchent avec une canne, utilisent un déambulateur ou se déplacent en fauteuil roulant.

Une autre perspective est d'inclure d'autres situations dangereuses qui peuvent se produire, comme des convulsions, des accidents vasculaires cérébraux, des étouffements, etc. L'idée est de collecter des données qui représentent ces situations dans des environnements réels afin de garantir une représentation précise des scénarios

dans lesquels le modèle pourrait être déployé.

Cependant, certaines des actions ou des événements qui doivent être saisis peuvent être difficiles à simuler ou à collecter. Par exemple, il peut être difficile de reproduire un accident vasculaire cérébral ou des convulsions dans un environnement contrôlé. Par conséquent, une solution potentielle pour ces données complexes consiste à explorer la création de données synthétiques imitant les caractéristiques des données réelles.

Chapitre 8

Diffusion des travaux

8.1 Publications

— C. Huyghe, N. Ihaddadene, T. Haessle , C. Djeraba : Reconnaissance d'actions basée sur des modèles de segmentation. 2020 Plate-Forme de l'Intelligence Artificielle (PFIA)

— C. Huyghe, N. Ihaddadene, T. Haessle , C. Djeraba : Human Action Recognition Based on Body Segmentation Models. 2021 International Conference on Content-based Multimedia Indexing. 1-4.

— C. Huyghe, N. Ihaddadene : A dataset of human actions for abnormal event recognition in assistive robotic environments. 2023 Accessibility, Vision, and Autonomy Meet, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

8.2 Conférence sans comité

— Catherine Huyghe, Atelier "Reconnaissance d'actions humaines dans un environnement robotique d'assistance à la personne" au Forum des Sciences Cognitives de Lille 2022 (fsc-lille.com).

Bibliographie

[1] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3) :1066–1078, 2020.

[2] Isabelle Robert-Bobée. Projections de population 2005-2050 : vieillissement de la population en france métropolitaine. *Economie et statistique*, 408(1) :95–112, 2007.

[3] Brigitte Croff. Vieillir chez soi en situation de perte d'autonomie : les enjeux d'un choix éclairé à l'horizon 2030. *Apres-demain*, 63(3) :23–25, 2022.

[4] Catherine Piguet, Marion Droz-Mendelzweig, and Maria Grazia Bedin. Vivre et vieillir à domicile, entre risques vitaux et menaces existentielles. *Gérontologie et société*, 39(1) :93–106, 2017.

[5] F Bloch. Les complications non traumatiques des chutes : des conséquences trop souvent négligées chez la personne âgée. *NPG Neurologie-Psychiatrie-Gériatrie*, 15(88) :188–190, 2015.

[6] G Navarro Ocampo, Vincent Bréjard, and A Bonnet. La chute chez le sujet âgé : de l'impact psychologique au travail psychique. *NPG Neurologie-Psychiatrie-Gériatrie*, 17(97) :42–50, 2017.

[7] V Cayado and R Chahbi. La perception du risque d'accident et de chute par des personnes âgées vivant à domicile : un arbitrage complexe ? *NPG Neurologie-Psychiatrie-Gériatrie*, 15(88) :194–199, 2015.

[8] Yannick Fouquet, Anne-Claire Marmilloud, and Véronique Chirié. Technologies pour la détection et l'alerte en cas de chute : état des lieux, limites et recommandations pour leur accompagnement et amélioration. *Terminal. Technologie de l'information, culture & société*, (116), 2015.

[9] Marion Pech, Helene Sauzeon, Thinhinane Yebda, Jenny Benois-Pineau, and Helene Amieva. Falls detection and prevention systems in home care for older adults : myth or reality ? *JMIR aging*, 4(4) :e29744, 2021.

[10] Maribel Pino, Sébastien Dacunha, Étienne Berger, Anna GOnÇALVES, and Anne-Sophie Rigaud. Intérêt de la robotique sociale et d'assistance auprès des sujets âgés. *Actualités Pharmaceutiques*, 60(611) :36–39, 2021.

[11] Salifu Yusif, Jeffrey Soar, and Abdul Hafeez-Baig. Older people, assistive technologies, and the barriers to adoption : A systematic review. *International journal of medical informatics*, 94 :112–116, 2016.

[12] Tenzin Wangmo, Mirjam Lipps, Reto W Kressig, and Marcello Ienca. Ethical concerns with the use of intelligent assistive technology : findings from a qualitative study with professional stakeholders. *BMC Medical Ethics*, 20(1) :1–11, 2019.

[13] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu. Human action recognition from various data modalities : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(03) :3200–3225, 2023.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[15] W Liu, D Anguelov, D Erhan, C Szegedy, S Reed, and CY Fu. Parsenet : Looking wider to see better. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1340–1348, 2016.

[16] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12) :2481–2495, 2017.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[21] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous convolution for semantic image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4) :10–1109, 2018.

[22] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4) :834–848, 2017.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[24] P Kalaivani and DS Vimala. Human action recognition using background subtraction method. *International Research Journal of Engineering and Technology (IRJET)*, 2(3) :1032–1035, 2015.

[25] Bohyung Han and Larry S Davis. Density-based multifeature background subtraction with support vector machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5) :1017–1023, 2011.

[26] Fida El Baf, Thierry Bouwmans, and Bertrand Vachon. A fuzzy approach for background subtraction. In *2008 15th IEEE International Conference on Image Processing*, pages 2648–2651. IEEE, 2008.

[27] Diana Farcas, Cristina Marghes, and Thierry Bouwmans. Background subtraction via incremental maximum margin criterion : a discriminative subspace approach. *Machine Vision and Applications*, 23(6) :1083–1101, 2012.

[28] Wenbo Zheng, Kunfeng Wang, and Fei-Yue Wang. A novel background subtraction algorithm based on parallel vision and bayesian gans. *Neurocomputing*, 394 :178–200, 2020.

[29] Olivier Barnich and Marc Van Droogenbroeck. Vibe : A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6) :1709–1724, 2010.

[30] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981.

[31] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3) :185–203, 1981.

[32] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[33] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow : Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013.

[34] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet : Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[35] Zachary Teed and Jia Deng. Raft : Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[36] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer : A transformer architecture for optical flow. *ECCV*, 2022.

[37] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow : Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022.

[38] Thomas Brox and Jitendra Malik. Large displacement optical flow : descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3) :500–513, 2010.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[40] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[41] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[43] Hei Law and Jia Deng. Cornernet : Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[44] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 850–859, 2019.

[45] Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. Human detection from images and videos : A survey. *Pattern Recognition*, 51 :148–175, 2016.

[46] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose : realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1) :172–186, 2019.

[47] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120 : A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10) :2684–2701, 2019.

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[49] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(5) :914–927, 2013.

[50] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

[51] Alexander Toshev and Christian Szegedy. Deeppose : Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[52] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe : Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[53] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 597–600, 2017.

[54] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6769–6778, 2017.

[55] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE transactions on pattern analysis and machine intelligence*, 40(7) :1555–1569, 2017.

[56] Liang Lin, Yiming Gao, Ke Gong, Meng Wang, and Xiaodan Liang. Graphonomy : Universal image parsing via graph reasoning and transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5) :2504–2518, 2020.

[57] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3185–3193, 2016.

[58] Zhong Li, Xin Chen, Wangyiteng Zhou, Yingliang Zhang, and Jingyi Yu. Pose2body : Pose-guided human parts segmentation. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 640–645. IEEE, 2019.

[59] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4) :677–691, 2017.

[60] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1) :221–231, 2012.

[61] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 1, 06 2014.

[62] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit : A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[63] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[64] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[65] Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matas. Systematic evaluation of convolution neural network advances on the imagenet. *Computer vision and image understanding*, 161 :11–19, 2017.

[66] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[67] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net : Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104. Springer, 2020.

[68] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets : Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[69] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. Db-lstm : Densely-connected bi-directional lstm for human action recognition. *Neurocomputing*, 444 :319–331, 2021.

[70] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[71] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets : New architecture and transfer learning for video classification, 2017.

[72] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[73] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco : Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.

[74] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[75] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[76] J. Carreira and A. Zisserman. Quo vadis, action recognition ? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[77] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d : Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020.

[78] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network : A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 802–810, Cambridge, MA, USA, 2015. MIT Press.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[80] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[81] Xuefan Zha, Wentao Zhu, Lv Xun, Sen Yang, and Ji Liu. Shifted chunk transformer for spatio-temporal representational learning. *Advances in Neural Information Processing Systems*, 34 :11384–11396, 2021.

[82] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[83] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18. Springer, 2022.

[84] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip : A new paradigm for video action recognition, 2021.

[85] Yinxiao Li, Zhichao Lu, Xuehan Xiong, and Jonathan Huang. Perf-net : Pose empowered rgb-flow net. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 513–522, 2022.

[86] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101 : A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012.

[87] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021.

[88] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *International Conference on Learning Representations*, 2023.

[89] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[90] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb : a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[91] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet : A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[92] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.

[93] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[94] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m : A large-scale video classification benchmark, 2016.

[95] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m : Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[96] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes : Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[97] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs : Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019.

[98] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[99] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[100] Huiwen Guo, Xinyu Wu, Nannan Li, Ruiqing Fu, Guoyuan Liang, and Wei Feng. Anomaly detection and localization in crowded scenes using short-term trajectories. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 245–249, 2013.

[101] Raul Igual, Carlos Medrano, and Inmaculada Plaza. Challenges, issues and trends in fall detection systems. *Biomedical engineering online*, 12(1) :66, 2013.

[102] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine*, 117(3) :489–501, 2014.

[103] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE journal of biomedical and health informatics*, 18(6) :1915–1922, 2014.

[104] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. Up-fall detection dataset : A multimodal approach. *Sensors*, 19(9), 2019.

[105] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Definition and performance evaluation of a robust svm based fall detection solution. In *2012 eighth international conference on signal image technology and internet based systems*, pages 218–224. IEEE, 2012.

[106] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. *DIRO-Université de Montréal, Tech. Rep*, 1350 :24, 2010.

[107] Greet Baldewijns, Glen Debard, Gert Mertes, Bart Vanrumste, and Tom Croonenborghs. Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. *Healthcare technology letters*, 3(1) :6–11, 2016.

[108] José Camilo Eraso Guerrero, Elena Muñoz España, Mariela Muñoz Añasco, and Jesús Emilio Pinto Lopera. Dataset for human fall recognition in an uncontrolled environment. *Data in Brief*, 45 :108610, 2022.

[109] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.